引文格式：熊文辉,田振昂,赵东保. 基于语义特征的多源兴趣点机器学习数据匹配方法[J]. 地理空间信息, 2026, 24(7): 96-100. doi:10. 3969/j. issn. 1672-4623.2026.07.018

基于语义特征的多源兴趣点机器学习数据匹配方法

熊文辉[1,2]，田振昂[1,2]，赵东保[1,2*]
（1. 华北水利水电大学 测绘与地理信息学院，河南 郑州 450046；2. 智慧地球重点实验室，北京 100094）

摘　要：为准确识别多源异构数据中的同名兴趣点地理实体，以多平台导航电子地图中的兴趣点数据为研究对象，提出了中心词相似度、词频相似度和名称干扰度3种兴趣点语义特征相似度，并结合4种常用语义特征相似度构建机器学习模型，实现了一种两阶段多源兴趣点数据匹配方法。以四川省成都市武侯区的高德地图、腾讯地图、天地图3个平台约40万个兴趣点数据为例开展数据匹配实验。结果表明，该方法可准确且快速地实现多源兴趣点间的匹配，构建的随机森林匹配分类模型的精确率为91.0%，召回率为95.0%，F_1-score为0.91，有效提升了POI数据匹配的效率与准确性。

关键词：兴趣点；空间数据匹配；相似度计算；随机森林
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2026）07-0096-05

Machine Learning Data Matching Method for Multi-source Points of Interest Based on Semantic Features

XIONG Wenhui[1,2], TIAN Zhen'ang[1,2], ZHAO Dongbao[1,2]
(1. College of Surveying and Geo-informatics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;
2. Key Laboratory of Smart Earth, Beijing 100094, China)

Abstract: In order to accurately identify the geographical entities of the same name of points of interest (POI) in multi-source heterogeneous data, taking the POI data in multi-platform navigation electronic maps as the research object, we proposed three semantic feature similarity of POI, namely, central word similarity, word frequency similarity and name interference degree, combined four commonly used semantic feature similarity to construct a machine learning model, and implemented a two-stage multi-source POI data matching method. We used a total of approximately 400 000 POI data from three platforms in Wuhou District, Chengdu City, Sichuan Province, China including Gaode Map, Tencent Map, and MapWorld to conduct a data matching experiment. The results show that the method can accurately and quickly match multi-sources POIs. The accuracy rate of random forest matching classification model constructed by the method is 91.0%, recall rate is 95.0%, and F_1-score is 0.91. This method can effectively improve the efficiency and accuracy of POI data matching.

Key words: POI, spatial data matching, similarity computation, random forest

地理空间信息能对地球表层的各类地理实体及其空间关联关系进行刻画，而兴趣点（POI）是其不可或缺的核心构成要素之一，被广泛应用于地图导航平台。不同来源的POI数据的位置和属性信息并不完全相同，因此如何匹配不同来源的POI数据成为一个重要的研究问题[1]。首先通过构建差异化的测算方法对任意一对待匹配 POI 点之间的相似性进行量化评估；再根据相似性度量结果，将这些点对划分为同名匹配点与非同名匹配点[2-4]。目前POI匹配通常综合利用欧氏距离、名称属性、中文地址、类别等空间和属性信息开展相似性度量。POI间的欧氏距离可构成距离相似度[5]；名称属性通过编辑距离[6]和Jaccard距离[7]等计算其相似度。仅依赖欧氏距离和名称属性的匹配方法的准确率仍存在可优化的空间[8]。已有学者[9-10]将POI数据的地址属性信息纳入匹配流程，同步构建对应的相似度测度指标完成匹配判定，以进一步提升匹配准确率。对于POI的中文地址，通常先采用中文分词算法进行分词，再利用编辑距离[11-12]或词袋相似度计算其相似性[13]。不同POI间的类别信息也是重要的匹配特征[14-15]。对于多种相似度的综合度量，部分学者[16-19]采用机器学习或深度学习方法进行POI匹配。尽管各类空间与属性特征相似度设计已在提升POI匹配准确性方面起到了关键作用，但仍存在进一步优化的空间：针对POI的文本属性信息，还可开展更深层次的语义结构分析与自然语言理解，进一步提升匹配准确率；需针对不同类型的POI，设计适配性的匹配

项目来源：智慧地球重点实验室开放基金资助项目（KF2023ZD02-04）；国家自然科学基金资助项目（41971346）。
第一作者简介：熊文辉（1999—），硕士研究生，从事空间数据融合及空间数据研究工作，E-mail：xwh2024@126.com。
通信作者：赵东保（1979—），博士，教授，从事空间数据融合及空间数据研究工作，E-mail：zdongbao@126.com。

方法。鉴于此，本文选取商业导航电子地图领域的高德地图、腾讯地图以及国家地理信息公共服务平台的天地图提供的POI数据作为研究样本，构建了一套融合逻辑规则与机器学习分类模型的多源POI匹配方法。

1　多源POI数据匹配方法框架

本文提出的多源POI匹配方法分为3个实施阶段，具体匹配流程见图1。在初始候选筛选阶段，一方面基于空间邻近特性为目标POI建立缓冲区，以此确定潜在匹配空间范围，筛选出该范围内的候选匹配POI；另一方面通过计算POI名称间的编辑距离[20]实现名称字符串相似度的量化计算，逐一遍历比对待匹配POI，选取与目标POI名称相似度最高的条目进行配对，最终形成POI待匹配点对。

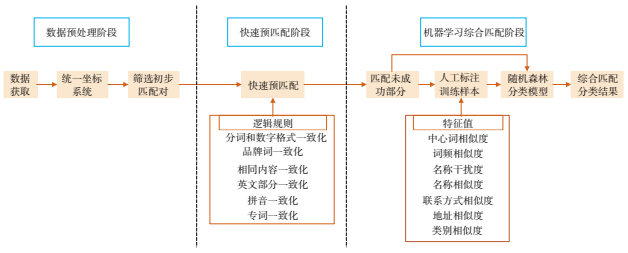


图1　多源POI匹配流程图

2　多源POI数据的两阶段匹配

2.1　基于逻辑规则的快速预匹配

在完成POI待匹配点对的初步筛选后，进行多源POI的快速预匹配，以POI的名称特征差异为依据设计了对应的逻辑规则，并作为匹配约束条件。具体的逻辑规则为：①分词与数字格式一致化，首先对POI待匹配点对执行统一分词操作，借助中文分词工具搭配停用词词典，完成无关停用词的过滤剔除，并将名称中包含的数字统一转换为小写格式，以保障待处理数据的一致性与标准化程度；然后计算该匹配点对双方名称的字符串相似度，若二者完全一致，可直接判定为同名匹配对，若相似度低于阈值，则初步将其归为非同名匹配对。②品牌词一致化，若两组待匹配POI的分词中存在同一知名品牌词汇，可直接判定为同名匹配对；若不存在，则初步判定为非同名匹配对。③相同内容一致化，先剔除匹配点对双方共有的分词内容，若剩余内容为空，则可直接判定为同名匹配对；若仍有剩余内容，则进入下一规则环节继续校验。④英文部分一致化，计算二者英文部分的字符串相似度，若相似度低于阈值，则直接判定为非同名匹配对；若相似度处于中间区间，则需做进一步的区分判断。⑤拼音一致化，首先将剩余的分词结果转换为拼音；再计算拼音内容字符串的相似度，若拼音内容完全一致，则判定为同名匹配对，若存在差异，则进入后续处理环节。⑥专词一致化，针对双方分词结果剔除公共部分后识别出的出现频率为1（即在全库中仅出现过一次）的专有名词，依托这些专有名词之间的相似度完成判定，若二者的相似度极低，则直接判定为非同名匹配对；其余情况则不作即时判定[21]。经过上述逻辑规则的精细化筛选处理，能快速完成不同来源POI数据中同名POI的识别与匹配，有效提升整体匹配流程的效率与准确性。

2.2　基于随机森林的综合匹配

1）多源POI数据语义特征相似度计算。本文采用7种POI匹配特征相似度的计算方法，其中名称相似度、联系方式相似度、地址相似度和类别相似度是基于语义特征的常用相似度。①名称相似度。名称是POI数据匹配中最核心且最具辨识度的属性字段，本文利用编辑距离算法[20]计算名称相似度，通过动态规划算法求得将一个字符串转换为另一个字符串所需的最少编辑次数，从而衡量两个字符串的相似性。②地址相似度。本文采用词袋模型构建词语向量表达，首先对地址属性信息进行分词处理并剔除停用词，再以余弦相似度作为衡量地址相似度的指标。③类别相似度。对不同平台的分类代码编码表进行人工映射对齐，判断二者类别是否存在对应关系，若对应则类别相似度为1，反之则为0。④联系方式相似度。以双方电话信息是否完全互相包含为判断依据，若完全互相包含则取值为1，否则为0。⑤中心词相似度。通过分析POI名称内部的句法依存关系，忽略修饰性附加成分，聚焦核心中心词，以此构建精度更高的相似度计算模型，如“精武合作运动空间”与“精武合作运动空间（华林店）”经处理后，剩余核心部分为“武合作空间”与“武合作空间店”，编辑距离由0.615提升至0.833，消除了文本中修饰成分的干扰，提升了匹配准确率。⑥词频相似度。在实际匹配中发现，POI名称经分词后，低频词在同名匹配判定中往往具有更强的区分能力[22]。首先对词频库中的词频进行归一化处理，再以词频的倒数为各词汇相似度的权重，以此体现不同词频对相似度判断的重要程度，如“极限（富森·美家居2号店建材）”与“精度（富森·美家居2号店建材）”的编辑距离高达0.87，但由于词频较低的“极限”和“精度”不同，可以确定二者为非同名POI。

$$S_F=\sum_{i}^{m}\sum_{j}^{n}\frac{\left(x_i+y_j\right)S}{2m\times n} \tag{1}$$

式中，x_i 为目标POI名称分词结果第 i 个词的词频；y_j 为待匹配POI名称分词结果第 j 个词的词频；S 为两个词之间的字符串相似度；m、n 分别为目标、待匹配POI名称中文分词数量。

⑦名称干扰度基于POI之间的空间距离计算，与POI数据的空间属性直接相关。在候选匹配点筛选阶段，对名称相似度进行排序，若存在多个名称相似度较接近的候选点，名称干扰度的计算公式为：

$$S_I = 1 - \frac{S_{\max}}{\sum_{i=1}^{n} S_i} \tag{2}$$

式中，$S_{\max}$ 为范围内POI名称相似度的最大值；$\sum_{i=1}^{n} S_i$ 为 n 个干扰点的名称相似度之和，干扰点可根据相似度差值阈值设置，本文将与名称相似度最大值相差10%以内的作为干扰点。

2）训练数据集构建。随机森林模型的构建与性能验证均依赖于人工标注数据集。首先从经过快速预匹配处理的数据集中每隔约100个数据随机抽取一个数据样本构建数据集；再人工审查这些数据，确定它们是否构成同名匹配对，并根据审查结果进行相应标注。

3）数据集特征筛选。由于并非所有特征都对模型精度具有正面影响[23]，因此建模前需进行特征筛选。同时，特征变量间的相关性过高可能会导致模型估计不准确，降低预测能力，模型构建更加繁琐，因此需对这些特征进行相关性分析。由特征间相关性的热力图（图2）可知，本文采用的7个特征变量之间无高度相关的特征。按照特征重要性从大到小顺序，每次增加一个特征反复建模，以精确率作为模型的评价指标，若精确率开始下降，则将新增特征视为噪声特征。RFE特征筛选结果见图3，可以看出，不断加入新特征后，评价分数无下降趋势，因此无需剔除任何特征。

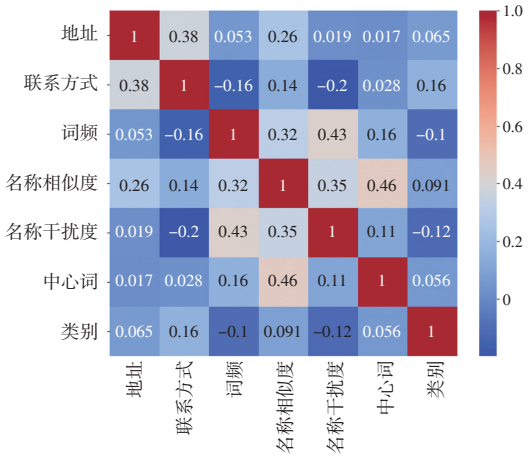


图2 特征相关性矩阵热力图

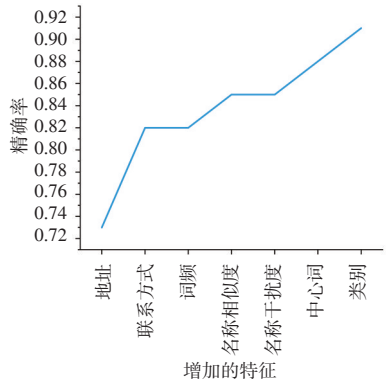


图3 RFE特征筛选

4）随机森林分类模型的建立。随机森林的核心思路是通过融合多棵决策树的预测结果来优化整体模型性能。超参数调优是模型构建过程中的关键环节，本文通过十折交叉验证方法确定最优超参数配置，即决策树数量为1 000，单棵决策树的最大深度为14，节点分裂可选用的最大特征数为9，叶子节点包含的最小样本数为4，节点分裂所需的最小样本数为10。在原始数据集的基础上，计算得到7种特征相似度，且经过特征筛选环节验证，这7种特征均对模型精度提升具有正向贡献。多源POI数据快速预匹配完成后，针对未成功匹配的数据，通过人工标注的方式划分训练集与测试集；再采用有放回抽样的方式构建基决策树；最终训练得到随机森林分类模型。

3 实例验证分析

为验证本文方法的可靠性以及是否可应用于实际数据生成与更新场景，本文选取了四川省成都市武侯区的两组实验数据：天地图POI数据（共98 872条）与腾讯地图POI数据（共165 521条），高德地图POI数据（共150 277条）与腾讯地图POI要素数据（共165 521条）开展多源POI数据匹配实验。结合快速预匹配的处理结果，人工抽取未完成预匹配的剩余数据，以此划分训练集与测试集完成模型构建。本文采用Python、C#语言进行算法编写，结合ArcGIS Engine二次开发接口完成模型构建。实验采用的硬件环境为：处理器Intel(R)Core(TM)i5、主频2.50 GHz、物理内存16 GB。

3.1 快速预匹配结果分析

数据预处理完成后，基于逻辑规则的快速预匹配实验结果见表1，其中预匹配率是指经逻辑规则成功判定为同名POI或非同名POI的样本数量占全部待处理数据总数的比值，可以看出，第二组数据的预匹配率比第一组高约10个百分点，该差异主要来源于不同地图服务平台自身的独特性以及各平台间的数据属性差

表1　快速预匹配结果

数据内容	同名POI/条	非同名POI/条	新增数据/条	未匹配成功/条	预匹配率/%
天地图：腾讯地图（98 872：165 521）	31 038	10 397	8 267	49 170	50.2
高德地图：腾讯地图（150 277：165 521）	81 275	9 814	435	58 753	60.9

异；两组数据中均有至少50%的POI能通过简单的逻辑规则快速完成匹配对应关系的识别，说明简单规则匹配方法能有效提升整体POI匹配流程的效率与精确度。

3.2　综合匹配结果分析

1）各种机器学习算法匹配结果的对比分析。为验证随机森林模型的分类准确性，本文将其与目前常用的机器学习模型进行对比实验。所有对比模型采用的训练集为人工标注构建的包含3 000条样本的标注数据集，其中标记为1的同名POI对1 506条，标记为0的非同名POI对1 494条。模型训练完成后，采用另一组人工标注的测试集进行精度验证，该测试集包含938条非同名POI对与1 062条同名POI对。实验选取精确率、召回率和F_1–score作为模型的性能评估指标，各模型的精度统计结果见表2，可以看出，多层感知器模型虽然取得了最高的精确率，但召回率不足0.8；而随机森林模型的召回率与F_1–score均为所有模型中最优，且该模型的3项评价指标全部处于0.9以上的水平。通过对比分析可知，在所有机器学习模型中，随机森林模型最适配本文提出的多源POI综合匹配方法；该模型与逻辑规则的快速预匹配环节相结合，不仅提升了匹配判断的处理效率，还进一步增强了整体匹配流程的准确性，适用于当前绝大多数电子地图平台的多源POI数据匹配任务。

表2　机器学习分类模型精度统计

模型	精确率	召回率	F_1–score
随机森林	0.910	0.950	0.909
支持向量机	0.928	0.825	0.873
多层感知器	0.931	0.792	0.856
逻辑回归	0.845	0.899	0.871
朴素贝叶斯	0.715	0.904	0.798

2）与现有POI匹配算法的对比分析。在POI匹配算法研究中采用的POI属性特征不尽相同，本文采用相同数据集对比代表文献，按照各方法中所用语义特征相似度建立随机森林模型，精度结果见表3。由于本文在筛选待匹配对时已采用POI的位置属性，因此采用除位置外的其他属性特征进行建模对比。结果表明，本文采用7种特征相似度构建的模型在3个评价指标中均表现最优，证明在使用相同数据和匹配方法的条件下，本文引入了更多更有效的特征相似度。

表3　各种方法所用相似度模型精度对比

方法	语义特征相似度	精确率	召回率	F_1–score
参考文献[17]	名称、类别	0.610	0.579	0.617
参考文献[11]	名称、地址	0.770	0.719	0.746
参考文献[14]	名称、地址、类别	0.820	0.757	0.763
参考文献[5]	名称、类别、分类、联系电话	0.870	0.881	0.873
本文方法	中心词、词频、名称干扰度、名称相似度、地址、类别、联系方式	0.910	0.950	0.909

4　结　论

本文以实现多源POI空间数据的自动匹配与融合为主要目标，从自然语言理解角度出发，着重分析了POI的名称信息，提出了中心词相似度、词频相似度和名称干扰度3种POI语义特征相似度，并与名称相似度、类别相似度、地址相似度和联系方式相似度等特征相似度相结合，实现了包含逻辑规则快速匹配和机器学习综合匹配两个阶段的多源POI数据匹配方法。在大规模实测高德地图、腾讯地图和天地图等数据集上的实验结果表明，逻辑规则快速匹配可准确判断50%以上的POI对，而剩余部分可采用机器学习综合匹配有效解决。对比其他常用的机器学习分类模型，本文构建的随机森林POI数据匹配模型的综合性能表现最优。与运用多种POI特征相似度的现有文献相比，本文提出的7种语义特征相似度涵盖了更多的POI语义特征，构建的多源POI随机森林模型精度更高。本文方法可有效提升多源POI匹配的准确性和效率，为多源空间数据匹配应用实践提供有力支持。

参考文献

[1] 薛冰,赵冰玉,李京忠. 地理大数据中POI数据质量的评估与提升方法[J]. 地理学报,2023,78(5):1290-1303

[2] 张巍,高新院,李瑞姗. 空间位置信息的多源POI数据融合[J]. 中国海洋大学学报(自然科学版),2014,44(7):111-116

[3] 高新院. 基于空间位置信息的多源POI数据融合问题的研究[D]. 青岛:中国海洋大学,2013

[4] Xavier E M A, Ariza-López F J, Urea-Cámara M A. A Survey of Measures and Methods for Matching Geospatial Vector Datasets[J]. ACM Computing Surveys, 2016, 49(2): 1-34

[5] 戴鹏. 基于实体匹配的互联网餐馆POI数据整合研究[J]. 测绘与空间地理信息,2024,47(3):54-57

[6] 张翼然,陶迎春,蔡彩,等. 基于位置和属性特征信息的多源数据融合[J]. 北京测绘,2018,32(9):1020-1023

[7] 孙晓霞. 融合语义和位置信息的兴趣点实体匹配方法[J]. 地理空间信息,2023,21(5):22-24

[8] Li C,Liu L,Dai Z,et al. Different Sourcing Point of Interest Matching Method Considering Multiple Constraints[J]. ISPRS International Journal of Geo-information,2020,9(4):214

[9] 姜松言,宋国民,贾奋励,等. 结合地名知识和位置信息的中文地名匹配方法[J]. 测绘通报,2022(5):120-125

[10] 彭雨龙,胡顺石,吴涛. 多策略中文地址匹配方法[J]. 测绘通报,2022(2):145-148

[11] 李鹏飞. 基于互联网地理信息的公共服务平台POI数据增量更新研究[D]. 兰州:兰州交通大学,2020

[12] 张寅宝,张威巍. 基于相似度计算的室内兴趣点匹配方法[J]. 测绘与空间地理信息,2017,40(2):5-8

[13] 邵蕾. 网络POI数据增量更新技术研究[D]. 兰州:兰州交通大学,2016

[14] 罗国玮,叶嘉媛,王金凤. 基于多特征相似性的多源POI匹配方法[J]. 测绘通报,2022(4):96-100

[15] Zhao J,Niu X,Cui Y,et al. POI Point Entity Matching and Fusion Based on Multi-similarity Calculation[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,2022,10:87-92

[16] Zhou Y, Wang M J, Zhang C, et al. A Points of Interest Matching Method Using a Multivariate Weighting Function with Gradient Descent Optimization [J]. Transactions in GIS,2020,25(1):359-381

[17] 田晨睿. 基于众源POI匹配与融合的实体库构建方法研究[D]. 武汉:武汉大学,2020

[18] 龙军. 基于角色标注的中文POI名称匹配的研究及原型系统实现[D]. 重庆:西南大学,2008

[19] Li P,Liu J,Luo A,et al. Deep Learning Method for Chinese Multi-source Point of Interest Matching[J]. Computers, Environment and Urban Systems,2022,96:101821

[20] 张胜楠. 基于编辑距离的字符串相似度算法研究[J]. 现代计算机,2023,29(14):23-26

[21] 田振昂. 多源兴趣点数据匹配与融合方法研究[D]. 郑州:华北水利水电大学,2024

[22] Hong X P,Huang W,Liu H T. The Structural Complexity of Chinese Words and Its Relationship with Word Frequency[J]. Journal of Quantitative Linguistics,2023,30(3/4):231-256

[23] 刘傲,赵东保,魏义长,等. 顾及时空特征的参考作物蒸散量集成学习估算[J]. 排灌机械工程学报,2024,42(2):179-186

(上接第21页)

[2] 张祖勋,陶鹏杰. 谈大数据时代的“云控制”摄影测量[J]. 测绘学报,2017,46(10):1238-1248

[3] Tao P,Lu L,Zhang Y,et al. On-orbit Geometric Calibration of the Panchromatic/Multispectral Camera of the ZY-1 02C Satellite Based on Public Geographic Data[J]. Photogrammetric Engineering and Remote Sensing, 2014, 80(6): 505-517

[4] 辛国栋,安文,黄令勇. 天绘一号卫星几何定标场建设与应用[J]. 测绘地理信息,2018,43(6):52-54

[5] Zhang Y, Zheng M, Xiong J, et al. On-orbit Geometric Calibration of ZY-3 Three-line Array Imagery with Multistrip Data Sets[J]. IEEE Transactions on Geoscience and Remote Sensing,2013,52(1):224-234

[6] Yang G B, Pi Y, Li X, et al. Integrated Geometric Self-calibration of Stereo Cameras Onboard the ZY-3 Satellite[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2020,162:173-183

[7] Zhang Y,Wang T,Zheng T,et al. On-orbit Geometric Calibration and Performance Validation of the GF-14 Stereo Mapping Satellite [J]. Remote Sensing,2023,15(17):4256

[8] 卢学良,王建荣,杨秀策,等. 高分十四号卫星双线阵相机高精度在轨几何标定[J]. 测绘学报,2023,52(1):15-21

[9] Cao H, Tao P, Li H, et al. Bundle Adjustment of Satellite Images Based on an Equivalent Geometric Sensor Model with Digital Elevation Model[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2019,156:169-183

[10] Hou Z, Liu Y, Zhang L, et al. 2OC: A General Automated Orientation and Orthorectification Method for Corona KH-4B Panoramic Imagery [J]. Remote Sensing,2023,15(21):5116

[11] 孙钰珊,张力,许彪,等. 资源三号卫星影像无控制区域网平差[J]. 遥感学报,2019,23(2):205-214

[12] Zhao X, Zhou Q, Dong J, et al. Digital Elevation Model-assisted Aerial Triangulation Method on an Unmanned Aerial Vehicle Sweeping Camera System[J]. The Photogrammetric Record,2022,37(178):208-227

[13] 刘学人. DSM控制的高分辨率卫星影像几何精度校正[J]. 地理空间信息,2023,21(6):82-85

[14] 孟小亮,胡予萱,王明霞,等. 多源空间数据辅助的云台控制摄像机几何定位方法[J]. 地理空间信息,2024,22(2):1-7

[15] 甘文祥,张远谊,李欣园. 一种航空影像建筑物检测的轻量化CNN建模方法[J]. 地理空间信息,2023,21(6):24-27

[16] 刘冬枝. 基于高分七号卫星影像生产建筑白模[J]. 地理空间信息,2023,21(11):104-106