引文格式:黄颖菁,张帆,李勇,等.城市影像的智能计算表征[J].武汉大学学报(信息科学版),2026,51(1):22-31.DOI:10.13203/j.whugis20240472

Citation:HUANG Yingjing,ZHANG Fan,LI Yong,et al.Intelligent Computational Representation of Urban Imagery[J].Geomatics and Information Science of Wuhan University,2026,51(1):22-31.DOI:10.13203/j.whugis20240472

城市影像的智能计算表征

黄颖菁[1] 张 帆[1] 李 勇[1,2] 邬 伦[1] 刘 瑜[1]

1 北京大学地球与空间科学学院遥感与地理信息系统研究所,北京,100871

2 香港科技大学土木及环境工程学系,香港,999077

摘 要:城市影像能够详尽刻画城市物理环境,支持从全球到微观层面的多尺度分析。基于高效的特征工程方法,从庞大且复杂的城市影像像素数据中提取高层次语义特征,用于模式识别和决策支持,一直是城市研究的重要方向。相较于传统的语义要素表征,表示学习支持下的计算表征方法能够从城市影像中学习高维深度特征,这些特征不仅提炼了更丰富的城市语义与结构信息,还促进了多模态数据的融合和更精准、更鲁棒的城市模型的构建。特别地,基于自监督学习的智能计算表征,能够在无需标注数据的情况下自动编码与城市任务相关的关键信息,进一步提升了城市影像分析的自动化水平。通过探讨城市影像智能计算表征的特点、发展历程及其可解释性,发现该方法有望显著提升城市智能化分析能力,从而为城市研究、规划、管理和可持续发展提供更精准的决策支持。

关键词:遥感影像;街景影像;计算表征;表示学习;深度特征

中图分类号:P208;P237 **文献标识码**:A **收稿日期**:2025-03-22

DOI:10.13203/j.whugis20240472 **文章编号**:1671-8860(2026)01-0022-10

Intelligent Computational Representation of Urban Imagery

HUANG Yingjing[1] *ZHANG Fan*[1] *LI Yong*[1,2] *WU Lun*[1] *LIU Yu*[1]

1 Insititute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China

2 Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China

Abstract: **Objectives**: Urban imagery, including high-resolution remote sensing imagery and street view imagery, provides a detailed representation of the physical environment of cities. It enables multi-scale analysis ranging from global perspectives to microscopic details. Extracting high-level semantic features from the vast and complex pixel data of urban imagery for applications in pattern recognition and decision-making support has long been a critical focus in urban studies. **Methods**: Traditional approaches, which progress from handcrafted shallow features like color histograms to semantic element representations, are often insufficient. These methods are limited by expert knowledge and fail to capture deep attributes or complex spatial relationships inherent in urban environments. In response, computational representation methods supported by representation learning can extract high-dimensional deep features from urban imagery. These features not only capture richer urban semantic and structural information but also facilitate more effective multimodal data integration by reducing semantic conflicts, thereby enabling the development of more accurate and robust urban models. Unlike handcrafted features, these computational representations are inherently data-driven and learn directly from the intrinsic structure and patterns within the raw data itself. Notably, intelligent computational representations based on self-supervised learning (SSL) stand out as a critical solution to the data-labeling bottleneck. Given the massive scale of urban imagery and the prohibitively high cost

基金项目:国家自然科学基金(42371468)。

第一作者:黄颖菁,博士,主要从事城市视觉智能研究。huangyingjing@stu.pku.edu.cn

通信作者:张帆,博士,助理教授。fanzhanggis@pku.edu.cn

of manual annotation, which often suffers from inconsistent quality, SSL advances the automation of urban imagery analysis by learning powerful representations from the data's own intrinsic properties, requiring no labeled datasets. We believe that the most effective SSL representations, moving beyond generic augmentations which are ill-suited for complex urban tasks, must be task-centric. This is achieved by autonomously encoding key information by designing specialized pretext tasks. These tasks utilize the unique attributes and structure of urban imagery to systematically learn distinct representations, such as isolating time-invariant features by ignoring dynamic elements for location recognition, versus capturing ambient socio-economic and cultural semantics for regional-level prediction tasks. **Results:** We explore the characteristics, evolutionary trajectory, and interpretability of these intelligent computational representations. Given that deep features often lack explicit semantic meaning, we emphasize interpretability analysis as essential for validating the effectiveness of the encoding process and understanding what information has been encoded. We discuss a two-fold approach: (1) Revealing the semantic content within features using methods like SemAxis or visualizing clustering and separation patterns in low-dimensional space via T-SNE and PacMap. (2) Discovering the model's focus using visualization techniques like Grad-CAM and Attention Maps to understand which image regions are critical for a decision. **Conclusions:** Consequently, these advancements offer more precise support for urban research, planning, and sustainable development. We also highlight future opportunities and challenges, including: (1) Developing robust multimodal alignment to bridge the gap between areal remote sensing features and point-based street view representations. (2) Integrating visual features with large language models (LLM) to summarize complex urban concepts, facts, and opinions. (3) Addressing the fundamental "black box" nature of deep features to advance AI for Science.

Key words: remote sensing imagery; street view imagery; computational representation; representation learning; deep feature

城市影像是一种描述城市物理环境的影像数据,包括行人视角的街景自然影像和自上而下视角的卫星遥感影像[1-2]。这些影像具有高时空分辨率,且能够深入刻画从城市整体格局到建筑街道细节的物理环境要素和空间结构。长期以来,这些数据为城市感知与分析应用提供了坚实而持久的技术与信息支撑,进而为城市规划和治理提供决策支持和政策指导[3-4]。随着感知技术的飞速发展,影像数据的获取与处理效率显著提升,不仅增加了利用遥感影像实现全球、城市群等区域乃至城市尺度多层次研究的可能性,而且也使得借助街景影像在微观层面近乎沉浸式地观察与评估城市环境的动态演化成为现实[5]。

对于规模持续增长且类型日益多样的城市影像数据,高效而精准地挖掘其中蕴含的城市复杂模式与规律,一直是城市研究领域的一个重要方向。由于原始像素数据结构复杂且信息冗余,未经处理的数据往往难以提供清晰的认知。因此,通过特征工程对影像进行统一且高度抽象的描述显得尤为必要。通过这一过程,零散的像素信息被转化为计算机可理解和处理的密集特征向量,为后续的模式识别、规律挖掘和决策支持提供了关键基础。

在早期阶段,城市影像的特征工程主要依赖经验驱动的手工设计浅层特征,包括颜色直方图、纹理描述子和形状特征等[6-7]。这些手工特征虽有效提升了分析的可行性与精度,但其表现受到专家经验与领域知识的限制,难以满足当下多样化、动态化的研究需求。近年来计算机视觉的快速发展,使研究者能够在更高层次的语义空间中对城市影像进行特征提取[8-10]。语义分割、目标检测等技术可直接识别并提取建筑、行人、道路、树木等具体城市要素,这些更富语义的特征已被证明在预测社会经济指标方面具有较大潜力[11]。

尽管语义要素表征在一定程度上缓解了传统特征提取的局限,但仍难以全面刻画城市环境的复杂性。一方面,现有的语义要素表征无法充分表达城市影像中的丰富信息。具体而言,单一的要素类别难以反映要素的深层属性与动态特征,例如植被类型可暗示区域的生态健康状况,人际互动方式则体现社会功能,甚至车辆的品牌与年份可折射社会经济结构[12]。然而,要全面获取此类细粒度的语义信息,往往需要大量专业标注与人工成本,这在实践中几乎无法实现[13-14]。此外,相似的要素构成与占比并不必然意味着相

同的空间关系与城市语义,例如整齐排列的行道树与零散分布的树林,虽然在绿视指数上可能相似,却传递出截然不同的环境含义。另一方面,现有的语义要素表征也难以与其他模态的表征进行有效对齐和融合。要素语义特征往往需要依赖于特定的标签或者具体的语义解释,不同的语义标签(如"车""行人""建筑"等)在不同模态(如图像、文本、激光雷达等)中的表现形式可能大相径庭,且同一语义在不同模态中的特征可能截然不同。因此,如何更加充分地刻画城市环境的复杂性,依然是当前城市影像特征工程亟待突破的难题。

近年来,计算表征的兴起为此提供了新契机。计算表征是指通过表示学习等计算方法,从原始数据中提取特征向量。这些特征向量不仅能够有效捕捉数据的关键属性和内在结构,还具备高度的可计算性,能够支持后续的计算任务,如数据分析、模式识别和复杂结构建模等。通过计算表征,数据中的复杂关系和潜在模式得以显式表达,从而为更高级的计算分析提供基础[15]。POI2Vec、Place2vec等方法[16-17]已成功应用于地理空间数据的计算表征,这些方法不仅能提取高级语义信息,提升地理建模的精度与鲁棒性,还能支持多模态数据的融合,减少了不同模态之间的语义冲突,从而实现在隐空间内的信息对齐,进一步推动地理空间人工智能(artificial intelligence, AI)的发展[18-20]。作为地理空间数据的重要组成部分,城市影像的特征工程同样得益于上述技术的发展,为城市研究提供了更灵活、有深度的分析工具与方法。

在此背景下,自监督学习(self-supervised learning, SSL)等新兴方法为城市影像的智能计算表征提供了创新的解决方案。通过SSL,模型能够充分利用数据自身的结构进行学习,从而在无需大量标注数据的情况下生成智能计算表征,并自主挖掘城市影像中的潜在特征和规律。因此,探索适用于城市影像的智能计算表征方法,不仅能够突破传统方法在数据依赖性和计算效率上的局限,还将为城市规划、交通管理、环境监测等领域的深入研究和广泛应用注入新的活力,成为推动城市研究及相关技术发展的关键方向。

在此基础上,为了有效编码城市影像中与城市任务相关的图像信息,通常需要构建专门的标注数据集以支持模型训练[21-22]。以场景识别任务为例,研究者常常需要构建大量的图像-位置对,以便训练模型聚焦于不同场景中的关键信息,如地标建筑、街道布局等[23]。这些数据集的构建虽然至关重要,但也面临诸多挑战:(1)手工标注过程通常是劳动密集型的,尤其是在城市影像的复杂背景下,标注工作可能涉及大量的地理和环境知识。(2)由于城市影像的多样性和复杂性,全面且精确的标注往往难以实现,导致标注质量不一致,进而影响模型的训练效果和泛化能力。

为应对这些挑战,研究者们逐渐转向SSL等新兴方法,旨在自动化地提取与城市任务相关的信息,从而实现城市影像的智能计算表征。通过这些方法,模型能够在无需大量标注数据的情况下自主发现城市影像中的潜在特征和规律,进而生成更加精确的任务表征。这种技术的应用不仅能够有效解决标注数据稀缺的问题,还能够提高模型在处理多样化城市任务时的适应性。因此,探索适用于城市影像的智能计算表征方法,将成为推动城市研究和相关应用发展的关键方向。

1 城市影像的特点

城市影像涵盖了遥感影像、街景影像、社交媒体影像以及视频监控影像等多种形式。这些影像提供了从宏观到微观的多维度观察手段,使研究者能够深入探索城市空间结构、社会经济活动、建筑美学及其随时间变化的动态过程。利用这些丰富的影像数据,城市研究者和规划者可以更全面地理解城市环境的复杂性,从而为规划、治理和可持续发展提供有力支持。在众多城市影像中,遥感影像和街景影像由于其高时空分辨率、广泛的覆盖范围以及较为便捷的获取方式,已成为城市影像研究中最常用的数据源,并在城市感知中发挥着重要作用。因此,本文以这两类影像为例,探讨城市影像的独特性及其与物体图像的区别。

遥感影像通过航空器或卫星搭载的传感设备从高空获取地表信息,已成为城市研究中不可或缺的重要数据来源。首先,遥感影像的地理覆盖范围极其广泛,全球多个国家和地区通过商业卫星、政府卫星(如陆地卫星(Landsat)系列和哨兵(Sentinel)系列)以及私人卫星(如Planet Labs和Maxar)获取连续且高频次的影像数据。近年来,亚米级分辨率的遥感影像数据逐渐普及,使研究人员能够清晰捕捉地表特征,如城市扩张、道路网络和植被变化[24-25]。然后,遥感影像具备

多光谱和高光谱成像能力,能够捕捉人眼不可见的电磁波段信息,如近红外、短波红外、热红外波段等。这些光谱数据已被广泛用于计算植被指数、监测水体质量(如悬浮物浓度、富营养化程度)以及进行土地覆盖分类[26-27]。最后,遥感影像数据的时间连续性在动态监测城市变化中有着重要优势,许多卫星平台(如 Sentinel-2 和 Landsat 8)能够以 5～16 d 的时间间隔重复采集同一区域的影像,使得长时间序列分析成为可能[28]。

街景影像主要通过谷歌地图、百度地图、腾讯地图等服务商的街景车沿城市路网进行拍摄[29],也包括众包街景影像平台用户上传的标准化图像[30]。首先,街景影像的覆盖范围极为广泛。根据文献[31]的研究,全球大部分人口聚集区已被街景影像覆盖。截至 2025 年 11 月,腾讯和百度的街景服务已涵盖中国 652 个城市,谷歌街景影像覆盖了七大洲的大部分城市,众包街景影像平台则存储了用户上传的超过 5 亿张街景图像。其次,街景影像的采样密度极高,能够无缝连接城市路网的各个部分,形成完整的视觉记录,这种高密度覆盖使研究人员能够在精细的空间尺度上分析城市的物理环境[32]。然后,街景影像具有极高的分辨率,以谷歌街景为例,其最高分辨率可达 16 384×8 192 像素,极大地提升了研究人员在提取和分析城市视觉元素时的精度。这种高分辨率使得建筑立面识别、商店招牌分析、街道家具统计等微观层面的研究成为可能[33]。最后,街景影像的获取方式便捷。研究者可以通过应用程序接口调用数据,并根据具体需求定制影像的拍摄位置、时间、视角及类型,从而满足多样化的应用需求。例如,研究者可以根据特定的研究目的,指定特定街区、不同时段或特定视角(如全景、水平视角)进行影像下载。

如图 1 所示,遥感影像和街景影像分别从鸟瞰和行人视角捕捉城市环境的信息,均是表征城市场景的语义。这些城市影像均是由多种要素组成,如建筑、道路、植被、车辆、行人等[34-35],这些要素的空间分布和组合方式能够反映城市的功能、结构和社会经济特征,从而支持场景识别、情感感知和社会经济指标预测等任务[23]。例如,在场景识别任务中,重点往往是捕捉城市影像中静态要素之间的关系与风格,如遥感影像中的建筑屋顶风格和路网结构,或街景影像中的街道布局等。而大多数经典计算机视觉的表征模型主要面向物体图像设计,通常聚焦于单一或少量主体(如特定物体或简单场景),并侧重物体识别、分类和检测等任务。这些模型在处理结构相对简单、元素相对独立的图像时效果良好,但在面对城市影像中多样且紧密关联的要素时,特别是在大规模场景理解和复杂任务需求下,往往显得力不从心。因此,传统的计算机视觉方法在应对城市影像分析时面临着显著挑战,需要发展全新的表征技术,以适应城市环境的多样性与复杂性。

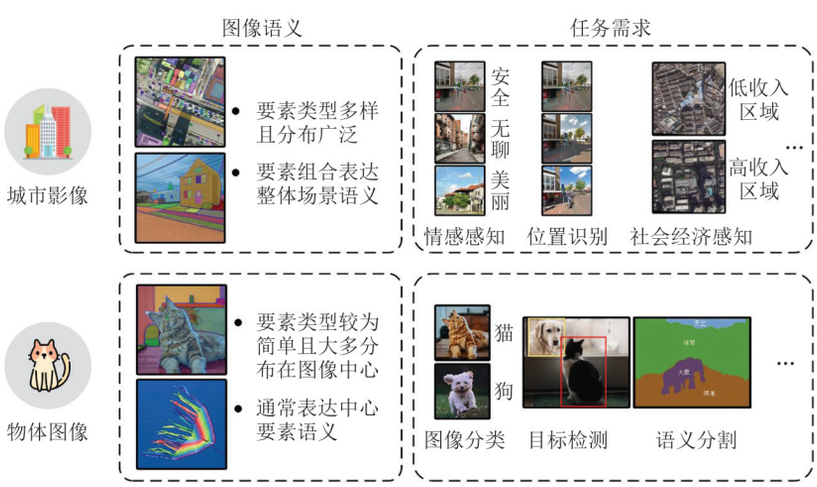


图 1　城市影像与物体图像的比较

Fig. 1　Comparison of Urban Imagery and Object Images

2　面向城市任务的城市影像计算表征

城市影像计算表征是实现城市感知与分析的核心环节,其目的是从多源影像数据中提取有用信息,以支持对城市空间结构、社会经济活动以及环境动态变化的定量分析。由于不同城市任务的需求差异,在设计计算表征方法时,需要根据具体的城市任务,聚焦与任务相关的特征,

从而确保提取的表征能够精准地反映与任务密切相关的城市要素。这意味着计算表征不仅要考虑城市场景的全貌,还要根据任务目标优先编码与之相关的特征,以提升表征结果在实际应用中的有效性和准确性。

计算表征方法可分为基于监督学习(supervised learning, SL)的传统方法和近年来发展迅速的SSL方法。SL方法依赖于大量标注数据来训练模型,而SSL方法则从未标注的数据中自动挖掘特征,能够在缺乏大规模标注数据的情况下进行高效的特征学习。

2.1 SL支持下的计算表征

SL方法依赖于标注数据,通过构建输入影像与目标变量之间的映射关系学习影像的特征表示。这类方法通常在大规模标注数据集上训练深度学习模型,以提取深层特征。在城市影像的应用中,SL方法根据具体任务需求构建相应的标注数据集,如地物分类、语义分割和目标检测等。对于高分辨率的遥感影像,构建语义分割数据集可训练模型识别道路、建筑和水体等类别[36]。遥感影像与文本配对的数据集则可用于训练类似CLIP的跨模态框架,实现影像与文本的对应[22]。对于街景影像,Places[37]和Cityscapes[13]等计算机视觉数据集因包含大多数城市场景,被广泛应用于街景影像的计算表征。此外,PlacePulse[21]数据集基于谷歌街景,标注了人类感知信息,建立了城市可视环境与非可视感知之间的联系。近期,文献[38]通过人工标注和计算机视觉技术的结合,构建了一个包含场景类型、语义分割和目标检测等信息的百万级街景影像数据集。

在这些大型标注数据集的支持下,研究者可以采用常用的神经网络架构(如卷积神经网络和Transformer)进行预训练,以获取适用于各类任务的城市影像计算表征。例如,文献[39]结合遥感影像和经济数据,利用卷积神经网络模型预测非洲地区的贫困分布;文献[40]基于PlacePulse数据集,构建了可以预测大范围人类感知的深度卷积神经网络预训练模型。尽管SL在许多任务中表现优异,但其受限于对高质量标注数据的依赖。随着影像数据规模的迅速增长,标注成本的提高成为一大挑战。此外,基于SL得到的深度特征往往局限于与标签相关的浅层表征,缺乏对数据深层内容的理解,使得这些特征难以泛化到与标签无关的城市任务。

2.2 SSL支持下的智能计算表征

SSL是一种无需大规模标注数据的计算表征方法,通过设计代理任务来学习影像的潜在特征表示,这些代理任务通常利用影像自身的内在属性,如预测影像块之间的关系、重建影像内容或对比不同视角的影像[41-42]。因为城市影像数据量巨大,且人工标注成本高昂,SSL在城市影像分析中具有重要意义。

虽然物体图像常用的旋转、翻转等数据增强方法也可作为城市影像自监督的代理任务[43],但这些方法通常倾向于编码图像中的目标主体或尽可能编码所有图像内容,难以满足特定的城市任务需求。不同的城市任务对影像信息的需求不同,例如针对社会经济或历史文化感知的任务,希望编码与周围一致的社会经济或历史文化等相关信息,而忽略局部细节;针对位置识别的任务,希望编码相对静态的城市环境要素,而忽略人流、车流和光线等动态因素。

城市影像特有的时空属性为设计面向城市任务的代理任务提供了重要支持[44-45]。针对遥感影像,文献[46]提出了地理位置感知的SSL方法,利用地理相邻区域的影像作为正样本,捕捉空间连续性特征。类似地,文献[47]开发了季节对比方法,通过利用遥感影像的季节性变化,设计时序对比学习任务,降低模型对季节和时间变化的敏感度。对于街景影像,文献[48]详细比较了自己对比、时序对比和空间近邻对比3种任务设计,揭示了不同设计适合不同类型的城市任务。具体来说,自己对比编码了全局信息,适用于对车辆、植被等动态要素敏感的城市感知任务,如人类感知;时序对比侧重于编码静态要素信息,忽略动态要素,适用于位置识别等任务;而空间近邻对比则关注场景氛围相关的语义,捕捉到与当前场景一致的社会经济、历史文化等信息,适用于社会经济指标预测等城市任务。

此外,城市影像常与其他数据源(如文本、地理信息、兴趣点数据)相关联,利用这些多模态数据设计代理任务,可以丰富特征表示的语义信息。例如,文献[49]提出了结合城市知识图谱的SSL方法,通过影像与知识图谱的关联关系,提升影像的语义表征能力;文献[50]将遥感影像与兴趣点数据关联,将城市功能相似的区域作为正样本,利用对比学习框架来学习影像的功能特征。这些方法充分利用了多模态数据的互补性,为城市任务提供了更全面的信息。

3　可解释性分析:理解计算表征

计算表征方法能够从城市影像中提取深度特征,但这些特征的每个维度通常缺乏显式的语义信息。因此,为了更有效地支持特定的城市感知任务,需要深入探究以下问题:基于不同城市影像的标注数据集或不同自监督代理任务所获得的深度特征,究竟编码了哪些信息?理解这些深度特征所包含的语义信息,不仅有助于验证计算表征过程的有效性,还能为下游任务的决策提供关键支持。具体来说,通过解析深度特征中可能反映的道路拓扑、建筑功能及环境要素等信息,可以判断这些特征是否已包含完成任务所需的关键内容,从而指导计算表征方法的进一步微调。与此同时,可解释的深度特征还能帮助决策者制定更有针对性的政策,例如在关注居民健康的情境下,通过分析深度特征对区域绿化度的编码,有针对性地提升绿化水平,为城市规划和管理提供科学依据。

城市影像计算表征的可解释性分析主要聚焦于两个方面:(1)揭示深度特征中蕴含的语义内容;(2)发现模型对图像中哪些区域更为关注。为了揭示深度特征中的语义内容,研究者常采用可解释性分析方法,如 SemAxis 和特征近邻等[51]。进一步地,研究者还会使用 t-SNE 或 PaCMAP 等降维技术[52-53],在低维空间中将深度特征的分布可视化,随后,通过观察这些特征的聚集和分离模式,可以对其进行语义解读。例如,文献[54]发现在模型训练过程中,深度特征从初始的分散状态逐渐形成不同的聚集模式,这表明模型区分不同流量模式的能力和记忆保持能力不断增强。

为了理解模型在处理城市图像时关注的区域,研究者通常通过分析网络的激活状态来进一步解析深度特征。例如,利用梯度加权类激活映射(gradient-weighted class activation mapping,Grad-CAM)等可视化技术[55],可以生成热力图,显示模型在进行分类决策时关注的图像区域[56]。利用注意力图等技术,可以展示模型在处理输入数据时对不同区域或特征的关注程度。例如,文献[57]使用 Grad-CAM 展示了相对于 BYOL 和 DenseCL,IndexNet 的深度特征能针对不同的对象而变化,从而有效解决遥感影像中的多对象问题。与 Grad-CAM 类似,注意力图通过可视化模型的注意力分配,帮助解释模型如何聚焦于输入的关键部分进行决策[58]。这些可视化结果不仅有助于解释模型的决策过程,还能揭示深度特征与城市环境要素之间的关联性。例如,文献[48]使用从自己对比、时序对比和空间近邻对比 3 种代理任务中提取的深度特征,并通过它们在特定的下游任务中的表现来体现编码能力。不仅如此,他们还对深度特征进行了多种可解释性分析,以直观展示不同深度特征所关注信息的差异。这种可解释性分析的结果如图 2 所示,它通过注意力图揭示了不同代理任务的训练模型在不同深度所关注信息的差异。例如,在深度为 1 时,自己对比和时序对比显示出更广泛的注意力分布,而空间近邻对比则更侧重于局部区域。这表明前两者在早期阶段会优先捕捉全局信息,而后者在初期则倾向于强调细节信息。然而,在深度为 11 时,自己对比会关注整个图像的全局信息,但更倾向于关注查询标记附近的区域;相比之下,时序对比中的查询 1 主要关注天空(静态区域),而放置在汽车(动态物体)上的查询 2 没有显示出对汽车的关注。这表明,时序对比倾向于通过忽略动态元素来学习时间不变特征。

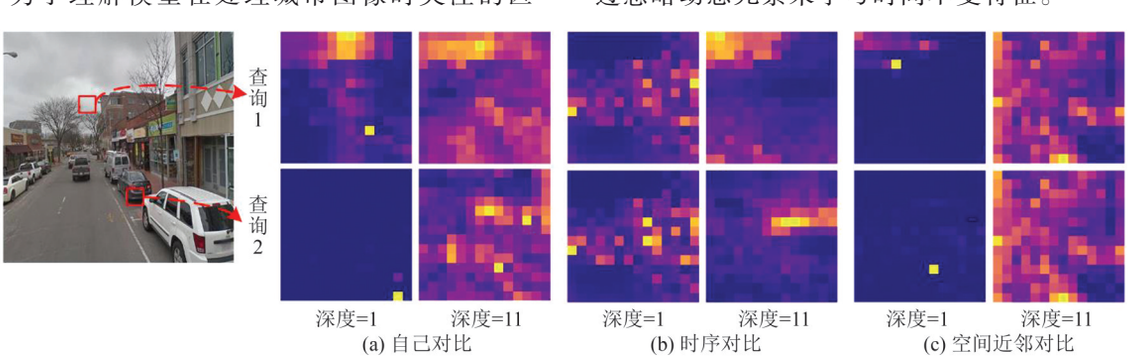


图 2　跨模型与跨深度可视化的两个查询的注意力图(引自文献[48])

Fig. 2　Attention Maps for Two Queries Visualized Across Models and Depths (Quoted from Reference [48])

通过这些方法,研究者能够更深入地理解深度特征在城市影像中的语义表达,从而为优化城市感知模型提供有价值的见解。例如,识别出模型在识别建筑物、道路或绿地时关注的具体图像

区域,可以指导特征提取和模型训练过程,提升模型在特定城市感知任务中的表现。此外,这种理解还可以帮助发现深度特征中的潜在偏差,确保模型在不同城市和环境中的泛化能力。未来,结合更多的可解释性分析方法和多模态数据源,将进一步增强对城市影像深度特征的理解,推动智能城市感知技术的发展。

4 城市影像计算表征的机遇与挑战

随着大语言模型(large language model, LLM)的快速发展,城市影像计算表征领域迎来了前所未有的机遇与挑战。LLM通过以语言为中心的框架,从海量数据中提炼了常识和世界观等基础知识,能够基于城市影像计算表征中的视觉信息,有效概括城市环境中的复杂概念、事实和观点[59]。同时,LLM在跨模态信息理解和生成方面展现出了巨大的潜力,它能够融合来自不同模态的信息,使得对城市的理解不再局限于单一的视角。然而,这也带来了新的挑战:为了将不同模态的数据有效地输入LLM,需要将这些数据对齐并编码到同一个特征空间,这对多模态对齐提出了更高的要求。不同城市空间数据在表征形式和分辨率上的差异,迫切需要更加高效的算法来确保数据的一致性和完整性。例如,遥感影像主要呈现大范围区域的环境信息,而街景影像则提供采样点附近的细节信息[54]。如何有效对齐遥感影像的面状深度特征与街景影像的点状深度特征,依然是一个亟待解决的难题[60]。

同时,城市影像计算表征的进展为AI驱动的科学研究带来了机遇与挑战。一方面,深度学习模型能够从城市影像中提取更深层次的特征,这不仅能提升具体任务的精度,还为探索更复杂的城市机理提供了可能。例如,深层次的空间和时间特征有助于揭示城市中的潜在规律,从而更好地理解城市环境的动态变化。另一方面,深度特征的可解释性仍然是AI驱动的科学研究面临的一大挑战。尽管现有的可解释性方法可以在一定程度上解释部分特征维度的信息,但计算表征中的大多数编码信息仍然处于黑箱状态。因此,如何有效解释这些深层特征,以便更好地支持科学探索和决策制定,将成为未来研究的关键方向。

5 结 语

本文探讨了城市影像与物体图像的显著区别,并分析了城市影像计算表征方法设计的特殊性。随着高时空分辨率影像数据的广泛获取,以及深度学习和SSL等技术的迅猛发展,城市影像的计算表征迎来了前所未有的机遇,为从海量城市影像中提取高维、语义丰富的特征提供了强大的支持。展望未来,研究应更加关注如何充分挖掘城市影像的时空特征和多模态属性,开发以SSL为基础的智能计算表征方法,以高效、自动化地编码与城市任务相关的关键信息,从而提升城市感知与分析的能力。此外,利用计算表征中蕴含的丰富信息,深入探索城市理论和机制的创新,也将是未来研究的关键方向。这些探索不仅有助于提升城市模型的精度和鲁棒性,还能为城市规划、管理与可持续发展提供更加精准的支持。

致谢:感谢北京大学高性能计算校级公共平台为本文研究提供的支持。

参 考 文 献

[1] 李德仁. 脑认知与空间认知:论空间大数据与人工智能的集成[J]. 武汉大学学报(信息科学版), 2018, 43(12): 1761-1767.
LI Deren. Brain Cognition and Spatial Cognition: On Integration of Geo-Spatial Big Data and Artificial Intelligence[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(12): 1761-1767.

[2] 李德仁, 沈欣. 论基于实景影像的城市空间信息服务:以影像城市·武汉为例[J]. 武汉大学学报(信息科学版), 2009, 34(2): 127-130.
LI Deren, SHEN Xin. Geo-Spatial Urban Information Services Based on Real Image: Take Image City·Wuhan as an Example[J]. *Geomatics and Information Science of Wuhan University*, 2009, 34(2): 127-130.

[3] 涂伟, 曹劲舟, 高琦丽, 等. 融合多源时空大数据感知城市动态[J]. 武汉大学学报(信息科学版), 2020, 45(12): 1875-1883.
TU Wei, CAO Jinzhou, GAO Qili, et al. Sensing Urban Dynamics by Fusing Multi-sourced Spatio-temporal Big Data[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(12): 1875-1883.

[4] ZHANG F, SALAZAR-MIRANDA A, DUARTE F, et al. Urban Visual Intelligence: Studying Cities with Artificial Intelligence and Street-Level Imagery[J]. *Annals of the American Association of Geographers*, 2024, 114(5): 876-897.

[5] BILJECKI F, ITO K. Street View Imagery in Ur-

ban Analytics and GIS: A Review[J]. *Landscape and Urban Planning*, 2021, 215: 104217.

[6] LOWE D G. Object Recognition from Local Scale-Invariant Features[C]//The 7th IEEE International Conference on Computer Vision, Kerkyra, Greece, 2002.

[7] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005.

[8] 邵振峰, 孙悦鸣, 席江波, 等. 智能优化学习的高空间分辨率遥感影像语义分割[J]. 武汉大学学报(信息科学版), 2022, 47(2): 234-241.
SHAO Zhenfeng, SUN Yueming, XI Jiangbo, et al. Intelligent Optimization Learning for Semantic Segmentation of High Spatial Resolution Remote Sensing Images[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(2): 234-241.

[9] 李彦胜, 张永军. 耦合知识图谱和深度学习的新一代遥感影像解译范式[J]. 武汉大学学报(信息科学版), 2022, 47(8): 1176-1190.
LI Yansheng, ZHANG Yongjun. A New Paradigm of Remote Sensing Image Interpretation by Coupling Knowledge Graph and Deep Learning[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(8): 1176-1190.

[10] 龚健雅, 张展, 贾浩巍, 等. 面向多源数据地物提取的遥感知识感知与多尺度特征融合网络[J]. 武汉大学学报(信息科学版), 2022, 47(10): 1546-1554.
GONG Jianya, ZHANG Zhan, JIA Haowei, et al. Multi-source Data Ground Object Extraction Based on Knowledge-Aware and Multi-scale Feature Fusion Network[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(10): 1546-1554.

[11] FAN Z Y, ZHANG F, LOO B P Y, et al. Urban Visual Intelligence: Uncovering Hidden City Profiles with Street View Images[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2023, 120(27): e2220417120.

[12] GEBRU T, KRAUSE J, WANG Y L, et al. Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods Across the United States[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, 114(50): 13108-13113.

[13] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[14] ROS G, SELLART L, MATERZYNSKA J, et al. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[15] BENGIO Y, COURVILLE A, VINCENT P. Representation Learning: A Review and New Perspectives[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8): 1798-1828.

[16] FENG S S, CONG G, AN B, et al. POI2Vec: Geographical Latent Representation for Predicting Future Visitors[C]//The 31st AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017.

[17] ZHAI W, BAI X Y, SHI Y, et al. Beyond Word2vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2vec and POIs[J]. *Computers, Environment and Urban Systems*, 2019, 74: 1-12.

[18] JANOWICZ K, GAO S, MCKENZIE G, et al. GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond[J]. *International Journal of Geographical Information Science*, 2020, 34(4): 625-636.

[19] 高松. 地理空间人工智能的近期研究总结与思考[J]. 武汉大学学报(信息科学版), 2020, 45(12): 1865-1874.
GAO Song. A Review of Recent Researches and Reflections on Geospatial Artificial Intelligence[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(12): 1865-1874.

[20] LI W W, HSU C Y. GeoAI for Large-Scale Image Analysis and Machine Vision: Recent Progress of Artificial Intelligence in Geography[J]. *ISPRS International Journal of Geo-Information*, 2022, 11(7): 385.

[21] DUBEY A, NAIK N, PARIKH D, et al. Deep Learning the City: Quantifying Urban Perception at a Global Scale[C]//The 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016.

[22] ZHANG Z L, ZHAO T C, GUO Y L, et al. RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-23.

[23] MANS LARSSON M, STENBORG E, HAM-

MARSTRAND L, et al. A Cross-Season Correspondence Dataset for Robust Semantic Segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019.

[24] TOWNSHEND J, JUSTICE C, LI W, et al. Global Land Cover Classification by Remote Sensing: Present Capabilities and Future Possibilities[J]. *Remote Sensing of Environment*, 1991, 35(2/3): 243-255.

[25] XIE Y C, SHA Z Y, YU M. Remote Sensing Imagery in Vegetation Mapping: A Review[J]. *Journal of Plant Ecology*, 2008, 1(1): 9-23.

[26] TUCKER C J. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation[J]. *Remote Sensing of Environment*, 1979, 8(2): 127-150.

[27] ZHANG X Y, DU S H, WANG Q. Integrating Bottom-up Classification and Top-Down Feedback for Improving Urban Land-Cover and Functional-Zone Mapping[J]. *Remote Sensing of Environment*, 2018, 212: 231-248.

[28] SETO K C, FRAGKIAS M, GÜNERALP B, et al. A Meta-Analysis of Global Urban Land Expansion[J]. *PLoS One*, 2011, 6(8): e23777.

[29] ANGUELOV D, DULONG C, FILIP D, et al. Google Street View: Capturing the World at Street Level[J]. *Computer*, 2010, 43(6): 32-38.

[30] NEUHOLD G, OLLMANN T, BULÒ S R, et al. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes[C]//IEEE International Conference on Computer Vision, Venice, Italy, 2017.

[31] GOEL R, GARCIA L M T, GOODMAN A, et al. Estimating City-Level Travel Patterns Using Street Imagery: A Case Study of Using Google Street View in Britain[J]. *PLoS One*, 2018, 13(5): e0196521.

[32] LI X J, ZHANG C R, LI W D, et al. Assessing Street-Level Urban Greenery Using Google Street View and a Modified Green View Index[J]. *Urban Forestry & Urban Greening*, 2015, 14(3): 675-685.

[33] SUN M R, ZHANG F, DUARTE F, et al. Understanding Architecture Age and Style Through Deep Learning[J]. *Cities*, 2022, 128: 103787.

[34] ZHOU B L, LAPEDRIZA A, XIAO J X, et al. Learning Deep Features for Scene Recognition Using Places Database[C]//The 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014.

[35] TIGHE J, NIETHAMMER M, LAZEBNIK S. Scene Parsing with Object Instance Inference Using Regions and Per-exemplar Detectors[J]. *International Journal of Computer Vision*, 2015, 112(2): 150-171.

[36] WANG J J, ZHENG Z, MA A L, et al. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation[C]//The 25th Neural Information Processing Systems Track on Datasets and Benchmarks, Online, 2021.

[37] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 Million Image Database for Scene Recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(6): 1452-1464.

[38] HOU Y J, QUINTANA M, KHOMIAKOV M, et al. Global Streetscapes: A Comprehensive Dataset of 10 Million Street-Level Images Across 688 Cities for Urban Science and Analytics[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 215: 216-238.

[39] JEAN N, BURKE M, XIE M, et al. Combining Satellite Imagery and Machine Learning to Predict Poverty[J]. *Science*, 2016, 353(6301): 790-794.

[40] ZHANG F, ZHOU B L, LIU L, et al. Measuring Human Perceptions of a Large-Scale Urban Region Using Machine Learning[J]. *Landscape and Urban Planning*, 2018, 180: 148-160.

[41] HE K, CHEN X, XIE S, et al. Masked Autoencoders Are Scalable Vision Learners[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Louisiana, USA, 2022.

[42] HE K M, FAN H Q, WU Y X, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020.

[43] LI T, XIN S D, XI Y X, et al. Predicting Multi-level Socioeconomic Indicators from Structural Urban Imagery[C]//The 31st ACM International Conference on Information & Knowledge Management, Atlanta, USA, 2022.

[44] LIAO C, HU H, YUAN X K, et al. BCE-Net: Reliable Building Footprints Change Extraction Based on Historical Map and Up-to-Date Images Using Contrastive Learning[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023, 201: 138-152.

[45] STALDER S, VOLPI M, BÜTTNER N, et al. Self-Supervised Learning Unveils Urban Change from Street-Level Images[J]. *Computers, Environ-*

ment and Urban Systems, 2024, 112: 102156.

[46] AYUSH K, UZKENT B, MENG C L, et al. Geography-Aware Self-Supervised Learning[C]//IEEE/CVF International Conference on Computer Vision, Online, 2021.

[47] MAÑAS O, LACOSTE A, GIRO-I-NIETO X, et al. Seasonal Contrast: Unsupervised Pre-training from Uncurated Remote Sensing Data[C]//IEEE/CVF International Conference on Computer Vision, Online, 2021.

[48] LI Y, HUANG Y J, ZHANG F. Learning Street View Representations Based on a Spatiotemporal Contrastive Learning Framework[J]. *Computers, Environment and Urban Systems*, 2026, 125: 102393.

[49] LIU Y, ZHANG X, DING J T, et al. Knowledge-Infused Contrastive Learning for Urban Imagery-Based Socioeconomic Prediction[C]//The ACM Web Conference, Austin, USA, 2023.

[50] LI T, XI Y X, WANG H D, et al. Learning Representations of Satellite Imagery by Leveraging Point-of-Interests[J]. *ACM Transactions on Intelligent Systems and Technology*, 2023, 14(4): 1-32.

[51] AN J S, KWAK H, AHN Y Y. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment[C]//The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018.

[52] MAATEN L, HINTON G. Visualizing Data Using t-SNE[J]. *Journal of Machine Learning Research*, 2008, 9(86): 2579-2605.

[53] WANG Y F, HUANG H Y, RUDIN C, et al. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization[J]. *Journal of Machine Learning Research*, 2021, 22(1): 9129-9201.

[54] ZHANG Y, LI Y, ZHANG F. Multi-level Urban Street Representation with Street-View Imagery and Hybrid Semantic Graph[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 218: 19-32.

[55] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.

[56] ZHANG F, WU L, ZHU D, et al. Social Sensing from Street-Level Imagery: A Case Study in Learning Spatio-Temporal Urban Mobility Patterns[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019, 153: 48-58.

[57] MUHTAR D, ZHANG X L, XIAO P F. Index Your Position: A Novel Self-Supervised Learning Method for Remote Sensing Images Semantic Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 4411511.

[58] CHEFER H, GUR S, WOLF L. Transformer Interpretability Beyond Attention Visualization[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021.

[59] HOU C, ZHANG F, LI Y, et al. Urban Sensing in the Era of Large Language Models[J]. *The Innovation*, 2025, 6(1): 100749.

[60] HUANG Y J, ZHANG F, GAO Y, et al. Comprehensive Urban Space Representation with Varying Numbers of Street-Level Images[J]. *Computers, Environment and Urban Systems*, 2023, 106: 102043.