Unlocking Location Intelligence: A Survey from Deep Learning to The LLM Era

Xixuan Hao[1], Yutian Jiang[1], Xingchen Zou[1], Jiabo Liu[1], Yifang Yin[2], Yuxuan Liang[1]
[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Institute for Infocomm Research (I^{2}R), A*STAR
{xhao390,xzou428}@connect.hkust-gz.edu.cn,{yutianjiang,jiaboliu.mail}@gmail.com,
yin_yifang@i2r.a-star.edu.sg,yuxliang@outlook.com

Abstract

Location Intelligence (LI), the science of transforming location-centric geospatial data into actionable knowledge, has become a cornerstone of modern spatial decision-making. The rapid evolution of Geospatial Representation Learning is fundamentally reshaping LI development through two successive technological revolutions: the deep learning breakthrough and the emerging large language model (LLM) paradigm. While deep neural networks (DNNs) have demonstrated remarkable success in automated feature extraction from structured geospatial data (e.g., satellite imagery, GPS trajectories), the recent integration of LLMs introduces transformative capabilities for cross-modal geospatial reasoning and unstructured geo-textual data processing. This survey presents a comprehensive review of geospatial representation learning across both technological eras, organizing them into a structured taxonomy based on the complete pipeline comprising: (1) data perspective, (2) methodological perspective and (3) application perspective. We also highlight current advancements, discuss existing limitations, and propose potential future research directions in the LLM era. This work offers a thorough exploration of the field and providing a roadmap for further inno- vation in LI. The summary of the up-to-date paper list can be found in https://github.com/CityMind-Lab/Awesome-Location-Intelligence and will undergo continuous updates.

CCS Concepts

• **Information systems → Spatial-temporal systems**.

Keywords

Location Intelligence, Geospatial Representation Learning, Location Embedding, Region Embedding, Multimodal Learning

1 Introduction

In the era of ubiquitous geospatial sensing and AI proliferation, the ability to extract actionable insights from location-centric data has become a critical driver of scientific discovery and societal transformation. This paradigm, termed **Location Intelligence (LI)**, which converts raw geographic observations into contextualized knowledge that enhances decision-making across environmental, urban, and human mobility domains [7, 93]. The exponential growth of multimodal geospatial data, from satellite imagery to GPS trajectories and geo-textual records, has catalyzed a fundamental shift in LI development. Traditional Geographic Information Systems (GIS), while effective for basic spatial operations [6, 62], now give way to Geospatial Artificial Intelligence (GeoAI) – a transformative fusion of deep learning and spatial science that unlocks unprecedented capabilities in modeling geographic complexity [35, 79].

At the core of Location Intelligence, **Geospatial Representation Learning** plays a pivotal role. This discipline aims to extract meaningful features and patterns from heterogeneous location-centric data, facilitating a profound comprehension of geographic processes and dynamics. Such learning method encompasses the incorporation of multidimensional information, including spatial relationships, temporal variations, and environmental factors. By constructing efficient location representations, researchers and practitioners are empowered to conduct comprehensive spatial analyses, make accurate predictions, and support informed decision-making [115, 124].

The evolution of LI has undergone two revolutionary phases. The first wave, driven by deep neural networks, established new benchmarks in spatial pattern recognition through architectures like CNNs for satellite image analysis [47], GNNs for urban network modeling [113], and RNNs for trajectory prediction [8]. These approaches demonstrated remarkable success in automating feature extraction from structured geospatial data (coordinates, trajectories, raster images), enabling applications ranging from climate change prediction [90] to intelligent transportation systems [111]. However, the emerging second wave – propelled by large language models (LLMs) and multimodal foundation models (FMs) – is redefining the boundaries of LI. Modern GeoAI systems now integrate unstructured textual data (social media geotags, administrative reports) with conventional spatial data streams through cross-modal alignment [124], while leveraging pretrained knowledge from LLMs to enhance spatial reasoning and human-AI collaboration in urban planning [23, 56]. Although Large Language Models (LLMs) have not yet been extensively adopted in LI, their proven efficacy in other disciplines (e.g., Embedded AI [20] and Mathematical reasoning [1]) highlights critical potentials that merit focused exploration and prospective analysis.

In Figure 1, we delineate the complete pipeline for geospatial representation learning with location-centric data, structured along three fundamental dimensions: data, methodology, and applications. This framework incorporates two distinct but complementary geospatial data paradigms: (1) location-level representations that capture point-based spatial entities and their local contextual information, and (2) region-level representations that model areal units with aggregated spatial characteristics. The three-dimensional architecture systematically addresses: (i) *the data perspective*, focusing on geospatial data acquisition, preprocessing, and feature engineering; (ii) *the methodological perspective*, encompassing model architecture design and representation learning techniques; and (iii) *the application perspective*, demonstrating the deployment of learned representations in various geospatial analysis tasks and decision support systems.

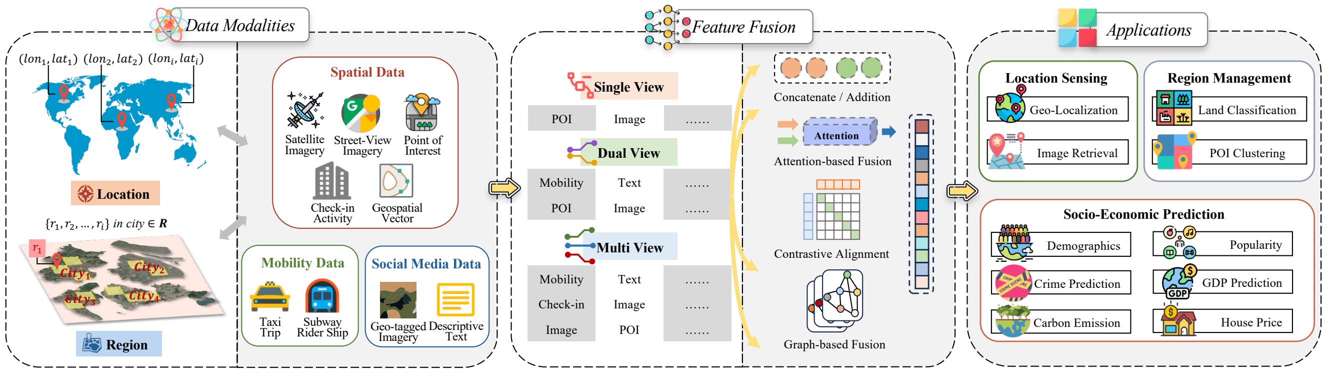

Figure 1: The complete pipeline of geospatial representation learning with location-centric data.

Motivations & Related Surveys. Although deep neural networks have been broadly applied in LI tasks, this rapidly expanding field still lacks a systematic review. As is shown in Table 1, Chen et al. [14] systematically review self-supervised learning in GeoAI across location- and region-level data. However, their framework is limited to geometric primitives (points, lines, polygons), omitting multimodal integration (e.g., visual/textual data) and lacking systematic evaluation of downstream applications. The focus of [63] is exclusively on location encoding, neglecting the fact that surrounding region embedding is also an important component in location intelligence. Similarly, [42] focuses solely on the application of Graph Neural Network architectures in urban computing, without discussing the significant contributions of other methods. [51] is confined to a specific subset of data, namely visual and mapping data, and therefore fails to comprehensively represent the broader scope of the GeoAI domain. Wang et al. [93] discusse the entire pipeline for spatio-temporal learning, including data, methodology, and application. However, limited by its content from five years ago, it has limited coverage of recent advancements, especially in the areas of Transformers [95], pretraining techniques [74], and large language models (LLMs) [18]. In contrast, our focus encompasses the complete pipeline in Figure 1 and propose a more fine-grained taxonomy to systematically review and summarize the current status of deep learning techniques for LI.

Our Contributions. The contributions of this work can be sumarized as the following three aspects. (1) A structured taxonomy. A comprehensive overview of the field is presented in Table 2, from the perspectives of three methodologies, with a particular focus on modality, coverage, and downstream tasks. (2) A comprehensive review. Building upon the proposed taxonomy, the current advancements in geospatial representation learning are comprehensively outlined, providing a systematic exploration of the field's progress. (3) Prospects for Future Directions in the LLM era. We discuss the remaining limitations of existing works and point out possible future directions in the LLM era.

Table 1: Comparison between our survey and related surveys.

Survey	Taxonomy	Data Coverage	Methodology
Chen et al. [14]	Data Type	Location,Region	Unsupervised Learning
Mai et al. [63]	Pipeline	Location	Supervised Learning
Jin et al. [42]	Pipeline	Spatio-temporal Data	Graph Neural Network
Li and Hsu [51]	Pipeline	Visual and Mapping Data	Supervised + Unsupervised Learning
Wang et al. [93]	Pipeline	Spatio-temporal Data	Supervised Learning
Ours	Pipeline	Location,Region	Supervised + Unsupervised Learning

2 Preliminaries

In this section, we introduce the basic concepts and provide an intuitive illustration of these concepts in Figure 2.

2.1 Region Representation Learning

Definition 1: Urban Region. A city can be partitioned into a set of urban regions $\mathcal{R} = \{r_1, r_2, \ldots, r_i, \ldots, r_N\}$ [58], where N denotes the number of sub-regions in the specific city, following various criteria: (1) road network layouts [27]. (2) administrative boundaries [94]. (3) subdivisions based on specific sizes and shapes (e.g., rectangular/hexagonal grids) [21, 92, 110].

Definition 2: Urban Region Attributes. Urban Region attributes are the inherent social and geographic characteristics of urban areas [114], which can be learned from multiple data modalities, such as POI, mobility, and urban layout (Sec. 3). An urban region can be comprehensively represented by a series of attributes with dimension d, denoted by $\mathcal{A}_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,j}, \ldots, a_{i,n}\}$, where i denotes the i-th sub-region, j represents the j-th attribute, and n denotes the number of attributes for the region.

Definition 3: Urban Region Embedding. The objective of urban region representation learning is to generate region embeddings with high generalizability by integrating regions and their attributes using various methodologies, such as Contrastive Learning and Graph Neural Networks [124]. The process can be expressed as $v_i = \boldsymbol{M}_R(\mathcal{A})$, where $\boldsymbol{M}_R$ refers to the corresponding networks. The distributed embedding v_i from each sub-region r_i collectively form the set of region embeddings with D dimension, which can be obtained as $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}, v_i \in \mathbb{R}^D$.

2.2 Location Representation Learning

Definition 1: Location. A location can be defined as a specific position or area in space that is identified by its geographical coordinates, denoted as $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}, l_i = [\lambda_i, \phi_i]$, where $\lambda_i \in [-\pi, \pi]$ represents the longitude and $\phi_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ represents the latitude, n indicates the number of locations.

Definition 2: Location Embedding. Location representation learning process is divided into two parts: coordinate vectorization and representation fusion. In coordinate vectorization, a coordinate encoder $\boldsymbol{E}(\mathcal{L}) : \mathbb{R}^{n\times 2} \rightarrow \mathbb{R}^{n\times d}$ aims to project discrete coordinates into d dimensional vectors [64, 107]. Since the semantic information of a geographic location does not originate from its numerical coordinate alone [31], a semantically meaningful location embedding

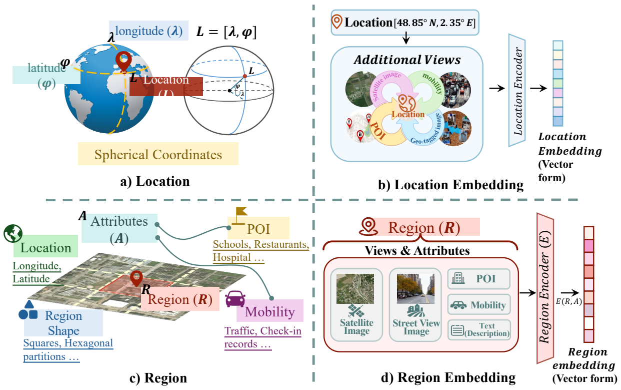

Figure 2: The concepts of location, location embedding, region, and region embedding in geospatial representation.

necessitates the integration of extrinsic semantics from other data modalities. In representation fusion, a model $M_{\mathcal{L}}$ is constructed to integrate coordinate vectors with various data modalities (e.g., visual images and texts) [33, 45, 65, 85] for capturing geographical and functional attributes of location across the globe.

3 Data Modality Perspective

This section provides a concise overview of the diverse data mobilities employed in geospatial representation across various locations and regions. These data types are categorized into four primary classes: **spatial data**, **mobility data**, **socio-economic attributes**, and **social media data**. Each category serves distinct purposes and contributes unique insights to the comprehensive understanding of geospatial dynamics.

3.1 Spatial Data

Spatial data identifies the geographic location and topography of objects distribution on the Earth's surface, which can be used for describing the details of location and region, including:

- **Points of Interest (POIs)** represent a collection of specific locations or sites of significance [96]. It can be denoted as $\mathcal{P}^r = \{p_1, p_2, \ldots, p_m\}$, where $\mathcal{P}^r$ represents a set of POIs, and m denotes the number of POIs within a region r. Formally, each POI, $p_i = [n_i, l_i, c_i, a_i]$, contains the name n_i, coordinates l_i, category c_i and additional attributes a_i, where the category is selected from a hierarchical taxonomy that includes major categories and corresponding subcategories.
- **Check-in Activities** as a region attribute [114], are generated by users at specific POIs [113], incorporating human activities and reflecting urban dynamics. Each check-in record can be formally represented as a triplet (u, p, t), where u denotes the user, p represents the POI, and t indicates the timestamp.
- **Satellite Imagery** provides a bird's-eye view of Earth's surface, capturing urban layouts, environments and building distribution [110]. Each city or region can be segmented into multiple satellite-image tiles, denoted by $\mathcal{SI} = \{I_1^{sa}, I_2^{sa}, \ldots, I_n^{sa}\}, I_i^{sa} \in \mathbb{R}^{H\times W\times C}$, where n denotes the total number of satellite images, and H, W, and C represent the height, width, and number of channels of each imagery, respectively.
- **Street View Imagery** is a photograph that can be captured from ground-level perspectives along streets and roads, providing a comprehensive visual clues at street level, denoted by $\mathcal{SV} = \{I_1^{sv}, I_2^{sv}, \ldots, I_m^{sv}\}, I_i^{sv} \in \mathbb{R}^{H\times W\times C}$, where m denotes the number of street view imagery. The definitions of H, W, and C are the same as those in satellite imagery.
- **Geospatial Vector** represents geographic features such as points, polylines, and polygons, capturing spatial relationships and attributes. In OpenStreetMap (OSM), these vectors are defined by nodes (i.e., p_i), ways (i.e., $w_i = [p_1, p_2, \ldots, p_m]$), relations (i.e., $e_i = [w_1, w_2, \ldots, w_n]$), and building footprints [4, 82].

3.2 Mobility Data

Mobility data reflects human's transitions behavior among POIs within cities [26, 70], recorded as a sequence of points with geographic coordinates and timestamps, which is often represented by taxi trip [21, 26, 116]. Taxi trip is generated within urban areas, typically related to urban dynamics, denoted by a set of trajectories $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$. Each trip τ_i contains $[p_s, p_e; t_s, t_e]$, where p_s and p_e represent the starting and ending geographic coordinates (i.e., latitude and longitude), and t_s and t_e display the starting and ending times of the trip, respectively [88]. Beyond taxi data, other sources such as subway ridership [77] are also leveraged to represent mobility.

3.3 Social Media Data

The proliferation of social media and location-based platforms (e.g., Twitter/X, Facebook) has fostered the crowdsourcing of extensive geospatial data through user-generated, geo-tagged content [124]. This multimodal social sensing data has fueled interdisciplinary advances in multimodal representation learning [16, 17].

- **Geo-tagged Imagery** refers to images linked with metadata from platforms like social media (e.g., YouTube, Facebook), open mapping services (e.g., Google Maps), which contains geographic coordinates, timestamps, identifiers, etc., [65, 66, 107, 109].
- **Descriptive Texts** sourced from online encyclopedias (e.g., websites [94], Wikipedia [84]), and generative models [30, 103], provide contextual details about locations and entities [78]. The emergence of large language models (LLMs)[18, 71] further enriches textual descriptions, advancing multimodal alignment and representation learning [56, 124].

3.4 Socio-Economic Attributes

Socio-economic attributes describe the social and economic aspects of urban development and geospatial dynamics, which are essential for understanding the interactions between population groups, economic sectors, and spatial distributions [91].

Common socio-economic attributes employed in geospatial representation tasks include *land usage* [11, 76, 81], *demographic data* [28], *crime statistics* [37], *income* [88], *poverty* [40], *check-in activities* [97], *nightlight imagery* [43], *house price* [30], and *carbon emissions* [103]. They often serve as predictive indicators in downstream tasks to assess the performance of geospatial representations. Other data such as the *number of takeaways (and reviews)* [99], *women's BMI* [48], and *Origin-Destination Employment Statistics (OSRM)* [59] also offer unique insights into urban development.

3.5 Overall

Figure 3 shows the frequency distribution of four data categories, highlighting three primary findings: (1) Spatial data dominates (67%), offering comprehensive insights via POIs and satellite imagery; (2) Mobility data ranks second (18%), with taxi trip patterns frequently representing movement; (3) Social media data, though comprising only 13%, holds substantial future potential for textual analysis, particularly with advancements in LLMs. Figure 4 further examines data usage in cities such as Beijing, New York, and Chicago, which leverage diverse datasets, including POIs, mobility patterns, and imagery. Developed cities receive more focus due to their well-established infrastructure and open data initiatives, while studies on African cities emphasize socio-economic metrics.

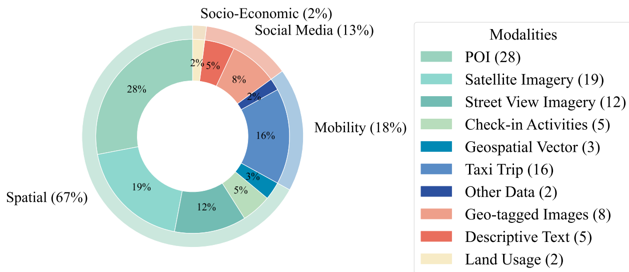

Figure 3: The usage frequency of data modalities during learning stage across four categories in the survey. Each category contains commonly used data modalities.

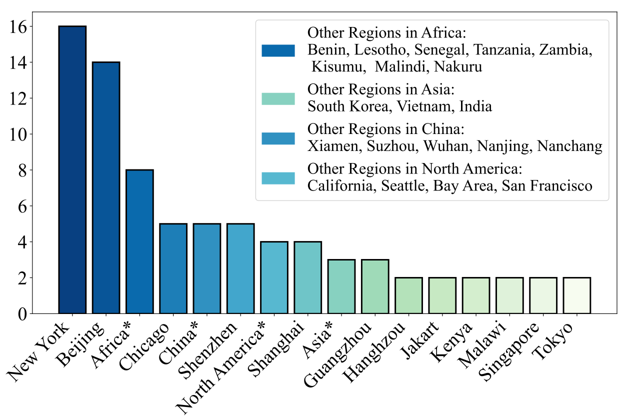

Figure 4: The dataset usage frequency across cities / countries in relevant papers. Popular cities are listed individually, while other cities within a country are aggregated and marked with * denoting "other regions" in annotations.

4 Methodology Perspective

4.1 Data-centric View

Cross-modal data integration [15, 115, 121, 124] is pivotal for shaping global or city-level geospatial embeddings. We systematically review recent advancements in geospatial representation learning (i.e., geospatial embedding), analyzing their evolution and classifying methods into **single-view**, **dual-view**, and **multiple-view** approaches based on the complexity of views in learning geospatial representations. A taxonomy of representative work is provided in Figure 5, with detailed summaries in Table 2. We organize our discussion from location- and region-level views respectively.

4.1.1 Single View. In this section, we systematically review *Single View Geospatial Embedding* methods, which project diverse spatial data modalities into a unified representational space.

- Single View Location Embedding. Single view location embedding refers to the process of extracting and representing information centered around specific **geographical coordinates**. Loc2Vec [80] represents one of the earliest efforts to captures location semantics using environmental contexts retrieved from GIS queries. Place2Vec [102] uses the distributional differences of POI types as semantic information to enhance place embeddings.

- Single View Region Embedding. Similarly, single-view region embeddings extract and refine the single-modal attributes associated with their corresponding urban regions.

- **1) Satellite Imagery**: Satellite visual data extraction is the most significant component in single-view region embedding due to its global coverage, accessibility [124], and ability to capture diverse physical, environmental, and socio-economic features. READ [27] pioneers a semi-supervised approach using the mean-teacher model [83] to analyze satellite imagery, with a specific emphasis on economic scales in South Korea. While [29] further integrates the partial order graph to cluster satellite imagery, thus facilitating the assessment of economic development.
- **2) Mobility**. Mobility Flow [88, 97, 105] captures geospatial semantics dynamically through movement patterns, revealing spatial interactions, temporal variations, and socio-economic traits. HDGE [88] employs a heterogeneous graph structure to simultaneously account for temporal dynamics and multi-hop location transitions. MGFN [97] prioritizes temporal information by constructing mobility graphs per timestep, aggregating similar graphs to extract varied mobility patterns.

4.1.2 Dual View. The dual view perspective integrates data from two distinct modalities to address the limitations of single-modality frameworks. For example, visual data provides detailed spatial representations, whereas mobility flow captures dynamic processes and temporal variations.

- Dual View Location Embedding. In summary, the dual view location embedding presents a **location + others** pattern, which integrates geographic location data with various multidimensional modalities, facilitating a deeper understanding of the complex interactions. In this pattern, *visual data* continues to play a predominant role. By integrating geographic location information with satellite imagery [2, 45, 65] and geo-tagged imagery [66, 85, 108, 109], we can get a comprehensive representation of land use types, changes in natural resources and socio-economic conditions.

In addition to visual data, the representation of *other forms of information* also merits considerable attention. MGeo [19], designed for query-POI matching, treats geographic context as an independent modality, emphasizing spatial relationships between a point and its surroundings. As large language models (LLMs) gain widespread attention [46, 119], GeoLLM [67] and LLMGeovec [33] both integrate nearby location information from OSM to construct textual prompts, enabling the extraction of geospatial knowledge.

- Dual View Region Embedding. From the perspective of explicit region characterization, the dual view exemplifies two distinctive developmental trajectories of model architectures: **POI +**

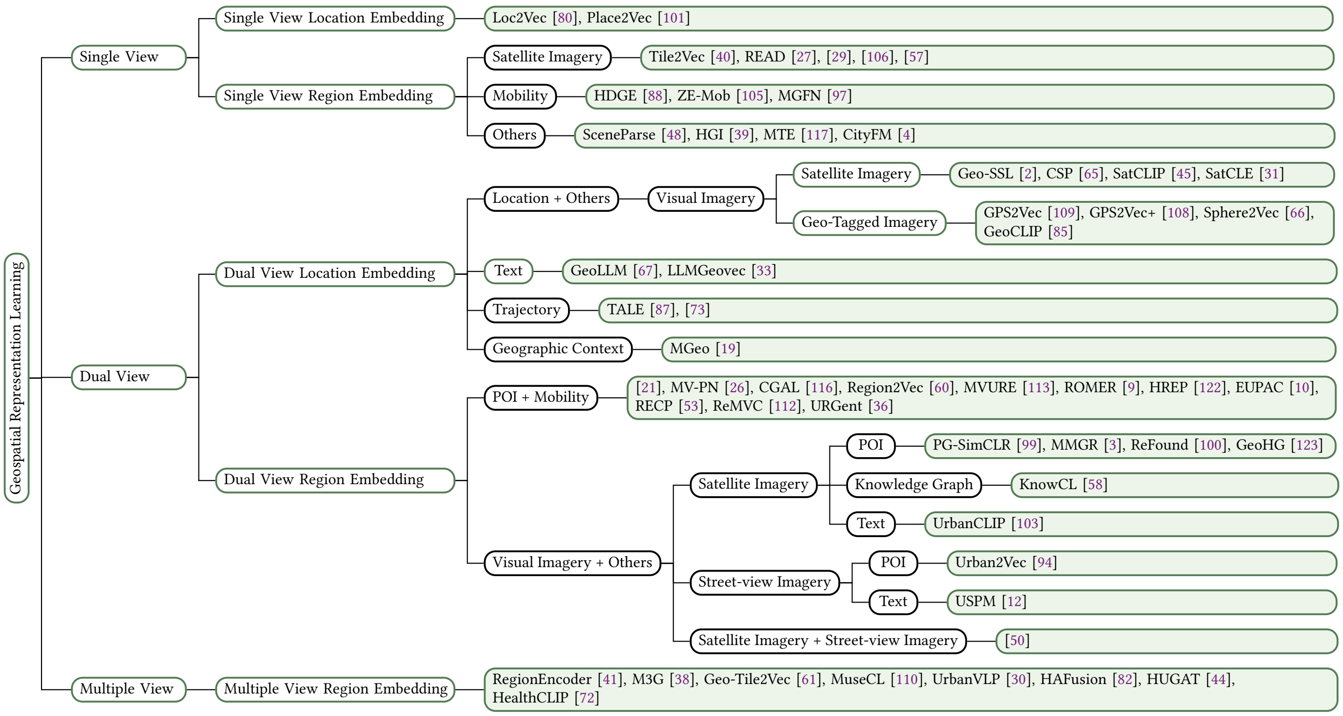

Figure 5: A taxonomy of representative works for geospatial representation learning.

Mobility [9, 10, 21, 26, 53, 60, 77, 92, 112, 113, 116, 120, 122] and **Visual Imagery + Others** [3, 12, 50, 58, 94, 99, 100, 104, 118]. Next, we discuss each of them in detail.

- **1) POI + Mobility.** In urban region embedding, POIs offer static functional attributes, while mobility data reflects dynamic activity patterns. Their integration facilitates the modeling of spatiotemporal characteristics, enhancing urban dynamic analysis. Following a comprehensive literature review, we summarize that the primary distinction among relevant works lies in the approaches employed for mobility graph construction. Studies such as [21, 26, 60, 116] utilize geographical distance and human mobility to construct multi-view graphs, which are subsequently integrated through techniques like autoencoders [21, 26] and graph convolutional networks (GCNs) [60, 116]. MVURE [113] and ROMER [9] integrate additional graph perspectives, such as check-in and source / destination graphs, employing attention mechanisms to fuse information across multiple modules. [10, 122] explore heterogeneous graphs in region embedding, using multiple edge types to represent diverse node relationships.
- **2) Visual Imagery + Others.** Here we discuss scenarios where satellite imagery and street-level imagery are each combined with other modalities as visual data.

  Satellite Imagery. The semantic richness of satellite imagery enables it to be enhanced by various modality data to precisely characterize regions. Compared to other modalities, ***POI*** data, characterized by its fine-grained categorization, strong spatial alignment, and ease of acquisition, can be effectively integrated with satellite imagery to enhance socio-economic representation. PG-SimCLR [99] and MMGR [3] are grounded in the representation of physical and geographic information derived from satellite imagery, leveraging POI categories as a quantitative proxy for human activity factors. ReFound [100] transforms POI data and satellite imagery into unified embeddings using knowledge distillation from multiple pre-trained foundation models, transferring their generalization capabilities to urban region modeling.

  KnowCL [58] pioneers the use of ***Knowledge Graphs (KGs)*** to model urban knowledge, by introducing image-KG pairs to enhance semantic-visual representation alignment through mutual information maximization. With the surge of LLMs across various domains [46, 119], the interpretability of text has garnered significant attention. UrbanCLIP [104] leverages ***LLM*** to generate text to describe the content of satellite imagery.

Figure 7 in Appendix illustrates the temporal development trends of dual view approaches. The **Street-view + Others** pattern exhibit similarities, which can be found in Appendix B.3.

4.1.3 Multiple View. Although the dual view paradigm currently dominates the geospatial representation learning field, the diverse and flexible data modalities in this domain still offer opportunities for enhanced representations through additional information. However, the trade-off between the cost of incorporating new modalities and the performance improvements must be carefully evaluated.

Researchers have creatively combined the two pipelines within the dual view framework. For example, RegionEncoder [41] initially integrates satellite imagery, POIs, and human mobility data to jointly learn region representations through GCN and denoising autoencoder. To more effectively model intra-relationships corresponding to POIs, both M3G [38] and Geo-Tile2Vec [61] incorporate street-level visual data in conjunction with POIs and mobility information. UrbanVLP [30] takes a different approach by utilizing textual descriptions as a substitute for the functionality of POIs and mobility data. It incorporates the web-scale knowledge compressed within LLMs into region embeddings with effective quality control.

	Model	Venue	Modality	Coverage	Downstream Task	Code
Single View	Loc2Vec [80]	Blog 2018	OSM	Global	Visualization	-
	Place2Vec [101]	SIGSPATIAL 2017	POI	Las Vegas	Hierarchy-based Evaluation | Binary HIT Evaluation | Ranking-based HIT Evaluation | Place Type Compression | Place Type Profiles	Link
	Tile2Vec [40]	AAAI 2019	Satellite Imagery	-	Land Cover Classification | Poverty Prediction | Health Index Prediction	Link
	READ [27]	AAAI 2020	Satellite Imagery	South Korea	Population Prediction | Age Prediction | Household Prediction | Income Prediction	Link
	[29]	KDD 2020	Satellite Imagery	Korea | Malawi | Vietnam	Economic Development Evaluation | Economic Visual Interpretation| Change Detection	Link
	[106]	Nat Commun 2020	Satellite Imagery	Benin | Lesotho | Malawi | Rwanda | Sierra | Leone Senegal | Tanzania | Zambia	Asset Wealth Estimation | Social Protection Program | Satellite-Estimated Wealth Distribution| Temperature Distribution	Link
	[57]	AAAI 2023	Satellite Imagery	Kisumu | Malindi | Nakuru | Kenya	Poverty Prediction	-
	HDGE [88]	CIKM 2017	Mobility	Chicago	Crime Prediction | Income Prediction | House Price Prediction	-
	ZE-Mob [105]	IJCAI 2018	Mobility	New York	Functional Region Clssification	-
	MGFN [97]	IJCAI 2022	Mobility	Beijing	Predicting Willingness to Pay | Spotting Vibrant Urban Communities	-
	SceneParse [48]	AAAI 2021	Geotagged Imagery	India | Kenya	Poverty Prediction | Population Prediction | Women’s BMI Prediction	Link
	HGI [39]	ISPRS 2023	POI	Shenzhen | Xiamen	Urban Functional Distributions | Population Density Prediction | House Price Prediction	Link
	MTE [117]	GISRS2024	Trajectory	Shenzhen	Similar Location Search | Land Use Classification | Population Density Estimation	Link
	CityFM [4]	CIKM2024	OSM	Singapore | Seattle | New York	Traffic Speed Inference | Building Functionality Classification | Population Density Estimation	-
Dual View	Geo-SSL [2]	ICCV 2021	Location + Satellite Imagery	Global (Europe | America)	Geotagged Image Classification	Link
	CSP [65]	ICML 2023	Location + Satellite Imagery	New York | Tokyo | Jakart | Beijing	Location Classification | Location Visitor Flow Prediction | Next Location Prediction	Link
	SatCLIP [45]	arxiv 2023	Location + Satellite Imagery	Global	Regression: Air Temperature, Elevation, Median Income, California Housing, Population, Density | Classification: Countries, iNaturalist, Biome, Ecoregions	Link
	SatCLE [31]	WWW 2025	Location + Satellite Imagery	Global	Regression: Population, Elevation, Carbon Emission | Classification: Countries, Land Vegetation	Link
	GPS2Vec [109]	IEEE TMM 2021	Location + Geo-Tagged Imagery	Global	Venue Semantic Annotation | Geotagged Image Classification | Next Location Prediction	-
	GPS2Vec+ [108]	ACM MM 2021	Location + Geo-Tagged Imagery	Global	Venue Semantic Annotation | Geotagged Image Classification	Link
	Sphere2Vec [66]	ISPRS 2023	Location + Geo-Tagged Imagery	Global	Geotagged Image Classification	Link
	GeoCLIP [85]	NeurIPS 2023	Location + Geo-Tagged Imagery	Global	Image to GPS	Link
	GeoLLM [67]	ICLR 2024	Location + Text	Global	Population Prediction| Asset Wealth prediction | Women Edu Prediction | Sanitation Prediction | Women BMI Prediction | Population Prediction | Income prediction | Hispanic Ratio Prediction | Home Value Prediction	Link
	LLMGeovec [33]	AAAI 2025	OSM + Text	Global	Geographic Prediction | Long-term Time series Forecasting | Graph-based Spatio-Temporal Forecasting	-
	TALE [87]	TKDE 2022	Location + Trajectory	New York | Tokyo | Jakart | Beijing	Location Classification | Location Visitor Flow Prediction | Next Location Prediction	Link
	[73]	ECML-PKDD 2023	Location + Trajectory	Global	Next Location Prediction | Land Usage Classification | Transportation Mode Classification	Link
	MGeo [19]	SIGIR 2023	Location + Geographic Context	Hangzhou	Query-POI Matching | Ranking task | Retrieval task	Link
	[21]	ICDM 2019	POI + Mobility	Beijing	House Sale Amount Prediction	-
	MV-PN [26]	AAAI 2019	POI + Mobility	Beijing	Regional Mobility Popularity	Link
	CGAL [116]	KDD 2019	POI + Mobility	Beijing	Regional Mobility Popularity	-
	Region2Vec [60]	CIKM 2022	POI + Mobility	New York	Region Clustering | Popularity Prediction | Crime Prediction	-
	MVURE [113]	IJCAI 2020	POI + Mobility	New York	Land Usage Classification | Crime Prediction	Link
	ROMER [9]	CIKM 2023	POI + Mobility	New York	Land Usage Classification | Check-in Prediction	-
	HREP [122]	AAAI 2023	POI + Mobility	New York	Land Use Classification | Crime Prediction	-
	EUPAC [10]	arxiv 2024	POI + Mobility	New York	Check-in Prediction | Crime Prediction | Land Usage Classification	-
	RECP [53]	AAAI 2024	POI + Mobility	New York	Land Usage Clustering | Region Popularity Prediction	-
	ReMVC [112]	TKDE 2022	POI + Mobility	New York	Land Usage Clustering | Region Popularity Prediction	-
	URGent [36]	IEEE TCSS 2022	POI + Mobility	Beijing | Hangzhou	Traffic Prediction	-
	RegionDCL [52]	KDD 2023	POI + OSM	Singapore | New York	Land Use Prediction | Population Density Estimation	Link
	PG-SimCLR [99]	WWW 2022	Satellite Imagery + POI	Beijing	Region Similarity Analysis | Socio-Economic Prediction	Link
	MMGR [3]	ISPRS 2023	Satellite Imagery + POI	Shanghai | Wuhan	Urban Function Mapping | Population Prediction | GDP Prediction	Link
	ReFound [100]	KDD 2024	Satellite Imagery + POI	Beijing | Shanghai | Guangzhou Suzhou | Shenzhen	Urban Village Detection | Commercial Activeness Prediction | Population Prediction	-
	GeoHG [123]	arxiv 2024	Satellite Imagery + POI	Beijing | Shanghai | Guangzhou | Shenzhen	Carbon Prediction | GDP Prediction | Population Prediction | NightLight Prediction | PM2.5 Prediction	-
	UrbanCLIP [103]	WWW 2024	Satellite Imagery + Text	Beijing | Shanghai | Guangzhou | Shenzhen	Carbon Prediction | GDP Prediction | Population Prediction	Link
	Urban2Vec [94]	AAAI 2020	Satellite Imagery + POI	Bay Area | Chicago | New York	Income Prediction | Education Prediction | Recial Diversity Prediction	Link
	USPM [12]	KDD 2024	Street-view Imagery + Text	Wuhan	Street Function Prediction | Socioeconomic Indicator Prediction	-
	[50]	CIKM 2022	Satellite Imagery + Street-view Imagery	Beijing	POIs Count | Commercial Activeness | Resident Consumption Population | Economic Activity	-
Multiple View	RegionEncoder [41]	CIKM 2019	Satellite Imagery + POI + Mobility	Chicago | New York	House Sale Prediction	Link
	M3G [38]	AAAI 2021	Street-view Imagery + POI + Mobility	Chicago | New York	Crime Prediction | Bike Flow Prediction | Average Personal Income Prediction	Link
	Geo-Tile2Vec [61]	ACM TSAS 2023	Street-view Imagery + POI + Mobility	Beijing | Nanjing | Nanchang	Land Use Classification | POIs Classification | Restaurant Average Price Prediction	-
	MuseCL [110]	IJCAI 2024	Satellite Imagery + Street-view Imagery + POI + Mobility	Beijing | Shanghai | New York	Land Usage Clustering | Popularity Prediction	Link
	UrbanVLP [30]	AAAI 2025	Satellite Imagery + Street-view Imagery + Text	Beijing | Shanghai | Guangzhou | Shenzhen	Carbon Prediction | GDP Prediction | Population Prediction | NightLight Prediction | House Price Prediction | POI Prediction	Link
	KnowCL [58]	WWW 2023	Satellite Imagery + Knowledge Graph	Beijing | Shanghai | New York	Population Prediction | Economy Prediction | Crime Prediction	Link
	HAFusion [82]	ICDE 2024	POI + Mobility + Land Usage	New York | Chicago | San Francisco	Crime Prediction | Check-in Prediction	Link
	HUGAT [44]	arxiv 2022	POI + Land Usage+ Check-in + Taxi Record	New York	Crime Prediction | Check-in Prediction	-

Table 2: A summary of deep learning-based works in geospatial representation learning.

In addition to POIs and mobility data, ***Land Usage*** also serves as an auxiliary feature that can provide significant semantic information for downstream socio-economic tasks. HAFusion [82] incorporates land usage as a third perspective, in addition to POIs and mobility, to comprehensively capture region features from three distinct angles. An attentive fusion mechanism is employed to facilitate information interaction across different modalities and effectively integrate these multi-view representations.

4.2 Representation Learning Methodology

4.2.1 Location Embedding Methodology.

- **1)** ***Contrastive Learning.*** As a classical unsupervised learning method, contrastive learning [13, 34] aims to learn effective feature representations by comparing similarities and differences between data samples. Its core idea is to enable the model to pull similar samples (positive pairs) closer in the feature space while pushing dissimilar samples (negative pairs) apart, thereby capturing the underlying structure of the data. In contrastive learning, the most commonly used loss function is the InfoNCE [74] loss, which is formulated as follows:

$$L = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(z_i, z_k)/\tau)}, \tag{1}$$

where (z_i, z_j) are the embeddings of a positive sample pair, (z_i, z_k) are the embeddings of a negative sample pair, and τ is the temperature coefficient.

In addition, the triplet loss is also widely utilized, which compares three samples (anchor, positive and negative sample) to optimize the distances within the feature space. The structure of triplets allows for more flexible sample selection, especially in the case of data imbalance.

$$L = \max\left(d(A, P) - d(A, N) + \mathrm{margin}, 0\right), \tag{2}$$

where A is the anchor point, P is the positive sample, N is the negative sample. $d(A, P)$ denotes the distance between the anchor point and the positive sample. $d(A, N)$ denotes the distance between the anchor point and the negative sample. Margin is a hyperparameter that ensures separation between positive and negative samples.

Contrastive Learning, owing to its conceptual simplicity and versatile applicability, has emerged as a pivotal approach for intra-modal representation learning and inter-modal information alignment. Particularly driven by the rise of image-text multimodal paradigms [74], it has become the method of choice for integrating visual data with other modalities. In the single-view scenario, Loc2Vec [80] employs a triplet loss framework to effectively encode geo-spatial relationships and semantic similarities that characterize the surroundings of a given location. Within the dual-view context, [2, 31, 45, 65, 85] all endow digital location coordinates with semantic information through contrastive learning between locations and visual imagery.

- **2)** ***Large Language Model.*** The global perception and vast knowledge storage capabilities of Large Language Models (LLMs) have rapidly established them as an important approach for data fusion [124]. In the domain of location embedding, due to the lack of foundation models and sufficient data volumes, existing work remains focused on extracting geographical information stored within LLMs, using various prompts and meta-information [33, 67]. GeoLLM [67] endows LLM with specific geo-contextual information from the vicinity of a location (obtained from OpenStreetMap) and then uses it to predict downstream indicators like population and income. Instead of directly predicting indicators, LLMGeovec [33] acquires intermediate embeddings from LLM, which are subsequently utilized to augment time series forecasting and spatial-temporal forecasting.

4.2.2 *Region Embedding Methodology.*

- **1)** ***Contrastive Learning.*** Similar to the *Location Embedding Methodology*, Contrastive Learning within the *Region Embedding Methodology* has been widely applied across works spanning single-view, dual-view, and multi-view perspectives. In the single-view scenario, Tile2vec [40] employs triplet loss to differentiate features of neighboring and non-neighboring satellite images within a single data domain. CityFM [4] utilizes three types of contrastive objects from OSM data—nodes, polylines, and polygons—as well as relational information.

  In the dual-view setting, PG-SimCLR [99] & MMGR [3], KnowCL [58], and UrbanCLIP [103] respectively represent satellite imagery undergoing contrastive learning to align and interact information with different modalities such as POI, Knowledge Graph, and Text, with similar model architectures. Similarly, Urban2Vec [94] and USPM [12] respectively perform contrastive learning between street view images and POIs as well as text.

  In multiple view, MuseCL [110] and UrbanVLP [30] jointly incorporate satellite and street-view imagery, where the former aligns heterogeneous urban data through contrastive learning with POI data and mobility patterns, while the latter establishes semantic correlations with synthesized textual descriptions via cross-modal alignment.

- **2)** ***Graph Neural Network.*** Graph Neural Networks (GNNs) are a type of deep learning model specifically designed for processing graph-structured data, capable of effectively capturing complex relationships between nodes and non-local dependencies. The core idea of GNNs is to aggregate neighborhood information and update node representations through *Message Passing*. The detailed theoretical introduction can be found in Appendix C.

  As we discussed previously, another mainstream approach for inter-modal integration is the combination of Mobility data. Given that the traffic patterns in mobility data, such as the tidal phenomena during morning and evening rush hours, are essentially spatiotemporal diffusion processes, a graph structure emerges as a natural and optimal choice. It can model the traffic flow propagation between regions through dynamically changing edge weights. The primary differences among various methods are reflected in their distinct graph construction approaches.

  Among them, MV-PN [26], CGAL [116], Region2Vec [60], MVURE [113], and ROMER [9] all adopt a multi-view graph framework to integrate region information from multiple complementary perspectives. Specifically, MV-PN [26] utilizes an autoencoder to encode the graph network, while CGAL [116] employs Graph Convolutional Networks (GCNs) and adversarial networks. In contrast, Region2Vec [60], MVURE [113], and ROMER [9] all use Graph Attention Networks (GATs) to adaptively encode region features. MGFN [97] considers the temporal dimension during graph construction, where mobility graphs from different time steps are jointly integrated to form a multi-temporal graph. HREP [122] explores the construction of heterogeneous graphs, in which nodes maintain multiple types of edges to represent diverse relationships.

5 Application Perspective

5.1 Socioeconomic Indicators Prediction

Regional indicators include various statistical measures of a geospatial area from both environmental and social perspectives, such as *regional GDP and poverty* [30, 40, 57, 78, 94, 106, 107, 123], *crime* [88], *mobility popularity* [41, 59, 110, 116], *house prices* [21, 41, 110], *population* [22, 30, 48, 50, 110, 123], and *air quality* [30, 123]. Traditionally, data collection has relied on extensive and costly field research [68, 124], such as population censuses and air quality monitoring. However, limited human resources and budgets often hinder comprehensive data collection [25, 75], prompting the development of statistical methods to estimate regional indicators and enhance estimation accuracy.

At an early stage, the application of geospatial representation was primarily limited to urban scales and a narrow range of indicators, due to the constraints of representation models and the availability of real-world datasets. Most researches initially focused on the prediction of regional **crime rates** and **house prices**. HDGE [88] utilized taxi mobility data to estimate these rates, while RegionEncoder [41] enhanced model performance by integrating regional satellite imagery and POI data. Additionally, the multiview graph structure proposed by Du et al. [21] improved the understanding of topical relationships between regions, resulting in better predictions of house sales. These approaches have effectively provided insights for urban planning and real estate investment.

With the advancement of dual view representation, applications have expanded to include other indicators like **popularity** prediction, quantified by regional check-in counts [59, 116], and the prediction of more economic indicators [57, 78, 94, 106, 107].

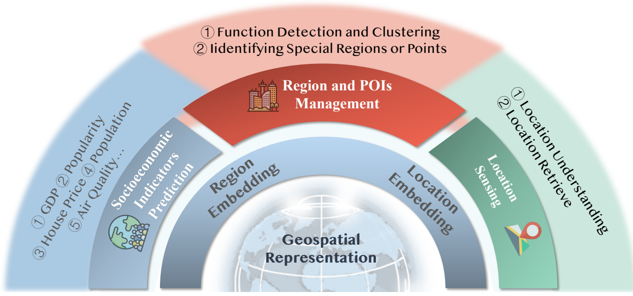

Figure 6: Taxonomy of Application for Geospatial Embedding.

To address diverse partial demands, geospatial representation in socioeconomic indicator prediction has evolved to encompass broader scales and tasks. Recent efforts, such as those by Fan et al. [22], Lee et al. [48], Li et al. [50], have demonstrated the potential for geospatial representation in predicting **poverty** and **population density** on a wider or even global scale. The fusion of multi-modal data and advanced representation learning has significantly contributed to improvements in generalization across applications. For instance, MuseCL [110] employs contrastive learning to combine features from satellite imagery, street-view imagery, POI, and mobility data, enhancing prediction accuracy. This contrastive learning approach has been utilized in several recent studies [30, 100, 103, 117, 123], improving accuracy in predictions of various regional indicators *(e.g., GDP, air quality, night light)* [117, 123]. SatCLIP [45] introduces a global location embedding to predict air temperature and population density at a global scale, while GeoLLM [67] further enhances the performance of global location embeddings by incorporating large language models (LLMs).

5.2 Region and POIs Management

The rapid dynamics of human activities render administrative boundaries and other manually designed boundaries or marks insufficient for meeting the real-world requirements of public services. The functional similarities and socio-economic connections between regions and locations are challenging to detect and quantify using traditional methods [124]. Location and region embeddings can facilitate the detection and management of POI and regional characteristics through a data-driven approach, such as automatically clustering regions into different functional groups [60, 98, 105, 114] and identifying special regions or locations [80, 92, 101, 117].

ZE-Mob [105] utilizes taxi mobility data to identify the urban functions of regions in New York City. Similarly, Wang et al. [92] employs the same type of data to detect popular zones in urban communities. Numerous unsupervised region embedding methods [3, 22, 54, 61, 112] have demonstrated impressive performance in land use clustering and detection. For example, ReFound [100] introduces a contrastive learning-based framework that efficiently detects urban villages in cities through satellite imagery and POIs.

5.3 Location Sensing

Geospatial embeddings essentially provide comprehensive information of geospace and contribute to advanced location intelligence applications like **location understanding** [73, 87] and **location retrieval** [2, 65, 66, 85, 107, 109], demonstrated in Figure 8. Location understanding help to gain more information from geospatial respective to enhance the useful features for various GeoAI tasks. For example, Wan et al. [87] enhances the representation of traffic trajectories with the pre-trained location embedding to improve the model's performance on visitor's flow prediction in cities. Park et al. [73] introduce location embedding to trajectory analysis to improve prediction accuracy of transportation modes. Many others researches like spatio-temporal prediction of urban traffic flow, regional climate variation and air quality also unitize geospatial embedding as useful basic information for learning the spatial correlations of the varies.

Location retrieval involves identifying the geographical origin of photographs or digital data, which is crucial for various location-based applications like navigation, tourism, and security [85]. The key for location retrieval is to gain clear and distinguishable location embedding for each geospatial point or region. GPS2Vec [109] and GPS2Vec+ [107] demonstrate the necessity of location embedding in geo-image classification and venue annotation. GeoCLIP [85] realizes the worldwide geo-localization through contrastive learning on CLIP framework [74].

6 Location Intelligence in the LLM Era

Despite the substantial progress in location intelligence researches recently, several persistent challenges remain, underscoring critical avenues for future exploration and innovation in the LLM era.

- Unified Benchmark, Data, and Downstream Tasks. In contrast to general domains such as Computer Vision (CV) [86] and Natural Language Processing (NLP) [69], a review of Geospatial Representation Learning reveals significant deficiencies in current location intelligence researches. These include the lack of standardized datasets, codebases, and evaluation metrics for downstream tasks, as shown in Table 2. The lack of uniformity complicates model comparisons and hampers the quantification of contributions from various data modalities, such as satellite imagery and mobility data. Additionally, suboptimal open-source practices, with many studies failing to provide comprehensive code and data, limit reproducibility and scalability. Constructing a unified benchmark encompassing data, models, and downstream tasks is a valuable direction for future development.

- Large-scale dataset and Foundation Model. Large-scale dataset and Foundation Model. In recent years, the remarkable success of Large Language Models (LLMs) [18, 71] has fundamentally transformed artificial intelligence research patterns, subsequently spurring extensive exploration of foundation models across various domains including meteorological, financial and medical area [5, 49, 89]. The daily generation of massive geospatial data in our cities presents opportunities and challenges for foundation model development, with data complexity and format variations hindering the construction of effective models [115]. Current approaches mainly involve fine-tuning LLMs [55, 56] or developing small to medium-sized pretrained models [30, 103], which are far from achieving Artificial General Intelligence (AGI) in the GeoAI domain [115].

A key challenge for future research is unifying diverse geospatial data modalities to construct large-scale datasets for training foundation models. While initiatives like CityBench [24] and CityGPT [23] have made strides using simulation platforms, a comprehensive solution still requires collaboration across the research community to leverage interdisciplinary expertise and innovative methods.

7 Conclusion

In conclusion, this survey highlights the critical role of deep learning in advancing geospatial representation learning. We provide a systematic and detailed overview of modern frameworks leveraging deep neural networks for this purpose, introducing a novel taxonomy organized along three methodological dimensions: modality, coverage, and downstream tasks. Representative studies are systematically reviewed according to this taxonomy, followed by a discussion of current limitations and promising future research directions in the LLM era.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, St. Julian's, Malta, 225–237. https://aclanthology.org/2024.eacl-srw.17/

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10181–10190.

[3] Lubin Bai, Weiming Huang, Xiuyuan Zhang, Shihong Du, Gao Cong, Haoyu Wang, and Bo Liu. 2023. Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. *ISPRS Journal of Photogrammetry and Remote Sensing* 201 (2023), 193–208.

[4] Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. 2024. City foundation models for learning general purpose representations from openstreetmap. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 87–97.

[5] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619, 7970 (2023), 533–538.

[6] Graeme Bonham-Carter. 1994. *Geographic information systems for geoscientists: modelling with GIS*. Number 13. Elsevier.

[7] Anasse Boutayeb, Iyad Lahsen-cherif, and Ahmed El Khadimi. 2024. A comprehensive GeoAI review: Progress, Challenges and Outlooks. *arXiv preprint arXiv:2412.11643* (2024).

[8] Samuele Capobianco, Leonardo M Millefiori, Nicola Forti, Paolo Braca, and Peter Willett. 2021. Deep learning methods for vessel trajectory prediction based on recurrent neural networks. *IEEE Trans. Aerospace Electron. Systems* 57, 6 (2021), 4329–4346.

[9] Weiliang Chan and Qianqian Ren. 2023. Region-wise attentive multi-view representation learning for urban region embedding. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3763–3767.

[10] Weiliang Chan, Qianqian Ren, and Jinbao Li. 2024. Enhanced Urban Region Profiling with Adversarial Self-Supervised Learning for Robust Forecasting and Security. *arXiv preprint arXiv:2402.01163* (2024).

[11] Bin Chen, Bing Xu, and Peng Gong. 2021. Mapping essential urban land use categories (EULUC) using geospatial big data: Progress, challenges, and opportunities. *Big Earth Data* 5, 3 (2021), 410–441.

[12] Meng Chen, Zechen Li, Weiming Huang, Yongshun Gong, and Yilong Yin. 2024. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 319–328.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[14] Yile Chen, Weiming Huang, Kaiqi Zhao, Yue Jiang, and Gao Cong. 2024. Self-supervised Learning for Geospatial AI: A Survey. *ArXiv* abs/2408.12133 (2024). https://api.semanticscholar.org/CorpusID:271924270

[15] Yile Chen, Weiming Huang, Kaiqi Zhao, Yue Jiang, and Gao Cong. 2024. Self-supervised Learning for Geospatial AI: A Survey. *arXiv preprint arXiv:2408.12133* (2024).

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[17] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. 2022. Vista: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5184–5193.

[18] DeepSeek-AI, Aixin Liu, Bei Feng, et al. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[19] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: Multi-modal geographic language model pre-training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 185–194.

[20] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. Article 340, 20 pages.

[21] Jiadi Du, Yunchao Zhang, Pengyang Wang, Jennifer Leopold, and Yanjie Fu. 2019. Beyond geo-first law: Learning spatial representations via integrated autocorrelations and complementarity. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 160–169.

[22] Zhuangyuan Fan, Fan Zhang, Becky PY Loo, and Carlo Ratti. 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences* 120, 27 (2023), e2220417120.

[23] Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. 2024. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. *arXiv preprint arXiv:2406.13948* (2024).

[24] Jie Feng, Jun Zhang, Junbo Yan, Xin Zhang, Tianjian Ouyang, Tianhui Liu, Yuwei Du, Siqi Guo, and Yong Li. 2024. CityBench: Evaluating the Capabilities of Large Language Model as World Model. *arXiv preprint arXiv:2406.13945* (2024).

[25] Jo Ellen Force and Gary E Machlis. 1997. The human ecosystem part II: Social indicators in ecosystem management. *Society & Natural Resources* 10, 4 (1997), 369–382.

[26] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 906–913.

[27] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. 2020. Lightweight and robust representation of economic scales from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 428–436.

[28] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to Score Economic Development from Satellite Imagery *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2970–2979. doi:10.1145/3394486.3403347

[29] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2970–2979.

[30] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2024. UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction. *arXiv preprint arXiv:2403.16831* (2024).

[31] Xixuan Hao, Wei Chen, Xingchen Zou, and Yuxuan Liang. 2025. Nature Makes No Leaps: Building Continuous Location Embeddings with Satellite Imagery from the Web. In *THE WEB CONFERENCE 2025*.

[32] Zellig S Harris. 1954. Distributional structure.

[33] Junlin He, Tong Nie, and Wei Ma. 2024. Geolocation representation from large language models are generic enhancers for spatio-temporal learning. *arXiv preprint arXiv:2408.12116* (2024).

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[35] Marco Helbich, Yao Yao, Ye Liu, Jinbao Zhang, Penghua Liu, and Ruoyu Wang. 2019. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environment international* 126 (2019), 107–117.

[36] Mingliang Hou, Feng Xia, Haoran Gao, Xin Chen, and Honglong Chen. 2022. Urban region profiling with spatio-temporal graph neural networks. *IEEE Transactions on Computational Social Systems* 9, 6 (2022), 1736–1747.

[37] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1423–1432.

[38] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y Ng, and Ram Rajagopal. 2021. Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond. *arXiv preprint arXiv:2105.02489* (2021).

[39] Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. 2023. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing* 196 (2023), 134–145.

[40] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3967–3974.

[41] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1993–2002.

[42] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. 2023. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[43] Souknilanh Keola, Magnus Andersson, and Ola Hall. 2015. Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. *World Development* 66 (2015), 322–334.

[44] Namwoo Kim and Yoonjin Yoon. 2022. Effective urban region representation learning using heterogeneous urban graph attention network (HUGAT). *arXiv preprint arXiv:2202.09021* (2022).

[45] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. 2023. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179* (2023).

[46] Pranjal Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review* 57, 10 (2024), 260.

[47] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[48] Jihyeon Lee, Dylan Grosz, Burak Uzkent, Sicheng Zeng, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Predicting livelihood indicators from community-generated street-level imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 268–276.

[49] Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. 2024. MarS: a Financial Market Simulation Engine Powered by Generative Foundation Model. *arXiv preprint arXiv:2409.07486* (2024).

[50] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. 2022. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 3282–3291.

[51] Wenwen Li and Chia-Yu Hsu. 2022. GeoAI for large-scale image analysis and machine vision: recent progress of artificial intelligence in geography. *ISPRS International Journal of Geo-Information* 11, 7 (2022), 385.

[52] Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban Region Representation Learning with OpenStreetMap Building Footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 1363–1373. doi:10.1145/3580305.3599538

[53] Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. 2024. Urban Region Embedding via Multi-View Contrastive Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8724–8732.

[54] Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. 2024. Urban Region Embedding via Multi-View Contrastive Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8724–8732.

[55] Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. 2024. OpenCity: Open Spatio-Temporal Foundation Models for Traffic Prediction. arXiv:2408.10269

[56] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5351–5362.

[57] Zhili Li, Yiqun Xie, Xiaowei Jia, Kara Stuart, Caroline Delaire, and Sergii Skakun. 2023. Point-to-region co-learning for poverty mapping at high resolution using satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14321–14328.

[58] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*. 4150–4160.

[59] Zhicheng Liu, Fabio Miranda, Weiting Xiong, Junyan Yang, Qiao Wang, and Claudio Silva. 2020. Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 808–816.

[60] Yan Luo, Fu-lai Chung, and Kai Chen. 2022. Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 4294–4298.

[61] Yan Luo, Chak-Tou Leong, Shuhai Jiao, Fu-Lai Chung, Wenjie Li, and Guoping Liu. 2023. Geo-Tile2Vec: A multi-modal and multi-stage embedding framework for urban analytics. *ACM Transactions on Spatial Algorithms and Systems* 9, 2 (2023), 1–25.

[62] David J Maguire. 1991. An overview and definition of GIS. *Geographical information systems: Principles and applications* 1, 1 (1991), 9–20.

[63] Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. 2022. A review of location encoding for GeoAI: methods and applications. *International Journal of Geographical Information Science* 36, 4 (2022), 639–673.

[64] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. 2020. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv preprint arXiv:2003.00824* (2020).

[65] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. 2023. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*. PMLR, 23498–23515.

[66] Gengchen Mai, Yao Xuan, Wenyun Zuo, Krzysztof Janowicz, and Ni Lao. 2022. Sphere2vec: Multi-scale representation learning over a spherical surface for geospatial predictions. *arXiv preprint arXiv:2201.10489* (2022).

[67] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213* (2023).

[68] André Mascarenhas, Pedro Coelho, Eduarda Subtil, and Tomás B Ramos. 2010. The role of common local indicators in regional sustainability assessment. *Ecological indicators* 10, 3 (2010), 646–656.

[69] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.

[70] National Association of City Transportation Officials and International Municipal Lawyers Association. 2019. Managing Mobility Data. https://nacto.org/wp-content/uploads/2019/05/NACTO_IMLA_Managing-Mobility-Data.pdf Accessed: 2025-01-11.

[71] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[72] Tianjian Ouyang, Xin Zhang, Zhenyu Han, Yu Shang, and Yong Li. 2024. Health CLIP: Depression Rate Prediction Using Health Related Features in Satellite and Street View Images. In *Companion Proceedings of the ACM on Web Conference 2024*. 1142–1145.

[73] Chung Park, Taesan Kim, Junui Hong, Minsung Choi, and Jaegul Choo. 2023. Pre-training contextual location embeddings in personal trajectories via efficient hierarchical location representations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 125–140.

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[75] Paul L Ringold, James Boyd, Dixon Landers, and Matt Weber. 2013. What data should we collect? A framework for identifying indicators of ecosystem contributions to human well-being. *Frontiers in Ecology and the Environment* 11, 2 (2013), 98–105.

[76] Avay Risal, Prem B Parajuli, Padmanava Dash, Ying Ouyang, and Anna Linhoss. 2020. Sensitivity of hydrology and water quality to variation in land use and land cover data. *Agricultural Water Management* 241 (2020), 106366.

[77] Gwanghwan Seong, Namwoo Kim, Seyun Kim, and Yoonjin Yoon. 2023. Multimodal Based Region Representation Learning Considering Mobility Data in Seoul. *Procedia Computer Science* 220 (2023), 251–258. doi:10.1016/j.procs.2023.03.153 The 14th International Conference on Ambient Systems, Networks and Technologies Networks (ANT) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40).

[78] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2698–2706.

[79] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*. 2618–2624.

[80] Vincent Spruyt. 2018. Loc2vec: Learning location embeddings with triplet-loss networks. *Sentiance web article: https://www. sentiance. com/2018/05/03/venue-mapping* (2018).

[81] Miriam Steurer and Caroline Bayr. 2020. Measuring urban sprawl using land use data. *Land Use Policy* 97 (2020), 104799.

[82] Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. 2024. Urban region representation learning with attentive fusion. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 4409–4421.

[83] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).

[84] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Learning to interpret satellite images in global scale using wikipedia. *arXiv preprint arXiv:1905.02506* (2019).

[85] Cepeda Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. 2024. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems* 36 (2024).

[86] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018, 1 (2018), 7068349.

[87] Huaiyu Wan, Yan Lin, Shengnan Guo, and Youfang Lin. 2021. Pre-training time-aware location embeddings from spatial-temporal trajectories. *IEEE Transactions on Knowledge and Data Engineering* 34, 11 (2021), 5510–5523.

[88] Hongjian Wang and Zhenhui Li. 2017. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 237–246.

[89] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975* (2023).

[90] Kun Wang, Hao Wu, Yifan Duan, Guibin Zhang, Kai Wang, Xiaojiang Peng, Yu Zheng, Yuxuan Liang, and Yang Wang. 2024. NuwaDynamics: Discovering and updating in causal spatio-temporal modeling. In *The Twelfth International Conference on Learning Representations*.

[91] Lingqi Wang, Shenjing He, Shiliang Su, Yu Li, Lirong Hu, and Guie Li. 2022. Urban neighborhood socioeconomic status (SES) inference: A machine learning approach based on semantic and sentimental analysis of online housing advertisements. *Habitat International* 124 (2022), 102572. doi:10.1016/j.habitatint.2022.102572

[92] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Xiaolin Li, and Dan Lin. 2018. Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 6 (2018), 1–28.

[93] Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering* 34, 8 (2020), 3681–3700.

[94] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1013–1020.

[95] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

[96] Wikipedia. 2024. Point of Interest — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Point_of_interest [Online; accessed 2025-01-11].

[97] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-Graph Fusion Networks for Urban Region Embedding. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 2312–2318. Main Track.

[98] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-graph fusion networks for urban region embedding. *arXiv preprint arXiv:2201.09760* (2022).

[99] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*. 3308–3316.

[100] Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. 2024. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3527–3538.

[101] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*. 1–10.

[102] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*. 1–10.

[103] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, 4006–4017.

[104] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*. 4006–4017.

[105] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. 2018. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

[106] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications* 11, 1 (2020), 2583.

[107] Yifang Yin, Ying Zhang, Zhenguang Liu, Yuxuan Liang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2021. Learning multi-context aware location representations from large-scale geotagged images. In *Proceedings of the 29th ACM International Conference on Multimedia*. 899–907.

[108] Yifang Yin, Ying Zhang, Zhenguang Liu, Yuxuan Liang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2021. Learning multi-context aware location representations from large-scale geotagged images. In *Proceedings of the 29th ACM International Conference on Multimedia*. 899–907.

[109] Yifang Yin, Ying Zhang, Zhenguang Liu, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2021. Gps2vec: Pre-trained semantic embeddings for worldwide gps coordinates. *IEEE Transactions on Multimedia* 24 (2021), 890–903.

[110] Xixian Yong and Xiao Zhou. 2024. MuseCL: Predicting Urban Socioeconomic Indicators via Multi-Semantic Contrastive Learning. *arXiv preprint arXiv:2407.09523* (2024).

[111] Haitao Yuan and Guoliang Li. 2021. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering* 6, 1 (2021), 63–85.

[112] Liang Zhang, Cheng Long, and Gao Cong. 2022. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9031–9036.

[113] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4431–4437.

[114] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4431–4437.

[115] Wejia Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. 2024. Urban Foundation Models: A Survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[116] Yunchao Zhang, Yanjie Fu, Pengyang Wang, Xiaolin Li, and Yu Zheng. 2019. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1700–1708.

[117] Yu Zhang, Weiming Huang, Yao Yao, Song Gao, Lizhen Cui, and Zhongmin Yan. 2024. Urban region representation learning with human trajectories: a multi-view approach incorporating transition, spatial, and temporal perspectives. *GIScience & Remote Sensing* 61, 1 (2024), 2387392.

[118] Yan Zhang, Yong Li, and Fan Zhang. 2024. Multi-level urban street representation with street-view imagery and hybrid semantic graph. *ISPRS Journal of Photogrammetry and Remote Sensing* 218 (2024), 19–32.

[119] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[120] Yunxiang Zhao, Jianzhong Qi, Bayu D Trisedya, Yixin Su, Rui Zhang, and Hongguang Ren. 2023. Learning region similarities via graph-based deep metric learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 10237–10250.

[121] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 1–55.

[122] Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. 2023. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4981–4989.

[123] Xingchen Zou, Jiani Huang, Xixuan Hao, Yuhao Yang, Haomin Wen, Yibo Yan, Chao Huang, and Yuxuan Liang. 2024. Learning Geospatial Region Embedding with Heterogeneous Graph. *arXiv preprint arXiv:2405.14135* (2024).

[124] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion* 113 (2025), 102606.

Appendix

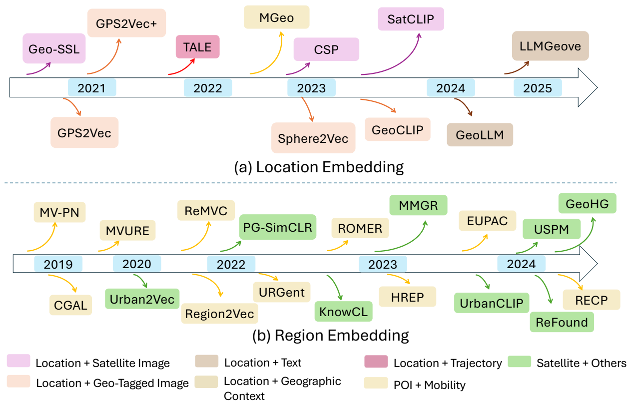

Figure 7: The roadmap of Dual View.

A Paper Orgnization

Apart from the Introduction, the rest of this paper is structured as follows: Section 2 introduces the definitions of various fundamental data formats and provides an overview of geospatial representation learning. Section 3 elaborates on a range of specific data modalities and their respective applications within the context of this work. In Section 4, a taxonomy of deep learning methodologies for geospatial representation learning is proposed, categorized into three perspectives: single-view, dual-view, and multi-view approaches. Section 5 consolidates a diverse array of application scenarios, while Section 6 highlights promising research directions and unresolved challenges for future exploration in the LLM era. Finally, Section 7 concludes this survey with a summary of key insights.

B More Details in Methodology Perspective

B.1 Other modalities in Single View Region Embedding

Other modalities [4, 39, 48, 117] that capture spatial contexts and dynamic characteristics have also been explored and utilized for modeling socio-economic attributes. However, these approaches have not yet gained widespread adoption. HGI [39] exclusively leverages POI data to model regional features, employing GCNs to hierarchically aggregate POI embeddings from the region to city level. MTE [117] represents trajectories via transition, spatial, and temporal views, effectively capturing socio-economic characteristics and aiding land use type prediction.

B.2 More information about POI + Mobility pattern in Dual View Region Embedding

Although most work is modeled based on graph structures, notable exceptions include ReCP [53] and ReMVC[112], which deviate from graph-based methods by representing both dynamic and static attributes of regions using POI distributions and inflow/outflow counts, enhanced by contrastive learning.

B.3 More information about Street-view Imagery pattern in Dual View Region Embedding

The advantage of street-view imagery lies in its ability to provide fine-grained semantic information at the location level. Urban2Vec [94] models POI data using a bag-of-words [32] approach and employs contrastive learning to align it with features derived from street-view imagery. Aiming at urban street profiling, USPM [12] integrates street imagery with textual information and employs semi-supervised graph learning based on spatial topology. [50] comprehensively investigates the distinct roles and complementary functions of visual modalities (e.g. satellite imagery and street-view imagery) at various levels in urban region representation learning.

C More Details in Representation Learning Methodology

C.1 More information about Graph Neural Network

Graph Convolutional Network (GCN) is one classical GNN that propagates information through a normalized adjacency matrix.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{3}$$

where $\tilde{A} = A + I$ denotes the adjacency matrix with added self-loops, $\tilde{D}$ is the degree matrix, $W^{(l)}$ represents the learnable parameters of $l - th$ layer, σ is the activation function.

Furthermore, Graph Attention Network (GAT) is another important type of GNN, which introduces an attention mechanism to compute the attention weights of nodes with respect to their neighbors.

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} W^{(l)} h_j^{(l)}\right), \tag{4}$$

where the attention coefficients are given by

$$\alpha_{ij} = \mathrm{softmax}\left(\mathrm{LeakyReLU}(a^T[Wh_i||Wh_j])\right), \tag{5}$$

where a is a learnable attention vector, and || denotes vector concatenation.

D More Details in Application Perspective

In figure 8, we provide an intuitive illustration of the concepts introduced in Section 5.3.

Figure 8: Application of Geospatial Embeddings for Location Sensing.