UNIVERSITY OF LEEDS

This is a repository copy of *Self-supervised representation learning for geospatial objects: A survey*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/226703/

Version: Accepted Version

Article:
Chen, Y., Huang, W. orcid.org/0000-0002-3208-4208, Zhao, K. et al. (2 more authors) (2025) Self-supervised representation learning for geospatial objects: A survey. Information Fusion, 123. 103265. ISSN 1566-2535

https://doi.org/10.1016/j.inffus.2025.103265

This is an author produced version of an article published in Information Fusion, made available under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Self-Supervised Representation Learning for Geospatial Objects: A Survey

Yile Chen[a], Weiming Huang[b], Kaiqi Zhao[c], Yue Jiang[a], Gao Cong[a]

[a]*College of Computing and Data Science, Nanyang Technological University, Singapore*
[b]*Department of Physical Geography and Ecosystem Science, Lund University, Sweden*
[c]*School of Computer Science, University of Auckland, New Zealand*

Abstract

The proliferation of various data sources in urban and territorial environments has significantly facilitated the development of geospatial artificial intelligence (GeoAI) across a wide range of geospatial applications. However, geospatial data, which is inherently linked to geospatial objects, often exhibits data heterogeneity that necessitates specialized fusion and representation strategies while simultaneously being inherently sparse in labels for downstream tasks. Consequently, there is a growing demand for techniques that can effectively leverage geospatial data without heavy reliance on task-specific labels and model designs. This need aligns with the principles of self-supervised learning (SSL), which has garnered increasing attention for its ability to learn effective and generalizable representations directly from data without extensive labeled supervision. This paper presents a comprehensive and up-to-date survey of SSL techniques specifically applied to or developed for geospatial objects in three primary vector geometric types: *Point*, *Polyline*, and *Polygon*. We systematically categorize various SSL techniques into predictive and contrastive methods, and analyze their adaptation to different data types for representation learning across various downstream tasks. Furthermore, we examine the emerging trends in SSL for geospatial objects, particularly the gradual advancements towards geospatial foundation models. Finally, we discuss key challenges in current research and outline promising directions for future investigation. By offering a structured analysis of existing studies, this paper aims to inspire continued progress in integrating SSL with geospatial objects, and the development of geospatial foundation models in a longer term.

Keywords: Geospatial artificial intelligence, Spatial data mining, Self-Supervised learning, Spatial representation learning, Geospatial foundation models

1. Introduction

The digitization of urban and territorial environments has significantly expanded the collection of extensive geospatial data associated with various objects on our planet, including road segments, buildings, neighborhoods, etc. The vast repository of data serves as the foundation of smart city applications, such as spatial keyword search [1, 2], location-based services [3, 4], geospatial knowledge graph [5], intelligent transportation systems [6, 7], and socioeconomic indicator prediction [8, 9]. Despite the richness of geospatial data, its effective mining and utilization remain challenging. In particular, a fundamental limitation arises from the task-specific nature of many deep learning models. These models are typically trained using supervised learning in specific tasks with abundant domain-specific labeled datasets (e.g., traffic records), which can be limited or costly to acquire and restricted by data privacy regulations [10]. Additionally, while various urban tasks exhibit intrinsic commonalities, such as the close relationship between population density and land use patterns, these models usually suffer from limited generalization across downstream applications due to their reliance on task-specific supervision signals and specialized model architectures. This lack of adaptability restricts the broader applicability of previous approaches across different geospatial analytics tasks, which entails the need for more flexible and generalizable learning paradigms for a variety of geospatial analyses.

In response to these challenges, self-supervised learning (SSL) [11] has emerged in recent years as a promising paradigm that reduces the dependency on annotated labels while producing task-agnostic and general-purpose data representations. The core principle of SSL is to extract transferrable knowledge from the target data through well-designed self-supervised tasks (i.e., pretext tasks), wherein the supervision signals are automatically generated from the data itself. SSL has achieved notable success across various domains and diverse data modalities, including images [12, 13], videos [14], language [15, 16], graphs [17, 18], time series [19], etc. For example, massive text corpora are structured in an autoregressive generation framework, which is well-suited for next token prediction for the training of large language models (LLMs) [20]. In addition, images are processed through data augmentation operations to produce multiple views, with models trained to produce invariant representations across views for computer vision tasks [12].

A key motivation for applying SSL techniques in the geospatial domain is to learn effective and generalizable representations (embeddings) for various forms of geospatial objects, such as points-of-interest (POI), road segments, and urban regions. These objects underpin a variety of human activities within urban environments, and therefore serve as fundamental analytical units for numerous urban analytical applications. For ex-

ample, road networks are critical infrastructures that support human movement activities within a city. As a result, a variety of tasks, such as the prediction of traffic speed, congestion, and destination, commonly regard individual road segments as analytical units. In these scenarios, SSL offers a powerful framework for deriving general-purpose representations of geospatial objects through fusing information from multiple perspectives, capturing both the intrinsic characteristics of each object and the complex interplay among different objects. The representations learned through SSL can be readily leveraged to train simpler models (e.g., a linear model) for various downstream tasks while maintaining effective performance.

Apart from the strong generalization capabilities, the interest in applying SSL techniques is also driven by the ability to work without extensive labeled datasets. This advantage addresses a long-standing challenge in the geospatial domain– the scarcity of task-specific labeled data. As a result, SSL presents a viable alternative to the conventional supervised learning paradigm, which necessitates the development of specialized deep learning models trained on sufficiently labeled data for each downstream application. By leveraging the self-supervised signals derived from data itself, SSL has the potential to better tackle diverse geospatial (urban) analytical applications.

However, geospatial objects, embedded within geographic spaces, exhibit various forms and spatial attributes while adhering to geographical principles, such as the Laws of Geography [21]. These spatial characteristics introduce significant challenges when applying standard SSL techniques developed for other domains, as they often fail to capture the intricate spatial semantics attached to geospatial objects. Besides, certain SSL components, such as data augmentation and pretext tasks, must be carefully designed and adapted to preserve the spatial and structural integrity of geospatial objects. Unlike in vision or language domains, where augmentations such as cropping, rotation, or synonym replacement are widely used, augmentations for spatial objects must ensure consistency with geographical relationships and dependencies. Moreover, geospatial objects are often associated with heterogeneous context information, capturing diverse yet complementary aspects of their intrinsic properties. Integrating this information requires the development of effective modeling and fusion strategies to produce high-quality data representations. Considering these unique challenges in the geospatial domain, recent research has introduced novel SSL techniques that incorporate spatial awareness and domain-specific constraints, ensuring the effective adaptation of SSL within the geospatial context.

Despite the growing body of literature, the application of SSL within the geospatial domain remains insufficiently discussed and summarized. To bridge this gap, this survey provides a comprehensive and systematic review of up-to-date SSL techniques tailored or developed for learning geospatial object representations, which in turn facilitate various geospatial analyses. In particular, we focus on three primary geospatial data types, categorized based on their geometric forms: points, polylines, and polygons. We adopt a structured framework to present the specialized SSL studies for these data types, focusing particularly on methods that operate independently of specific tasks and supervised settings. For each data type, we analyze how SSL techniques encode intrinsic attributes while integrating heterogeneous context information associated with geospatial objects. Besides, we review studies that utilize SSL to fuse multiple geospatial data types for learning representations. Building on the insights from the surveyed studies, we identify and analyze emerging trends in the development of geospatial foundation models. Furthermore, since SSL techniques serve as a universal paradigm that can also be utilized to enhance the supervised models as auxiliary objects in geospatial applications such as spatio-temporal forecasting, we provide a brief overview of task-specific SSL implementations applied to several domain-specific scenarios, supplementing the research landscape of SSL in this field. This survey aims to cover a wide range of model scopes, from specialized SSL models to advanced foundation models, applied to different geospatial data types and analyses. The main contributions of this survey are summarized as follows:

- We present a detailed and up-to-date review of SSL techniques for geospatial objects, focusing on three types of geospatial data, mainly in urban environments: *Point*, *Polyline*, and *Polygon (Region)*. To the best of our knowledge, this work is the first to systematically discuss SSL techniques tailored for learning representations in the geospatial domain.
- We introduce a comprehensive and structured way for specialized SSL models designed for the studied data types. Our categorization includes an analysis of intrinsic characteristics and heterogeneous context information within each data type, encompassing discussion on both predictive and contrastive SSL implementations.
- We review several recent advancements based on foundation models and task-specific SSL techniques for geospatial data, providing insights into the emerging trends.
- We discuss several key challenges for SSL in geospatial domain, and propose potential future directions to advance the related research.

Related Surveys and Our Distinction: Several recent surveys have discussed the application of SSL, mainly focusing on other domains, such as general SSL [11], SSL for computer vision [13], graphs [17, 18], time series [19], and recommender systems [22]. However, despite SSL's growing importance in data-driven analytics, its adaptation in the geospatial domain remains underexplored. Recognizing this limitation, our survey presents a comprehensive and systematic review of SSL tailored for geospatial data, which serves as the foundations in numerous downstream applications to enhance urban intelligence. On the other hand, several recent surveys have paid attention to spatio-temporal data analytics with varying emphases, such as trajectory data mining [23, 24], urban foundation models [25], geo-location encoding [26], and supervised or generative deep learning [27, 28]. While these works contribute to the broader field of GeoAI and may overlap with

some studies discussed in this paper, our survey distinguishes itself by structuring knowledge from the unique perspective of SSL. Specifically, we provide a detailed summary of SSL developments across different geospatial data types, systematically categorizing existing methodologies and synthesizing their implementation in geospatial contexts.

Paper Structure: The rest of this survey is organized as follows. Section 2 provides definitions, preliminary concepts, and background knowledge necessary for the subsequent sections. Section 3, 4, and 5 look into the details of specialized SSL techniques applied to data with three distinct geometric types: points, polylines and polygons, respectively. For each data type, representative data instances are further elaborated in the context of SSL, including POI for points, trajectory and road network for polylines, and region for polygons. Section 6 presents SSL techniques for the integration of multiple data types. Section 7 presents our discussion on the emerging trends based on the development of geospatial foundation models, and provides an overview for task-specific SSL techniques for geospatial data. Section 8 provides several promising future research directions. Finally, Section 9 concludes this paper.

2. Definition and Background

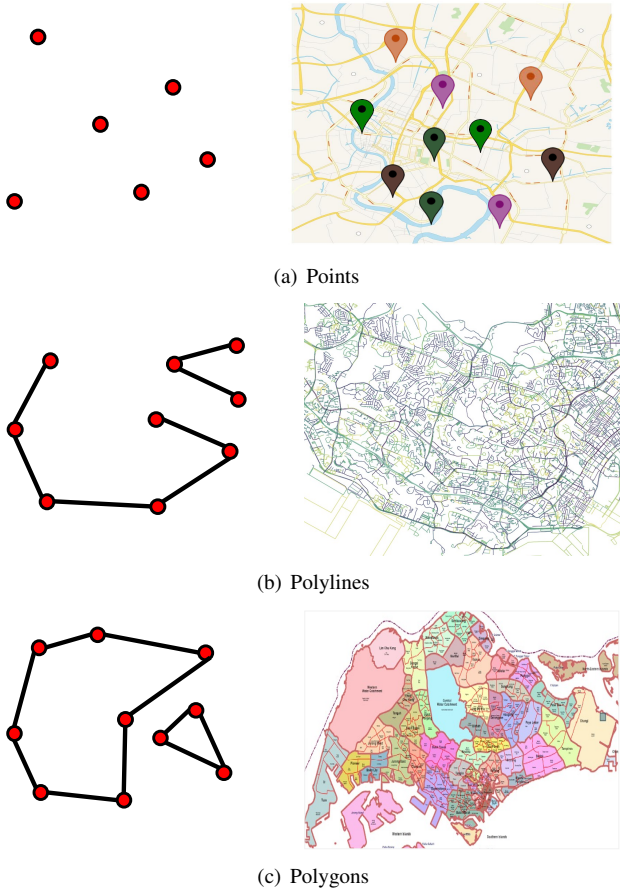
In this section, we introduce the definition of three types of geospatial vector data examined in this paper. Then we present the paradigms of SSL applied to geospatial objects based on the traits of pretext tasks. Last, we discuss preliminaries on typical models that encode these geospatial data types.

2.1. Definition of Geospatial Data Types

In this survey, we adopt the widely recognized classification scheme in spatial database research, which categorizes geospatial objects by their geometric representations into three types: *Point*, *Polyline*, and *Polygon (Region)* [29]. This scheme is well-suited to our study, as each type corresponds to distinct real-world geospatial phenomena that require different representation strategies. The three data types are illustrated in Figure 1. The formal definitions of the three data types are provided as follows.

Definition 1 (Point). A geospatial point object is represented as $p = (l, x)$, where l denotes the geographical coordinates, and x refers to the associated features of the point, such as attributes or readings. This data type indicates the spatial locations equipped with contextual information, applicable to data instances such as POIs and sensor measurements.

Definition 2 (Polyline). A geospatial polyline object is defined as a sequence of connected line segments, represented by a list of vertices $\mathcal{L} = [(l_1, x_1), ..., (l_n, x_n)]$, where l_i denotes the geographical coordinates of the i-th point, and x_i denotes its associated features, such as timestamps or semantic tags. This data type captures the sequential nature and directionality of spatial paths, applicable to data instances such as trajectories and road networks.



(a) Points

(b) Polylines

(c) Polygons

Figure 1: Three types of geospatial objects and their data instances.

Definition 3 (Polygon). A geospatial polygon object is defined as a closed shape (region) consisting of a sequence of line segments that connect to enclose an area, denoted by $\mathcal{B} = [(l_1, l_2), ..., (l_{n-1}, l_n), (l_n, l_1)]$, where l_i represents the geographical coordinates of the i-th vertex. The vertices are connected sequentially, with the first vertex repeated as the last to close the polygon. By default, the line segments of a polygon are arranged to ensure that it does not intersect with itself, maintaining a simple, closed loop. This data type can be employed to describe administrative regions or subzones in urban spaces. Therefore, research on polygons often extends its focus beyond its geometric configuration, placing emphasis on the semantic patterns exhibited by objects within or close to its enclosed area.

2.2. Paradigms for Self-supervised Learning

Geospatial data instances in practical scenarios can be conceptualized as a combination of geometric forms with intrinsic attributes (e.g., spatial coordinates) and associated context information (e.g., textual content). Therefore, SSL process in geospatial context aims to integrate the geometric form of the target geospatial object and its contextual information to derive effective representations. These representations are learned through the training objectives (pretext tasks), where supervision signals are automatically extracted from the geospatial data itself, eliminating the need for additional label annotations. Based on the design of pretext tasks, SSL techniques for geospa-

tial objects can be divided into two categories: predictive and contrastive methods.

2.2.1. Predictive Methods

Predictive methods employ pretext tasks that are formulated as prediction problems, with objectives derived from the original data instances. Specifically, these methods involve tasks like the reconstruction of corrupted geospatial objects using a subset of available data, or the prediction of auxiliary labels that are readily extracted from the attributes or structures of geospatial objects. They can be formulated as:

$$f_\theta^*, p_\phi^* = \arg\min_{f_\theta, p_\phi} \mathcal{L}_{pre}\left(p_\phi\left(f_\theta(\mathcal{D}, \mathcal{D}_c)\right), \mathcal{D}_t\right) \tag{1}$$

where f_θ and p_ϕ represent the geospatial encoder and the pretext decoder respectively. The geospatial encoder, which is introduced in Section 2.3, is responsible for deriving representations for geospatial objects, while the pretext decoder is typically a lightweight structure, such as a shallow multi-layer perceptron (MLP), to map the representations to the space of the prediction objectives. $\mathcal{D}$ denotes the target geospatial objects, which can be of any data types. $\mathcal{D}_c$ is the context information associated with $\mathcal{D}$, and $\mathcal{D}_t$ denotes the prediction objectives, which could either be the original data instance for reconstruction tasks, or additional features excluded from the encoder input for auxiliary label prediction. $\mathcal{L}_{pre}$ is the loss function that measures the prediction error, such as cross-entropy loss or mean squared error (MSE) loss.

2.2.2. Contrastive Methods

Contrastive methods are based on the principle of maximizing agreement between different views generated from the same data instance. Specifically, these methods aim to pull closer the representations of positive view pairs, which are derived from various data augmentation operations of the same data instance, while pushing apart the representations of negative view pairs from different data instances. They can be formulated as:

$$f_\theta^*, p_\phi^* = \arg\min_{f_\theta, p_\phi} \mathcal{L}_{con}\left(p_\phi\left(f_\theta(\tilde{\mathcal{D}}^1, \tilde{\mathcal{D}}_c^1)\right), p_\phi\left(f_\theta(\tilde{\mathcal{D}}^2, \tilde{\mathcal{D}}_c^2)\right)\right) \tag{2}$$

where f_θ and p_ϕ represent the geospatial encoder and the pretext decoder respectively. The pretext decoder is usually a projection head for linear transformation [12]. $\tilde{\mathcal{D}}^1$ and $\tilde{\mathcal{D}}^2$ are two distinct views generated from the target geospatial objects $\mathcal{D}$, which can belong to any geospatial data types, and $\tilde{\mathcal{D}}_c^1$ and $\tilde{\mathcal{D}}_c^2$ are the context information associated with these respective views. $\mathcal{L}_{con}$ is the contrastive loss function that quantifies the degree of agreement, typically measured by mutual information estimator [11], such as InfoNCE [30], JS divergence [31] and triplet loss [32].

2.3. Preliminaries on Geospatial Encoder

Given the diversity of geospatial data types discussed in this survey, each with its unique geometric (locational) and intrinsic properties, the utilized geospatial encoder f_θ would be varied to accommodate their distinct characteristics. Therefore, we provide a brief introduction on several neural network modules frequently employed or adapted as geospatial encoders.

2.3.1. Graph Neural Networks

Graph Neural Networks (GNNs) [33] correspond to a type of neural network architectures designed to process graph-structured data, aiming to capture the complex relationships and structures within the graph. GNNs employ message-passing operations iteratively on the graph, where the representation of a node v is updated through interactions with its neighbors. This process can be expressed as:

$$h_v^{(l)} = \mathcal{F}^{(l)}\left(h_v^{(l-1)}, \mathrm{AGG}^{(l)}\left(\left\{h_u^{(l-1)}\right\}_{u\in\mathcal{N}_v}\right)\right) \tag{3}$$

where $h_v^{(l)}$ indicates the representation of v at layer l and $\mathcal{N}_v$ denotes the neighbors of v. $\mathrm{AGG}^{(l)}$ is the message aggregation function at layer l, which collects and combines node features, and potentially edge features, from the neighbors, and $\mathcal{F}^{(l)}$ is the function that updates the representation of v based on the aggregated information. For geospatial objects, GNNs are frequently utilized to model discrete objects, enabling the capture of complex relationships among them.

2.3.2. Sequential Models

Sequential models are designed to process input data composed of sequences, which include domains such as time series, text, audio, and video. Over the past decade, neural network architectures have exhibited exceptional performance in sequence modeling due to their capability of capturing dependencies effectively. The general process can be described as:

$$[h_1, ..., h_n] = \mathcal{F}\left([x_1, ..., x_n]; \Theta\right) \tag{4}$$

where $[x_1, ..., x_n]$ denotes the input sequence, $[h_1, ..., h_n]$ denotes the hidden representations output by the sequential model $\mathcal{F}$, which is parameterized by Θ. Recurrent Neural Networks (RNNs) accomplish this by recursively processing the current input along with previous elements of the sequence, where the previous elements are encoded into internal hidden states, leading to several model variants, with GRU [34] and LSTM [35] being the most notable ones. In recent years, the Transformer architecture [36] has revolutionized sequence modeling by handling historical sequences in a parallel manner, instead of the recursive approaches. Meanwhile, it demonstrates superiority of modeling pairwise relationships between any two positions in a sequence through self-attention mechanism. For geospatial objects, sequential models are particularly valuable for modeling trajectories or data instances that are built to consider the sequential dependencies.

2.3.3. Pre-trained Models

The evolution of advanced sequential models, especially those based on the Transformer architecture, have marked the milestone in the development of pre-trained language models. One popular paradigm is to adhere to the principles set by BERT [37]. These models [38, 39, 40, 41] leverage large-scale datasets

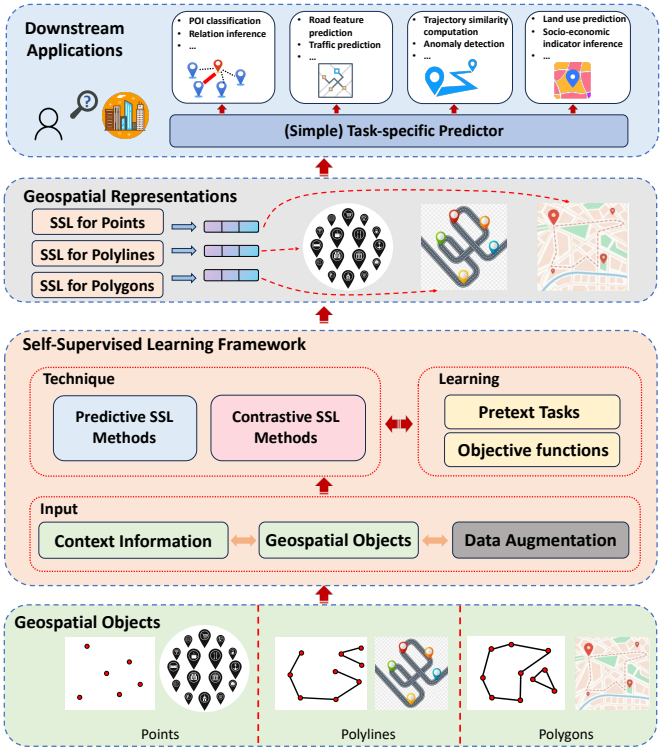
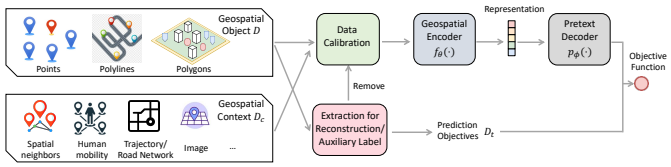
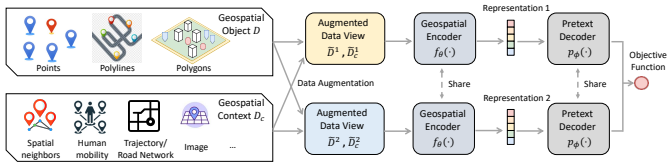


Figure 2: Overview of SSL framework for geospatial objects.



(a) Predictive SSL methods



(b) Contrastive SSL methods

Figure 3: Illustration of two types of SSL methods for geospatial objects.

to employ the training objective of masked language modeling (MLM). This process enables the acquisition of rich and transferable representations, which can be fine-tuned for specific tasks with less labeled data. Another paradigm is to employ training with autoregressive language modeling, exemplified by ChatGPT and its counterparts [42, 43, 44, 45, 46]. These large language models (LLMs) demonstrate strong capabilities in language understanding and reasoning, transforming various tasks into the process of autoregressive language generation. The robust performance of these two paradigms has led to the widespread use of pre-trained language models to encode textual information associated with geospatial data or to adapt their training objectives to develop specialized representations. On the other hand, in scenarios where vision information, such as street view images, is associated with geospatial elements like polylines or polygons, pre-trained visual models are also utilized. These include CNN-based models [47, 48] and recent Transformer-based models [49, 50] trained on large-scale image datasets. When both textual and visual data sources are available, visual-language pre-trained models such as CLIP and its variants [51, 52] are employed to synergize the semantics of the two data modalities, enhancing the interpretability and utility of combined data sources in geospatial applications.

It is important to note that these representative models are not mutually exclusive, and they can be employed concurrently to encode geospatial objects through model combination or to incorporate diverse contextual information. In the following sections, we adopt a structured analytical approach for each geospatial data type under consideration. We first present the encoding methods for intrinsic attributes that reflect the fundamental characteristics of the data instance. Then we discuss methods to integrate and fuse heterogeneous context information to enhance the extraction of meaningful insights and knowledge for the studied geospatial object. Finally, we introduce the downstream applications that benefit from the derived geospatial representations.

2.4. *Overview and Taxonomy*

Figure 2 provides an overview of the Self-Supervised Learning (SSL) framework for geospatial objects, illustrating the pipeline from raw data to downstream applications. At the foundation, geospatial objects are categorized into three primary geometric data types–points, polylines, and polygons–each corresponding to different data instances such as POIs, road segments, trajectories, and urban regions. These objects and their context information and augmented variations serve as input to the SSL framework. Within this framework, predictive methods focus on recovering missing or masked attributes, while contrastive methods aim to learn representations by distinguishing between similar and dissimilar instances. These techniques are guided by specific pretext tasks and objective functions, such as the cross-entropy for attribute classification, or the InfoNCE loss for contrastive learning. Lightweight task-specific models then utilize the learned representations to support various geospatial applications, such as POI classification, traffic prediction, anomaly detection, socio-economic inference, etc.

The taxonomy of SSL methods discussed in the paper is presented in Figure 4. It is organized according to data instance and the contextual information exploited for representation learning. The taxonomy comprises four types of geospatial data instances, each grouped according to the specific context information they utilize. We classify SSL methods within each branch into predictive, contrastive, or hybrid (predictive & contrastive) methods. To enhance conceptual clarity, Figure 3 provides a schematic illustration for predictive and contrastive SSL methods, as introduced in Section 2.2, applied across different geospatial objects. This illustration outlines the core modeling workflows and highlights the general design principles underlying each SSL paradigm. Building on these paradigms, each

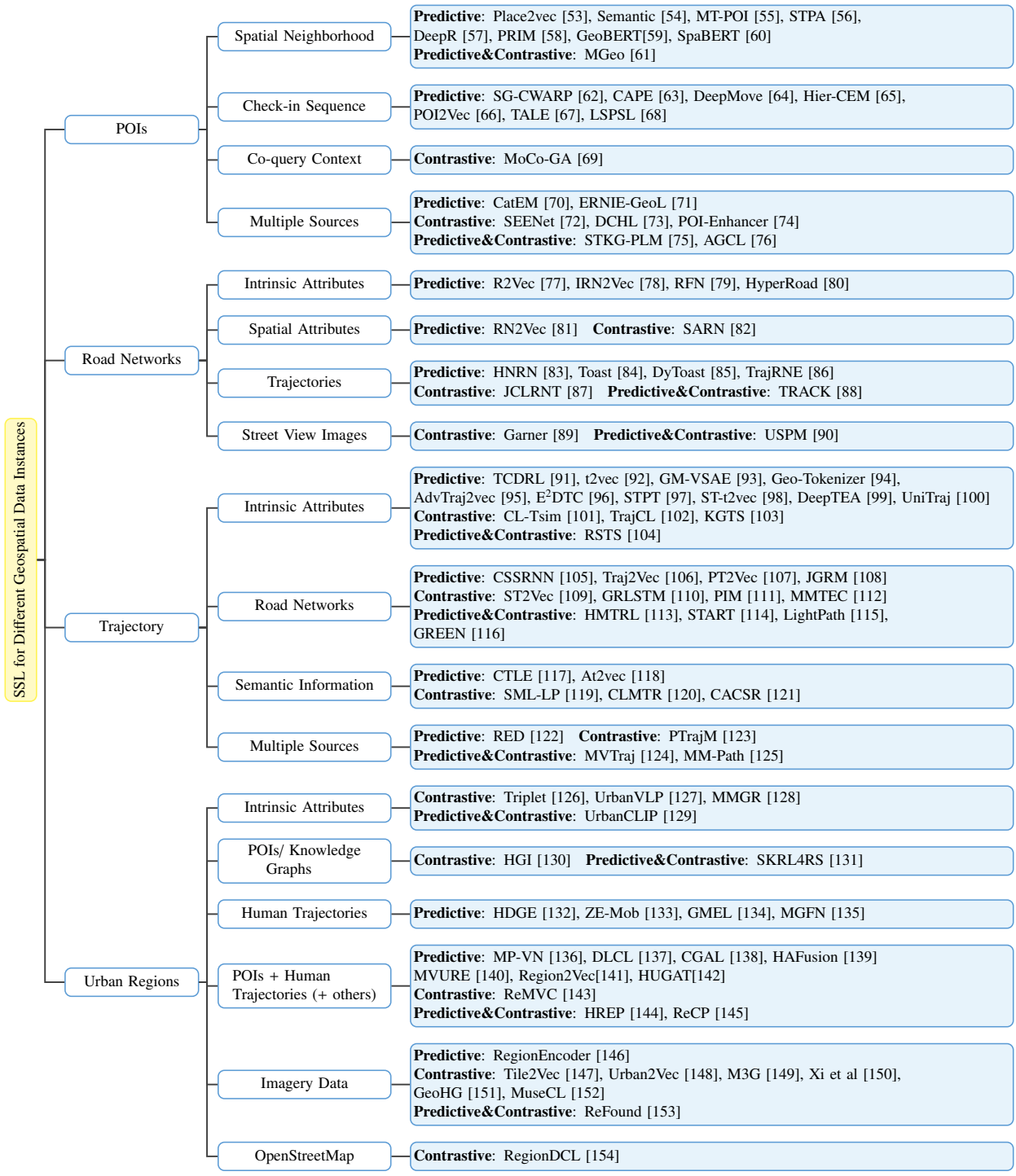


Figure 4: Taxonomy of SSL methods for geospatial data instances discussed in the paper. It is categorized by the type of context information leveraged to for encoding each data instance.

geospatial data type incorporates tailored designs for the constituent modules and data sources within its respective learning pipeline. In the following sections, we present a detailed discussion of the methods under each branch of the taxonomy.

3. Points

Points are the most fundamental geospatial objects, forming the basic component of polylines and polygons. Each point is associated with a geo-location and associated features, such as text descriptions and categories. Examples of point data include location-based point sensors, road network intersections, and points-of-interest (POIs). While extensive research has been

conducted on modeling point data for forecasting tasks related to traffic, weather, and environment [155, 156], most studies adopt an end-to-end training method to improve forecasting accuracy, with no emphasis on representation learning. Recent research on self-supervised representation learning techniques for point data focuses primarily on POI data due to their rich semantic information. Consequently, this survey focuses on POI data, which plays a crucial role in understanding user mobility and urban functionality. The surveyed studies for POI data are listed in Table 1.

3.1. Points of Interest

3.1.1. Intrinsic Attributes

POIs refer to the semantic locations in location-based services that users might want to visit. Examples of POIs include restaurants, stores, schools, etc. Each POI is associated with geographical coordinates l (i.e., geo-location) and usually a number of features $x = \{name, categories, reviews\}$, including the POI's name, one or multiple categories, and possibly a set of user reviews.

Since the geo-location and features are in different formats, existing methods [61, 5, 60] often employ separate encoders to extract essential information from geo-location [26], categories, and other text features [37]. The encoded features are fused in a subsequent model, such as multi-layer perceptrons (MLP) or attention-based models, to obtain the POI representations [61, 5, 60]. The objective of SSL for POI is to learn an encoder f_θ that can acquire a d-dimension latent representation of any input POI p, denoted by $\mathbf{e}_p = f_\theta(p) \in \mathbb{R}^d$, such that p's geo-location l and features x, as well as p's context are well preserved in $\mathbf{e}_p$. Based on this objective, we note that scenarios of (next) POI recommendation [157] are more related to specific tasks with supervised signals rather than the SSL setting. Therefore, we do not focus on these methods in this section.

3.1.2. Context Information

The intrinsic attributes only include the information of individual POIs. POIs are often located in spatial environments where diverse types of POIs are present at different distance ranges. Besides, users' check-ins or queries related to POIs indicate the dependency on the urban functions of distinct POIs. Consequently, most SSL research for POIs has been dedicated to modeling the rich context information of POIs, including spatial neighborhoods, check-in sequences, co-query context, and temporal context.

Spatial Neighborhood. To effectively capture the spatial context of a POI, Skip-gram [53], GNNs [56, 57, 58] and masked language modeling (MLM) [59, 61, 60, 71] have been explored. The skip-gram methods consider the spatial neighbors as the context of a POI, analogous to the surrounding words in Word2vec [158]. They minimize the errors in predicting the spatial context using the target POI embeddings. For instance, Place2vec [53] partitions the spatial neighborhood of a target POI into equal-distance bins and calculates a bin-wise boosting factor to increase the occurrence of popular POI categories in each bin. In this way, POI categories with a higher boosting factor will contribute more to the skip-gram learning objective. In this vein, Huang et al. [54] propose to preserve the hierarchical structure of POI categories in learning POI category embeddings. For example, the categories Japanese Restaurants and Chinese Restaurants should be similar in the embedding space due to their resemblance of functions and the fact that they often share the same generic category, e.g., Food. They employ Laplacian eigenmaps as regularization terms to pull together the POI categories that share the same ancestor categories. This method is later extended in [55] and [70] to consider categories with close semantics in different years and the check-in sequences, respectively.

GNN-based methods construct a graph where two POIs are connected by an edge if they are close in spatial. Leveraging the ability of GNNs to acquire the local structure of a graph, these methods encode the spatial neighborhood of a POI into its latent representation. Predictive SSL methods have been proposed to predict either the intrinsic attributes of the POI or the relation between POIs. STPA [56] constructs a Delaunay triangulation graph based on the distance between POIs. Each node in the graph is a POI, represented by a one-hot encoding of its category. GNN is applied to acquire a target POI's representation by aggregating the category information from its neighbors in the Delaunay triangulation graph. Subsequently, a predictive objective is employed to predict the category of the target POI, given its latent representation. For objectives that predict the relations (competitive or complementary) between POIs, the relations between POIs are constructed based on certain heuristic rules. DeepR [57] builds the competitive relations between two POIs if they are spatially close and frequently co-occur in the same query. DeepR employs a Heterogeneous POI Information Network (HPIN) to represent POI, brands, aspects, and their relations, and a spatial adaptive GNN to acquire the POI representations from neighbors. Given the representations of two POIs, it predicts whether they are in a competitive relation. PRIM [58] constructs competitive relations between different categories of POIs if they frequently co-occur in the same query. Likewise, it constructs complementary relations between POIs in the same category that frequently co-occur in the same query. To predict the relations, PRIM employs the POI representations obtained through a GNN that gives importance to the neighbors based on spatial distance and feature similarity.

Besides, numerous studies leverage the MLM idea applied in pre-trained language models to learn POI representations [59, 60, 61]. The common idea behind MLM-based methods is to construct pseudo sentences from the spatial neighborhoods and then apply MLMs to the pseudo sentences. GeoBERT [59] divides the digital map into grids and proposes two methods to create a pseudo sentence from POIs residing in each grid cell. The first method creates the pseudo sentence by finding the shortest path between the two farthest POIs, passing through all POIs in the same grid cell. The second method returns the ordered sequence of POIs by their distance to the grid center. With the obtained pseudo sentences, GeoBERT treats each POI as a token and adopts the same training mechanism as existing MLMs – predicting a masked POI in the pseudo sentence based on others. SpaBERT [60] creates pseudo sentences by concatenating the names of neighboring POIs. For each POI, SpaBERT

Table 1: A summary of the surveyed papers on self-supervised learning for **points-of-interest**.

Method	Type	Data Augmentation	Context Information	Objective Function	Pretext Tasks
Place2vec [53]	Predictive	Boosting the occurrence of popular neighbors	Spatial neighbors	Skipgram	Category prediction
Semantic [54]	Predictive	Random walk sampling	Spatial neighbors	Skipgram	Category prediction
MT-POI [55]	Predictive	Random walk sampling	Spatial neighbors	Skipgram	Category prediction
STPA [56]	Predictive	None	Spatial neighbors	Cross-Entropy	Category prediction
DeepR [57]	Predictive	None	Spatial neighbors	Cross-Entropy	POI relation prediction
PRIM [58]	Predictive	None	Spatial neighbors	Cross-Entropy	POI relation prediction
GeoBERT[59]	Predictive	Neighbors ordered by distance Word masking	Spatial neighbors	Cross-Entropy	Masked word prediction
SpaBERT [60]	Predictive	Shortest paths in grid cells, POI sequences ordered by the distance to grid center	Spatial neighbors	Cross-Entropy	Masked POI prediction
MGeo [61]	Predictive, Contrastive	Neighbors ordered by distance POI attribute masking	Spatial neighbors	Cross-Entropy, KL divergence	POI attribute prediction, Distance contrast
SG-CWARP [62]	Predictive	None	Check-in sequence	Ranking loss, Skipgram	POI prediction
CAPE [63]	Predictive	None	Check-in sequence	Skipgram	POI ID and text prediction
DeepMove [64]	Predictive	Construction of origin POI pairs with the same destination	Check-in sequence	Skipgram	Origin POI prediction
Hier-CEM [65]	Predictive	Hierarchical extension of context categories	Check-in sequence	CBOW	Next category prediction
POI2Vec [66]	Predictive	Binary region tree construction	Check-in sequence	CBOW	POI prediction
TALE [67]	Predictive	Binary temporal tree construction	Check-in sequence	CBOW	POI prediction
LSPSL [68]	Predictive	POI attribute masking	Check-in sequence	Cross-Entropy	Masked POI attribute prediction
MoCo-GA [69]	Contrastive	Samples from previous training steps	Co-query	InfoNCE	Location-location, Text-location, Text-text contrast
SEENet [72]	Contrastive	Grid cell augmentation	Temporal context, Spatial neighbors, Co-query, check-in sequence	JS divergence	Time-aware POI relation prediction
CatEM [70]	Predictive	None	Spatial neighbors, Check-in sequence	Skipgram	Category prediction
DCHL [73]	Contrastive	Multi-view hypergraph construction	Spatial neighbors Check-in sequence	InfoNCE	POI-POI contrast
STKG-PLM [75]	Predictive, Contrastive	Spatio-temporal knowledge graph construction	Spatial neighbors Check-in sequence	InfoNCE Cross-Entropy	POI-POI contrast POI prediction
AGCL [76]	Predictive, Contrastive	Spatio-temporal knowledge graph, co-occurrence frequency graph construction	Spatial neighbors Check-in sequence	InfoNCE Cross-Entropy	POI-POI constrast POI prediction
POI-Enhancer [74]	Contrastive	Geography view, Sequence-time view, Functional view construction	Spatial neighbors Check-in sequence	InfoNCE	POI-POI contrast
ERNIE-GeoL [71]	Predictive	Heterogeneous graph, Random walk sampling, Word substitution and masking	Co-query context, Spatial neighbors	Cross-Entropy	Masked word prediction, Geocoding prediction

finds its neighboring POIs using Geohash and sorts them by their distance to the POI. Subsequently, SpaBERT masks some or all words of a certain POI in a pseudo-sentence and predicts the masked words based on the remaining words in the sentence. In this way, the learned POI representation preserves essential information for generating the words that appear in the nearby POIs, thus capturing the underlying correlations between geo-locations and text. To better leverage the spatial context and support query-POI matching, MGeo [61] designs several geospatial encoders to encode the spatial attributes of a POI, including ID, shape, map position, and relative location to neighboring spatial objects extracted from OpenStreetMap [159]. Subsequently, MGeo employs both predictive and contrastive SSL objectives to train these encoders. The predictive SSL objective minimizes the loss of predicting the masked attribute (e.g., shape) using other attributes. The contrastive objective minimizes the difference between the actual spherical distance and the distance computed by the representations of all pairs of POIs in the same batch. After training the encoders, MGeo fuses the text and spatial information through another MLM that predicts the masked words or spatial attributes of a POI.

Check-in Sequence. Location-based services allow users to share their visits at POIs. Such visits are often referred to as "check-ins". Each check-in record is a triplet of (u, p, t), denoting a user u visiting a POI p at a certain time t. The check-in records for a specific user can be arranged in chronological order, forming a check-in sequence. The check-in sequences produced by users often exhibit strong dependencies between POIs.

Most existing studies [62, 66, 63, 64, 65] employ skip-gram, contextual bag-of-word (CBOW), which are three typical predicative objectives, to the check-in sequences. Specifically, SGCWARP [62] learns POI representations to predict other POIs within the same check-in sequences. As a result, the co-occurren ces of POIs in the same check-in sequences are preserved in the POI representations. Building upon this, CAPE [63] learns POI representations that effectively predict IDs and texts of other POIs in the same sequences. In this way, the text information in the sequential context is encoded in the learned POI representations. Unlike previous methods that predict all POIs visited within a close time period, DeepMove [64] focuses more on the intent of the trip and considers only the origin and destination POIs of a trip. Utilizing an origin POI, DeepMove predicts other origin POIs with the same destination POI. Consequently, POIs with similar travel intents are projected into a close region in the learned representation space. Apart from predictive methods that employ the skip-gram objective, SSL methods with the CBOW objective have been proposed to predict a check-in POI based on its context. These methods often focus on the modeling of spatio-temporal context in the check-in sequences. Hier-CEM [65] extends the check-in sequence context by associating each check-in POI with the ancestor categories. POI2Vec [66] introduces a binary region tree to enhance the modeling of check-in sequence context with spatial proximity. TALE [67] extends the binary region tree in POI2Vec with time information, producing a temporal tree that splits the check-ins into different time slots. In this way, TALE captures both spatial and temporal proximity in the context of a check-in.

Motivated by Masked Language Models (MLM), some recent studies propose to learn POI representation by predicting the masked check-in in a sequence. LSPSL [68] obtains a check-in's representation by a self-attention network, aggregating information from its attributes (e.g., location). The attribute embeddings are then pre-trained by employing each check-in's representation to predict the attributes of other check-in records in the same trajectory.

Co-query Context. In digital map services, users can submit text queries to search for POIs. The POIs clicked by a user in the same query are often correlated. Motivated by this, several studies have been dedicated to incorporating the co-query context in learning POI representations. ERNIE-GeoL [71] constructs a heterogeneous graph containing query and POI nodes to capture the correlation between POIs in the same query. POIs are connected by a typed edge if they are clicked successively in the same query or co-locate in the same geographical region. A POI node is connected to its historical query nodes. On the heterogeneous graph, ERNIE-GeoL runs a random walk algorithm to create node sequences, on which it applies two predictive SSL objectives to learn the representations of queries and POIs. Specifically, the first SSL objective minimizes the prediction errors for the masked words of POI or query nodes in the node sequence. The second SSL objective minimizes the error of predicting the geocoding in the discrete global grid system [160]. Besides, to enhance the learning of the text features of POIs, ERNIE-GeoL augments the text attribute of POIs by randomly swapping words with misspelled words or random words. POI relationship prediction methods, such as SEENet [72], use the POI queries to construct competitive and complementary relations between POIs and employ GNNs to obtain POI representations that preserve the co-query information. Unlike the previous predictive SSL methods, MoCo-GA [69] employs momentum contrastive learning to ensure the current POI embeddings are closer to the query embeddings than those obtained in the previous training steps.

Temporal Context. POIs may exhibit distinct relations at different times. For instance, a restaurant and a coffee shop might be competitive during breakfast and complimentary during lunch time. Motivated by this, SEENet [72] proposes to learn POI representations for different time slots. It constructs a dynamic graph of POIs with changing relations between POIs and adapts GNN to capture the intra-time and inter-time correlations between POIs. SEENet optimizes two SSL objectives simultaneously. In the first SSL objective, SEENet divides the map into grid cells and considers POI at time t and its corresponding grid cell at time $t-1$ as a positive sample. In contrast, negative samples are created by randomly replacing the grid cell with a random grid cell. Subsequently, it adopts a contrastive loss [161] to maximize mutual information based on the positive and negative samples. In the second SSL objective, SEENet predicts the existence of a relation between two consecutive time slots. The two SSL objective functions enable SEENet to encode time-specific features and the evolving relations between POIs in the POI representations.

Multiple Context Sources. Since POIs are often located within

complex spatiotemporal contexts, extensive research highlights the value of incorporating multiple contextual sources when learning POI representations. CatEM [70] represents the category co-occurrence in check-in sequences by a Point-wise Mutual Information (PMI) matrix and captures the spatial neighborhood by a category proximity matrix. CatEM learns POI representations by simultaneously optimizing the matrix factorization loss on the PMI matrix and minimizing the distance between embeddings of nearby POIs. In this way, CatEM incorporates the spatial neighborhood context and the check-in sequence context. DCHL [73] constructs hypergraphs from three different views, including geographical view, collaborative view, and transitional view, to incorporate the spatial neighborhood and sequential information of check-in sequences at the same time. To learn POI representations, DCHL employs a cross-view contrastive learning technique that takes the same POI from the different views as positive pairs while different POIs as negative pairs. STKG-PLM [75] models the POI data with a spatio-temporal knowledge graph, defining spatial neighborhoods and check-in context by pre-defined relations. It employs a knowledge graph encoder to obtain POI representations and the InfoNCE [12] objective for minimizing the gap between POI representations checked in by the same user while maximizing those checked in by different users. Similar to STKG-PLM, AGCL [76] constructs a spatio-temporal knowledge graph to capture the spatial neighborhood and transitional relations between POIs. Different from STKG-PLM, it constructs a co-occurrence frequency graph to incorporate the frequency of all users who visit one POI after another. To learn POI representations, AGCL employs both intra-graph and inter-graph contrastive learning objectives. The intra-graph contrastive objective ensures that POIs with high correlations within and across graphs are positioned closely in the embedding space. POI-Enhancer [74] proposes a POI embedding model to extract semantic information of POIs from large language models and employs a multi-view contrastive learning method to learn POI representations. It creates a geography view, a sequence-time view, and a functional view of POIs to incorporate the spatial, sequential, and semantic context of POIs. POIs that are located in a close region are considered positive samples in the geography view, and those that share the same category and visit patterns are considered positive samples in the functional view. In the sequence-time view, POIs visited in a close time period are regarded as positive samples. With these three views, POI-Enhancer learns a comprehensive POI representation by optimizing the InfoNCE loss. While most studies incorporate the spatial context and the sequential context of POIs, some research has been dedicated to exploring other combinations of context sources. For instance, ERNIE-GeoL [71] incorporates multiple context sources by defining a heterogeneous graph with three types of edges, i.e., Origin-to-Destination, POI co-location, and POI co-query. SEENet [72] constructs a dynamic graph based on the spatial proximity and the temporal relation between POIs to incorporate spatiotemporal, sequential, and co-query contexts.

3.1.3. Applications

The POI representations acquired from the SSL methods can be applied to various downstream applications. Most POI representations can be used to predict specific POI attributes, such as category [56] and geo-location [61]. Besides, POI presentations can be directly used or used with a subsequent sequential model to predict the next POI visited by a user [62, 66, 63]. The POI representations can also be used to answer POI queries by matching POI representations and the query representation [71]. Furthermore, the representations of multiple POIs can be used for more complex geospatial data mining tasks. For instance, POI relation prediction methods [57, 58, 72] use the representations of two POIs to predict their competitive or complementary relations. Likewise, the POI representations are pivotal in matching POIs from different databases for entity resolution [162]. In addition, increasing studies have utilized POI embeddings in an aggregated manner for the inference of land use [54] or land use change [55] across years. The aforementioned applications demonstrate that the SSL methods can effectively encode POIs and their contexts, thereby helping to improve the accuracy of various GeoAI tasks.

4. Polylines

Polylines serve as an important spatial data type, widely employed across diverse GeoAI applications to depict a range of features, such as contour lines, road segments, and trajectories. SSL for polylines aims to derive effective representations for such data instances, with a particular emphasis on road segments and trajectories due to their prominence in existing literature. While both road networks and trajectories can be geometrically represented as polylines, they differ fundamentally in structure and semantics. Road networks typically denote static infrastructure with topological consistency and connectivity, whereas trajectories capture dynamic movements, often annotated with timestamps and user-specific metadata. This divergence in semantics and application motivates a distinct categorization for these two data instances in our survey, enabling a more precise analysis of SSL methods tailored to each context. Consequently, we will separately discuss these two instances of polylines, and present the specialized SSL techniques tailored to each data instance. The surveyed studies for road networks are listed in Table 2.

4.1. Road Networks

4.1.1. Intrinsic Attributes

Road networks are composed of connected road segments, each of which is represented as a polyline. Since the fundamental analytical units in road network studies are typically individual road segments rather than enclosed regions, road networks are classified under the category of polylines. Conceptually, road networks can be represented as graphs, allowing for the modeling and analysis of their structural and topological properties. In this regard, two primary graph-based perspectives are adopted in the research. In the first perspective, road segments themselves are treated as nodes, and the connections between

these segments are treated as graph edges. In the alternative perspective, road intersections serve as nodes, while the road segments linking these intersections are treated as edges.

The graph-based formulation from these two perspectives naturally represents the topological structure of road networks. Additionally, various attributes on road networks are integrated as node or edge features within the formulation, such as geolocations, road attributes, and intersection characteristics. As a result, existing studies on SSL for road networks generally follow the principles of SSL for graph data. The objective is to learn geospatial encoder f_θ to obtain node representations $\{\mathbf{e}_v\}_{v\in\mathcal{V}} \in \mathbb{R}^d$ (i.e., $\{\mathbf{e}_v\}_{v\in\mathcal{V}} = f_\theta(\mathcal{G})$, where $\mathcal{G}$ and $\mathcal{V}$ denote the road network graph and its nodes, by employing either graph perspective in a self-supervised manner.

Early approaches generally adopt predictive SSL with various training objectives. Initial exploration into this field [77] applies node2vec [163], a self-supervised graph representation learning method based on skip-gram training, directly to road networks. It predicts road segments within a context window derived from random walks sampled on road networks [158]. The derived representations are demonstrated to be effective on several road classification tasks. Building upon this groundwork, IRN2Vec [78] targets at learning representations for intersections by treating intersections as graph nodes. This method refines node2vec by employing shortest-path random walk sampling and selecting positive intersection pairs based on defined road path distances for skip-gram training. Moreover, it leverages intersection-specific attributes (e.g., intersection tags and types), to enrich the selection of positive road segment pairs. Another research line explores GNN with reconstruction-based objectives. For example, RFN [164, 79] applies GNN to both perspectives of the road network graphs, fusing features from both road segments and intersections, including zone categories, road types and intersection angles, to derive representations capable of reconstructing the original graphs. HyperRoad [80] expands upon the vanilla graph structure by constructing a corresponding hypergraph, where hyperedges represent the road segments within the polygons produced through a map segmentation algorithm [165]. GNN is further enhanced by a dual-channel attention mechanism that operates on both the original graph and the hypergraph. This method is trained to not only construct both the original graph and its corresponding hypergraph, but also to perform a hyperedge classification task.

4.1.2. Context Information

Apart from the topological structure and the attributes associated with road segments and intersections, road networks contain rich context information that can be leveraged to enhance the extraction of semantic knowledge. We introduce several types of context information utilized in existing methods, where road segments are mainly regarded as graph nodes.

Spatial Attributes. Different from conventional graphs that

Table 2: A summary of the surveyed papers on self-supervised learning for **road networks**.

Method	Type	Data Augmentation	Context Information	Objective Function	Pretext Tasks
R2vec[77]	Predictive	Random walk sampling	None	Skipgram	Context neighbor prediction
IRN2Vec [78]	Predictive	Random walk sampling	Road attributes	Cross-Entropy	Road attribute prediction
RFN [79]	Predictive	None	Road attributes	Cross-Entropy	Graph reconstruction
RN2Vec [81]	Predictive	Random walk sampling	Road attributes Spatial features	Skipgram	Road attribute prediction
HyperRoad [80]	Predictive	Hypergraph construction	Road attributes	Cross-Entropy, Skipgram	Graph&Hypergraph reconstruction, Road attribute prediction, Hyperedge Classification
SARN [82]	Contrastive	Edge perturbation	Spatial features, Road attributes	InfoNCE	Road-road, Road-region contrast
HNRN [83]	Predictive	None	Trajectory	Cross-Entropy	Graph reconstruction
Toast [84]	Predictive	Road segment masking, Random walk sampling, Similar trajectory generation	Trajectory	Cross-Entropy, Skipgram	Context neighbor&attribute prediction, Masked road recovery, Trajectory discrimination
DyToast [85]	Predictive	Road segment masking, Random walk sampling, Similar trajectory generation	Trajectory	Cross-Entropy Skipgram	Time-aware context neighbor &attribute prediction, Masked road recovery, Trajectory discrimination
TRACK [88]	Predictive, Contrastive	Road segment masking, Traffic state masking, Similar trajectory generation	Trajectory	Cross-Entropy, MAE, InfoNCE	Time-aware masked road& traffic state recovery, Traffic state-road matching, Trajectory discrimination
TrajRNE [86]	Predictive	Transition view construction	Trajectory	Cross-Entropy, MAE	Graph reconstruction, Road attribute prediction
JCLRNT [87]	Contrastive	Transition view construction, Detour generation	Trajectory	JS divergence	Road-road, Road-trajectory, Trajectory-trajectory contrast
USPM [90]	Predictive, Contrastive	None	SVI	InfoNCE, Cross-Entropy	Road-image contrast, Road attribute prediction
Garner [89]	Contrastive	Multi-view graph construction	SVI	JS divergence	Road-road contrast

primarily focus on connectivity, road networks are inherently defined within a geospatial context, exhibiting spatial features such as coordinates, lengths, and angles which are beneficial in modeling efforts [79, 80]. Accordingly, spatial information has been effectively utilized as context knowledge in various methods. RN2Vec [81] extends IRN2Vec to obtain representations for both intersections and road segments by selecting spatially nearby positive pairs among roads and intersections in skip-gram training. SARN [82] constructs a weighted adjacency matrix that reflects the spatial proximity among road segments. The matrix is inferred from both road connectivity and the spatial and angular distances between road segments. Utilzing this matrix, SARN employs a GNN-based contrastive learning approach [166] with graph augmentation techniques, deriving similar representations for the identical road segments from two distinct augmented graphs.

Trajectories. Trajectories traveled on road networks act as valuable data sources that provide travel-related semantics beyond the topological structure. HNRN [83] develops a hierarchical GNN framework to model relationships from individual road segments to structural regions and further extending to functional zones. It involves incorporating trajectories to derive connectivity matrix in structural regions, and aims to reconstruct the connectivity matrix for the base-level road segments derived from the higher level of region and functional representations. Toast [84] is the initial effort to explicitly utilize detailed trajectories to enhance road network representation learning through predictive SSL. It equips skip-gram training with additional traffic and attribute prediction for context neighbors. Besides, it proposes two novel trajectory pre-training tasks within BERT framework [37]: route recovery, which recovers a sequence of masked road segments, and trajectory discrimination, which assesses whether a trajectory is authentic or simulated through random walk sampling on the road network graph. These tasks enable the encoding of transition patterns and long-term dependencies intrinsic in road networks. Dy-Toast [85] further extends this method by incorporating temporal consideration into the representations. It modifies the skip-gram training and the BERT framework by integrating parameterized trigonometric functions to capture dynamics and evolution in time dimension. Similarly, TRACK [88] considers temporal dynamics for road segment representations by integrating trajectory data-based transition probabilities into the GAT attention mechanism while modeling traffic states for road networks. Building on Toast pretext tasks, TRACK further employs masked traffic state imputation and prediction tasks for training dynamic road segment representations. Besides, it further incorporates a co-attentional transformer encoder with a gravity-based attention mechanism and a trajectory-traffic state matching task, ensuring mutual reinforcement between these two data sources. JCLRNT [87] proposes three types of objectives based on contrastive SSL: road-road, trajectory-trajectory and road-trajectory contrastive loss. Each type is designed to differentiate entity pairs that share relatedness (e.g., road segments frequently traveled within trajectories) with those that do not. TrajRNE [86] utilizes the road transition matrix derived from trajectories in GNN aggregation function, and employs the objective of reconstructing the original topological structure of road networks from this transition matrix.

Street View Images. Street view images (SVIs), which are available from various map services, provide high-resolution visual perspectives of road networks. These images capture the surroundings and configurations of different road segments, inherently encoding rich urban semantics and insights. USPM [90] utilizes pre-trained image encoders to extract the representations for these images and derives a road segment representation aggregated from all associated images. The road segment representation and each of its associated images are treated as a positive pair for contrastive learning. Moreover, USPM further enhances the representations by incorporating textual descriptions of the images and applies GNN on the topological graph of road networks. The final representations are trained with the objective of road attribute prediction. Garner [89] extends beyond the idea that nearby road segments should exhibit similar representations. It seeks to also derive similar representations for road segments that display similar geographical configurations as evidenced by street view images. To achieve these two goals, after extracting and aggregating representations from image encoders for each road segment, Garner employs a dual contrastive learning objective, in which GNNs are applied to distinguish a road segment representation within three graphs, where edges represent the topological structure, nearest neighbors and similar configurations respectively.

4.1.3. Applications

The representations of road networks, derived from either predictive or contrastive SSL objectives, provide task-agnostic inputs that can be directly utilized or fine-tuned with geospatial encoder, for downstream applications. For example, they are directly applicable in classifying attributes of road segments or intersections with simple models (e.g., MLP), such as road type [86], lane number [80], and intersection tags [78, 81]. Besides, these representations are similarly utilized to infer the traffic status of road segments, including metrics like average speed [84, 87] and speed limits [79]. They also support the efficient vector computation for road network-based queries, such as calculating shortest path distances [82]. Furthermore, road segment representations serve as the effective start point for training models that involve map-matched trajectories, for applications such as destination prediction [83] and anomalous sub-trajectory detection [167]. By enabling diverse analytical tasks related to road networks, these representations offer substantial potential to improve understanding and decision-making within the domain of transportation infrastructure.

4.2. Trajectory

4.2.1. Intrinsic Attributes

A trajectory, composed of a sequence of sampled points that represent the path of a moving object, can essentially be conceptualized as a polyline. Based on the definition, the context feature associated with each point in trajectory data instance includes timestamps, along with potentially additional textual

Table 3: A summary of the surveyed papers on self-supervised learning for **trajectory**.

Method	Type	Data Augmentation	Context Information	Objective Function	Pretext Tasks
TCDRL[91]	Predictive	Sliding window feature extraction	Grid partition	Cross-Entropy	Trajectory reconstruction
t2vec [92]	Predictive	Point dropping, Spatial distortion	Grid partition	Spatial-aware Cross-Entropy	Trajectory reconstruction
GM-VSAE [93]	Predictive	None	Grid partition	Cross-Entropy	Variational trajectory reconstruction
Geo-Tokenizer [94]	Predictive	None	Multi-scale grid partition	Multi-scale Cross-Entropy	Autoregressive next grid prediction
AdvTraj2vec [95]	Predictive	Point dropping, Embedding perturbation	Grid partition	Spatial-aware Cross-Entropy, Adversarial loss	Trajectory reconstruction, Adversarial learning
E^2DTC [96]	Predictive	Point dropping, Spatial distortion	Grid partition	Spatial-aware Cross-Entropy, Triplet loss, KL divergence	Trajectory reconstruction, Triplet margin similarity, Self-clustering
STPT [97]	Predictive	None	Grid partition	Cross-Entropy	Sub-trajectory discrimination
ST-t2vec [98]	Predictive	Point dropping, Spatial&Temporal distortion	3D temporal grid partition	MSE	Trajectory reconstruction, Pairwise&Pattern similarity
RSTS [104]	Predictive, Contrastive	Point dropping, Spatial&Temporal distortion	3D temporal grid partition	Spatial-aware Cross-Entropy, Triplet loss	Trajectory reconstruction, Triplet margin similarity
DeepTEA [99]	Predictive	None	Grid partition, Historical traffic	Cross-Entropy	Variational trajectory reconstruction
UniTraj [100]	Predictive	Point dropping	Point coordinates	Cross-Entropy	Trajectory reconstruction
CL-Tsim [101]	Contrastive	Point dropping, Spatial distortion	Grid partition	InfoNCE	Trajectory-trajectory contrast
TrajCL [102]	Contrastive	Point dropping, Spatial distortion, Trajectory trimming & simplification	Grid partition, Point coordinates	InforNCE	Trajectory-trajectory contrast
KGTS [103]	Contrastive	Grid deletion, Grid alternation	Grid partition	InfoNCE	Trajectory-trajectory contrast
CSSRNN [105]	Predictive	Map matching	Road networks	Cross-Entropy	Autoregressive road prediction
Traj2Vec [106]	Predictive	Map matching	Road networks	Cross-Entropy, MSE	Trajectory reconstruction, Travel time estimation
PT2Vec [107]	Predictive	Map matching, Road network partition	Road networks	Cross-Entropy	Trajectory reconstruction
JGRM [108]	Predictive	Map matching, Road segment masking	Road networks	Cross-Entropy	Masked road recovery, Trajectory discrimination
ST2Vec [109]	Contrastive	Map matching	Road networks	Triplet loss	Triplet margin similarity
GRLSTM [110]	Contrastive	Map matching	Road networks	Triplet loss	Triplet margin similarity
PIM [111]	Contrastive	Map matching	Road networks	JS divergence	Road-road, Road-trajectory contrast
MMTEC [112]	Contrastive	Map matching, Continuous Trajectory	Road networks	Maximize entropy encoding	Trajectory-trajectory contrast
HMTRL [113]	Predictive, Contrastive	Map matching	Road networks	Cross-Entropy, MSE	Masked attribute prediction, Trajectory-trajectory contrast
START [114]	Predictive, Contrastive	Map matching, Dropout, Temporal distortion, Trajectory trimming, Road segment masking	Road networks	Cross-Entropy, InfoNCE	Masked road recovery, Trajectory-trajectory contrast
LightPath [115]	Predictive, Contrastive	Map matching, Road segment masking	Road networks	Cross-Entropy	Masked road recovery, Multi-view trajectory matching
GREEN [116]	Predictive, Contrastive	Map matching, Road segment masking	Grid partition, Road networks	Cross-Entropy, InfoNCE	Masked road recovery, Road-grid view trajectory contrast
CTLE [117]	Predictive	Point masking, Time masking	Location semantics	Cross-Entropy	Masked location&time recovery
At2vec [118]	Predictive	Point dropping	Location semantics	Spatial-temporal -activity-aware Cross-Entropy	Trajectory reconstruction
SML-LP [119]	Contrastive	Point dropping, Location alternation	Location semantics	InfoNCE	Location embedding contrast

Continued on next page

Table 3 A summary of the surveyed papers on self-supervised learning for trajectory (continued).

Method	Type	Data Augmentation	Context Information	Objective Function	Pretext Tasks
CLMTR [120]	Contrastive	Point dropping, Spatial distortion, Trajectory trimming & simplification	Location semantics	InfoNCE	Spatial-temporal -textual feature contrast, Trajectory-trajectory contrast
CACSR [121]	Contrastive	Location embedding & latent space perturbation	Location semantics	InfoNCE	Latent representation contrast
MVTraj [124]	Predictive, Contrastive	Map matching, Road segment masking, Point dropping	Grid partition, Road networks, Location semantics	Cross-Entropy, InfoNCE	Masked grid/road recovery, Multi-view trajectory cross contrast
PTrajM [123]	Contrastive	Map matching, Point allocation	Road networks, Location semantics	InfoNCE	Trajectory-poi view contrast, Trajectory-road view contrast
RED [122]	Predictive	Map matching	Road networks, User information	Cross-Entropy	Masked road recovery, Next segment prediction
MM-Path [125]	Predictive, Contrastive	Map matching, Road segment masking	Road networks, Road images	Triple loss, InfoNCE, Cross-Entropy	Masked road recovery, Multi-granularity road-image view contrast, Fused road-image view contrast

content, and semantic tags, etc. The objective of SSL for trajectory data is to learn a representation produced by geospatial encoder f_θ for any given trajetocry $\mathcal{T}$: $\mathbf{e}_T = f_\theta(\mathcal{T}) \in \mathbb{R}^d$. The sequential nature of a trajectory, marked by its spatio-temporal point sequence, represents its most fundamental intrinsic attributes. Accordingly, sequential models are used to model trajectory data, effectively capturing the transition patterns and long-term dependencies inherent in this data instance. The surveyed studies for trajectory are listed in Table 3.

For predictive SSL methods, trajectories are typically trained to reconstruct their original sequence from a corrupted version of the input. For example, TCDRL [91] proposes to extract local features within sliding windows applied on trajectory records, such as speeds and rate of turns. These features are then processed using a sequence-to-sequence encoder-decoder architecture [168] based on RNN to reconstruct the local features for each window. t2vec [92] refines this reconstruction process by aligning with the tokenization paradigm in natural language. It partitions the geospatial space into regular and square-sized grids and match the coordinates to their corresponding grids (tokens). Moreover, t2vec introduces downsampling and point distortion in the input sequence, and aims to reconstruct the original trajectory with spatial proximity aware loss that penalizes more for the predicted tokens that deviate significantly from the correct grids. In such an encoder-decoder framework, the vector produced by encoder part can be treated as the trajectory representation.

This framework is enhanced by integrating a variety of modifications in terms of model architecture and loss functions, such as integrating via variatioanl inference [93] or self-attention [169]. Moreover, Geo-Tokenizer [94] reduces the number of grids to be trained by representing a location as a combination of multiple shared grids at several granular scales. It utilizes the objective of predicting the grids for the next token at each scale in a hierarchical way. Furthermore, AdvTraj2vec [95] aims to learn more robust trajectory representations through adversarial training. It adds perturbations to the token embeddings of the input sequence, with the magnitude of these perturbations guided by generative adversarial network [170] to ensure the effects are neither too large or too small. E^2DTC incorporates self-training via soft cluster assignments [96] as an auxiliary loss into the reconstruction process. STPT [97] performs a sub-trajectory discrimination loss that differentiates whether pairs of sub-trajectory representations originate from the same source.

In addition to considering spatial dimension associated with trajectories, several SSL methods leverage temporal dimension into the reconstruction process. The method in [98] expands the square-sized spatial partitions to 3D spatio-temporal grids for each point. It accordingly adapts the reconstruction loss to account for temporal effects, imposing heavier penalties for reconstructed results with larger temporal discrepancies. Similarly, RSTS [104] also employs 3D spatio-temporal grids and applies a linear combination of spatial and temporal distances for reconstructed results as the final loss. Besides, DeepTEA [99] utilizes Convolutional-LSTM [171] to model historical traffic conditions reflected in holistic trajectories, thus enhancing the variational inference by providing additional hints into dynamic patterns for each trajectory. Recently, UniTraj [100] introduces a large-scale, global trajectory dataset to train a universal trajectory model capable of generalizing across diverse tasks and regions, thus utilizing pure spatial coordinates without partitioning. To manage variations in sampling rates, it employs dynamic resampling and utilizes multiple masking strategies, including random, block, key point, and last-N masking, for data augmentation. The model adopts an encoder-decoder architecture with Rotary Position Embedding [172] to model trajectory sequences, and aims to predict masked segments for training.

In the category of contrastive SSL methods, standard mutual information maximization objectives are applied to produce similar/dissimilar representations for views derived from the same/different trajectories with various data augmentation operations. CL-Tsim [101] primarily uses point distortion to generate positive pairs of trajectories for contrastive learning. Expanding on this, TrajCL [102] introduces three additional operations, namely point masking, trajectory truncating and trajectory simplification, to enhance the diversity of the patterns for positive pairs. Besides, it proposes a dual-feature self-attention-based encoder to process not only the grid sequence, but also the spatial attributes for points, such as coordinates, angles and lengths. KGTS [103] further implements GNN to consider the interactions of neighboring grids, and includes grid deletion and movement operations to both the entire trajectories and partial trajectories to create positive pairs.

4.2.2. Context Information

In addition to the intrinsic spatial and temporal attributes, we discuss two distinct scenarios where trajectories are modeled under specific constraints and characteristics with further context information.

Road Networks. While trajectories offer travel-related semantics for road networks, road networks in turn serve as complementary elements that impose latent topological constraints for trajectories. Specifically, when path information alongside road networks is emphasized, trajectories are usually map-matched to road networks [173, 174]. This process transforms the trajectory from a sequence of points to a sequence of road segments for subsequent modeling.

Several predictive SSL methods for map-matched trajectories utilize objectives similar to those used in trajectories with grid partitions, such as the reconstruction of road segments [106, 107] and next road prediction [105]. Besides, tailored objectives are devised to effectively incorporate the knowledge within road networks. For example, JGRM [108] aims to recover the road segments masked from the complete path. It also considers the original point sequence and adopts another objective of differentiating whether pairs of representations are derived from the same trajectory with both the road segment sequence and the point sequence.

For contrastive SSL methods applied to map-matched trajectories, ST2Vec [109] adopts co-attention mechanism to merge spatial and temporal embeddings, followed by LSTM sequence modeling. This model employs energy-based margin functions

(i.e., triplet loss) to enforce higher similarities for positive pairs which follow similar routes. GRLSTM [110] combines GNN and graph embedding techniques [175] to handle sequence inputs and augments LSTM with residual connections to generate trajectory representations. It employs similar triplet loss at both the point level and the trajectory level. In addition, PIM [111] treats trajectories that share the same source and destination as positive pairs, and aims to maximize mutual information over these pairs. MMTEC [112] utilizes both a discrete sequence model based on Transformer, and a continuous model formulated as neural controlled differential equation to generate representations from two views. It applies a contrastive learning objective of maximizing entropy coding between two views.

Moreover, both contrastive and predictive SSL can be simultaneously employed with a single framework. HMTRL [113] employs the strategies to predict road attributes and traverse time for road segments, as well as employing full trajectory and its sub-trajectory for contrastive learning. START [114] incorporates the trajectory pattern in GNN aggregation and enhances the Transformer with a time-sensitive self-attention mechanism. Furthermore, it utilizes masked road segment recovery as the predictive task, and enhances its contrastive learning aspect by employing four distinct data augmentation operation – trajectory trimming, masking, temporal shifting, and dropout– to generate positive pairs. Similarly, LightPath [115] utilizes masked road path recovery as its predictive task. For its contrastive aspect, LightPath achieves pairwise matching through the use of dual encoders and varied dropping ratios to generate positive pairs from the same trajectories. GREEN [116] integrates a grid encoder based on CNN and a road encoder based on GNN to encode trajectories from two perspectives. It employs contrastive training to align the representations from both encoders for the same trajectory while using masked road segment recovery as a predictive task.

Semantic Information. Trajectories may carry rich semantic meaning when composed of check-in sequence at specific POIs, which provide concrete location names and related activities (i.e., categories) but result in sparser sequences. Several SSL methods are developed to tackle semantic trajectories. CTLE [117] designs two predictive tasks, namely masked location recovery and masked hour prediction, with a sinusoidal temporal encoding technique incorporated into Transformer to derive representations. At2vec [176, 118] utilizes an encoder-decoder framework for sequence reconstruction, and adopts multi-level attention to consider the importance of semantic information at different locations. SML-LP [119] applies trajectory augmentation techniques that modify several locations within close spatial and temporal proximity to form positive pairs, and aims to maximize the mutual information between the LSTM-derived hidden representations and the future location representations. CLMTR [120] encodes trajectory by leveraging textual descriptions of POIs processed with BERT alongside spatial and temporal embeddings to capture geographical proximity and periodicity. It employs contrastive learning at two levels: intra-trajectory, where textual features are contrasted with fused spatiotemporal features within the same trajectory, and inter-trajectory, where nearest-neighbor trajectories serve as positive pairs while distant ones act as negatives. Furthermore, CACSR [121] innovates by generating challenging positive and negative pairs in the representation space, rather than input sequence, via adversarial perturbations. We note that while SSL techniques are integrated into the modeling of semantic trajectories in other studies [177, 178], the main target and its objective derives from supervised applications. Therefore, these specific studies are not further discussed in our context.

Multiple Context Sources. The extensive studies have revealed that different types of context information provide distinct and complementary insights for trajectory data representation. For instance, spatial coordinates and grid partitions offer high-resolution positional accuracy, while road networks impose movement constraints and reveal structural mobility patterns. Additionally, location semantics provide valuable information regarding regional functionalities and human activity patterns. To this end, recent research has increasingly focused on integrating multiple context sources simultaneously, recognizing that their combination enables a more comprehensive and multi-dimensional understanding of trajectory data. MVTraj [124] integrates context information from multiple views, including GPS trajectories, road networks and POIs. It employs three encoders to capture point sequences, road network paths, and grid-based semantic features, utilizing a hierarchical multi-modal interaction with twelve attention streams for cross-view fusion. The model incorporates masked part recovery as a predictive task and applies contrastive learning to align trajectory representations across different views. PTrajM [123] utilizes the Mamba state-space model [179] to capture continuous movement behaviors and infer travel purposes from road network and POI semantic views by assigning data points to POIs. It extracts road-based and POI-based trajectory representations using pre-trained textual embeddings and Transformer-based encoders. A contrastive learning approach is then employed to align trajectory representations extracted from the state-space model with their corresponding road network and POI-based representations. Several studies have expanded trajectory representation learning by integrating additional contextual information. RED [122] incorporates user information alongside spatial and temporal encodings to enhance trajectory representation within a road network context. It employs dual Transformer models for next-segment prediction and masked road segment prediction, adapting the latter to selectively mask non-crucial path segments as a pretext task. MM-Path [125] combines road network-based trajectories with corresponding road segment images across three granularities: road intersection level, road sub-path level, and entire path level. It employs Transformer-based encoders for both modalities and introduces a GNN-based cross-modal fusion to integrate image and road network representations. The model utilizes masked road segment recovery as a predictive task and applies contrastive learning by aligning representations across modalities and granularities.

4.2.3. Applications

Trajectory representations can be directly utilized or fine-tuned for downstream applications. These representations, regardless of their specific context information, facilitate a variety of operations due to their vectorized format, including similarity computation [92, 180] and clustering [96, 91]. Moreover, they can be utilized to support diverse applications, such as transportation mode prediction [94], driver status inference [97], and anomalous trajectory detection [93]. In scenarios involving map-matched trajectories, these representations prove particularly valuable in applications focused on path analysis, such as travel time estimation [114], path ranking [115], and destination prediction [112]. Semantic trajectory representations trained with SSL methods are typically fine-tuned to enhance performance in next location prediction [117], and can be uniquely applied in trajectory user-linking task [121, 119]. These applications demonstrate the broad utility and adaptability of SSL for trajectories in advancing the understanding and analysis of mobility patterns and intelligent transportation systems.

5. Polygons

Polygonal representations offer detailed 2D shapes of geospatial objects. In principle, many geospatial objects like POIs and road segments could be represented this way, while the effort to collect precise polygonal data limits its use. In this survey, we focus on urban regions, i.e., small urban areas which serve as the analytical units for a wide array of urban analytical tasks, e.g., for land use, population density, and house price.

Urban regions are partitioned, e.g., by road networks, grids, or administrative boundaries like Singapore Subzones [181] and NYC Census Tracts [182]. These regions serve as "containers" for data in various modalities, entailing either intrinsic properties of regions, or interrelatedness with other regions. Different types of spatial piror, e.g., spatial proximity, is also crucial, to incorporate the inherent relations entailed by geographic locations. Most studies utilize multiple data perspectives, e.g., the region embedding studies [136, 137, 138, 144, 141, 143, 145, 139] use POIs, human trajectories, and spatial proximity. The modeling of different perspectives is often intertwined, so we present each study holistically, covering the entire analytical pipeline. To this end, we categorize the studies according to the data modalities used. See Table 4 for surveyed studies.

5.1. Urban regions

5.1.1. Intrinsic Attributes

Among various data modalities, POIs and imagery data, including remote sensing (RS) images and street view images (SVIs), are most commonly utilized to reflect the intrinsic properties of each region. POIs capture socioeconomic factors, while images reflect the overall physical appearance from a human or aerial perspective. Several studies have explored these data sources to learn region representations in a self-supervised manner, sometimes without much consideration of contextual information.

In a pioneering study on POI-based region representation learning, [126] treats each region as an "image" where some pixels are filled with POIs. This approach allows regions to be processed by a CNN, with POI category information, represented by one-hot encoding, serving as "pixel values". The model is trained with the objective of a triplet loss. For each anchor region, its augmentation generated by random removal, addition, and shifting of POIs serves as a positive sample, while negative samples are non-overlapping regions or augmentations with larger perturbations (hard negative samples). UrbanCLIP [129] utilizes RS images and LLM to learn region representations. For each image, it generates a detailed description using a pre-trained LLM, where detailed and specific prompts oriented to urban infrastructures are found to be beneficial. The method then trains an image encoder and a text encoder using contrastive learning and auto-regressive text generation. Building upon this work, UrbanVLP [127] employs both SVIs and RS images. It proposes to filter the LLM-generated text by a quality metric CycleScore, to avoid low-quality text generation. This approach fuses remote sensing and street view images within each region, contrasting them with generated texts at both image and token levels. In addition, MMGR [128] proposes to fuse RS images and POIs for learning intrinsic attributes of urban regions. It extends the idea in [12] by carrying out contrastive learning between multiple augmentations of each image, as well as between images and POIs using the method proposed in [54].

5.1.2. Context information

Incorporating diverse sources of context information to model the semantics of urban regions has become a standard practice. The contextual information often comes from connectivity patterns exhibited in human trajectories, spatial proximity, temporal patterns, and other types of property similarities (e.g., from land use data and knowledge graphs). The primary incentive for utilizing context information is its ability to significantly enhance the quality of learned region representations, especially when the expressiveness of data modalities representing intrinsic attributes, such as POIs, is limited. We organize the subsequent content based on the data modalities utilized to learn representations for urban regions, allowing for a focused discussion on how different types of data contribute to the modeling process.

POIs/Knowledge Graphs. Based on the pioneering work [126], SKRL4RS [131] further enhances the information of spatial entities by incorporating the rich context information from two well-established knowledge graphs, YAGO and DBpedia. Instead of one-hot category encoding, the spatial entities in the knowledge graphs are transformed into representations that encapsulate the relatedness derived from their hierarchical categories (ontologies) and their proximity within the knowledge graph. In this way, the semantics of spatial entities within regions are better captured. For example, the similarity between a Japanese restaurant and a Korean restaurant is much larger than that between a factory and a Korean restaurant. In another research line, HGI [130] extends beyond the POI category

Table 4: A summary of the surveyed papers on self-supervised learning for **urban regions**.

Method	Type	Data Source	Data Augmentation	Context Information	Objective Function	Pretext Tasks
Triplet[126]	Contrastive	POI	POI removal, addition, and shifting	None	Triplet loss	Region-region contrast
SKRL4RS[131]	Predictive, Contrastive	Knowledge graph	Spatial entity shifting	Knowledge graph	Triplet loss	Region-region contrast
HGI[130]	Contrastive	POI	POI and region graph	Spatial proximity, City overall context	MI maximization	POI-region contrast, Region-city contrast
HDGE[132]	Predictive	Human trajectory	Heterogeneous region graph	Human mobility, Spatial proximity, Temporal pattern	Skipgram, KL divergence	Reconstruction of human mobility patterns
ZE-Mob[133]	Predictive	Human trajectory	Human mobility event	Human mobility, Spatial proximity	Skipgram	Context neighbor prediction
GMEL[134]	Predictive	Human trajectory	Region graph	Human mobility, Spatial proximity	MSE	Commuting flow and in/out flow prediction
MGFN[135]	Predictive	Human trajectory	Region graph via clustering	Human mobility	KL divergence	Reconstruction of human mobility patterns
MP-VN[136]	Predictive	Human trajectory, POI	POI graph	Human mobility, Spatial proximity, POi distribution similarity	MSE	Graph reconstruction
DLCL[137]	Predictive	Human trajectory, POI	POI and region graph	Human mobility, Spatial proximity	Cross-Entropy	Region graph reconstruction
CGAL[138]	Predictive	Human trajectory, POI	POI and region graph	Human mobility, Spatial proximity	MSE, adversarial loss	Reconstruction of node feature and graph structure
HAFusion[139]	Predictive	Human trajectory, POI, Land use	None	Human mobility, POIs	MSE, KL divergence	Reconstruction of human mobility patterns, POI, and land use feature similarity
MVURE[140]	Predictive	Human trajectory, POI, User check-in	Multi-view graph	Human mobility, POIs, User check-in similarity	MSE, KL divergence	Reconstruction of human mobility patterns, POI, and check-in distributions
HREP[144]	Predictive, Contrastive	Human trajectory, POI	Heterogeneous region graph	Human mobility, Spatial proximity, POIs	KL divergence, Triplet loss, MSE	Region-region contrast, Reconstruction of human mobility patterns and POI distributions
HUGAT[142]	Predictive	Human trajectory, POI, Land use	Construction of heterogeneous urban graph and meta-path	Human mobility, Spatial proximity, Temporal pattern	KL divergence, MSE	Reconstruction of human mobility patterns, check-in, and land use distributions
Region2Vec[141]	Predictive	Human trajectory, POI	Multi-view region graph	Human mobility, Spatial proximity, POIs	KL divergence, MSE	Reconstruction of human mobility patterns, geospatial adjacency and POI distributions
ReMVC[143]	Contrastive	Human trajectory, POI	POI insertion, deletion, and replacement Trajectory heatmap perturbation	Human mobility, POIs	InfoNCE	Region-region contrast
ReCP[145]	Contrastive, Predictive	Human trajectory, POI	POI insertion, deletion, and replacement Trajectory heatmap perturbation	Human mobility, POIs	InfoNCE, MI maximization, Conditional entropy	Region-region contrast, Region feature reconstruction
Tile2Vec[147]	Contrastive	RS	None	Spatial proximity	Triplet	Region-region contrast
RegionEncoder[146]	Predictive	RS, Human trajectory, POI	Image noising, region graph	Human mobility	MSE, KL divergence, Cross-Entropy	Reconstruction of images and human mobility patterns , Multi-view graph discrimination
Urban2Vec[148]	Contrastive	SVI, POI	None	Spatial proximity	Triplet	Image- and region-level contrast

Continued on next page

Table 4 A summary of the surveyed papers on self-supervised learning for urban regions (continued).

Method	Type	Data Source	Data Augmentation	Context Information	Objective Function	Pretext Tasks
M3G[149]	Contrastive	Human trajectory, POI, SVI	None	Spatial proximity	Triplet	Region-SVI, region-POI, region-region contrast
Xi et al[150]	Contrastive	RS, POI	None	Spatial proximity	NT_Xent	Region-region contrast
UrbanVLP[127]	Contrastive	RS, SVI	Textual descrptions of images	None	InfoNCE	Image- and token-level Image-text alignment
MMGR[128]	Contrastive	RS, POI	RS image augmentation, POI graph	None	InfoNCE	Region-region contrast
UrbanCLIP[129]	Predictive, Contrastive	RS	Textual descrptions of images	None	InfoNCE, Cross-Entropy	Image-text alignment, Text autoregressive prediction
GeoHG[151]	Contrastive	RS, POI	Heterogenous graph	Spatial adjacency, Region feature similarity	InfoNCE	Graph contrast
ReFound[153]	Predictive, Contrastive	RS, POI	Foundation model distillation and text generation	None	KL divergence Cross-entropy	Knowledge distillation Masked data modeling Cross-modal spatial alignment
MuseCL[152]	Contrastive	RS, SVI, POI	None	Region feature similarity	Triplet InfoNCE	Image-level contrast Image-text contrast
GeoHG[151]	Contrastive	RS, POI	Heterogenous graph	Spatial adjacency, region feature similarity	InfoNCE	Graph contrast
RegionDCL[154]	Contrastive	OSM building footprints, POI	Point injection, Building removal	Spatial proximity	InfoNCE, Triplet loss	Region-region contrast

embedding technique in [54] by further applying GNN aggregation process to escalate the representations to region and city levels. This model is optimized by maximizing the mutual information among the POI-region-city hierarchy.

Human Trajectories. Human mobility data, i.e., trajectories, has been a popular data source for learning region representations, mainly in merit of the rich region-level connectivity patterns exhibited from massive trajectories. HDGE [132] is the first to leverage such data sources to consider both temporal dynamics and multi-hop transition patterns between regions. Specifically, it defines a flow graph where each node represents a region at a certain time point. Nodes in the flow graph are connected by two types of edges based on human flow between regions and spatial adjacency. The spatial adjacency edges are built to mitigate the data sparsity problem in human trajectories, i.e., to gauge the location when a user's location is not recorded. The model is trained to reconstruct the transition probability between regions, i.e., minimizing the KL-divergence between the skip-gram probability in the graph and the empirical transition probability observed from the trajectory dataset.

Later, ZE-Mob [133] defines several human mobility events pertaining to regions, time, and movement mode (i.e., departure/arrival). In this way, region representations can be learned similarly in Word2vec [158] by skip-gram objective. Besides, it enriches the objective by integrating the importance of region origin-destination pairs based on popularity and distance. GMEL [134] constructs a region adjacency graph and subsequently employs two graph attention networks to model the two types of representations for regions functioning as travel origins and destinations. The model is trained with the pretext tasks of predicting human flow and predicting in/out flow. MGFN [135] regards human movements in each time slot as a mobility graph and defines several graph distance measures to cluster the graphs into several mobility patterns, e.g., distance between mean or variance of edge weights and human flow imbalance. A hierarchical clustering method is then applied to distill these into a reduced number of region graphs that represent specific mobility patterns. Within each mobility pattern graph, message passing is performed. Besides, inter-pattern attention mechanisms are employed to fuse various mobility patterns for the final region representations. The model is trained with the objective of reconstructing region-level human mobility patterns.

POIs + Human Trajectories (+others). Many region representation studies integrate POIs with human trajectories, as these data sources naturally complement each other. POIs describe the range of activities within a region, while human trajectories reveal inter-regional connections. This integration is often enhanced with additional data, such as land use. For instance, MP-VN [136] constructs two POI graphs to capture both static and mobility patterns among POI types, based on mobility connectivity and geographic distances. These graphs are sent to an autoencoder for learning region representations via graph reconstruction, while also incorporating spatial proximity and functional similarity derived from POIs. Building on this, DLCL [137] employs an adversarial autoencoder to learn from these two POI graphs. CGAL [138] enhances this model by introducing collective adversarial learning, using an assemble-disassemble strategy where fused region representations are disaggregated to reconstruct the original graphs. It captures region similarities based on POI distributions, textual information, and temporal patterns from human trajectories. HAFusion [139] further utilizes human trajectories, POIs, and land use data to model urban regions. Each region is characterized by mobility features, POI categories, and land use counts, with an attention-based encoder capturing intra- and inter-view region correlations, while a fusion module integrates multi-view embeddings and captures higher-order correlations.

Another line of research extensively employs GNNs. MVURE [140] uses human trajectories, POIs, and check-ins to construct four graph views, encoding each view with a graph attention network. It implements cross-view information sharing through attention mechanisms and fuses the views using adaptive weighting, aiming to reconstruct mobility patterns and region relatedness from POIs and check-ins. HREP [144] enhances this by defining multiple edge types from human mobility, POIs, and geographic context, with source and target edges based on trajectories, POI similarities, and geographic adjacency. It learns region representations using an attention-based GNN and three objectives: geographic proximity loss, mobility reconstruction, and POI correlation reconstruction. Additionally, HREP uses prompt learning (prefix-tuning) to adapt learned representations for downstream tasks, a method borrowed from NLP. In addition, HUGAT [142] utilizes POI check-in trajectories and land use data for region representation learning. It defines five types of nodes, e.g., regions and POI categories, and two types of edges for spatial and temporal relations. This method then constructs a heterogeneous information network and designs five meta-paths for interesting relations like regions that are popular destinations at the same time. A heterogeneous graph attention network is employed to derive region representations by minimizing the difference between estimated and actual mobility patterns of regions, check-in distributions, and land use distribution. Region2Vec [141] constructs a multi-graph using human trajectories, spatial adjacency, and POI distribution embeddings from a knowledge graph. GNN and attention-based fusion layers are employed to create region representations, with objectives of reconstructing mobility patterns, maintaining spatial adjacency, and preserving POI distribution similarities.

The third line employing POIs and human trajectories focuses on the application of the contrastive learning paradigm. ReMVC [143] is a dual-view approach integrating POIs and human mobility data. For the POI view, it uses region-level POI category proportions as raw features, and performs data augmentation through random POI insertion, deletion, and replacement. These augmented POI views of regions serve as positive samples, whereas POI data from differing regions are used as negative samples for the contrastive learning. For the human mobility view, it constructs two heatmaps to represent the source and destination patterns of each region, followed by the augmentation of Gaussian noise injection to form positive samples for the contrastive learning. Furthermore, an inter-view contrastive learning objective is employed to ensure that repre-

sentations of a region from different perspectives are similar, while the representations of different regions are distinguishable. ReCP [145] develops a fusion technique for multiple information views from an information theory perspective. Apart from maximizing the mutual information (consistency) shared between different views, this method implements a dual prediction strategy to minimize the conditional entropy between representations from different views, thereby reducing the inconsistency between views.

Imagery Data. Imagery data, including RS images and SVIs, have long been established to represent the physical appearance of urban environments from both aerial and ground-level human perspectives [183, 184]. Such visual appearances of urban environments can be used to partially reflect the socioeconomic factors in cities. Therefore, it has become increasingly popular to utilize imagery data for learning region representations, often in conjunction with additional data sources and context information, such as spatial proximity.

Tile2Vec [147] generates representations based on RS images patches for square-shaped regions. It employs a CNN model to encode remote sensing images and a triplet loss to ensure that patches which are geographically adjacent are also close in the embedding space. RegionEncoder [146] leverages RS images, POIs, and human trajectories for learning region representations. Initially, it employs a denoising autoencoder to extract representations from RS images. Following this, a region graph is constructed, utilizing region-level POI features as node attributes and human trajectory data to define inter-region connectivity through edges. The model is trained with two reconstruction losses, and a cross-modal alignment objective. The study in [150] adopts a contrastive learning strategy to maximize the similarity of representations derived from adjacent RS images and those that have similar POI distributions. The representations learned from the two pathways of each region are adaptively fused using learnable weights in downstream tasks.

Urban2Vec [148] derives the initial region representations by averaging all the SVI representations obtained from an image encoder for each region. It subsequently fine-tunes this image encoder with a spatial proximity-based triplet loss to bring SVIs that are spatially close together in the embedding space. Besides, this method integrates the POIs by further generating the POI representation of a region that encapsulate all the words associated with the POIs in each region. It introduces another triplet loss designed to merge POI information into the region representations by minimizing the distance between each region's representation and its corresponding POI representation. M3G [149] extends Urban2Vec by simultaneously utilizing SVIs, POIs, and human trajectories. M3G enhances the model by conducting region-level contrastive learning, selecting other regions that are either spatially close or connected through human mobility as positive samples. In another work, GeoHG [151] used the semantic segmentation features of the European Space Agency images system for processing remote sensing images, and constructed a heterogeneous graph to capture the spatial adjacency and similarities at the region level reflected from the similarities in environmental and socioeconomic characteristics. The graph model was pretrained using contrastive learning.

Another interesting region embedding study using both street view and satellite images, in conjunction with POIs and human trajectories is MuseCL [152]. The method carries out contrastive learning at multiple levels. First, contrast was conducted for street view images based on travel pattern similarity reflected form human trajectories. Likewise, contrast was carried out for remote sensing images based on POI similarity. The rationale of such pairing is that street view and mobility data entail human movement patterns, while remote sensing and POI data both capture the built environment and land use. The final region embeddings are obtained through fusing street view and remote sensing features, and supplemented with textual information from POIs.

With the rapid advancements in vision foundation models that are pretrained with vast amounts of images or image-text pairs, leveraging the power of foundation models for region embedding has become a trend. To this end, [153] proposed ReFound with POIs and remote sensing images. One major argument here is the distillation of knowledge contained in foundation models is beneficial. To this end, the POI embeddings are enforced to be close to the embeddings of LLM-generated text, image embeddings are pushed towards the representations generated by vision foundation models, and the similarities between image-POI pairs also imitate the similarities derived from a vision-language model. ReFound is finally pretrained by combining distillation objectives, masked data modeling, and cross-modal modality alignment.

OpenStreetMap. A new trend in region representation is to utilize data from OpenStreetMap (OSM) [159], a global-scale open geospatial dataset contributed by a community of mappers. OSM offers a extensive repository of geospatial entities such as building footprints and road networks, serving as valuable and accessible resources for learning effective region representations. RegionDCL [154] extracts building footprints and POIs from OSM, and begins by encoding the shape information from building footprints using a CNN-based model to generate initial region features. Furthermore, RegionDCL addresses the challenge of empty areas, which are spaces not explicitly represented in discrete geospatial vector data but are physically present. To effectively represent these empty areas in the final region embeddings, this method employs Poisson Disk Sampling to fill these gaps. Each building group–small regions partitioned by road networks–is then processed through a distance-biased Transformer. This Transformer is trained using building group-level contrastive learning. Subsequently, these building groups are aggregated into larger regions for a second round of contrastive learning, employing a triplet loss with an adaptive margin to refine the final region embeddings.

5.1.3. Applications

As urban regions serve as a critical analytical scale for various urban analyses and prediction tasks, the learned region representations are used in a diverse range of downstream tasks.

Commonly, these region embeddings are utilized by integrating them as frozen inputs into shallow task predictors, such as MLP, for making task-specific inferences. Additionally, several studies have enhanced the utilization of region representations in downstream tasks through advanced methods like prompt learning [144] and adaptive multi-view fusion [150]. From the perspective of downstream tasks, the predominant tasks involve predicting various socioeconomic indicators in cities. These include region-level attributes such as land use/urban function/land cover [133, 143, 145], population density [148, 130], house prices [146, 129], average income [149, 132], check-in counts [136, 138], crime rates [135, 132], service call volumes [139], GDP [128, 129], nighttime lighting [127], takeaway order volumes [150], health indices [147], and so on. Moreover, region representations are extensively utilized for tasks like similar region search[126, 131] and region or land use clustering [141] in a fully unsupervised manner, which has significant practical implications for real-world urban planning and management. Furthermore, region representations can benefit studies on human mobility in cities, such as predicting human flow and bike flow [134, 142]. Overall, the application of SSL-based region representations has proven to be diverse and effective across many critical urban analytical tasks, leading to substantial real-world benefits. As urban environments continue to expand and evolve, the role of robust region representations in tackling complex urban challenges will become increasingly crucial, paving the way for smarter, more resilient cities.

6. Multi-type Learning

While SSL techniques have demonstrated promising performance in producing representations for geospatial objects, these methods focus on separately deriving representations for individual geospatial data types. This paradigm does not consider the complex interactions and potential synergies among various geospatial data types. To this end, recent studies propose multi-type learning methods, which involve the joint modeling for multiple geospatial data types, as a step towards the development of geospatial foundation models.

One research direction involves the construction of geospatial knowledge graphs to systematically represent various geospatial objects as nodes and their relationships with diverse data sources as edges. By structuring geospatial entities within a graph-based framework, knowledge graphs enable the encoding of complex spatial, semantic, and contextual relationships among various data types. Once constructed, graph embedding techniques can be applied to derive representations that capture both the intrinsic properties of geospatial objects and their interactions within the broader spatial environment. WorldKG[185] aligns geospatial objects from OpenStreetMap with ontologies in existing knowledge graphs such as Wikidata. It structures geospatial knowledge by defining graph nodes based on geospatial objects, semantic classes, and object attributes. In contrast, frameworks such as UUKG [186], UrbanKG [187], and UrbanFlood [188] develop customized graph schemas tailored to specific research perspectives, such as user aspect [187], hierarchical spatial resolution [186], or application-oriented [188] geospatial analysis. Following the construction of these knowledge graphs, various graph embedding techniques [189] can be employed to generate representations for each node. These embeddings facilitate downstream geospatial tasks by leveraging the rich structural and semantic information encoded within the knowledge graph framework from multiple data types.

Besides, other studies formulate this problem through a heterogeneous learning framework using contrastive SSL. HOME-GCL [190] aims to derive representations for both road segments and regions. Specifically, it constructs a heterogeneous graph that incorporates multi-view intra-entity relationships based on geographical distance, functionality, and human mobility records, as well as inter-entity connections based on topological containment. A heterogeneous GNN is then applied to aggregate features among different entities. Subsequently, intra-level contrastive learning is employed to distinguish entities from the same object after graph augmentation, while inter-level contrastive learning differentiates between connected and disparate object. CityFM [191] simultaneously considers three data types. It encodes the textual information for each entity from these data types using pre-trained language models and employs contrastive learning to contrast an entity to its spatial neighbors. Besides, it incorporates visual content from regions, applying visual encoders to derive corresponding representations, which are then contrasted with textual representations. Lastly, CityFM utilizes contrastive learning to encourage road segments with similar traffic patterns to develop similar representations. For multi-type learning studies, the derived representations encode complementary interactions from multiple data types. As a result, these representations are effectively utilized in downstream applications, particularly those involving multiple data types, such as site selection.

7. Discussion

Sections 3-6 have presented specialized SSL techniques designed for different geospatial data types. Through an extensive review of prior studies, we observe a notable evolution in SSL methodologies for geospatial representation learning, characterized by the following key trends:

1. **Transition from Single-Source to Multi-Source Context Integration**: Early research predominantly derived representations from a single source of context information. However, recent advancements increasingly emphasize multi-source and multi-modal integration, leveraging diverse and complementary context signals to better capture intrinsic spatial patterns in geospatial data.

2. **Transition from Naive Models to Hybrid Architectures**: Initial approaches typically employ simple models (e.g., Word2vec or RNN variants) for geospatial representation learning. Over time, research has shifted towards hybrid models that combine multiple and complex architectural components, enabling more expressive modeling of geospatial context information.

3. **Influence of Foundation Models from Visual and Language Domains**: The rapid progress in visual and language foundation models has significantly influenced geospatial representation learning. Recent studies increasingly incorporate these foundation models as integral components or explore methods to align the implementation of these pre-trained models in the context of geospatial data.

While the studies belonging to the first two research trends have been well-established and detailed in preceding sections, the emergence of geospatial foundation models remains an evolving area with no fixed formulation and is explored from various perspectives. Recognizing this new emerging trend, we provide an overview of recent efforts in developing geospatial foundation models from multiple perspectives.

Additionally, while the primary focus of this survey is to summarize SSL-based studies that adopt a task-agnostic approach to geospatial representation learning, another important research direction involves task-specific SSL techniques. Unlike general-purpose SSL, these methods employ self-supervised objectives as auxiliary learning signals to enhance supervised tasks in geospatial applications. To offer a broader perspective, we also present a brief overview of research efforts that incorporate task-specific SSL into several typical geospatial applications.

7.1. A Step Towards Geospatial Foundation Models

Foundation models, which are fundamentally rooted in the SSL paradigm with pretext tasks such as autoregressive modeling in language domains and masked patch recovery in vision domains, have emerged as a transformative trend across various domains, including GeoAI. Foundation models have demonstrated significant advancements in understanding, reasoning, and generative capabilities across diverse data modalities, including language [43], images [51, 192], graphs [193, 194], and time series [195, 196]. Inspired by the success, researchers have increasingly explored the development of geospatial foundation models primarily based on language models from various perspectives.

One perspective involves enhancing existing foundation models with geospatial knowledge. Empirical studies have shown that LLMs, such as the GPT series, already possess a certain level of geospatial knowledge [197, 198, 199]. To further strengthen their geospatial knowledge understanding, researchers have proposed the development of geospatial language models [200, 201, 202] by fine-tuning general-purpose LLMs within an autoregressive SSL paradigm. These models are trained on domain-specific corpora, including research publications and Wikipedia pages related to geography, urban planning, and spatial sciences, enabling them to perform better on geospatial tasks such as geospatial question answering.

Beyond continued fine-tuning, another perspective explores leveraging LLMs for complex geospatial problem-solving through planning, decision-making, and pipeline orchestration. LLMs are increasingly utilized as intelligent agents capable of organizing systematic workflows for geospatial tasks. This is achieved through task decomposition, strategic tool selection, and model integration, where LLMs interact with external spatiotemporal models and analytical tools to process queries and generate informed responses. Notable examples include UrbanLLM [203], GeoGPT [204], PlanGPT [205], and MapGPT [206], which are designed to answer geospatial queries by organizing the workflow of solving these queries by calling external functions (e.g., ArcGIS) and integrating outputs from them. While these approaches contribute to the broader development of geospatial foundation models, they are often not explicitly designed for direct representation learning of geospatial objects.

Moreover, several studies have explored to enhance the interaction between LLMs and particular geospatial objects through targeted alignment strategies by SSL. For example, LAMP [207] aims to infuse fine-grained knowledge about POI into LLMs for a specific city, subsequently facilitating POI-related applications in a conversational manner. To achieve this, it structures several POI search tasks with their ground truth data as templates within the SSL corpus to fine-tune the LLM. Consequently, LAMP can solve several POI applications, such as route recommendation and location search, by formulating them as query-response processes within LLM framework. For polyline object, TrajFM [208] develops a trajectory modeling framework that pre-trains LLMs from scratch in an autoregressive manner, integrating spatial, temporal, and POI modalities for each trajectory data point. To enhance task transferability, TrajFM employs a trajectory masking and recovery scheme, which unifies various trajectory-related task generation processes by masking and reconstructing trajectory sub-segments and modalities. This enables the pre-trained model to generalize across diverse trajectory-based tasks. TrajCogn [209] transforms trajectory features into structured, natural language-like inputs by mapping these features into descriptive words, allowing LLMs to process trajectory data effectively. The model is then fine-tuned using a cross-reconstruction pretext task, enabling generalizable utility for downstream applications such as travel time estimation, destination prediction, and trajectory similarity search. For polygon object, GeoLLM [210] proposes harness the capabilities of LLMs to encode geospatial knowledge and capture regional features effectively. Specifically, it pinpoints targeted region coordinates on the map, extracts the corresponding address that contains place names from the neighborhood level up to the country, and identifies the ten nearest places. The regional coordinates with its geospatial context are then formatted into templated textual prompts as input to the LLM. After processing the prompts, the LLM is designed to be automatically fine-tuned using response variables, such as various socio-economic indicators. UrbanGPT [211] introduces a SSL instruction-tuning paradigm that seeks to align the dependencies of time and space, with the language space of LLMs. This method constructs prompts that combine textual descriptions with representations obtained from spatio-temporal learning models for LLM. The resulting output representations encapsulate both semantic information and relevant time-space dependencies for geospatial regions.

7.2. Task-specific SSL techniques

As previously discussed, task-specific SSL techniques serve as auxiliary objectives to enhance performance in specific geospatial tasks. These methods are particularly relevant for data instances such as grids (rasters) and point sensors, where various types of measurements (e.g., flows, speeds) are recorded over time, forming a collection of time series from these instances that exhibit both spatial and temporal interdependencies.

Therefore, a notable application of task-specific SSL applied to these data instances is in the application of spatio-temporal forecasting based on the collected records. For example, models like UniST [155], W-MAE [156], and GPT-ST [212] are utilized for grids or sensor points, employing a strategy of reconstructing masked features as self-supervised pre-training method to learn dynamic dependencies. The learned parameters are then fine-tuned for specific spatio-temporal forecasting datasets. Besides, SSL techniques are also integrated within multi-task learning frameworks for spatio-temporal forecasting. UrbanSTC [213] applies contrastive learning to identify grid regions in both spatial and temporal dimension with similar patterns. STGCL [214] explores various contrastive learning schemes with in the framework of spatio-temporal GNN, such as node-level and graph-level contrasts, providing some insights into effective integration strategies. ST-SSL [215] performs the adaptive augmentation over the traffic flow graph at both attribute- and structure-levels within the spatio-temporal GNN framework. It introduces two SSL auxiliary tasks to supplement the main traffic forecasting task to account for spatial and temporal heterogeneity. SSTBAN [216] incorporates a masked auto-encoder module to reconstruct masked spatio-temporal patches, thus deriving more robust representations to support the forecasting task. CL4ST [217] develops a meta view generator to automatically construct node and edge augmentation views for contrastive learning in a data-driven manner. Moreover, several studies also incorporate contrastive SSL into the scenario of POI recommendation [218, 219, 220]. These diverse methods demonstrate the versatility and potential of task-specific SSL techniques in enhancing the performance for various downstream applications regarding geospatial data.

8. Future Research Directions

In this section, we identify critical problems of existing SSL methods for geospatial objects, and outline several promising research directions for future exploration in this domain.

Selection of pretext tasks and data augmentation. Pretext tasks and data augmentation techniques play a crucial role in the effectiveness of SSL. Existing SSL methods for geospatial objects usually draw inspiration from the domains of computer vision and graph learning, employing heuristic adaptations tailored to geospatial contexts. As a result, the selection of pretext tasks and data augmentation strategies can vary significantly across different geospatial SSL implementations. While there have been efforts to systematically assess the effectiveness of different pretext tasks and data augmentation techniques in other domains [221, 222, 223], these findings may not directly translate to geospatial SSL due to the unique characteristics and diversity of geospatial data types. Therefore, there is a pressing need to investigate the impact of different pretext task and data augmentation selections specifically within the geospatial domain. Future research should explore whether certain pretext tasks or augmentation techniques offer distinct advantages for geospatial representation learning. This inquiry should ideally be guided by theoretical analyses and comprehensive empirical evaluations to ensure the robustness and generalizability of SSL techniques for geospatial objects.

Benchmarking SSL for geospatial objects. The establishment of standardized benchmarks is essential for advancing SSL techniques in geospatial domain. Unlike domains such as computer vision or natural language processing, where benchmark datasets are widely available and standardized, SSL methods for geospatial objects are currently evaluated on datasets that are often task-specific and unique to individual studies. The absence of unified benchmarking datasets hinders consistent and reliable comparisons across different methods, making it difficult to systematically assess their effectiveness and generalizability. Future research should prioritize the development of well-structured benchmarking frameworks that facilitate comprehensive evaluations of SSL models for geospatial objects. This requires a concerted effort to curate and release open-access benchmark datasets [224] that encompass diverse geospatial data types, heterogeneous contextual information, and multiple downstream tasks. Additionally, the establishment of automated benchmarking platforms would significantly advance this domain by enabling standardized evaluations of SSL models. Such platforms could provide pre-configured training pipelines, fair comparison protocols, and leaderboards, similar to benchmark suites in other AI disciplines.

Enhanced multi-modality fusion. With the rapid expansion of location-based services and spatial crowdsourcing, data from diverse sources attached to geospatial objects, previously difficult to access, becomes increasingly available, including street view images, textual comments, and videos. Several efforts have been made to integrate multi-modality models to enhance the performance of geospatial applications [225, 226, 227]. However, there remains significant untapped potential in advancing model capabilities for geospatial tasks. Integrating multi-modality data sources for geospatial objects presents both challenges and opportunities. Models designed for multi-modality data fusion must handle the discrepancies in scale, resolution, and relevance across different data sources. Future research can explore novel architectures, pretext tasks, and fusion techniques that effectively leverage the complementary information from different data sources. For example, adapting models like CLIP [51] with geospatial awareness. It would also be interesting to study how to balance the contributions of different modalities, aiming to achieve more generalizable and robust representations in real-world scenarios.

Geospatial foundation models and LLM adaptation. As discussed in Section 7, LLMs and multi-type pre-trained models opens up exciting possibilities for adaptation to the geospa-

tial context, serving as a universal basis for various geospatial applications. In terms of pathways in pre-training geospatial foundation models, future research could focus on creating SSL techniques that can be efficiently trained on massive-scale geospatial datasets [228, 229]. These models could learn generalized representations of geospatial features, patterns, and relationships, forming a foundation to be fine-tuned for specific tasks. On the other hand, adapting LLMs to the geospatial context requires novel learning paradigms that align spatial relationships and geographic context within the language space. For instance, there is a critical need to develop cross-modal training techniques that effectively bridge textual, visual, and spatial data for enhancing the applicability in more diverse scenarios and tasks in multiple modalities.

Privacy and vulnerability. The inherent nature of geospatial data, which often contains sensitive information about individuals and urban infrastructure, raises significant privacy and data vulnerability concerns [230]. For instance, aggregated trajectories, when analyzed with certain approaches, can expose individual's routes and private locations [231]. To this end, privacy-preserving techniques, such as differential privacy, could be considered within the SSL framework, especially the pre-training corpus. Moreover, federated learning approaches [232, 233] can be explored, which enable the collaborative training of models on sensitive geospatial data while avoiding the direct sharing of raw data, thus maintaining privacy. Furthermore, geospatial SSL models, like other machine learning models, are susceptible to adversarial attacks where input data is manipulated to induce model errors, which are particularly problematic in urban decision-making process. Therefore, robust SSL techniques can be developed to resist such data poisoning attacks or adversarial examples [234, 235].

9. Conclusion

This paper provides a comprehensive overview of self-supervised learning for geospatial objects in the domain of GeoAI. We develop a structured framework and introduce a systematic taxonomy that organizes SSL based on three geospatial data types and two methodology categories. For each data type, we offer detailed descriptions of methods, summarize their key features within the SSL component, and discuss their downstream applications, along with the utilization of multi-type SSL techniques. We further present studies on the emerging trends and task-specific SSL techniques. Finally, we outline several promising directions for the research in the future. As this domain continues to expand and evolve, we hope that the discussion presented in this paper will contribute to the future advancements of GeoAI.

References

[1] Z. Chen, L. Chen, G. Cong, C. S. Jensen, Location- and keyword-based querying of geo-textual data: a survey, VLDB J. 30 (4) (2021) 603–640.

[2] G. Cong, C. S. Jensen, D. Wu, Efficient retrieval of the top-k most relevant spatial web objects, Proc. VLDB Endow. 2 (1) (2009) 337–348.

[3] Y. Liu, T. N. Pham, G. Cong, Q. Yuan, An experimental evaluation of point-of-interest recommendation in location-based social networks, Proc. VLDB Endow. 10 (10) (2017) 1010–1021.

[4] Y. Chen, C. Long, G. Cong, C. Li, Context-aware deep model for joint mobility and time prediction, in: WSDM, 2020, pp. 106–114.

[5] P. Balsebre, D. Yao, G. Cong, W. Huang, Z. Hai, Mining geospatial relationships from text, Proc. ACM Manag. Data 1 (1) (2023) 93:1–93:26.

[6] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, A. K. Qin, A survey on modern deep neural network for traffic prediction: Trends, methods and challenges, IEEE Trans. Knowl. Data Eng. 34 (4) (2022) 1544–1561.

[7] H. Miao, Y. Zhao, C. Guo, B. Yang, Z. Kai, F. Huang, J. Xie, C. S. Jensen, A unified replay-based continuous learning framework for spatio-temporal prediction on streaming data, in: ICDE, 2024.

[8] G. Chi, H. Fang, S. Chatterjee, J. E. Blumenstock, Microestimates of wealth for all low-and middle-income countries, Proceedings of the National Academy of Sciences 119 (3) (2022).

[9] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, M. Burke, Using publicly available satellite imagery and deep learning to understand economic well-being in africa, Nature communications 11 (1) (2020) 2583.

[10] Eu gdpr regulation (2016).
URL `https://gdpr-info.eu`

[11] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Trans. Knowl. Data Eng. 35 (1) (2023) 857–876.

[12] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: ICML, 2020, pp. 1597–1607.

[13] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 4037–4058.

[14] M. C. Schiappa, Y. S. Rawat, M. Shah, Self-supervised learning for videos: A survey, ACM Comput. Surv. 55 (13s) (2023) 288:1–288:37.

[15] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, et al., A survey of large language models, CoRR abs/2303.18223 (2023).

[16] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, CoRR abs/2402.06196 (2024).

[17] Y. Liu, M. Jin, S. Pan, C. Zhou, Y. Zheng, F. Xia, P. S. Yu, Graph self-supervised learning: A survey, IEEE Trans. Knowl. Data Eng. 35 (6) (2023) 5879–5900.

[18] L. Wu, H. Lin, C. Tan, Z. Gao, S. Z. Li, Self-supervised learning on graphs: Contrastive, generative, or predictive, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 4216–4235.

[19] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, et al., Self-supervised learning for time series analysis: Taxonomy, progress, and prospects, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, et al., Language models are few-shot learners, in: NeurIPS, 2020.

[21] A.-X. Zhu, M. Turner, How is the third law of geography different?, Annals of GIS 28 (1) (2022) 57–67.

[22] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, Z. Huang, Self-supervised learning for recommender systems: A survey, IEEE Trans. Knowl. Data Eng. 36 (1) (2024) 335–355. doi:10.1109/TKDE.2023.3282907.
URL `https://doi.org/10.1109/TKDE.2023.3282907`

[23] W. Chen, Y. Liang, Y. Zhu, Y. Chang, K. Luo, H. Wen, L. Li, et al., Deep learning for trajectory data management and mining: A survey and beyond, CoRR abs/2403.14151 (2024).

[24] S. Wang, Z. Bao, J. S. Culpepper, G. Cong, A survey on trajectory data management, analytics, and learning, ACM Comput. Surv. 54 (2) (2022) 39:1–39:36.

[25] W. Zhang, J. Han, Z. Xu, H. Ni, H. Liu, H. Xiong, Towards urban general intelligence: A review and outlook of urban foundation models, CoRR abs/2402.01749 (2024).

[26] G. Mai, K. Janowicz, Y. Hu, S. Gao, B. Yan, R. Zhu, L. Cai, N. Lao, A review of location encoding for geoai: methods and applications, International Journal of Geographical Information Science 36 (2021) 639 – 673.

[27] S. Wang, J. Cao, P. S. Yu, Deep learning for spatio-temporal data mining: A survey, IEEE Trans. Knowl. Data Eng. 34 (8) (2022) 3681–3700.

[28] Q. Zhang, H. Wang, C. Long, L. Su, X. He, J. Chang, T. Wu, H. Yin, S. Yiu, Q. Tian, C. S. Jensen, A survey of generative techniques for spatial-temporal data mining, CoRR abs/2405.09592 (2024).

[29] R. H. Güting, An introduction to spatial database systems, the VLDB Journal 3 (1994) 357–399.

[30] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018).

[31] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, R. D. Hjelm, Deep graph infomax, in: ICLR, 2019.

[32] Y. Jiao, Y. Xiong, J. Zhang, Y. Zhang, T. Zhang, Y. Zhu, Sub-graph contrast for scalable self-supervised graph representation learning, in: ICDM, 2020, pp. 222–231.

[33] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems 32 (1) (2020) 4–24.

[34] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP, 2014, pp. 1724–1734.

[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[37] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186.

[38] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: ICLR, 2020.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[40] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: ACL, 2020, pp. 7871–7880.

[41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.

[42] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, et al., Training language models to follow instructions with human feedback, in: NeurIPS, 2022.

[43] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023).

[44] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J. Alayrac, R. Soricut, A. Lazaridou, O. Firat, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, CoRR abs/2403.05530 (2024).

[45] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, et al., Llama 2: Open foundation and fine-tuned chat models, CoRR abs/2307.09288 (2023).

[46] A. Anthropic, The claude 3 model family: Opus, sonnet, haiku, Claude-3 Model Card 1 (2024).

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[48] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, CoRR abs/1704.04861 (2017).

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: ICCV, 2021, pp. 9992–10002.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021, pp. 8748–8763.

[52] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: ICML, Vol. 162, 2022, pp. 12888–12900.

[53] B. Yan, K. Janowicz, G. Mai, S. Gao, From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts, in: SIGSPATIAL, 2017.

[54] W. Huang, L. Cui, M. Chen, D. Zhang, Y. Yao, Estimating urban functional distributions with semantics preserved poi embedding, International Journal of Geographical Information Science 36 (10) (2022) 1905–1930.

[55] Y. Yao, Q. Zhu, Z. Guo, W. Huang, Y. Zhang, X. Yan, A. Dong, Z. Jiang, H. Liu, Q. Guan, Unsupervised land-use change detection using multi-temporal poi embedding, International Journal of Geographical Information Science 37 (11) (2023) 2392–2415.

[56] D. Zhang, R. Xu, W. Huang, K. Zhao, M. Chen, Towards an integrated view of semantic annotation for pois with spatial and textual information, in: IJCAI, 2023, pp. 2441–2449.

[57] S. Li, J. Zhou, T. Xu, H. Liu, X. Lu, H. Xiong, Competitive analysis for points of interest, in: KDD, 2020, p. 1265–1274.

[58] Y. Chen, X. Li, G. Cong, C. Long, Z. Bao, S. Liu, W. Gu, F. Zhang, Points-of-interest relationship inference with spatial-enriched graph neural networks, Proc. VLDB Endow. 15 (3) (2021) 504–512.

[59] Y. Gao, Y. Xiong, S. Wang, H. Wang, Geobert: Pre-training geospatial representation learning on point-of-interest, Applied Sciences 12 (24) (2022).

[60] Z. Li, J. Kim, Y.-Y. Chiang, M. Chen, SpaBERT: A pretrained language model from geographic data for geo-entity representation, in: EMNLP Findings, 2022, pp. 2757–2769.

[61] R. Ding, B. Chen, P. Xie, F. Huang, X. Li, Q. Zhang, Y. Xu, Mgeo: Multi-modal geographic language model pre-training, in: SIGIR, 2023, p. 185–194.

[62] X. Liu, Y. Liu, X. Li, Exploring the context of locations for personalized location recommendations, in: IJCAI, 2016, p. 1188–1194.

[63] B. Chang, Y. Park, D. Park, S. Kim, J. Kang, Content-aware hierarchical point-of-interest embedding model for successive poi recommendation, in: IJCAI, 2018, p. 3301–3307.

[64] Y. Zhou, Y. Huang, Deepmove: Learning place representations through large scale movement data, in: 2018 IEEE International Conference on Big Data (IEEE Big Data), 2018, pp. 2403–2412.

[65] M. Chen, L. Zhu, R. Xu, Y. Liu, X. Yu, Y. Yin, Embedding hierarchical structures for venue category representation, ACM Trans. Inf. Syst. 40 (3) (2021).

[66] S. Feng, G. Cong, B. An, Y. M. Chee, Poi2vec: geographical latent representation for predicting future visitors, in: AAAI, 2017, p. 102–108.

[67] H. Wan, Y. Lin, S. Guo, Y. Lin, Pre-training time-aware location embeddings from spatial-temporal trajectories, IEEE Transactions on Knowledge and Data Engineering 34 (11) (2022) 5510–5523.

[68] S. Jiang, W. He, L. Cui, Y. Xu, L. Liu, Modeling long- and short-term user preferences via self-supervised learning for next poi recommendation, ACM Trans. Knowl. Discov. Data 17 (9) (Jun. 2023).

[69] Y. Qiang, J. Zheng, L. Wu, H. Wen, J. Lou, M. Deng, A momentum contrastive learning framework for query-poi matching, in: 2024 IEEE International Conference on Data Mining (ICDM), 2024, pp. 833–838.

[70] J. Bing, M. Chen, M. Yang, W. Huang, Y. Gong, L. Nie, Pre-trained semantic embeddings for poi categories based on multiple contexts, IEEE Transactions on Knowledge and Data Engineering 35 (9) (2023) 8893–8904.

[71] J. Huang, H. Wang, Y. Sun, Y. Shi, Z. Huang, A. Zhuo, S. Feng, Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps, in: KDD, 2022, p. 3029–3039.

[72] S. Li, J. Zhou, J. Liu, T. Xu, E. Chen, H. Xiong, Multi-temporal relationship inference in urban areas, in: KDD, 2023, p. 1316–1327.

[73] Y. Lai, Y. Su, L. Wei, T. He, H. Wang, G. Chen, D. Zha, Q. Liu, X. Wang, Disentangled contrastive hypergraph learning for next poi recommendation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, 2024, p. 1452–1462.

[74] J. Cheng, J. Wang, Y. Zhang, J. Ji, Y. Zhu, Z. Zhang, X. Zhao, Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning, in: Proceedings of the AAAI conference on artificial intelligence, 2025, pp. 1–12.

[75] W. Chen, H. Huang, Z. Zhang, T. Wang, Y. Lin, L. Chang, H. Wan, Next-poi recommendation via spatial-temporal knowledge graph contrastive learning and trajectory prompt, IEEE Transactions on Knowledge and Data Engineering (2025) 1–14.

[76] X. Rao, R. Jiang, S. Shang, L. Chen, P. Han, B. Yao, P. Kalnis, Next point-of-interest recommendation with adaptive graph contrastive learning, IEEE Transactions on Knowledge and Data Engineering 37 (3) (2025) 1366–1379.

[77] T. S. Jepsen, C. S. Jensen, T. D. Nielsen, K. Torp, On network embedding for machine learning on road networks: A case study on the danish road network, in: IEEE International Conference on Big Data (IEEE BigData), 2018, pp. 3422–3431.

[78] M. Wang, W. Lee, T. Fu, G. Yu, Learning embeddings of intersections on road networks, in: SIGSPATIAL, 2019, pp. 309–318.

[79] T. Jepsen, C. S. Jensen, T. D. Nielsen, Relational fusion networks: Graph convolutional networks for road networks, IEEE Transactions on Intelligent Transportation Systems 23 (1) (2020) 418–429.

[80] L. Zhang, C. Long, Road network representation learning: A dual graph-based approach, ACM Transactions on Knowledge Discovery from Data 17 (9) (2023) 1–25.

[81] M.-X. Wang, W.-C. Lee, T.-Y. Fu, G. Yu, On representation learning for road networks, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (1) (2020) 1–27.

[82] Y. Chang, E. Tanin, X. Cao, J. Qi, Spatial structure-aware road network embedding via graph contrastive learning, in: EDBT, 2023, pp. 144–156.

[83] N. Wu, X. W. Zhao, J. Wang, D. Pan, Learning effective road network representation with hierarchical graph neural networks, in: KDD, 2020, pp. 6–14.

[84] Y. Chen, X. Li, G. Cong, Z. Bao, C. Long, Y. Liu, A. K. Chandran, R. Ellison, Robust road network representation learning: When traffic patterns meet traveling semantics, in: CIKM, 2021, pp. 211–220.

[85] Y. Chen, X. Li, G. Cong, Z. Bao, C. Long, Semantic-enhanced representation learning for road networks with temporal dynamics, CoRR abs/2403.11495 (2024).

[86] S. Schestakov, P. Heinemeyer, E. Demidova, Road network representation learning with vehicle trajectories, in: PAKDD, 2023, pp. 57–69.

[87] Z. Mao, Z. Li, D. Li, L. Bai, R. Zhao, Jointly contrastive representation learning on road network and trajectory, in: CIKM, 2022, pp. 1501–1510.

[88] C. Han, J. Wang, Y. Wang, Y. Xie, H. Lin, C. Li, J. Wu, Bridging traffic state and trajectory for dynamic road network and trajectory representation learning, in: AAAI, 2025.

[89] H. Zhou, W. Huang, Y. Chen, T. He, G. Cong, Y. S. Ong, Road network representation learning with the third law of geography, in: NeurIPS, 2024.

[90] M. Chen, Z. Li, W. Huang, Y. Gong, Y. Yin, Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology, in: KDD, 2024.

[91] D. Yao, C. Zhang, Z. Zhu, J. Huang, J. Bi, Trajectory clustering via deep representation learning, in: IJCNN, 2017, pp. 3880–3887.

[92] X. Li, K. Zhao, G. Cong, C. S. Jensen, W. Wei, Deep representation learning for trajectory similarity computation, in: ICDE, 2018, pp. 617–628.

[93] Y. Liu, K. Zhao, G. Cong, Z. Bao, Online anomalous trajectory detection with deep generative sequence modeling, in: ICDE, 2020, pp. 949–960.

[94] C. Park, T. Kim, J. Hong, M. Choi, J. Choo, Pre-training contextual location embeddings in personal trajectories via efficient hierarchical location representations, in: ECML PKDD 2023, 2023, pp. 125–140.

[95] Q. Jing, S. Liu, X. Fan, J. Li, D. Yao, B. Wang, J. Bi, Can adversarial training benefit trajectory representation?: An investigation on robustness for trajectory similarity computation, in: CIKM, 2022, pp. 905–914.

[96] Z. Fang, Y. Du, L. Chen, Y. Hu, Y. Gao, G. Chen, E^2dtc: An end to end deep trajectory clustering framework via self-training, in: ICDE, 2021, pp. 696–707.

[97] M. Hu, Z. Zhong, X. Zhang, Y. Li, Y. Xie, X. Jia, X. Zhou, J. Luo, Self-supervised pre-training for robust and generic spatial-temporal representations, in: ICDM, 2023, pp. 150–159.

[98] D. A. Tedjopurnomo, X. Li, Z. Bao, G. Cong, F. M. Choudhury, A. K. Qin, Similar trajectory search with spatio-temporal deep representation learning, ACM Trans. Intell. Syst. Technol. 12 (6) (2021) 77:1–77:26.

[99] X. Han, R. Cheng, C. Ma, T. Grubenmann, Deeptea: Effective and efficient online time-dependent trajectory outlier detection, Proc. VLDB Endow. 15 (7) (2022) 1493–1505.

[100] Y. Zhu, J. J. Yu, X. Zhao, X. Wei, Y. Liang, Unitraj: Learning a universal trajectory foundation model from billion-scale worldwide traces, CoRR abs/2411.03859 (2024).

[101] L. Deng, Y. Zhao, Z. Fu, H. Sun, S. Liu, K. Zheng, Efficient trajectory similarity computation with contrastive learning, in: CIKM, 2022, pp. 365–374.

[102] Y. Chang, J. Qi, Y. Liang, E. Tanin, Contrastive trajectory similarity learning with dual-feature attention, in: ICDE, 2023, pp. 2933–2945.

[103] Z. Chen, D. Zhang, S. Feng, K. Chen, L. Chen, P. Han, S. Shang, KGTS: contrastive trajectory similarity learning over prompt knowledge graph embedding, in: AAAI, 2024, pp. 8311–8319.

[104] Z. Chen, K. Li, S. Zhou, L. Chen, S. Shang, Towards robust trajectory similarity computation: Representation-based spatio-temporal similarity quantification, World Wide Web Journal 26 (4) (2023) 1271–1294.

[105] H. Wu, Z. Chen, W. Sun, B. Zheng, W. Wang, Modeling trajectories with recurrent neural networks, in: IJCAI, 2017, pp. 3083–3090.

[106] T. Fu, W. Lee, Trembr: Exploring road networks for trajectory representation learning, ACM Trans. Intell. Syst. Technol. 11 (1) (2020) 10:1–10:25.

[107] J. Li, M. Wang, L. Li, K. Xin, W. Hua, X. Zhou, Trajectory representation learning based on road network partition for similarity computation, in: DASFAA, 2023, pp. 396–413.

[108] Z. Ma, Z. Tu, X. Chen, Y. Zhang, D. Xia, G. Zhou, Y. Chen, Y. Zheng, J. Gong, More than routing: Joint GPS and route modeling for refine trajectory representation learning, in: WWW, 2024, pp. 3064–3075.

[109] Z. Fang, Y. Du, X. Zhu, D. Hu, L. Chen, Y. Gao, C. S. Jensen, Spatio-temporal trajectory similarity learning in road networks, in: A. Zhang, H. Rangwala (Eds.), KDD, 2022, pp. 347–356.

[110] S. Zhou, J. Li, H. Wang, S. Shang, P. Han, GRLSTM: trajectory similarity computation with graph-based residual LSTM, in: AAAI, 2023, pp. 4972–4980.

[111] S. B. Yang, C. Guo, J. Hu, J. Tang, B. Yang, Unsupervised path representation learning with curriculum negative sampling, in: IJCAI, 2021, pp. 3286–3292.

[112] Y. Lin, H. Wan, S. Guo, J. Hu, C. S. Jensen, Y. Lin, Pre-training general trajectory embeddings with maximum multi-view entropy coding, IEEE Transactions on Knowledge and Data Engineering (2023) 1–15.

[113] H. Liu, J. Han, Y. Fu, Y. Li, K. Chen, H. Xiong, Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training, VLDB J. 32 (2) (2023) 325–342.

[114] J. Jiang, D. Pan, H. Ren, X. Jiang, C. Li, J. Wang, Self-supervised trajectory representation learning with temporal regularities and travel semantics, in: ICDE, 2023, pp. 843–855.

[115] S. B. Yang, J. Hu, C. Guo, B. Yang, C. S. Jensen, Lightpath: Lightweight and scalable path representation learning, in: KDD, 2023, pp. 2999–3010.

[116] S. Zhou, S. Shang, L. Chen, P. Han, C. S. Jensen, Terra: A multimodal spatio-temporal dataset spanning the earth, in: KDD, 2025.

[117] Y. Lin, H. Wan, S. Guo, Y. Lin, Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction, in: AAAI, 2021, pp. 4241–4248.

[118] A. Liu, Y. Zhang, X. Zhang, G. Liu, Y. Zhang, Z. Li, L. Zhao, Q. Li, X. Zhou, Representation learning with multi-level attention for activity trajectory similarity computation, IEEE Trans. Knowl. Data Eng. 34 (5) (2022) 2387–2400.

[119] F. Zhou, Y. Dai, Q. Gao, P. Wang, T. Zhong, Self-supervised human mobility learning for next location prediction and trajectory classification, Knowl. Based Syst. 228 (2021) 107214.

[120] A. Liang, B. Yao, J. Xie, W. Zheng, Y. Shen, Q. Ge, Clmtr: a generic framework for contrastive multi-modal trajectory representation learning, GeoInformatica (2024) 1–21.

[121] L. Gong, Y. Lin, S. Guo, Y. Lin, T. Wang, E. Zheng, Z. Zhou, H. Wan, Contrastive pre-training with adversarial perturbations for check-in sequence representation learning, in: AAAI, 2023, pp. 4276–4283.

[122] S. Zhou, S. Shang, L. Chen, C. S. Jensen, P. Kalnis, Red: Effective trajectory representation learning with comprehensive information, Proc. VLDB Endow. 18 (2) (2025) 80–92.

[123] Y. Lin, Y. Liu, Z. Zhou, H. Wen, E. Zheng, S. Guo, Y. Lin, H. Wan, Ptrajm: Efficient and semantic-rich trajectory learning with pretrained trajectory-mamba, CoRR abs/2408.04916 (2024).

[124] T. Qian, J. Li, Y. Chen, G. Cong, T. Sun, F. Wang, Y. Xu, Context-enhanced multi-view trajectory representation learning: Bridging the gap through self-supervised models, CoRR abs/2410.13196 (2024).

[125] R. Xu, H. Cheng, C. Guo, H. Gao, J. Hu, S. B. Yang, B. Yang, Mm-path: Multi-modal, multi-granularity path representation learning, CoRR abs/2411.18428 (2024).

[126] Y. Liu, K. Zhao, G. Cong, Efficient similar region search with deep metric learning, in: KDD, 2018, pp. 1850–1859.

[127] X. Hao, W. Chen, Y. Yan, S. Zhong, K. Wang, Q. Wen, Y. Liang, Urbanvlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction, arXiv preprint arXiv:2403.16831 (2024).

[128] L. Bai, W. Huang, X. Zhang, S. Du, G. Cong, H. Wang, B. Liu, Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois, ISPRS Journal of Photogrammetry and Remote Sensing 201 (2023) 193–208.

[129] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, Y. Liang, Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 4006–4017.

[130] W. Huang, D. Zhang, G. Mai, X. Guo, L. Cui, Learning urban region representations with pois and hierarchical graph infomax, ISPRS Journal of Photogrammetry and Remote Sensing 196 (2023) 134–145.

[131] X. Jin, B. Oh, S. Lee, D. Lee, K.-H. Lee, L. Chen, Learning region similarity over spatial knowledge graphs with hierarchical types and semantic relations, in: CIKM, 2019, pp. 669–678.

[132] H. Wang, Z. Li, Region representation learning via mobility flow, in: CIKM, 2017, pp. 237–246.

[133] Z. Yao, Y. Fu, B. Liu, W. Hu, H. Xiong, Representing urban functions through zone embedding with human mobility patterns, in: IJCAI, 2018.

[134] Z. Liu, F. Miranda, W. Xiong, J. Yang, Q. Wang, C. Silva, Learning geo-contextual embeddings for commuting flow prediction, in: AAAI, 2020, pp. 808–816.

[135] S. Wu, X. Yan, X. Fan, S. Pan, S. Zhu, C. Zheng, M. Cheng, C. Wang, Multi-graph fusion networks for urban region embedding, arXiv preprint arXiv:2201.09760 (2022).

[136] Y. Fu, P. Wang, J. Du, L. Wu, X. Li, Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations, in: AAAI, 2019, pp. 906–913.

[137] J. Du, Y. Zhang, P. Wang, J. Leopold, Y. Fu, Beyond geo-first law: Learning spatial representations via integrated autocorrelations and complementarity, in: ICDM, 2019, pp. 160–169.

[138] Y. Zhang, Y. Fu, P. Wang, X. Li, Y. Zheng, Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning, in: KDD, 2019, pp. 1700–1708.

[139] F. Sun, J. Qi, Y. Chang, X. Fan, S. Karunasekera, E. Tanin, Urban region representation learning with attentive fusion, in: ICDE, 2024, pp. 4409–4421.

[140] M. Zhang, T. Li, Y. Li, P. Hui, Multi-view joint graph representation learning for urban region embedding, in: IJCAI, 2021, pp. 4431–4437.

[141] Y. Luo, F.-l. Chung, K. Chen, Urban region profiling via multi-graph representation learning, in: CIKM, 2022, pp. 4294–4298.

[142] N. Kim, Y. Yoon, Effective urban region representation learning using heterogeneous urban graph attention network, arXiv preprint arXiv:2202.09021 (2022).

[143] L. Zhang, C. Long, G. Cong, Region embedding with intra and inter-view contrastive learning, IEEE Transactions on Knowledge and Data Engineering 35 (9) (2022) 9031–9036.

[144] S. Zhou, D. He, L. Chen, S. Shang, P. Han, Heterogeneous region embedding with prompt learning, in: AAAI, 2023, pp. 4981–4989.

[145] Z. Li, W. Huang, K. Zhao, M. Yang, Y. Gong, M. Chen, Urban region embedding via multi-view contrastive prediction, in: AAAI, 2024, pp. 8724–8732.

[146] P. Jenkins, A. Farag, S. Wang, Z. Li, Unsupervised representation learning of spatial data via multimodal embedding, in: CIKM, 2019, pp. 1993–2002.

[147] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, S. Ermon, Tile2vec: Unsupervised representation learning for spatially distributed data, in: AAAI, 2019, pp. 3967–3974.

[148] Z. Wang, H. Li, R. Rajagopal, Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding, in: AAAI, 2020, pp. 1013–1020.

[149] T. Huang, Z. Wang, H. Sheng, A. Y. Ng, R. Rajagopal, Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond, arXiv preprint arXiv:2105.02489 (2021).

[150] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma, P. Hui, Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3308–3316.

[151] X. Zou, J. Huang, X. Hao, Y. Yang, H. Wen, Y. Yan, C. Huang, Y. Liang, Learning geospatial region embedding with heterogeneous graph, arXiv preprint arXiv:2405.14135 (2024).

[152] X. Yong, X. Zhou, Musecl: predicting urban socioeconomic indicators via multi-semantic contrastive learning, in: IJCAI, 2024, pp. 7536–7544.

[153] C. Xiao, J. Zhou, Y. Xiao, J. Huang, H. Xiong, Refound: Crafting a foundation model for urban region understanding upon language and visual foundations, in: KDD, 2024, pp. 3527–3538.

[154] Y. Li, W. Huang, G. Cong, H. Wang, Z. Wang, Urban region representation learning with openstreetmap building footprints, in: KDD, 2023, pp. 1363–1373.

[155] Y. Yuan, J. Ding, J. Feng, D. Jin, Y. Li, Unist: A prompt-empowered universal model for urban spatio-temporal prediction, in: KDD, 2024.

[156] X. Man, C. Zhang, C. Li, J. Shao, W-MAE: pre-trained weather model with masked autoencoder for multi-variable weather forecasting, CoRR abs/2304.08754 (2023).

[157] M. A. Islam, M. M. Mohammad, S. S. S. Das, M. E. Ali, A survey on deep learning based point-of-interest (POI) recommendations, Neurocomputing 472 (2022) 306–325.

[158] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013, pp. 3111–3119.

[159] OpenStreetMap, `https://www.openstreetmap.org` (2017).

[160] K. Sahr, D. White, A. J. Kimerling, Geodesic discrete global grid systems, Cartography and Geographic Information Science 30 (2) (2003) 121–134.

[161] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, R. D. Hjelm, Deep graph infomax., ICLR 2 (3) (2019) 4.

[162] P. Balsebre, D. Yao, G. Cong, Z. Hai, Geospatial entity resolution, in: Proceedings of the ACM Web Conference, 2022, p. 3061–3070.

[163] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: KDD, 2016, pp. 855–864.

[164] T. S. Jepsen, C. S. Jensen, T. D. Nielsen, Graph convolutional networks for road networks, in: SIGSPATIAL, 2019, pp. 460–463.

[165] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and pois, in: KDD, 2012, pp. 186–194.

[166] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, in: NeurIPS, 2020.

[167] Q. Zhang, Z. Wang, C. Long, C. Huang, S. Yiu, Y. Liu, G. Cong, J. Shi, Online anomalous subtrajectory detection on road networks with deep reinforcement learning, in: ICDE, pp. 246–258.

[168] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: NIPS, 2014, pp. 3104–3112.

[169] P. Yang, H. Wang, Y. Zhang, L. Qin, W. Zhang, X. Lin, T3S: effective representation learning for trajectory similarity computation, in: ICDE, 2021, pp. 2183–2188.

[170] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[171] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: NIPS, 2015, pp. 802–810.

[172] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063.

[173] C. Yang, G. Gidofalvi, Fast map matching, an algorithm integrating hidden markov model with precomputation, International Journal of Geographical Information Science 32 (3) (2018) 547–570.

[174] P. Newson, J. Krumm, Hidden markov map matching through noise and sparseness, in: SIGSPATIAL, 2009, pp. 336–343.

[175] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: AAAI, 2014, pp. 1112–1119.
[176] Y. Zhang, A. Liu, G. Liu, Z. Li, Q. Li, Deep representation learning of activity trajectory similarity computation, in: ICWS, 2019, pp. 312–319.
[177] C. Duan, W. Fan, W. Zhou, H. Liu, J. Wen, Clsprec: Contrastive learning of long and short-term preferences for next POI recommendation, in: CIKM, 2023, pp. 473–482.
[178] Z. Jia, Y. Fan, J. Zhang, C. Wei, R. Yan, X. Wu, Improving next location recommendation services with spatial-temporal multi-group contrastive learning, IEEE Transactions on Services Computing 16 (5) (2023) 3467–3478.
[179] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, CoRR abs/2312.00752 (2023).
[180] Y. Chang, E. Tanin, G. Cong, C. S. Jensen, J. Qi, Trajectory similarity measurement: An efficiency perspective, Proc. VLDB Endow. 17 (2024).
[181] Singapore subzones.
URL https://data.gov.sg/dataset/master-plan-2019-subzone-boundary-no-sea
[182] Nyc census tracts.
URL https://www.nyc.gov/site/planning/data-maps/open-data/census-download-metadata.page
[183] J. E. Patino, J. C. Duque, A review of regional science applications of satellite remote sensing in urban settings, Computers, Environment and Urban Systems 37 (2013) 1–17.
[184] Z. Fan, F. Zhang, B. P. Loo, C. Ratti, Urban visual intelligence: Uncovering hidden city profiles with street view images, Proceedings of the National Academy of Sciences 120 (27) (2023).
[185] A. Dsouza, N. Tempelmeier, R. Yu, S. Gottschalk, E. Demidova, Worldkg: A world-scale geographic knowledge graph, in: CIKM, 2021, pp. 4475–4484.
[186] Y. Ning, H. Liu, H. Wang, Z. Zeng, H. Xiong, UUKG: unified urban knowledge graph dataset for urban spatiotemporal prediction, in: NeurIPS, 2023.
[187] Y. Liu, J. Ding, Y. Fu, Y. Li, Urbankg: An urban knowledge graph system, ACM Trans. Intell. Syst. Technol. 14 (4) (2023) 60:1–60:25.
[188] Y. Wang, F. Ye, B. Li, G. Jin, D. Xu, F. Li, Urbanfloodkg: An urban flood knowledge graph system for risk assessment, in: CIKM, 2023, pp. 2574–2584.
[189] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Trans. Knowl. Data Eng. 29 (12) (2017) 2724–2743.
[190] J. Jiang, Y. Yang, J. Wang, J. Wu, Jointly learning representations for map entities via heterogeneous graph contrastive learning, CoRR abs/2402.06135 (2024).
[191] P. Balsebre, W. Huang, G. Cong, Y. Li, City foundation models for learning general purpose representations from openstreetmap, in: CIKM, 2024.
[192] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, Trans. Mach. Learn. Res. (2022).
[193] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, J. Wen, Structgpt: A general framework for large language model to reason over structured data, in: EMNLP, 2023, pp. 9237–9251.
[194] H. Wang, S. Feng, T. He, Z. Tan, X. Han, Y. Tsvetkov, Can language models solve graph problems in natural language?, in: NeurIPS, 2023.
[195] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P. Chen, Y. Liang, Y. Li, S. Pan, Q. Wen, Time-llm: Time series forecasting by reprogramming large language models, in: ICLR, 2024.
[196] T. Zhou, P. Niu, X. Wang, L. Sun, R. Jin, One fits all: Power general time series analysis by pretrained LM, in: NeurIPS, 2023.
[197] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, N. Lao, On the opportunities and challenges of foundation models for geospatial artificial intelligence, CoRR abs/2304.06798 (2023).
[198] W. Gurnee, M. Tegmark, Language models represent space and time, in: ICLR, 2024.
[199] J. Roberts, T. Lüddecke, S. Das, K. Han, S. Albanie, GPT4GEO: how a language model sees the world's geography, CoRR abs/2306.00020 (2023).
[200] C. Deng, T. Zhang, Z. He, Q. Chen, Y. Shi, Y. Xu, L. Fu, W. Zhang, X. Wang, C. Zhou, Z. Lin, J. He, K2: A foundation language model for geoscience knowledge understanding and utilization, in: WSDM, 2024, pp. 161–170.
[201] Z. Li, W. Zhou, Y. Chiang, M. Chen, Geolm: Empowering language models for geospatially grounded language understanding, in: EMNLP, 2023, pp. 5227–5240.
[202] Y. Zhang, Z. Wang, Z. He, J. Li, G. Mai, J. Lin, C. Wei, W. Yu, Bb-geogpt: A framework for learning a large language model for geographic information science, Inf. Process. Manag. 61 (5) (2024) 103808.
[203] Y. Jiang, Q. Chao, Y. Chen, X. Li, S. Liu, G. Cong, Urbanllm: Autonomous urban activity planning and management with large language models, in: Findings of EMNLP 2024, 2024, pp. 1810–1825.
[204] Y. Zhang, C. Wei, S. Wu, Z. He, W. Yu, Geogpt: Understanding and processing geospatial tasks through an autonomous GPT, CoRR abs/2307.07930 (2023).
[205] H. Zhu, W. Zhang, N. Huang, B. Li, L. Niu, Z. Fan, T. Lun, Y. Tao, J. Su, Z. Gong, C. Fang, X. Liu, Plangpt: Enhancing urban planning with tailored language model and efficient retrieval, CoRR abs/2402.19273 (2024).
[206] Y. Zhang, Z. He, J. Li, J. Lin, Q. Guan, W. Yu, Mapgpt: an autonomous framework for mapping by integrating large language model and cartographic tools, Cartography and Geographic Information Science 51 (6) (2024) 717–743.
[207] P. Balsebre, W. Huang, G. Cong, LAMP: A language model on the map, CoRR abs/2403.09059 (2024).
[208] Y. Lin, T. Wei, Z. Zhou, H. Wen, J. Hu, S. Guo, Y. Lin, H. Wan, Trajfm: A vehicle trajectory foundation model for region and task transferability, CoRR abs/2408.15251 (2024).
[209] Z. Zhou, Y. Lin, H. Wen, S. Guo, J. Hu, Y. Lin, H. Wan, Plm4traj: Cognizing movement patterns and travel purposes from trajectories with pre-trained language models, CoRR abs/2405.12459 (2024).
[210] R. Manvi, S. Khanna, G. Mai, M. Burke, D. B. Lobell, S. Ermon, Geollm: Extracting geospatial knowledge from large language models, in: ICLR, 2024.
[211] Z. Li, L. Xia, J. Tang, Y. Xu, L. Shi, L. Xia, D. Yin, C. Huang, Urbangpt: Spatio-temporal large language models, CoRR abs/2403.00813 (2024).
[212] Z. Li, L. Xia, Y. Xu, C. Huang, GPT-ST: generative pre-training of spatio-temporal graph neural networks, in: NeurIPS, 2023.
[213] H. Qu, Y. Gong, M. Chen, J. Zhang, Y. Zheng, Y. Yin, Forecasting fine-grained urban flows via spatio-temporal contrastive self-supervision, IEEE Trans. Knowl. Data Eng. 35 (8) (2023) 8008–8023.
[214] X. Liu, Y. Liang, C. Huang, Y. Zheng, B. Hooi, R. Zimmermann, When do contrastive learning signals help spatio-temporal graph forecasting?, in: SIGSPATIAL, 2022, pp. 5:1–5:12.
[215] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, Y. Zheng, Spatio-temporal self-supervised learning for traffic flow prediction, in: AAAI, 2023, pp. 4356–4364.
[216] S. Guo, Y. Lin, L. Gong, C. Wang, Z. Zhou, Z. Shen, Y. Huang, H. Wan, Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting, in: ICDE, 2023, pp. 1585–1596.
[217] J. Tang, L. Xia, J. Hu, C. Huang, Spatio-temporal meta contrastive learning, in: CIKM, 2023, pp. 2412–2421.
[218] Q. Gao, J. Hong, X. Xu, P. Kuang, F. Zhou, G. Trajcevski, Predicting human mobility via self-supervised disentanglement learning, IEEE Trans. Knowl. Data Eng. 36 (5) (2024) 2126–2141.
[219] Q. Gao, W. Wang, K. Zhang, X. Yang, C. Miao, T. Li, Self-supervised representation learning for trip recommendation, Knowl. Based Syst. 247 (2022) 108791.
[220] S. Jiang, W. He, L. Cui, Y. Xu, L. Liu, Modeling long- and short-term user preferences via self-supervised learning for next POI recommendation, ACM Trans. Knowl. Discov. Data 17 (9) (2023) 125:1–125:20.
[221] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, M. Soljacic, Equivariant contrastive learning, CoRR abs/2111.00899 (2021).
[222] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, in: NeurIPS, 2020.
[223] T. Xiao, X. Wang, A. A. Efros, T. Darrell, What should not be contrastive in contrastive learning, in: ICLR, 2021.
[224] W. Chen, X. Hao, Y. Wu, Y. Liang, Terra: A multimodal spatio-temporal dataset spanning the earth, in: NeurIPS, 2024.
[225] W. Huang, J. Wang, G. Cong, Zero-shot urban function inference with

street view images through prompting a pretrained vision-language model, Int. J. Geogr. Inf. Sci. 38 (7) (2024) 1414–1442.
[226] V. V. Cepeda, G. K. Nayak, M. Shah, Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, in: NeurIPS, 2023.
[227] S. Xu, C. Zhang, L. Fan, G. Meng, S. Xiang, J. Ye, Addressclip: Empowering vision-language models for city-wide image address localization, arXiv preprint arXiv:2407.08156 (2024).
[228] T. Yao, X. Yi, D. Z. Cheng, F. X. Yu, T. Chen, A. K. Menon, L. Hong, E. H. Chi, S. Tjoa, J. J. Kang, E. Ettinger, Self-supervised learning for large-scale item recommendations, in: CIKM, 2021, pp. 4321–4330.
[229] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, J. Huang, Self-supervised graph transformer on large-scale molecular data, in: NeurIPS, 2020.
[230] J. Rao, S. Gao, G. Mai, K. Janowicz, Building privacy-preserving and secure geospatial artificial intelligence foundation models (vision paper), in: SIGSPATIAL, 2023, pp. 41:1–41:4.
[231] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, D. Jin, Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data, in: WWW, 2017, pp. 1241–1250.
[232] J. Feng, C. Rong, F. Sun, D. Guo, Y. Li, PMF: A privacy-preserving human mobility prediction framework via federated learning, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (1) (2020) 10:1–10:21.
[233] Z. Liu, H. Miao, Y. Zhao, C. Liu, K. Zheng, H. Li, Lighttr: A lightweight framework for federated trajectory recovery, in: ICDE, 2024.
[234] Y. Lun, H. Miao, J. Shen, R. Wang, X. Wang, S. Wang, Resisting tul attack: balancing data privacy and utility on trajectory via collaborative adversarial learning, GeoInformatica (2023) 1–21.
[235] F. Xu, Y. Li, Z. Tu, S. Chang, H. Huang, No more than what I post: Preventing linkage attacks on check-in services, IEEE Trans. Mob. Comput. 20 (2) (2021) 620–633.