Self-supervised Learning for Geospatial AI: A Survey

Yile Chen, Weiming Huang, Kaiqi Zhao, Yue Jiang, Gao Cong

***Abstract*—The proliferation of geospatial data in urban and territorial environments has significantly facilitated the development of geospatial artificial intelligence (GeoAI) across various urban applications. Given the vast yet inherently sparse labeled nature of geospatial data, there is a critical need for techniques that can effectively leverage such data without heavy reliance on labeled datasets. This requirement aligns with the principles of self-supervised learning (SSL), which has attracted increasing attention for its adoption in geospatial data. This paper conducts a comprehensive and up-to-date survey of SSL techniques applied to or developed for three primary data (geometric) types prevalent in geospatial vector data: points, polylines, and polygons. We systematically categorize various SSL techniques into predictive and contrastive methods, discussing their application with respect to each data type in enhancing generalization across various downstream tasks. Furthermore, we review the emerging trends of SSL for GeoAI, and several task-specific SSL techniques. Finally, we discuss several key challenges in the current research and outline promising directions for future investigation. By presenting a structured analysis of relevant studies, this paper aims to inspire continued advancements in the integration of SSL with GeoAI, encouraging innovative methods to harnessing the power of geospatial data.**

***Index Terms*—Geospatial artificial intelligence, self-supervised learning, spatial representation learning, contrastive learning.**

I. Introduction

The digitization of urban and territorial environments has significantly enhanced the collection of extensive data in geospatial space, including user check-ins, traffic records, building footprints, trajectories, etc. Such 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]. This vast pool of data, while rich in potential, has long faced limitations and challenges for analysis and utilization for the development of geospatial artificial intelligence (GeoAI) [10], especially for the advanced models in the context of deep learning. In particular, the models are usually trained on specific tasks with abundant labeled data (e.g., traffic records), which can be both limited and costly to be accessible, especially considering data privacy regulations [11]. Additionally, despite the commonalities shared by various urban tasks, such as the close relationship between population density and land use, these models usually suffer from limited generalization across downstream tasks due to their reliance on task-specific supervision signals.

In recent years, self-supervised learning (SSL) [12] has emerged as a paradigm that can reduce the dependency on annotated labels while producing generic and task-agnostic 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 [13], [14], videos [15], language [16], [17], graphs [18], [19], time series [20], etc. For example, massive text corpora are structured in an autoregressive way, which is suitable for next token prediction for the training of large language models (LLMs) [21]. Images can be processed through data augmentation operations to produce multiple views, with the models trained to maintain invariance across views [13].

A primary motivation of applying SSL techniques for GeoAI is to learn effective and generalizable representations (embeddings) for various forms of geospatial objects, such as 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 various urban analytical tasks. For example, road networks support most human mobility activities within a city. As a result, a range of tasks, such as the predictions of traffic speed, congestion, and destination, commonly regard individual road segments as analytical units. In this case, SSL serves as a compelling framework to derive general-purpose representations for each type of geospatial data, capturing both the properties of each object as well as the complex interplay among different objects. The derived representations can be readily utilized to train simpler models for various downstream tasks while maintaining effective performance.

Apart from the strong generalization capabilities, the interest of applying SSL techniques is also driven by their intrinsic advantages to operate without extensive labeled datasets. This characteristic aligns seamlessly with addressing a common challenge in the geospatial domain—the scarcity of labeled data. Consequently, SSL offers a viable alternative to the conventional approach of developing specific deep learning models for each urban analytical task with sufficient labeled data in a supervised learning paradigm. However, geospatial data, embedded within geographical space, presents various forms with unique spatial features and adheres to certain geographical principles, such as the Laws of Geography [22]. As

Yile Chen, Yue Jiang, and Gao Cong are with the College of Computing and Data Science, Nanyang Technological University, Singapore. Email: {yile001@e., yue013@e., gaocong@}.ntu.edu.sg

Weiming Huang is with the Department of Physical Geography and Ecosystem Science, Lund University, Sweden. Email: weiming.huang@nateko.lu.se

Kaiqi Zhao is with the School of Computer Science, University of Auckland, New Zealand. Email: kaiqi.zhao@auckland.ac.nz

a result, SSL techniques commonly applied in other domains often fall short in capturing the spatial semantics attached to geospatial data. Furthermore, certain SSL operations utilized in other domains, such as data augmentation, need to be carefully designed to reflect the characteristics of different geospatial objects.

Despite the clear distinction and a growing body of literature, the application of SSL within the domain of GeoAI, remains insufficiently discussed and explored. To bridge this gap, this survey provides a comprehensive and systematic review of the up-to-date SSL techniques tailored for geospatial data. In particular, we focus on three primary data types depending on their geometric forms, i.e., points, polylines, and polygons, which are prevalent in urban contexts. We organize a structured way to present the specialized SSL studies related to these data types, focusing particularly on those that operate independently of specific primary tasks and supervised settings. Furthermore, considering the rapid development in this domain, we discuss several emerging trends that have recently appeared, including multi-type learning and adaptation to language space. We also provide a brief review of task-specific SSL techniques applied to geospatial data. This survey aims to cover a wide range of model scopes, from specialized SSL models to advanced multi-modality models, applied to different geospatial data types. The main contributions of this survey are summarized as follows:

- We present a detailed and up-to-date review of SSL techniques for geospatial data, focusing on three types of geospatial vector data in urban environments: points, polylines, and polygons. To the best of our knowledge, this work is the first to systematically discuss SSL techniques tailored for geospatial data.
- We introduce a comprehensive and structured taxonomy for specialized SSL models designed for the studied geospatial data types. Our categorization includes an analysis of intrinsic characteristics and contextual information within each data type, encompassing discussion on both predictive and contrastive SSL implementations.
- We review several recent advancements and task-specific SSL techniques for geospatial data, providing insights into the emerging and promising trends applied to GeoAI.
- We discuss several key challenges for SSL in GeoAI, and propose promising future directions to advance this domain.

Related Surveys and Our Distinction: Several recent surveys have discussed the application of SSL, mainly focusing on other domains, such as general SSL [12], SSL for computer vision [14], graph data [18], [19], time series [20], and recommender systems [23]. Recognizing the lack of dedicated SSL exploration within the domain of GeoAI, this paper presents a comprehensive and systematic review of SSL tailored for geospatial data, which serves as the foundations in numerous downstream applications. On the other hand, several recent surveys have paid attention to spatio-temporal data analytics with various emphases, such as trajectory data mining [24], [25], urban foundation models [26], geo-location encoding [27], and supervised/generative deep learning [28], [29]. These surveys generally fall within the domain of GeoAI, and may overlap with some studies discussed in this paper. However, this survey is distinctive as it pioneers in structuring the knowledge from the unique perspective of SSL, detailing its developments across different geospatial data types and providing novel insights into the applications of SSL in geospatial contexts.

Paper Structure: The rest of this survey is organized as follows. Section II provides definitions, preliminary concepts, and background knowledge necessary for the subsequent sections. Section III, IV, and V look into the details of specialized SSL techniques applied to geospatial 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 VI highlights emerging trends in this domain, such multi-type learning and adaptation to language space, and provides an overview for task-specific SSL techniques for geospatial data. Section VII discusses several promising future research directions. Finally, Section VIII concludes this paper.

II. Definition and Background

In this section, we first 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 some preliminaries on typical models that encode these geospatial data types.

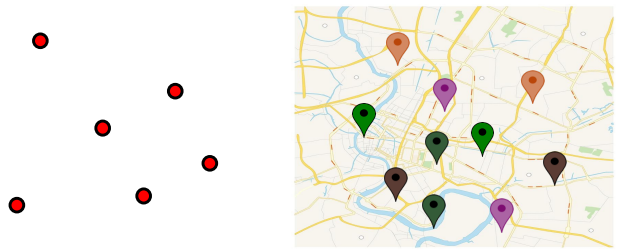
A. Definition of Geospatial Data Types

The diverse and rich data sources collected in urban spaces, which display unique geospatial characteristics, can be mainly categorized into three data types in terms of their geometric representations: points, polylines and polygons [30]. These data types are illustrated in Fig. 1. Then we provide formal definitions for each of these types to establish a clear understanding necessary for subsequent discussions and analyses.

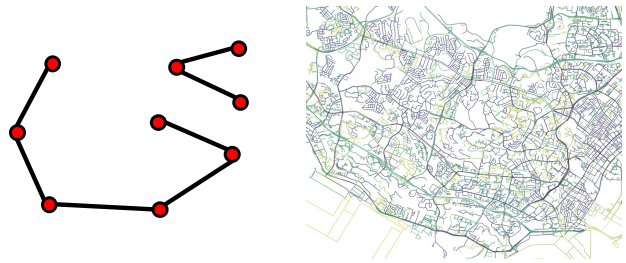
Definition 1 (**Point**)**.** A point in the geospatial context is represented as $p = (l, x)$, where l denotes the geographical coordinates, and x refers to the associated features of the point, such as attributes or readings. This data type indicates the spatial locations equipped with contextual information,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

Definition 2 (**Polyline**)**.** A polyline in the geospatial context 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.

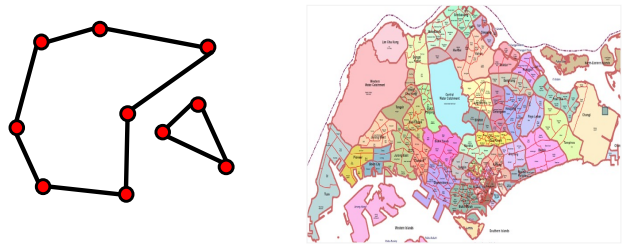
Definition 3 (**Polygon**)**.** A polygon in the geospatial context is defined as a closed shape consisting of a sequence of line segments that connect to enclose an area, denoted by



(a) Points



(b) Polylines



(c) Polygons

Fig. 1. Three types of geospatial objects and their data instances.

$\mathcal{B} = [(l_1, l_2), ..., (l_{n-1}, l_n), (l_n, l_1)]$, where l_i represents the geographical coordinates of the i-th vertex. The vertices are sequentially linked, with the last vertex reconnecting to the first vertex to complete 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. In urban contexts, 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 its enclosed area.

B. Paradigms for Self-supervised Learning

SSL techniques for geospatial objects can be divided into two categories: predicative and contrastive, according to the design of pretext tasks.

1) Predicative 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 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. $\mathcal{D}$ 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) 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. $\tilde{\mathcal{D}}^1$ 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 [12], such as InfoNCE [31], JS divergence [32] and triplet loss [33].

C. 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 that are frequently employed or adapted as geospatial encoders.

1) Graph Neural Networks: Graph Neural Networks (GNNs) [34] 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 $N(v)$ 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, GNN

are frequently utilized to model discrete objects, enabling the capture of complex relationships among them.

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:

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

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 [35] and LSTM [36] being the most notable ones. In recent years, the Transformer architecture [37] 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.

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 [38]. These models [39]–[42] leverage large-scale datasets 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 (i.e., next token prediction), exemplified by ChatGPT and its counterparts [43]–[47]. 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 [48], [49] and recent Transformer-based models [50], [51] 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 [52], [53] are employed to synergize the semantics of the two data modalities, enhancing the interpretability and utility of combined data sources in geospatial applications.

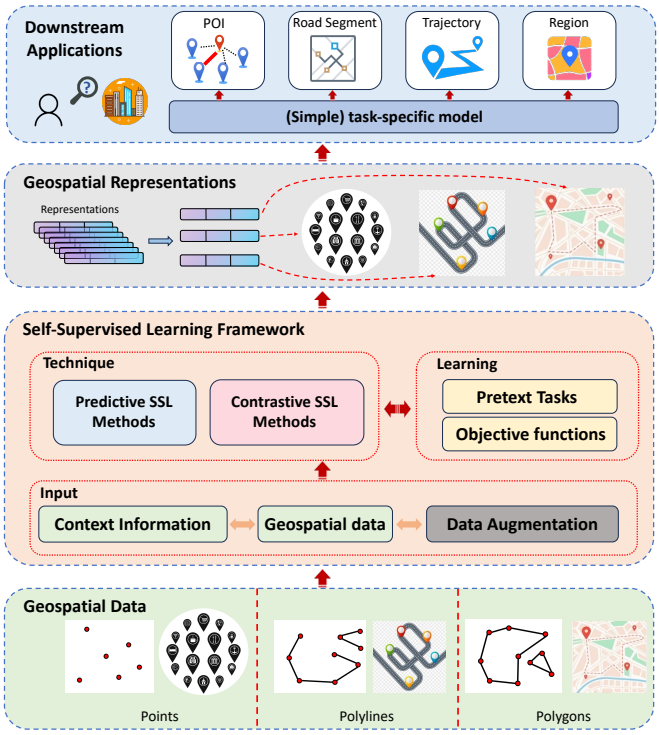


Fig. 2. Overview of SSL framework for geospatial data.

In the following sections, for each data type under consideration, our analysis adopts a structured way. We first present the encoding methods for intrinsic attributes that reflect the fundamental characteristics of the data instance. Then we discuss the context information leveraged to enhance the extraction of meaningful insights and knowledge. Finally, we introduce the downstream applications that benefit from the derived geospatial representations. The overview of applying SSL on geospatial data is illustrated in Fig. 2.

III. 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. Recent studies have focused on learning effective representations for points with SSL techniques, enabling better accuracy in downstream GeoAI applications. As the most representative type of geospatial point data, points-of-interests (POIs) are significant in understanding user mobility and urban functionality. This section will primarily discuss POIs and the corresponding SSL techniques designed for them. The surveyed studies for road networks are listed in Table I

A. Points-of-Interest

1) Intrinsic Attributes: POIs refer to the semantic locations in location-based services that users might be interested in visiting. 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 [5], [54], [55] often employ separate encoders to extract essential information from geo-location [27], categories, and other text features [38]. The encoded features are fused in a subsequent model, such as multi-layer perceptrons (MLP) or attention-based models, to obtain the POI representations [5], [54], [55]. 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 [56] 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.

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 [57], GNNs [58]–[60] and masked language modeling (MLM) [54], [55], [61], [62] have been explored. The skip-gram methods consider the spatial neighbors as the context of a POI, analogous to the surrounding words in Word2vec [63]. They minimize the errors in predicting the spatial context using the target POI embeddings. For instance, Place2vec [57] 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. [64] 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. In this regard, 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 [65] and [66] 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. To learn the POI representation in a self-supervised manner, predictive SSL methods have been exploited to predict either the intrinsic attributes of the POI or the relation between POIs. STPA [58] 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 [59] builds the competitive relations between two POIs if they are close in spatial 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 [60] 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 [54], [55], [61]. The common idea behind MLM-based methods is to construct pseudo sentences from the spatial neighborhoods and then apply MLMs on the pseudo sentences. GeoBERT [61] 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 [55] creates pseudo sentences by concatenating the names of neighboring POIs. For each POI, SpaBERT 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 [54] 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 [67]. 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.

Existing studies [68]–[72] employ skip-gram or contextual bag-of-word (CBOW), which are two typical predicative objectives, to the check-in sequences. Specifically, C-WARP [68] learns POI representations to predict other POIs within the same check-in sequences. As a result, the co-occurrences of POIs in the same check-in sequences are preserved in the POI representations. Building upon this, CAPE [70] learns POI representations that effectively predict both 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 [71] 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. Hier-CEM [72] extends the check-in sequence context by associating each check-in POI with the ancestor categories. POI2Vec [69] introduces a binary region tree to enhance the modeling of check-in sequences with spatial proximity.

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 [62] 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 [73]. 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, including DeepR [59] and SEENet [74], 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.

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 lunchtime. Motivated by this, SEENet [74] 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 [75] 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.

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 [58] and geo-location [54]. Besides, POI presentations can be directly used or used with a subsequent sequential model to predict the next POI visited by a user [68]–[70]. The POI representations can also be used to answer POI queries by matching POI representations and the query representation [62]. Furthermore, the presentations of multiple POIs can be used for more complex geospatial data mining tasks. For instance, POI relation prediction methods [59], [60], [74] 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 [76]. In addition, increasing studies have utilized POI embeddings in an aggregated manner for the inference of land use [64] or land use change [65] 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.

IV. Polylines

Polylines serve as an important spatial data type, widely employed across diverse geospatial data mining 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. We will discuss these two representative instances of polylines in a geospatial context, and present the specialized SSL techniques tailored to each data instance. The surveyed studies for road networks are listed in Table II

A. Road Networks

1) Intrinsic Attributes: Road networks are composed of connected road segments, where each segment is characterized

TABLE I
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 [57]	Predictive	Boosting the occurrence of popular neighbors	Spatial neighbors	Skipgram	Category prediction
Semantic [64]	Predictive	Random walk sampling	Spatial neighbors	Skipgram	Category prediction
MT-POI [65]	Predictive	Random walk sampling	Spatial neighbors	Skipgram	Category prediction
CatEM [66]	Predictive	None	Spatial neighbors, Check-in sequence	Skipgram	Category prediction
STPA [58]	Predictive	None	Spatial neighbors	Cross-Entropy	Category prediction
DeepR [59]	Predictive	None	Spatial neighbors Co-query context	Cross-Entropy	POI relation prediction
PRIM [60]	Predictive	None	Spatial neighbors	Cross-Entropy	POI relation prediction
GeoBERT [61]	Predictive	Neighbors ordered by distance Word masking	Spatial neighbors	Cross-Entropy	Masked word prediction
SpaBERT [55]	Predictive	Shortest paths in grid cells, POI sequences ordered by the distance to grid center	Spatial neighbors	Cross-Entropy	Masked POI prediction
MGeo [54]	Predictive, Contrastive	Neighbors ordered by distance POI attribute masking	Spatial neighbors	Cross-Entropy, KL divergence	POI attribute prediction, Distance contrast
C-WARP [68]	Predictive	None	Check-in sequence	Ranking loss, Skipgram	POI prediction
CAPE [70]	Predictive	None	Check-in sequence	Skipgram	POI ID and text prediction
DeepMove [71]	Predictive	Construction of origin POI pairs with the same destination	Check-in sequence	Skipgram	Origin POI prediction
Hier-CEM [72]	Predictive	Hierarchical extension of context categories	Check-in sequence	CBOW	Next category prediction
POI2Vec [69]	Predictive	Binary region tree construction	Check-in sequence	CBOW	POI prediction
ERNIE-GeoL [62]	Predictive	Heterogeneous graph, Random walk sampling, Word substitution and masking	Co-query context, Spatial neighbors	Cross-Entropy	Masked word prediction, Geocoding prediction
SEENet [74]	Contrastive	Grid cell augmentation	Co-query, Time, Spatial neighbors	JS divergence	Time-aware POI relation prediction

as a polyline. As a collection of polylines, road networks are usually concepturalized as a graph. 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 geo-locations, 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 [78], 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 [63]. The derived representations are demonstrated to be effective on several road classification tasks. Building upon this groundwork, IRN2Vec [79] 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 [80], [81] 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 [82] 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 [83]. 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.

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 three types of context information utilized in existing methods, where road segments are mainly regarded as graph nodes.

Spatial Features. Different from conventional graphs that 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 [81], [82]. Accordingly, spatial information has been effectively utilized as context knowledge in various

TABLE II
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]	predicative	Random walk sampling	None	Skipgram	Context neighbor prediction
IRN2Vec [79]	predicative	Random walk sampling	Road attributes	Cross-Entropy	Road attribute prediction
RFN [81]	predicative	None	Road attributes	Cross-Entropy	Graph reconstruction
RN2Vec [84]	predicative	Random walk sampling	Road attributes Spatial features	Skipgram	Road attribute prediction
HyperRoad [82]	predicative	Hypergraph construction	Road attributes	Cross-Entropy, Skipgram	Graph&Hypergraph reconstruction, Road attribute prediction, Hyperedge Classification
SARN [85]	contrastive	Edge perturbation	Spatial features, Road attributes	InfoNCE	Road-road, Road-region contrast
HNRN [86]	predicative	None	Trajectory	Cross-Entropy	Graph reconstruction
Toast [87]	predicative	Road segment masking, Random walk sampling	Trajectory	Cross-Entropy Skipgram	Context neighbor&attribute prediction, Masked road recovery, Trajectory discrimination
DyToast [88]	predicative	Road segment masking, Random walk sampling	Trajectory	Cross-Entropy Skipgram	Time-aware context neighbor &attribute prediction, Masked road recovery, Trajectory discrimination
TrajRNE [89]	predicative	Transition view construction	Trajectory	Cross-Entropy, MAE	Graph reconstruction, Road attribute prediction
JCLRNT [90]	contrastive	Transition view construction, Detour generation	Trajectory	JS divergence	Road-road, Road-trajectory, Trajectory-trajectory contrast
USPM [91]	predicative, contrastive	None	SVI	InfoNCE, Cross-Entropy	Road-image contrast, Road attribute prediction
Garner [92]	contrastive	Multi-view graph construction	SVI	JS divergence	Road-road contrast

methods. RN2Vec [84] 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 [85] 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 [93] 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 [86] 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 [87] is the initial effort to explicitly utilize detailed trajectories to enhance road network representation learning through predicative SSL. It equips skip-gram training with additional traffic and attribute prediction for context neighbors. Besides, it proposes two novel trajectory pretraining tasks within BERT framework [38]: route recovery, which recovers a sequence of masked road segments, and trajectory discrimination, which assesses whether a trajectory is authentic or simulated. These tasks enable the encoding of transition patterns and long-term dependencies intrinsic in road networks. DyToast [88] 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. JCLRNT [90] 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 [89] 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 [91] 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 [92] 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.

3) Applications: The representations of road networks, derived from either predicative 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 [89], lane number [82], and intersection tags [79], [84]. Besides, these representations are similarly utilized to infer the traffic status of road segments, including metrics like average speed [87], [90] and speed limits [81]. They also support the efficient vector computation for road network-based queries, such as calculating shortest path distances [85]. Furthermore, road segment representations serve as the effective start point for training in models that involve map-matched trajectories, for applications such as destination prediction [86] and anomalous sub-trajectory detection [94]. By enabling a diverse range of analytical tasks related to road networks, these representations offer substantial potential to improve understanding and decision-making within the domain of transportation infrastructure.

B. Trajectory

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 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 road networks are listed in Table III

In the category of predicative SSL methods, trajectories are typically trained to reconstruct their original sequence from a corrupted version of the input. For example, TC-DRL [95] 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 [96] based on RNN to reconstruct the local features for each window. t2vec [97] 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. Specifically, model adaptations include refinements via variatioanl inference [98] or self-attention mechanisms [99]. Geo-Tokenizer [100] reduces the number of grids to be trained by representing a location as a combination of multiple shared grids at several granular scales. It then utilizes the objective of predicting the grids for the next token at each scale in a hierarchical way. Furthermore, AdvTraj2vec [101] 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 (GAN) [102] to ensure the effects are neither too large or too small. E^2DTC incorporates self-training via soft cluster assignments [103] as an auxiliary loss into the reconstruction process. STPT [104] 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 [105] 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 [106] 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 [107] utilizes Convolutional-LSTM [108] to model historical traffic conditions reflected in holistic trajectories, thus enhancing the variational inference by providing additional hints into dynamic patterns for each trajectory.

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 [109] primarily uses point distortion to generate positive pairs of trajectories for contrastive learning. Expanding on this, TrajCL [110] 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 [111] 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.

2) Context Information: In addition to the spatial and temporal attributes intrinsic to trajectories, 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 seman-

TABLE III
A SUMMARY OF THE SURVEYED PAPERS ON SELF-SUPERVISED LEARNING FOR **TRAJECTORY**.

Method	Type	Data Augmentation	Context Information	Objective Function	Pretext Tasks
TCDRL [95]	predicative	Sliding window feature extraction	Grid partition	Cross-Entropy	Trajectory reconstruction
t2vec [97]	predicative	Point dropping, Spatial distortion	Grid partition	Spatial-aware Cross-Entropy	Trajectory reconstruction
GM-VSAE [98]	predicative	None	Grid partition	Cross-Entropy	Variational trajectory reconstruction
Geo-Tokenizer [100]	predicative	None	Multi-scale grid partition	Multi-scale Cross-Entropy	Autoregressive next grid prediction
AdvTraj2vec [101]	predicative	Point dropping, Embedding perturbation	Grid partition	Spatial-aware Cross-Entropy, Adversarial loss	Trajectory reconstruction, Adversarial learning
E^2DTC [103]	predicative	Point dropping, Spatial distortion	Grid partition	Spatial-aware Cross-Entropy, Triplet loss, KL divergence	Trajectory reconstruction, Triplet margin similarity, Self-clustering
STPT [104]	predicative	None	Grid partition	Cross-Entropy	Sub-trajectory discrimination
ST-t2vec [105]	predicative	Point dropping, Spatial&Temporal distortion	3D temporal grid partition	MSE	Trajectory reconstruction, Pairwise&Pattern similarity
RSTS [106]	predicative, contrastive	Point dropping, Spatial&Temporal distortion	3D temporal grid partition	Spatial-aware Cross-Entropy, Triplet loss	Trajectory reconstruction, Triplet margin similarity
DeepTEA [107]	predicative	None	Grid partition, Historical traffic	Cross-Entropy	Variational trajectory reconstruction
CL-Tsim [109]	contrastive	Point dropping, Spatial distortion	Grid partition	InfoNCE	Trajectory-trajectory contrast
TrajCL [110]	contrastive	Point dropping, Spatial distortion, Trajectory trimming&simplification	Grid partition, Point coordinates	InforNCE	Trajectory-trajectory contrast
KGTS [111]	contrastive	Grid deletion, Grid alternation	Grid partition	InfoNCE	Trajectory-trajectory contrast
CSSRNN [84]	predicative	Map matching	Road networks	Cross-Entropy	Autoregressive road prediction
Traj2Vec [112]	predicative	Map matching	Road networks	Cross-Entropy, MSE	Trajectory reconstruction, Travel time estimation
PT2Vec [113]	predicative	Map matching, Road network partition	Road networks	Cross-Entropy	Trajectory reconstruction
JGRM [114]	predicative	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 [115]	contrastive	Map matching	Road Networks	Triplet loss	Triplet margin similarity
PIM [116]	contrastive	Map matching	Road networks	JS divergence	Road-road, Road-trajectory contrast
MMTEC [117]	contrastive	Map matching, Continuous Trajectory	Road networks	Maximize entropy encoding	Trajectory-trajectory contrast
HMTRL [118]	predicative, contrastive	Map matching	Road networks	Cross-Entropy, MSE	Masked attribute prediction, Trajectory-trajectory contrast
START [119]	predicative, contrastive	Map matching, Dropout, Temporal distortion, Trajectory trimming, Road segment masking	Road networks	Cross-Entropy, InfoNCE	Masked road recovery, Trajectory-trajectory contrast
LightPath [120]	predicative, contrastive	Map matching, Road segment masking	Road networks	Cross-Entropy	Masked road recovery, Multi-view trajectory matching
CTLE [121]	predicative	Location masking, Time masking	Location semantics	Cross-Entropy	Masked location&time recovery
At2vec [122]	predicative	Location dropping	Location semantics	Spatial-temporal-activity-aware Cross-Entropy	Trajectory reconstruction
SML-LP [123]	contrastive	Location dropping, Location alternation	Location semantics	InfoNCE	Location embedding contrast
CACSR [124]	contrastive	Location embedding perturbation, Latent space perturbation	Location semantics	InfoNCE	Latent representation contrast

tics 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 [125], [126]. This process transforms the trajectory from a sequence of points to a sequence of road segments for subsequent modeling.

Several predicative SSL methods for map-matched trajectories utilize objectives similar to those used in trajectories with grid partitions, such as the reconstruction of road segments [112], [113] and next road prediction [127]. Besides, tailored objectives are devised to effectively incorporate the knowledge within road networks. For example, JGRM [114] 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 [128] 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 [115] combines GNN and graph embedding techniques [129] 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 [116] treats trajectories that share the same source and destination as positive pairs, and aims to maximize mutual information over these pairs. MMTEC [117] 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 different 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 [118] 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 [119] incorporates trajectory pattern into GNN aggregation and enhances Transformer with time-aware self-attention mechanism. Furthermore, it utilizes masked road segment recovery as the predicative 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 [120] 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.

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 [121] designs two predicative tasks, namely masked location recovery and masked hour prediction, with a sinusoidal temporal encoding technique incorporated into Transformer to derive representations. At2vec [122], [130] 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 [123] 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. Furthermore, CACSR [124] 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 [131], [132], the main target and its objective derives from supervised applications. Therefore, these specific studies are not further discussed in our context.

3) Applications: Similar to the representations derived for road networks, 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 [97], [133] and clustering [95], [103]. Moreover, they can be utilized to support diverse applications, such as transportation mode prediction [100], driver status inference [104], and anomalous trajectory detection [98]. In scenarios involving map-matched trajectories, these representations prove particularly valuable in applications focused on path analysis, such as travel time estimation [119], path ranking [120], and destination prediction [117]. In addition, semantic trajectory representations trained with SSL methods are typically fine-tuned to enhance performance in next location prediction [121], and can be uniquely applied in trajectory user-linking task [123], [124]. 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.

V. POLYGONS

A wide range of geospatial objects are often represented with polygonal geometries. This type provides a detailed depiction of the shapes of geospatial objects in 2D planes. In principle, from a 2D perspective, many other geospatial objects discussed in this paper, such as POIs and road segments, can also be represented as polygons. However, due to the substantial effort required to collect and curate objects with precise polygonal geometries, only a few geospatial data types are represented this way. Within the scope of this survey, a predominant object type that has garnered significant interest in an SSL context is the modeling of small urban areas, also known as urban regions.

Urban regions have been an ideal analytical scale for surveying and predicting various socio-economic and demographic characteristics in our cities [134]. These regions are essentially small polygonal areas partitioned by specific methods, such as fine-grained areas defined by road networks, fixed-size grids, or administrative units like Singapore Subzones [135] and NYC census Tracts [136]. Regardless of the partitioning method, modeling urban regions primarily relies on other source data modalities that reflect the intrinsic or contextual properties of these regions. Thus, a region can be viewed as a "container" of various human activities and data entities from different modalities. Some data modalities, such as POIs and street view images (SVIs), capture the intrinsic properties of the regions, while others, like human trajectories, describe region traits from a connectivity perspective, illustrating how people traverse through regions. In addition, spatial proximity is an important contextual factor, as spatially closer regions tend to be more similar.

With such diverse information perspectives used to model urban regions in an SSL context, most studies incorporate multiple perspectives. For example, a strand of region embedding studies [137]–[144] utilizes both POIs and human trajectories, sometimes supplemented by the consideration of spatial proximity. The modeling methods of different perspectives (data modalities) are often intertwined and challenging to disentangle. In this section, our goal is to present each study holistically, rather than fragmenting the analysis into fragmented components. This means that if a study models certain contextual information, the entire analytical pipeline of that study will be discussed collectively. The surveyed studies for urban regions are listed in Table IV.

A. Urban regions

1) Intrinsic Attributes: Among various data modalities, POIs and imagery data, including remote sensing (RS) images and 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

TABLE IV
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 [145]	Contrastive	POI	POI removal, addition, and shifting	None	Triplet loss	Region-region contrast
SKRL4RS [146]	Predictive, Contrastive	Knowledge graph	Spatial entity shifting	Knowledge graph	Triplet loss	Region-region contrast
HGI [147]	Contrastive	POI	POI and region graph	Spatial proximity, City overall context	MI maximization	POI-region contrast, Region-city contrast
HDGE [148]	Predictive	Human trajectory	Heterogeneous region graph	Human mobility, Spatial proximity, Temporal pattern	Skipgram, KL divergence	Reconstruction of human mobility patterns
ZE-Mob [149]	Predictive	Human trajectory	Human mobility event	Human mobility, Spatial proximity	Skipgram	Context neighbor prediction
GMEL [150]	Predictive	Human trajectory	Region graph	Human mobility, Spatial proximity	MSE	Commuting flow and in/out flow prediction
MGFN [151]	Predictive	Human trajectory	Region graph via clustering	Human mobility	KL divergence	Reconstruction of human mobility patterns
MP-VN [137]	Predictive	Human trajectory, POI	POI graph	Human mobility, Spatial proximity, POI distribution similarity	MSE	Graph reconstruction
DLCL [138]	Predictive	Human trajectory, POI	POI and region graph	Human mobility, Spatial proximity	Cross-Entropy	Region graph reconstruction
CGAL [139]	Predictive	Human trajectory, POI	POI and region graph	Human mobility, Spatial proximity	MSE, adversarial loss	Reconstruction of node feature and graph structure
MVURE [152]	Predictive	Human trajectory, POI, User check-in	Multi-view graph	Human mobility, POI distribution similarity, User check-in similarity	MSE, KL divergence	Reconstruction of human mobility patterns, POI, and check-in distributions
HREP [140]	Predictive, Contrastive	Human trajectory, POI	Heterogeneous region graph	Human mobility, Spatial proximity	KL divergence, Triplet loss, MSE	Region-region contrast, Reconstruction of human mobility patterns and POI distributions
HUGAT [153]	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	KL divergence, MSE	Reconstruction of human mobility patterns, geospatial adjacency and POI distributions
ReMVC [142]	Contrastive	Human trajectory, POI	POI insertion, deletion, and replacement Trajectory heatmap perturbation	Human mobility	InfoNCE	Region-region contrast
ReCP [143]	Contrastive, Predictive	Human trajectory, POI	POI insertion, deletion, and replacement Trajectory heatmap perturbation	Human mobility	InfoNCE, MI maximization, Conditional entropy	Region-region contrast
HAFusion [144]	Predictive	Human trajectory, POI, Land use	None	Human mobility	MSE, KL divergence	Reconstruction of human mobility patterns, POI, and land use feature similarity
Tile2Vec [154]	Contrastive	RS	None	Spatial proximity	Triplet	Region-region contrast
RegionEncoder [155]	Predictive	RS data, 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 [156]	Contrastive	SVI, POI	None	Spatial proximity	Triplet	Image- and region-level contrast
M3G [157]	Contrastive	Human trajectory, POI, SVI	None	Spatial proximity	Triplet	Region-SVI, region-POI, region-region contrast
Xi et al [158]	Contrastive	RS, POI	None	Spatial proximity	NT_Xent	Region-region contrast
MMGR [159]	Contrastive	RS, POI	RS image augmentation, POI graph	None	InfoNCE	Region-region contrast
UrbanCLIP [160]	Predictive, Contrastive	RS	Textual descrptions of images	None	InfoNCE, Cross-Entropy	Image-text alignment, Text autoregressive prediction
UrbanVLP [161]	Contrastive	RS, SVI	Textual descrptions of images	None	InfoNCE	Image- and token-level Image-text alignment
RegionDCL [162]	Contrastive	OSM building footprints, POI	Point injection, Building removal	Spatial proximity	InfoNCE, Triplet loss	Region-region contrast

these data sources to learn region representations in a self-supervised manner, sometimes with minimal consideration of contextual information.

In a pioneering study on POI-based region representation learning, [145] 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 [160] 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 [161] 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 [159] proposes to fuse RS images and POIs for learning intrinsic attributes of urban regions. It extends the idea in [13] by carrying out contrastive learning between multiple augmentations of each image, as well as between images and POIs using the method proposed in [64].

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 [145], SKRL4RS [146] 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 [147] extends beyond the POI category embedding technique in [64] 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 [148] 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 [149] 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 [63] by skip-gram objective. Besides, it enriches the objective by integrating the importance of region origin-destination pairs based on popularity and distance. GMEL [150] 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 [151] 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). A large portion of region representation learning studies incorporate POIs alongside human trajectories, as these two data sources naturally complement each other. POIs outline the range of human activities within each region, while human trajectories illustrate the interconnections and relatedness between regions. This integration is sometimes augmented with additional data sources, such as land use data, to further enhance the effectiveness of the learned region representations. In this regard, MP-VN [137] constructs two POI graphs to capture both static and mobility patterns among various POI types, based on their mobility connectivity patterns and average geographic distances. The two graphs are flattened as input to an autoencoder designed to learn region representations through graph reconstruction. Besides, the model incorporates the inter-region relations considering both spatial proximity and functional similarity derived from POIs to enhance the learning process. Built on this, DLCL [138] employs an adversarial autoencoder to learn representations of the two POI graphs in [137]. CGAL [139] further enhances the model by introducing a collective adversarial learning approach instead of autoencoder reconstruction. Specifically, it implements an assemble-disassemble strategy in which the assemble step fuses the two graph views into region representations, and the disassemble step then disaggregates the fused representations to reconstruct the original two graph views. It also incorporates two adversarial components to capture the similarities between regions based on POI distributions, textual information, and temporal patterns reflected from human trajectories. This model is trained to minimize the reconstruction loss of the adversarial framework. HAFusion [144] utilizes human trajectories, POIs, and land use data to model urban regions. Each region is characterized by a human mobility feature derived from outflow records, a POI category feature, and a land use feature quantifying the counts of land use zones in different categories. This method employs an attention-based encoder to capture region correlations within a single view and across different views. Furthermore, it applies another attention-based fusion module that integrates multi-view embeddings and captures higher-order correlations among regions.

Another line of research extensively emoplys GNNs. MVURE [152] utilizes human trajectories, POIs, and user check-ins to construct four different graph views. It then employs graph attention networks to encode each view and implements cross-view information sharing with attention

mechanisms as well as multi-view fusion using adaptively weighted combination of different views. The model is trained through objectives of reconstructing human mobility patterns and region relatedness reflected by POIs and check-ins. HREP [140] further considers several types of graph edges using human mobility data, POIs, and geographic context. This method constructs source and target edges based on human trajectories, POI edges derived from the similarities of POIs shared by different regions, and geographic context edges based on spatial adjacency. Using this heterogeneous graph structure, region representations are learned through attention-based GNN method by three objectives: a geographic proximity loss, a mobility reconstruction loss, and a POI correlation reconstruction loss. Additionally, HREP employs prompt learning (prefix-tuning) to adapt the learned representations for various downstream urban analytical tasks, a technique adapted from natural language processing. In addition, HUGAT [153] utilizes POI check-in trajectories and land use data for region representation learning. It defines five types of nodes, such regions and POI categories, and two types of edges representing spatial and temporal relations. This method then constructs a heterogeneous information network and designs five meta-paths to involve interesting relationships between regions, such as identifying regions that are popular destinations at the same time. A heterogeneous graph attention network [163] 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] utilizes inter-region human trajectories, spatial adjacency, and POI distributions as three similarity measures to construct a multi-graph for regions. Specifically, human mobility similarity is reflected by the co-occurrence frequency between regions, and POI distribution similarity is measured using the embeddings from a POI knowledge graph. This method utilizes GNN and fusion layers based on attention mechanisms to generate comprehensive region representations, with the training objectives of reconstructing human mobility patterns, preserving geospatial adjacency, and POI distribution similarity.

The third line of research employing POIs and human trajectories focuses on the application of contrastive learning paradigm. ReMVC [142] 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 representations of a region from different perspectives are similar, while the representations of different regions are distinguishable. ReCP [143] 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 [164], [165]. 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 [154] 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 [155] 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 three specific objectives: a reconstruction loss for the denoising autoencoder, a reconstruction loss for human mobility patterns, and a binary cross-entropy loss used by a discriminator to verify whether RS image representations and region graph node representations correspond to the same geographical region. The study in [158] adopts a contrastive learning strategy to maximize the similarity of representations derived from RS images that are geographically adjacent and those that have similar POI distributions. The representations learned from the two pathways, including spatial proximity and POI distribution similarity, of each region are adaptively fused using learnable weights in downstream tasks. Urban2Vec [156] derives the initial region representations by averaging all the SVI representations obtained from a CNN-based model 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 [157] extends Urban2Vec by simultaneously utilizing SVIs, POIs, and human trajectories. The techniques for handling SVIs and POIs are similar to those in Urban2Vec. Additionally, 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. The model is trained using a combination of triplet losses: region-SVI, region-POI, and region-region, which consider both spatial proximity and human mobility connections.

OpenStreetMap data. A new trend in region representation is to utilize data from OpenStreetMap (OSM) [67], 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 [162] 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.

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 [140] and adaptive multi-view fusion [158]. 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 [142], [143], [149], population density [147], [156], house prices [155], [160], average income [148], [157], check-in counts [137], [139], crime rates [148], [151], service call volumes [144], GDP [159], [160], nighttime lighting [161], takeaway order volumes [158], health indices [154], and so on. Moreover, region representations are extensively utilized for tasks like similar region search [145], [146] 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 [150], [153].

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. Such applications not only improve the accuracy of urban analyses but also support more informed decision-making in urban planning and policy-making. For example, accurate predictions of socioeconomic indicators can inform better resource allocation, enhance public services, and facilitate targeted interventions in sectors like healthcare, education, and transportation. Additionally, the capability to identify and cluster similar regions supports strategic site selection, the development of cohesive urban zones, optimization of land use, and promotion of sustainable development. 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.

VI. Discussion

Sections III-V have presented specialized SSL techniques designed for each geospatial data type. In this section, we extend the discussion by presenting several aspects not covered in the preceding sections. Specifically, we present some emerging trends in SSL techniques for geosaptial AI, including multi-type learning, and the adaptation to textual space via LLMs. Moreover, we provide a brief introduction of research that adopts task-specific SSL for geospatial applications.

A. Multi-type Learning

While SSL techniques have demonstrated promising performance in GeoAI, 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 learning of representations for multiple geospatial data types, as a step towards the development of geospatial foundation models. For example, several studies [166]–[168] approach this goal by utilizing a variety of entities such as POIs, regions, and their associated content like street view images, remote sensing images, and user data, to construct an urban knowledge graph. The edges between these entities are established by considering factors such as topological relations, human mobility patterns, semantic relations, and similarities in POI distributions across regions. Then knowledge graph embedding techniques are applied to derive the representations for multiple geospatial data types.

Besides, other studies formulate this problem through a heterogeneous learning framework using contrastive SSL. HOME-GCL [169] 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 [170] 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.

B. Adaptation to Language Space

The integration of LLMs, which are fundamentally rooted in the SSL paradigm, represents a burgeoning trend across various domains, including GeoAI. LLMs have demonstrated not only enhanced language understanding, reasoning, and generation capabilities, but also remarkable adaptability to various other data modalities, including images [52], [171], graphs [172], [173], and time series [174], [175]. Furthermore, LLMs have been validated to possess a certain level of knowledge in geospatial domain [176]–[178]. Based on these capabilities, several studies have started to explore the application of LLMs to geospatial data types within language space for GeoAI. These efforts typically employ alignment strategies to facilitate language-based interactions with geospatial data through targeted SSL fine-tuning of LLMs.

GeoLLM [179] 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 socioeconomic indicators. LAMP [180] 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, this method 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. UrbanGPT [181] 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.

C. Task-specific SSL techniques

Although the primary focus of this survey is on deriving generalizable representations of various geospatial data types for broad application across multiple downstream tasks, there are numerous studies that employ more task-specific SSL techniques for pre-training or as supplementary objectives to enhance the main task. These specialized approaches tailor SSL techniques to meet the unique demands and intricacies of specific geospatial applications.

A prominent example of task-specific SSL is in spatio-temporal forecasting tasks. For example, models like UniST [182], W-MAE [183], and GPT-ST [184] are utilized for grid regions 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 [185] applies contrastive learning to identify grid regions in both spatial and temporal dimension with similar patterns. STGCL [186] 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 [187] 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 [188] incorporates a masked auto-encoder module to reconstruct masked spatio-temporal patches, thus deriving more robust representations to support the forecasting task. CL4ST [189] 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 [190]–[192]. These diverse methods demonstrate the versatility and potential of task-specific SSL techniques in enhancing the performance for various downstream applications regarding geospatial data.

VII. 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 in GeoAI usually draw inspiration from the domains of CV 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 combining different pretext tasks and data augmentation techniques in other domains [193]–[195], these results 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 rigorously investigate the impact of different pretext task and data augmentation selections specifically within the geospatial domain. This investigation should ideally be supported by theoretical analysis and supplemented by comprehensive empirical evaluations to ensure robustness. Moreover, it is beneficial to construct standard benchmarks for geospatial

SSL. This would not only facilitate the comparison of different SSL approaches but also provide valuable guidance for the development of future research efforts in various downstream applications.

Multi-modality data 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. While there are several efforts to integrate multi-modality data to enhance the performance of geospatial applications [196]–[198], only few have explored the ideas of SSL [91], [92]. 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. Based on this, 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 [52] 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 VI, LLMs and multi-type pre-trained models opens up exciting possibilities for adaptation to the geospatial 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 [199], [200]. 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.

Privacy and vulnerability in geospatial SSL. The inherent nature of geospatial data, which often contains sensitive information about individuals and urban infrastructure, raises significant privacy and data vulnerability concerns. For instance, aggregated trajectories, when analyzed with certain approaches, can expose individual's routes and private locations [201]. To this end, privacy-preserving techniques, such as differential privacy, could be considered within the SSL framework. Moreover, federated learning approaches [202], [203] 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 [204], [205].

Dynamic model learning and updating. Most existing SSL techniques for geospatial data employ the setting of learning static representations. However, as urban environments continuously evolve, the characteristics of geospatial objects may experience distribution shifts. Therefore, it is desirable to develop more flexible SSL techniques that possess the ability of dynamically updating in response to new environmental conditions. One promising direction is the integration of continual learning mechanisms into SSL techniques [206]. This would allow models to incorporate new information and adapt to changes without the need for complete retraining, while also preventing catastrophic forgetting of previously learned knowledge. Furthermore, investigating the active learning principles with SSL techniques could lead to more efficient updating strategies [207]. By identifying and utilizing the most informative new samples for model adaptation, these methods could significantly reduce the computational and data demands for model updating.

VIII. Conclusion

This paper provides a comprehensive overview of self-supervised learning 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. We further present studies on the emerging trends for GeoAI 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 in GeoAI.

References

[1] Z. Chen, L. Chen, G. Cong, and C. S. Jensen, "Location- and keyword-based querying of geo-textual data: a survey," *VLDB J.*, vol. 30, no. 4, pp. 603–640, 2021.
[2] G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 337–348, 2009.
[3] Y. Liu, T. N. Pham, G. Cong, and Q. Yuan, "An experimental evaluation of point-of-interest recommendation in location-based social networks," *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1010–1021, 2017.
[4] Y. Chen, C. Long, G. Cong, and 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, and Z. Hai, "Mining geospatial relationships from text," *Proc. ACM Manag. Data*, vol. 1, no. 1, pp. 93:1–93:26, 2023.
[6] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, and A. K. Qin, "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1544–1561, 2022.
[7] H. Miao, Y. Zhao, C. Guo, B. Yang, Z. Kai, F. Huang, J. Xie, and 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, and J. E. Blumenstock, "Microestimates of wealth for all low-and middle-income countries," *Proceedings of the National Academy of Sciences*, vol. 119, no. 3, 2022.

[9] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature communications*, vol. 11, no. 1, p. 2583, 2020.

[10] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri, "Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 625–636, 2020.

[11] "Eu gdpr regulation," 2016. [Online]. Available: https://gdpr-info.eu

[12] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2023.

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

[14] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, 2021.

[15] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 288:1–288:37, 2023.

[16] 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*, vol. abs/2303.18223, 2023.

[17] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *CoRR*, vol. abs/2402.06196, 2024.

[18] Y. Liu, M. Jin, S. Pan, C. Zhou, Y. Zheng, F. Xia, and P. S. Yu, "Graph self-supervised learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5879–5900, 2023.

[19] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4216–4235, 2023.

[20] 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.

[21] 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.

[22] A.-X. Zhu and M. Turner, "How is the third law of geography different?" *Annals of GIS*, vol. 28, no. 1, pp. 57–67, 2022.

[23] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 335–355, 2024. [Online]. Available: https://doi.org/10.1109/TKDE.2023.3282907

[24] 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*, vol. abs/2403.14151, 2024.

[25] S. Wang, Z. Bao, J. S. Culpepper, and G. Cong, "A survey on trajectory data management, analytics, and learning," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 39:1–39:36, 2022.

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

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

[28] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3681–3700, 2022.

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

[30] S. Gao, Y. Hu, and W. Li, "Handbook of geospatial artificial intelligence," 2023.

[31] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[32] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *ICLR*, 2019.

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

[34] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

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

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

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

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

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

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

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

[43] 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.

[44] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.

[45] 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*, vol. abs/2403.05530, 2024.

[46] 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*, vol. abs/2307.09288, 2023.

[47] A. Anthropic, "The claude 3 model family: Opus, sonnet, haiku," *Claude-3 Model Card*, vol. 1, 2024.

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

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

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

[51] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 9992–10 002.

[52] 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.

[53] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, vol. 162, 2022, pp. 12 888–12 900.

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

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

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

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

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

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

[60] Y. Chen, X. Li, G. Cong, C. Long, Z. Bao, S. Liu, W. Gu, and F. Zhang, "Points-of-interest relationship inference with spatial-enriched graph neural networks," *Proc. VLDB Endow.*, vol. 15, no. 3, pp. 504–512, 2021.

[61] Y. Gao, Y. Xiong, S. Wang, and H. Wang, "Geobert: Pre-training geospatial representation learning on point-of-interest," *Applied Sciences*, vol. 12, no. 24, 2022.

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

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

[64] W. Huang, L. Cui, M. Chen, D. Zhang, and Y. Yao, "Estimating urban functional distributions with semantics preserved poi embedding," *International Journal of Geographical Information Science*, vol. 36, no. 10, pp. 1905–1930, 2022.

[65] Y. Yao, Q. Zhu, Z. Guo, W. Huang, Y. Zhang, X. Yan, A. Dong, Z. Jiang, H. Liu, and Q. Guan, "Unsupervised land-use change detection using multi-temporal poi embedding," *International Journal of Geographical Information Science*, vol. 37, no. 11, pp. 2392–2415, 2023.

[66] J. Bing, M. Chen, M. Yang, W. Huang, Y. Gong, and L. Nie, "Pre-trained semantic embeddings for poi categories based on multiple contexts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8893–8904, 2023.

[67] OpenStreetMap, https://www.openstreetmap.org, 2017.

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

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

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

[71] Y. Zhou and 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.

[72] M. Chen, L. Zhu, R. Xu, Y. Liu, X. Yu, and Y. Yin, "Embedding hierarchical structures for venue category representation," *ACM Trans. Inf. Syst.*, vol. 40, no. 3, 2021.

[73] K. Sahr, D. White, and A. J. Kimerling, "Geodesic discrete global grid systems," *Cartography and Geographic Information Science*, vol. 30, no. 2, pp. 121–134, 2003.

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

[75] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax." *ICLR*, vol. 2, no. 3, p. 4, 2019.

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

[77] T. S. Jepsen, C. S. Jensen, T. D. Nielsen, and 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] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*, 2016, pp. 855–864.

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

[80] T. S. Jepsen, C. S. Jensen, and T. D. Nielsen, "Graph convolutional networks for road networks," in *SIGSPATIAL*, 2019, pp. 460–463.

[81] T. Jepsen, C. S. Jensen, and T. D. Nielsen, "Relational fusion networks: Graph convolutional networks for road networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 418–429, 2020.

[82] L. Zhang and C. Long, "Road network representation learning: A dual graph-based approach," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 9, pp. 1–25, 2023.

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

[84] M.-X. Wang, W.-C. Lee, T.-Y. Fu, and G. Yu, "On representation learning for road networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 1, pp. 1–27, 2020.

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

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

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

[88] Y. Chen, X. Li, G. Cong, Z. Bao, and C. Long, "Semantic-enhanced representation learning for road networks with temporal dynamics," *CoRR*, vol. abs/2403.11495, 2024.

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

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

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

[92] H. Zhou, W. Huang, Y. Chen, T. He, G. Cong, and Y.-S. Ong, "Road network representation learning with the third law of geography," *CoRR*, vol. abs/2406.04038, 2024.

[93] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *NeurIPS*, 2020.

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

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

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

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

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

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

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

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

[102] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

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

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

[105] D. A. Tedjopurnomo, X. Li, Z. Bao, G. Cong, F. M. Choudhury, and A. K. Qin, "Similar trajectory search with spatio-temporal deep representation learning," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, pp. 77:1–77:26, 2021.

[106] Z. Chen, K. Li, S. Zhou, L. Chen, and S. Shang, "Towards robust trajectory similarity computation: Representation-based spatio-temporal similarity quantification," *World Wide Web Journal*, vol. 26, no. 4, pp. 1271–1294, 2023.

[107] X. Han, R. Cheng, C. Ma, and T. Grubenmann, "Deeptea: Effective and efficient online time-dependent trajectory outlier detection," *Proc. VLDB Endow.*, vol. 15, no. 7, pp. 1493–1505, 2022.

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

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

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

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

[112] T. Fu and W. Lee, "Trembr: Exploring road networks for trajectory representation learning," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 1, pp. 10:1–10:25, 2020.

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

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

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

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

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

[118] H. Liu, J. Han, Y. Fu, Y. Li, K. Chen, and H. Xiong, "Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training," *VLDB J.*, vol. 32, no. 2, pp. 325–342, 2023.

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

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

[121] Y. Lin, H. Wan, S. Guo, and 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.

[122] A. Liu, Y. Zhang, X. Zhang, G. Liu, Y. Zhang, Z. Li, L. Zhao, Q. Li, and X. Zhou, "Representation learning with multi-level attention for activity trajectory similarity computation," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 5, pp. 2387–2400, 2022.

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

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

[125] C. Yang and G. Gidofalvi, "Fast map matching, an algorithm integrating hidden markov model with precomputation," *International Journal of Geographical Information Science*, vol. 32, no. 3, pp. 547–570, 2018.

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

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

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

[129] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *AAAI*, 2014, pp. 1112–1119.

[130] Y. Zhang, A. Liu, G. Liu, Z. Li, and Q. Li, "Deep representation learning of activity trajectory similarity computation," in *ICWS*, 2019, pp. 312–319.

[131] C. Duan, W. Fan, W. Zhou, H. Liu, and J. Wen, "Clsprec: Contrastive learning of long and short-term preferences for next POI recommendation," in *CIKM*, 2023, pp. 473–482.

[132] Z. Jia, Y. Fan, J. Zhang, C. Wei, R. Yan, and X. Wu, "Improving next location recommendation services with spatial-temporal multi-group contrastive learning," *IEEE Transactions on Services Computing*, vol. 16, no. 5, pp. 3467–3478, 2023.

[133] Y. Chang, E. Tanin, G. Cong, C. S. Jensen, and J. Qi, "Trajectory similarity measurement: An efficiency perspective," *Proc. VLDB Endow.*, vol. 17, 2024.

[134] A. D. Singleton and P. Longley, "The internal structure of greater london: a comparison of national and regional geodemographic models," *Geo: Geography and Environment*, vol. 2, no. 1, pp. 69–87, 2015.

[135] "Singapore subzones." [Online]. Available: https://data.gov.sg/dataset/master-plan-2019-subzone-boundary-no-sea

[136] "Nyc census tracts." [Online]. Available: https://www.nyc.gov/site/planning/data-maps/open-data/census-download-metadata.page

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

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

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

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

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

[142] L. Zhang, C. Long, and G. Cong, "Region embedding with intra and inter-view contrastive learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9031–9036, 2022.

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

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

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

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

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

[148] H. Wang and Z. Li, "Region representation learning via mobility flow," in *CIKM*, 2017, pp. 237–246.

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

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

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

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

[153] N. Kim and Y. Yoon, "Effective urban region representation learning using heterogeneous urban graph attention network," *arXiv preprint arXiv:2202.09021*, 2022.

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

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

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

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

[158] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma, and 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.

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

[160] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and 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.

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

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

[163] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *WWW*, 2019, pp. 2022–2032.

[164] J. E. Patino and J. C. Duque, "A review of regional science applications of satellite remote sensing in urban settings," *Computers, Environment and Urban Systems*, vol. 37, pp. 1–17, 2013.

[165] Z. Fan, F. Zhang, B. P. Loo, and C. Ratti, "Urban visual intelligence: Uncovering hidden city profiles with street view images," *Proceedings of the National Academy of Sciences*, vol. 120, no. 27, 2023.

[166] A. Dsouza, N. Tempelmeier, R. Yu, S. Gottschalk, and E. Demidova, "Worldkg: A world-scale geographic knowledge graph," in *CIKM*, 2021, pp. 4475–4484.

[167] Y. Liu, J. Ding, Y. Fu, and Y. Li, "Urbankg: An urban knowledge graph system," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 4, pp. 60:1–60:25, 2023.

[168] Y. Ning, H. Liu, H. Wang, Z. Zeng, and H. Xiong, "UUKG: unified urban knowledge graph dataset for urban spatiotemporal prediction," in *NeurIPS*, 2023.

[169] J. Jiang, Y. Yang, J. Wang, and J. Wu, "Jointly learning representations for map entities via heterogeneous graph contrastive learning," *CoRR*, vol. abs/2402.06135, 2024.

[170] P. Balsebre, W. Huang, G. Cong, and Y. Li, "City foundation models for learning general purpose representations from openstreetmap," in *CIKM*, 2024.

[171] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, 2022.

[172] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J. Wen, "Structgpt: A general framework for large language model to reason over structured data," in *EMNLP*, 2023, pp. 9237–9251.

[173] H. Wang, S. Feng, T. He, Z. Tan, X. Han, and Y. Tsvetkov, "Can language models solve graph problems in natural language?" in *NeurIPS*, 2023.

[174] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P. Chen, Y. Liang, Y. Li, S. Pan, and Q. Wen, "Time-llm: Time series forecasting by reprogramming large language models," in *ICLR*, 2024.

[175] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One fits all: Power general time series analysis by pretrained LM," in *NeurIPS*, 2023.

[176] 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, and N. Lao, "On the opportunities and challenges of foundation models for geospatial artificial intelligence," *CoRR*, vol. abs/2304.06798, 2023.

[177] W. Gurnee and M. Tegmark, "Language models represent space and time," in *ICLR*, 2024.

[178] J. Roberts, T. Lüddecke, S. Das, K. Han, and S. Albanie, "GPT4GEO: how a language model sees the world's geography," *CoRR*, vol. abs/2306.00020, 2023.

[179] R. Manvi, S. Khanna, G. Mai, M. Burke, D. B. Lobell, and S. Ermon, "Geollm: Extracting geospatial knowledge from large language models," in *ICLR*, 2024.

[180] P. Balsebre, W. Huang, and G. Cong, "LAMP: A language model on the map," *CoRR*, vol. abs/2403.09059, 2024.

[181] Z. Li, L. Xia, J. Tang, Y. Xu, L. Shi, L. Xia, D. Yin, and C. Huang, "Urbangpt: Spatio-temporal large language models," *CoRR*, vol. abs/2403.00813, 2024.

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

[183] X. Man, C. Zhang, C. Li, and J. Shao, "W-MAE: pre-trained weather model with masked autoencoder for multi-variable weather forecasting," *CoRR*, vol. abs/2304.08754, 2023.

[184] Z. Li, L. Xia, Y. Xu, and C. Huang, "GPT-ST: generative pre-training of spatio-temporal graph neural networks," in *NeurIPS*, 2023.

[185] H. Qu, Y. Gong, M. Chen, J. Zhang, Y. Zheng, and Y. Yin, "Forecasting fine-grained urban flows via spatio-temporal contrastive self-supervision," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8008–8023, 2023.

[186] X. Liu, Y. Liang, C. Huang, Y. Zheng, B. Hooi, and R. Zimmermann, "When do contrastive learning signals help spatio-temporal graph forecasting?" in *SIGSPATIAL*, 2022, pp. 5:1–5:12.

[187] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," in *AAAI*, 2023, pp. 4356–4364.

[188] S. Guo, Y. Lin, L. Gong, C. Wang, Z. Zhou, Z. Shen, Y. Huang, and H. Wan, "Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting," in *ICDE*, 2023, pp. 1585–1596.

[189] J. Tang, L. Xia, J. Hu, and C. Huang, "Spatio-temporal meta contrastive learning," in *CIKM*, 2023, pp. 2412–2421.

[190] Q. Gao, J. Hong, X. Xu, P. Kuang, F. Zhou, and G. Trajcevski, "Predicting human mobility via self-supervised disentanglement learning," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 5, pp. 2126–2141, 2024.

[191] Q. Gao, W. Wang, K. Zhang, X. Yang, C. Miao, and T. Li, "Self-supervised representation learning for trip recommendation," *Knowl. Based Syst.*, vol. 247, p. 108791, 2022.

[192] S. Jiang, W. He, L. Cui, Y. Xu, and L. Liu, "Modeling long- and short-term user preferences via self-supervised learning for next POI recommendation," *ACM Trans. Knowl. Discov. Data*, vol. 17, no. 9, pp. 125:1–125:20, 2023.

[193] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljacic, "Equivariant contrastive learning," *CoRR*, vol. abs/2111.00899, 2021.

[194] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *NeurIPS*, 2020.

[195] T. Xiao, X. Wang, A. A. Efros, and T. Darrell, "What should not be contrastive in contrastive learning," in *ICLR*, 2021.

[196] W. Huang, J. Wang, and G. Cong, "Zero-shot urban function inference with street view images through prompting a pretrained vision-language model," *Int. J. Geogr. Inf. Sci.*, vol. 38, no. 7, pp. 1414–1442, 2024.

[197] V. V. Cepeda, G. K. Nayak, and M. Shah, "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization," in *NeurIPS*, 2023.

[198] S. Xu, C. Zhang, L. Fan, G. Meng, S. Xiang, and J. Ye, "Addressclip: Empowering vision-language models for city-wide image address localization," *arXiv preprint arXiv:2407.08156*, 2024.

[199] 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, and E. Ettinger, "Self-supervised learning for large-scale item recommendations," in *CIKM*, 2021, pp. 4321–4330.

[200] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, "Self-supervised graph transformer on large-scale molecular data," in *NeurIPS*, 2020.

[201] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data," in *WWW*, 2017, pp. 1241–1250.

[202] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, "PMF: A privacy-preserving human mobility prediction framework via federated learning," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 10:1–10:21, 2020.

[203] Z. Liu, H. Miao, Y. Zhao, C. Liu, K. Zheng, and H. Li, "Lighttr: A lightweight framework for federated trajectory recovery," in *ICDE*, 2024.

[204] Y. Lun, H. Miao, J. Shen, R. Wang, X. Wang, and S. Wang, "Resisting tul attack: balancing data privacy and utility on trajectory via collaborative adversarial learning," *GeoInformatica*, pp. 1–21, 2023.

[205] F. Xu, Y. Li, Z. Tu, S. Chang, and H. Huang, "No more than what I post: Preventing linkage attacks on check-in services," *IEEE Trans. Mob. Comput.*, vol. 20, no. 2, pp. 620–633, 2021.

[206] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, 2024.

[207] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 180:1–180:40, 2022.