Annals of GIS

ISSN: 1947-5683 (Print) 1947-5691 (Online) Journal homepage: www.tandfonline.com/journals/tagi20

Representation learning for geospatial data

Yu Liu, Xuechen Wang, Yidan Wang, Fei Huang, Yingjing Huang, Yong Li, Weiyu Zhang, Shuhui Gong, Gengchen Mai, Yao Yao, Yang Yue, Haifeng Li & Fan Zhang

To cite this article: Yu Liu, Xuechen Wang, Yidan Wang, Fei Huang, Yingjing Huang, Yong Li, Weiyu Zhang, Shuhui Gong, Gengchen Mai, Yao Yao, Yang Yue, Haifeng Li & Fan Zhang (2025) Representation learning for geospatial data, Annals of GIS, 31:4, 557-583, DOI: 10.1080/19475683.2025.2552157

To link to this article: https://doi.org/10.1080/19475683.2025.2552157

Published online: 17 Sep 2025.

ANNALS OF GIS
2025, VOL. 31, NO. 4, 557–583
https://doi.org/10.1080/19475683.2025.2552157

Taylor & Francis
Taylor & Francis Group

Representation learning for geospatial data

Yu Liu[a,b,c], Xuechen Wang[a], Yidan Wang[d], Fei Huang[e], Yingjing Huang[a], Yong Li[d], Weiyu Zhang[a], Shuhui Gong[f], Gengchen Mai[g], Yao Yao[h,i], Yang Yue[j], Haifeng Li[k] and Fan Zhang[a]

[a]Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China; [b]Ordos Research Institute of Energy, Peking University, Ordos, China; [c]Southwest United Graduate School, Kunming, China; [d]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; [e]Department of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China; [f]School of Artificial Intelligence, China University of Geosciences, Beijing, China; [g]SEAI Lab, Department of Geography and the Environment, the University of Texas at Austin, Austin, TX, USA; [h]UrbanComp Lab, School of Geography and Information Engineering, China University of Geosciences, Wuhan, China; [i]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, China; [j]Thrust of Urban Governance and Design, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; [k]School of Geosciences and Info-Physics, Central South University, Changsha, China

ABSTRACT

This paper reviews representation learning for geospatial data, focusing on methods for automatically extracting meaningful features from diverse data types. By simplifying tasks and improving accuracy, representation learning has emerged as a powerful tool for geospatial analysis. Due to its generalizability and scalability, representation learning provides an effective approach to processing geospatial data, which is inherently diverse and unstructured. We summarize the representation learning methods for different geospatial data types, including locations, points of interest (POIs), trajectories, spatial interactions, remote sensing imagery, and street view imagery. Treating each data type as a distinct modality, we emphasize the potential of multi-modal representation learning to advance the understanding of geographical phenomena and propose an LLM-guided framework as a potential solution. The review concludes by highlighting the need for further research to improve multi-modal data alignment and enhance the interpretability of feature representations, particularly in complex and dynamic geographical environments.

ARTICLE HISTORY
Received 18 February 2025
Accepted 19 August 2025

KEYWORDS
Representation learning; geospatial data; multi-modal representation learning

1. Introduction

Representation learning is the process of automatically uncovering meaningful and effective feature representations from raw data. This approach addresses a fundamental challenge in machine learning: raw data, such as images and text, often contain complex and entangled features that are difficult for models to process directly (Bengio, Courville, and Vincent 2013). By automatically extracting and transforming raw data into useful representations, it simplifies the learning task for predictive models, enhancing their performance and accuracy. Compared with traditional feature engineering that requires manual feature design, representation learning offers two key advantages – generalizability and scalability – making it a foundational pillar of contemporary artificial intelligence systems. This field has been transformed by two major breakthroughs: (1) the Transformer architecture, which employs self-attention mechanisms to capture contextual dependencies and enhance feature representation (Vaswani et al. 2017), and (2) scalable self-supervised learning (SSL) techniques, which extract transferable knowledge from large-scale unlabelled datasets (T. Chen et al. 2020; Devlin et al. 2019; He et al. 2020). These innovations have laid the groundwork for foundational models (Bommasani et al. 2021) – large-scale, pre-trained models capable of learning universal representations for a wide range of tasks. Notable examples include GPT-4, which demonstrates advanced language understanding (OpenAI et al. 2024); CLIP, which aligns visual and linguistic modalities (Radford

CONTACT Yu Liu liuyu@urban.pku.edu.cn Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China

et al. 2021); and AlphaFold2, which achieves groundbreaking performance in protein structure prediction (Jumper et al. 2021).

Geospatial data are unique for several reasons. First, they encompass diverse data types that are often unstructured. Traditional manual feature extraction methods struggle to capture complex features and implicit patterns (Du et al. 2019). Second, they are influenced by various spatial effects (Y. Liu et al. 2024), including: (1) spatial dependence (Tobler 1970), the principle that nearby things are more related; (2) spatial heterogeneity (Goodchild 2004), which reflects the non-uniform distribution of geographic forms and processes; (3) distance decay (Fotheringham 1981), the reduction in interaction strength as distance increases; and (4) scale effects (Openshaw 1984), where spatial patterns vary across analytical scales. Lastly, they cover both physical and human phenomena, which follow different laws but are coupled tightly (Goodchild and Li 2021). Representation learning offers a promising approach by enabling the automatic discovery of hidden geographic patterns and spatial effects within observed data, which are crucial for constructing meaningful and high-quality representations of geographical units.

Given that locational information is essential to geospatial data, the features to be learned are in general associated with particular geographical units (or places). Therefore, the inputs for geospatial representation learning consist of heterogeneous raw data related to geographical units. Such raw data describe physical geographical attributes (e.g. terrain and land cover), human behaviours (e.g. movement trajectories), and built environmental elements (e.g. buildings and streets). Each data type can be viewed as a modality and corresponds to certain representation learning methods. In this manner, geospatial representation learning forms a foundation for geospatial artificial intelligence (GeoAI). In this review, we systematically summarize representation learning methods for heterogeneous geospatial data, categorizing them along two key dimensions: (1) spatial (geometric locations) and (2) semantic (including POIs, trajectories, spatial interactions/flows, and remote sensing/street view imagery), with further differentiation based on their static/dynamic nature and physical/human-oriented attributes (Figure 1). These data types capture complementary spatiotemporal phenomena and are fundamental to geographic analysis in GIS. Developing specialized representation learning methods for each data type while modelling their interrelationships is therefore critical. We argue that multi-modal representation learning, by integrating diverse modalities into a unified embedding space, has the potential to enhance geographic understanding and improve downstream applications. When combined with large language models (LLMs) for cross-modal alignment and fusion, it offers a promising framework for analysing geographical units characterized by multiple data sources.

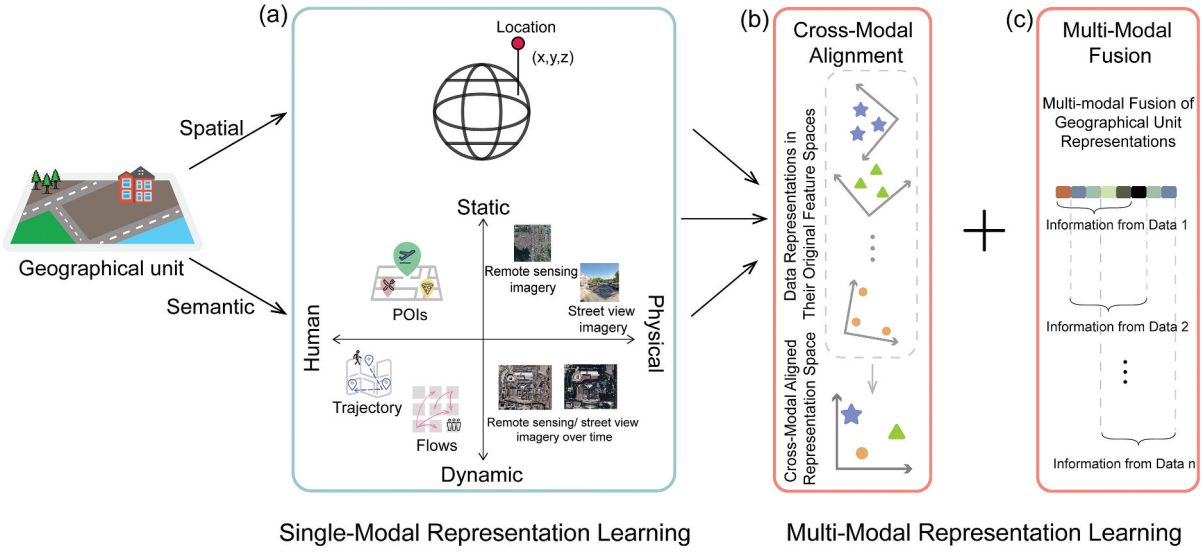

Figure 1. Framework for multi-modal geospatial representation learning. (a) Input modalities include spatial data (geometric locations), human-oriented data (POIs, trajectories, flows), and physical observations (remote/street-view imagery), processed separately for each modality. (b) Cross-modal alignment transforms heterogeneous data sources (data 1,. . .,n) from their native feature spaces into a unified representation space. (c) Multi-modal fusion integrates complementary spatiotemporal phenomena, enabling information fusion from diverse data sources.

The contributions of this survey are as follows. First, we systematically review the representation learning for typical data in geospatial data, including locations, POIs, trajectories, spatial interaction networks, remotely sensed imagery, and street view imagery (Sections 2.1-2.6). Second, we discuss the general principles for multi-modal representation learning in the LLM era in Section 3.1. Lastly, to our knowledge, this is the first survey to discuss multi-modal representation learning for geospatial data with the language-centred interactive paradigm, which we call multi-modal interactive representation (MMIR) (see Section 3.2). This survey aims to bridge two key audiences: geospatial researchers entering machine learning and machine learning experts exploring the geospatial domain. For the former, it demonstrates how representation learning addresses geospatial-specific challenges while highlighting practical applications. For the latter, it clarifies unique opportunities for innovation through spatially aware architectures and evaluation metrics. Readers will gain an overview of the methods, a roadmap for multi-modal learning with LLMs, and insights into open challenges for future research.

2. Single-modal representation learning

In the traditional paradigm, representation learning for different types of geospatial data typically involves learning representations for each geospatial modality separately and then fusing them into a unified representation. In Sections 2.1–2.6, we review representation learning methods for individual geospatial modalities (Figure 1 (a) and Table 1).

2.1. Representation learning for locations

Location representation learning focuses on encoding a location within a manifold space into a high-dimensional vector (so-called location embedding) or decoding a high-dimensional vector back into a location in the manifold space. We usually refer to the former models as location encoders (Cole et al. 2023; Klemmer et al. 2025; Mac Aodha, Cole, and Perona 2019; Mai et al. 2022; Mai, Janowicz, Yan, et al. 2020; Mai, Lao, et al. 2023) and the latter as location decoders (Dufour et al. 2025; S. Luo and Hu 2021; Mai et al. 2024; Z. Wang et al. 2025). The manifold space may be a 2D Euclidean space (e.g. a 2D space defined by a projection coordinate system), a 3D Euclidean space, a spherical surface (e.g. the Earth's surface), or any other space defined by a manifold.

Until now, different location representation learning models have been developed for different manifold spaces. However, several properties are expected to be satisfied regardless of the nature of manifold spaces: 1) Distance preservation: The distance between two location embeddings in the representation space should be proportional to their corresponding locations' distance in the original manifold space (Mai et al. 2022; Mai, Xuan, et al. 2023); 2) Bijection: A one-to-one mapping must exist between locations in the original manifold space and their corresponding representations in the embedding space; 3) Learning-friendly space: The location embedding space should be more learning-friendly to downstream machine learning models (e.g. neural networks); 4) Direction awareness: Locations oriented in similar directions should have more similar embeddings than those facing markedly different directions (Mai et al. 2022); 5) Inductive encoder: Location encoders should employ inductive learning through neural network models. Unlike transductive approaches that memorize fixed embeddings for specific locations (as in word embeddings (Mikolov, Sutskever, et al. 2013) or knowledge graph embeddings (Bordes et al. 2013; Cai et al. 2019; Schlichtkrull et al. 2018), these models should learn a mapping function capable of representing arbitrary locations in the embedding space, which enables generalization to unseen locations (Cole et al. 2023). Among these properties, distance preservation is the most important, as it ensures that neural architectures incorporating distance-preserved location representations can capture spatial dependence across geographic entities. Similarly, direction-aware location representations are critical when the underlying geographic distribution exhibits anisotropic patterns (R. Zhu, Janowicz, and Mai 2019). Note that many location representation learning models can satisfy some but not all properties listed above.

Location encoders aim to represent a location within a manifold space as a high-dimensional vector. According to the nature of the manifold space, we can classify the current location encoders into two categories: 2D location encoders and 3D location encoders. 2D location encoders operate in 2D projected spaces, such as projected coordinate systems. Examples are Wrap (Cole et al. 2023; Mac Aodha, Cole, and

Table 1. Summary of data, model architecture, learning paradigm and major applications for single-modal representation learning.

Data	Model architecture	Learning paradigm	Major applications
Location	• Encoder: $e = NN(PE(x); \theta)$ • $PE()$: a deterministic function that transforms the location into a high-dimensional vector • $NN(;\theta)$: a learnable neural network which is usually implemented as a fully connected layer or a multilayer perceptron (MLP) with θ denoting the learnable parameters • Decoder: no common neural architecture	Supervised/Self-supervised	• Species fine-grained recognition (Mac Aodha, Cole, and Perona 2019 Mai, Janowicz, Yan, et al. 2020; Mai, Lao, et al. 2023) • Satellite image classification (Klemmer et al. 2025; Rußwurm et al. 2024) • Geographic question answering (Mai, Janowicz, Cai, et al. 2020) • Environmental variable prediction, sustainability index prediction (N. Wu et al. 2024) • Image geolocalization (Vivanco Cepeda, Nayak, and Shah 2023; Z. Wang et al. 2025)
POIs	• POI sequence: Word2Vec, LSTM, MLM • POI graph: GNNs	Supervised/ Unsupervised	• Personalized recommendation systems (Lai et al. 2024; J. Zhang and Ma 2024) • Urban functional distribution identification (J. Fan and Thakur 2023; K. Liu et al. 2020) • Socioeconomic analysis (Bai et al. 2023; Chen, Zhao, et al. 2022; F. Huang, Lv, and Yue 2024)
Trajectory	• Sequences: RNNs/LSTM, Transformer • Graphs: GNNs • Textual descriptions: Word2Vec, LLMs	Supervised/ Unsupervised	• Trajectory similarity computation (Y. Chang et al. 2023; X. Li et al. 2018; D. Yao et al. 2022) • Trajectory prediction (Alahi et al. 2016; Y. Yao, Guo, et al. 2023; Y. Zhang et al. 2024) • Pattern mining (Y. Chen et al. 2021; Haydari et al. 2024; P. Wang et al. 2019)
Spatial interaction network (flow matrix)	• Spatiotemporal context: Word2Vec, GNN • Flow allocation modelling: MLP	Supervised/ Unsupervised	• Land use classification (N. Kim and Yoon 2025; Z. Yao et al. 2018; Y. Zhou and Huang 2018) • Socioeconomic status prediction (N. Kim and Yoon 2025; Y. Luo, Chung, and Chen 2022; H. Wang and Li 2017; X. Wang, Chen, and Liu 2024; M. Zhang et al. 2020) • Flow prediction (N. Kim and Yoon 2025)
Remote sensing imagery	• CNN, Transformer	Supervised/Self-supervised	• Land cover classification (Hong et al. 2019) • Change detection (P. Chen et al. 2022) • Scene recognition (C. Luo, Jin, and Sun 2019) • Socio-demographic variable estimation (Neal et al. 2022; Rolf et al. 2021)
Street view imagery	• CNN, Transformer	Supervised/Self-supervised	• Building style and age identification (Sun et al. 2022) • Road quality assessment (Chacra and Zelek 2018) • Street store-type classification (Noorian, Psyllidis, and Bozzon 2019) • Socioeconomic indicator prediction (Li et al. 2025)

Perona 2019), RBF (Mai, Janowicz, Yan, et al. 2020), Space2Vec (Mai, Janowicz, Yan, et al. 2020), SpaBERT (Z. Li et al. 2022, Z. Li et al. 2023), and GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023). In contrast, 3D location encoders handle locations defined in 3D space. Several location encoders like XYZ (Mai, Lao, et al. 2023) and NeRF (Mildenhall et al. 2022) operate directly in unrestricted 3D Euclidean space, while others such as Sphere2Vec (Mai, Xuan, et al. 2023) and Spherical Harmonics (Rußwurm et al. 2024) are specifically designed for spherical surfaces embedded within 3D space. These location encoders share a common neural architecture:

$$\boldsymbol{e} = NN(PE(\boldsymbol{x}); \theta),$$

where x is the input location, which can be a 2D projection coordinate, or geographic coordinates, or coordinates in a 3D space used by XYZ and NERF. $PE()$ is a deterministic function that transforms the location into a high-dimensional vector. $NN(;\theta)$ is a learnable neural network which is usually implemented as a fully connected layer or a multilayer perceptron with θ denoting the learnable parameters. Note that there is no location encoder as a global winner for all tasks, i.e. different location encoders are suitable for different

geospatial tasks. Please refer to Mai et al. (2022) for a comprehensive survey about location encoders and TorchSpatial (N. Wu et al. 2024) as a Python library of their implementations.

Location decoders can be considered as the reverse operation of location encoders which receive much less attention from the GeoAI community. Although in some pioneering works, a simple multi-layer perceptron is used to regress one pair of coordinates as the initial implementation of location decoders (Rao et al. 2020) in city-scale tasks, this regression-based approach has proven ineffective for global-scale applications like image geolocalization (Seo et al. 2018; Vo, Jacobs, and Hays 2017; M. Wu and Huang 2022; Z. Zhou et al. 2024). Instead, one common way is to change the location prediction task into a location cell classification task – dividing the Earth into non-hierarchical/hierarchical grid cells and predicting cell IDs instead of raw coordinates (Izbicki, Papalexakis, and Tsotras 2020; Seo et al. 2018; Vo, Jacobs, and Hays 2017; M. Wu and Huang 2022). An alternative way is a retrieval-based approach. Researchers established a gallery of locations (or image-location pairs) to find the location (or image-location pair) in the gallery that matches the input query image most and return the corresponding locations (Tian, Chen, and Shah 2017; Vivanco Cepeda, Nayak, and Shah 2023; H. Yang, Lu, and Zhu 2021; Z. Zhou et al. 2024; S. Zhu, Yang, and Chen 2021). Several recent works also use diffusion models to decode locations either for predicting the next location of a trajectory (Y. Zhu et al. 2023) or for image geolocalization (Dufour et al. 2025; Z. Wang et al. 2025). Currently, there is no common neural architecture shared by different research, making it one of the most promising research directions in spatial representation learning.

Location representation learning models can be trained in both supervised and self-supervised manners. Many pioneering works in this domain train location encoders in a supervised learning manner for specific downstream tasks. Recently, one promising research direction in location representation learning is conducting self-supervised learning (SSL) between location and other data modalities. For instance, CSP employs a CLIP-like self-supervised learning objective to contrast location embeddings with image embeddings using geo-tagged data (including species occurrence records and satellite imagery). This approach has demonstrated effectiveness across multiple downstream tasks, including fine-grained species recognition and satellite image classification (Mai, Lao, et al. 2023). SatCLIP uses a similar contrastive learning objective between location embedding and satellite image embedding and shows promising performances on multiple geo-aware image classification and regression tasks such as air temperature prediction, elevation prediction, socio-economic factor prediction, species image classification, etc (Klemmer et al. 2025). Similarly, GeoCLIP employs this location-image contrastive learning framework for geolocalization tasks (Vivanco Cepeda, Nayak, and Shah 2023). GAIR, a recent geo-foundation model, expands this idea by conducting contrastive learning across three data modalities – locations, remote sensing images, and street view images – to achieve state-of-the-art performance on 10+ downstream tasks (Z. Liu et al. 2025). How to leverage location representation learning to form a spatially explicit SSL objective for geo-foundation model pre-training has consequently emerged as a crucial direction for future research directions.

Since locations are the fundamental georeference for all geospatial data, location representation learning has been widely used in various downstream tasks, including species fine-grained recognition (Mac Aodha, Cole, and Perona 2019; Mai, Janowicz, Yan, et al. 2020; Mai, Lao, et al. 2023), satellite image classification (Klemmer et al. 2025; Rußwurm et al. 2024), geographic question answering (Mai, Janowicz, Cai, et al. 2020), environmental variable prediction, sustainability index prediction (N. Wu et al. 2024), image geolocalization (Vivanco Cepeda, Nayak, and Shah 2023; Z. Wang et al. 2025), health outcome prediction (J. Zhang et al. 2025), trajectory imputation (Yang, Yao, Whalen and Mai, 2025), traffic violation prediction (Yang, Yao, Roozkhosh, Liu and Mai, 2025), among others. Despite these successes, several challenges remain. First, while the distance preservation property enables the learned location representations to effectively capture spatial dependencies among geographic features, how to use these representations to capture spatial heterogeneity remains an ongoing research direction. Second, although advanced location encoders like Space2Vec and Sphere2Vec already use inductive multi-scale representations to capture spatial patterns across different scales, designing effective inductive multi-scale location decoders remains a challenging and open research problem. Lastly, while many current location representation learning methods achieve strong in-domain performance, they often generalize poorly to new geographic areas. Developing models with better spatiotemporal generalizability thus remains a key challenge for future GeoAI research.

2.2. Representation learning for POIs

POIs serve as geographic space mappings of urban functions and human activities, reflecting socioeconomic vitality and complex human behaviour patterns. POI representation learning involves quantifying and abstracting the inherent deep semantic information to transform POIs into high-dimensional vectors (Figure 2). The vectors effectively capture complex characteristics, such as functional distribution, spatio-temporal relationships, and social interaction patterns among POIs (Bing et al. 2023; X. Liu, Andris, and Rahimi 2019). As such, POI representation supports a wide range of tasks, including urban functional area identification (Bing et al. 2023; X. Liu, Andris, and Rahimi 2019), land use change detection (Y. Yao, Zhu et al. 2023), and personalized recommendations (Lai et al. 2024; J. Zhang and Ma 2024).

To incorporate POI data into deep learning models and capture their spatial characteristics, existing studies commonly employ sequence-based and graph-based structures to model relationships among POIs. Sequence-based methods draw inspiration from natural language processing (NLP) by treating POIs as 'words' and constructing meaningful POI sequences. These sequences encode geographic proximity or user behaviour patterns, allowing models to implicitly learn spatial and temporal dependencies. Graph-based methods represent POIs as nodes in a graph, with edge weights typically defined by geographic distance and interaction features. These structured approaches provide richer contextual information and spatial constraints, enhancing the modelling of spatial distribution patterns and semantic features.

The sequence-based method proposed by Y. Yao et al. (2017) was an early example of POI representation (Niu and Silva 2021). They generated POI sequences by considering the shortest path and then inputted these sequences into the Word2Vec model to produce category representation vectors. As shown in Figure 2, the sequence-based approach focuses on two core elements: (1) how to construct effective sequences and (2) how to extract POI representations using sequence models. Strategies for sequence construction typically consider the shortest path (Y. Yao et al. 2017), geographic proximity (B. Yan et al. 2017; Zhai et al. 2019), and users' historical visits to POI locations (Cao et al. 2020) to generate sequence data with spatiotemporal semantics. The sequence data reflects the direct relationships between POIs and introduces information on users' spatial behaviour and temporal changes. To extract meaningful representations from these sequences, models such as Word2Vec (B. Chang et al. 2018; T. Li et al. 2025), LSTM (F. Yu et al. 2020), and the masked language model (MLM) (J. Huang et al. 2022) are employed to uncover spatiotemporal dependencies and implicit semantic information.

Although sequence-based POI representation methods are effective, they face inherent limitations in capturing complex spatial structures and interactions among POIs. To address these shortcomings, graph-based POI representation methods have emerged as a more robust solution. Graph-based POI representation learning treats POIs as nodes in a graph and uses features such as distances and interactions between POIs as edge weights (Gong et al. 2024; W. Huang et al. 2022) to construct the graph structure. Graph Neural Networks (GNNs) are then applied to capture spatial dependencies and relationships between POIs in two-dimensional geographic space. GNNs also aggregate neighbourhood information to enhance the ability to express both local and global features of POI representations (Xu et al. 2022). A recent study (W. Huang et al. 2023) shows that graph-based POI representation methods not only consider the category information of POIs but also integrate geographic neighbourhood information to achieve single POI representation learning.

POI representation learning has demonstrated significant application potential across various fields. In personalized recommendation systems, POI representation leverages users' historical behaviours and the spatial relationships of POIs (E. Wang et al. 2023; Zheng and Zhou 2024) to provide accurate recommendation services. In urban functional distribution identification, POI representation analyses the spatial distribution of POIs (Qin et al. 2022) and scene categories (Chen, Zhu, et al. 2022), enabling the effective identification of urban functional zones and potential development trends. When combined with multimodal data, such as remote sensing images (Bai et al. 2023) and human movement (Chen, Zhao, et al. 2022; F. Huang, Lv and Yue 2024), POI representation provides richer information support for socioeconomic analysis. Additionally, POI representation plays a crucial role in large geographic models. POI representation provides spatial context information, including spatial location, function type, and the relationships between POIs and their surrounding environment (P. Li et al. 2024), thereby supporting model training. The rich semantic information provided by POI representation enhances the capabilities of geographic models in spatial analysis and intelligent geographic reasoning.

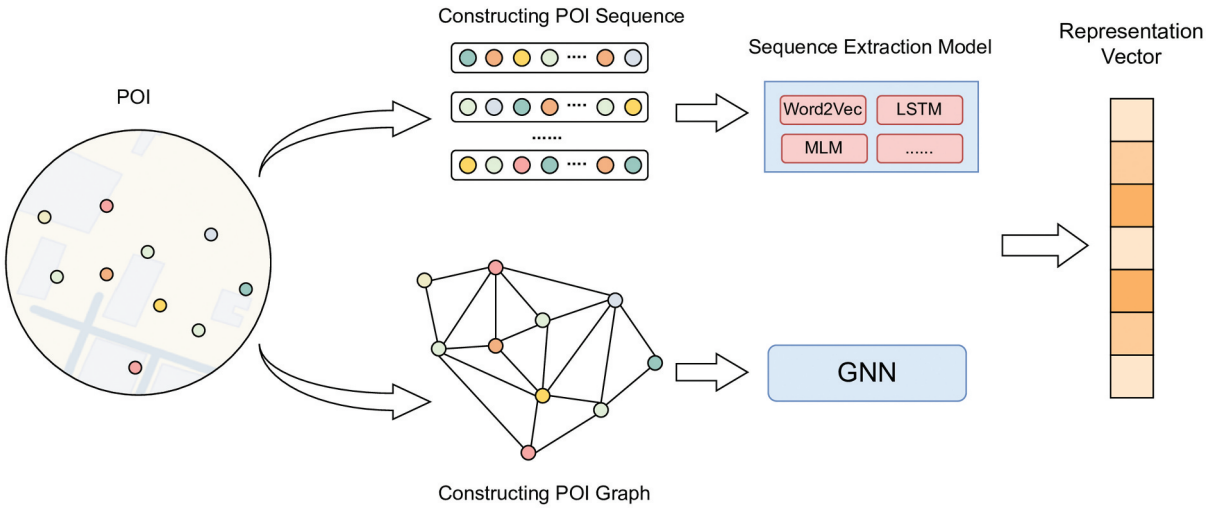

Figure 2. Methods for learning representations from POIs. The upper part uses a sequence-based approach, where the sequence is constructed and fed into a feature extraction module. The lower part uses a graph-based approach, where a graph is constructed and input into a GNN. Both methods output representation vectors.

Despite the critical role of POI representation, current evaluation frameworks lack standardized criteria. Most mainstream approaches rely on dimensionality reduction visualizations for qualitative assessments or indirectly evaluate representation quality through customized downstream tasks. Such practices limit comparative analysis and compromise the objectivity and generalizability of evaluations. To advance the research and improve comparability, it is crucial to establish standardized and quantifiable evaluation systems for POI representations.

2.3. Representation learning for trajectories

Trajectory data extend beyond static spatial coordinates by incorporating temporal dynamics and spatial transitions, reflecting complex interactions between moving objects and their surrounding environments (Hägerstrand 1970). This information, encompassing 'where', 'when', and 'why', emphasizes the goal of trajectory representation learning: encoding spatiotemporal relationships and contextual semantics into high-dimensional vector embeddings (F. Huang, Lv, and Yue 2024; Rao, Gao, and Zhu 2023). Trajectories present heterogeneity, correlation, and irregularity across space and time. Mainstream approaches tackle these challenges by imposing distinct structural assumptions on the data, thereby conceptualizing trajectories primarily as sequences, graphs, or textual descriptions. Each of these frameworks imposes unique assumptions about the organization of features and relationships, ultimately influencing the learned representations.

Viewing trajectories as temporally ordered location sequences, i.e. $T = \langle (l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n) \rangle$, prioritizes the temporal progression where the location and state at any point are influenced by preceding ones. This perspective models the temporal correlation of movement, framing it as a time-series problem. Early studies often employed manual feature extraction (e.g. time intervals, speed, visit frequency) followed by sequence models like seq2seq (Damiani et al. 2020; D. Yao et al. 2017), but this heavily relied on predefined features. Modern methods directly process ordered timestamp-location pairs using models such as RNNs or Transformers, automatically capturing non-linear dependencies (see Figure 3(a)). RNNs naturally model the Markovian assumptions (and their extensions) inherent in movement patterns through their recurrent states, making them effective at capturing properties like travel time and distance (Alahi et al. 2016; X. Li et al. 2018; H. Zhang et al. 2020). This 'process-first' viewpoint excels at modelling inertia in spatial sequences $T^{(l)}$ (X. Jiang et al. 2017) but struggles with capturing long-range dependencies, especially in sparse data. Transformers address these limitations through self-attention mechanisms that explicitly model non-adjacent relationships. For instance, TrajFormer (Liang et al. 2022) enhances position encoding by considering continuous spatiotemporal intervals between tokens to capture the irregularity of trajectories. Similarly,

TrajBERT (Si et al. 2024) employs Transformer encoder to learn mobility patterns bi-directionally with temporal refinement. The primary training objective of Transformer-based frameworks is focused on self-supervised learning tasks, such as next-token prediction or masked token reconstruction, to uncover the underlying structural (Y. Chang et al. 2023; Fang et al. 2022; J. Jiang et al. 2023). As a result, effective tokenization of spatial and temporal dimensions is essential. This involves encoding spatial structures (e.g. road networks) into spatial tokens $T^{(l)}$ and discretizing continuous and periodic temporal dimensions into temporal tokens $T^{(t)}$.

Trajectories within relational structures (graphs) shift focus to the spatial context and connectivity. In graph-based methods (see Figure 3(b)), trajectories are constructed as, or projected onto, graphs $\mathcal{G}$ that are then fed into models like GNNs. These models address spatial heterogeneity and correlation, encoding relationships between disparate locations (e.g. POIs and road intersections) and leveraging network topology to define spatial proximity. The specific graph construction determines which spatiotemporal contextual elements are explicitly represented: (1) Trajectory Matching on Road Networks: Matching trajectories onto road networks emphasizes the modelling of hierarchical spatial structure, constraining movement and informing path choices (Y. Chen et al. 2021; Fu and Lee 2020; S. Zhou et al. 2023). In this approach, GNNs learn representations sensitive to network topology and reachability. (2) Individual Trajectory Graphs: Constructing individual trajectory graphs captures personal mobility preferences in space (P. Wang et al. 2019; D. Yao et al. 2022). Here, nodes represent attributes of locations or regions (e.g. POI categories or place functions), while edges encode learned transition weights, such as travel frequency or distance between locations. (3) Dynamic Spatial Graphs: Dynamic Spatial Trajectory Graphs account for the temporal evolution of spatial relationships (e.g. traffic flows, land-use changes). Frameworks like (Dai et al. 2021) and (W. Yu and Wang 2023) model this through time-varying graph structures $P(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n | t_1, t_2, \ldots, t_n)$, explicitly capturing spatiotemporal interactions. Furthermore, combining GNNs and Transformers is an intuitive strategy: GNNs can learn spatial embeddings (e.g. from road networks) associated with trajectory points, which then serve as input for temporal modelling by Transformers (J. Jiang et al. 2023; Zhang, Yu, and Zhu 2024). However, it is worth noting that spatiotemporal dependencies might remain implicitly encoded, particularly in conditional modelling approaches (F. Huang, Lv, and Yue 2024).

Human mobilities signify activities, goals, and interactions with meaningful places. From this perspective, trajectories are treated more as narratives of activity rich with semantic content that links to POI types, activity labels, or social comments. Early NLP-inspired methods treated trajectories as sequences of 'place words' (e.g. POI categories), using techniques like Word2Vec to learn semantic embeddings (Murray et al. 2023; Yao, Guo et al. 2023; F. Zhou et al. 2019). The advent of LLMs, with their vast world knowledge and sophisticated language understanding, has significantly advanced this view. As shown in Figure 3(c), by translating trajectory data into textual descriptions, LLMs can be employed to generate deep semantic representations, capturing nuanced activity understanding, contextual reasoning, and even inferring intent (Haydari et al. 2024; Li et al. 2024; Zhang , Amiri, Liu, Zhao, and Zuefle 2024). However, scaling diverse spatiotemporal information into a purely semantic space presents challenges. The key lies in accurately

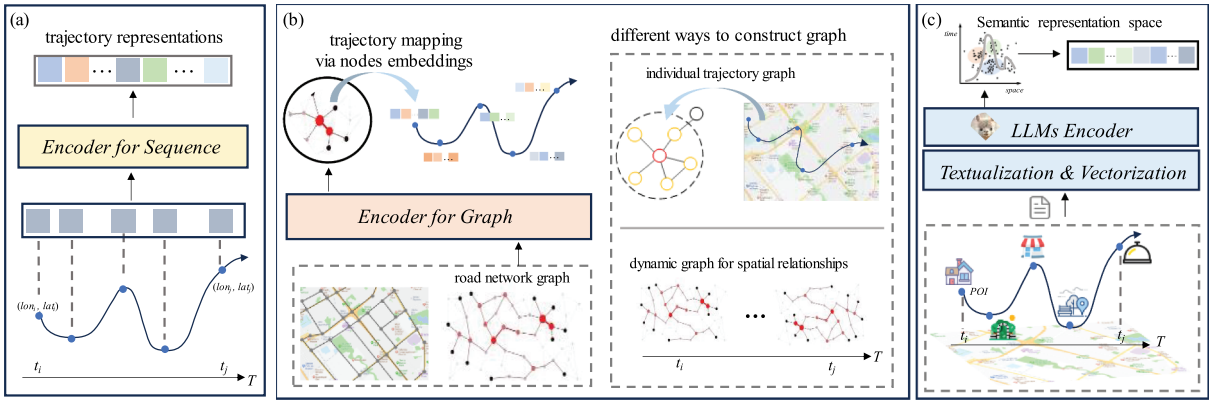

Figure 3. Trajectory representation learning across three data structures: sequences, graphs, and semantic text.

transforming continuous spatiotemporal data into discrete textual tokens compatible with LLMs, while addressing inherent model biases and minimizing potential 'semantic gaps' (Vafa et al. 2024).

In summary, although advanced deep learning models have significantly propelled trajectory representation learning, a fundamental challenge remains: achieving a deeper understanding of the intrinsic spatiotemporal structure and semantic meaning of trajectories, as this foundation ultimately determines the quality of representations. From a data epistemology perspective (Kitchin 2013), revisiting foundational questions is pertinent: how do human activities unfold across time and space, and how do trajectory representations help us understand this implicit process (Couclelis 1986; S. Gao 2024; Tuan 1979)? The path forward thus hinges on unified frameworks that not only align spatial, temporal, and semantic information but are, more importantly, grounded in the definition of space-time units (tokens) and their linkages that fully respect the data's intrinsic irregularity, correlation, and heterogeneity. Such integration is key to developing more holistic representations and achieving deeper insights into complex processes like human-environment interactions.

2.4. Representation learning for spatial interaction networks

Spatial interactions represent the supply–demand relationships between places, driven by people's decision-making processes and manifested through flows of people, goods, ideas, and more (Anderson 2011; Fotheringham and O'Kelly 1989; Simini et al. 2012; J. Wang 2017; Wilson 1967). The characteristics of a place determine its ability to generate (nodal propulsiveness) or attract (nodal attractiveness) bi-directional flows, while spatial impedances – such as distance and travel costs – acts as a barrier to flows between places (Anderson 2011; Fotheringham and O'Kelly 1989; J. Wang 2017; Wilson 1967). These flows form weighted networks embedded in geographic space, with places as nodes and flow volumes between them as edge weights (Barthélemy 2011; Batty 2013; Louail et al. 2015). Spatial interaction networks are essential for understanding places, as they capture inter-place relationships and connect distant locations (Batty 2013; Y. Liu et al. 2024). Many GeoAI tasks, such as socioeconomic prediction (Z. Fan et al. 2023; Rolf et al. 2021) and flow generation (Simini et al. 2021), rely on a comprehensive characterization of places (S. Gao 2024), with spatial interaction data offering a promising path forward. However, traditional feature engineering faces significant challenges, particularly due to the distance-decay effect (M. Zhang et al. 2020; Fotheringham 1981), where interactions weaken as the distance between places increases. Accurately quantifying the effective distance (e.g. spatial impedance parameters) and modelling appropriate decay functions remain complex and unresolved tasks.

Representation learning for spatial interaction networks offers distinct advantages by automatically addressing the complexities of the distance-decay effect and generating multi-scale, multi-faceted embeddings that capture diverse flow semantics aligned with varying place characteristics. For instance, commuting flows can encode fine-grained income variations (Kreindler and Miyauchi 2023), while trade flows reflect macroeconomic indicators such as regional GDP (Helpman, Melitz, and Rubinstein 2008). Typically, representation learning for spatial interaction networks employs static flow matrices (e.g. origin-destination data) as input observations, with the core objective of inferring latent driving factors. These factors may be derived explicitly by modelling the underlying interaction mechanisms or implicitly through data-driven techniques such as network embedding methods. Given a spatial interaction network among N places, representation learning generates two distinct embeddings for each place: one as an origin and the other as a destination (Figure 4). The origin embedding captures information related to nodal propulsiveness, while the destination embedding encapsulates nodal attractiveness. For example, in trade flow networks, the origin embedding may reflect exporter-specific factors such as firm productivity, while the destination embedding incorporates importer-related features like market size (e.g. GDP, population) or consumer preferences (Helpman, Melitz, and Rubinstein 2008).

Existing approaches to learning representations from spatial interaction networks can be categorized into two paradigms based on their incorporation of domain knowledge: process-agnostic methods that rely exclusively on data-driven patterns, and process-explicit methods that incorporate domain knowledge. These approaches generally operate under two fundamental assumptions: (1) places exhibiting similar interaction patterns should possess similar representations, and (2) place representations can be derived by modelling flow allocation probability distributions. The first assumption is inspired by word and network

representation learning, where words or nodes with similar contexts are assumed to have similar representations (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013; Tang et al. 2015). By constructing spatiotemporal contexts from spatial interaction networks, place representations can be inferred similarly. Research has demonstrated the usefulness of representations learned from these methods, such as using representations learned from taxi flows to perform various downstream tasks like land use classification and crime rate predictions (N. Kim and Yoon 2025; Y. Luo, Chung, and Chen 2022; H. Wang and Li 2017; Z. Yao et al. 2018; M. Zhang et al. 2020; Y. Zhou and Huang 2018). The second assumption builds upon single-constrained spatial interaction models (Fotheringham 1983; Fotheringham and O'Kelly 1989), where outflows from a given origin are allocated to destinations based on destination attractiveness and impedance, while inflows to a destination are determined by origin propulsiveness and impedance. The representation learning framework inverts this logic: it derives (1) destination attractiveness and spatial impedance from observed outflow allocation probability distributions and (2) origin propulsiveness and spatial impedance from inflow allocation probability distributions. This approach has been validated through experiments on synthetic flow data and has demonstrated empirical effectiveness in predicting income levels, housing prices, and inter-place distances across multiple scales in commuting datasets (X. Wang, Chen, and Liu 2024).

Integrating the objective functions derived from these two assumptions enables effective incorporation of flow semantics into geospatial multi-modal representation learning. However, existing methods still exhibit limitations. As (X. Wang, Chen, and Liu 2024) have demonstrated, applying generic network representation learning techniques to spatial interaction networks is equivalent to optimizing only outflow allocation objectives, thereby neglecting origin propulsiveness. Furthermore, while the second assumption is grounded in domain knowledge, current implementations oversimplify spatial interaction mechanisms by relying solely on inner product operations. Empirical evidence suggests that modelling interaction dynamics through non-linear modules could yield significant performance improvements (Simini et al. 2021) – a direction meriting deeper exploration. Moreover, current research has decomposed the impedance variable into place representations, complicating the interpretation of the distance decay effect. Last but not least, the inherent multi-scale nature of spatial interactions, where different interaction types typically manifest at distinct geographic scales, presents an additional fundamental challenge for robust representation learning.

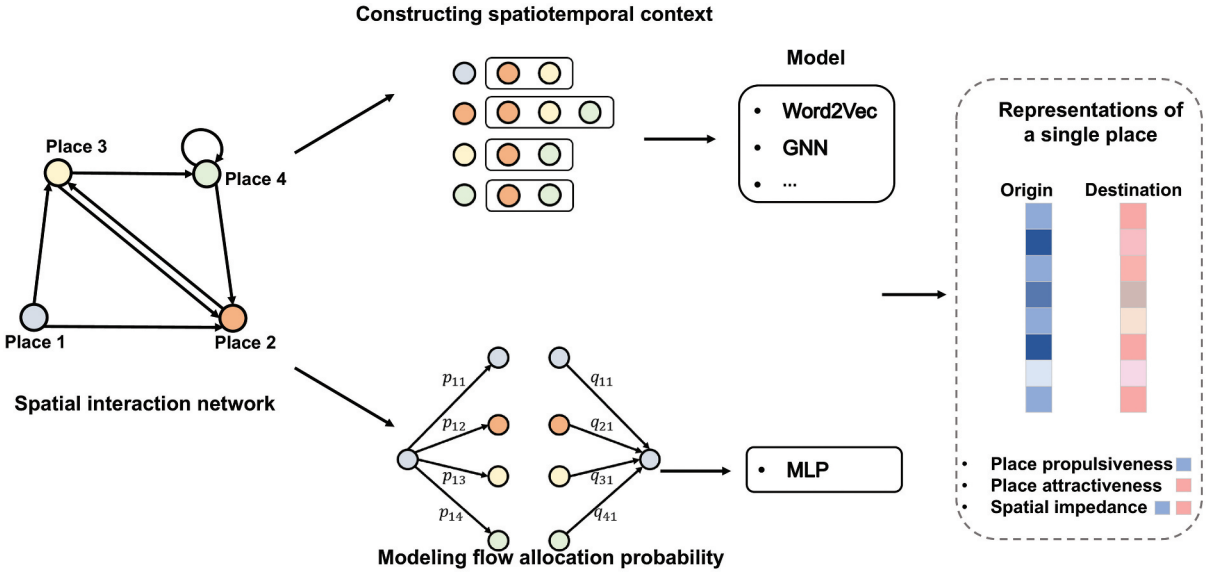

Figure 4. Representation learning for spatial interaction networks. Input: spatial interaction networks represented by static flow matrices (e.g. origin-destination data). Objective: infer latent driving factors through either (1) implicit pattern extraction via data-driven approaches (e.g. word or network embedding techniques) or (2) explicit modelling of interaction mechanisms. Output: two representations for each place (origin and destination embeddings) capturing latent driving factors.

2.5. Representation learning for remote sensing imagery

Remote sensing imagery is characterized by high spatial resolution, high temporal resolution, and spatio-temporal continuity, with diverse and heterogeneous ground objects. While remote sensing data provide rich temporal, spatial, and spectral information, such information is typically challenging to utilize directly without proper processing (Xue et al. 2025). Representation learning enables the automatic extraction of features from high-dimensional and complex remote sensing data. The features extracted from representation learning could not only encompass the information captured by traditional handcrafted features but also reveal underlying coordination patterns and deep semantic structures, thereby enhancing the performance of downstream tasks (Tao et al. 2023).

Supervised feature learning (SFL) is a classic approach in remote sensing representation learning, relying on large-scale annotated datasets. Its performance on downstream tasks heavily depends on the quantity and quality of the labels. To address this need, several high-quality annotated datasets have been developed, such as DOTA for object detection (Xia et al. 2018), RS5M for matching remote sensing imagery with text (Z. Zhang et al. 2024), and SatlasPretrain for various tasks (Bastani et al. 2023). Advances in architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Vision Transformers (ViTs) have further enabled SFL to achieve remarkable performance in downstream tasks, including land cover classification (Hong et al. 2019), change detection (P. Chen et al. 2022) and scene recognition (C. Luo, Jin, and Sun 2019). However, most representations learned by SFL are directly related to the labels, which limits their generalization ability. Typically, it requires re-labelling the data and retraining the model to adapt to new tasks.

Compared to SFL, self-supervised feature learning (SSFL) reduces reliance on labelled data by extracting feature representations from large volumes of unlabelled remote sensing imagery. Existing SSFL methods can be broadly categorized into three types: generative, predictive, and contrastive, each suited for different tasks (Y. Wang et al. 2022). Figure 5 illustrates the general structures of the three main types of SSFL models. The core objective of generative methods is to reconstruct the input from a compressed representation. To achieve this goal, the model is compelled to retain the understanding of spatial continuity and local texture consistency within geographic data, thereby enabling the extraction of hidden geographic patterns and spatial effects. Generative models, represented by autoencoders (Bank, Koenigstein, and Giryes 2023) and their variants (He et al. 2022; Kingma and Welling 2013), focus on reconstructing input data to derive meaningful features, which have demonstrated strong performance in low-level visual tasks such as denoising (X. Wang et al. 2022), unmixing (Hong et al. 2022), and image fusion (Rajaei, Abiri, and Helfroush 2024).

Predictive methods emphasize learning semantic context features, such as spatial and spectral relationships, making them well-suited for tasks like remote-sensing image rotation prediction (Ji et al. 2022) and multispectral feature prediction (X. Yang et al. 2022). DeepCluster (Caron et al. 2018) is a representative predictive method. The core idea is to iteratively perform feature extraction and clustering. It uses a neural network to extract image features, applies K-means clustering to assign pseudo-labels, and then trains the network using these pseudo-labels as supervision, progressively enhancing the quality of feature representations. This approach has been applied to estimate socio-demographic variables, such as population, from remote sensing imagery, providing frequent and reliable local population estimates (Neal et al. 2022).

Contrastive methods in remote sensing can address different geographic tasks by designing specific types of positive and negative sample pairs. For example, using images from nearby locations as positive pairs encourages spatial dependency, making neighbouring regions more similar in representation and better reflecting the spatial structure of geographic space. Studies have applied classic contrastive models like SimCLR (T. Chen et al. 2020) and Moco (He et al. 2020) to remote sensing tasks such as scene interpretation and classification (Tao et al. 2022). Some approaches leverage the unique spatiotemporal attributes of remote sensing imagery. They construct spatial location-based contrasts (Jean et al. 2019) and time series-based contrasts (H. Huang et al. 2022) to support tasks such as semantic segmentation (H. Li et al. 2022) and image classification (Guan and Lam 2022). Multi-modal contrastive methods compare positive samples from multi-modal data of the same scene against negative samples from different scenes. Notable studies have explored cross-modal matching, such as remote sensing imagery with text (Yuan et al. 2022), audio (Heidler et al. 2023), and Synthetic Aperture Radar (SAR) images (Jain, Schoen-Phelan, and Ross 2022). These

approaches produce feature representations rich in implicit information, thereby enhancing the performance of downstream tasks.

In summary, remote-sensing image representation learning based on SFL and SSFL has witnessed rapid development in recent years, showing great potential in handling complex and diverse data. SSFL, with its low reliance on labelled data and strong generalization, has gradually become the main trend for future development. Among the three categories of SSFL models, contrastive models have shown superior performance in recent years. Their ability to directly learn highly discriminative features makes them well-suited for a wide range of downstream tasks. Compared to generative models, contrastive approaches typically offer stronger generalization capabilities, and they also exhibit greater training stability than predictive methods. Additionally, the emergence of LLMs has also spurred interest in Large Remote Sensing Models (Hu et al. 2025; F. Liu et al. 2024; Z. Zhang et al. 2024). The multi-modal learning and cross-modal understanding capabilities of large models enhance the feature representation of remote sensing images. This results in representations that are richer in information and better suited for various downstream tasks. This creates unprecedented opportunities for remote sensing pretraining and alignment with cross-modal data, advancing representation learning in remote sensing towards greater intelligence.

2.6. Representation learning for street view imagery

Street view imagery is captured along urban road networks using either dedicated vehicles or crowdsourced devices, providing high-resolution images from a pedestrian's perspective (Biljecki and Ito 2021; F. Zhang et al. 2024). Compared to the macro-level depiction of the urban physical environment provided by remote sensing imagery (Y. Huang, Sanatani, et al. 2025), street view imagery offers detailed representations that enable fine-grained analyses, such as identifying building styles and ages (Sun et al. 2022), assessing road quality (Chacra and Zelek 2018), and classifying types of street stores (Noorian, Psyllidis, and Bozzon 2019). To effectively extract meaningful information from such complex, multidimensional pixel data contained in these images for subsequent analysis and decision-making, robust representation learning methods are essential (see Figure 6). Deep learning has emerged as the core methodology for representation learning in street view imagery. By employing multi-layer neural network architectures as feature extractors, such as CNNs and Transformers, deep learning methods can progressively extract visual information – from low-level edges and textures to high-level semantic features, such as street layouts, street furniture, and human

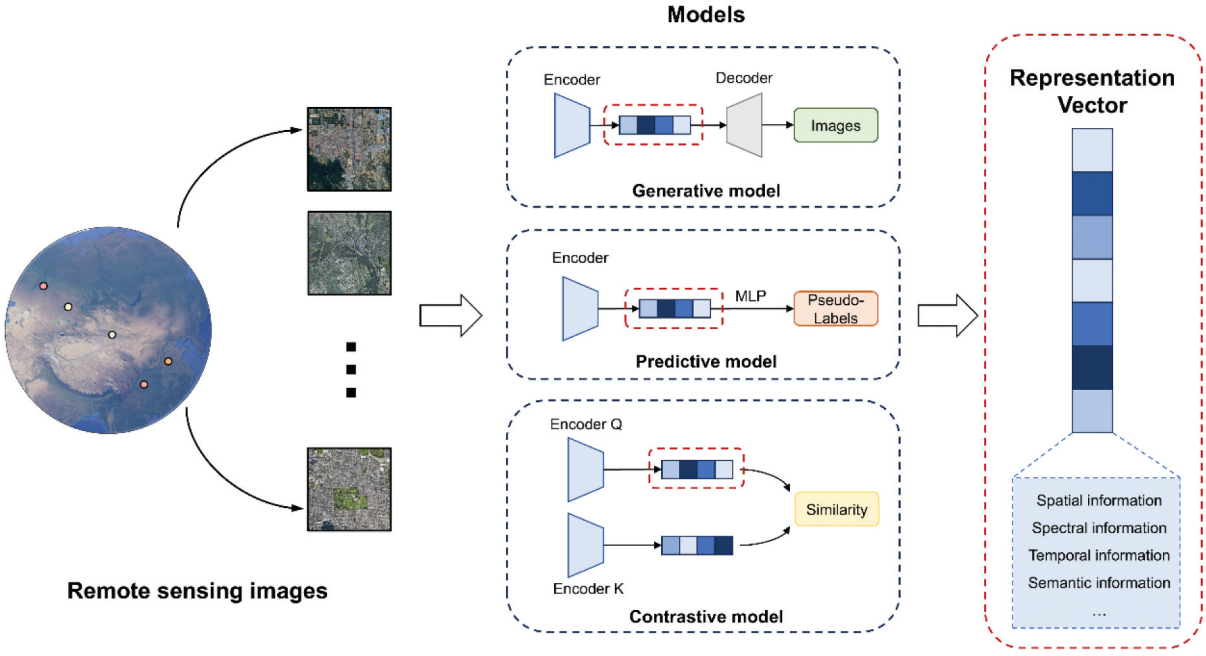

Figure 5. Representation learning for remote sensing imagery, illustrating the general structures of the three main types of self-supervised learning models: generative, predictive, and contrastive models.

activity (Y. Huang et al. 2023; Y. Huang, F. Zhang et al. 2025). Specifically, similar to representation learning for remote sensing imagery, representation learning for street view imagery can also be categorized into supervised and self-supervised approaches.

The availability of large-scale annotated, scene-centric datasets, such as Places and Place Pulse, has greatly facilitated the supervised learning of street view imagery (Hou et al. 2024; F. Zhang et al. 2018; B. Zhou et al. 2018). While these datasets are labelled for specific tasks, such as scene classification in Places and human perception evaluation in Place Pulse, the representation vectors learned from them encode rich and diverse information about the urban physical environment, extending beyond the scope of their original annotations. For instance, the work of (Y. Huang et al. 2023) demonstrates that leveraging a model pre-trained on the Places dataset for feature extraction effectively captures the semantic information of urban scenes, thereby enabling a comprehensive representation of urban areas.

The advent of self-supervised learning, combined with vast quantities of unlabelled street view imagery, has provided new vitality into the field of representation learning (Stalder et al. 2024). Without requiring manual annotations, self-supervised methods rely on carefully designed pretext tasks to automatically extract representation vectors, while maintaining flexibility for adaptation to downstream tasks (Z. Wang, Li, and Rajagopal 2020). In the context of street view imagery, most current approaches adopt contrastive learning frameworks. A fundamental component of these frameworks is the construction of positive and negative sample pairs, which has traditionally been guided by principles like self-transformations. To capture the unique characteristics of geographic space more effectively, recent approaches have begun to integrate geographical principles. For example, the principle of spatial autocorrelation is utilized to refine positive sampling, based on the premise that geographically proximate images are more likely to be semantically similar. A recent study (Y. Li et al. 2025) systematically compares these three self-supervised strategies and reveals their applicability to different urban tasks. 'Self-contrast' emphasizes global information and is thus suitable for tasks involving abundant dynamic elements and human perception; 'temporal contrast' focuses on static features, facilitating stable representations for tasks such as place recognition; and 'spatial contrast' captures the socioeconomic and cultural ambiences shared by neighbouring scenes, making it especially beneficial for macro-level analyses like socioeconomic indicator prediction.

In sum, as annotated data accumulate and self-supervised learning methods develop, street view imagery representation learning has established a clear methodological continuum from traditional supervised models to flexible and varied self-supervised approaches, thereby laying a robust and enriched theoretical groundwork for a wide range of urban environment analysis.

3. Multi-modal representation learning

3.1. The development phase

Multi-modal representation learning aims to integrate geospatial data from various modalities, such as trajectory, imagery, and spatial network data. Facing challenges inherent in geospatial data, such as its heterogeneity, high dimensionality, and complex interconnections, the main purposes of this approach are as follows: a) Enhancing Robustness and Generalization: By combining information from multiple sources, the model can improve its robustness and generalization capabilities in geospatial analysis tasks. This means that the model can perform better across a variety of different situations and data types, reducing overfitting to a single data source. b) Enriching Understanding of Geographical Entities and Processes: Multi-modal learning allows for a richer interpretation of geographical objects and processes by providing multiple perspectives on the same object. This can lead to deeper insights into the dynamics and characteristics of geographical features. Supported by self-supervised representation learning techniques, multi-modal learning has become a hot research topic in the field of geospatial sciences. The development of this field can generally be divided into three stages (Figure 7):

3.1.1 Independent Modal Modelling

In this stage, feature extraction neural networks for different modalities operate independently, often aligning modalities through contrastive learning. For instance, models like CLIP (Radford et al. 2021) use separate visual and textual encoders, learning semantic consistency across modalities via a contrastive loss,

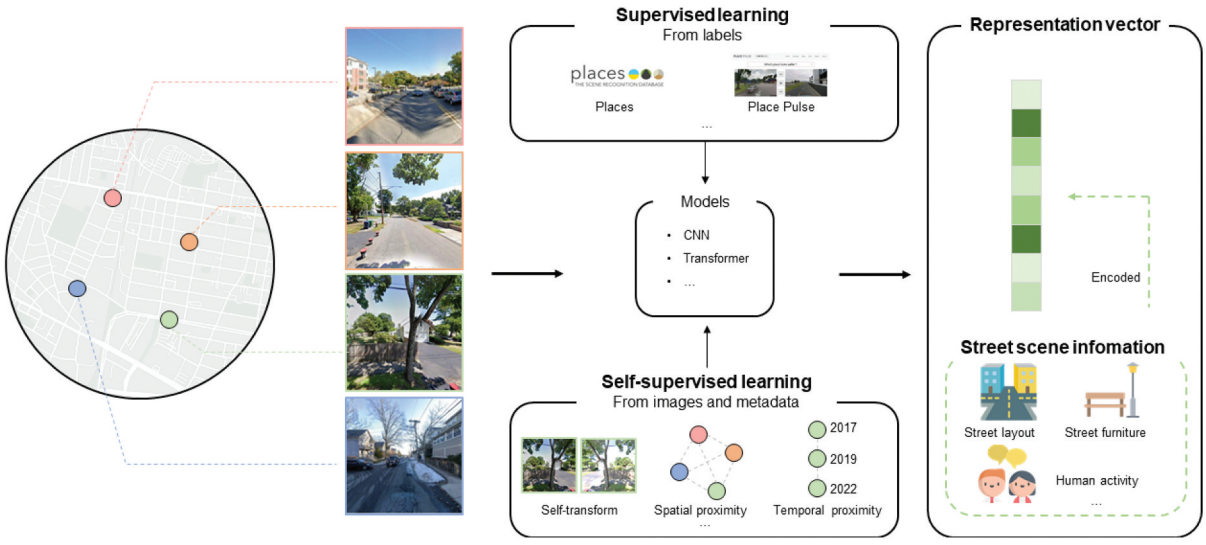

Figure 6. Methods for learning representations from street view imagery. Representation vectors can be extracted either through supervised learning based on large annotated datasets, or through self-supervised learning that leverages the intrinsic properties of the image itself or associated metadata.

achieving cross-modal alignment. In geospatial applications, models like UrbanVLP (Hao et al. 2024) and UrbanCLIP (Y. Yan et al. 2024) utilize contrastive learning between street view and remote sensing images to enhance urban scene analysis capabilities. These methods, by employing independent encoders and contrastive learning, effectively address challenges like geospatial data heterogeneity, sparsity, and specific modal feature extraction, thereby improving the accuracy of fine-grained classification and economic assessment in complex urban environments. For example, UrbanCLIP has been applied to urban land use classification and street-level socioeconomic status estimation in major Chinese cities, demonstrating significant improvements in accuracy compared to unimodal baselines. GeoJEPA (Lundqvist and Delvret 2025) uses masked learning to pretrain on a large dataset of OpenStreetMap attributes, geometries, and aerial images to generate multi-modal representations for geospatial data.

3.1.2 Unified Modal Modelling

At this stage, a shared network architecture is employed to extract features from multiple modalities, deepening modal alignment through a combination of contrastive and masked learning. Models like BLIP (J.Li et al. 2022) introduce masked learning into the contrastive framework, enhancing the interaction and alignment between modalities. Models such as FLAVA (Singh et al. 2022) and ViLT (W. Kim, Son, and Kim 2021), based on a shared Transformer backbone, perform modal alignment and integration within a unified structure, significantly enhancing the representation of cross-modal features. Unified modal modelling, through shared architectures and multi-task learning, more deeply fuses multi-modal geospatial information. It excels at capturing complex correlations between geospatial features of varying scales and granularities, thereby improving performance in tasks such as urban understanding and environmental monitoring. Unified Modal Modelling methods have demonstrated the potential of unified multi-modal architectures for a range of GeoAI tasks, including urban scene understanding, cross-modal geospatial representation, and environmental monitoring. For example, PDFM (Agarwal et al. 2024) has been applied in real-world scenarios such as public health monitoring, retail site selection, climate risk assessment, and socioeconomic indicator mapping, demonstrating the practical value of multi-modal GeoAI frameworks.

3.1.3 LLM Enhancement

With the rise of LLMs, the latest multi-modal learning frameworks utilize LLMs as core modules, implementing transformation alignment through adapter layers (Yin et al. 2024). This method maps features from different modalities to a unified textual latent space, leveraging the large-scale parameters of LLMs to learn efficient mappings and alignments between modalities. Models like Flamingo (Alayrac et al. 2022) and

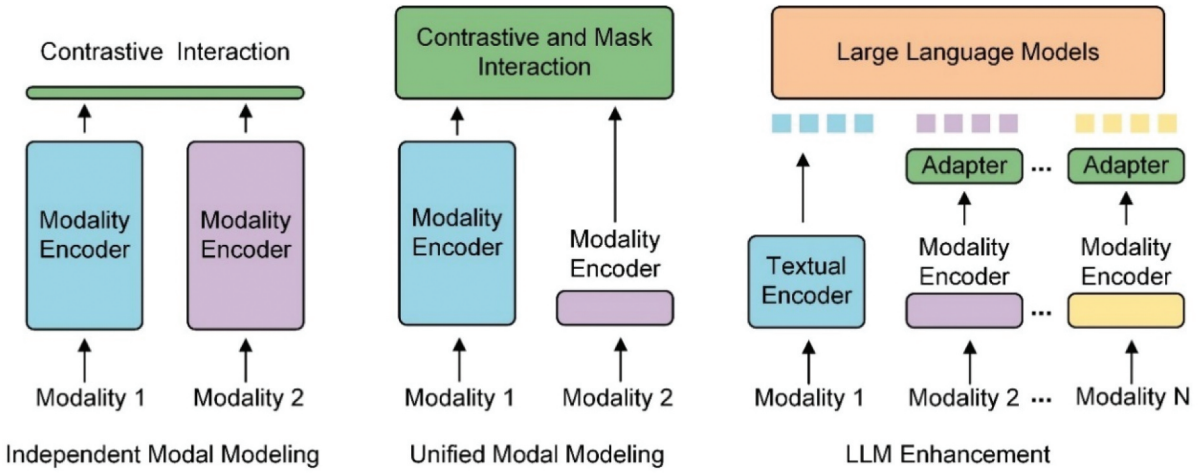

Figure 7. Developmental stages of multi-modal representation learning.

LLaMA-Adapter (R.Zhang et al. 2024), by combining the generative and reasoning capabilities of LLMs, strengthen the alignment and interaction between modalities. LLM enhancement frameworks leverage the powerful reasoning and knowledge integration capabilities of LLMs to bridge the semantic gap between modalities. They utilize pre-trained knowledge to interpret complex geospatial semantics (i.e. detailed meanings) and analyse spatiotemporal dynamics with higher precision. For example, AllSpark (Shao et al. 2025) integrates spatiotemporal information, supporting more modalities and handling complex spatiotemporal scenarios. In the geospatial domain, this transformational alignment not only improves cross-modal synergy but also facilitates real-world applications. For example, Hu et al. (2025) contribute a high-quality, human-annotated remote sensing captioning dataset (RSICap) and a comprehensive benchmark (RSIEval), enabling the development and evaluation of large vision-language models specifically tailored for remote sensing image understanding. These advances have significant potential in practical GeoAI applications, including disaster monitoring, land use mapping, and environmental change detection, where the integration and reasoning of multi-modal data are essential.

3.2. Multi-modal interactive representation (MMIR)

Geospatial data originate from diverse spatiotemporal processes, offering complementary multi-modal information. While fusing these modalities into a unified representation for each geographical unit could theoretically support various geospatial tasks, conventional fusion approaches (e.g. simple concatenation of representations) often fail to capture the nuanced relationships between modalities. A more promising direction involves developing frameworks that enable dynamic, context-aware interactions between modalities – where each modality adaptively aligns and contributes its strengths based on the specific task requirements. This paradigm mirrors the behaviour of complex adaptive systems, where components interact organically while maintaining their distinct identities. For instance, in geographic question answering, the relevant modalities and their interactions may vary significantly depending on whether the question concerns urban infrastructure, environmental patterns, or socioeconomic factors.

A fundamental challenge in multi-modal fusion lies in establishing a unified feature space that effectively accommodates all modalities (Baars 1993; Huh et al. 2024). To address this, we propose Multi-modal Interactive Representation (MMIR), a novel paradigm that maps each modality's representation into a shared language space via LLMs (Figure 8). Unlike traditional fusion approaches that merge modalities into a single representation, MMIR enables: 1) Dynamic inter-modal communication through intrinsic semantic links, 2) Maximal preservation of each modality's distinct features, and 3) Task-driven reasoning where LLMs selectively leverage modalities based on contextual needs. The rationale for using language as the unifying modality stems from two key reasons. First, language inherently encapsulates vast amounts of geospatial and human experiential knowledge – from qualitative descriptions of places to subjective perceptions of space (Adams and McKenzie 2013; Tuan

1979). This makes it uniquely capable of providing semantic grounding for abstract concepts across modalities through shared meanings. Second, recent advancements in large language models (LLMs), such as GPT-4 (OpenAI et al. 2024) and DeepSeek R1 (DeepSeek-AI et al. 2025), have endowed these systems with the technical capability to decode and systematically organize linguistic knowledge. Simultaneously, they function as adaptive interfaces for cross-modal alignment and contextual reasoning. A growing body of evidence suggests that pretrained LLMs exhibit surprising versatility across other modalities. For instance, Sharma et al. (2024) demonstrated that LLMs trained exclusively on language data possess a rich understanding of visual structures. Similarly, Pang et al. (2024) found that LLMs trained solely on text data can serve as effective representations for purely visual tasks using a simple yet effective approach: leveraging a frozen transformer block from pretrained LLMs as a constituent encoder layer to process visual tokens directly. In the domain of visual generation, LLMs have shown remarkable capabilities to enhance captions with visual structures (e.g. bounding boxes) and improve generation quality (Betker et al. 2023; Lian, Shi et al. 2023; Lian, Li et al. 2023). Ngo and Kim (2024) further demonstrated that auditory models can achieve approximate alignment with LLMs through a simple linear transformation, while Ng et al. (2023) highlighted the effectiveness of pretrained LLMs in tasks such as facial motion prediction.

Practically, our MMIR framework extends the Language as Reference Framework (LaRF) (Shao et al. 2025) by using linguistic structures to unify spatiotemporal modalities in a shared representation space while preserving their distinct features. MMIR's key strength lies in its potential for generalization, especially for spatial and temporal knowledge transfer across different geographic regions and time periods. This is enabled by the language's inherent capacity to abstractly represent and recompose real-world knowledge, along with its interpretability that allows precise control over contextual specifications. For instance, when predicting flood risks in a data-scarce region, the system can specify conditions such as 'low-lying urban area near river systems during monsoon season' and retrieve similar patterns in real-world knowledge. Language models can further leverage reasoning capabilities (DeepSeek-AI et al. 2025) to not only better analogize similar scenarios but also demonstrate the rationale behind model predictions. For example, the system might explain that 'elevated flood risk arises from factors such as topographic vulnerability, urban impermeability, and hydrological proximity, as observed in similar metropolitan areas'. By utilizing linguistic structures as a cross-modal bridge, combined with knowledge abstraction and reasoning capabilities, the framework can generalize to novel scenarios beyond its training distribution. This enables it to overcome data limitations through compact yet expressive semantic encoding. While this approach shows promise in addressing spatial-data challenges – including modality imbalance, cross-region transfer, and temporal shifts – its practical efficacy awaits further validation in real-world implementations.

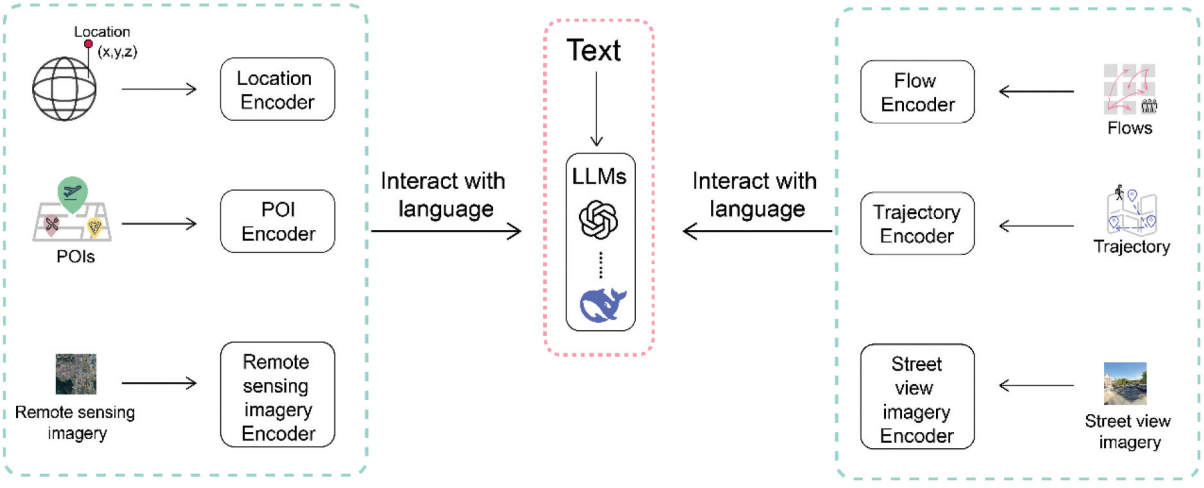

Figure 8. The MMIR framework: language-space based dynamic interactive representation for multi-modal fusion.

4. Conclusions

Geographical representation learning has emerged as a powerful approach to addressing fundamental challenges in spatiotemporal systems. By leveraging diverse data sources and advanced algorithms, it facilitates a comprehensive understanding of geographical phenomena. Recent research has made significant strides in geographical unit representation learning, particularly in single-modal feature extraction and multi-modal fusion. Through the integration of heterogeneous data sources, such as satellite imagery, mobility patterns, and POI distributions, these methods enable richer and more holistic representations of geographical units. However, critical challenges remain, including (1) achieving effective cross-modal data alignment and (2) modelling relationships between inter-dimensional features.

Current approaches typically generate geographical unit representations through non-interactive fusion (e.g. late concatenation) of unimodal features, rather than performing semantically aligned fusion. While existing representations demonstrate effectiveness in applications like land-use classification and urban morphology analysis, they often inadequately capture the intricate spatial relationships between geographical units. Current models successfully extract discrete features of individual units, yet they still struggle with two key limitations: (1) cross-modal data alignment challenges and (2) limited interpretability of their high-dimensional feature representations. Addressing these challenges will enable geographical unit representation learning to better cope with the dynamic and multi-scale nature of complex geographical environments.

Building upon recent advances in LLMs, we have proposed the MMIR framework as a promising new paradigm for multi-modal geospatial intelligence. By harnessing the language's inherent abstraction capabilities to dynamically connect and compose across modalities, MMIR offers a potential pathway towards more generalizable and data-efficient systems – enabling reasoning about novel scenarios through semantic recombination while working within practical data constraints. Though challenges like modality imbalance and temporal shifts require further investigation, we believe this language-anchored approach may open new directions for developing more adaptable and interpretable multi-modal architectures that better capture the complexity of real-world environments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Yu Liu http://orcid.org/0000-0002-0016-2902

References

Adams, B., and G. McKenzie. 2013. "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions." In *Crowdsourcing Geographic Knowledge*, edited by D. Sui, S. Elwood, and M. Goodchild, 201–221. Dordrecht: Springer Netherlands.

Agarwal, M., M. Sun, C. Kamath, A. Muslim, P. Sarker, J. Paul, and G. Prasad. 2024. "General geospatial inference with a population dynamics foundation model." *arXiv preprint arXiv:2411.07207*. https://arxiv.org/abs/2411.07207

Alahi, A., K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. 2016. "Social LSTM: Human Trajectory Prediction in Crowded Spaces." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 961–971.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, et al. 2022. "Flamingo: A Visual Language Model for Few-Shot Learning." In *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 23716–23736. Vol. 35. Red Hook, NY, USA: Curran Associates, Inc.

Anderson, J. E. 2011. "The Gravity Model." *Annual Review of Economics* 3 (1): 133–160. https://doi.org/10.1146/annurev-economics-111809-125114.

Baars, B. J. 1993. *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.

Bai, L., W. Huang, X. Zhang, S. Du, G. Cong, H. Wang, and B. Liu. 2023. "Geographic Mapping with Unsupervised Multi-Modal Representation Learning from VHR Images and POIs." *ISPRS Journal of Photogrammetry & Remote Sensing* 201:193–208. https://doi.org/10.1016/j.isprsjprs.2023.05.006.

Bank, D., N. Koenigstein, and R. Giryes. 2023. "Autoencoders." In L. Rokach, O. Maimon, & E. Shmueli (Eds.), *Machine Learning for Data Science Handbook*. pp. 353–353. Springer, Berlin. https://doi.org/10.1007/978-3-031-24628-9_16.

Barthélemy, M. 2011. "Spatial Networks." *Physics Reports* 499 (1–3): 1–101. https://doi.org/10.1016/j.physrep.2010.11.002.

Bastani, F., P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi. 2023. "Satlaspretrain: A Large-Scale Dataset for Remote Sensing Image Understanding." *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*: 16726–16736. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/iccv51070.2023.01538.

Batty, M. 2013. *The New Science of Cities*. Cambridge, MA: MIT press.

Bengio, Y., A. Courville, and P. Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 35 (8): 1798–1828. https://doi.org/10.1109/tpami.2013.50.

Betker, J., G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, and A. Ramesh. 2023. "Improving Image Generation with Better Captions." *Computer Science* 2 (3): 8. https://cdn.openai.com/papers/dall-e-3.pdf.

Biljecki, F., and K. Ito. 2021. "Street View Imagery in Urban Analytics and GIS: A Review." *Landscape and Urban Planning* 215:104217. https://doi.org/10.1016/j.landurbplan.2021.104217.

Bing, J., M. Chen, M. Yang, W. Huang, Y. Gong, and L. Nie. 2023. "Pre-Trained Semantic Embeddings for POI Categories Based on Multiple Contexts." *IEEE Transactions on Knowledge and Data Engineering* 35 (9): 8893–8904. https://doi.org/10.1109/TKDE.2022.3218851.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, et al. 2021. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258*. https://arxiv.org/abs/2108.07258.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. "Translating Embeddings for Modeling Multi-Relational Data." In *Advances in Neural Information Processing Systems*, edited by C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 2787–2795. Vol. 26. Red Hook, NY, USA: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

Cai, L., B. Yan, G. Mai, K. Janowicz, and R. Zhu. 2019. "TransGCN: Coupling transformation assumptions with graph convolutional networks for link prediction." In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19)* (pp. 131–138). New York, NY, USA: ACM. https://doi.org/10.1145/3360901.3364441.

Cao, H., F. Xu, J. Sankaranarayanan, Y. Li, and H. Samet. 2020. "Habit2Vec: Trajectory Semantic Embedding for Living Pattern Recognition in Population." *IEEE Transactions on Mobile Computing* 19 (5): 1096–1108. https://doi.org/10.1109/TMC.2019.2902403.

Caron, M., P. Bojanowski, A. Joulin, and M. Douze. 2018. "Deep Clustering for Unsupervised Learning of Visual Features." In *Proceedings of the European Conference on Computer Vision (ECCV)*, edited by V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss, 132–149. Germany: Springer. https://doi.org/10.1007/978-3-030-01264-9_9.

Chacra, D. A., and J. Zelek. 2018. "Municipal Infrastructure Anomaly and Defect Detection." In *2018 26th European Signal Processing Conference (EUSIPCO)*, 2125–2129. Piscataway, NJ, USA: IEEE. https://doi.org/10.23919/EUSIPCO.2018.8553322.

Chang, B., Y. Park, D. Park, S. Kim, and J. Kang. 2018. "Content-Aware Hierarchical Point-of-Interest Embedding Model for Successive POI Recommendation." In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3301–3307. California, USA: International Joint Conferences on Artificial Intelligence (IJCAI) Organization. https://doi.org/10.24963/ijcai.2018/458.

Chang, Y., J. Qi, Y. Liang, and E. Tanin. 2023. "Contrastive Trajectory Similarity Learning with Dual-Feature Attention." In *Proceedings of the 2023 IEEE 39th International Conference on Data Engineering*, 2933–2945. Piscataway, New Jersey, USA: IEEE. https://doi.org/10.1109/ICDE55515.2023.00224.

Chen, M., Y. Zhao, Y. Liu, X. Yu, and K. Zheng. 2022. "Modeling Spatial Trajectories with Attribute Representation Learning." *IEEE Transactions on Knowledge and Data Engineering* 34 (4): 1902–1914. https://doi.org/10.1109/TKDE.2020.3001025.

Chen, M., L. Zhu, R. Xu, Y. Liu, X. Yu, and Y. Yin. 2022. "Embedding Hierarchical Structures for Venue Category Representation." *ACM Transactions on Information Systems* 40 (3): 1–29. https://doi.org/10.1145/3478285.

Chen, P., B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li. 2022. "FCCDN: Feature Constraint Network for VHR Image Change Detection." *ISPRS Journal of Photogrammetry & Remote Sensing* 187:101–119. https://doi.org/10.1016/j.isprsjprs.2022.02.021.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. 2020. "A Simple Framework for Contrastive Learning of Visual Representations." In *Proceedings of the 37th International Conference on Machine Learning*, edited by H. Daumé III and A. Singh, Vol. 119, 1597–1607. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v119/chen20j.html.

Chen, Y., X. Li, G. Cong, Z. Bao, C. Long, Y. Liu, A. K. Chandran, and R. Ellison. 2021. "Robust Road Network Representation Learning: When Traffic Patterns Meet Traveling Semantics." In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, edited by G. Demartini, H. Liu and W. Wang, 211–220. Piscataway, NJ, USA: ACM. https://doi.org/10.1145/3459637.3482293.

Cole, E., G. Van Horn, C. Lange, A. Shepard, P. Leary, P. Perona, S. Loarie, and O. Mac Aodha. 2023. "Spatial Implicit Neural Representations for Global-Scale Species Mapping." In *Proceedings of the 40th International Conference on Machine Learning*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Vol. 202, 6320–6342. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v202/cole23a.html.

Couclelis, H. 1986. "Artificial Intelligence in Geography: Conjectures on the Shape of Things to Come." *Professional Geographer* 38 (1): 1–11. https://doi.org/10.1111/j.0033-0124.1986.00001.x.

Dai, S., J. Wang, C. Huang, Y. Yu, and J. Dong. 2021. "Temporal Multi-View Graph Convolutional Networks for Citywide Traffic Volume Inference." In *2021 IEEE International Conference on Data Mining (ICDM)*, edited by J. Bailey, P. Miettinen, Y. S. Koh, D. Tao and X. Wu, 1042–1047. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/ICDM51629.2021.00120.

Damiani, M. L., A. Acquaviva, F. Hachem, and M. Rossini. 2020. "Learning Behavioral Representations of Human Mobility." In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, edited by Y. Huang, S. Newsam and L. Xiong, 367–376. New York, NY, USA: ACM. https://doi.org/10.1145/3397536.3422255.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, et al. 2025. "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning". arXiv preprint arXiv:2501.12948. https://arxiv.org/abs/2501.12948.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by J. Burstein, C. Doran and T. Solorio, Vol. 1, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Du, L., X. You, K. Li, L. Meng, G. Cheng, L. Xiong, and G. Wang. 2019. "Multi-Modal Deep Learning for Landform Recognition." *ISPRS Journal of Photogrammetry & Remote Sensing* 158:63–75. https://doi.org/10.1016/j.isprsjprs.2019.09.018.

Dufour, N., V. Kalogeiton, D. Picard, and L. Landrieu. 2025. "Around the world in 80 timesteps: A generative approach to global visual geolocation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 23016–23026. Piscataway, NJ, USA: IEEE.

Fan, J., and G. Thakur. 2023. "Towards POI-Based Large-Scale Land Use Modeling: Spatial Scale, Semantic Granularity, and Geographic Context." *International Journal of Digital Earth* 16 (1): 430–445. https://doi.org/10.1080/17538947.2023.2174607.

Fan, Z., F. Zhang, B. P. Y. Loo, and C. Ratti. 2023. "Urban Visual Intelligence: Uncovering Hidden City Profiles With Street View Images." *Proceedings of the National Academy of Sciences of the United States of America* 120 (27): e2220417120. https://doi.org/10.1073/pnas.2220417120.

Fang, Z., Y. Du, X. Zhu, D. Hu, L. Chen, Y. Gao, and C. S. Jensen. 2022. "Spatio-Temporal Trajectory Similarity Learning in Road Networks." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, edited by A. Zhang and H. Rangwala, 347–356. New York, NY, USA: ACM. https://doi.org/10.1145/3534678.3539375.

Fotheringham, A. S. 1981. "Spatial Structure and Distance-Decay Parameters." *Annals of the Association of American Geographers* 71 (3): 425–436. https://doi.org/10.1111/j.1467-8306.1981.tb01367.x.

Fotheringham, A. S. 1983. "Some Theoretical Aspects of Destination Choice and Their Relevance to Production-Constrained Gravity Models." *Environment & Planning A: Economy & Space* 15 (8): 1121–1132. https://doi.org/10.1068/a151121.

Fotheringham, A. S., and M. E. O'Kelly. 1989. *Spatial Interaction Models: Formulations and Applications*. Netherlands: Kluwer Academic Publishers.

Fu, T.-Y., and W.-C. Lee. 2020. "Trembr: Exploring Road Networks for Trajectory Representation Learning." *ACM Transactions on Intelligent Systems and Technology* 11 (1): 10:1–10:25. https://doi.org/10.1145/3361741.

Gao, S. 2024. "Artificial intelligence and human geography." In The Encyclopedia of Human Geography, edited by B. Warf, 1–7. Springer, Berlin.

Gong, Z., C. Wang, Y. Chen, B. Liu, P. Zhao, and Z. Zhou. 2024. "Learning Spatial Interaction Representation with Heterogeneous Graph Convolutional Networks for Urban Land-Use Inference." *International Journal of Geographical Information Science* 38 (11): 2235–2271. https://doi.org/10.1080/13658816.2024.2379473.

Goodchild, M. F. 2004. "The Validity and Usefulness of Laws in Geographic Information Science and Geography." *Annals of the Association of American Geographers* 94 (2): 300–303. https://doi.org/10.1111/j.1467-8306.2004.09402008.x.

Goodchild, M. F., and W. Li. 2021. "Replication Across Space and Time Must Be Weak in the Social and Environmental Sciences." *Proceedings of the National Academy of Sciences of the United States of America* 118 (35): e2015759118. https://doi.org/10.1073/pnas.2015759118.

Guan, P., and E. Y. Lam. 2022. "Cross-Domain Contrastive Learning for Hyperspectral Image Classification." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5528913): 1–13. https://doi.org/10.1109/TGRS.2022.3176637.

Hägerstrand, T. 1970. "What About People in Regional Science?" *Papers of the Regional Science Association* 24 (1): 6–21. https://doi.org/10.1007/BF01936872.

Hao, X., W. Chen, Y. Yan, S. Zhong, K. Wang, Q. Wen, and Y. Liang. 2024. "UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (27): 28061–28069. https://doi.org/10.1609/aaai.v39i27.35024.

Haydari, A., D. Chen, Z. Lai, M. Zhang, and C.-N. Chuah. 2024. "MobilityGPT: Enhanced Human Mobility Modeling with a GPT Model" arXiv preprint arXiv:2402.03264. https://doi.org/10.48550/arXiv.2402.03264.

He, K., X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. 2022. "Masked Autoencoders Are Scalable Vision Learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15979–15988. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/CVPR52688.2022.01553.

He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by C. Liu, G. Mori, K. Saenko and S. Savarese. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/cvpr42600.2020.00975.

Heidler, K., L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu. 2023. "Self-Supervised Audiovisual Representation Learning for Remote Sensing Data." *International Journal of Applied Earth Observation and Geoinformation* 116:103130. https://doi.org/10.1016/j.jag.2022.103130.

Helpman, E., M. Melitz, and Y. Rubinstein. 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes." *Quarterly Journal of Economics* 123 (2): 441–487. https://doi.org/10.1162/qjec.2008.123.2.441.

Hong, D., L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, and B. Zhang. 2022. "Endmember-Guided Unmixing Network (EGU-Net): A General Deep Learning Framework for Self-Supervised Hyperspectral Unmixing." *IEEE Transactions on Neural Networks and Learning Systems* 33 (11): 6518–6531. https://doi.org/10.1109/TNNLS.2021.3082289.

Hong, D., N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu. 2019. "Learnable Manifold Alignment (LeMA): A Semi-Supervised Cross-Modality Learning Framework for Land Cover and Land Use Classification." *ISPRS Journal of Photogrammetry & Remote Sensing* 147:193–205. https://doi.org/10.1016/j.isprsjprs.2018.10.006.

Hou, Y., M. Quintana, M. Khomiakov, W. Yap, J. Ouyang, K. Ito, Z. Wang, T. Zhao, and F. Biljecki. 2024. "Global Streetscapes: A Comprehensive Dataset of 10 Million Street-Level Images Across 688 Cities for Urban Science and Analytics." *ISPRS Journal of Photogrammetry & Remote Sensing* 215:216–238. https://doi.org/10.1016/j.isprsjprs.2024.06.023.

Hu, Y., J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li. 2025. "RSGPT: A Remote Sensing Vision Language Model and Benchmark." *ISPRS Journal of Photogrammetry & Remote Sensing* 224:272–286. https://doi.org/10.1016/j.isprsjprs.2025.03.028.

Huang, F., J. Lv, and Y. Yue. 2024. "Jointly Spatial-Temporal Representation Learning for Individual Trajectories." *Computers, Environment and Urban Systems* 112:102144. https://doi.org/10.1016/j.compenvurbsys.2024.102144.

Huang, H., Z. Mou, Y. Li, Q. Li, J. Chen, and H. Li. 2022. "Spatial-Temporal Invariant Contrastive Learning for Remote Sensing Scene Classification." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/LGRS.2022.3173419.

Huang, J., H. Wang, Y. Sun, Y. Shi, Z. Huang, A. Zhuo, and S. Feng. 2022. "Ernie-Geol: A Geography-and-Language Pre-Trained Model and Its Applications in Baidu Maps." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, edited by A. Zhang and H. Rangwala, 3029–3039. New York, NY, USA: ACM. https://doi.org/10.1145/3534678.3539021.

Huang, W., L. Cui, M. Chen, D. Zhang, and Y. Yao. 2022. "Estimating Urban Functional Distributions with Semantics Preserved POI Embedding." *International Journal of Geographical Information Science* 36 (10): 1905–1930. https://doi.org/10.1080/13658816.2022.2040510.

Huang, W., D. Zhang, G. Mai, X. Guo, and L. Cui. 2023. "Learning Urban Region Representations with POIs and Hierarchical Graph Infomax." *ISPRS Journal of Photogrammetry & Remote Sensing* 196:134–145. https://doi.org/10.1016/j.isprsjprs.2022.11.021.

Huang, Y., R. P. Sanatani, C. Liu, Y. Kang, F. Zhang, Y. Liu, F. Duarte, and C. Ratti. 2025. "No "True" Greenery: Deciphering the Bias of Satellite and Street View Imagery in Urban Greenery Measurement." *Building & Environment* 269:112395. https://doi.org/10.1016/j.buildenv.2024.112395.

Huang, Y., F. Zhang, Y. Gao, W. Tu, F. Duarte, C. Ratti, D. Guo, and Y. Liu. 2023. "Comprehensive Urban Space Representation with Varying Numbers of Street-Level Images." *Computers, Environment and Urban Systems* 106:102043–102043. https://doi.org/10.1016/j.compenvurbsys.2023.102043.

Huang, Y., F. Zhang, L. Wu, and Y. Liu. 2025. "Measuring Urban Physical Environments Using Image Deep Features." *Cities* 166:106196. https://doi.org/10.1016/j.cities.2025.106196.

Huh, M., B. Cheung, T. Wang, and P. Isola. 2024. "The Platonic Representation Hypothesis." In *Proceedings of the 41st International Conference on Machine Learning*, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Vol. 235, 20617–20642. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v235/huh24a.html.

Izbicki, M., E. E. Papalexakis, and V. J. Tsotras. 2020. "Exploiting the Earth's Spherical Geometry to Geolocate Images." *Lecture Notes in Computer Science* 11907:3–19. https://doi.org/10.1007/978-3-030-46147-8_1.

Jain, P., B. Schoen-Phelan, and R. Ross. 2022. "Self-Supervised Learning for Invariant Representations from Multi-Spectral and SAR Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 15:7797–7808. https://doi.org/10.1109/JSTARS.2022.3204888.

Jean, N., S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. 2019. "Tile2Vec: Unsupervised representation learning for spatially distributed data." *In Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (1): 3967–3974. Palo Alto, CA: AAAI Press. https://doi.org/10.1609/aaai.v33i01.33013967.

Ji, H., Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei. 2022. "Few-Shot Scene Classification of Optical Remote Sensing Images Leveraging Calibrated Pretext Tasks." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5625513): 1–13. https://doi.org/10.1109/TGRS.2022.3184080.

Jiang, J., D. Pan, H. Ren, X. Jiang, C. Li, and J. Wang. 2023. "Self-Supervised Trajectory Representation Learning with Temporal Regularities and Travel Semantics." In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, edited by L. Chen, S. Manegold and S. Mehrotra, 843–855. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/ICDE55515.2023.00070.

Jiang, X., E. N. de Souza, A. Pesaranghader, B. Hu, D. L. Silver, and S. Matwin. 2017. "TrajectoryNet: An Embedded GPS Trajectory Representation for Point-Based Classification Using Recurrent Neural Networks." Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering, 192–200. IBM Corp. https://doi.org/10.5555/3172795.3172817.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kim, N., and Y. Yoon. 2025. "Effective Urban Region Representation Learning Using Heterogeneous Urban Graph Attention Network (HUGAT)." *IEEE Access* 13:102602–102612. https://doi.org/10.1109/ACCESS.2025.3577202.

Kim, W., B. Son, and I. Kim. 2021. "VILT: Vision-and-Language Transformer Without Convolution or Region Supervision." In *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, Vol. 139, 5583–5594. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v139/kim21k.html .

Kingma, D. P., and M. Welling. 2022. "*Auto-Encoding Variational Bayes*." arXiv preprint arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114.

Kitchin, R. 2013. "Big Data and Human Geography: Opportunities, Challenges and Risks." *Dialogues in Human Geography* 3 (3): 262–267. https://doi.org/10.1177/2043820613513388.

Klemmer, K., E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm. 2025. "SATCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery." *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (4): 4347–4355. https://doi.org/10.1609/aaai.v39i4.32457.

Kreindler, G. E., and Y. Miyauchi. 2023. "Measuring Commuting and Economic Activity Inside Cities with Cell Phone Records." *Review of Economics and Statistics* 105 (4): 899–909. https://doi.org/10.1162/rest_a_01085.

Lai, Y., Y. Su, L. Wei, T. He, H. Wang, G. Chen, D. Zha, Q. Liu, and X. Wang. 2024. "Disentangled Contrastive Hypergraph Learning for Next POI Recommendation." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, edited by C. Hauff, G. Zuccon and Y. Zhang, 1452–1462. New York, NY, USA: ACM. https://doi.org/10.1145/3626772.3657726.

Li, H., Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao. 2022. "Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5618014): 1–14. https://doi.org/10.1109/TGRS.2022.3147513.

Li, J., D. Li, C. Xiong, and S. Hoi. 2022. "BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation." In *Proceedings of the 39th International Conference on Machine Learning*, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Vol. 162, 12888–12900. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v162/li22n.html.

Li, P., M. De Rijke, H. Xue, S. Ao, Y. Song, and F. D. Salim. 2024. "Large Language Models for Next Point-of-Interest Recommendation." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Washington DC, USA, edited by G. Zuccon, G. H. Yang and Y. Fang, 1463–1472. New York, NY, USA: ACM.

Li, T., Q. Feng, B. Niu, B. Chen, F. Yan, J. Gong, and J. Liu. 2025. "Mapping Urban Villages Based on Point-of-Interest Data and a Deep Learning Approach." *Cities* 156:105549. https://doi.org/10.1016/j.cities.2024.105549.

Li, X., K. Zhao, G. Cong, C. S. Jensen, and W. Wei. 2018. "Deep Representation Learning for Trajectory Similarity Computation." In *IEEE International Conference on Data Engineering*, edited by B. C. Ooi, P. K. Chrysanthis and J. Dittrich, 617–628. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/icde.2018.00062.

Li, Y., Y. Huang, G. Mai, and F. Zhang. 2025. "Learning street view representations with spatiotemporal contrast." *arXiv preprint arXiv:2502.04638*. https://arxiv.org/abs/2502.04638.

Li, Z., J. Kim, Y. Y. Chiang, and M. Chen. 2022. "SPABERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation." In *Findings of the Association for Computational Linguistics: EMNLP 2022*, edited by Y. Goldberg, Z. Kozareva and Y. Zhang, 2757–2769. Stroudsburg, PA, USA: Association for Computational Linguistics.

Li, Z., L. Xia, J. Tang, Y. Xu, L. Shi, L. Xia, D. Yin, and C. Huang. 2024. "Urbangpt: Spatio-Temporal Large Language Models." In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, edited by F. Chierichetti, D. Koutra and R. Kumar, 5351–5362. New York, NY, USA: ACM. https://doi.org/10.1145/3637528.3671578.

Li, Z., W. Zhou, Y. Y. Chiang, and M. Chen. 2023. "GEOLM: Empowering Language Models for Geospatially Grounded Language Understanding." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, edited by H. Bouamor, J. Pino and K. Bali, 5227–5240. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lian, L., B. Li, A. Yala, and T. Darrell. 2023. "LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models." *arXiv preprint arXiv:2305.13655*. https://arxiv.org/abs/2305.13655.

Lian, L., B. Shi, A. Yala, T. Darrell, and B. Li. 2023. "LLM-grounded video diffusion models." arXiv preprint arXiv:2309.17444. https://arxiv.org/abs/2309.17444.

Liang, Y., K. Ouyang, Y. Wang, X. Liu, H. Chen, J. Zhang, Y. Zheng, and R. Zimmermann. 2022. "Trajformer: Efficient Trajectory Classification With Transformers." In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, New York, NY, USA, edited by M. Al Hasan and L. Xiong, 1229–1237. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3511808.3557481.

Liu, F., D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou. 2024. "RemoteClip: A Vision Language Foundation Model for Remote Sensing." *IEEE Transactions on Geoscience & Remote Sensing* 62 (5622216): 1–16. https://doi.org/10.1109/TGRS.2024.3390838.

Liu, K., L. Yin, F. Lu, and N. Mou. 2020. "Visualizing and Exploring POI Configurations of Urban Regions on POI-Type Semantic Space." *Cities* 99:102610. https://doi.org/10.1016/j.cities.2020.102610.

Liu, X., C. Andris, and S. Rahimi. 2019. "Place Niche and Its Regional Variability: Measuring Spatial Context Patterns for Points of Interest with Representation Learning." *Computers, Environment and Urban Systems* 75:146–160. https://doi.org/10.1016/j.compenvurbsys.2019.01.011.

Liu, Y., S. Wang, X. Wang, Y. Zheng, X. Chen, Y. Xu, and C. Kang. 2024. "Towards Semantic Enrichment for Spatial Interactions." *Annals of GIS* 30 (2): 151–166. https://doi.org/10.1080/19475683.2024.2324392.

Liu, Z., F. Zhang, J. Jiao, N. Lao, and G. Mai. 2025. "GAIR: Improving multimodal geo-foundation model with geo-aligned implicit representations." *arXiv preprint arXiv:2503.16683*. https://doi.org/10.48550/arXiv.2503.16683.

Louail, T., M. Lenormand, M. Picornell, O. García Cantú, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy. 2015. "Uncovering the Spatial Structure of Mobility Networks." *Nature Communications* 6 (1): 6007. https://doi.org/10.1038/ncomms7007.

Lundqvist, T., and L. Delvret. 2025. "GeoJEPA: Towards eliminating augmentation- and sampling bias in multimodal geospatial learning." *arXiv preprint arXiv:2503.05774*. https://doi.org/10.48550/arXiv.2503.05774.

Luo, C., L. Jin, and Z. Sun. 2019. "MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition." *Pattern Recognition* 90:109–118. https://doi.org/10.1016/j.patcog.2019.01.020.

Luo, S., and W. Hu. 2021. "Diffusion Probabilistic Models for 3D Point Cloud Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, edited by D. Forsyth, G. Gkioxari, T. Tuytelaars, R. Yang and J. Yu, 2837–2845. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/CVPR46437.2021.00286.

Luo, Y., F.-L. Chung, and K. Chen. 2022. "Urban Region Profiling via Multi-Graph Representation Learning." In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, edited by M. A. Hasan and L. Xiong, 4294–4298. New York, NY, USA: ACM. https://doi.org/10.1145/3511808.3557720.

Mac Aodha, O., E. Cole, and P. Perona. 2019. "Presence-Only Geographical Priors for Fine-Grained Image Classification." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, edited by K. M. Lee, D. Forsyth, M. Pollefeys and X. Tang, 9595–9605. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/ICCV.2019.00969.

Mai, G., K. Janowicz, L. Cai, R. Zhu, B. Regalia, B. Yan, M. Shi, and N. Lao. 2020. "Se-KGE: A Location-Aware Knowledge Graph Embedding Model for Geographic Question Answering and Spatial Semantic Lifting." *Transactions in GIS* 24 (3): 623–655. https://doi.org/10.1111/tgis.12629.

Mai, G., K. Janowicz, Y. Hu, S. Gao, B. Yan, R. Zhu, L. Cai, and N. Lao. 2022. "A Review of Location Encoding for GeoAI: Methods and Applications." *International Journal of Geographical Information Science* 36 (4): 639–673. https://doi.org/10.1080/13658816.2021.2004602.

Mai, G., K. Janowicz, B. Yan, R. Zhu, L. Cai, and N. Lao. 2020. "Multi-Scale Representation Learning for Spatial Feature Distributions Using Grid Cells." In *International Conference on Learning Representations*, edited by D. Song, K. Cho and M. White. OpenReview https://openreview.net/forum?id=rJljdh4KDH.

Mai, G., N. Lao, Y. He, J. Song, and S. Ermon. 2023. "CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations." In *Proceedings of the 40th International Conference on Machine Learning*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Vol. 202, 23498–23515. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v202/mai23a.html.

Mai, G., Y. Xuan, W. Zuo, Y. He, J. Song, S. Ermon, K. Janowicz, and N. Lao. 2023. "Sphere2Vec: A General-Purpose Location Representation Learning Over a Spherical Surface for Large-Scale Geospatial Predictions." *ISPRS Journal of Photogrammetry & Remote Sensing* 202:439–462. https://doi.org/10.1016/j.isprsjprs.2023.06.016.

Mai, G., X. Yao, Y. Xie, J. Rao, H. Li, Q. Zhu, Z. Li, and N. Lao. 2024. "SRL: Towards a General-Purpose Framework for Spatial Representation Learning." In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, edited by M. Nascimento, L. Xiong and A. Züfle, 465–468. New York, NY, USA: ACM. https://doi.org/10.1145/3678717.3691246.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, edited by C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, 3111–3119. Vol. 26. Red Hook, NY, USA: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. 2022. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." *Communications of the ACM* 65 (1): 99–106.

Murray, D., J. Yoon, S. Kojaku, R. Costas, W.-S. Jung, S. Milojević, and Y.-Y. Ahn. 2023. "Unsupervised Embedding of Trajectories Captures the Latent Structure of Scientific Migration." *Proceedings of the National Academy of Sciences* 120 (52): e2305414120. https://doi.org/10.1073/pnas.2305414120.

Neal, I., S. Seth, G. Watmough, and M. S. Diallo. 2022. "Census-Independent Population Estimation Using Representation Learning." *Scientific Reports* 12 (1): 5185. https://doi.org/10.1038/s41598-022-08935-1.

Ng, E., S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar. 2023. "Can Language Models Learn to Listen?" In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, edited by L. Agapito, Y. Furukawa, K. Grauman, K. He and I. Laptev, 10083–10093. Piscataway, NJ, USA

Ngo, J., and Y. Kim. 2024. "What Do Language Models Hear? Probing for Auditory Representations in Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by L.-W. Ku, A. Martins and V. Srikumar, 5435–5448. Stroudsburg, PA, USA: Association for Computational Linguistics.

Niu, H., and E. A. Silva. 2021. "Delineating Urban Functional Use from Points of Interest Data with Neural Network Embedding: A Case Study in Greater London." *Computers, Environment and Urban Systems* 88:101651. https://doi.org/10.1016/j.compenvurbsys.2021.101651.

Noorian, S. S., A. Psyllidis, and A. Bozzon. 2019. "St-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery." In *Web Engineering. ICWE 2019. Lecture Notes in Computer Science*, edited by M. Bakaev, F. Frasincar, and I. Ko, Vol. 11496, Cham, Switzerland. Springer. https://doi.org/10.1007/978-3-030-19274-7_3.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, et al. 2024. "GPT-4 technical report." arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774.

Openshaw, S. 1984. *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Pang, Z., Z. Xie, Y. Man, and Y.-X. Wang. 2024. "Frozen Transformers in Language Models Are Effective Visual Encoder Layers." In *Proceedings of the International Conference on Learning Representations*, edited by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, 894–916. OpenReview https://proceedings.iclr.cc/paper_files/paper/2024/file/03cd3cf3f74d4f9ce5958de269960884-Paper-Conference.pdf.

Qin, Q., S. Xu, M. Du, and S. Li. 2022. "Identifying Urban Functional Zones by Capturing Multi-Spatial Distribution Patterns of Points of Interest." *International Journal of Digital Earth* 15 (1): 2468–2494. https://doi.org/10.1080/17538947.2022.2160841.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research*, edited by M. Meila and T. Zhang, 139, 8748–8763. Massachusetts, USA: PMLR. https://proceedings.mlr.press/v139/radford21a.html.

Rajaei, A., E. Abiri, and M. S. Helfroush. 2024. "Self-Supervised Spectral Super-Resolution for a Fast Hyperspectral and Multispectral Image Fusion." *Scientific Reports* 14 (1): 29820. https://doi.org/10.1038/s41598-024-81031-8.

Rao, J., S. Gao, Y. Kang, and Q. Huang. 2020. "LSTM-TrajGAN: A deep learning approach to trajectory privacy protection." In Proceedings of the 11th International Conference on Geographic Information Science (GIScience 2021) - Part I, edited by K. Janowicz and J. A. Verstegen, 177, 12, 1–17. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany. https://doi.org/10.4230/LIPIcs.GIScience.2021.I.12.

Rao, J., S. Gao, and S. Zhu. 2023. "CATS: Conditional Adversarial Trajectory Synthesis for Privacy-Preserving Trajectory Data Publication Using Deep Learning Approaches." *International Journal of Geographical Information Science* 37 (12): 2538–2574. https://doi.org/10.1080/13658816.2023.2262550.

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang. 2021. "A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery." *Nature Communications* 12 (1): 4392. https://doi.org/10.1038/s41467-021-24638-z.

Rußwurm, M., K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia. 2024. "Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks." In *Proceedings of the Twelfth International Conference on Learning Representations*, edited by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan and Y. Sun. OpenReview https://openreview.net/forum?id=PudduufFLa.

Schlichtkrull, M., T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. 2018. "Modeling Relational Data With Graph Convolutional Networks." In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, edited by A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, and A. Hollink, L. Sabou, 593–607. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-93417-4_38.

Seo, P. H., T. Weyand, J. Sim, and B. Han. 2018. "CPlanet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps." *Lecture Notes in Computer Science*: 544–560. https://doi.org/10.1007/978-3-030-01249-6_33.

Shao, R., C. Yang, Q. Li, L. Xu, X. Yang, X. Li, M. Li, et al. 2025. "Allspark: A Multimodal Spatiotemporal General Intelligence Model With Ten Modalities via Language as a Reference Framework." *IEEE Transactions on Geoscience & Remote Sensing* 63 (5606620): 1–20. https://doi.org/10.1109/TGRS.2025.3561307.

Sharma, P., T. R. Shaham, M. Baradad, S. Fu, A. Rodriguez-Munoz, S. Duggal, and A. Torralba. 2024. "A Vision Check-Up for Language Models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, edited by Z. Akata, D. Crandall, A. Farhadi, R. Pless, I. Sato and J. Wu, 14410–14419. Piscataway, NJ, USA: IEEE.

Si, J., J. Yang, Y. Xiang, H. Wang, L. Li, R. Zhang, B. Tu, and X. Chen. 2024. "TrajBERT: BERT-Based Trajectory Recovery with Spatial-Temporal Refinement for Implicit Sparse Trajectories." *IEEE Transactions on Mobile Computing* 23 (5): 4849–4860. https://doi.org/10.1109/TMC.2023.3297115.

Simini, F., G. Barlacchi, M. Luca, and L. Pappalardo. 2021. "A Deep Gravity Model for Mobility Flows Generation." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-26752-4.

Simini, F., M. C. González, A. Maritan, and A.-L. Barabási. 2012. "A Universal Model for Mobility and Migration Patterns." *Nature* 484 (7392): 96–100. https://doi.org/10.1038/nature10856.

Singh, A., R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. 2022. "FLAVA: A Foundational Language and Vision Alignment Model." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by R. Chellappa, 15617–15629. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/CVPR52688.2022.01519.

Stalder, S., M. Volpi, N. Büttner, S. Law, K. Harttgen, and E. Süel. 2024. "Self-Supervised Learning Unveils Urban Change from Street-Level Images." *Computers, Environment and Urban Systems* 112:102156. https://doi.org/10.1016/j.compenvurbsys.2024.102156.

Sun, M., F. Zhang, F. Duarte, and C. Ratti. 2022. "Understanding Architecture Age and Style through Deep Learning." *Cities* 128:103787. https://doi.org/10.1016/j.cities.2022.103787.

Tang, J., M. Wang, M. Qu, M. Zhang, J. Yan, and Q. Mei. 2015. "LINE: Large-scale information network embedding." In *Proceedings of the 24th international conference on World Wide Web (WWW '15)*, 1067–1077. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741093.

Tao, C., J. Qi, M. Guo, Q. Zhu, and H. Li. 2023. "Self-Supervised Remote Sensing Feature Learning: Learning Paradigms, Challenges, and Future Works." *IEEE Transactions on Geoscience & Remote Sensing* 61 (5610426): 1–26. https://doi.org/10.1109/TGRS.2023.3276853.

Tao, C., J. Qi, W. Lu, H. Wang, and H. Li. 2022. "Remote Sensing Image Scene Classification With Self-Supervised Paradigm Under Limited Labeled Samples." *IEEE Geoscience & Remote Sensing Letters* 19 (8004005): 1–5. https://doi.org/10.1109/LGRS.2020.3038420.

Tian, Y., C. Chen, and M. Shah. 2017. "Cross-View Image Matching for Geo-Localization in Urban Environments." In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by A. van den Hengel, J. Kosecka, I. Aloimonos and M. Hebert, 1998–2006. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/CVPR.2017.216.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (Suppl.): 234–240. https://doi.org/10.2307/143141.

Tuan, Y.-F. 1979. "Space and Pl`Manistic Perspective." In *Philosophy in Geography*, edited by S. Gale and G. Olsson, 387–427. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9394-5_19.

Vafa, K., J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan. 2024. "Evaluating the World Model Implicit in a Generative Model." In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Vol. 37 26941–26975. Red Hook, NY, USA: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/2f6a6317bada76b26a4f61bb70a7db59-Paper-Conference.pdf.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Red Hook, NY, USA: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vivanco Cepeda, V., G. K. Nayak, and M. Shah. 2023. "GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization." In *Advances in Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, 36, 8690–8701. Red Hook, NY, USA: Curran Associates, Inc.

Vo, N., N. Jacobs, and J. Hays. 2017. "Revisiting IM2GPS in the Deep Learning Era." In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, edited by A. Zisserman, A. Fitzgibbon and S. Savarese, 2640–2649. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/ICCV.2017.286.

Wang, E., Y. Xu, Y. Yang, Y. Jiang, F. Yang, and J. Wu. 2023. "Zone-Enhanced Spatio-Temporal Representation Learning for Urban POI Recommendation." *IEEE Transactions on Knowledge and Data Engineering* 35 (9): 9628–9641. https://doi.org/10.1109/TKDE.2023.3243239.

Wang, H., and Z. Li. 2017. "Region Representation Learning via Mobility Flow." In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, edited by D. Lo, M. Winslett and C. Zhang, 237–246. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3132847.3133006.

Wang, J. 2017. "Economic Geography: Spatial Interaction." In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, edited by D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu and R. A. Marston, 1–4. Chichester, West Sussex, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118786352.wbieg0641.

Wang, P., Y. Fu, H. Xiong, and X. Li. 2019. "Adversarial Substructured Representation Learning for Mobile User Profiling." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, edited by A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi and G. Karypis, 130–138. New York, NY, USA: ACM. https://doi.org/10.1145/3292500.3330869.

Wang, X., H. Chen, and Y. Liu. 2024. "Learning Place Representations from Spatial Interactions." *International Journal of Geographical Information Science* 38 (6): 1065–1090. https://doi.org/10.1080/13658816.2024.2332908.

Wang, X., Z. Luo, W. Li, X. Hu, L. Zhang, and Y. Zhong. 2022. "A Self-Supervised Denoising Network for Satellite-Airborne-Ground Hyperspectral Imagery." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5503716): 1–16. https://doi.org/10.1109/TGRS.2021.3064429.

Wang, Y., C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu. 2022. "Self-Supervised Learning in Remote Sensing: A Review." *IEEE Geoscience and Remote Sensing Magazine* 10 (4): 213–247. https://doi.org/10.1109/MGRS.2022.3198244 .

Wang, Z., H. Li, and R. Rajagopal. 2020. "Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding." *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (1): 1013–1020. https://doi.org/10.1609/aaai.v34i01.5450.

Wang, Z., J. Zhang, Z. Zhou, Q. Cao, N. Wu, Z. Liu, L. Mu, Y. Song, Y. Xie, N. Lao, and G. Mai. 2025. "LocDiffusion: Identifying locations on Earth by diffusing in the Hilbert space." *arXiv preprint arXiv:2503.18142*. https://doi.org/10.48550/arXiv.2503.18142.

Wilson, A. 1967. "A Statistical Theory of Spatial Distribution Models." *Transportation Research* 1 (3): 253–269. https://doi.org/10.1016/0041-1647(67)90035-4.

Wu, M., and Q. Huang. 2022. "IM2City: Image Geo-Localization via Multi-Modal Learning." In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, edited by S. Newsam, G. Mai, D. Lunga, B. Martins, L. Yang and S. Gao, 50–61. New York, NY, USA: ACM. https://doi.org/10.1145/3557918.3565868.

Wu, N., Q. Cao, Z. Wang, Z. Liu, Y. Qi, J. Zhang, J. Ni, et al. 2024. *TorchSpatial: A Location Encoding Framework and Benchmark for Spatial Representation Learning*. Edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang. Vol. 37. Red Hook, NY, USA: Curran Associates, Inc.

Xia, G.-S., X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. 2018. "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, edited by D. Forsyth, I. Laptev, A. Oliva and D. Ramanan, Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/cvpr.2018.00418 .

Xu, Y., B. Zhou, S. Jin, X. Xie, Z. Chen, S. Hu, and N. He. 2022. "A Framework for Urban Land Use Classification by Integrating the Spatial Context of Points of Interest and Graph Convolutional Neural Network Method." *Computers, Environment and Urban Systems* 95:101807. https://doi.org/10.1016/j.compenvurbsys.2022.101807.

Xue, Z., G. Yang, X. Yu, A. Yu, Y. Guo, B. Liu, and J. Zhou. 2025. "Multimodal Self-Supervised Learning for Remote Sensing Data Land Cover Classification." *Pattern Recognition* 157:110959. https://doi.org/10.1016/j.patcog.2024.110959.

Yan, B., K. Janowicz, G. Mai, and S. Gao. 2017. "From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts." In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, edited by E. Hoel. New York, NY, USA: ACM. https://doi.org/10.1145/3139958.3140054.

Yan, Y., H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, and Y. Liang. 2024. "UrbanClip: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web." In *Proceedings of the ACM Web Conference 2024*, edited by H. W. Lauw, I. Keidar and Y. Maarek, 4006–4017. New York, NY, USA: ACM. https://doi.org/10.1145/3589334.3645378.

Yang, H., X. Lu, and Y. Zhu. 2021. "Cross-View Geo-Localization with Layer-to-Layer Transformer." In *Advances in Neural Information Processing Systems*, edited by M. Ranzato and A. Beygelzimer and Y. Dauphin and P.S. Liang and J. Wortman Vaughan, 29009–29020. Vol. 34. Red Hook, NY, USA: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/f31b20466ae89669f9741e047487eb37-Paper.pdf.

Yang, H., A. Yao, C. C. Whalen, and G. Mai. 2025. "BERT4Traj: Transformer-Based Trajectory Reconstruction for Sparse Mobility Data." In *Proceedings of the 13th International Conference on Geographic Information Science (GIScience 2025)*, edited by K. Sila-Nowicka, A. Moore and D. O'Sullivan Christchurch. Leibniz International Proceedings in Informatics (LIPIcs), vol. 346, pp. 8:1–8:9. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.GIScience.2025.8.

Yang, H., X. A. Yao, F. Roozkhosh, R. Liu, and G. Mai. 2025. "From Theory to Deep Learning: Understanding the Impact of Geographic Context Factors on Traffic Violations." *Computers, Environment and Urban Systems* 119:102268. https://doi.org/10.1016/j.compenvurbsys.2025.102268.

Yang, X., W. Cao, Y. Lu, and Y. Zhou. 2022. "Self-Supervised Learning with Prediction of Image Scale and Spectral Order for Hyperspectral Image Classification." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5545715): 1–15. https://doi.org/10.1109/TGRS.2022.3225663.

Yao, D., H. Hu, L. Du, G. Cong, S. Han, and J. Bi. 2022. "Trajgat: A Graph-Based Long-Term Dependency Modeling Approach for Trajectory Similarity Computation." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, edited by A. Zhang and H. Rangwala, 2275–2285. New York, NY, USA: ACM. https://doi.org/10.1145/3534678.3539358.

Yao, D., C. Zhang, Z. Zhu, J. Huang, and J. Bi. 2017. "Trajectory Clustering via Deep Representation Learning." 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA. 3880–3887. https://doi.org/10.1109/IJCNN.2017.7966345.

Yao, Y., Z. Guo, C. Dou, M. Jia, Y. Hong, Q. Guan, and P. Luo. 2023. "Predicting Mobile Users' Next Location Using the Semantically Enriched Geo-Embedding Model and the Multilayer Attention Mechanism." *Computers, Environment and Urban Systems* 104:102009. https://doi.org/10.1016/j.compenvurbsys.2023.102009.

Yao, Y., X. Li, X. Liu, P. Liu, Z. Liang, J. Zhang, and K. Mai. 2017. "Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model." *International Journal of Geographical Information Science* 31 (4): 825–848. https://doi.org/10.1080/13658816.2016.1244608.

Yao, Y., Q. Zhu, Z. Guo, W. Huang, Y. Zhang, X. Yan, A. Dong, Z. Jiang, H. Liu, and Q. Guan. 2023. "Unsupervised Land-Use Change Detection Using Multi-Temporal POI Embedding." *International Journal of Geographical Information Science* 37 (11): 2392–2415. https://doi.org/10.1080/13658816.2023.2257262.

Yao, Z., Y. Fu, B. Liu, W. Hu, and H. Xiong. 2018. Representing Urban Functions through Zone Embedding with Human Mobility Patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. California, USA: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/545.

Yin, S., C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. 2024. "A Survey on Multimodal Large Language Models." *National Science Review* 11 (12). https://doi.org/10.1093/nsr/nwae403.

Yu, F., L. Cui, W. Guo, X. Lu, Q. Li, and H. Lu. 2020. "A category-aware deep model for successive POI recommendation on sparse check-in data." In *Proceedings of The Web Conference 2020*, edited by I. King, R. Socher and J. Tang. Taipei Taiwan: 1264–1274. New York, NY, USA: ACM.

Yu, W., and G. Wang. 2023. "Graph Based Embedding Learning of Trajectory Data for Transportation Mode Recognition by Fusing Sequence and Dependency Relations." *International Journal of Geographical Information Science* 37 (12): 2514–2537. https://doi.org/10.1080/13658816.2023.2268668.

Yuan, Z., W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, and X. Sun. 2022. "Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information." *IEEE Transactions on Geoscience & Remote Sensing* 60 (5620616): 1–16. https://doi.org/10.1109/TGRS.2022.3163706.

Zhai, W., X. Bai, Y. Shi, Y. Han, Z.-R. Peng, and C. Gu. 2019. "Beyond Word2Vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2Vec and POIs." *Computers, Environment and Urban Systems* 74:1–12. https://doi.org/10.1016/j.compenvurbsys.2018.11.008.

Zhang, F., A. Salazar-Miranda, F. Duarte, L. Vale, G. Hack, M. Chen, Y. Liu, M. Batty, and C. Ratti. 2024. "Urban Visual Intelligence: Studying Cities with Artificial Intelligence and Street-Level Imagery." *Annals of the American Association of Geographers* 114 (5): 876–897. https://doi.org/10.1080/24694452.2024.2313515.

Zhang, F., B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti. 2018. "Measuring Human Perceptions of a Large-Scale Urban Region Using Machine Learning." *Landscape and Urban Planning* 180:148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020.

Zhang, H., X. Zhang, Q. Jiang, B. Zheng, Z. Sun, W. Sun, and C. Wang. 2020. "Trajectory Similarity Learning with Auxiliary Supervision and Optimal Matching." In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, edited by C. Bessiere, 3209–3215. California, USA: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/444.

Zhang, J., and W. Ma. 2024. "Hybrid Structural Graph Attention Network for POI Recommendation." *Expert Systems With Applications* 248:123436. https://doi.org/10.1016/j.eswa.2024.123436.

Zhang, J., L. Mu, D. Zhang, Z. Chen, J. Rajbhandari-Thapa, J. A. Pagán, and Y. Li, G. Mai, Z. Zhou. 2025. "Space: A Spatial Counterfactual Explainable Deep Learning Model for Predicting Out-of-Hospital Cardiac Arrest Survival Outcome." *International Journal of Geographical Information Science*: 1–32. https://doi.org/10.1080/13658816.2024.2443757.

Zhang, M., T. Li, Y. Li, and P. Hui, 2020. "Multi-view joint graph representation learning for urban region embedding." In C. Bessiere (Ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, (pp. 4431–4437). International Joint Conferences on Artificial Intelligence Organization. Special track on AI for CompSust and Human well-being. https://doi.org/10.24963/ijcai.2020/611.

Zhang, R., J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao. 2024. "Llama-Adapter: Efficient Fine-Tuning of Large Language Models with Zero-Initialized Attention." In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=d4UiXAHN2W.

Zhang, Y., W. Huang, Y. Yao, S. Gao, L. Cui, and Z. Yan. 2024. "Urban Region Representation Learning with Human Trajectories: A Multi-View Approach Incorporating Transition, Spatial, and Temporal Perspectives." *GIScience and Remote Sensing* 61 (1): 2387392. https://doi.org/10.1080/15481603.2024.2387392.

Zhang, Y., W. Yu, and D. Zhu. 2024. "Next Track Point Prediction Using a Flexible Strategy of Subgraph Learning on Road Networks." *International Journal of Geographical Information Science* 38 (10): 1939–1964. https://doi.org/10.1080/13658816.2024.2358527.

Zhang, Z., H. Amiri, Z. Liu, L. Zhao, and A. Zuefle. 2024. "Large language models for spatial trajectory patterns mining." *In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection*, edited by Z. Liu, L. Liu, A. Züfle and H. Ning, 52–55. New York, NY, USA: ACM. https://doi.org/10.1145/3681765.3698467.

Zhang, Z., T. Zhao, Y. Guo, and J. Yin. 2024. "RS5M and GeoRSClip: A Large Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing." *IEEE Transactions on Geoscience & Remote Sensing* 62 (5642123): 1–23. https://doi.org/10.1109/TGRS.2024.3510781.

Zheng, Y., and X. Zhou. 2024. "Modeling Multi-Factor User Preferences Based on Transformer for Next Point of Interest Recommendation." *Expert Systems With Applications* 255:124894. https://doi.org/10.1016/j.eswa.2024.124894.

Zhou, B., A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. 2018. "Places: A 10 Million Image Database for Scene Recognition." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 40 (6): 1452–1464. https://doi.org/10.1109/tpami.2017.2723009.

Zhou, F., X. Yue, G. Trajcevski, T. Zhong, and K. Zhang. 2019. "Context-Aware Variational Trajectory Encoding and Human Mobility Inference." In The World Wide Web Conference (WWW '19), 3469–3475. New York, NY, USA: ACM. 3469–3475. https://doi.org/10.1145/3308558.3313608.

Zhou, S., J. Li, H. Wang, S. Shang, and P. Han. 2023. "GRLSTM: Trajectory Similarity Computation With Graph-Based Residual LSTM." In *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (4): 4972–4980. California, USA: AAAI Press. https://doi.org/10.1609/aaai.v37i4.25624.

Zhou, Y., and Y. Huang. 2018. "DeepMove: Learning Place Representations Through Large Scale Movement Data." In *2018 IEEE International Conference on Big Data (Big Data)*, edited by J.-Y. Nie, D.-N. Yang and V. Kantere, 24032412. Piscataway, NJ, USA: IEEE.

Zhou, Z., J. Zhang, Z. Guan, M. Hu, N. Lao, L. Mu, S. Li, and G. Mai. 2024. "Img2Loc: Revisiting Image Geolocalization Using Multi-Modality Foundation Models and Image-Based Retrieval-Augmented Generation." In *Proceedings of the 47th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, edited by G. Zuccon, G. H. Yang and Y. Fang, 2749–2754. New York, NY, USA: ACM. https://doi.org/10.1145/3626772.3657673.

Zhu, R., K. Janowicz, and G. Mai. 2019. "Making Direction a First-Class Citizen of Tobler's First Law of Geography." *Transactions in GIS* 23 (2): 398–416. https://doi.org/10.1002/tgis.12550.

Zhu, S., T. Yang, and C. Chen. 2021. "VIGOR: Cross-View Image Geo-Localization Beyond One-to-One Retrieval." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by D. Forsyth, G. Gkioxari, T. Tuytelaars, R. Yang, and J. Yu, 3640–3649. Piscataway, NJ, USA: IEEE.

Zhu, Y., Y. Ye, S. Zhang, X. Zhao, and J. Yu. 2023. "DiffTraj: Generating GPS Trajectory with Diffusion Probabilistic Model." In *Advances in Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Vol. 36, 65168–65188. Red Hook, NY, USA: Curran Associates, Inc.