Transportation Research Part C 178 (2025) 105252

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Activity-aware urban area embedding with contrastive learning for intelligent transportation systems applications

Gen Li [a], Tao Feng [a,*], Dan He [b], Li Yan [c], Jiwon Kim [b]

[a] Hiroshima University, Graduate School of Advanced Science and Engineering, Urban and Data Science Lab, Higashi-Hiroshima, Hiroshima, Japan

[b] The University of Queensland, School of Civil Engineering, Brisbane, Queensland, Australia

[c] University of Technology Sydney, Faculty of Engineering and Information Technology, Sydney, New South Wales, Australia

ARTICLE INFO

Keywords:
Contrastive learning
Representation learning
Urban area embedding
Activity-aware semantic
Traffic prediction

ABSTRACT

Embedding is a machine learning technique that represents data entities as continuous vector representations, capturing the underlying semantic relationships between them. Urban area embedding applies this concept to urban regions, representing each area as a vector that encapsulates its key characteristics. These embeddings enable models to better understand the relationships between different urban areas, facilitating applications such as traffic management, urban planning, and resource allocation. In this paper, we propose a comprehensive framework called **AUAEC** (Activity-aware Urban Area Embedding with Contrastive Learning) that integrates diverse open datasets including Location-Based Social Network (LBSN) check-ins, taxi flow data, and Points of Interest (POI) to produce enriched and context-aware region embeddings. To capture both mobility patterns and activity-aware semantics of LBSN users, we apply spatial interpolation based on road network, coupled with activity vector construction to represent user daily activity and movement patterns. To refine these embeddings into comprehensive urban regional representations, the AUAEC incorporates two complementary contrastive learning strategies: View-wise Contrastive Learning, which aligns representations across multiple data views, and Activity-aware Contrastive Learning, which captures inter-region relationships based on activity-aware semantics. The resulting embeddings are evaluated across four critical ITS tasks including land use distribution classification, traffic incident prediction, public transport delay prediction and traffic volume prediction using real-world data. Our approach demonstrates promising results, outperforming state-of-the-art solutions and highlighting the superiority of AUAEC in providing robust, contextual representations of urban areas for ITS and urban planning applications.

1. Introduction

As cities grow more complex, effective traffic management and Intelligent Transportation Systems (ITS) increasingly depend on an in-depth understanding of how urban areas function and interact. Urban area embedding offers a novel approach to meeting these challenges by integrating dynamic and functional interactions into city representations. Unlike traditional methods, such as static boundary definitions or administrative zones—embedding techniques provide a more flexible, data-driven means of characterizing

* Corresponding author.
E-mail addresses: d223296@hiroshima-u.ac.jp (G. Li), taofeng@hiroshima-u.ac.jp (T. Feng), d.he@uq.edu.au (D. He), li.yan-7@student.uts.edu.au (L. Yan), jiwon.kim@uq.edu.au (J. Kim).
URL: https://home.hiroshima-u.ac.jp/taofeng/ (T. Feng).

https://doi.org/10.1016/j.trc.2025.105252
Received 5 December 2024; Received in revised form 20 May 2025; Accepted 21 June 2025
Available online 6 July 2025

urban areas. Specifically, urban area embeddings transform rich spatial and mobility data, such as vehicle flows and Points of Interest (POIs), into compact vector representations that capture latent relationships and functional connections between regions. These embeddings not only reflect static characteristics but also dynamic patterns, such as commuting flows and recreational activity distributions, which are crucial for understanding the underlying mechanisms of urban mobility. This shift in perspective has several clear benefits for downstream ITS tasks. Through embedding, we gain a refined toolset for land use classification, enabling better identification of how different regions contribute to traffic demand. Furthermore, these embeddings enhance traffic pattern prediction, helping to anticipate congestion and optimize resource allocation. By representing dynamic relationships between areas, urban area embeddings empower adaptive traffic control, congestion mitigation strategies, and demand-responsive routing, all of which are vital for improving urban mobility. In short, the integration of dynamic, functional interactions into urban area representations provides a more comprehensive framework that not only bridges the gap between raw data and actionable insights, but also paves the way for more effective, intelligent transportation solutions.

The academic community has long sought effective methods to represent urban areas using geospatial data. Household travel surveys, which often rely on respondents' reporting of trips and activities, are both costly and susceptible to biases (Müggenburg, 2021). In recent years, methods have been developed to disaggregate spatial data and create spatial zones based on demographic, economic, and ethnographic similarities (Zhong et al., 2014). Early approaches utilized multivariate techniques, drawing from social area analysis, transport studies, and geographical analysis to capture urban area characteristics. With advancements in data collection technologies, sensors, GPS, and Volunteered Geographic Information (VGI), such as Open Street Map (OSM) data, provide detailed descriptions of urban areas. These data sources enable mapping of geographic features and human activities with unprecedented granularity. However, many basic approaches primarily focus on directly linking geographic features, traffic patterns (Zhang et al., 2022b), or Call Detail Records (CDR) (Sun et al., 2022) to their corresponding locations. For example, traffic congestion may be analyzed based solely on road networks within a specific area, or activity hotspots identified from CDRs may be limited to individual urban zones. Such methods often overlook the interdependencies and relationships between regions, such as commuting patterns, cross-regional functional complementarities, or shared infrastructure usage (Huang et al., 2022), which are critical for a holistic understanding of urban systems. Graph-based methods partially address this by representing urban areas as nodes and modeling their interactions through information propagation. Techniques such as Graph Partitioning (Zhong et al., 2014) and Markov Chains (Lin et al., 2017) have been applied to better capture these relationships. More recently, Graph Neural Networks (GNNs) have been adopted to model spatial dependencies and neighborhood interactions; however, their reliance on static graph structures limits their ability to represent dynamic inter-regional interactions and poses challenges in scalability due to computational demands. Additionally, urban areas often exhibit highly imbalanced distributions of activities, as seen in Zipf's law (Niu and Silva, 2021), certain areas show significantly higher activity levels, a characteristic also observed in natural language distributions. Inspired by parallels with Natural Language Processing (NLP), recent research has explored NLP-based techniques such as word embeddings and sequence modeling to capture the complex interdependencies within urban environments (Fan and Thakur, 2023; Yao et al., 2017; Zhai et al., 2019). By treating areas and mobility patterns analogously to words and sentences, these models effectively capture both spatial and temporal dependencies and contextual nuances in human mobility. However, integrating diverse and heterogeneous data sources into these models remains an open challenge, suggesting a critical gap in fully harnessing region embedding techniques for complex multi-dimensional urban data. To address this gap, contrastive learning has emerged as a powerful technique for aligning diverse data views while preserving critical interdependencies. By maximizing the agreement between similar samples and minimizing conflicts across different data sources, contrastive learning enables the creation of embeddings that retain rich semantic insights across multi-dimensional urban data. Recent studies (Chen et al., 2023b; Lan et al., 2024; Yan et al., 2024) have demonstrated the effectiveness of contrastive learning in enhancing downstream traffic tasks, such as trajectory prediction, multi-modal recommendation, and driver behavior modeling.

In this study, we propose a **Activity-aware Urban Area Embedding** framework to generate enriched regional embeddings by integrating multiple urban data sources. Our approach utilizes human check-in data, POI distributions, and taxi traffic flows, each offering distinct insights into urban dynamics. Human check-in data captures mobility patterns and provides clues about the types of activities prevalent in specific areas (Hong et al., 2023; Lu et al., 2024), while POI distributions reflect the functional landscape of areas. Taxi traffic flows meanwhile reveal movement intensity and connectivity between areas, essential for understanding inter-area relationships. By combining these data types, our framework is designed to capture both the functional roles of areas and the dynamics of human movement and activities, producing embeddings that are both comprehensive and contextually rich for urban ITS analysis.

While integrating multi-source urban data offers significant potential for creating meaningful urban area embeddings, it presents two primary challenges: (1). *Multi-View Fusion:* Combining diverse data sources such as human check-ins, POI distributions, and taxi flows is challenging due to their varying formats, granularities, and temporal patterns. Standard fusion methods like feature concatenation or averaging often fall short, as they treat each data source independently, missing the complementary insights that arise from interactions between sources. More advanced methods including attention mechanisms and graph-based fusion address some of these limitations but encounter scalability issues when handling large-scale urban data. Additionally, graph-based fusion is highly sensitive to graph construction quality, inaccuracies in graph structure can introduce biases, compromising the model's ability to capture authentic inter-area relationships. (2). *Effective Extraction of Human Activity Semantics:* Understanding urban area characteristics requires capturing the underlying human activity semantics, as these activities reveal the functional identities and activity patterns within each area. Privacy restrictions, however, limit direct access to detailed socio-demographic data, pushing researchers to rely on mobility data to approximate activity patterns. While sequence analysis techniques, including NLP-inspired models, can analyze these patterns, they often capture only superficial co-occurrences of activities rather than the socio-behavioral

motivations behind them. Moreover, mobility data is frequently sparse and unevenly distributed, leading to incomplete or biased representations of human activities. Although data augmentation techniques can help fill these gaps, integrating additional features or synthetic data brings challenges related to generalizability and the risk of overfitting.

To address the challenges in creating meaningful urban area embeddings, we propose AUAEC for aligning multi-view data, managing data sparsity, and capturing activity-aware semantics. (1) *Multi-View Data Alignment Using Contrastive Learning:* Urban data often comes from multiple sources (e.g., POI distributions, traffic flows, Location-Based Social Network (LBSN) check-ins) that differ in format, scale, and temporal patterns. Traditional fusion methods such as concatenation and averaging treat each data source independently, overlooking the complementary information that could be leveraged to capture nuanced regional characteristics. Without effective alignment, data from one source may overshadow or conflict with others, leading to incomplete or biased embeddings. Motivated by this, we use contrastive learning, a powerful machine learning approach that trains models by comparing pairs of data points to learn representations that maximize similarity for related pairs (positive pairs) and minimize similarity for unrelated pairs (negative pairs). This framework is particularly effective for aligning diverse data sources because it emphasizes the relationships within and between data views, ensuring that the learned representations are both discriminative and meaningful. Contrastive learning enables the integration of diverse data by identifying and aligning complementary insights from heterogeneous data views, including POI data, taxi traffic flows, and human mobility patterns, while simultaneously reducing conflicts or redundancies between them. By constructing positive and negative pairs, the model learns to distinguish genuine urban area representations from artificially constructed ones, optimizing the alignment of data views. This results in comprehensive embeddings that reflect the multifaceted nature of urban areas. The ability of contrastive learning to effectively align similar and dissimilar pairs allows us to model complex interdependencies across data views, ensuring that the resulting embeddings capture both shared and unique characteristics of urban areas.

(2) *Addressing Sparsity in Human Activity Data:* Urban data is often sparse and fragmented, both spatially and temporally. Temporal sparsity, in particular, arises when certain times of day, days of the week, or longer periods have few recorded check-ins, leading to incomplete representations of human activity patterns. For example, check-in data may heavily favor peak hours while leaving early mornings, late nights, or off-peak periods underrepresented. This uneven distribution limits the ability to understand how activities continuously flow across urban regions and reduces the effectiveness of embeddings meant to capture regional dynamics. To overcome spatial sparsity, we apply H3[1] spatial indexing based on road network architecture to generate interpolated travel links. This spatial interpolation enriches the check-in data by filling gaps between observed points, effectively creating a more continuous and comprehensive representation of movement across regions (Chen et al., 2024). To address temporal sparsity, we build multi-day activity vectors for individual users based on Location-Based Social Network (LBSN) data. Each user's weekday and weekend activity patterns are represented as separate vectors, with each dimension corresponding to a specific hour of the day. Clustering these user-level activity vectors allows us to identify recurring behavioral patterns, grouping individuals who exhibit similar activity habits. These clusters are then used to generate aggregate representations at the H3-cell level. Specifically, for each H3 cell, we count the number of users belonging to each cluster, effectively forming an activity-aware semantic profile for that area. This approach ensures that both spatial and temporal patterns are consistently captured and integrated into the embeddings. Additionally, using the H3-cell grid framework provides scalability and standardization, making the approach well-suited for large-scale urban datasets.

(3) *Semantic Enrichment of Urban area Embeddings with Activity-aware Differentiation:* Accurately representing urban area characteristics requires capturing the underlying motivations behind human activities. While areas may share similar POI distributions or mobility patterns, they can differ significantly in the activity patterns they support. Traditional data-driven methods often fall short in this regard, as privacy concerns limit access to detailed socio-demographic data, resulting in superficial interpretations of activity patterns and a lack of insights into their broader social and functional contexts. To address these limitations, we construct unified activity category vectors from human check-in data to capture diverse human activity patterns. These vectors are aggregated into an activity pattern distance matrix, which reflects activity-based similarities and differences across areas, thereby enriching our embeddings with meaningful semantic information. To further refine these embeddings based on activity pattern distinctions, we employ Triplet Contrastive Learning. This approach leverages anchor, positive, and negative samples to enforce differentiation in activity patterns between regions. As a result, our method effectively distinguishes urban areas with similar activity volumes but divergent multi-day activity patterns of individuals engaged in these activities. This differentiation highlights semantic distinctions rooted in the social, functional, and mobility contexts unique to each region. The motivation behind this approach is to enhance our embeddings by capturing not only quantitative data but also the qualitative, semantic nuances that define each urban area. By integrating these distinctions, we provide a more comprehensive and accurate representation of urban areas, enabling deeper insights into their unique characteristics.

In summary, our work has the following primary contributions:

- By applying H3-cell level interpolation for spatial sparsity and aggregating clustered LBSN data to manage temporal sparsity, we ensure that activity-aware urban area patterns are consistently represented, providing a continuous and scalable solution for data sparsity challenges.
- Our use of contrastive learning enables effective alignment across heterogeneous urban data views, capturing complementary insights from each source and producing embeddings that holistically represent area characteristics.

[1] https://h3geo.org/.

- Incorporating activity-aware pattern semantics through unified activity vectors and Triplet Contrastive Learning, we produce urban area embeddings that differentiate them based on social, functional, and mobility-driven characteristics, resulting in more meaningful and contextually aware representations of urban areas.
- We demonstrate the utility and robustness of our approach through four urban ITS analysis tasks: land use classification, public transport delay prediction, traffic incident prediction and traffic volume prediction, highlighting its versatility for diverse urban planning and traffic management applications.

The remainder of this paper is organized as follows: In Section 2, we review the **Related Work**, covering key studies on urban area embedding, and activity semantics extraction in urban data analysis. Section 3 provides the **Preliminaries and Problem Statement**, where we define the essential concepts and formally outline the problem we aim to solve. Section 4 details our **Methodology**, describing the proposed framework, its components, and the steps involved in addressing the challenges of multi-view fusion and activity-aware differentiation. In Section 5, we present the **Experiments** conducted to evaluate the effectiveness of our approach across multiple downstream tasks, followed by a discussion of the results. Finally, Section 6 concludes the paper, summarizing our contributions and suggesting directions for future research.

2. Related work

2.1. Urban area embedding for traffic applications

The effective representation of urban areas is a foundation for transportation applications including traffic prediction, congestion analysis, and transit optimization. In existing literature, static geographic features such as land use, buildings, and satellite imagery are typically used to represent areas. While remote sensing data provides detailed sources for extracting physical characteristics of the Earth's surface, the character of areas is often closely related to socio-economic factors (Fan and Thakur, 2023), which are difficult to extract from static data alone. With the growing popularity of location-based services, crowdsourced data such as social media, POIs, and geo-tagged images has shown promise for understanding urban activities due to its high granularity and temporal resolution. Many studies have treated urban areas as documents for topic extraction using NLP techniques, including LDA models (Yuan et al., 2012; Du et al., 2020; Gao et al., 2017), Doc2Vec (Niu and Silva, 2021), and TF-IDF (Liu et al., 2020a). Additionally, Word2vec (Mikolov et al., 2013), a classic word embedding algorithm, has been widely applied to urban area embedding (Fan and Thakur, 2023; Yao et al., 2017; Liu et al., 2020b; Hu et al., 2020), relying on the co-occurrence of POIs within areas to construct sentences for embedding.

Human mobility data, such as taxi flows and LBSN check-ins, has increasingly been used to reveal characteristics and model relationships between areas, providing a dynamic perspective that is critical for traffic management (Hu et al., 2021; Zhang et al., 2019; Zhai et al., 2019). Origin–Destination (OD) data is frequently used to depict the closeness between areas (Liu et al., 2024), while Hidden Markov Models (HMM) (Tu et al., 2017) and transition graphs (Wang and Li, 2017) constructed from OD data have been used to model transition probabilities between areas, capturing dynamic patterns like diurnal (day-night) patterns (Xia et al., 2019). Many methods still rely on word2vec-based approaches to embed trajectories, treating them as sequences. These sequences can include elements like POIs, road segments (Hu et al., 2021), and areas (Chen et al., 2020), providing rich context for understanding urban mobility (Crivellari and Beinat, 2019; Tian et al., 2022; Hu et al., 2020).

Recent advancements in urban area embedding have leveraged regional relationships to address the challenges of modeling spatial and temporal dynamics for transportation systems. Early works, such as HDGE (Wang and Li, 2017) and ZE-Mob (Yao et al., 2018a), used multi-graph learning and contextual enhancements to improve area embeddings. Wu et al. (2022) extended this idea by introducing Multi-Graph Fusion Networks(MGFN), which employ mobility graph distance (MGD) to cluster areas based on a comprehensive representation of mobility patterns.

To further exploit intra- and inter-area relationships, more sophisticated methods have emerged. One approach involves adversarial learning: models such as CGAL (Zhang et al., 2019) and MP-VN (Fu et al., 2019) introduced adversarial learning and VAE-based encoding to integrate intra- and inter-area relationships, ensuring stability and robustness in the resulting spatial embeddings. Another effective approach is contrastive learning, which originated from the broader machine learning field and has been notably advanced by models like UrbanCLIP (Huang et al., 2024). Contrastive learning has been proven to effectively align different views of data in embedding spaces. In urban area embedding, ReMVC (Zhang et al., 2023) employed contrastive learning combined with multi-graph fusion to enhance cross-view consistency and better capture complex relationships. Similarly, MetaRSTP (Chen et al., 2023a) applied adaptive meta-learning to mitigate negative transfer issues in spatio-temporal predictions, emphasizing the ongoing trend towards multi-view learning, semantic alignment, and contrastive techniques to generate robust and generalizable urban area embeddings. To the best of our knowledge, very limited work has explored how to comprehensively align multi-view urban data by fully leveraging activity-aware semantics and intra- and inter-area dynamics in a unified contrastive learning framework.

2.2. Semantic extraction of human mobility data

Current methods for defining the semantics of human mobility are diverse, often implicitly embedded in certain contexts or patterns within the data (Luca et al., 2021). These may include individual preferences, geographic connectivity, activity patterns, and vehicle trajectories. The semantic analysis of human activities involves analyzing data generated by these activities to understand and interpret human decision-making within specific contexts. Therefore, semantics play a crucial role in modeling and prediction

problems based on human activities. Some studies consider choice preferences as semantics. For example, Geo-Teaser (Zhao et al., 2017) examines preferences for specific POIs, preferences for combinations of POI categories (Yang et al., 2018), or activity habits. Behavior2vector (Liu et al., 2022b) constructs a k-partite graph to incorporate OD information, extracting user preferences for travel modes in different spatiotemporal contexts as semantics to predict travel mode choices. Sem-LSTM uses an LSTM model to extract user preferences (frequency and timing of visits to areas of interest) from trajectories to predict the next visit location. In POI recommendation systems, preference semantics are widely used, such as long-term and short-term preferences for POIs (Zhang et al., 2022a; Sun et al., 2020), and the trade-off between commuting distance and favorableness (Qin et al., 2023). Other studies consider travel purpose as semantics, for example, Feng and Timmermans (2015), DAGE-A (Liao et al., 2023), Trip2Vec (Chen et al., 2019), DeepMove (Feng et al., 2018) and Gstp2Vec (Liu et al., 2022a) analyze the purpose of trips to infer semantics. Additionally, some studies use the macroscopic characteristics of areas or stations as semantics. For instance, studies define urban functions as semantics to categorize city functions, while others analyze mobility patterns at subway stations (Zhuang et al., 2020) to classify them. The geographic environment's role in traffic has also been explored, with Liu et al. (2020a) approaches considering human cognition of subway stations as semantics for spatial analysis of the Shenzhen metro area. Recent advancements in mobility research have focused on understanding and leveraging activity patterns to promote sustainable transportation and improve mobility systems. These efforts include analyzing activity chains to classify individuals based on their embedded activity patterns through trajectory data (Cao et al., 2020), leveraging peer pressure via game theory to encourage environmentally-conscious travel behaviors (Feygin and Pozdnoukhov, 2018), utilizing spatio-temporal contexts and land-use functions with self-attentional neural networks for precise activity location prediction (Hong et al., 2023), and employing multi-view classification frameworks to robustly analyze and categorize complex traffic states across datasets (Sharma et al., 2024).

The process of semantic extraction and analysis can be divided into several key directions: NLP, temporal, spatial, and spatio-temporal joint methods. In the NLP dimension, some studies utilize natural NLP techniques to extract semantic information from user-generated text data, such as social media posts and reviews, for location recommendation and behavior prediction. Additionally, more advanced approaches involve converting travel plans into natural language descriptions for semantic extraction, such as using large language models (LLMs) with prompts to extract semantics from language descriptions. In the temporal dimension, time series data is leveraged to analyze and extract temporal patterns of human activities. Techniques such as ARIMA models, seasonal decomposition, RNNs, and LSTMs are used to analyze time series data, capture temporal dependencies and periodicity, and thereby understand and predict human behavior. In the spatial dimension, graph theory or grid-based methods are employed to understand spatial relationships. For instance, graph embedding techniques embed spatial POIs or areas into low-dimensional vector spaces to capture spatial relationships and connectivity. Techniques like GraphSAGE and GAT, which use graph neural networks, build graph structures between areas or POIs to capture spatial relationships and patterns. Additionally, recent advancements have introduced hybrid graphs and heterogeneous hypergraphs that integrate multiple modalities of geographic information, placing users, POIs, and other elements into a unified system. Grid-based spatial analysis techniques like ST-ResNet divide urban space into grids to analyze the activity patterns and relationships within each grid cell. In the spatio-temporal joint dimension, hierarchical grid division and spatio-temporal interpolation techniques allow point-to-point interaction between non-adjacent locations and non-consecutive check-ins to understand deeper associations between user behaviors and locations. Joint spatio-temporal feature modeling, for example, uses GAT combined with encoder–decoder structures to simultaneously capture spatio-temporal features. Additionally, attention-based recurrent networks employ attention mechanisms in recurrent neural networks to model the spatio-temporal state transitions of human mobility data, thereby extracting activity semantics. Also, some studies use HMM to model the spatio-temporal state transitions of mobility data, extracting and analyzing activity semantics. By employing these techniques and methods across various dimensions, researchers can more comprehensively and accurately extract and analyze the semantics of human activities, facilitating a wide range of practical applications such as location recommendation, behavior prediction.

The process of extracting semantics can easily lead to the loss of truly useful information due to several reasons: data sparsity, which makes it difficult to extract semantic information; the high dimensionality of human activity data such as multi-dimensional spatiotemporal data, which complicates dimensionality reduction or embedding; the presence of significant noise in real-world data, which interferes with semantic extraction; and the lack of sufficient contextual information, making it challenging for models to accurately understand and extract semantics. The current academic approaches to address these challenges primarily include multi-task learning, GAN, attention mechanisms, hierarchical embedding, and pre-trained models. Multi-task learning enhances the semantic retention capability of the main task by defining related auxiliary tasks, such as activity prediction, to preserve the semantic information associated with user activities and locations (Chen et al., 2023c). Adversarial training includes POI syntax tree encoding adversarial learning (Huang et al., 2022), which brings embeddings of the same category closer to preserve category semantics, and autoencoders combined with adversarial learning strategies to explore POI category information at all levels (Xu et al., 2023), emphasizing important POIs within areas and preserving useful information at the category syntax level. Attention mechanisms dynamically focus on the important parts of the input data to reduce information loss, while hierarchical embedding captures and preserves semantics at different hierarchical levels. Pre-trained models leverage the rich semantic information learned from vast amounts of data by using large-scale pre-trained models and fine-tuning them. By employing these techniques and strategies, researchers can better preserve semantic information during the extraction process, thereby improving the accuracy and effectiveness of applications.

2.3. Research gaps

The significance of the current work lies in its ability to address key research gaps in urban area representation and transportation analysis. Traditional methods often rely on static geographic features or simple POI co-occurrence patterns. Even advanced approaches, including graph-based methods and NLP-inspired embeddings, struggle to integrate diverse data sources and align them effectively. Additionally, existing techniques frequently overlook the role of activity-aware semantics and the nuanced interdependencies between areas, leaving gaps in their ability to model spatio-temporal dynamics comprehensively. This work directly tackles these challenges by proposing a unified framework that aligns multi-source urban data, including check-ins, POIs, and taxi flows, through contrastive learning. By doing so, it not only fills the void left by static, single-view approaches but also introduces a means to embed activity-aware semantics, improving both the interpretability and generalizability of the resulting representations. In doing so, the research sets a foundation for more effective urban ITS applications, such as land use classification, traffic delay prediction, and congestion mitigation, thereby addressing long-standing limitations and advancing the state-of-the-art in urban data analysis.

3. Preliminaries and problem statement

Definition 1 (*Urban Area*)**.** An area is defined as a geographic region represented by an H3-cell grid, which allows consistent and scalable analysis across different spatial scales. Each area, denoted as $h \in H$, serves as the basic unit for data aggregation and embedding in this framework.

Definition 2 (*LBSN Data*)**.** Building on the definition of urban areas, LBSN data captures user check-ins derived from the trip records of online users. These check-ins provide insights into the movement dynamics and behaviors of individuals in urban spaces, aggregated at the area level. The entire check-in dataset is denoted as M:

$$M = \{m_0, m_1, \dots, m_{|M|-1}\} \tag{1}$$

Each record m includes the following components:

$$m = (I_m, T_m, C_m, L_m), \quad \forall m \in M \tag{2}$$

where I_m is the user identity information, T_m denotes the timestamp at which the check-in occurred, C_m represents the category of check-in (POI category), and L_m is the location of the check-in. Together, these components capture the who, when, what, and where of human mobility within urban areas, providing a comprehensive view of mobility patterns.

Definition 3 (*Taxi Data*)**.** In addition to LBSN data, taxi data includes records of taxi OD trips, specifying the origin and destination locations as well as timestamps. This data characterizes the dynamic flow between areas, capturing inter-area relationships and traffic patterns over time. The entire taxi dataset is denoted as N:

$$N = \{n_0, n_1, \dots, n_{|N|-1}\} \tag{3}$$

Each trip record n contains the following information:

$$n = (T_n, O_n, D_n, LO_n, LD_n), \quad \forall n \in N \tag{4}$$

where T_n is the timestamp of the trip, O_n and D_n are the identifiers for the origin and destination areas, respectively, and $LO_n = (o_{n,x}, o_{n,y})$ and $LD_n = (d_{n,x}, d_{n,y})$ represent the coordinates of the origin and destination.

Definition 4 (*POI Data*)**.** Complementing the LBSN and taxi data, the POI dataset describes the static distribution of POIs within each area h, including different types of locations such as restaurants, shopping malls, and workplaces. In this study, the POI categories share the same set C_m as defined in the LBSN dataset. The entire POI dataset is denoted as P.

3.1. Problem statement

Urban areas are characterized by diverse types of human activities, infrastructure, and spatial interaction. To understand these dynamics, it is critical to represent different aspects of urban activity in a unified manner. The datasets described above capture unique information about the behavior and structure of urban areas, providing perspectives that together offer a holistic view of urban dynamics. For urban area H, given LBSN data M, taxi OD data N, and POI data P, the goal is to learn a distributed, low-dimensional embedding for each urban area h, denoted as $\mathcal{E}$. The embeddings can be formally defined as:

$$\mathcal{E} = \{\vec{e}_0, \vec{e}_1, \dots, \vec{e}_{|H|-1}\}, \quad \vec{e}_i \in \mathbb{R}^{D_e} \tag{5}$$

where $\vec{e}_i$ represents the embedding of the ith area, and D_e is the dimension of the embedding space. In this D_e-dimensional space, the embeddings capture and preserve key characteristics such as human mobility as well as static POI features, effectively representing the functional, social, and dynamic aspects of each area. These embeddings can then be used for various downstream tasks, such as predicting land use patterns, assessing traffic flow, or understanding socioeconomic factors in different areas.

3.2. Language models for urban sequential data processing

To solve sequence data embedding, we leverage Transformer (Vaswani, 2017), a language model to perform Masked Language Model (MLM) to learn trajectory embeddings of LBSN users, capturing spatial correlations in their check-in data. It utilizes a multi-head self-attention mechanism to calculate the importance of different parts of the input H3 sequence, expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V represent the Query, Key, and Value matrices respectively, derived from the input H3-cell sequence. The scaling factor $\sqrt{d_k}$ is used to stabilize gradients during training (Vaswani, 2017). The attention mechanism enables the model to focus on both local (adjacent H3-cells, capturing fine-grained spatial continuity) and global relationships check-in travel behaviors. Then fully connected feed-forward layers are applied to further refine the learned representations at each H3-cell. These layers consist of two linear projections with ReLU activations, operating on each H3-cell in the trajectory sequence independently:

$$\text{FFN}(x) = \text{ReLU}(xW_1)W_2 \tag{7}$$

where W_1 and W_2 are the learnable weight matrices. The fully connected feed-forward layers further refine the learned representations, and the final output layer is used to predict the masked tokens in the MLM task. The model aims to predict the masked H3-cell indices in the trajectory sequences. Cross-entropy loss function is often used on the masked positions while ignoring the padded ones.

On the other hand, the Long Short-Term Memory (LSTM) network is used to process taxi OD in-out flow data. It consists of memory cells that maintain internal states, governed by input, forget, and output gates. The key equations for the LSTM cell are:

$$i_t = \sigma(U_i x_t + V_i \tilde{h}t - 1 + b_i) \tag{8}$$

$$f_t = \sigma(U_f x_t + V_f \tilde{h}t - 1 + b_f) \tag{9}$$

$$o_t = \sigma(U_o x_t + V_o \tilde{h}t - 1 + b_o) \tag{10}$$

$$c_t = f_t \odot ct - 1 + i_t \odot \tanh(U_c x_t + V_c \tilde{h}_{t-1} + b_c) \tag{11}$$

$$\tilde{h}_t = o_t \odot \tanh(c_t) \tag{12}$$

In these equations, x_t is the input vector at time step t, and $\tilde{h}_{t-1}$ is the hidden state from the previous time step. i_t, f_t, and o_t are the input, forget, and output gates, while c_t is the cell state that maintains memory over time. U, V, and b are learnable parameters, σ is the sigmoid function, and tanh is the hyperbolic tangent function. The gates control information flow, allowing the LSTM to selectively retain or forget information to effectively capture long-term dependencies in the OD in-out flow data.

3.3. Contrastive learning

Contrastive learning is a method used to maximize mutual information between different data views. The goal is to align similar representations while distinguishing dissimilar ones. This approach is particularly useful for multi-view data, where different types of information (e.g., spatial, temporal, and functional features) need to be integrated into a unified embedding. Formally, contrastive learning aims to bring embeddings from similar instances closer together in the embedding space while pushing embeddings from different instances apart.

A common formulation of contrastive loss is based on the InfoNCE framework, which is defined as follows:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{13}$$

where z_i and z_j are embeddings of the same instance from different views, $\text{sim}(z_i, z_j)$ denotes the similarity metric, τ is a temperature parameter, and N represents the number of negative samples. This approach allows the model to learn discriminative features effectively across different views of the data, providing a unified representation that captures both shared and unique information. While the InfoNCE loss is a common choice in contrastive learning, the proposed framework employs two specialized approaches tailored to the characteristics of urban area embeddings: 1. View-wise Contrastive Learning which uses a binary cross-entropy loss to align embeddings across views, and 2. Activity-aware Contrastive Learning which uses a triplet loss to capture inter-area relationships based on activity semantics. These methods are specifically designed to address the multi-view and multi-area challenges in urban data analysis.

4. Methodology

4.1. Overview

As illustrated in Fig. 1, our proposed framework consists of three major components:

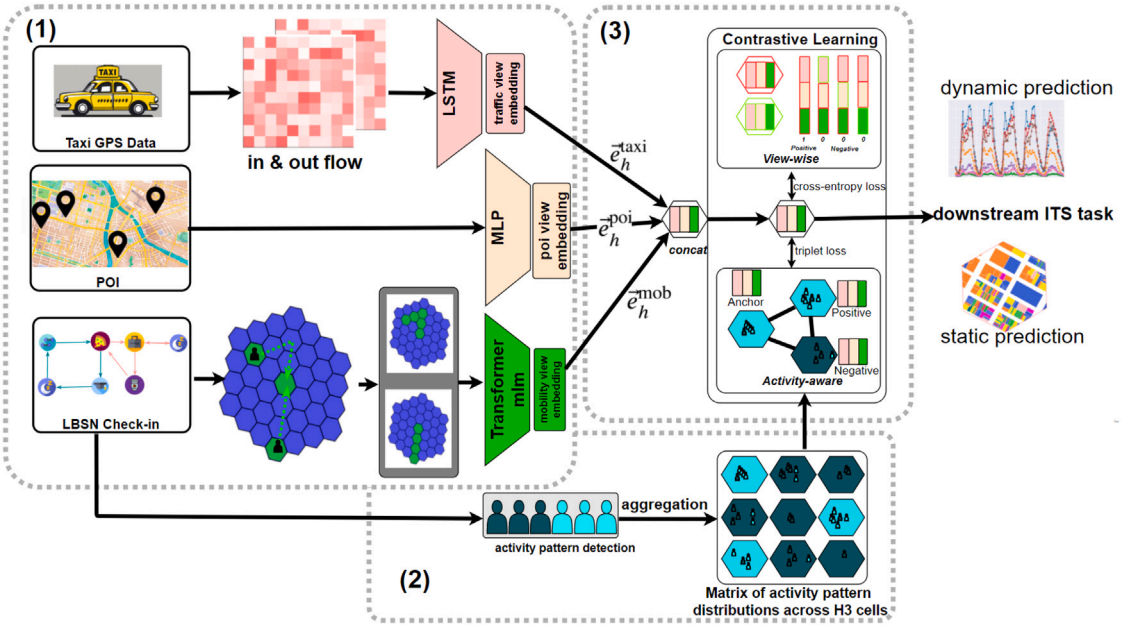

Fig. 1. Framework of our model.

1. **Initial Embedding Generation**: This stage captures three complementary views of urban regions (traffic dynamics, POI distributions, and mobility behavior), by leveraging taxi GPS data, POI data, and LBSN (Location-Based Social Network) check-in records.
2. **Activity Pattern Detection**: Based on LBSN check-in data, this module detects region-level activity patterns and aggregates them to form a distribution matrix that reflects the likelihood of each H3 cell belonging to different activity clusters.
3. **Contrastive Learning**: This component includes two submodules: *View-aware Contrastive Learning*, which aligns embeddings across different data modalities within the same region, and *Activity-aware Contrastive Learning*, which encourages embeddings of regions with similar activity profiles to be closer, using a triplet loss guided by the activity distribution matrix.

In detail, the embedding generation step (Component 1) processes taxi GPS data to compute hourly in-flow and out-flow statistics for each H3 cell. These are then encoded using an LSTM network to obtain the traffic view embedding ($\vec{e}_h^{\text{taxi}}$). Concurrently, POI data is encoded with an MLP to produce the POI view embedding ($\vec{e}_h^{\text{poi}}$). For mobility patterns, LBSN check-ins are interpolated along the road network to form daily trip chains. A Transformer-based Masked Language Model (MLM) is trained on these sequences to extract the mobility view embedding ($\vec{e}_h^{\text{mob}}$). In Component 2, the LBSN-derived trip chains are used to detect typical activity sequences. Each H3 cell is then assigned a probability distribution over these patterns, forming a matrix of activity pattern distributions across regions. Component 3 applies contrastive learning in two ways. First, *View-aware Contrastive Learning* uses a cross-entropy loss to encourage consistency across the traffic, POI, and mobility embeddings within the same H3 cell. Second, *Activity-aware Contrastive Learning* uses a triplet loss that brings together the embeddings of cells with similar activity distributions while pushing apart those with dissimilar patterns. The concatenated embeddings are ultimately used for downstream ITS tasks, including both dynamic prediction (e.g., traffic flow forecasting) and static analysis (e.g., land use classification).

4.2. Multi-view initial embedding

4.2.1. LBSN view embedding

LBSN check-in trajectories provide valuable insight into human mobility patterns, effectively linking various urban areas and illuminating the spatial relationships and interconnectedness that shape urban environments. These trajectories, which detail individual movements across geographic locations, are represented as sequences of visited H3-cells—referred to here as "mobility H3-cell chains". This representation draws a parallel to sequences of words in natural language processing (NLP), where each H3-cell is analogous to a word, and the trajectory forms a meaningful "sentence" of human movement. By organizing the data in this way, we can leverage advanced sequence modeling techniques traditionally applied in NLP. To enhance the quality of mobility sequences derived from LBSN check-in data, we interpolate the original check-in points along the road network to generate continuous H3-cell sequences (as illustrated in Fig. 2). This step refines the granularity of the input data by filling in gaps between observed locations, producing more complete and realistic representations of individual travel paths. These interpolated sequences are aligned to the H3 spatial grid and serve as the basis for constructing mobility patterns, as shown in the bottom-left region of Fig. 1 (Component 1).

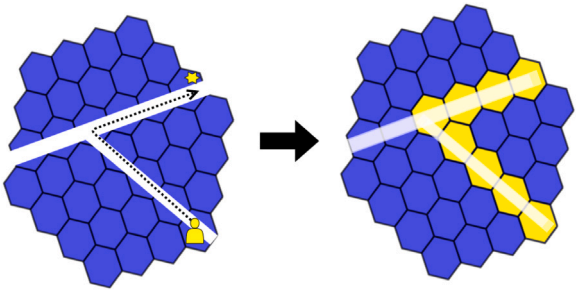

Fig. 2. Interpolation of H3-cells along the route between check-ins.

Next, as depicted in the green Transformer block in Fig. 1, a Transformer-based Masked Language Model (MLM) is employed to encode the enriched H3-cell sequences into mobility view embeddings ($\vec{e}_h^{\text{mob}}$). The MLM randomly masks a subset of H3-cells within each sequence and learns to predict their identities based on the surrounding context. This self-supervised learning strategy enables the model to capture the sequential and spatial dependencies inherent in human mobility patterns. The use of interpolated sequences provides the MLM with additional contextual cues, improving its capacity to model both common and subtle travel behaviors across the urban landscape. These embeddings serve as a core input to the model's contrastive learning stage (Fig. 1, Component 3), and contribute to downstream tasks such as activity pattern detection (Component 2), land use classification, and traffic flow prediction. By integrating spatial continuity and contextual representation, this module strengthens the model's ability to understand and differentiate region-level mobility dynamics. The objective function for the MLM model is defined as:

$$\mathcal{L}_{\text{MLM}} = -\sum_{j \in Q} \log P(h_j | h_{-j}), \tag{14}$$

where Q denotes the set of H3-cells that are randomly masked during training. The term h_j represents a specific masked H3-cell, while h_{-j} comprises the surrounding H3-cells in the same trajectory that serve as the context for predicting h_j. The primary goal of the MLM model is to learn an embedding space where each H3-cell's position and spatial relationships are well captured. By optimizing the objective function $\mathcal{L}_{\text{MLM}}$, the model effectively leverages human mobility patterns to encode the underlying spatial dependencies between different urban areas. To achieve this, a Transformer-based architecture is employed. Transformers are particularly suitable due to their self-attention mechanism, which enables the model to account for long-range dependencies and complex spatial–temporal interactions within the sequence of H3-cells. During training, certain H3-cells are masked, and the model learns to predict these masked cells using the surrounding context. This approach ensures that the learned embeddings not only capture local spatial features but also reflect broader patterns of human movement across the urban environment. Once the model is trained, the embedding of any H3-cell, denoted as h, can be obtained using the trained Transformer-based MLM. This process is described as follows:

$$\vec{e}_h^{\text{mob}} = \text{TransformerMLM}(h, h_{-j}), \tag{15}$$

where $\text{TransformerMLM}(h, h_{-j})$ denotes the Transformer-based masked language model applied to the target H3-cell h given its surrounding context h_{-j}. Consequently, the resulting embeddings effectively encode fine-grained urban mobility patterns and interactions across urban areas.

4.2.2. Taxi and POI view embedding

Taxi Origin–Destination (OD) data provides valuable insights into movement patterns between H3-cells by recording the origin and destination of each taxi trip. This information captures the intensity and directional characteristics of traffic flow across urban regions. As illustrated in Fig. 1, Component 1 (top-left section), we process the raw Taxi OD data to compute hourly in-flow and out-flow statistics for each H3 cell. For a given H3 region H_i, we aggregate the number of taxi trips entering (*inflow*) and leaving (*outflow*) during each hour of the day. This results in a 24-dimensional in-out flow vector, denoted as Flow_h, which captures the temporal dynamics of taxi traffic throughout a typical day. These hourly sequences are then fed into an LSTM encoder (top-center block of Fig. 1) to generate the traffic view embedding $\vec{e}_h^{\text{taxi}}$. This method preserves fine-grained temporal patterns, allowing the model to differentiate between peak and off-peak travel periods and understand daily variations in mobility demand. The structured nature of the in-out flow vector also facilitates efficient integration with other data views and downstream tasks, such as congestion detection or demand forecasting. By leveraging this temporally-resolved representation in combination with a sequential encoder, we capture both the volume and rhythm of urban traffic in a computationally effective manner. The resulting embeddings are passed to the contrastive learning module (Fig. 1, Component 3), where they are aligned with other view embeddings for joint representation learning.

To learn meaningful embeddings from these sequences, we employ an LSTM-based flow encoder. The model consists of two independent LSTMs: one dedicated to processing the pickup (outflow) sequences and another to handle the dropoff (inflow) sequences. Each LSTM extracts temporal patterns by processing its respective sequence and produces a final hidden state. These two hidden states, which capture the essential flow dynamics of the area, are then concatenated into a single vector. To project this

combined representation into a fixed-dimensional embedding space, a fully connected (FC) layer is applied. This procedure can be expressed as follows:

$$\vec{e}_h^{\text{taxi}} = \text{FC}(\text{Concat}(\text{LSTM}_{\text{pickup}}(\text{Flow}_h^{\text{pickup}}), \text{LSTM}_{\text{dropoff}}(\text{Flow}_h^{\text{dropoff}}))) \tag{16}$$

Here, $\text{Flow}_h^{\text{pickup}}$ and $\text{Flow}_h^{\text{dropoff}}$ represent the hourly sequences of taxi pickups (outflows) and dropoffs (inflows), respectively. The resulting embedding $\vec{e}_h^{\text{taxi}}$ captures both temporal dependencies and directional flow characteristics, offering a nuanced understanding of each area's taxi traffic patterns.

The POI data provides a complementary perspective by highlighting the variety and density of services, amenities, and activities available in different urban areas. This data is categorized into predefined types, and each area h is characterized by a vector of POI counts per category. Formally, the POI distribution for area h is represented as:

$$poi\_dist^{(h)} = \{\text{Count}_{c_1}^{(h)}, \text{Count}_{c_2}^{(h)}, \ldots, \text{Count}_{c_m}^{(h)}\} \tag{17}$$

where $c_i \in C_m$ denotes a specific POI category, and $\text{Count}_{c_i}^{(h)}$ is the number of POIs of category c_i in area h. This vectorized representation captures the distribution of different POI types, offering insight into the functional composition of urban areas—ranging from residential and commercial zones to cultural and recreational spots. To generate a low-dimensional embedding that encapsulates the key characteristics of the POI distribution, we employ a Multi-Layer Perceptron (MLP). The MLP processes the categorical count vector and projects it into a more compact embedding space. The resulting embedding is expressed as:

$$\vec{e}_h^{\text{poi}} = \text{MLP}(poi\_dist^{(h)}) \tag{18}$$

4.3. Activity-aware semantic classification

Human activities inherently carry semantic information that reflects the purpose and timing of interactions with various urban spaces. For instance, visiting a workplace at 10:00 AM and 1:00 PM, followed by returning to a residence at 5:00 PM, reveals a distinct daily routine. Such patterns, when aggregated across individuals, enable the inference of dominant activity characteristics within a region, such as a predominantly working population, even if the POI data does not explicitly identify the area as a workplace. By examining the relationships between POIs and the temporal sequences in which they are visited, the model encodes the functional roles that different regions play in urban life. This approach is conceptually similar to the way of how semantics are handled in language models. In language models, semantic meaning emerges from patterns of word co-occurrence and contextual usage. Words that frequently appear together or in comparable syntactic environments develop embeddings that reflect their semantic relationships. In urban activity analysis, POIs and their visit times serve a similar function to tokens in a sentence, forming sequences that provide context. By analyzing these sequences, the model identifies latent relationships and captures the higher-level meanings associated with various urban activities, such as work, commuting, and leisure. In both domains, the underlying methodology is to uncover latent semantic structure from sequential data. While language models rely on textual data to derive word embeddings, the proposed framework learns representations from temporal activity patterns. This approach enables the model to construct embeddings that encapsulate not only what activities occur in a region, but also when and why these activities happen, thereby providing a richer semantic understanding of human mobility.

To derive the semantic features of regional visits, visitors to a given area are first classified by their activity patterns. For each individual i, their weekday activity pattern vector is denoted as AP_i^{WD}, and their weekend activity pattern vector is denoted as AP_i^{WE}. Each individual's activity pattern is represented by two 24-dimensional vectors – one for weekdays and one for weekends – that encode the dominant POI category visited during each hour:

$$\text{AP}_i^{W} = \left[c_i^1, c_i^2, \ldots, c_i^{24}\right] \tag{19}$$

where W indicates whether the vector corresponds to a weekday or weekend, and c_i^t represents the most frequently visited POI category by individual i during the tth hour of the day. This representation is inspired by the method described in Cao et al. (2020), where daily check-in data for each user, grouped by POI category, is used to infer activity types. To group individuals with similar activity patterns, K-means clustering is applied separately to the weekday and weekend activity pattern vectors. This results in distinct weekday and weekend cluster assignments:

$$\begin{aligned} &\text{K-means}(\{\text{AP}^{\text{WD}}\}) \rightarrow \{C_1^{\text{WD}}, C_2^{\text{WD}}, \ldots, C_{S_{\text{WD}}}^{\text{WD}}\} \\ &\text{K-means}(\{\text{AP}^{\text{WE}}\}) \rightarrow \{C_1^{\text{WE}}, C_2^{\text{WE}}, \ldots, C_{S_{\text{WE}}}^{\text{WE}}\} \end{aligned} \tag{20}$$

where S_{WD} and S_{WE} are the number of weekday and weekend clusters, respectively. By combining these cluster assignments, each individual is assigned a "combined label" that indicates their weekday and weekend cluster membership:

$$\{(C_i^{\text{WD}}, C_j^{\text{WE}}) \mid i \in \{1, 2, \ldots, S_{\text{WD}}\}, j \in \{1, 2, \ldots, S_{\text{WE}}\}\} \tag{21}$$

This two-step clustering approach—first separately for weekdays and weekends, then combining labels—has several advantages. It distinguishes behavioral differences between weekday and weekend activity patterns, which are often characterized by different routines and visit distributions. It also makes the clustering process more interpretable, revealing distinct patterns that may otherwise be obscured if weekdays and weekends were clustered together. This method yields $S_{\text{WD}} \times S_{\text{WE}}$ unique combined labels that capture

a more detailed view of individuals' activity characteristics. For each H3-cell h, the number of visitors belonging to each combined label is aggregated to form an activity-aware semantic visit frequency vector:

$$\mathbf{visit}^{(h)} \in \mathbb{R}^{S_{\mathrm{WD}} \times S_{\mathrm{WE}}} = \left[\sum_{(i,j)} \mathrm{count}((C_i^{\mathrm{WD}}, C_j^{\mathrm{WE}}) \in h) \right] \tag{22}$$

where $\mathrm{count}((C_i^{\mathrm{WD}}, C_j^{\mathrm{WE}}) \in h)$ is the number of individuals in H3-cell h whose weekday activity pattern belongs to cluster C_i^{WD} and whose weekend activity pattern belongs to cluster C_j^{WE}. This vector encapsulates the activity pattern composition of the area, reflecting the semantic behavior of its visitors. Using these activity-aware semantic visit frequency vectors, a similarity matrix is constructed to compare the activity patterns between H3-cells. For two H3-cells h_i and h_j, the cosine similarity between their visit frequency vectors is defined as:

$$\mathrm{sim}(h_i, h_j) = \cos(\mathbf{visit}^{(h_i)}, \mathbf{visit}^{(h_j)}) = \frac{\mathbf{visit}^{(h_i)} \cdot \mathbf{visit}^{(h_j)}}{\|\mathbf{visit}^{(h_i)}\| \|\mathbf{visit}^{(h_j)}\|} \tag{23}$$

where $\mathbf{visit}^{(h_i)} \cdot \mathbf{visit}^{(h_j)}$ denotes the dot product of the vectors, and $\|\mathbf{visit}^{(h_i)}\|$ and $\|\mathbf{visit}^{(h_j)}\|$ are their Euclidean norms. The resulting similarity matrix $\mathbf{S}$ is of size $H \times H$, where H is the total number of H3-cells. Each entry in this matrix is given by:

$$\mathbf{S}_{ij} = \mathrm{sim}(h_i, h_j) \tag{24}$$

This matrix provides a pairwise similarity score for all H3-cells, allowing the identification of areas with similar activity patterns based on the distribution of their visitors' combined activity cluster labels.

4.3.1. Contrastive learning

Contrastive learning serves as a powerful tool in our proposed model to enhance the quality and expressiveness of area embeddings derived from different data sources. While the initial embeddings from POI, Taxi, and LBSN mobility data ($\vec{e}_h^{\mathrm{poi}}$, $\vec{e}_h^{\mathrm{taxi}}$, and $\vec{e}_h^{\mathrm{mob}}$, respectively) independently capture valuable urban characteristics, each view alone provides a limited and partial depiction of the complex urban environment. Therefore, to produce comprehensive and robust area representations suitable for downstream urban analytic tasks, it is crucial to effectively align and integrate these multiple data views. Contrastive learning is leveraged precisely for this integration task, by encouraging embeddings to capture shared information across data modalities and to reflect meaningful spatial–temporal relationships between urban areas.

As described in Section 4.2, the initial embeddings $\vec{e}_h^{\mathrm{mob}}$, $\vec{e}_h^{\mathrm{taxi}}$, and $\vec{e}_h^{\mathrm{poi}}$ provide individual representations of mobility, traffic, and POI data for each area. However, these embeddings need to be further refined to capture a unified, comprehensive representation of urban regions. To achieve this, our model leverages **View-wise Contrastive Learning** and **Activity-aware Contrastive Learning**, two complementary approaches that help integrate and align these embeddings while preserving essential inter-area relationships.

The goal of View-wise Contrastive Learning is to align the embeddings from different views (POI, Taxi, and Mobility) within the same area. This is accomplished by training the model to differentiate between "real" combinations of embeddings that belong to the same area and "fake" combinations generated by swapping out one of the views with an embedding from a different area. Specifically, for a given area, a positive sample is created by concatenating the corresponding POI, Taxi, and Mobility embeddings. Negative samples are generated by replacing one of these embeddings with a randomly selected embedding from a predefined pool of alternatives.

The training process involves assigning a probability score to each concatenated embedding, where a higher score indicates that the combination is genuine. Positive samples are assigned a label of 1, while negative samples are assigned a label of 0. The model optimizes a cross-entropy-based loss function:

$$\mathcal{L}_{\mathrm{view}} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \tag{25}$$

where y_i is the true label (0 or 1) for the ith sample, and $\hat{p}_i$ is the predicted probability. A multi-layer perceptron (MLP) with three layers is used to calculate these probabilities. By guiding the model to correctly identify and align the embeddings corresponding to the same area, this process enables each view's embedding to retain unique information while also integrating effectively with the other views, thereby producing more robust area-level representations.

While View-wise Contrastive Learning focuses on aligning embeddings within a single area, Activity-aware Contrastive Learning examines relationships between areas. It does so by leveraging activity pattern semantics derived from the similarity matrix $\mathbf{S}$ (Eq. (24)), which measures how closely two areas' activity patterns resemble each other. This component uses triplet loss, a technique that encourages embeddings of similar areas to be closer together in the embedding space while pushing embeddings of dissimilar areas apart.

A triplet consists of three components: an anchor embedding $\vec{e}_a$, a positive embedding $\vec{e}_p$, and a negative embedding $\vec{e}_n$. The anchor represents the target area, the positive is chosen from the $\theta\%$ most similar areas (as defined by $\mathbf{S}$), and the negative is chosen from the $\theta\%$ least similar areas. The triplet loss is defined as:

$$\mathcal{L}_{\mathrm{triplet}} = \max(0, \|\vec{e}_a - \vec{e}_p\|^2 - \|\vec{e}_a - \vec{e}_n\|^2 + \alpha) \tag{26}$$

where α is a margin that ensures the positive sample is sufficiently closer to the anchor than the negative sample. The embeddings $\vec{e}_a$, $\vec{e}_p$, and $\vec{e}_n$ are formed by concatenating the three view embeddings:

$$\vec{e} = \mathrm{concat}\left(\left[\vec{e}^{\mathrm{mob}}, \vec{e}^{\mathrm{taxi}}, \vec{e}^{\mathrm{poi}}\right]\right) \tag{27}$$

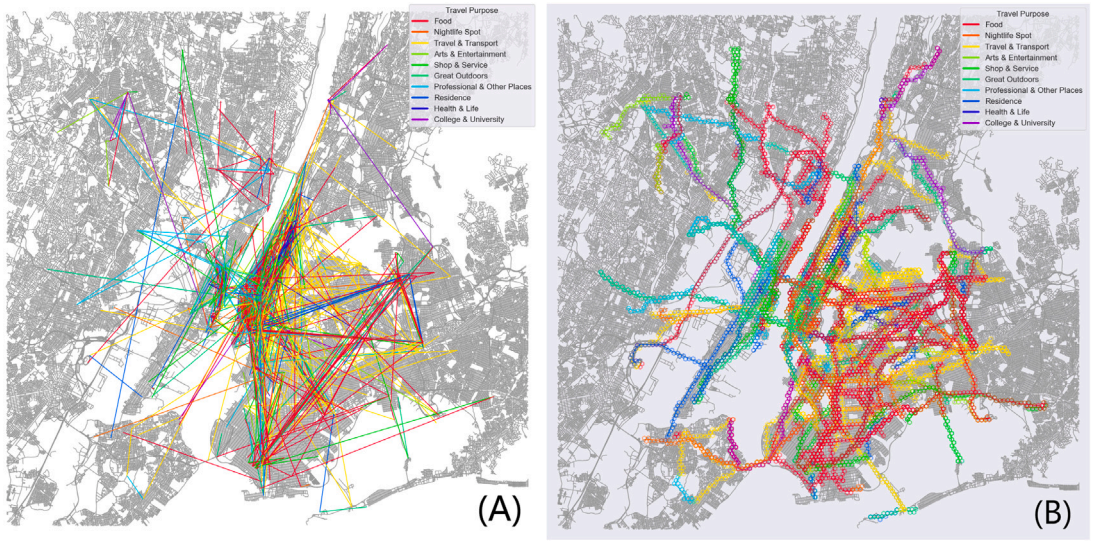

Fig. 3. (A) Check-in trajectories of 100 randomly selected LBSN users. (B) Interpolated trajectories for the same 100 users.

To train the model, the triplet loss is combined with an information loss term, resulting in the overall objective:

$$\mathcal{L} = \beta \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{view}} \tag{28}$$

where β is a regularization parameter that balances the two components. This joint objective ensures that the embeddings are not only well-aligned across views but also capture meaningful relationships between areas.

Our decision to pretrain the LBSN view embedding before contrastive learning stems from the nature of the data and the roles of the different learning objectives. LBSN data is characterized by complex temporal patterns and activity-based semantics that are not immediately apparent in raw form. Using MLM pretraining as an initial step allows the model to uncover these intrinsic patterns without interference from other modalities. This dedicated phase ensures that the LBSN embedding is well-structured and semantically rich, providing a strong foundation for subsequent alignment with POI and Taxi embeddings. By training the embeddings in stages, we maintain stable optimization dynamics and ensure that each view contributes its unique strengths to the unified embedding space. In addition, separating the training phases helps reduce noise and conflicts between objectives. MLM pretraining focuses exclusively on capturing the sequence-like nature of LBSN data, while contrastive learning emphasizes inter-view consistency and holistic area representations. Trying to optimize all these objectives simultaneously could result in suboptimal convergence or diluted signal from the more nuanced LBSN patterns. This multi-phase training strategy ensures that the final embeddings are not only robust and informative but also well-aligned and ready to support a wide range of urban analysis tasks.

After training, the final embedding for each area h is computed as:

$$\vec{e}_h = \text{contrastive\_learning}(\vec{e}_h^{\text{mob}}, \vec{e}_h^{\text{taxi}}, \vec{e}_h^{\text{poi}}) \tag{29}$$

and the complete set of embeddings across all areas is given by:

$$\mathcal{E} = \{\vec{e}_0, \vec{e}_1, \ldots, \vec{e}_{|H|-1}\} \tag{30}$$

This combined approach provides a unified, robust representation for each urban area, suitable for downstream applications and further analysis.

5. Experiments

To evaluate the effectiveness of our model in traffic applications, we conduct experiments on four downstream tasks: land use distribution classification, car accident prediction, bus delay prediction, and traffic volume counts prediction. These tasks are directly relevant to optimizing transportation systems, as they address challenges such as understanding urban traffic patterns, mitigating road accidents, improving the reliability of public transit services, and improving traffic flow management. The experiments are designed to comprehensively test our model's performance to capture spatial, temporal, and activity-aware semantic patterns in urban settings by comparing it against baseline models. For each task, relative datasets are preprocessed and aggregated on H3-cell level first to ensure consistency in spatial representation. Then downstream experimental setup follows a supervised learning framework where H3-cell level labels are used as targets to train the model. We evaluate the model performance using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics, followed (Huang et al., 2021).

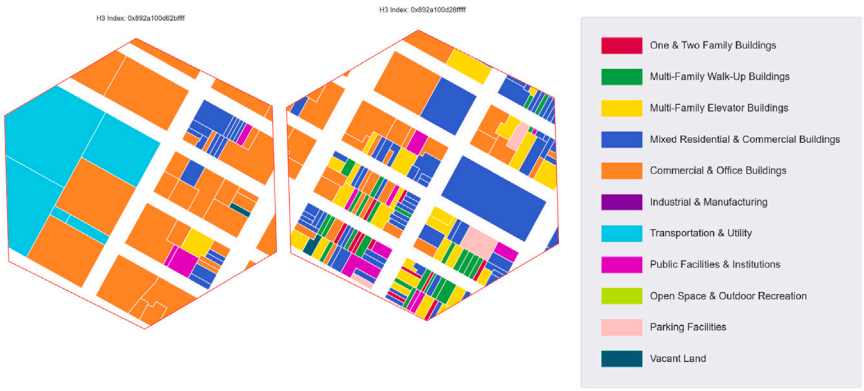

Fig. 4. Land Use Data of two H3 cells in resolution 9.

5.1. Data description and preprocessing

To prepare the input features for area embedding, we utilize several datasets relevant to urban analysis across New York City, focusing on the boroughs of Manhattan, Brooklyn, Bronx, Queens, and Staten Island. These datasets are chosen due to their open availability and richness in capturing diverse urban characteristics. The input data includes NYC taxi OD data,[2] OSM POIs,[3] and Foursquare check-ins.[4] (Yang et al., 2014) For evaluation, we process land use data from PLUTO,[5] traffic accident data from Motor Vehicle Collisions,[6] bus delay data from NYC MTA[7] and Automated Traffic Volume Counts[8] as H3-cell level labels to assess the effectiveness of the embeddings. All datasets are processed to the corresponding H3-cells at resolution 9, where each cell covers approximately 0.23 km^2 each, for example PLUTO land use data as shown in Fig. 4.For detailed descriptions of the feature dimensions and their corresponding views, please refer to Table 2.

Specifically, the NYC taxi dataset provides detailed information about taxi trips, including temporal and spatial data for both pickup and drop-off events. This information is mapped to H3-cells at an hourly resolution which allows to construct a 24-h H3-cell level heatmap. To get regional POI distribution, amenities within the study area are extracted and categorized into 10 groups to reduce sparsity. For LBSN data, user trajectories are processed by mapping check-in activities into 10 categories. Routes between check-ins are then reconstructed using the Google Maps API[9] to derive reasonable paths between check-in points. Subsequently, the interpolation method from H3 pandas[10] is used to identify the H3-cells along these routes, as shown in Fig. 3. This step assigns meaningful travel purposes to the H3-cells along the routes, which are then constructed as H3-cell trajectories. It is important to note that the LBSN check-in dataset includes data not only from NYC but also from some areas of New Jersey across the Hudson River. To ensure effective and complete mobility representation, these areas outside the primary research scope are initially retained but are subsequently removed after generating the embeddings through the MLM process. Similarly, the evaluation datasets are also mapped to the same H3-cell level. The PLUTO dataset contains 11 types of land use distributed across New York City. It is provided in geometric shapefile format, allowing us to calculate the area of each land use type within each H3-cell. For traffic accident data, each collision report includes the latitude and longitude of the incident location, enabling aggregation of incidents at the H3-cell level. For the MTA delay data, each record includes vehicle IDs, estimated arrival times, and scheduled arrival times, allowing for the calculation of delay durations. Each day is divided into four distinct time periods, and the cumulative delay in each H3-cell is sum over this four time periods, as shown in Fig. 6, representing long-term bus delay patterns. If a vehicle appears multiple times within the same H3-cell during one specific time period in one day, the average delay is calculated. For dynamic traffic prediction, we use automated traffic volume count data. A sliding time window of 13 h is employed, where the model predicts the sensor count for the next hour based on the data from the previous 12 h. If any data is missing within a 13-h window, the entire window is omitted from the analysis to ensure the model is trained and evaluated only on complete temporal sequences. Table 1 provides a summary of the characteristics of each dataset.

For transformer MLM model, trajectory data is split into overlapping segments with a window size of 50 and a stride of 25, result in 4758 segments. Each segment has 15% of its H3-cell indices randomly replaced with a mask token ([MASK]). The model

[2] https://www.nyc.gov/site/tlc/businesses/yellow-cab.page.
[3] https://www.openstreetmap.org/.
[4] https://sites.google.com/site/yangdingqi/home/foursquare-dataset.
[5] http://opendata.cityofnewyork.us/.
[6] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.
[7] https://www.kaggle.com/datasets/stoney71/new-york-city-transport-statistics.
[8] https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt/about_data.
[9] https://developers.google.com/maps.
[10] https://h3-pandas.readthedocs.io/en/latest/.

Table 1
Summary of datasets.

Dataset	Description	Usage
NYC taxi	Around 30 million taxi trip records during four months in 2012.	Taxi OD heatmap view embeddings
OSM	Road network and POI data from OpenStreetMap provide information on the street layout, road connectivity, and POIs.	POI view embeddings and road topology for trajectory interpolation
Foursquare	227428 Check-in data from Foursquare including timestamps and locations from April 2012 to 16 February 2013.	LBSN check-in trajectory mobility view embeddings
PLUTO	The 2012 NYC Primary Land Use Tax Lot Output provides detailed information on land use and administrative boundaries in New York City. This dataset includes 856826 data on land use categories, zoning, and building characteristics.	H3-cell label : land use distribution
Motor Vehicle Collisions	100545 motor vehicle crash events of New York from all police reports in 2012.	H3-cell label : traffic accident counts
NYC MTA	6730436 NYC MTA bus location data in approximately 10-min intervals, including route, bus stop, and scheduled arrival times to assess buses' on-time performance.	H3-cell label : bus delay
Automated Traffic Volume Counts	Traffic volume data collected by Automated Traffic Recorders (ATR) at bridge crossings and roadways in New York City for the years 2012 and 2013, containing a total of 5,264,239 records.	H3-cell label : traffic patterns

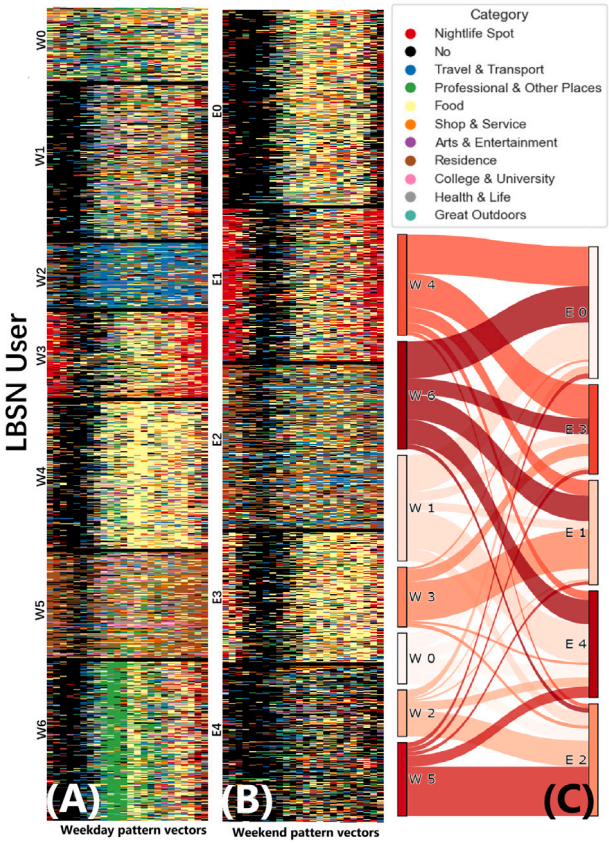

Fig. 5. Different group of LBSN and their activity pattern vectors in (A) weekday and (B) weekend. (C) shows the transfer of weekday and weekend patterns.

consists of 2 layers of transformer encoder with 4 heads in the multi-head attention mechanism, a hidden dimension of 128, and an embedding dimension of 64, as described in Table 3. Training uses the Cross-Entropy Loss function, ignoring padding tokens.

To derive the area activity-aware similarity matrix **S**, we first use the LBSN check-in dataset M to generate activity pattern vectors for both weekdays and weekends, as described in Eq. (19). These activity pattern vectors are visualized in Fig. 5 panels (A) and (B), representing weekday and weekend patterns respectively. The data is clustered into seven distinct groups for weekdays and five groups for weekends. From this visualization, we can clearly identify different activity patterns. For example, individuals in group *W6* tend to work throughout the day with a break for lunch, which identifies them as workers. In contrast, group *W3* exhibits a pronounced nightlife pattern. Fig. 5(C) illustrates the transition between weekday and weekend lifestyle styles. Notably,

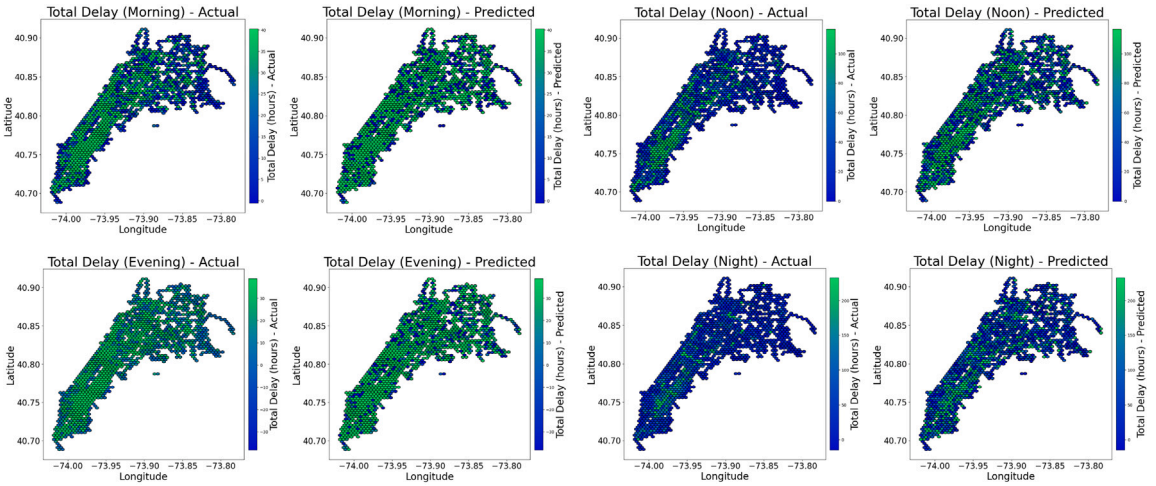

Fig. 6. Actual and predicted public traffic delays across different time periods (Morning, Noon, Evening, Night), showcasing the spatial and temporal variations captured by the model.

Table 2
Summary of feature views for H3-cells at resolution 9.

Feature	Shape	Description
Taxi in/out flow view	$2 \times 24 \times H$	Matrix representing hourly in/out flows between areas, capturing dynamic traffic patterns over 24 h.
POI view	$H \times 10$	Distribution vector across 10 POI categories for each area, indicating types of activities and services available.
LBSN check-in trajectory mobility view	$H \times 128$	Transformer-based MLM embedding derived from LBSN check-ins, capturing mobility patterns along interpolated trajectories.
Land use distribution	$H \times 11$	Distribution vector for each area across 11 land use categories, providing insights into area land usage patterns.
Traffic accident volume	$H \times 1$	Total count of traffic accidents recorded in each area.
Bus delay	$H \times 4$	Matrix representing bus delays for four time periods
Traffic volume	$H \times 12$	Time series of hourly traffic volume measurements for each H3, using 12 historical time points to predict the next time point.

individuals in group *W3*, those with a strong nightlife component, often maintain similar behaviors on both weekdays and weekends, transitioning to group *E1* on weekends. Meanwhile, a significant portion of individuals from group *W6* (identified as workers) tend to adopt a more nightlife-oriented behavior on weekends. Finally, using Eqs. (22) and (23), we derive the area activity pattern similarity matrix **S**. This matrix provides a quantitative representation of how similar different areas are based on the activity patterns observed in the data.

5.2. Model setup

The core of our model is the contrastive learning component, which begins with View-wise Contrastive Learning across three distinct feature views for each area: POI, Taxi, and Mobility data. By leveraging these views, the model learns embeddings that capture both the unique and shared characteristics of urban areas. Each positive sample is constructed by concatenating the true POI, Taxi, and Mobility embeddings for a specific area. Negative samples, on the other hand, are generated by replacing one of these three embeddings with an embedding from another area. Specifically, for each negative sample, one of the three embeddings (POI, Taxi, or Mobility) is substituted with an embedding randomly selected from the same batch in training. This approach ensures that the negative sample still contains a plausible embedding structure while making it different from the true representation of the original area. By contrasting these positive and negative samples, the model learns to distinguish genuine area embeddings from artificially constructed ones, ultimately enhancing the quality and robustness of the learned representations. We use Eq. (25) to train the model, with positive samples labeled as 1 and negative samples as 0. The contrastive learning is guided by a three-layer MLP, which outputs a probability score indicating the likelihood that the combined embedding belongs to a genuine area. This method promotes embeddings that align with real-world area patterns, capturing both unique and shared characteristics across views.

The second, Activity-aware Contrastive Learning captures inter-area relationships by leveraging the area activity pattern similarity matrix **S**. In this setup, each area is treated as an anchor, with positive and negative samples selected to reflect activity-based spatial distinctions. Specifically, the positive sample for each anchor area is chosen from the top 25% of areas most similar

Table 3
Parameter configuration.

Module	Parameter	Value
MLM	Window size	50
	Stride	25
	Mask rate	15%
	Transformer layers	2
	Transformer heads	4
	Transformer hidden dimension	128
	Embedding dimension	128
Contrastive learning	View-wise MLP layers	3
	View-wise MLP hidden dimension	128
	View-wise negative samples	5
	Activity-aware triplet loss margin α	0.3
	Weight of the view-wise loss in the overall objective β	0.3
Traffic view embedding	Input dimension	24
	Output dimension	32
	Pick up & drop off LSTM layers	2
	LSTM hidden dimension	64
POI view embedding	Input dimension	8
	Output dimension	32
	Hidden dimension	128
	Activation	ReLU
Mobility view embedding	Input dimension	24
	Output dimension	32
	Hidden dimension	128
	Activation	ReLU

to the anchor based on **S**, while the negative sample is randomly selected from the bottom 25% least similar areas. A triplet loss function (Eq. (26)) with margin $\alpha = 0.3$ is applied to ensure that the anchor embedding is closer to the positive than to the negative sample. This encourages the embeddings to capture activity-related spatial patterns by positioning areas with similar activity pattern closer together, while distinguishing those with contrasting activity patterns.

Finally, losses from view-wise and inter-area contrastive are combined according to Eq. (28) with $\beta = 0.3$. The training is 100 epochs and learning rate = 0.001. To evaluate the learned embeddings on a downstream task, we use a Random Forest Regressor model for land use distribution classification, car accident prediction, bus delay prediction, and LSTM for traffic volume counts prediction, with all results reported as the average of a 5-fold cross-validation. Specifically, the input for traffic volume counts prediction additionally includes 'day of the week', 'hour', and 'sensor ID', along with the traffic volume at each timestep. The parameter configuration for our model is shown in Table 3.

5.3. Baselines

We compare our model with several region embedding baselines from simple POI features to gaph representation methods, to recent model which leverage contrastive learning

- **POI (TF-IDF)**: A method that represents urban areas based on the frequency and importance of different types of POIs within each H3-cell.
- **DeepWalk** (Perozzi et al., 2014): A graph-based embedding method that learns node representations through random walks, typically used for prediction tasks. Here, we construct a Delaunay Triangulation Graph on H3-cells to represent spatial relationships between regions. DeepWalk is configured with a walk length of 30, 10 walks per node, an embedding dimension of 128, and a context window size of 5.
- **Node2Vec** (Grover and Leskovec, 2016): An extension of DeepWalk that uses biased random walks to better capture network structure. We use a return parameter of 1.0 and an in-out parameter of 2.0 to balance breadth-first and depth-first search. Other configurations are the same as those of DeepWalk.
- **GMEL** (Liu et al., 2020c): This model leverages geographic context to predict commuting flows by constructing a geo-adjacency network and using graph attention networks (GAT) to capture spatial correlations. Specifically, two separate GATs are used to encode supply and demand characteristics into embedding space. For this implementation, we use H3-cell level data with taxi OD instead of the origin data used in the original paper. The input features are based on the distribution of POIs across H3-cells. The embedding size of both GATs is 128, and the number of GAT layers is 2, as suggested by prior work.
- **ZE-Mob** (Yao et al., 2018b): A framework for urban area embedding that extracts mobility patterns from taxi trajectories and learns zone embeddings through the co-occurrence of origin–destination zones. It incorporates temporal and directional context into the embedding process by adding these factors as events to the word representation. Additionally, ZE-Mob adjusts co-occurrence weights using a gravity model to reduce the error between estimated and actual travel volumes through a heuristic optimization process. The embedding dimension is set to 128 in our implementation, while other settings remain the same as those in the original paper.

Table 4
Model performance comparison for four tasks.

Model	Land use distribution		Traffic accident		MTA delay		Traffic volume prediction	
	MAE	RMSE	RMSE	MAE	RMSE	MAE	RMSE	MAE
POI (TF-IDF)	0.035	0.080	16.131	13.200	89.030	52.691	122.802	22.781
DeepWalk	0.063	0.129	17.176	14.177	73.855	41.852	117.800	21.502
Node2Vec	0.080	0.183	17.342	15.214	67.276	39.588	117.662	21.087
GMEL	0.029	0.075	15.898	12.088	67.214	38.204	130.910	22.662
ZE-Mob	0.028	0.071	16.143	12.859	65.163	36.004	114.253	20.374
ReMVC	0.013	0.068	15.621	11.496	72.900	40.496	118.334	21.279
Ours	**0.011**	**0.057**	**15.510**	**11.200**	**62.781**	**34.224**	**108.862**	**19.717**

- **ReMVC** (Zhang et al., 2023): A state-of-the-art method in contrastive learning for area embedding, employs two key strategies: intra-view comparison, where regions within the same view are compared to extract effective representations, and inter-view comparison, where an area is compared with itself across different views to enable cross-view information sharing.

5.4. Model performance

The results in Table 4 demonstrate performance variations across models in the tasks of land use distribution classification, car accident prediction, MTA bus delay prediction, and traffic volume prediction. Overall, our model performs exceptionally well in all tasks, achieving the lowest RMSE and MAE values across the board.

Specifically, for land use distribution classification, using only POI data outperforms both DeepWalk and Node2Vec, with an RMSE of 0.080 and an MAE of 0.035. This indicates that unsupervised learning based solely on graph topology is not well-suited for land use classification. However, GMEL and ZE-Mob achieve better results, suggesting that even without explicit geographic features, incorporating mobility data can significantly improve prediction accuracy. Consequently, it is unsurprising that REMVC, which integrates both POI and mobility data, performs second best, achieving an RMSE of 0.068, less than half of the worst-performing model, and is only slightly behind our model among the evaluated models.

In traffic accident prediction, although the overall performance is similar to the first task, the differences between models are smaller, with Node2Vec once again performing the worst. Interestingly, TF-IDF on POI achieves a better RMSE compared to ZE-Mob, despite having a larger MAE. This indicates that while TF-IDF on POI effectively captures some of the variability, its predictions tend to exhibit larger individual errors compared to ZE-Mob. GMEL rises to third place, suggesting that area traffic flow regression through multi-task learning is beneficial in car accident prediction. The performance of REMVC is also very close to LEMC in this task, further demonstrating the potential of incorporating multi-sources data.

Although REMVC is the second-best model after our model in the first two tasks, it performs poorly compared to several other models in the MTA Bus Delay Prediction task. The use of POI data alone proves insufficient, as indicated by poor results, with an MAE of 52.691 and an RMSE of 89.030, almost 50% higher than our model. Conversely, ZE-Mob performs well, which can be attributed to its effective weighting of origin–destination zones, leading to better performance in predicting bus delays.

In traffic volume prediction, the results in Table 4 reveal that our model achieves the lowest RMSE and MAE among all evaluated approaches, demonstrating its strong capability to handle this dynamic prediction task. While most baseline methods, including POI-based TF-IDF and graph-based embeddings, show relatively high error metrics, our approach consistently outperforms them. Notably, the GMEL model, which performs well in other tasks, falls short here, highlighting the challenge of leveraging multi-graph fusion for highly dynamic and time-sensitive data. ZE-Mob, although better than some models like POI (TF-IDF) and DeepWalk, still shows higher error metrics in comparison to our approach. By contrast, our model ability to effectively incorporate and align diverse data sources ensures that it can capture the temporal and spatial variations of traffic flows, resulting in more accurate volume predictions.

5.5. Case study

To visually evaluate the model's performance, we conducted four case studies covering key ITS tasks: land use classification, motor vehicle collision prediction, MTA delay prediction, and automated traffic volume prediction.

In the MTA delay prediction case, we analyzed public transit delays in Manhattan and the Bronx (Fig. 6). The results, averaged across five cross-validation folds, demonstrated the model's ability to identify morning rush hour delays, midday traffic reduction, and widespread evening delays caused by increased traffic volume. At night, the model accurately captured occasional severe delays in the city's core, while also minimizing underestimation of minor delays. This pattern highlights the model's capability to understand and replicate delay trends over time.

Finally, in the automated traffic volume case (Fig. 7), we found that integrating our embeddings into the LSTM model consistently improved test MAE across individual sensors compared to models using only ID embeddings. This clearly illustrates the advantage of our approach in enhancing prediction accuracy for dynamic traffic volume data.

In the land use classification case (Fig. 8), we examined three H3 cells with sparse POI data and three cells with a typical amount of POI data. By comparing land use proportions inferred from POI data alone to those predicted through our embeddings, we observed that when POI data was absent, the model struggled to infer land use proportions accurately. However, for the three

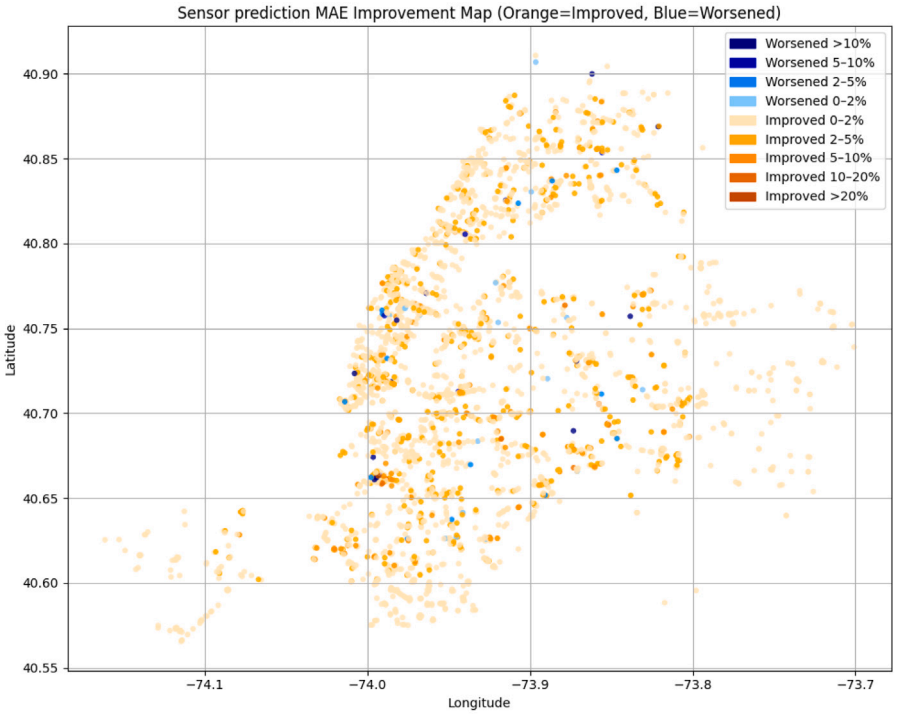

Fig. 7. Integrating our H3 embedding into the LSTM model led to improvements in the test MAE for traffic volume prediction for most of sensors.

POI-sparse cells, our model effectively inferred land use proportions even with zeroed input vectors. In cases with sufficient POI input, the model delivered more accurate predictions. For instance, the task of predicting 'residence' remained challenging, but our model's accuracy outperformed the POI-only baseline. One consistent observation was that both approaches tended to overestimate 'Vacant Land' proportions while underestimating 'Other' land uses.

In the motor vehicle collision case (Fig. 9), the visualizations revealed that our model successfully captured collision patterns, enhancing predictive accuracy and interpretability. The results showed a higher concentration of collisions in the city center, and a clear alignment between collision occurrences and road layouts in other areas—despite the fact that our model does not explicitly use road layout or traffic flow data. This demonstrates the model's ability to implicitly learn these patterns and reflect them in its predictions.

5.6. Ablation study

As our proposed approach comprises five major components, we perform an ablation study to verify the necessity of each component in the same three downstream tasks. The components evaluated include taxi OD view embedding, poi view embedding, LBSN mobility view embedding, view-wise contrastive learning, activity-aware contrastive learning. Each component was removed individually to analyze its impact on the model's performance, as measured by Mean Absolute Error MAE and RMSE. The resulting variants include:

- **wo-Taxi(woT)**: Model without taxi data embeddings. This variant assesses the importance of capturing the in-out flow of areas, representing the relationship of H3-cells in hourly flow.
- **wo-POI (woP)**: A model without POI view embedding. This illustrates how statistical POI features can enhance effective area embedding.
- **wo-LBSN (woL)**: A model without LBSN mobility embedding. This variant highlights the importance of capturing human activity sequence relationships along roads.
- **wo-View-wise Contrastive Learning (woV)**: A model without view-wise contrastive learning will lose alignment within the same region and fail to capture the comparative differences between different regions.
- **wo-Activity-aware Contrastive Learning (woE)**: A model without activity-aware contrastive learning will not capture the characteristics of H3-cell distributions associated with different activity pattern groups.

Table 5 presents the performance of each model variant. The results indicate that our model when deprived of different components experiences varying degrees of performance degradation across tasks. For land use distribution, the POI view is crucial,

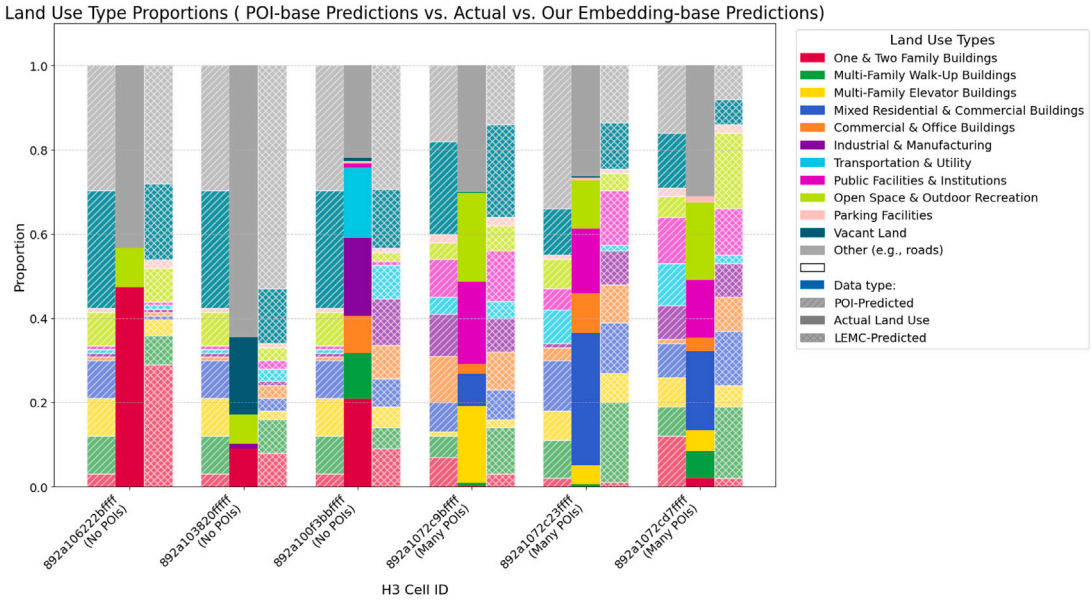

Fig. 8. Comparisons of different land use classification approaches across multiple H3 cells. Each cell displays three adjacent bar configurations: POI-based predictions, actual land use distributions, and our embedding-based model predictions.

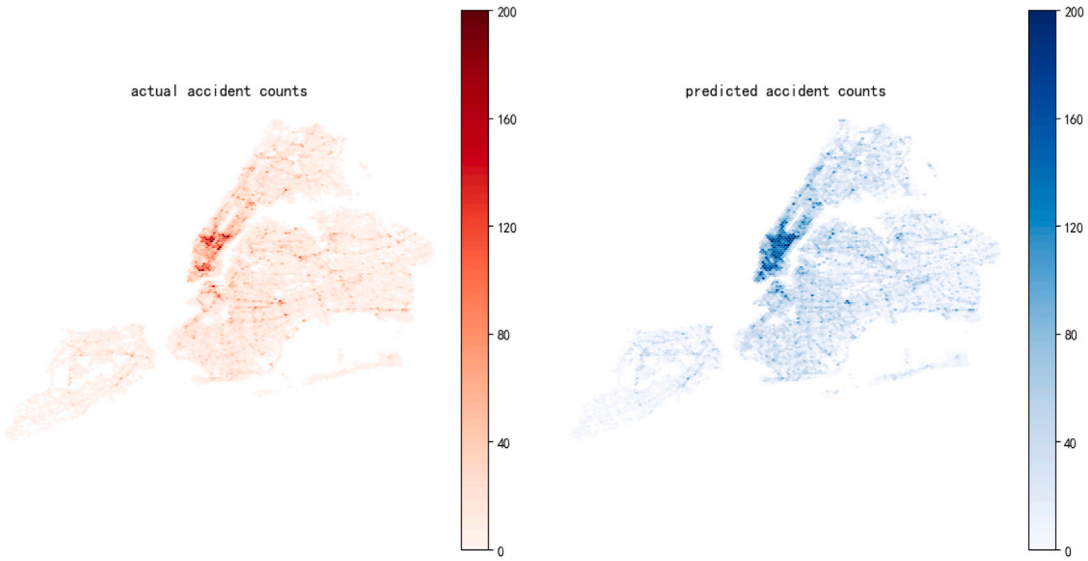

Fig. 9. Comparison of observed traffic accident frequencies with model predictions across H3 cells, illustrating the spatial accuracy of the accident prediction model in diverse urban contexts.

as evidenced by woP exhibiting the highest RMSE of 0.086 and MAE of 0.039. Similarly, the poor performance of woV underscores the importance of view-wise contrastive learning in aligning features within the same area and distinguishing differences across areas. In contrast, the performance of woT and woL degrades only slightly, suggesting that the taxi view and LBSN mobility embeddings are less critical for this specific task. For traffic accident prediction, woT and woV show the largest performance declines, indicating that taxi OD mobility information and feature alignment are critical for accurate predictions. Conversely, woE exhibits minimal performance degradation, suggesting that the activity-aware contrastive learning contributes less significantly to this task. For MTA delay forecasting, the removal of the taxi view (woT) and view-wise contrastive learning (woV) leads to the most significant performance increases in RMSE and MAE, underscoring the importance of these components in capturing temporal and spatial patterns associated with delays. Interestingly, woP performs nearly as well as the complete model, indicating that the POI view is less influential for delay prediction. In summary, the results reveal that the importance of individual components varies across tasks. While the POI view and view-wise contrastive learning are essential for land use distribution classification, taxi OD

Table 5
Ablation study results.

Model	Land use distribution		Traffic accident		MTA delay		Traffic volume prediction	
	MAE	RMSE	RMSE	MAE	RMSE	MAE	RMSE	MAE
woT	0.012	0.059	17.279	12.250	65.486	36.820	129.786	24.854
woP	0.039	0.086	15.828	11.608	62.900	34.421	110.200	20.397
woL	0.012	0.058	15.983	12.710	63.102	34.933	108.889	19.965
woV	0.029	0.073	17.342	16.234	65.270	36.572	111.767	20.780
woE	0.014	0.068	15.695	11.458	63.117	35.989	108.903	20.051
Ours	**0.011**	**0.057**	**15.510**	**11.200**	**62.781**	**34.224**	**108.862**	**19.717**

mobility information and feature alignment play dominant roles in traffic accident prediction and MTA bus delay forecasting. These findings emphasize the complementary nature of our model's components and their synergistic contributions to overall model performance. In the traffic prediction task, the results indicate that removing certain components from the model leads to only moderate performance degradation. Notably, excluding the LBSN mobility embeddings (woL) and the activity-aware contrastive learning (woE) results in a relatively small increase in RMSE and MAE, suggesting that these components have a lesser impact on dynamic traffic volume prediction. In contrast, the removal of the taxi view (woT) and the POI view (woP) leads to more substantial increases in error. This highlights the greater importance of these two components in capturing the patterns and dynamics required for accurate traffic volume prediction.

6. Conclusion

In this study, we introduced the Activity-aware Urban Area Embedding framework, a novel approach explicitly tailored to understanding urban transportation characteristics through the integration of three critical data views: taxi Origin–Destination (OD) flow, Points of Interest (POI) distributions, and Location-Based Social Network (LBSN) check-in trajectory mobility. These data sources were harmonized within an H3-cell level scalable spatial framework, effectively handling large-scale urban datasets and enabling detailed spatial granularity. The proposed method advances urban embedding research in several significant ways.

- Firstly, our utilization of H3-cell level interpolation and aggregation of long-term activity pattern vectors derived from LBSN data directly addresses the challenges of spatial and temporal sparsity inherent in urban mobility datasets. This methodological improvement enables richer, more accurate representations of urban mobility and long-term activity patterns, outperforming conventional methods reliant solely on static administrative boundaries or simplistic grid-based aggregations.
- Secondly, our adoption of contrastive learning represents a meaningful departure from traditional unsupervised learning methods commonly employed in the literature. Contrastive learning effectively integrates diverse urban mobility data views by explicitly distinguishing and aligning complementary insights from heterogeneous sources. This produces more holistic embeddings that simultaneously reflect transportation patterns, social interactions, and mobility dynamics. By actively distinguishing genuine urban area representations from artificially constructed negative samples, contrastive learning substantially enhances the robustness and discriminative capability of the generated embeddings compared to existing embedding solutions.
- Thirdly, the inclusion of activity-aware semantics, extracted from unified LBSN activities and refined through Triplet Contrastive Learning, generates contextually nuanced embeddings. These embeddings capture subtle distinctions among urban areas, identifying not only traffic functionality but also the underlying social activities and dynamic human mobility characteristics. This comprehensive representation addresses a notable limitation observed in prior literature, where urban embeddings often overlook social and semantic contextualization, resulting in less interpretable and less actionable insights.

The implications of the proposed Activity-aware Urban Area Embedding framework are broad and impactful. Compared to existing methods, which typically employ single-source data embeddings or simpler fusion techniques, our multi-view, contrastive-learning-driven approach delivers superior interpretability and practical utility. The integration of multiple complementary views yields a more comprehensive understanding of urban dynamics, greatly enhancing the accuracy and reliability of downstream ITS tasks such as land-use classification, traffic flow prediction, adaptive routing, and congestion mitigation. Moreover, AUAEC demonstrates a scalable and robust framework applicable to large-scale urban datasets, thereby offering city planners and transport authorities valuable tools for making informed decisions and implementing effective interventions. By leveraging advanced semantic models, future ITS solutions can provide more interpretable, context-aware insights that reflect real-world mobility behaviors and societal trends.

Future research could explore integrating multimodal approaches to address challenges in ITS by providing a deeper understanding of regional traffic behavior. By combining contextual information and human activity semantics, researchers can uncover critical drivers of traffic patterns, such as mobility preferences. Semantic understanding can be achieved through self-supervised learning applied to large-scale unstructured data, such as textual reviews or human trajectories, and by leveraging large language models (LLMs) with carefully crafted prompts to generate rational synthetic data. These models can enhance observational data by providing semantically meaningful representations that reflect real-world transportation contexts. For practical ITS applications, advanced frameworks inspired by models like Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) can integrate spatial–temporal traffic data with semantic insights, aligning multimodal representations in a shared embedding space. By leveraging

CLIP's ability to connect diverse data modalities such as textual semantics and visual-spatial traffic data, this approach enables the creation of unified embeddings that effectively capture complex multimodal relationships. Our framework, which generates enriched, context-aware urban region embeddings, has the potential to enhance emerging task such as LLM Retrieval-Augmented Generation (RAG) for urban tasks such as traffic management and urban planning. This methodology holds the potential to improve ITS capabilities, including traffic flow prediction, dynamic routing, and demand-responsive transportation planning, by delivering more interpretable and accurate models of traffic behavior.

CRediT authorship contribution statement

Gen Li: Writing – original draft, Methodology, Visualization, Investigation, Writing – review & editing, Resources, Conceptualization, Software, Data curation, Validation, Formal analysis. **Tao Feng:** Validation, Funding acquisition, Supervision, Formal analysis, Writing – review & editing, Methodology, Project administration, Conceptualization, Resources, Data curation. **Dan He:** Visualization, Validation, Writing – review & editing, Conceptualization, Methodology. **Li Yan:** Software, Visualization. **Jiwon Kim:** Validation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Chinese Scholarship Council (CSC) under Grant No. 202308050123.

Data availability

Data will be made available on request.

References

Cao, H., Xu, F., Sankaranarayanan, J., Li, Y., Samet, H., 2020. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. IEEE Trans. Mob. Comput. 19 (5), 1096–1108. http://dx.doi.org/10.1109/TMC.2019.2902403, URL: https://ieeexplore.ieee.org/document/8656580/.

Chen, K., Han, J., Feng, S., Zhu, M., Yang, H., 2023b. Region-aware hierarchical graph contrastive learning for ride-hailing driver profiling. Transp. Res. Part C: Emerg. Technol. 156, 104325.

Chen, W., Huang, C., Yu, Y., Jiang, Y., Dong, J., 2024. Trajectory-user linking via hierarchical spatio-temporal attention networks. ACM Trans. Knowl. Discov. Data 18 (4), 1–22.

Chen, C., Liao, C., Xie, X., Wang, Y., Zhao, J., 2019. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. Pers. Ubiquitous Comput. 23 (1), 53–66. http://dx.doi.org/10.1007/s00779-018-1175-9, URL: http://link.springer.com/10.1007/s00779-018-1175-9.

Chen, J., Liu, T., Li, R., 2023a. Region profile enhanced urban spatio-temporal prediction via adaptive meta-learning. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. ACM, Birmingham United Kingdom, pp. 224–233. http://dx.doi.org/10.1145/3583780.3615027, URL: https://dl.acm.org/doi/10.1145/3583780.3615027.

Chen, H., Wang, D., Liu, C., 2020. Towards semantic travel behavior prediction for private car users. In: 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems. HPCC/SmartCity/DSS, IEEE, Yanuca Island, Cuvu, Fiji, pp. 950–957. http://dx.doi.org/10.1109/HPCC-SmartCity-DSS50907.2020.00127, URL: https://ieeexplore.ieee.org/document/9408042/.

Chen, Y., Xie, N., Xu, H., Chen, X., Lee, D.-H., 2023c. A multi-context aware human mobility prediction model based on motif-preserving travel preference learning. IEEE Trans. Intell. Transp. Syst. 1–14. http://dx.doi.org/10.1109/TITS.2023.3314281, URL: https://ieeexplore.ieee.org/document/10268661/. GSCC: 0000000.

Crivellari, A., Beinat, E., 2019. From motion activity to geo-embeddings: Generating and exploring vector representations of locations, traces and visitors through large-scale mobility data. ISPRS Int. J. Geo-Inf. 8 (3), 134. http://dx.doi.org/10.3390/ijgi8030134, URL: https://www.mdpi.com/2220-9964/8/3/134.

Du, Z., Zhang, X., Li, W., Zhang, F., Liu, R., 2020. A multi-modal transportation data-driven approach to identify urban functional zones: An exploration based on Hangzhou City, China. Trans. GIS 24 (1), 123–141. http://dx.doi.org/10.1111/tgis.12591, URL: https://onlinelibrary.wiley.com/doi/10.1111/tgis.12591.

Fan, J., Thakur, G., 2023. Towards POI-based large-scale land use modeling: spatial scale, semantic granularity, and geographic context. Int. J. Digit. Earth 16 (1), 430–445. http://dx.doi.org/10.1080/17538947.2023.2174607, URL: https://www.tandfonline.com/doi/full/10.1080/17538947.2023.2174607.

Feng, J., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., Jin, D., 2018. DeepMove: Predicting human mobility with attentional recurrent networks. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18. ACM Press, Lyon, France, pp. 1459–1468. http://dx.doi.org/10.1145/3178876.3186058, URL: http://dl.acm.org/citation.cfm?doid=3178876.3186058. GSCC: 0000541.

Feng, T., Timmermans, H.J., 2015. Detecting activity type from GPS traces using spatial and temporal information. Eur. J. Transp. Infrastruct. Res. 15 (4), 662–674.

Feygin, S., Pozdnoukhov, A., 2018. Peer pressure enables actuation of mobility lifestyles. Transp. Res. Part C: Emerg. Technol. 87, 26–45.

Fu, Y., Wang, P., Du, J., Wu, L., Li, X., 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. Proc. AAAI Conf. Artif. Intell. 33 (01), 906–913. http://dx.doi.org/10.1609/aaai.v33i01.3301906, URL: https://ojs.aaai.org/index.php/AAAI/article/view/3879.

Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467. http://dx.doi.org/10.1111/tgis.12289, URL: https://onlinelibrary.wiley.com/doi/10.1111/tgis.12289.

Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 855–864.

Hong, Y., Zhang, Y., Schindler, K., Raubal, M., 2023. Context-aware multi-head self-attentional neural network model for next location prediction. Transp. Res. Part C: Emerg. Technol. 156, 104315.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., Gong, X., 2021. Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach. Comput. Environ. Urban Syst. 87, 101619. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101619, URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971521000260.

Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., Cui, H., 2020. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. Comput. Environ. Urban Syst. 80, 101442. http://dx.doi.org/10.1016/j.compenvurbsys.2019.101442, URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971519302625.

Huang, W., Cui, L., Chen, M., Zhang, D., Yao, Y., 2022. Estimating urban functional distributions with semantics preserved POI embedding. Int. J. Geogr. Inf. Sci. 36 (10), 1905–1930. http://dx.doi.org/10.1080/13658816.2022.2040510, URL: https://www.tandfonline.com/doi/full/10.1080/13658816.2022.2040510.

Huang, W., Wang, J., Cong, G., 2024. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. Int. J. Geogr. Inf. Sci. 1–29. http://dx.doi.org/10.1080/13658816.2024.2347322, URL: https://www.tandfonline.com/doi/full/10.1080/13658816.2024.2347322.

Huang, H., Yang, X., He, S., 2021. Multi-head spatio-temporal attention mechanism for urban anomaly event prediction. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (3), 1–21.

Lan, Z., Ren, Y., Yu, H., Liu, L., Li, Z., Wang, Y., Cui, Z., 2024. Hi-SCL: Fighting long-tailed challenges in trajectory prediction with hierarchical wave-semantic contrastive learning. Transp. Res. Part C: Emerg. Technol. 165, 104735.

Liao, C., Chen, C., Guo, S., Wang, L., Gu, F., Luo, J., Xu, K., 2023. Enriching large-scale trips with fine-grained travel purposes: A semi-supervised deep graph embedding framework. IEEE Trans. Intell. Transp. Syst. 24 (11), 13228–13239. http://dx.doi.org/10.1109/TITS.2022.3203464, URL: https://ieeexplore.ieee.org/document/9894369/.

Lin, X., Li, H., Zhang, Y., Gao, L., Zhao, L., Deng, M., 2017. A probabilistic embedding clustering method for urban structure detection. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-2/W7, 1263–1268. http://dx.doi.org/10.5194/isprs-archives-XLII-2-W7-1263-2017, URL: https://isprs-archives.copernicus.org/articles/XLII-2-W7/1263/2017/.

Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., Silva, C., 2020c. Learning geo-contextual embeddings for commuting flow prediction. Proc. AAAI Conf. Artif. Intell. 34 (01), 808–816. http://dx.doi.org/10.1609/aaai.v34i01.5425, URL: https://ojs.aaai.org/index.php/AAAI/article/view/5425.

Liu, K., Qiu, P., Gao, S., Lu, F., Jiang, J., Yin, L., 2020a. Investigating urban metro stations as cognitive places in cities using points of interest. Cities 97, 102561. http://dx.doi.org/10.1016/j.cities.2019.102561, URL: https://linkinghub.elsevier.com/retrieve/pii/S0264275119304809. GSCC: 0000039.

Liu, Y., Wu, F., Lyu, C., Liu, X., Liu, Z., 2022b. Behavior2vector: Embedding users' personalized travel behavior to vector. IEEE Trans. Intell. Transp. Syst. 23 (7), 8346–8355. http://dx.doi.org/10.1109/TITS.2021.3078229, URL: https://ieeexplore.ieee.org/document/9439960/.

Liu, X., Wu, M., Peng, B., Huang, Q., 2022a. Graph-based representation for identifying individual travel activities with spatiotemporal trajectories and POI data. Sci. Rep. 12 (1), 15769. http://dx.doi.org/10.1038/s41598-022-19441-9, URL: https://www.nature.com/articles/s41598-022-19441-9.

Liu, K., Yin, L., Lu, F., Mou, N., 2020b. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. Cities 99, 102610. http://dx.doi.org/10.1016/j.cities.2020.102610, URL: https://linkinghub.elsevier.com/retrieve/pii/S0264275119310637. GSCC: 0000068.

Liu, C., Zhang, H., Zhu, G., Guan, H., Kwong, S., 2024. Exploring trajectory embedding via spatial-temporal propagation for dynamic region representations. Inform. Sci. 668, 120516. http://dx.doi.org/10.1016/j.ins.2024.120516, URL: https://linkinghub.elsevier.com/retrieve/pii/S0020025524004298.

Lu, Q.-L., Qurashi, M., Antoniou, C., 2024. A two-stage stochastic programming approach for dynamic OD estimation using LBSN data. Transp. Res. Part C: Emerg. Technol. 158, 104460.

Luca, M., Barlacchi, G., Lepri, B., Pappalardo, L., 2021. A survey on deep learning for human mobility. URL: http://arxiv.org/abs/2012.02825. arXiv:2012.02825 [cs].

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. URL: http://arxiv.org/abs/1301.3781. arXiv:1301.3781 [cs].

Müggenburg, H., 2021. Beyond the limits of memory? The reliability of retrospective data in travel research. Transp. Res. Part A: Policy Pr. 145, 302–318.

Niu, H., Silva, E.A., 2021. Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. Comput. Environ. Urban Syst. 88, 101651. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101651, URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971521000582.

Perozzi, B., Al-Rfou, R., Skiena, S., 2014. Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 701–710.

Qin, Y., Wang, Y., Sun, F., Ju, W., Hou, X., Wang, Z., Cheng, J., Lei, J., Zhang, M., 2023. DisenPOI: Disentangling sequential and geographical influence for point-of-interest recommendation. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. pp. 508–516. http://dx.doi.org/10.1145/3539597.3570408, URL: http://arxiv.org/abs/2210.16591. GSCC: 0000013 arXiv:2210.16591 [cs].

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Sharma, S., Nayak, R., Bhaskar, A., 2024. Multi-view feature engineering for day-to-day joint clustering of multiple traffic datasets. Transp. Res. Part C: Emerg. Technol. 162, 104607.

Sun, Z., Peng, Z., Yu, Y., Jiao, H., 2022. Deep convolutional autoencoder for urban land use classification using mobile device data. Int. J. Geogr. Inf. Sci. 36 (11), 2138–2168. http://dx.doi.org/10.1080/13658816.2022.2105848, URL: https://www.tandfonline.com/doi/full/10.1080/13658816.2022.2105848.

Sun, K., Qian, T., Chen, T., Liang, Y., Nguyen, Q.V.H., Yin, H., 2020. Where to go next: Modeling long- and short-term user preferences for point-of-interest recommendation. Proc. AAAI Conf. Artif. Intell. 34 (01), 214–221. http://dx.doi.org/10.1609/aaai.v34i01.5353, URL: https://ojs.aaai.org/index.php/AAAI/article/view/5353. GSCC: 0000233.

Tian, C., Zhang, Y., Weng, Z., Gu, X., Chan, W.K.V., 2022. Learning large-scale location embedding from human mobility trajectories with graphs. In: 2022 International Joint Conference on Neural Networks. IJCNN, pp. 1–8. http://dx.doi.org/10.1109/IJCNN55064.2022.9892698, URL: http://arxiv.org/abs/2103.00483. arXiv:2103.00483 [cs].

Tu, W., Cao, J., Yue, Y., Shaw, S.-L., Zhou, M., Wang, Z., Chang, X., Xu, Y., Li, Q., 2017. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. Int. J. Geogr. Inf. Sci. 31 (12), 2331–2358. http://dx.doi.org/10.1080/13658816.2017.1356464, URL: https://www.tandfonline.com/doi/full/10.1080/13658816.2017.1356464.

Vaswani, A., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst..

Wang, H., Li, Z., 2017. Region representation learning via mobility flow. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, Singapore Singapore, pp. 237–246. http://dx.doi.org/10.1145/3132847.3133006, URL: https://dl.acm.org/doi/10.1145/3132847.3133006.

Wu, S., Yan, X., Fan, X., Pan, S., Zhu, S., Zheng, C., Cheng, M., Wang, C., 2022. Multi-graph fusion networks for urban region embedding. URL: http://arxiv.org/abs/2201.09760. arXiv:2201.09760 [cs].

Xia, T., Yu, Y., Xu, F., Sun, F., Guo, D., Jin, D., Li, Y., 2019. Understanding urban dynamics via state-sharing hidden Markov model. In: The World Wide Web Conference. ACM, San Francisco CA USA, pp. 3363–3369. http://dx.doi.org/10.1145/3308558.3313453, URL: https://dl.acm.org/doi/10.1145/3308558.3313453.

Xu, R., Huang, W., Zhao, J., Chen, M., Nie, L., 2023. A spatial and adversarial representation learning approach for land use classification with POIs. ACM Trans. Intell. Syst. Technol. 14 (6), 1–25. http://dx.doi.org/10.1145/3627824, URL: https://dl.acm.org/doi/10.1145/3627824.

Yan, H., Liao, Y., Ma, Z., Ma, X., 2024. Improving multi-modal transportation recommendation systems through contrastive De-biased heterogenous graph neural networks. Transp. Res. Part C: Emerg. Technol. 164, 104689.

Yang, D., Zhang, D., Zheng, V.W., Yu, Z., 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Trans. Syst. Man Cybern.: Syst. 45 (1), 129–142.

Yang, W., Zhao, Y., Zheng, B., Liu, G., Zheng, K., 2018. Modeling travel behavior similarity with trajectory embedding. In: Pei, J., Manolopoulos, Y., Sadiq, S., Li, J. (Eds.), Database Systems for Advanced Applications. In: Lecture Notes in Computer Science, vol. 10827, Springer International Publishing, Cham, pp. 630–646. http://dx.doi.org/10.1007/978-3-319-91452-7_41, URL: http://link.springer.com/10.1007/978-3-319-91452-7_41.

Yao, Z., Fu, Y., Liu, B., Hu, W., Xiong, H., 2018a. Representing urban functions through zone embedding with human mobility patterns. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, pp. 3919–3925. http://dx.doi.org/10.24963/ijcai.2018/545, URL: https://www.ijcai.org/proceedings/2018/545.

Yao, Z., Fu, Y., Liu, B., Hu, W., Xiong, H., 2018b. Representing urban functions through zone embedding with human mobility patterns. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI-18.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 31 (4), 825–848. http://dx.doi.org/10.1080/13658816.2016.1244608, URL: https://www.tandfonline.com/doi/full/10.1080/13658816.2016.1244608.

Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Beijing China, pp. 186–194. http://dx.doi.org/10.1145/2339530.2339561, URL: https://dl.acm.org/doi/10.1145/2339530.2339561.

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., Gu, C., 2019. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Comput. Environ. Urban Syst. 74, 1–12. http://dx.doi.org/10.1016/j.compenvurbsys.2018.11.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971518303582.

Zhang, Y., Fu, Y., Li, X., Wang, P., Zheng, Y., 2019. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, Anchorage AK USA, pp. 1700–1708. http://dx.doi.org/10.1145/3292500.3330972, URL: https://dl.acm.org/doi/10.1145/3292500.3330972.

Zhang, L., Long, C., Cong, G., 2023. Region embedding with intra and inter-view contrastive learning. IEEE Trans. Knowl. Data Eng. 35 (9), 9031–9036. http://dx.doi.org/10.1109/TKDE.2022.3220874, URL: https://ieeexplore.ieee.org/document/9973276/.

Zhang, L., Sun, Z., Wu, Z., Zhang, J., Ong, Y.S., Qu, X., 2022a. Next point-of-interest recommendation with inferring multi-step future preferences. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, pp. 3751–3757. http://dx.doi.org/10.24963/ijcai.2022/521, URL: https://www.ijcai.org/proceedings/2022/521. GSCC: 0000011.

Zhang, Y., Zheng, X., Helbich, M., Chen, N., Chen, Z., 2022b. City2vec: Urban knowledge discovery based on population mobile network. Sustain. Cities Soc. 85, 104000. http://dx.doi.org/10.1016/j.scs.2022.104000, URL: https://linkinghub.elsevier.com/retrieve/pii/S2210670722003201.

Zhao, S., Zhao, T., King, I., Lyu, M.R., 2017. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In: Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion. ACM Press, Perth, Australia, pp. 153–162. http://dx.doi.org/10.1145/3041021.3054138, URL: http://dl.acm.org/citation.cfm?doid=3041021.3054138.

Zhong, C., Arisona, S.M., Huang, X., Batty, M., Schmitt, G., 2014. Detecting the dynamics of urban structure through spatial network analysis. Int. J. Geogr. Inf. Sci. 28 (11), 2178–2199. http://dx.doi.org/10.1080/13658816.2014.914521, URL: http://www.tandfonline.com/doi/full/10.1080/13658816.2014.914521.

Zhuang, D., Hao, S., Lee, D.-H., Jin, J.G., 2020. From compound word to metropolitan station: Semantic similarity analysis using smart card data. Transp. Res. Part C: Emerg. Technol. 114, 322–337. http://dx.doi.org/10.1016/j.trc.2020.02.017, URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X18318175.