Article:
Chen, M., Jia, H., Li, Z. et al. (2026) Region Embedding With Adaptive Correlation Discovery for Predicting Urban Socioeconomic Indicators. IEEE Transactions on Knowledge and Data Engineering, 38 (2). pp. 1280-1291. ISSN: 1041-4347

https://doi.org/10.1109/tkde.2025.3631025

Region Embedding with Adaptive Correlation Discovery for Predicting Urban Socioeconomic Indicators

Meng Chen *Member, IEEE*, Hongwei Jia, Zechen Li, Weiming Huang, Kai Zhao, Yongshun Gong *Member, IEEE*, Haoran Xu, and Hongjun Dai

Abstract—A recent trend in urban computing involves utilizing multi-modal data for urban region embedding, which can be further expanded in a variety of downstream urban sensing tasks. Many previous studies rely on multi-graph embedding techniques and follow a two-stage paradigm: first building a k-nearest neighbor graph based on fixed region correlations for each view, and then blending multi-view information in a posterior stage to learn region representations. However, multi-graph construction and multi-graph representation learning are not associated in most existing two-stage studies, and the relationship between them is not leveraged, which can provide complementary information to each other. In this paper, we unify these two stages into one by constructing learnable weighted complete graphs of regions and propose a new one-stage Region Embedding method with Adaptive region correlation Discovery (READ). Specifically, READ comprises three modules, including a disentangled region feature learning module utilizing a city-context Transformer to encode regions' semantic and mobility features, and an adaptive weighted multi-graph construction module that builds multiple complete graphs with learnable weights based on disentangled features of regions. In addition, we propose a multi-graph representation learning module to yield effective region representations that integrate information from multiple graphs. We conduct thorough experiments on three downstream tasks to assess READ. Experimental results demonstrate that READ considerably outperforms state-of-the-art baseline methods in urban region embedding.

Index Terms—region embedding; human mobility; trajectory; POI data; urban profiling

1 Introduction

The proliferation of geo-tagged urban sensing data, including points-of-interest (POIs) and human trajectories, has ushered in new opportunities in the field of urban sensing [1], [3], [16], [18], [22], [29]. Among a variety of data sources, POIs inherently reflect human behavior and the socioeconomic perspectives of cities, while human trajectories offer direct insights into complex human mobility patterns and linkages between different regions in a city. This wealthy urban data provides insights for numerous urban planning and management tasks and thus leads to an increasing interest in data mining and urban computing fields. Particularly, a recently popular practice is to partition a city into numerous fine-grained regions and utilize multi-modal urban sensing data to learn the latent representations of these regions. The pre-trained low-dimensional vectors of regions provide valuable insights into regional configurations, structures, and interconnections, and can be used for various downstream tasks, such as land usage and socioeconomic prediction [10], [13], [19], [25].

In this context, multi-graph embedding methods have been widely used for tackling the problem of urban region embedding due to the representation capability for encoding multi-view information [15], [31], [33]. These approaches typically integrate both semantic features (reflecting static infrastructure and intended functions via POI distributions) and mobility features (capturing human movement patterns) to comprehensively characterize urban regions. The learning process involves two consecutive stages: static graph construction (usually using k-nearest neighbor; KNN) followed by multi-graph representation learning. Such a paradigm is demonstrated in Figure 1, where, it (1) separately computes region correlations using the Cosine similarity of hand-crafted raw features (e.g., semantic and mobility) and connects each region with its k-nearest neighbors (regions) to construct a static KNN graph for each view, and (2) leverages graph embedding methods (e.g., graph attention network [20]) to learn a single-view representation for each region and fuses multiple representations of a region to yield the final multi-view region representation. In this case, gradient back-propagation only occurs in the second stage to optimize region representations.

While such methods have been shown to be effective in certain analyses, they still grapple with major limitations inherent in the two-stage architecture.

- The optimization process, performed via gradient back-

- *Meng Chen, Zechen Li, Yongshun Gong, Haoran Xu, and Hongjun Dai are with the School of Software, Shandong University, Jinan, China. Email: mchen@sdu.edu.cn; lizc@mail.sdu.edu.cn; ysgong@sdu.edu.cn; hr_xu1990@sdu.edu.cn; dahogn@sdu.edu.cn.*
- *Zechen Li is also with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen, China.*
- *Hongwei Jia is with the Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Japan. Email: jia.hongwei.tkb_ct@u.tsukuba.ac.jp.*
- *Weiming Huang is with the School of Geography, University of Leeds, Leeds, UK. Email: W.Huang@leeds.ac.uk.*
- *Kai Zhao is with Walmart AI lab, California, USA. Email: kaizhaofrank@gmail.com.*
- *Corresponding author: Hongjun Dai.*

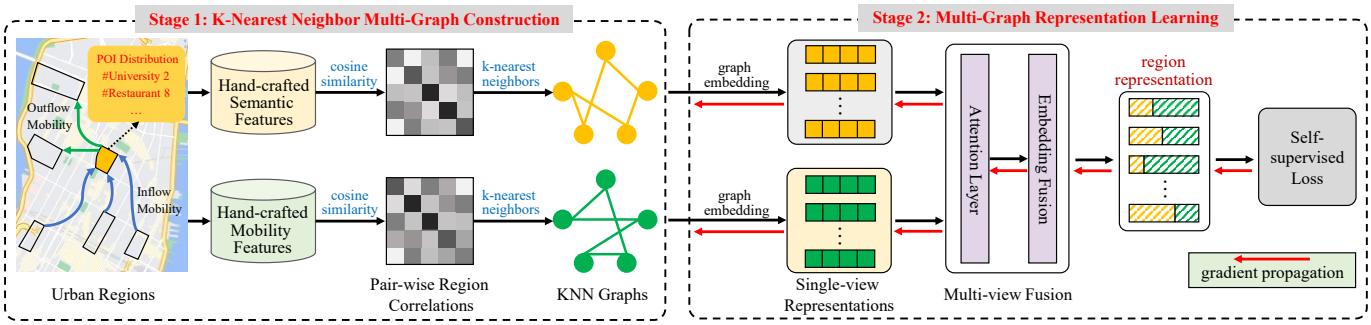


Fig. 1: Illustration of the existing multi-graph region embedding paradigm with two separate stages.

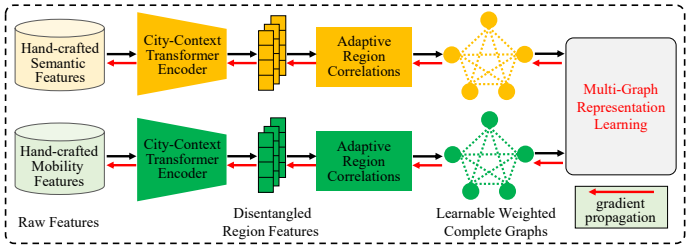


Fig. 2: Our end-to-end framework for region embedding.

propagation, is confined to the second stage, where latent region representations are generated based on a predefined KNN graph structure. This implies that the graph structure, including node initial features and link weights, remains fixed throughout the learning process. Consequently, the expressiveness of the learned region representations is limited, inducing such two-stage methods to fall short in capturing the nuanced and complex relations among regions. In this context, we argue that it is beneficial to adaptively discover the graph structure (i.e., inter-region correlation strengths) on-the-fly.

- The often-utilized raw region features, such as POI distributions, fall short in capturing the disparity among different regions. For example, two commercial areas, one in a downtown setting and the other in a suburban location, frequently exhibit significantly different POI distributions (particularly in terms of absolute POI counts). In this case, relying on raw features using absolute counts to measure pair-wise region correlations can be misleading and may cause their final representations to diverge substantially. Therefore, it is important to develop a feature encoder that accounts for urban disparity, capturing accurate region correlations that incorporate local contexts.

To overcome the above shortcomings, we present a new end-to-end Region Embedding model with Adaptive region correlation Discovery (READ), which embeds the adaptive graph construction with the multi-graph representation learning task in a streamlined and unified framework, where the gradient back-propagation can flow from the self-supervised signals all the way to the raw features, as illustrated in Figure 2. The connection is established by building a complete graph (with region correlations serving as edge weights) among all regions, where edge weights are measured using disentangled region features and adjusted during the training process, so as to discover optimal inter-region correlations through the self-supervision guidance.

We mainly develop two new components for adaptive graph construction: a city-context Transformer encoder that disentangles a region's unique features from the common features in its context, so as to emphasize its uniqueness while side-slide the commonalities shared by many regions; the second component, adaptive region correlation discovery, uses the generated disentangled features to calculate pairwise correlations between regions and form multiple complete graphs with these correlations serving as learnable edge weights. Finally, we utilize multi-graph embedding learning techniques to learn latent region representations from the constructed complete graphs across multiple views, employing both semantic and mobility reconstruction training objectives.

Our contributions are outlined as follows:

- We propose an end-to-end framework for generating urban region representations by exploring adaptive region correlation discovery using disentangled region features derived from human mobility and POI data. Unlike existing two-stage methods that rely on static KNN graphs and overlook the intricate relationships among regions in their latent representations, our approach adaptively learns inter-region correlation strengths from self-supervised signals to generate more expressive region representations.
- We develop a city-context Transformer encoder to encode raw region features, which disentangles the unique characteristics of regions from the common and less informative features considering the specific context of each region. This approach aids in accurately capturing pair-wise region correlations that reflect local contexts.
- We conduct extensive experiments to evaluate the proposed READ with real-world datasets. The results demonstrate that READ exhibits significant performance gains over baseline methods based on the paired t-test on three downstream tasks including land usage clustering, region popularity prediction, and region crime prediction. Data and source codes are available at https://github.com/AIMUrban/READ.

2 Related Work

Recent urban studies have explored learning region representations using diverse urban data sources, including POIs, human trajectories, satellite and street view images. These

studies can be broadly categorized into two groups based on their encoding mechanisms: MLP-based methods and graph learning methods.

2.1 MLP-Based Methods

MLP-based methods employ multi-layer perceptrons (MLPs) to encode raw region features, often enhanced by self-supervised contrastive learning for embedding refinement. A common approach involves leveraging POI data, where regions are treated as ”images” filled with POIs, and convolutional neural networks (CNNs) are used to learn latent region representations [14]. Other methods focus on urban dynamics, deriving region representations from human mobility data. For instance, Wang and Li [21] model temporal dynamics and multi-hop transitions between regions by constructing a flow graph and minimizing the Kullback-Leibler (KL) divergence between predicted and empirical transition probabilities. Similarly, Yao et al. [28] adopt a Word2vec skip-gram objective to learn region representations, defining human mobility events based on region, time, and movement mode.

Given the complementary nature of POIs and human trajectories, many studies integrate these two data sources within the contrastive learning paradigm. For example, Zhang et al. [30] propose a framework with intra-view and inter-view contrastive learning modules. The intra-view module enhances region distinctiveness, while the inter-view module aligns representations across different perspectives to ensure consistency. Building on this, Li et al. [11] introduce a dual-view contrastive learning method grounded in information theory. Instead of merely aligning representations across views, they maximize mutual information between views while employing a dual prediction strategy to minimize conditional entropy, effectively reducing cross-view inconsistencies.

Imagery data, such as satellite and street view images, capture urban environments from aerial and ground-level perspectives, offering rich visual context for region representation learning. These methods often combine imagery data with additional sources, such as spatial proximity or POIs, to enhance representation quality. For instance, Jean et al. [7] generate region representations by encoding satellite image patches using a CNN model. Similarly, Wang et al. [24] derive region embeddings by averaging street view image representations obtained from a CNN-based model, further enriching the embeddings with POI-derived text-based semantic context. With the emergence of large language models (LLMs), a growing research direction explores their potential for region representation learning. For example, Yan et al. [27] leverage satellite images and LLMs to generate detailed textual descriptions for images, employing contrastive learning and auto-regressive text generation to align visual and textual features in the embedding space. Extending this work, Hao et al. [6] incorporate both street view and satellite images, introducing CycleScore—a quality metric to filter LLM-generated text—ensuring higher-quality textual descriptions for more reliable region representation learning.

2.2 Graph Learning Methods

Leveraging the powerful capabilities of graph representations, recent studies have adopted graph learning techniques for region representation learning . For instance, Fu et al. [5] construct two POI graphs to model static and mobility connectivity patterns, incorporating geographic distances. These graphs are flattened and fed into an autoencoder, which learns region representations through graph reconstruction. Building on this, Zhang et al. [32] enhance the approach by replacing autoencoder reconstruction with a collective adversarial learning framework, improving representation quality. Further advancements integrate multiple data sources, such as human trajectories, POIs, and user check-ins, to learn region representations capturing multi-view information. For example, Zhang et al. [31] create four distinct graph views and employ graph attention networks (GATs) to encode each view. They facilitate cross-view information sharing through attention mechanisms and perform multi-view fusion using an adaptively weighted combination of the views. Similarly, Luo et al. [15] construct a multi-graph for regions based on inter-region human trajectories, spatial adjacency, and POI distributions, capturing diverse similarity measures. Additionally, recent methods aim to improve region embeddings by capturing both intra-view and cross-view correlations, as well as incorporating higher-order relationships in fused representations. For example, studies such as [4], [12], [17], [26] explore advanced techniques to model complex dependencies and enhance the quality of region representations.

Another line of research focuses on heterogeneous graph structures, which incorporate multiple types of urban data to derive region representations. For instance, Kim and Yoon [8] construct a heterogeneous information network with five node types and two edge types, and employ a heterogeneous graph attention network [23] to learn region representations. Zhou et al. [33] extend this approach by constructing heterogeneous graphs with edges derived from human mobility data, POIs, and geographic context. They first learn relation-specific region representations using attention-based graph learning and then combine them to generate comprehensive region representations.

We emphasize that the unique design of our proposed end-to-end method lies in its tailored disentangled region feature learning and adaptive region correlation discovery for urban regions. Unlike prior studies that rely on static region correlation graphs based on nearest neighbors derived from raw features, our approach adaptively discovers and leverages region correlations, providing a deeper understanding of regional dynamics and interactions, which is critical for downstream urban analysis tasks.

3 Problem Formulation

This research aims to learn region representations by leveraging the urban graph structure and its properties, thereby benefiting multiple downstream urban socioeconomic indicator prediction tasks. Urban region attributes, particularly those derived from Point-of-Interest (POI) data, reflect social characteristics. Residents' mobility drives regional interactions. Combining POI data with human mobility informa-

tion yields a multi-view dataset that provides rich insights into the urban regions. These are defined below.

Definition 1 (**Urban Region**). *Each region r_i is represented as an irregular shape defined by a set of boundary points $r_i = \langle b_1, b_2, \cdots \rangle$, with each boundary point b specified by its latitude and longitude coordinates.*

Definition 2 (**Geo-Tagged POI Data**). *A POI p_i consists of the latitude and longitude information as well as a category label (e.g., Restaurant, University).*

Definition 3 (**Human Mobility Data**). *Human mobility is the collection of individual trips in urban areas, represented as $\mathcal{M} = \{m_1, m_2, \cdots\}$. Each trip is defined as $m_i = \langle r_s, r_e \rangle$, where r_s and r_e are the starting and ending regions.*

Definition 4 (**Urban Region Embedding**). *Given a city consisting of n disjoint regions $\mathcal{R} = \{r_1, r_2, ..., r_n\}$, the objective of urban region embedding is to acquire a vector representation $\mathbf{E}_i \in \mathbb{R}^d$ for each r_i that captures the interrelations among regions based on geo-tagged region, POI and mobility data, where d is the embedding size.*

4 METHODOLOGY

4.1 Framework Overview

The framework of READ is shown in Figure 3. READ takes geo-tagged POI data, region data, and human mobility data as input, and adaptively encodes various region correlations based on semantic and mobility features to produce region representations through multi-graph representation learning. It starts with a newly designed city-context Transformer encoder to process the raw region features to generate the disentangled features. Then, an adaptive region correlation discovery module calculates pair-wise correlations between regions based on the disentangled features and builds multiple complete graphs with learnable edge weights. Finally, a multi-graph representation learning module is used to learn comprehensive region representations that integrate information from multiple graphs with semantic and mobility losses. The proposed READ allows the acquisition of region representations that incorporate adaptive accurate region correlation discovery and gradient back-propagation from self-supervised losses to raw features.

4.2 Disentangled Region Feature Learning

4.2.1 Region Raw Features

Using geo-tagged region, POI, and human mobility data, we initially build semantic and mobility features to depict correlations among regions from various perspectives. Our focus is on developing semantic features that reflect the functionality of each region and establishing mobility correlations between regions by considering their roles as starting and ending regions in trips.

Semantic features. With geo-tagged POI and region data, we associate each POI p_i with its respective region based on its geographic location. We then compute the semantic feature of each region based on the POI data. Formally, the semantic feature of r_i is denoted as $\mathbf{F}_i^{sem} \in \mathbb{R}^{N_c}$, where N_c represents the total number of POI categories and each dimension in $\mathbf{F}_i^{sem}$ corresponds to the number of POIs with a specific category in the region r_i. Such a semantic feature depicts the POI distribution in each region, thereby revealing the region's functionality.

Mobility features. With human mobility and geo-tagged region data, we construct two types of mobility features (the outflow feature and the inflow feature) for regions to capture the accessibility between regions. First, we compute the number of trips between r_i and r_j in the mobility dataset $\mathcal{M}$, $N_{r_j}^{r_i} = |\{(r_i, r_j) \in \mathcal{M}\}|$, where $|\cdot|$ counts the set size. Then, we compute the transition frequency distribution from r_i to all regions as the outflow feature of r_i, $\mathbf{F}_i^{out} = \{N_{r_j}^{r_i} / (\sum_{r \in \mathcal{R}} N_r^{r_i})\}_{j=1}^n$. Similarly, we calculate the transition frequency vector from all regions to r_i as the inflow feature, $\mathbf{F}_i^{in} = \{N_{r_i}^{r_j} / (\sum_{r \in \mathcal{R}} N_{r_i}^{r})\}_{j=1}^n$.

4.2.2 City-Context Transformer Encoder

Raw region features, such as POI distributions and mobility flows, often contain ubiquitous or trivial signals shared across many regions. While these features capture general urban trends, directly using them for modeling inter-region correlations may obscure the distinctive semantics of individual regions. To this end, we propose a city-context Transformer encoder that learns region-specific representations by separating each region's unique features from the globally shared patterns. Different from the vanilla Transformer that directly aggregates attention-weighted features from all regions, our encoder diminishes the influence of surrounding urban areas to reinforce the unique characteristics of each region within its specific context.

Intuitively, assuming there are multiple regions in a commercial functional area, the unique characteristics of each region can easily be buried into the overall commercial trend in its context. In this regard, we mitigate this problem by taking away the commonality within a geographic context, so as to reveal the uniqueness of each region. Specifically, our approach involves calculating the weighted average of all the regions' features to create city-context features for each region, and then computing the difference between the region's features and its corresponding city-context features as the disentangled features. We consider all regions in the city as the context of a region, as a region can impact both nearby and distant regions according to functionality and mobility.

Formally, we represent various raw features as $\mathbf{F}^{sem}$, $\mathbf{F}^{out}$, and $\mathbf{F}^{in}$, with $\mathbf{F}^v$ denoting the v-th kind of region features, and generate the new disentangled features as

$$\begin{aligned} \mathbf{Q} = \mathbf{F}^v \mathbf{W}_Q^v, \mathbf{K} = \mathbf{F}^v \mathbf{W}_K^v, \mathbf{V} = \mathbf{F}^v \mathbf{W}_V^v, \\ \boldsymbol{\beta}^v = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}} / \sqrt{D}), \\ \Delta \mathbf{F}^v = \mathbf{V} - \boldsymbol{\beta}^v \mathbf{V}, \\ \widetilde{\mathbf{F}}^v = \mathrm{FC}\left(\mathrm{LN}\left(\Delta \mathbf{F}^v\right)\right) + \Delta \mathbf{F}^v, \end{aligned} \tag{1}$$

where $\mathbf{W}_Q^v$, $\mathbf{W}_K^v$, and $\mathbf{W}_V^v$ are projection matrices, D represents the dimension of the key matrix of $\mathbf{W}_K^v$, β^v is the attention weight, $\mathrm{LN}(\cdot)$ is a layer norm operation, and $\mathrm{FC}(\cdot)$ is a fully connected network. Here, $\boldsymbol{\beta}^v \mathbf{V}$ denotes the aggregated city context for each region. By subtracting this from the original feature $\mathbf{V}$, we obtain the residual feature $\Delta \mathbf{F}^v$, which serves as the disentangled representation. This operation allows the model to focus on region-level differences rather than broadly shared patterns. It is important to

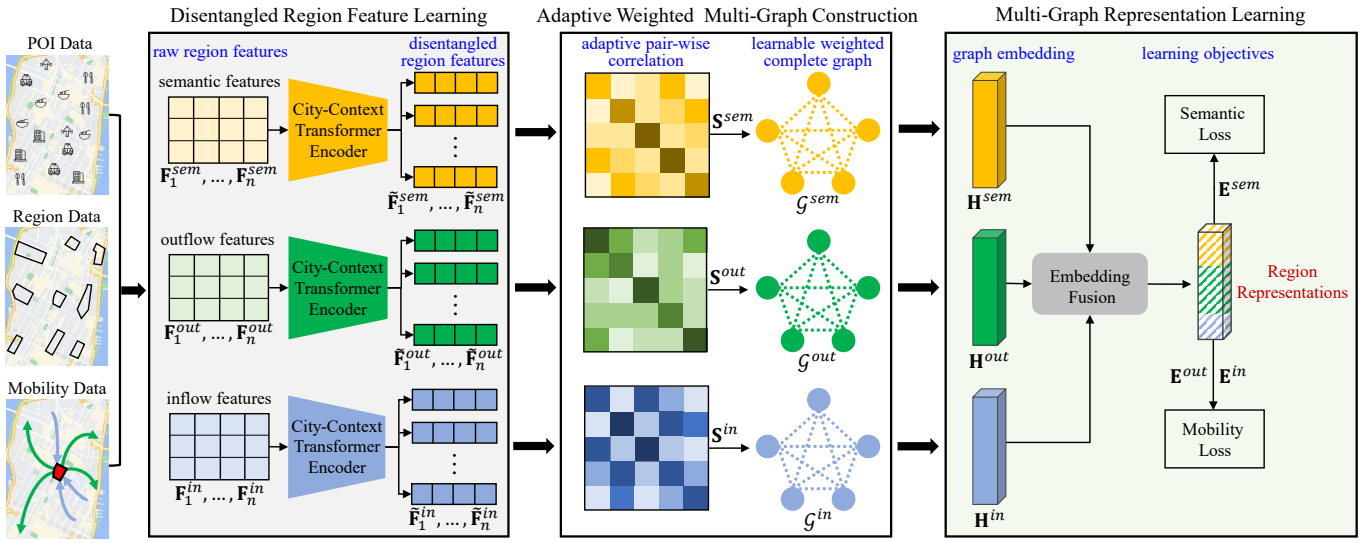


Fig. 3: Overview of the READ framework. The framework consists of three key components: (1) a disentangled region feature learning module, where a city-context Transformer encoder isolates region-specific features by removing shared contextual patterns; (2) an adaptive weighted multi-graph construction module, which leverages these disentangled features to dynamically compute region correlations and construct complete graphs with learnable edge weights; and (3) a multi-graph representation learning module that integrates multi-view signals via GCN and cross-view attention, jointly optimized under semantic and mobility supervision.

note that β^v adjusts the weights of all regions to each region adaptively. In contrast to the vanilla Transformer, which directly sets $\Delta\mathbf{F}^v = \beta^v\mathbf{V}$ as the updated representation, our subtraction-based mechanism $\Delta\mathbf{F}^v = \mathbf{V} - \beta^v\mathbf{V}$ explicitly filters out shared context and improves the isolation of discriminative information. Additionally, we utilize a multi-head attention mechanism to enhance performance.

4.3 Adaptive Region Correlation Discovery

Current methods assess region correlations through the Cosine similarity of manually designed features and create a static KNN graph for each view, which does not effectively capture the complex relationships among regions. We propose two methods to compute pairwise correlations using disentangled features and construct a complete graph for each view, adaptively updating correlation strengths during region representation learning.

4.3.1 Pair-wise Region Correlations

Cosine correlation. Using the disentangled features $\widetilde{\mathbf{F}}^v$, we calculate the correlation $\mathbf{S}^v_{ij}$ between r_i and r_j as

$$\mathbf{S}^v_{ij} = \max(\text{Sim}(\widetilde{\mathbf{F}}^v_i, \widetilde{\mathbf{F}}^v_j), 0), \tag{2}$$

where $\text{Sim}(\cdot)$ is the function for calculating the cosine similarity between $\widetilde{\mathbf{F}}^v_i$ and $\widetilde{\mathbf{F}}^v_j$. We assume a positive correlation between any two regions and therefore use the max operation to compute $\mathbf{S}^v_{ij}$.

Attention-based correlation. Using the disentangled features $\widetilde{\mathbf{F}}^v$, we employ an attention network to adaptively combine the two features of r_i and r_j to compute the pairwise correlation.

$$\begin{aligned} \alpha_m &= \mathbf{c} \cdot \tanh(\mathbf{A} \cdot \widetilde{\mathbf{F}}^v_m + \mathbf{b}), m \in \{i, j\} \\ \mathbf{S}^v_{ij} &= \frac{\exp(\alpha_i)}{\sum_{m\in\{i,j\}} \exp(\alpha_m)}, \end{aligned} \tag{3}$$

where $\mathbf{c}$, $\mathbf{b}$, and $\mathbf{A}$ are learnable parameters, and $\tanh$ introduces non-linearity. This structure constitutes a one-layer feedforward attention network with learnable parameters.

4.3.2 Complete Graph Construction

After computing the pair-wise region correlation, we construct complete graphs separately using the region correlations obtained from each view. Generally, let $\mathbf{S}^v$ denote the region correlation from a certain view. We build a graph $\mathcal{G}(\mathcal{V}; \mathbf{S}^v)$, where $\mathcal{V} = \{r_i\}^n_{i=1}$ denotes n regions that serve as nodes, and $\mathbf{S}^v$ denotes the adjacency matrix. Note that $\mathbf{S}^v$ is dynamic because the value of $\widetilde{\mathbf{F}}^v$ is learnable. Consequently, we construct the weighted complete graphs $\mathcal{G}^{sem}$, $\mathcal{G}^{out}$, and $\mathcal{G}^{in}$ based on the region correlations $\mathbf{S}^{sem}$, $\mathbf{S}^{out}$, and $\mathbf{S}^{in}$.

4.4 Multi-Graph Representation Learning

4.4.1 Graph Embedding

Based on these weighted complete graphs, we employ the message-passing mechanism [9] in graph convolutional networks (GCN) to learn the latent representation of each node (i.e., region). Formally, in each graph, we apply a GCN encoder to generate the new feature vector for each region,

$$\mathbf{H}^v = \text{Relu}((\mathbf{D}^v)^{-\frac{1}{2}}\mathbf{S}^v(\mathbf{D}^v)^{-\frac{1}{2}}\widetilde{\mathbf{F}}^v\mathbf{\Theta}), \tag{4}$$

where $\mathbf{D}^v$ is the degree matrix of $\mathbf{S}^v$, and $\mathbf{\Theta}$ is a linear transformation with learnable parameters. We finally obtain the output node representations from graphs $\mathcal{G}^{sem}$, $\mathcal{G}^{out}$, and $\mathcal{G}^{in}$ as $\mathbf{H}^{sem}$, $\mathbf{H}^{out}$, and $\mathbf{H}^{in}$.

4.4.2 Embedding Fusion

To encourage collaboration and information sharing among different graphs, we utilize an attention-based fusion method following [31] to effectively propagate knowledge

across the region representations. Formally, given the representations $\mathbf{H}^{sem}$, $\mathbf{H}^{out}$, and $\mathbf{H}^{in}$, we employ self-attention to compute the new representation $\widehat{\mathbf{H}}^v$,

$$\widehat{\mathbf{H}}^v = \text{Self-Attention}(\mathbf{H}^{sem}, \mathbf{H}^{out}, \mathbf{H}^{in}), \tag{5}$$

where Self-Attention$(\cdot)$ is the attention operation. Further, we combine each kind of region representation and the attentional representation with a weighted function as

$$\widetilde{\mathbf{H}}^v = \eta\widehat{\mathbf{H}}^v + (1-\eta)\mathbf{H}^v, v \in \{sem,\ out,\ in\}, \tag{6}$$

where η is a weight balancing the two components. Subsequently, we adaptively fuse the representations to generate the final region representations $\mathbf{E}$,

$$\begin{aligned} \mathbf{E} &= \sum_{v\in\{sem,\ out,\ in\}} w_v\widetilde{\mathbf{H}}^v, \\ w_v &= \text{softmax}(\text{LeakyRelu}(\widetilde{\mathbf{H}}^v)), \end{aligned} \tag{7}$$

where w_v is the weight of the v-th kind of representation and LeakyRelu$(\cdot)$ is an activation function.

4.5 Learning Objectives

Inspired by Chen et al. [2], we design various types of training tasks based on semantic and mobility features, i.e., mobility prediction and semantic relation reconstruction. As these training tasks are feature-specific, we average the final region representation $\mathbf{E}$ and the feature-specific representation (with global information) $\widetilde{\mathbf{H}}^v$ (i.e., $\mathbf{E}^v = (\mathbf{E} + \widetilde{\mathbf{H}}^v)/2$) to generate $\mathbf{E}^{sem}$, $\mathbf{E}^{out}$, and $\mathbf{E}^{in}$.

4.5.1 Mobility Prediction

Our objective is to predict the ending region given the starting region, or conversely, by utilizing the region representations $\mathbf{E}^{out}$ and $\mathbf{E}^{in}$. This task captures the human movement patterns, which inherently reflect functional or habitual correlations between regions (e.g., commuting patterns). Given a specific starting region r_j, the distribution of the ending region r_k is calculated as

$$P_{out}(r_j \to r_k) = \frac{\exp(\mathbf{E}_j^{out^{\mathrm{T}}} \cdot \mathbf{E}_k^{in})}{\sum_{z=1}^{n}\exp(\mathbf{E}_j^{out^{\mathrm{T}}} \cdot \mathbf{E}_z^{in})}. \tag{8}$$

Similarly, we calculate the distribution of the starting region r_j for a given ending region r_k as

$$P_{in}(r_j \to r_k) = \frac{\exp(\mathbf{E}_j^{out^{\mathrm{T}}} \cdot \mathbf{E}_k^{in})}{\sum_{z=1}^{n}\exp(\mathbf{E}_z^{out^{\mathrm{T}}} \cdot \mathbf{E}_k^{in})}. \tag{9}$$

Based on the human mobility dataset $\mathcal{M}$, the learning objective is defined as

$$\begin{aligned} \mathcal{L}^{\text{mobility}} = \frac{1}{|\mathcal{M}|}\sum_{(r_j,r_k)\in\mathcal{M}} [&-\log P_{out}(r_j \to r_k) \\ &- \log P_{in}(r_j \to r_k)]. \end{aligned} \tag{10}$$

TABLE 1: Data description.

	NYC Data	SF Data
# Regions	270	175
# POIs	16,925	28,578
# Taxi trips	10,919,198	357,749
# Check-ins	108,849	87,750
# Crime records	67,985	48,489

4.5.2 Semantic Relation Reconstruction

To enable the model to capture contextual semantic similarity between regions, we further propose a task that involves the reconstruction of semantic correlations using the corresponding region representations. The learning objective is formulated as

$$\mathcal{L}^{\text{semantic}} = \frac{1}{n^2}\sum_{j,k}(\mathbf{S}_{jk}^{\text{semantic}} - \mathbf{E}_j^{sem^{\mathrm{T}}} \cdot \mathbf{E}_k^{sem})^2, \tag{11}$$

where $\mathbf{S}_{j,k}^{\text{semantic}}$ is computed as the cosine similarity between the raw semantic features $\mathbf{F}_j^{sem}$ and $\mathbf{F}_k^{sem}$.

Finally, we combine the above objectives and obtain the overall loss function,

$$\mathcal{L} = \lambda\mathcal{L}^{\text{mobility}} + (1-\lambda)\mathcal{L}^{\text{semantic}}, \tag{12}$$

where λ is the weight balancing different components of the loss. The objective function can be optimized using the stochastic gradient descent method. Once optimized, the learned region representations $\mathbf{E}$ can be used in various urban downstream tasks, e.g., land usage clustering, region popularity prediction, and region crime prediction.

5 Experiments

5.1 Experimental Settings

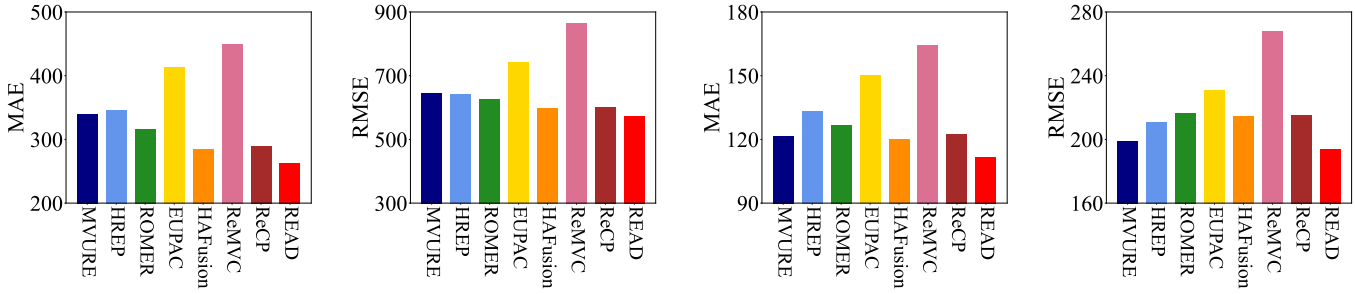
Datasets. We utilize real-world data from New York City[1] (NYC) and San Francisco[2] (SF). Datasets consist of city-defined regions, POIs and taxi trip records (pickup/drop-off locations) for learning region representations. Additionally, check-in and crime records are used for region popularity and crime prediction, respectively [11], [17]. A detailed description of the datasets is presented in Table 1.

Model Parameters. All model parameters are randomly initialized and trained from scratch. The raw semantic and mobility features are used solely as inputs to the network. Specifically, the semantic feature of each region, $\mathbf{F}_i^{sem}$, is represented as a vector of POI counts across the predefined categories, with a dimensionality of 9 for the NYC dataset and 26 for the SF dataset. The mobility features consist of an outflow vector $\mathbf{F}_i^{out}$ and an inflow vector $\mathbf{F}_i^{in}$, derived by normalizing inter-region transition counts from the raw mobility data. These vectors have dimensionalities of 270 and 175 for the NYC and SF datasets, respectively, corresponding to the total number of regions in each city. The dimension of region representations is set at 128. In the city-context Transformer encoder, we set the hidden size at 128 and the number of heads at 8. In the graph embedding module, we set the number of GCN layers at 2 and the hidden size at 128; in the embedding fusion module, we set η at 0.2. We set λ in the final objective loss at 0.5.

1. https://opendata.cityofnewyork.us
2. https://datasf.org/opendata/

TABLE 2: Performance comparison of different methods on the NYC data, where the performance improvements of READ are compared with the best of these baseline methods, marked by the asterisk.

Method	Land Usage Clustering		Region Popularity Prediction		Region Crime Prediction	
	ARI↑	F-measure↑	MAE↓	RMSE↓	MAE↓	RMSE↓
MV-PN	0.033 ± 0.01	0.068 ± 0.01	291.26 ± 20.63	433.96 ± 19.75	126.66 ± 1.05	181.54 ± 1.23
CGAL	0.057 ± 0.05	0.089 ± 0.05	294.91 ± 17.40	439.47 ± 7.15	130.19 ± 0.91	185.32 ± 1.27
MVURE	0.398 ± 0.03	0.412 ± 0.03	231.61 ± 8.18	341.97 ± 11.31	118.13 ± 6.85	166.25 ± 9.45
HREP	0.452 ± 0.02	0.459 ± 0.02	220.57 ± 10.15	325.12 ± 14.32	117.84* ± 7.91	166.37* ± 7.67
ROMER	0.433 ± 0.01	0.453 ± 0.01	231.13 ± 9.98	344.17 ± 19.67	121.24 ± 6.52	175.84 ± 11.98
EUPAC	0.477 ± 0.06	0.487 ± 0.06	222.02 ± 22.79	326.85 ± 31.51	118.82 ± 14.84	168.73 ± 22.49
HAFusion	0.417 ± 0.01	0.436 ± 0.01	191.99* ± 6.28	290.83* ± 4.91	129.39 ± 8.44	177.72 ± 10.45
ReMVC	0.456 ± 0.04	0.464 ± 0.04	280.41 ± 18.12	399.79 ± 18.72	138.22 ± 8.39	194.01 ± 9.39
ReCP	0.490* ± 0.01	0.508* ± 0.01	194.76 ± 12.64	296.60 ± 10.56	137.41 ± 13.03	190.55 ± 15.08
READ	**0.508 ± 0.02**	**0.525 ± 0.02**	**182.47 ± 12.11**	**277.83 ± 15.53**	**104.05 ± 5.55**	**151.32 ± 6.08**
Improvements	**3.67%**	**3.35%**	**4.96%**	**4.47%**	**11.70%**	**9.05%**



(a) Popularity Prediction (b) Crime Prediction

Fig. 4: Performance comparison on the SF data.

To optimize our model, we adopt Adam and initialize the learning rate at 0.001 with a linear decay. We implement READ and baselines with PyTorch 1.12.0 on a server with NVIDIA GeForce RTX 2080 Ti.

Baselines. We compare READ's performance against state-of-the-art region embedding methods that utilize both POI and human mobility data.

- **MV-PN** leverages the mobility connectivities between POIs to construct static graphs within each region. These graphs are then flattened and concatenated as initial region vectors and inputted into an AutoEncoder to learn the final region embeddings.
- **CGAL** extends MV-PN and employs adversarial learning technique to convert graphs into region embedding.
- **MVURE** constructs multiple static graphs by computing region correlations using POI and mobility features and connecting each region with its k-nearest neighbors, and then applies graph attention networks to learn single-view representations, which are integrated to form a unified multi-view representation of each region.
- **HREP** constructs a heterogeneous graph using multiple data sources to form different types of edges and generates relation-specific region embeddings, which are integrated in the second stage.
- **ROMER** constructs a complete graph that connects any two regions and employs graph attention networks to learn single-view representations, and then integrates them into a unified region representation.
- **EUPAC** constructs a heterogeneous graph like HREP, to generate relation-specific region representations, and then applies adversarial contrastive learning on these representations to boost model performance.
- **HAFusion** enhances region embedding by modeling the correlations of different regions in various views and the higher-order correlations between regions in the fused region representations.
- **ReMVC** maps raw region features directly to a latent space and learns region representations through both intra-view and inter-view contrastive learning objectives.
- **ReCP** learns region representations through intra-view and inter-view contrastive learning, with an enhanced inter-view module to improve consistency across views.

5.2 Comparison with Baselines

5.2.1 *Land Usage Clustering*

Given that the New York City (NYC) dataset is the only one containing land use labels, we focus our land usage clustering analysis on this dataset. Following the methodology outlined in [30], we adopt the district divisions defined by the community boards as the ground truth, partitioning the Manhattan borough into 29 distinct districts. Using the learned representations of these regions, we apply k-means clustering with $k = 29$ to group regions into clusters, as suggested in [11]. The underlying assumption is that regions sharing the same land usage type will be grouped into the same cluster. To evaluate the effectiveness of this clustering, we employ two widely used metrics: the Adjusted Rand Index (ARI) and the F-measure, as described in [11], [31]. All methods are tested on the same dataset, and each experiment is repeated five times to ensure reliability. The results, presented in Table 2, report the mean values along with their standard deviations.

Our analysis reveals several key insights. First, methods such as MV-PN and CGAL exhibit relatively poor performance. This is likely because these approaches primarily focus on capturing point-of-interest (POI) and mobility features within individual regions, without adequately considering the dynamic interactions between different regions. In contrast, methods like MVURE, HREP, ROMER, EUPAC, and HAFusion demonstrate superior performance. These techniques explicitly model the relationships between regions and leverage graph attention networks to integrate information from multiple perspectives, resulting in more accurate clustering outcomes.

Second, methods such as ReMVC and ReCP employ multi-layer perceptron (MLP) architectures to encode region features and utilize contrastive learning to model both intra-region and inter-region relationships. These approaches yield promising results, highlighting the importance of capturing both local and global region characteristics.

Finally, our proposed READ emerges as the top-performing method, surpassing all baseline approaches. By disentangling region features and adaptively discovering region correlations, READ achieves significant improvements over the best baseline (ReCP), with average gains of 3.67% in ARI and 3.35% in F-measure. Furthermore, statistical validation through paired t-tests confirms that READ's improvements are statistically significant, with a p-value of less than 0.01. This underscores the robustness and effectiveness of READ in capturing the complex relationships inherent in urban region data.

5.2.2 Region Popularity Prediction

To further assess the quality of the learned region representations, we conduct experiments on the region popularity prediction task. We measure the popularity of a region by aggregating the check-in counts within that region, following the methodology described in [30]. These aggregated counts serve as the ground truth for region popularity. Using the learned region representations as input features, we train a Ridge regression model to predict popularity. The performance of the model is evaluated using two widely adopted metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). To ensure robust evaluation, we employ 5-fold cross-validation on both the New York City (NYC) and San Francisco (SF) datasets. The results are summarized in Table 2 and illustrated in Figure 4(a).

Our experimental findings demonstrate that the proposed READ method consistently outperforms all baseline approaches. On the NYC dataset, READ achieves average improvements of 4.96% in MAE and 4.47% in RMSE compared to the best-performing baseline, HAFusion. The superior performance of READ can be attributed to its novel one-stage paradigm, which adaptively discovers region correlations and integrates them into the learning process. This approach enables the model to generate more informative and discriminative region representations, ultimately leading to better prediction accuracy.

5.2.3 Region Crime Prediction

We further conduct experiments on the crime prediction task, following [31], [33]. This task involves predicting the number of criminal incidents occurring within each region. Using the learned region representations as input features, we train a Ridge regression model to perform the prediction. The model's performance is also assessed using MAE and RMSE. The results for both New York City (NYC) and San Francisco (SF) datasets are reported in Table 2 and visualized in Figure 4(b).

The experimental results demonstrate that READ consistently outperforms all baseline methods. On the NYC dataset, READ achieves average improvements of 11.70% in MAE and 9.05% in RMSE compared to the best-performing baseline, HREP. By explicitly modeling the relationships between regions and extracting meaningful features, READ generates more accurate and informative region representations, which are crucial for tasks such as crime prediction.

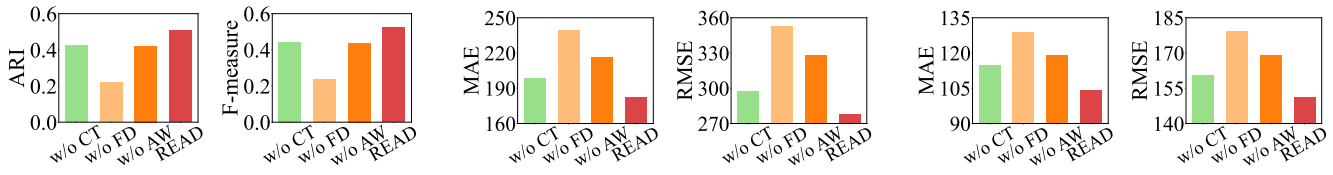
These findings further validate the effectiveness of READ's methodology in learning high-quality region embeddings. The consistent performance gains across multiple tasks and datasets underscore the robustness and versatility of READ as a framework for urban region analysis. These results underscore the potential of READ as a powerful tool for a wide range of urban computing applications requiring accurate and interpretable region embeddings, such as popularity and crime prediction and beyond.

5.3 Ablation Study and Parameter Sensitivity

To investigate the contribution of each module in READ to the quality of region representations, we design three variants of the model and conduct a comprehensive ablation study. These variants are constructed by systematically removing or modifying key components of READ, allowing us to assess their individual impact on performance.

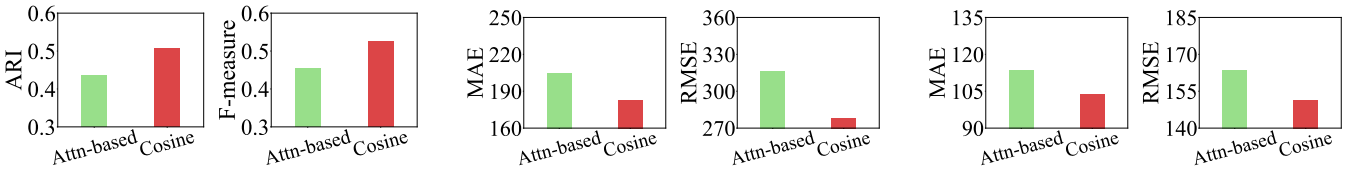
- **READ w/o CT**: We replace the city-context Transformer encoder with a simple multi-layer perceptron (MLP) layer to encode raw features. This modification helps evaluate the importance of the Transformer-based encoder in capturing contextual information.
- **READ w/o FD**: We replace the city-context Transformer encoder with a standard Vanilla Transformer that directly aggregates attention-weighted region features without performing feature disentanglement. Specifically, we remove the subtraction of the city-context features and set the updated representation as $\boldsymbol{\beta}^v\mathbf{V}$ instead of $\mathbf{V} - \boldsymbol{\beta}^v\mathbf{V}$. This ablation is designed to assess the role of feature disentangling in enhancing region representations.
- **READ w/o AW**: We perform feature disentangling based on geographic neighbors, where the disentangled features are computed as $\Delta\mathbf{F}_i^v = \mathbf{F}_i^v - \sum_{r_j \in \mathcal{N}_i^{geo}} \mathbf{F}_j^v / |\mathcal{N}_i^{geo}|$. Here, $\mathcal{N}_i^{geo}$ represents the set of geographic neighbors of region r_i. Static KNN graphs are then constructed using $\Delta\mathbf{F}_i^v$, removing the adaptive weighting mechanism. This variant helps evaluate the importance of adaptive region correlation discovery.

The results of READ and its variants are illustrated in Figure 5. We observe that both READ w/o CT and READ w/o FD exhibit significantly poorer performance compared to the full READ model. This confirms the effectiveness of the city-context Transformer encoder and the feature disentangling mechanism in improving the quality of region representations. In particular, the degradation observed in READ w/o FD underscores the importance of explicitly



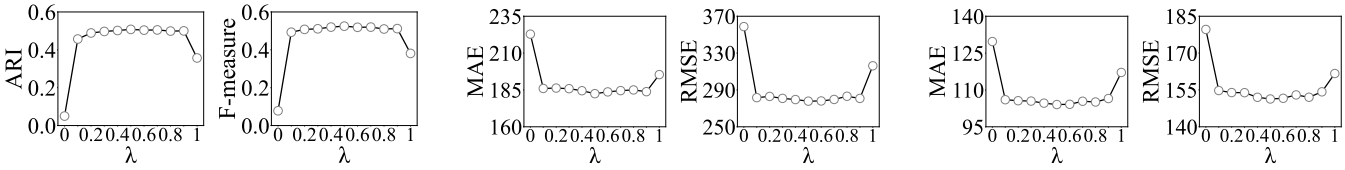
(a) Land Usage Clustering (b) Region Popularity Prediction (c) Region Crime Prediction

Fig. 5: Performance comparison of different variants on the NYC data.



(a) Land Usage Clustering (b) Region Popularity Prediction (c) Region Crime Prediction

Fig. 6: Performance comparison of different pair-wise region correlations on the NYC data.



(a) Land Usage Clustering (b) Region Popularity Prediction (c) Region Crime Prediction

Fig. 7: Parameter analysis across three downstream tasks on the NYC data.

disentangling region-specific features from shared urban patterns. Without the subtraction mechanism in our city-context Transformer, the model tends to retain common signals shared by many regions, which reduces the ability to capture region-specific features and degrades the quality of the learned representations. Furthermore, while READ w/o AW incorporates feature disentangling, its performance remains inferior to READ, highlighting the critical role of adaptive region correlation discovery in capturing complex inter-region relationships.

Next, we evaluate the impact of correlation metrics on model performance. Figure 6 demonstrates that cosine similarity consistently surpasses attention-based correlation across all tasks. The superiority of cosine similarity stems from its simplicity, interpretability, and geometric properties: it provides a direct, bounded measure of angular distance between region vectors, effectively capturing fine-grained semantic relationships. In contrast, attention-based correlation employs learnable parameters and nonlinear transformations, which increase model complexity and may increase noise sensitivity and training instability.

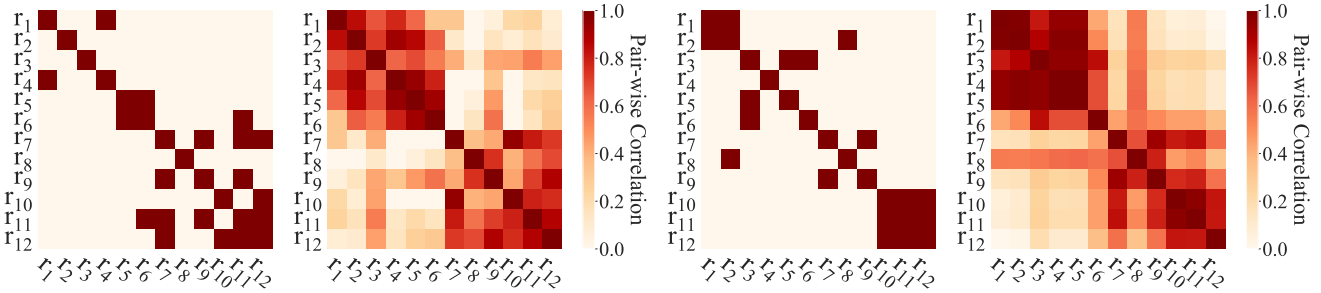
Finally, we examine the sensitivity of the model to λ, which controls the contribution of different components to the total loss. By varying λ from 0 to 1 in increments of 0.1, we observe that using only semantic or mobility loss results in suboptimal performance. Increasing λ from 0 to 0.5 leads to significant performance improvements, while further increases beyond 0.5 cause a decline in performance. This suggests that an optimal balance between semantic and mobility features is crucial for achieving the best results. Similar trends are observed in the ablation study and parameter sensitivity analysis on the San Francisco (SF) dataset, further validating the robustness of our findings.

Metric	Value
Training Time (per epoch)	0.19 seconds
Inference Time (land usage clustering)	0.24 seconds
Inference Time (popularity prediction)	2.66 milliseconds
Inference Time (crime prediction)	2.32 milliseconds
Parameter Count	$\sim$785k

TABLE 3: Training time, inference efficiency, and model size of our READ framework.

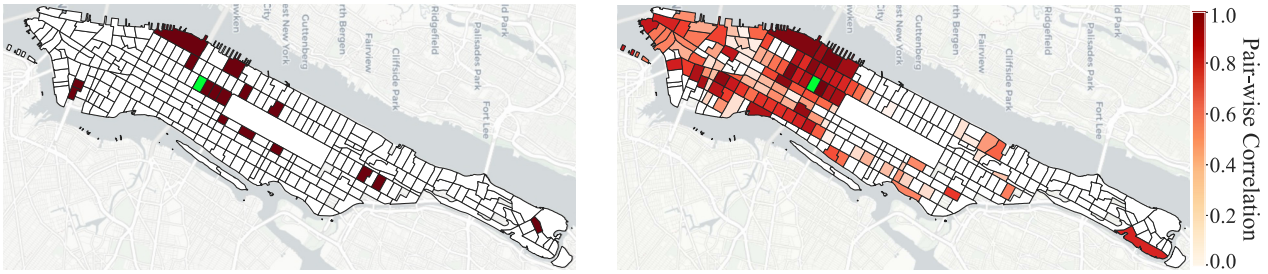
5.4 Model Complexity and Efficiency Analysis

We analyze both the theoretical complexity and empirical efficiency of the READ framework to assess its practicality in real-world urban applications. Theoretically, the overall training complexity of READ is $\mathcal{O}(|\mathcal{M}| \cdot Nd + N^2d + Nd^2)$, where N is the number of regions, d is the embedding dimension, and $|\mathcal{M}|$ is the number of mobility records. Specifically, the feature disentanglement module based on a modified Transformer involves attention computation and residual extraction with $\mathcal{O}(N^2d + Nd^2)$ complexity. Pair-wise region correlation (via cosine similarity or attention) adds $\mathcal{O}(N^2d)$. The GCN updates on three complete graphs contribute $\mathcal{O}(N^2d)$, while the cross-view embedding fusion module also incurs $\mathcal{O}(N^2d + Nd^2)$ due to attention-based integration. Regarding the loss functions, mobility prediction introduces $\mathcal{O}(|\mathcal{M}| \cdot Nd)$ complexity, and semantic recon-



(a) KNN (Semantic) (b) READ (Semantic) (c) KNN (Mobility) (d) READ (Mobility)

Fig. 8: Region correlation comparison between KNN and our proposed READ: $r_1 - r_6$ represent regions from the first land use label, and $r_7 - r_{12}$ from the second.



(a) KNN (b) READ

Fig. 9: Region correlation comparison between KNN and READ: a case study taking Times Square as the target region.

struction adds another $\mathcal{O}(N^2 d)$. This level of complexity is typical for models involving full pairwise interactions, and remains tractable for city-scale datasets. This level of complexity is typical for graph-based models with full pairwise interactions and remains tractable for city-scale datasets.

To complement the theoretical analysis, we report empirical results including the training time, inference time, and parameter counts in Table 3. All measurements are averaged over five runs on the NYC dataset. The reported training time refers to one complete epoch of end-to-end training, excluding offline preprocessing steps. For inference time, since READ is an unsupervised representation learning framework, its runtime depends on downstream usage. To provide a representative reference, we measure the inference time in three downstream tasks: popularity prediction, crime rate estimation (both using ridge regression), and land use clustering (using k-means). The READ model consists of $\sim$ 785k trainable parameters, mainly distributed across the city-context Transformer, GCN layers, and the cross-modality fusion module.

5.5 Visualization of Region Correlations

To demonstrate the effectiveness of our adaptive method, we compare visualizations of region correlations derived from a KNN-based approach and our proposed method READ. The KNN-based method calculates cosine similarity between regions based on raw semantic and mobility features, creating a graph connecting each region to its k most similar neighbors. Conversely, READ calculates pairwise correlations using our novel disentangled features, constructing complete graphs with learnable edge weights.

To compare the two methods, we randomly select 12 regions, with the first six regions r_1 to r_6 sharing a same land use label and the latter six regions r_7 to r_{12} sharing another land use type. It is worth noting that the region similarity comparison is performed among all regions in the study area, not only among the 12 selected regions. From Figures 8(a) and 8(c), we observe that the KNN-based method only partially captures the similarity between regions with the same land use label, for example, among the 12 regions, the region r_1 is only related to r_4 using semantic features and KNN. The connections with other similar regions are overlooked, e.g., with r_2, r_3, r_5 and r_6. In contrast, as shown in Figures 8(b) and 8(d), our method READ is able to capture the inter-region correlations in a comprehensive manner, effectively leading to larger correlations among regions sharing the same land use label, and smaller similarities with regions of different land use types. By adaptively assigning correlation weights, our approach captures a more flexible and comprehensive set of inter-region relationships. This enables smoother information flow and a more nuanced understanding of urban structures.

To illustrate the advantages of our method in a more concrete manner, we conduct a case study focusing on Times Square in NYC as the target area and visualize the similar regions found by both the methods KNN and READ. We analyze how both methods establish connections with other regions in New York City. As shown in Figure 9, we construct a heatmap showing the correlation strength between the target region (colored green) and all other regions. The KNN-based method connects the target region only to a small subset of predefined similar regions based on raw semantic features, such as those containing Carnegie Hall

and David Zwirner Gallery. In contrast, our method extends beyond these connections by also assigning significant attention to additional regions, including those containing the Chrysler Building, Madison Square Garden, and the Empire State Building, which share similar characteristics with the target region. This broader attention distribution underscores the advantage of our approach in capturing a more comprehensive set of inter-region relationships.

6 Conclusion and Future Work

In this paper, we propose a Region Embedding method with Adaptive region correlation Discovery named READ to generate the pre-trained region representations. READ consists of three modules: a disentangled region feature learning module that uses a city-context Transformer encoder to learn disentangled region features, an adaptive weighted multi-graph construction module to calculate pair-wise region correlations and build multiple complete graphs with learnable edge weights, and a multi-graph representation learning module to learn comprehensive region representations that integrate information from multiple graphs. We conduct comprehensive experiments on three downstream tasks to evaluate the proposed READ model. Experimental results demonstrate that READ outperforms state-of-the-art region embedding methods, proving that adaptive region correlation discovery with disentangled region features is crucial for effective region embedding.

While the proposed READ demonstrates significant advancements in urban region embedding, it encounters the limitation that the learned urban region representations are not adaptable to varying sizes of regions and randomly designated regions (e.g., drawn by users). Future work to address this limitation is desirable.

7 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62572274, the Key Scientific and Technological Innovation Project of Shandong Province under Grant No. 2024CXGC010113 and 2024CXG010213, and the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources.

References

[1] J. Bing, M. Chen, M. Yang, W. Huang, Y. Gong, and L. Nie. Pretrained semantic embeddings for poi categories based on multiple contexts. *IEEE Transactions on Knowledge and Data Engineering*, 35(09):8893–8904, 2023.

[2] J. Chen, T. Liu, and R. Li. Region profile enhanced urban spatio-temporal prediction via adaptive meta-learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 224–233, 2023.

[3] M. Chen, Y. Zhao, Y. Liu, X. Yu, and K. Zheng. Modeling spatial trajectories with attribute representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1902–1914, 2020.

[4] M. Deng, C. Chen, W. Zhang, J. Zhao, W. Yang, S. Guo, H. Pu, and J. Luo. Hyperregion: Integrating graph and hypergraph contrastive learning for region embeddings. *IEEE Transactions on Mobile Computing*, 2024.

[5] Y. Fu, P. Wang, J. Du, L. Wu, and X. Li. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 906–913, 2019.

[6] X. Hao, W. Chen, Y. Yan, S. Zhong, K. Wang, Q. Wen, and Y. Liang. Urbanvlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction. *arXiv e-prints*, pages arXiv–2403, 2024.

[7] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

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

[9] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

[10] T. Li, S. Xin, Y. Xi, S. Tarkoma, P. Hui, and Y. Li. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3282–3291, 2022.

[11] Z. Li, W. Huang, K. Zhao, M. Yang, Y. Gong, and M. Chen. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8724–8732, 2024.

[12] Y. Liu, J. Ding, Y. Fu, and Y. Li. Urbankg: An urban knowledge graph system. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–25, 2023.

[13] Y. Liu, X. Zhang, J. Ding, Y. Xi, and Y. Li. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, pages 4150–4160, 2023.

[14] Y. Liu, K. Zhao, and G. Cong. Efficient similar region search with deep metric learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1850–1859, 2018.

[15] Y. Luo, F.-l. Chung, and K. Chen. Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4294–4298, 2022.

[16] X. Pan, X. Cai, S. Xu, Y. Zhang, P. Nie, and X. Yuan. Geoco: geographical correlation enhanced network for poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[17] F. Sun, J. Qi, Y. Chang, X. Fan, S. Karunasekera, and E. Tanin. Urban region representation learning with attentive fusion. In *2024 IEEE 40th International Conference on Data Engineering*, pages 4409–4421. IEEE, 2024.

[18] T. Sun, K. Zhao, and M. Chen. Human-ai interaction: Human behavior routineness shapes ai performance. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8476 – 8487, 2024.

[19] J. Tang, H. Zhang, B. Zhang, J. Jin, and Y. Lyu. Spemi: Normalizing spatial imbalance with spatial eminence transformer for citywide region embedding. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 92–95, 2022.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[21] H. Wang and Z. Li. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 237–246, 2017.

[22] L. Wang, S. Wu, Q. Liu, Y. Zhu, X. Tao, and M. Zhang. Bi-level graph structure learning for next poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

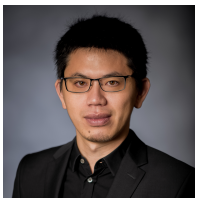
[23] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu. Heterogeneous graph attention network. In *the World Wide Web Conference*, pages 2022–2032, 2019.

[24] Z. Wang, H. Li, and R. Rajagopal. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1013–1020, 2020.

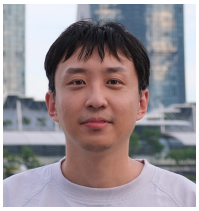
[25] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma, and P. Hui. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*, pages 3308–3316, 2022.

[26] Z. Xu and X. Zhou. Cgap: Urban region representation learning with coarsened graph attention pooling. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 7518–7526, 2024.

[27] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and Y. Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017, 2024.
[28] Z. Yao, Y. Fu, B. Liu, W. Hu, and H. Xiong. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3919–3925, 2018.
[29] C. Zhang, K. Zhao, and M. Chen. Beyond the limits of predictability in human mobility prediction: context-transition predictability. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4514–4526, 2022.
[30] L. Zhang, C. Long, and G. Cong. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, (01):1–6, 2022.
[31] M. Zhang, T. Li, Y. Li, and P. Hui. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4431–4437, 2021.
[32] Y. Zhang, Y. Fu, P. Wang, X. Li, and Y. Zheng. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1700–1708, 2019.
[33] S. Zhou, D. He, L. Chen, S. Shang, and P. Han. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4981–4989, 2023.



Kai Zhao is currently a research scientist and tech lead at the Walmart AI lab. Previously he was a professor in machine learning at Georgia State University. His research interests are AI and machine learning. Before coming to GSU, Dr. Kai Zhao worked as a post-doc at New York University 2016-2017. He received his Ph.D. in Computer Science from the University of Helsinki, Finland in 2015. Dr. Kai Zhao has authored or co-authored over 40 publications in top AI journals and conferences.



Weiming Huang is a Lecturer at the School of Geography, University of Leeds, UK. He obtained his Ph.D. in Geographical Information Science from Lund University, Sweden, and was a Wallenberg Postdoctoral Fellow at Nanyang Technological University and Lund University. His research interests include spatial data mining and geospatial foundation models.

Meng Chen is currently an associate professor in the School of Software, Shandong University, China. He received his Ph.D. degree in computer science and technology in 2016 from Shandong University, China. He worked as a Postdoctoral fellow from 2016 to 2018 in the School of Information Technology, York University, Canada. His research interest is in the area of spatio-temporal data mining and urban computing. He has published over 40 papers in prestigious journals and conferences in data mining field such as *TKDE*, *TOIS*, *KDD*, and *ICML*.

Yongshun Gong is a Professor in the School of Software, Shandong University, China. He received his Ph.D. degree from University of Technology Sydney in 2020. His principal research interest relies in the data science and machine learning. He has published above 30 papers at the first-tier venues.

Hongwei Jia is currently pursuing a Ph.D. degree at University of Tsukuba, Japan. He received his M.S. degree in artificial intelligence in 2025 from Shandong University, China. His research interest is in the area of spatio-temporal data mining. He has published papers in prestigious journals and conferences such as *TKDE*, *ICML* and *IJCAI*.

Haoran Xu received his Ph.D. degree in the School of Computer Science and Technology from Shandong University, China, in 2021. Currently, he is working as a postdoc at Shandong University. His research interests include data management and analysis.

Zechen Li is currently pursuing a Ph.D. degree in the School of Software at Shandong University, China. She received her B.S. degree in software engineering in 2023 from Shandong University, China. Her research interest is in the area of spatio-temporal data mining. She has published papers in prestigious journals and conferences such as *TKDE*, *ICML*, *KDD* and *AAAI*.

Hongjun Dai received his Ph.D. degree in the School of Software from Zhejiang University, Hangzhou, China, in 2007. Currently, he is working as a professor at Shandong University, Jinan. His research interests include AI inspried IoT, optimizations of Operating System.