

# Urban Region Embedding via Multi-View Contrastive Prediction

Zechen Li<sup>1</sup>, Weiming Huang<sup>2</sup>, Kai Zhao<sup>3</sup>, Min Yang<sup>1</sup>, Yongshun Gong<sup>1</sup>, Meng Chen<sup>1\*</sup>

<sup>1</sup> School of Software, Shandong University

<sup>2</sup> School of Computer Science and Engineering, Nanyang Technological University

<sup>3</sup> Robinson College of Business, Georgia State University

lizechenn@gmail.com, weiming.huang@ntu.edu.sg, kzhao4@gsu.edu, myang3@sdu.edu.cn, yongshun2512@hotmail.com, mchen@sdu.edu.cn

## Abstract

Recently, learning urban region representations utilizing multi-modal data (information views) has become increasingly popular, for deep understanding of the distributions of various socioeconomic features in cities. However, previous methods usually blend multi-view information in a posterior stage, falling short in learning coherent and consistent representations across different views. In this paper, we form a new pipeline to learn consistent representations across varying views, and propose the multi-view Contrastive Prediction model for urban Region embedding (ReCP), which leverages the multiple information views from point-of-interest (POI) and human mobility data. Specifically, ReCP comprises two major modules, namely an intra-view learning module utilizing contrastive learning and feature reconstruction to capture the unique information from each single view, and inter-view learning module that perceives the consistency between the two views using a contrastive prediction learning scheme. We conduct thorough experiments on two downstream tasks to assess the proposed model, i.e., land use clustering and region popularity prediction. The experimental results demonstrate that our model outperforms state-of-the-art baseline methods significantly in urban region representation learning.

## Introduction

A deep understanding of the spatial distribution of various socioeconomic factors in cities such as land use or population distribution, is important for urban planning and management. In recent years, an increasingly popular trend in the community of urban computing has been to partition a city into numerous regions and utilize various urban sensory data to learn the latent representations of the regions, which can subsequently be used in varying urban sensing tasks, e.g., land usage clustering, house price prediction, and population density inference (Liu et al. 2021; Li et al. 2022; Liu et al. 2023; Huang et al. 2023; Xu et al. 2023b; Li et al. 2023). This trend can also be attributed to the prosperity of mobile sensing technologies, which has led to the rapid accumulation of urban sensing data, such as human trajectories or points-of-interest (POIs) (Chen, Yu, and Liu 2018; Zhang, Zhao, and Chen 2022; Xu et al. 2023a; Zhang et al. 2023). Such various urban data provide more opportunities for tackling the problem of region representation learning.

\*Corresponding author.

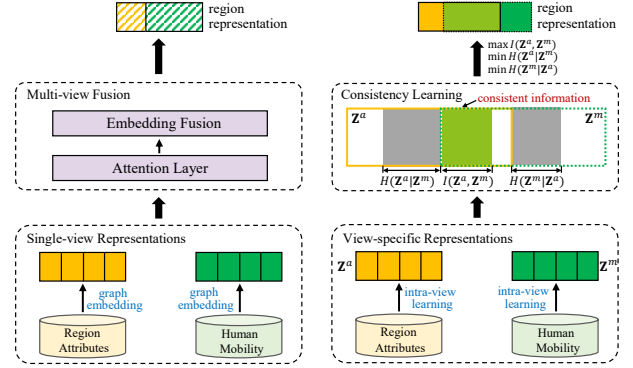


Figure 1: Illustration of (a) multi-view fusion paradigm and our proposed (b) consistency learning paradigm for region embedding. In the right figure, the solid and dotted rectangles denote the region representations  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$  from the attribute and mobility views, respectively. The mutual information  $I(\mathbf{Z}^a, \mathbf{Z}^m)$  (chartreuse area) quantifies the amount of information shared by  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$ ; the conditional entropy  $H(\mathbf{Z}^a|\mathbf{Z}^m)$  (grey area) quantifies the amount of information of  $\mathbf{Z}^a$  conditioned on  $\mathbf{Z}^m$ . To learn consistent region representations across different views, it is encouraged to maximize  $I(\mathbf{Z}^a, \mathbf{Z}^m)$  and minimize  $H(\mathbf{Z}^a|\mathbf{Z}^m)$  and  $H(\mathbf{Z}^m|\mathbf{Z}^a)$ .

Many previous studies have attempted to learn region representations by utilizing human mobility data. For instance, Wang et al. (Wang and Li 2017) construct flow graphs and spatial graphs using taxi flow data and propose a graph embedding method to learn region representations. Yao et al. (Yao et al. 2018) extract human mobility patterns from taxi trajectories, and model the co-occurrence of origin-destination regions to learn region representations. The above methods merely rely on single-view data, which offers a limited perspective of regions and fails to provide a comprehensive representation. Further, recent studies (Zhang et al. 2021; Luo, Chung, and Chen 2022; Zhou et al. 2023) propose learning region representations through integrating data in multiple modalities, thus forming multiple information views. In this context, the technical focus of recent region embedding studies has shifted towards the fusion between multiple information views, where they usually follow the same pipeline: separate single-view represen-

tation followed by multiple-view fusion. Such a pipeline is demonstrated in Figure 1(a), where, it (1) separately models each information view (usually with a graph structure) and learns multiple single-view representations for each region, and (2) leverages certain fusion techniques (e.g., based on attention mechanisms) to blend multiple representations and yield the final multi-view region representation.

The previous multi-view region embedding methods have been effective in certain analyses, but they come with a notable shortcoming: neglecting the information consistency across different views when generating the final region representation. Intuitively, the information carried by multiple views of a region is highly correlated, and thus their representations should be consistent. For example, an entertainment region could contain multiple bars and restaurants (region attribute view based on POIs), as well as a large number of nighttime mobility flows (human mobility view). Both views can reflect the intrinsic characteristics of this region (i.e., entertainment function). If we manage to leverage such correlation, it could be served as the constraint during the process of learning representations for each view, and enable the knowledge of transferring from one view to the other. Ultimately, the multi-view representations would become highly consistent and naturally fused.

Following the ideas above, we present a new pipeline - consistency learning paradigm - for multi-view region embedding from an information theory perspective (Tsai et al. 2021; Lin et al. 2021), where the multi-view representations are naturally fused through exchanging information between views along with learning view-specific region representations, rather than treating fusion as a posterior process. This new pipeline is shown in Figure 1(b). Given two view-specific region representations  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$  (where they are from the region attribute view and the human mobility view, respectively), we maximize the mutual information  $I(\mathbf{Z}^a, \mathbf{Z}^m)$  to increase the amount of the shared information (consistency) in the region representations of the two views. We also minimize the conditional entropy  $H(\mathbf{Z}^a|\mathbf{Z}^m)$  and  $H(\mathbf{Z}^m|\mathbf{Z}^a)$  to diminish the inconsistent information across the two views and improve the consistency further.

Based on the consistency learning paradigm, we propose a multi-view Contrastive Prediction model for urban Region embedding (ReCP), which can effectively enhance the consistency of region representations across different views. ReCP consists of two major components: intra-view learning and inter-view learning. In the intra-view learning component, to learn view-specific region representations, we compare each region with other dissimilar ones to embed the region into a latent space via contrastive learning; in the meantime, we also utilize autoencoders to capture view-specific region features for different views, which helps avoid model falling into a trivial solution. In the inter-view learning component, to learn the cross-view consistency of region representations, we design inter-view contrastive learning by maximizing  $I(\mathbf{Z}^a, \mathbf{Z}^m)$  and dual prediction between views by minimizing  $H(\mathbf{Z}^a|\mathbf{Z}^m)$  and  $H(\mathbf{Z}^m|\mathbf{Z}^a)$ .

To summarize, our contributions are as follows:

- We form a new pipeline following a consistency learning paradigm, to study the urban region embedding problem

by exploring the consistency across different views, using both human mobility and POI data. Different from existing multi-view region embedding methods which adopt the attention mechanisms to fuse representations of different views, we propose to learn consistent multi-view representations of regions by increasing the amount of shared information across multiple views from the information entropy perspective.

- We design the inter-view contrastive learning and dual prediction processes to diminish the inconsistent information across views and learn an informative and consistent region representation between different views, achieved by maximizing the mutual information among different views and minimizing the conditional entropy among them.
- We conduct extensive experiments to evaluate our model with real-world datasets. The results demonstrate that the proposed ReCP outperforms existing methods on two downstream tasks by a margin. Data and source code are available at <https://anonymous.4open.science/r/ReCP>.

## Problem Formulation

We formulate the urban region representation learning problem with the following definitions:

**Definition 1 (Urban Region)** *A city can be partitioned into  $n$  disjoint urban regions, denoted as  $\mathcal{R} = \{r_1, r_2, \dots, r_n\}$ .*

**Definition 2 (Region Attributes)** *In this study, region attributes are defined as inherent geographic features of regions. Specifically, we consider Point of Interest (POI) categories as region attributes following (Zhang, Long, and Cong 2022; Fu et al. 2019). These region attributes are represented as a set  $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_n\}$ , where  $\mathbf{A}_i \in \mathbb{R}^F$  and  $F$  represents the total number of POI categories. Each dimension in  $\mathbf{A}_i$  corresponds to the number of POIs with a specific category in the region  $r_i$ .*

**Definition 3 (Human Mobility)** *For a region  $r_i$ , we define its outflow feature  $\mathbf{S}_i^{j,t}$  as the number of trips made by all individuals originating from region  $r_i$  and destined for region  $r_j$  during a specific time interval  $t$ . Consequently, we generate a collection of outflow features based on the mobility data encompassing all regions within the set  $\mathcal{R}$ . This collection is represented as  $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_n\}$ , where  $\mathbf{S}_i \in \mathbb{R}^M$ . Here,  $M$  is calculated as the product of the number of regions,  $n$ , and the number of time intervals,  $N_t$ , within a day, for instance, 24. Similarly, by considering  $r_i$  as the destination region and the other regions  $r_j$  as the source regions, we can obtain an inflow feature vector, denoted as  $\mathbf{D}_i$ , and finally obtain a collection  $\mathcal{D} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n\}$  of inflow features for all regions.*

**Problem 1 (Region Representation Learning)** *Given the attribute features  $\mathcal{A}$ , outflow features  $\mathcal{S}$ , and inflow features  $\mathcal{D}$  of  $n$  regions, our objective is to acquire a collection of low-dimensional embeddings  $\mathcal{E} = \{\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_n\}$ , to serve as the latent representation for each region.*

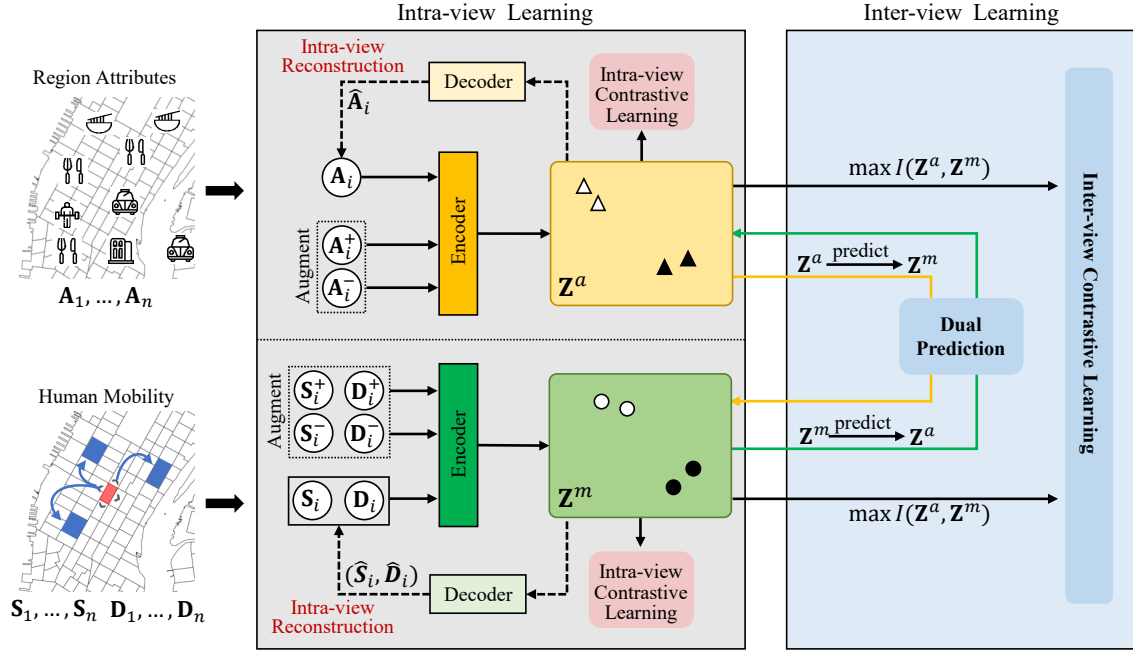


Figure 2: The framework of ReCP.

## Methodology

The framework of ReCP is illustrated in Figure 2, which includes two major components: 1) intra-view learning: for both region attribute and human mobility view, it captures the representative features of each region by intra-view contrastive learning to learn view-specific representations. Additionally, feature reconstruction is designed within each view to recover the original feature of the region, which helps avoid a trivial solution; 2) inter-view learning: within the same region, it integrates representations from different views through two learning objectives: inter-view contrastive learning is used to enhance the consistency across different views, and dual prediction is introduced to further diminish the inconsistent information between views.

### Intra-view Learning

Initially, we learn view-specific region representations based on the region attribute features  $\mathcal{A}$  and the mobility features  $\mathcal{S}$  and  $\mathcal{D}$ , respectively. Within each view, we learn the latent representation for each region by employing intra-view contrastive learning, i.e., we compare each region with others to highlight distinctive features within each region. Additionally, we design a within-view reconstruction loss to avoid the trivial solution.

**Intra-view Contrastive Learning** To learn region representations within each view, we design an intra-view contrastive learning module, which compares each region with others. For a given region  $r_i$ , we have three types of region features, including the attribute feature  $\mathbf{A}_i$ , outflow feature  $\mathbf{S}_i$ , and inflow feature  $\mathbf{D}_i$ . For simplicity, let  $\mathbf{X}_i^v$  denote the raw feature for the  $v$ -th view. For a target region  $r_i$ , we define its positive set as  $\mathcal{P}_i^v = \{\mathbf{X}_1^v, \mathbf{X}_2^v, \dots, \mathbf{X}_K^v\}$ , where

$\mathbf{X}_1^v, \mathbf{X}_2^v, \dots, \mathbf{X}_K^v$  are positive samples obtained through the data augmentation function following (Zhang, Long, and Cong 2022), and  $K$  is the number of positive samples. The negative set  $\mathcal{N}_i^v$  is defined as  $\mathcal{N}_i^v = \{\mathbf{X}_t^v | t \neq i\}$ , which contains features of regions except  $r_i$ .

We then map the raw features of regions into a latent representation,

$$\mathbf{Z}_i^v = E^{(v)}(\mathbf{X}_i^v), \quad (1)$$

where  $E^{(v)}$  denotes the encoder for the  $v$ -th view. In practice, we simply implement it as a fully connected neural network. As a result, we obtain three types of region representations,  $\mathbf{Z}_i^a$ ,  $\mathbf{Z}_i^s$  and  $\mathbf{Z}_i^d$ . Further, we compute the region representation  $\mathbf{Z}_i^m$  of the human mobility view as the average of  $\mathbf{Z}_i^s$  and  $\mathbf{Z}_i^d$ , i.e.,  $\mathbf{Z}_i^m = (\mathbf{Z}_i^s + \mathbf{Z}_i^d) / 2$ . To maximize the similarity of positive pairs while minimizing the similarity of negative pairs, the contrastive learning loss for the  $v$ -th view is defined as follows,

$$\mathcal{L}_{cl}^v = \sum_{r_i \in \mathcal{R}} \left[ -\log \sum_{k=1}^K \exp\left(\frac{\mathbf{Z}_i^v \cdot \mathbf{Z}_k^v}{\tau}\right) + \log\left(\sum_{k=1}^K \exp\left(\frac{\mathbf{Z}_i^v \cdot \mathbf{Z}_k^v}{\tau}\right) + \sum_{t=1}^{|\mathcal{N}_i^v|} \exp\left(\frac{\mathbf{Z}_i^v \cdot \mathbf{Z}_t^v}{\tau}\right)\right) \right], \quad (2)$$

where  $\tau$  is the temperature parameter and  $\mathcal{R}$  is the set of regions. Further, the intra-view contrastive learning loss across all views is formulated as

$$\mathcal{L}_{cl}^{intra} = \mu \mathcal{L}_{cl}^a + \mathcal{L}_{cl}^m. \quad (3)$$

where  $\mu$  is the parameter controlling the balance between the attribute view and the mobility view.

**Intra-view Reconstruction** Given the feature  $\mathbf{X}_i^v$  for the  $v$ -th view of the region  $r_i$ , we further optimize the latent region representations via an autoencoder and define the reconstruction loss  $\mathcal{L}_{rec}^v$  as

$$\mathcal{L}_{rec}^v = \sum_{r_i \in \mathcal{R}} \left\| \mathbf{X}_i^v - D^{(v)}(E^{(v)}(\mathbf{X}_i^v)) \right\|_2^2, \quad (4)$$

where  $E^{(v)}$  is the same as that in Equation (1) and  $D^{(v)}$  is the decoder for the  $v$ -th view to reconstruct the region features. Specifically, we employ a fully connected network to implement  $D^{(v)}$ , which shares the same number of layers and hidden sizes as  $E^{(v)}$ . Note that the autoencoder structure is helpful to avoid the trivial solution.

The total reconstruction loss across all views is

$$\mathcal{L}_{rec}^{intra} = \mu \mathcal{L}_{rec}^a + \mathcal{L}_{rec}^m, \quad (5)$$

where  $\mu$  is the same weight parameter as that in Equation (3). So far, we obtain two types of view-specific region representations  $\mathbf{Z}_i^a$  and  $\mathbf{Z}_i^m$  from the region attribute and human mobility views.

### Inter-view Learning

Different views of a region provide valuable information for describing the region, often offering complementary insights. To learn consistent and informative representations across different views, we employ inter-view contrastive learning to improve collaboration and information exchange between the views, achieved by maximizing the mutual information among different views. Additionally, dual prediction between two views is leveraged to reduce the impact of inconsistent information between the views by minimizing the conditional entropy across them.

**Inter-view Contrastive Learning** In the latent embedding space, we conduct contrastive learning to learn consistent representations shared across different views, as recent contrastive learning studies (He et al. 2020; Lin et al. 2021) have shown that consistency could be learned by maximizing the mutual information of different views. Formally, given the two representations  $\mathbf{Z}_i^a$  and  $\mathbf{Z}_i^m$  of region  $r_i$ , we maximize the mutual information between  $\mathbf{Z}_i^a$  and  $\mathbf{Z}_i^m$  from different views:

$$\mathcal{L}_{cl}^{inter} = - \sum_{r_i \in \mathcal{R}} [I(\mathbf{Z}_i^a, \mathbf{Z}_i^m) + \alpha(H(\mathbf{Z}_i^a) + H(\mathbf{Z}_i^m))], \quad (6)$$

where  $I(\cdot)$  represents mutual information,  $H(\cdot)$  denotes information entropy, and the parameter  $\alpha$  controls the balance between mutual information and information entropy. Note that the maximization of  $H(\mathbf{Z}_i^a)$  and  $H(\mathbf{Z}_i^m)$  also helps prevent trivial solutions where all regions are represented by the same representation. Based on the definition of mutual information,  $I(\cdot)$  is defined as

$$I(\mathbf{Z}_i^a, \mathbf{Z}_i^m) = P(\mathbf{Z}_i^a, \mathbf{Z}_i^m) \log \frac{P(\mathbf{Z}_i^a, \mathbf{Z}_i^m)}{P(\mathbf{Z}_i^a)P(\mathbf{Z}_i^m)}, \quad (7)$$

where  $P(\mathbf{Z}_i^a, \mathbf{Z}_i^m)$  represents the joint probability distribution of  $\mathbf{Z}_i^a$  and  $\mathbf{Z}_i^m$ . To represent the joint probability distribution, we employ a softmax function to transform the region representations  $\mathbf{Z}_i^a \in \mathbb{R}^d$  and  $\mathbf{Z}_i^m \in \mathbb{R}^d$  (where  $d$  is the

dimension of region representations) with

$$\mathbf{B}_i^a = \text{softmax}(\mathbf{Z}_i^a), \mathbf{B}_i^m = \text{softmax}(\mathbf{Z}_i^m), \quad (8)$$

where  $\mathbf{B}_i^a \in \mathbb{R}^d$  and  $\mathbf{B}_i^m \in \mathbb{R}^d$  can be interpreted as the probability distributions. Considering the entire set  $\mathcal{R}$  containing  $n$  regions, we define the matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$  as the joint probability distribution of  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$ ,

$$\mathbf{M} = \frac{1}{n} \sum_{i=1}^n \mathbf{B}_i^a (\mathbf{B}_i^m)^T. \quad (9)$$

We denote the element located at the  $r$ -th row and the  $r'$ -th column of the matrix as  $\mathbf{M}_{rr'}$ , and the sum of the elements in matrix  $\mathbf{M}$  along the  $r$ -th row (the  $r'$ -th column) as  $\mathbf{M}_r$  ( $\mathbf{M}_{r'}$ ).  $\mathbf{M}_{rr'}$  represents the joint probability, while  $\mathbf{M}_r$  and  $\mathbf{M}_{r'}$  represent the marginal probability, respectively. Then we could compute the mutual information  $I(\mathbf{Z}^a, \mathbf{Z}^m)$  as follows,

$$I(\mathbf{Z}^a, \mathbf{Z}^m) = \sum_{r=1}^d \sum_{r'=1}^d \mathbf{M}_{rr'} \log \frac{\mathbf{M}_{rr'}}{\mathbf{M}_r \cdot \mathbf{M}_{r'}}. \quad (10)$$

Information entropy  $H(\mathbf{Z}_i^v)$  is defined as follows,

$$H(\mathbf{Z}_i^v) = -P(\mathbf{Z}_i^v) \log P(\mathbf{Z}_i^v), \quad (11)$$

where  $v \in \{a, m\}$ . Following the above definition of  $\mathbf{M}$ ,  $H(\mathbf{Z}_i^v)$  could be computed as

$$\begin{aligned} H(\mathbf{Z}^a) &= - \sum_{r=1}^d \mathbf{M}_r \log \mathbf{M}_r, \\ H(\mathbf{Z}^m) &= - \sum_{r'=1}^d \mathbf{M}_{r'} \log \mathbf{M}_{r'}. \end{aligned} \quad (12)$$

Combining Equations (6), (10), and (12), the inter-view contrastive learning loss is formulated as

$$\mathcal{L}_{cl}^{inter} = - \sum_{r=1}^d \sum_{r'=1}^d \mathbf{M}_{rr'} \ln \frac{\mathbf{M}_{rr'}}{\mathbf{M}_r^{\alpha+1} \cdot \mathbf{M}_{r'}^{\alpha+1}}. \quad (13)$$

where  $\alpha$  is the weight parameter defined in the Equation (6).

**Inter-view Dual Prediction** To further diminish the inconsistency across different views, we predict the view-specific region representation by minimizing the conditioned entropy. Formally, given the region representations  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$ , we minimize the conditional entropy  $H(\mathbf{Z}^p | \mathbf{Z}^q)$ , where  $p = a, q = m$  or  $p = m, q = a$ . On one hand,  $\mathbf{Z}^q$  contains nearly all the information required to represent the  $p$ -th view if  $\mathbf{Z}^q$  can perfectly predict  $\mathbf{Z}^p$  for any  $(\mathbf{Z}^p, \mathbf{Z}^q) \sim P_{\mathbf{Z}^p, \mathbf{Z}^q}$ . On the other hand,  $\mathbf{Z}^q$  diminishes the inconsistent information within the  $q$ -th view if  $\mathbf{Z}^p$  can perfectly predict  $\mathbf{Z}^q$  under the constraint where  $I(\mathbf{Z}^p, \mathbf{Z}^q)$  is maximized. Mathematically,  $H(\mathbf{Z}^p | \mathbf{Z}^q)$  is defined as

$$H(\mathbf{Z}^p | \mathbf{Z}^q) = -\mathbb{E}_{P_{\mathbf{Z}^p, \mathbf{Z}^q}} [\log P(\mathbf{Z}^p | \mathbf{Z}^q)]. \quad (14)$$

To minimize  $H(\mathbf{Z}^p | \mathbf{Z}^q)$ , a common approach is to assume a variational distribution  $Q(\mathbf{Z}^p | \mathbf{Z}^q)$  for  $\mathbf{Z}^p$  and  $\mathbf{Z}^q$ . Specially, we present to maximize  $\mathbb{E}_{P_{\mathbf{Z}^p, \mathbf{Z}^q}} [\log Q(\mathbf{Z}^p | \mathbf{Z}^q)]$ ,

which serves as a lower bound of  $\mathbb{E}_{P_{\mathbf{Z}^p, \mathbf{Z}^q}} [\log P(\mathbf{Z}^p | \mathbf{Z}^q)]$ .  $Q(\cdot | \cdot)$  can be any distribution such as Gaussian or Laplacian. In this work, we simply adopt the Gaussian distribution  $N(\mathbf{Z}^p | F^{(q)}(\mathbf{Z}^q), \sigma \mathbf{I})$ , where  $F^{(q)}(\cdot)$  represents a parameterized function mapping  $\mathbf{Z}^q$  to  $\mathbf{Z}^p$ , and  $\sigma \mathbf{I}$  denotes the variance matrix. By ignoring the constants derived from the Gaussian distribution, maximizing  $\mathbb{E}_{P_{\mathbf{Z}^p, \mathbf{Z}^q}} [\log Q(\mathbf{Z}^p | \mathbf{Z}^q)]$  is equivalent to minimizing

$$\mathbb{E}_{P_{\mathbf{Z}^p, \mathbf{Z}^q}} \left\| \mathbf{Z}^p - F^{(q)}(\mathbf{Z}^q) \right\|_2^2. \quad (15)$$

Then the dual prediction loss can be formulated as

$$\mathcal{L}_{dp}^{inter} = \sum_{r_i \in \mathcal{R}} \left\| \mathbf{Z}_i^m - F^{(a)}(\mathbf{Z}_i^a) \right\|_2^2 + \left\| \mathbf{Z}_i^a - F^{(m)}(\mathbf{Z}_i^m) \right\|_2^2.$$

Here,  $F^{(a)}$  and  $F^{(m)}$  are respectively implemented as fully-connected networks, with each layer followed by a batch normalization layer and a ReLU layer. Note that the above loss may lead to model collapse without the intra-view reconstruction loss (Equation (4)), i.e.,  $\mathbf{Z}_i^a$  and  $\mathbf{Z}_i^m$  from different views become equivalent to the same constant.

Finally, the inter-view learning loss is defined as

$$\mathcal{L}_{inter} = \mathcal{L}_{dp}^{inter} + \mathcal{L}_{cl}^{inter}. \quad (16)$$

## Model Training

The final objective function is defined as

$$\mathcal{L} = \mathcal{L}_{inter} + \lambda_1 \mathcal{L}_{cl}^{intra} + \lambda_2 \mathcal{L}_{rec}^{intra}, \quad (17)$$

where  $\lambda_1$  and  $\lambda_2$  are parameters controlling the weights of different losses. After learning the latent representations  $\mathbf{Z}^a$  and  $\mathbf{Z}^m$ , we simply concatenate them as the final multi-view region representation, i.e.,  $\mathbf{E}_i = \mathbf{Z}_i^a \parallel \mathbf{Z}_i^m$ .

## Experiments

We start by presenting the experimental settings, followed by the evaluation of the learned region representations on two popular downstream tasks: land use clustering and region popularity prediction.

### Experimental Settings

**Datasets.** We collect a diverse set of real-world data from NYC Open Data<sup>1</sup> and use the Manhattan borough as the study area. We partition Manhattan into 270 regions based on the city boundaries designed by the US Census Bureau<sup>2</sup>. As for the human mobility data, we employ complete taxi trip records from February 2014 as our training data. We utilize the NYC check-in and POI data provided by (Yang et al. 2014) for our model training and the popularity prediction task. The detailed description of datasets is shown in Table 1. Based on these data, we construct the region features including  $\mathcal{A}$ ,  $\mathcal{S}$ , and  $\mathcal{D}$  for model training.

**Model Parameters.** In our experiments, the dimension of region representations is set to 96. In the intra-view reconstruction module, we set the number of layers at 3 and the

Table 1: Data description (K=10<sup>3</sup>, M=10<sup>6</sup>).

Dataset	Description
Regions	270 regions divided by streets in Manhattan
Taxi trips	10M taxi trips during February, 2014
POI data	10K POIs with 244 categories
Check-in data	100K check-in records

hidden size at 128 for the encoder  $E^{(v)}$  and decoder  $D^{(v)}$ ; in the intra-view contrastive learning module, following the settings in (Zhang, Long, and Cong 2022), we set the number of positive samples for region attribute and human mobility data at 3 and 4, and the parameter  $\mu$  controlling the balance between different views at 0.0001. In the inter-view dual prediction module, we set the number of layers at 3 and the hidden size at 96 for  $F^{(a)}$  and  $F^{(m)}$ ; in the inter-view contrastive learning module, we set the parameter  $\alpha$  at 9. We set the hyper-parameters  $\lambda_1$  and  $\lambda_2$  in the final objective loss at 1. Note that the optimal model parameters are selected using grid search with a small but adaptive step size. To optimize our model, we adopt Adam and initialize the learning rate at 0.01 with a linear decay.

**Baselines.** We compare the performance of ReCP with several state-of-the-art region embedding methods.

- **HDGE.** (Wang and Li 2017) constructs flow graphs and spatial graphs using taxi data and learns region representations with graph embedding methods.
- **ZE-Mob.** (Yao et al. 2018) models co-occurrence patterns between regions from mobility data to learn region representations.
- **MV-PN.** (Fu et al. 2019) models both inter-region and intra-region information to construct multi-view POI-POI networks within each region.
- **CGAL.** (Zhang et al. 2019) extends MV-PN and incorporates the spatial structure and spatial autocorrelation among regions to learn region representations.
- **MVURE.** (Zhang et al. 2021) learns region representations by cross-view information sharing and multi-view fusion with human mobility and region attributes.
- **MGFN.** (Wu et al. 2022) designs multi-level cross-attention mechanisms to extract region representations from multiple mobility patterns.
- **ReMVC.** (Zhang, Long, and Cong 2022) learns region representations through both intra-view and inter-view contrastive learning modules.
- **HREP.** (Zhou et al. 2023) constructs heterogeneous graphs and uses relation-aware graph embedding to learn region representations.

### Land Usage Clustering

We use the district division by the community boards (Berg 2007) as ground truth and divide the Manhattan borough into 29 districts, following the settings in (Zhang, Long, and Cong 2022). We cluster regions into groups by  $k$ -means

<sup>1</sup><https://opendata.cityofnewyork.us>

<sup>2</sup><https://www.census.gov/data.html>

Table 2: Performance comparison on two downstream tasks, where the performance improvements of ReCP are compared with the best of these baseline methods, marked by the asterisk.

Method	Land Usage Clustering			Region Popularity Prediction		
	NMI	ARI	F-measure	MAE	RMSE	R <sup>2</sup>
HDGE	0.469 ± 0.01	0.095 ± 0.01	0.117 ± 0.01	334.43 ± 10.17	474.94 ± 9.49	0.079 ± 0.04
ZE-Mob	0.437 ± 0.02	0.071 ± 0.01	0.097 ± 0.01	282.42 ± 13.71	418.02 ± 12.69	0.286 ± 0.04
MV-PN	0.407 ± 0.01	0.036 ± 0.01	0.070 ± 0.01	291.17 ± 16.54	435.23 ± 16.52	0.226 ± 0.06
CGAL	0.414 ± 0.08	0.059 ± 0.06	0.091 ± 0.06	351.10 ± 51.20	486.96 ± 52.58	0.021 ± 0.20
MVURE	0.735 ± 0.01	0.400 ± 0.02	0.415 ± 0.02	236.25 ± 7.86	347.01 ± 11.70	0.508 ± 0.03
MGFN	0.748 ± 0.01	0.424 ± 0.03	0.437 ± 0.03	240.37 ± 11.99	354.24 ± 17.14	0.487 ± 0.05
ReMVC	0.761* ± 0.02	0.455* ± 0.04	0.462* ± 0.04	283.02 ± 18.03	406.25 ± 18.00	0.325 ± 0.06
HREP	0.757 ± 0.01	0.448 ± 0.03	0.457 ± 0.03	217.52* ± 10.98	318.41* ± 14.54	0.585* ± 0.04
<b>ReCP</b>	<b>0.780 ± 0.01</b>	<b>0.483 ± 0.01</b>	<b>0.499 ± 0.02</b>	<b>195.16 ± 18.70</b>	<b>291.19 ± 20.04</b>	<b>0.652 ± 0.05</b>
<b>Improvements</b>	<b>2.50%</b>	<b>6.15%</b>	<b>8.01%</b>	<b>10.28%</b>	<b>8.55%</b>	<b>11.45%</b>

clustering ( $k = 29$ ), using region representations as inputs. The regions with the same land usage type are expected to be assigned to the same cluster. The experimental results are evaluated using three metrics: Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F-measure following (Yao et al. 2018; Zhang et al. 2021). We assess all the methods using the same dataset and conduct 10 runs to report the mean value with the standard deviation in Table 2. From the results, we observe that:

- HDGE and ZE-Mob exhibit relatively inferior performance as they merely model co-occurrence patterns using human mobility data. MGFN demonstrates better performance than HDGE and ZE-Mob, as it designs a deep model based on cross-attention mechanisms to capture complex mobility patterns from spatial-temporal human mobility data.
- The methods that model multi-view information generally achieve satisfactory results, validating the importance of effectively integrating multi-view information for region embedding. Specifically, MV-PN and CGAL exhibit poor performance as they simply combine region representations from two views, lacking the deep interaction between views; MVURE and HREP design attention-based mechanisms to fuse the multi-view information, consequently yielding superior performance; ReMVC adopts contrastive learning to model intra-view and inter-view information and also obtains good results.
- The proposed ReCP outperforms all these baselines, as it explores the consistency across different views in region embedding. Compared with ReMVC, ReCP achieves average improvements of 2.50%, 6.15%, and 8.01% in terms of NMI, ARI, and F-measure, respectively. Moreover, the results of the superiority paired t-test indicate that the improvement of ReCP over the baselines is statistically significant, with a  $p$ -value less than 0.01.

### Region Popularity Prediction

Another commonly-compared downstream task to evaluate the region representations is popularity prediction, where we aggregate the check-in counts within each region as the ground truth of popularity following (Yang et al. 2014;

Zhang, Long, and Cong 2022). We take region representations as input and train the Lasso regression model. The evaluation results including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R<sup>2</sup>) are obtained by 5-fold cross-validation, as reported in Table 2. From the results, we observe that the multi-view fusion methods including MVURE and HREP achieve decent performance, which further validates the necessity of integrating multi-view information in region embedding. ReCP performs the best among all methods, e.g., compared to HREP, ReCP achieves average improvements of 10.28%, 8.55%, and 11.45% in terms of MAE, RMSE, and R<sup>2</sup>. These results indicate that it is an effective way to learn better region representations by utilizing the new pipeline following the consistency learning paradigm.

### Ablation Study and Parameter Analysis

**Ablation study** We design four variants to explore how each module of ReCP affects the clustering and regression performance:

- ReCP w/o CL: we remove the intra-view contrastive learning loss.
- ReCP w/o Rec: we remove the intra-view reconstruction loss and only use the encoder to extract features.
- ReCP w/o IV: we remove the inter-view learning module and directly concatenate region representations from the two views without the constraint of consistency learning.
- ReCP w/o DP: we remove the inter-view dual prediction loss.

From the results in Figure 3, we observe that:

1) ReCP w/o CL achieves the lowest performance in both tasks, indicating that the intra-view contrastive learning loss is a crucial component in our model for learning view-specific feature representations of regions.

2) ReCP w/o Rec achieves worse performance than ReCP, supporting the aforementioned claim that the intra-view reconstruction loss could help prevent the model from converging to a trivial solution.

3) ReCP demonstrates an improvement of 29.84% (in terms of ARI) and 4.00% (in terms of R<sup>2</sup>) when compared to

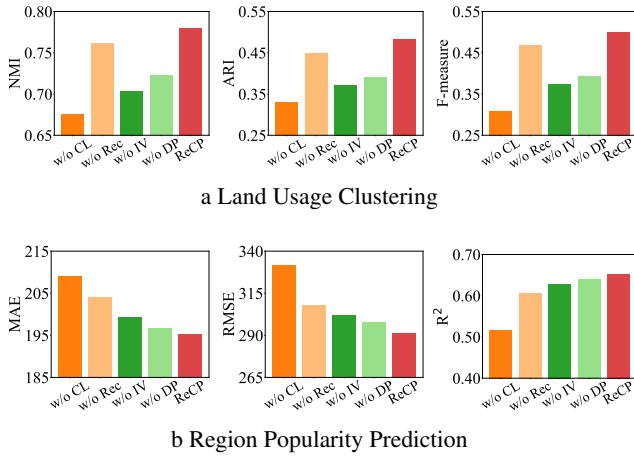


Figure 3: Performance comparison of different modules.

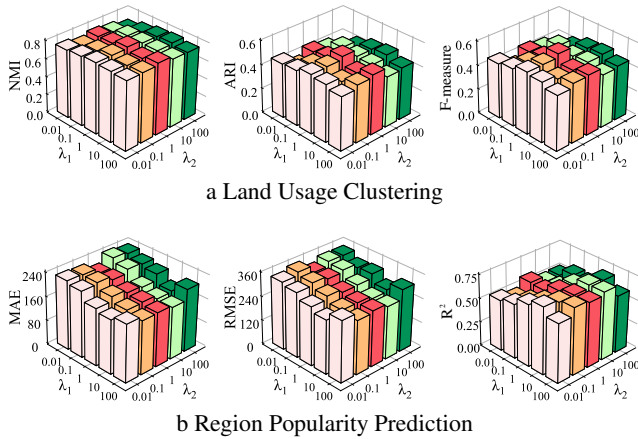


Figure 4: Parameter analysis on both downstream tasks.

ReCP w/o IV. This finding suggests that the proposed inter-view learning module effectively leverages the multi-view information and highlights the importance of consistency learning across different views.

4) ReCP w/o DP outperforms ReCP w/o IV but performs worse than ReCP, indicating that both the inter-view contrastive learning loss (which maximizes the mutual information between views) and the inter-view dual prediction loss (which minimizes the conditional entropy across them) are important for learning multi-view region representations.

**Parameter sensitivity** The parameters  $\lambda_1$  and  $\lambda_2$  govern the weighting of various losses. We vary their values within the range of  $\{0.01, 0.1, 1, 10, 100\}$  to assess the impact of  $\lambda_1$  and  $\lambda_2$  on the model performance. As depicted in Figure 4, ReCP achieves satisfactory performance when we set both  $\lambda_1$  and  $\lambda_2$  at 1.

## Related Work

Traditional methods for region embedding typically utilize human mobility data to analyze the transition patterns between urban regions. These methods are often based on the

word2vec framework and learn the latent representations of regions (Wang and Li 2017; Yao et al. 2018). In a similar vein, Wu et al. (Wu et al. 2022) incorporate mobility graphs with spatio-temporal similarity as mobility patterns, and propose multi-level cross-attention mechanisms to extract comprehensive region representations from these patterns. Additionally, some studies focus on leveraging the inherent attributes of regions to learn latent representations. For instance, Zhang et al. (Zhang et al. 2019) construct multiple spatial graphs to represent the geographic structure of regions. By transforming the region embedding problem into a graph embedding problem, they primarily capture the spatial structure within regions and the spatial autocorrelation between regions. Another approach, proposed by Wang et al. (Wang, Li, and Rajagopal 2020), involves mining street views and textual information of POIs within regions to learn representations.

Moreover, there have been studies that learn region representations by incorporating both attribute features within regions and mobility patterns between regions. For instance, Fu et al. (Fu et al. 2019) propose an autoencoder framework that effectively captures inter-region correlations and intra-region structural information during the process of region embedding. Zhang et al. (Zhang et al. 2021) model multi-view region correlations by leveraging human mobility data and inherent region attributes, and employ a graph attention mechanism to acquire region representations from each view of the established correlations. Zhou et al. (Zhou et al. 2023) learn relation-specific region representations from various types of relations in a heterogeneous graph constructed using human mobility, POI data, and geographic neighbors of regions. They devise an attention-based fusion technique to integrate shared information among different types of correlations. Additionally, Zhang et al. (Zhang, Long, and Cong 2022) introduce a multi-view region embedding model based on contrastive learning, which incorporates an intra-view contrastive learning module to discern distinct representations and an inter-view contrastive learning module to facilitate the transfer of knowledge across multiple views.

## Conclusion

In this paper, we form a new pipeline based on the consistency learning paradigm for multi-view region embedding. Under the hood, we propose a multi-view Contrastive Prediction model for urban Region embedding (ReCP) by exploring the consistency across two views, leveraging both POI and human mobility data. The ReCP model consists of two modules: an intra-view learning module that utilizes contrastive learning and feature reconstruction which learn region representations specific to each view, and an inter-view learning module utilizing a contrastive prediction learning scheme that enhances the consistency between two views. To evaluate the effectiveness of our proposed model, we conduct comprehensive experiments on two downstream tasks: land use clustering and region popularity prediction. The experimental results demonstrate that the proposed ReCP model outperforms state-of-the-art embedding methods, proving that retaining consistency across views is pivotal for effective region embedding.

## References

- Berg, B. F. 2007. *New York City Politics: Governing Gotham*. Rutgers University Press.
- Chen, M.; Yu, X.; and Liu, Y. 2018. PCNN: Deep convolutional networks for short-term traffic congestion prediction. *IEEE Transactions on Intelligent Transportation Systems*, 19(11): 3550–3559.
- Fu, Y.; Wang, P.; Du, J.; Wu, L.; and Li, X. 2019. 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, 906–913.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.
- Huang, W.; Zhang, D.; Mai, G.; Guo, X.; and Cui, L. 2023. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 134–145.
- Li, T.; Xin, S.; Xi, Y.; Tarkoma, S.; Hui, P.; and Li, Y. 2022. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3282–3291.
- Li, Y.; Huang, W.; Cong, G.; Wang, H.; and Wang, Z. 2023. Urban Region Representation Learning with OpenStreetMap Building Footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1363–1373.
- Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11174–11183.
- Liu, C.; Yang, Y.; Yao, Z.; Xu, Y.; Chen, W.; Yue, L.; and Wu, H. 2021. Discovering urban functions of high-definition zoning with continuous human traces. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1048–1057.
- Liu, Y.; Zhang, X.; Ding, J.; Xi, Y.; and Li, Y. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, 4150–4160.
- Luo, Y.; Chung, F.-I.; and Chen, K. 2022. Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4294–4298.
- Tsai, Y.-H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2021. Self-supervised learning from a multi-view perspective. In *Proceedings of the International Conference on Learning Representations, 2021*.
- Wang, H.; and Li, Z. 2017. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 237–246.
- Wang, Z.; Li, H.; and Rajagopal, R. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1013–1020.
- Wu, S.; Yan, X.; Fan, X.; Pan, S.; Zhu, S.; Zheng, C.; Cheng, M.; and Wang, C. 2022. Multi-graph fusion networks for urban region embedding. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2312–2318.
- Xu, R.; Chen, M.; Gong, Y.; Liu, Y.; Yu, X.; and Nie, L. 2023a. TME: Tree-guided multi-task embedding learning towards semantic venue annotation. *ACM Transactions on Information Systems*, 41(4).
- Xu, R.; Huang, W.; Zhao, J.; Chen, M.; and Nie, L. 2023b. A spatial and adversarial representation learning approach for land use classification with POIs. *ACM Transactions on Intelligent Systems and Technology*, 14(6): 1–25.
- Yang, D.; Zhang, D.; Zheng, V. W.; and Yu, Z. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1): 129–142.
- Yao, Z.; Fu, Y.; Liu, B.; Hu, W.; and Xiong, H. 2018. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.
- Zhang, C.; Zhao, K.; and Chen, M. 2022. Beyond the limits of predictability in human mobility prediction: context-transition predictability. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4514–4526.
- Zhang, D.; Xu, R.; Huang, W.; Zhao, K.; and Chen, M. 2023. Towards an integrated view of semantic annotation for POIs with spatial and textual information. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2441–2449.
- Zhang, L.; Long, C.; and Cong, G. 2022. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*.
- Zhang, M.; Li, T.; Li, Y.; and Hui, P. 2021. 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*, 4431–4437.
- Zhang, Y.; Fu, Y.; Wang, P.; Li, X.; and 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*, 1700–1708.
- Zhou, S.; He, D.; Chen, L.; Shang, S.; and Han, P. 2023. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4981–4989.