Region Embedding with Intra and Inter-View Contrastive Learning

Liang Zhang, Cheng Long, and Gao Cong

Abstract—Unsupervised region representation learning aims to extract dense and effective features from unlabeled urban data. While some efforts have been made for solving this problem based on multiple views, existing methods are still insufficient in extracting representations in a view and/or incorporating representations from different views. Motivated by the success of contrastive learning for representation learning, we propose to leverage it for multi-view region representation learning and design a model called ReMVC (Region Embedding with Multi-View Contrastive Learning) by following two guidelines: i) comparing a region with others within each view for effective representation extraction and ii) comparing a region with itself across different views for cross-view information sharing. We design the intra-view contrastive learning module which helps to learn distinguished region embeddings and the inter-view contrastive learning module which serves as a soft co-regularizer to constrain the embedding parameters and transfer knowledge across multi-views. We exploit the learned region embeddings in two downstream tasks named land usage clustering and region popularity prediction. Extensive experiments demonstrate that our model achieves impressive improvements compared with seven state-of-the-art baseline methods, and the margins are over 30% in the land usage clustering task.

Index Terms—Contrastive learning, region representation, multi-view representation, urban computing

1 Introduction

As basic elements of a city, regions are areas where people live, work, and entertain. Studying region representations can help us explore the functions and properties of cities, which is beneficial for many downstream applications such as land usage clustering [1] and popularity prediction [2], and ultimately makes our cities smarter. In recent years, with the pervasiveness of mobile sensing technologies, a large amount of urban data such as Point-Of-Interests (POIs) and human mobility records become available. They provide great opportunities for addressing the region representation learning problem jointly from multi-views, where one view called the *POI view* is based on POIs and the other one called the *mobility view* is based on human mobility records. Many existing studies have been conducted on the region representation learning task based on these two views [2], [3], [1], [4], [5]. They typically tackle the problem following two steps: intra-view feature extraction and multi-view embedding fusion.

For intra-view feature extraction, two strategies have been studied. i) One is to construct a region graph based on region correlations and learn representations with graph embedding methods [1], [4]. However, the graph is usually defined based on heuristic measures, which encode the region features manually instead of a learnable encoder. ii) The other strategy addresses the mentioned shortcomings to some extent by learning a parameterized encoder to reconstruct the region features [3]. Nevertheless, the feature reconstruction strategy may focus on region details with little representative information. For example, considering a case where two regions both contain several residential buildings but serve as different functions reflected by some discriminate POIs such as bars and parks. With perfect POI reconstructions, these residential POIs would dominate the model learning and then generate similar representations for these two regions (with different urban functions).

For the multi-view embedding fusion, existing works exploit either concatenation [2], [3] or weighted average strategies [1]. However, the observations from different views should be highly correlated for the same region while less correlated for different regions, and the existing strategies can not preserve the cross-view correlations explicitly.

Recently, the contrastive learning method has achieved great success in computer vision and natural language processing domains, demonstrating its superiority over other methods [6]. Motivated by its success, we propose to learn region representations by introducing comparison, which can help to capture abstract features and complex structures. Principally, a region may include some distinguished properties such as special POIs, which should be emphasised for intra-view feature extraction. We can capture this aspect easily by adopting the first guideline which is to *compare a region with others within each view*.

Furthermore, for a region, the multi-view data is generated due to the intrinsic region properties. For example, an entertainment region often includes certain POI types such as bars (in the POI view) and specific human mobility patterns such as large-scale flows in the evening (in the mobility view). Intuitively, only reasonable representations

• *L. Zhang, C. Long and G. Cong are with the Department of Computer Science, Nanyang Technological University, Singapore. E-mail: liang012@e.ntu.edu.sg, {c.long, gaocong}@ntu.edu.sg*
Corresponding author: Cheng Long

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). This research is also supported by the Ministry of Education, Singapore, under its Academic Research Fund (Tier 2 Award MOE-T2EP20221-0013). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

from multi-views can match each other. Therefore, we propose another guideline for multi-view embedding fusion which is to *compare a region with itself across different views*.

By following the two guidelines, in this paper, we propose a general representation learning framework called ReMVC (Region Embedding with Multi-View Contrastive Learning) for region embedding based on two views, namely the POI view and the mobility view. For each view, we design the *intra-view contrastive learning* module to extract region representations by maximizing the discrepancy between a region and other dissimilar regions. Furthermore, we design the *inter-view contrastive learning* module to preserve the proposed cross-view correlations between different views, and the feature learning in one view can constrain the learning in the other view. By using one view as the label for another, our model can also be regarded as transferring knowledge along bi-directions. Each view can learn how to represent the region by treating another view as the reference. In summary, we make the following contributions:

- We design a multi-task learning framework called ReMVC which mainly consists of (1) an intra-view contrastive learning module for discriminate representations extraction; (2) and an inter-view contrastive learning module for inter-view cooperation and information sharing. In fact, we are aware of no existing methods which use contrastive learning for both intra and inter-view representation learning.
- We conduct extensive experiments and the results demonstrate that our model outperforms existing methods on two downstream tasks, and the margins are over 30% in the land usage clustering task. Our codes and data are available online [1].

2 Related Work

Region Representation Learning. Recently, a few studies propose to learn generic region representations from unlabeled urban data. To this end, the geo-tagged social media data and satellite images are exploited firstly. The study in [7] exploits topic models to process twitter texts and the extracted topics can serve as region representation. Following studies extend the idea by using deep learning methods. For instance, TNE [8] defines a graph learning framework for geo-tagged social media data. Tile2Vec [9] regards a region as satellite images and learns region embeddings with image representation learning methods. M2G [10] proposes a multi-task learning framework based on both geo-tagged check-ins and satellite images.

Other studies propose approaches that merely use easily available data including human mobility patterns and POI attributes to learn representations of regions. Along this line, [4] and [11] construct transition graphs between regions based on human mobility patterns and use path sampling and AutoEncoder to extract region representations. CGAL [3] considers both mobility connectivity and geographic distance to construct POI-POI graphs for each region and develops an "encode-decode" based framework. [1] further extends CGAL by replacing the AutoEncoder with cross-view GNNs. However, these approaches follow the similar learning strategy by using a *reconstruction* framework and a simple embedding fusion method, which may lead to ineffective region representations, as explained in the introduction section. Different from these studies, we introduce the contrastive learning for both intra and inter-view to enhance discriminate embedding learning and multi-view cooperation.

Contrastive Learning. Contrastive learning is an emerging machine learning paradigm which aims to learn data representations form unlabeled raw data [12]. The typical contrastive models learn data representation through comparing similar sample pairs with dissimilar ones and adopting the Noise Contrastive Estimation (NCE) objective [13]. It is commonly used for visual and video feature extraction [14]. For example, simply asking the network to recognize the image orientation and invariance to data augmentation [12] can learn high-quality image representations which can be used in downstream tasks such as image classification. Nowadays, contrastive learning has also been applied to other domains such as graph embedding and point set representations [15]. For instance, based on the mutual information between each node and the local graph structure, pre-trained GNNs [16] can learn robust node and graph representations. In this paper, we propose a contrastive learning framework for region presentation learning based on POI data and mobility data, which is the first of it kind.

Multi-view Representation Learning. There is a rich literature of multi-view representation learning: most of the methods were designed for specific views of data such as images [17], graphs [18] and heterogeneous graphs [19], yet few is generally applicable across different views of data. These methods can be roughly categorized as *representation alignment methods* and *representation fusion methods* [20]: the former learn multi-view representations jointly with the alignment among different views used as regularizers/constraints and the latter learn representations of individual views separately and then fuse the representations as those for the multi-views.

Our method is a representation alignment method and differs from existing methods in: (1) it uses contrastive learning for both intra-view and inter-view representations learning for the first time; and (2) it is tailored for region data with POI view and mobility view.

3 Problem Formulation

Following existing studies [1], we partition the space into a set of L non-overlapping regions, denoted by $\mathcal{R} = \{r_1, r_2, ..., r_L\}$. In this paper, we consider two views for region representation learning, namely the POI view which is based on the POIs in a region and the mobility view which is based on human mobility records.

POI view. The region attributes are inherent geographic features of each region. A certain type of attribute can be represented as $\mathbf{f}_k$ for the k-th region, where $\mathbf{f}_k \in \mathbb{R}^F$ and F is the number of feature dimensions. In this paper, we use POI categories as region attributes by following [2], [1]. Each dimension in $\mathbf{f}_k$ denotes the ratio of POIs with the corresponding category in the k-th region.

Mobility view. We define the mobility flow data as a set $\mathcal{T} = \{t_1, t_2, ..., t_n\}$ of trips. Each trip t_i is a tuple consisting of the source and destination locations of the

1. https://github.com/Liang-NTU/ReMVC

Fig. 1: The architecture of ReMVC.

trip as well as the time information, which is denoted as (r_s, r_d, h). We map the source and destination to their corresponding regions. Then we can define two heat maps $\mathbf{MS}_k$ and $\mathbf{MD}_k \in \mathbb{R}^{H\times L}$ for the k-th region, where H is the number of time slices and L is the number of regions. The element $MS_k^{i,j}$ for a region r_k is defined as $MS_k^{i,j} = |\{(r_s, r_d, h)|h = i, r_s = r_j, r_d = r_k\}|$. Similarly, by specifying r_k as the source and other regions r_j as the destination, we can obtain another matrix $\mathbf{MD}_k$ for the k-th region.

Region representation learning problem. Given heat maps $\mathbf{MS}$, $\mathbf{MD}$ and region attributes $\mathbf{f}$ for regions in $\mathcal{R}$, the problem is to learn a latent embedding of each region, which is denoted as $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_L|\mathbf{e}_k \in \mathbb{R}^d\}$ with d as the embedding size.

4 Methodology

In this section, we propose a novel deep neural network called ReMVC (Region Embedding with Multi-View Contrastive Learning) to cope with the region representation learning problem. The overview of the model architecture is shown in Figure 1, which has two important modules: the *intra-view contrastive learning* to discriminate representations extracted from both POI and human mobility views, and the *inter-view contrastive learning* to deal with effective multi-view embedding fusion. Our key innovation lies in incorporating contrastive learning in both views by following two guidelines. The first guideline is to compare a region with other regions within each view, which is followed by the intra-view contrastive learning module. The second guideline is to compare a region with itself across different views, which is followed by the inter-view contrastive learning module.

4.1 Intra-view Contrastive Learning

Consider the POI view first. The POI features to be encoded are defined as $\mathbf{f}_k$ for the k-th region r_k. We define an augmentation function $\varphi(\mathbf{f}_k, a)$, where a is sampled from a set of data augmentation strategies $\mathcal{A}_p$. We design three POI level augmentations for $\mathcal{A}_p$: (1) Random Insertion: insert for each POI in region r_k a random POI with probability p; (2) Random Deletion: drop each POI with probability p; and (3) Random Replacement: replace each POI with a random one with probability p.

Then the positive region set $\mathcal{P}_k$ of r_k is defined as $\mathcal{P}_k = \{\mathbf{f}_{k,a}|a \in \mathcal{A}_p\}$ where the positive samples $\mathbf{f}_{k,a}$ are region augmentations $\varphi(\mathbf{f}_k, a)$, which is based on the principle that a region is similar to itself with some augmentations. The negative set $\mathcal{N}_k$ is defined as $\mathcal{N}_k = \{\mathbf{f}_n|n \neq k\}$ with $|\mathcal{N}_k| = N_p$, where N_p is the predefined negative size for the POI view and negative samples $\mathbf{f}_n$ are randomly sampled based on the POI feature distance distribution. For a candidate region r_n, its sampling probability is defined as the normalization of its POI feature distance from the anchor region r_k. The greater the POI feature distance, the more likely the region r_n is sampled as a negative sample for the POI view.

For the k-th region, we then design a POI encoder G^p to transform the coarse-grained POI features $\mathbf{f}_k$ to a latent embedding $\mathbf{z}_k^p$. In this paper, we exploit the most commonly used structure Multi-Layer Perceptions (MLPs) to implement G^p following the region representation studies in [2], [11] and [3]. Note that other feature extractors can possibly be used in our framework. Then, we have,

$$\mathbf{z}_k^p = G^p(\mathbf{f}_k) = \text{MLP}(\mathbf{f}_k). \tag{1}$$

Given the POI encoder G^p, we obtain $\mathbf{z}_k^p = G^p(\mathbf{f}_k)$ and $\mathbf{z}_{k,a}^p = G^p(\mathbf{f}_{k,a})$ where $\mathbf{z}_k^p$ and $\mathbf{z}_{k,a}^p$ denote the learned representations for the k-th region r_k in the POI view and its feature augmentation, respectively. We then follow the first guideline and contrast the two groups of embeddings by adopting the famous InfoNCE loss [21].

$$\mathcal{L}_{poi} = \sum_{r_k\in\mathcal{R}} \left[-\log \sum_{a\in\mathcal{A}_p} D^{ia}(\mathbf{z}_k^p, \mathbf{z}_{k,a}^p) + \log\left(\sum_{a\in\mathcal{A}_p} D^{ia}(\mathbf{z}_k^p, \mathbf{z}_{k,a}^p) + \sum_{n\in\mathcal{N}_k} D^{ia}(\mathbf{z}_k^p, \mathbf{z}_n^p) \right) \right]. \tag{2}$$

Here, the $\mathbf{z}_{k,a}^p$ is the positive embedding and $\mathbf{z}_n^p$ denotes the negative embedding. The function D^{ia} is an intra-view discriminator that takes two vectors as inputs and then scores the similarity between them. We simply implement it as the inner product as follows, where τ refers to the temperature parameter.

$$D^{ia}(\mathbf{z}_k^p, \mathbf{z}_{k,a}^p) = \exp\left(\frac{\mathbf{z}_k^p \cdot \mathbf{z}_{k,a}^p}{\tau}\right). \tag{3}$$

Similarly, we define the augmentation function $\varphi(\mathbf{MS}_k, \mathbf{MD}_k, a)$ and transformation set $\mathcal{A}_m$ for the mobility view, where $\mathcal{A}_m$ is implemented based on an operation called Noise Injection: add to each element in heatmaps $\mathbf{MS}_k$ and $\mathbf{MD}_k$ a Gaussian noise with mean 0 and variance σ. We then define a positive set $\mathcal{P}_k = \{(\mathbf{MS}_{k,a}, \mathbf{MD}_{k,a})|a \in \mathcal{A}_m\}$ which is obtained based on region augmentations and a negative set $\mathcal{N}_k = \{(\mathbf{MS}_n, \mathbf{MD}_n)|n \neq k\}$ with $|\mathcal{N}_k| = N_m$, which is randomly sampled. Similarly, we define the sampling probability of each candidate region based on its mobility feature distance from the anchor region r_k. N_m is the predefined negative size for mobility view. The mobility view encoder G^m is also implemented as an MLP network as follows.

$$\begin{aligned}\mathbf{z}_k^m &= G^m(\mathbf{MS}_k, \mathbf{MD}_k) \\ &= \text{Avg}(\text{MLP}(\mathbf{MS}_k), \text{MLP}(\mathbf{MD}_k)).\end{aligned} \tag{4}$$

Afterwards, based on the learned positive embedding $\mathbf{z}_{k,a}^m = G^m(\mathbf{MS}_{k,a}, \mathbf{MD}_{k,a})$ and negative embedding $\mathbf{z}_n^m = G^m(\mathbf{MS}_n, \mathbf{MD}_n)$, the intra-view contrastive signals for mobility view can be formulated as follows.

$$\mathcal{L}_{mob} = \sum_{r_k \in \mathcal{R}} \left[-\log \sum_{a \in \mathcal{A}_m} D^{ia}(\mathbf{z}_k^m, \mathbf{z}_{k,a}^m) + \log \left(\sum_{a \in \mathcal{A}_m} D^{ia}(\mathbf{z}_k^m, \mathbf{z}_{k,a}^m) + \sum_{n \in \mathcal{N}_k} D^{ia}(\mathbf{z}_k^m, \mathbf{z}_n^m) \right) \right]. \tag{5}$$

4.2 Inter-view Contrastive Learning

We learn two groups of view-specific region embeddings separately via an intra-view contrastive learning module. To enhance the cooperation between different views and capture a region's intrinsic properties, we follow the second guideline, which is to compare a region with itself based on different views. With this component, the feature learning in one view can constrain the learning in the other view, and hence it serves as a soft co-regularizer for the entire model. Conceptually, only good representations in both views can match each other.

Recall that, for the k-th region r_k, we obtain its region representations in each view as $\mathbf{z}_k^p$ and $\mathbf{z}_k^m$. Intuitively, the two embeddings can be the ground-truth of each other for self-supervised learning, and this one-to-one mapping can be regarded as the label augmentation. If two region representations correspond to those of the same region, we label this pair as a positive pair, otherwise a negative pair. Specifically, for the k-th region, the positive pair should be the representations for the same region within different views formulated as ($\mathbf{z}_k^p$, $\mathbf{z}_k^m$). Meanwhile, by fixing the embedding of r_k in one view and sampling other region representations from another view, we can obtain two groups of representations as $\mathcal{Z}^p = \{\mathbf{z}_n^p | n \neq k\}$ with $|\mathcal{Z}^p| = N_i$ and $\mathcal{Z}^m = \{\mathbf{z}_n^m | n \neq k\}$ with $|\mathcal{Z}^m| = N_i$, where N_i is the predefined negative size for inter-view. Then the negative pairs are extracted as ($\mathbf{z}_k^p$, $\mathbf{z}_n^m$) and ($\mathbf{z}_n^p$, $\mathbf{z}_k^m$). The InfoNCE loss function for inter-view is defined as follows.

$$\mathcal{L}_{inter} = \sum_{r_k \in \mathcal{R}} \left[-\log D^{ir}(\mathbf{z}_k^p, \mathbf{z}_k^m) + \log \left(\begin{array}{l} D^{ir}(\mathbf{z}_k^p, \mathbf{z}_k^m) + \sum_{\mathbf{z}_n^m \in \mathcal{Z}^m} D^{ir}(\mathbf{z}_k^p, \mathbf{z}_n^m) \\ \qquad + \sum_{\mathbf{z}_n^p \in \mathcal{Z}^p} D^{ir}(\mathbf{z}_n^p, \mathbf{z}_k^m) \end{array} \right) \right]. \tag{6}$$

Here, D^{ir} is an inter-view discriminator that evaluates the matching scores between two input region embeddings, which is different from D^{ia} in the intra-view contrastive learning for similarity measurement. Specifically, we implement it with a fully connected neural network by formulating the cross-view matching as a classification task.

$$D^{ir}(\mathbf{z}_k^p, \mathbf{z}_k^m) = \exp(\mathrm{ReLU}(\mathbf{W}(\mathbf{z}_k^p \parallel \mathbf{z}_k^m) + \mathbf{b})). \tag{7}$$

Here, $\mathbf{z}_k^p$ and $\mathbf{z}_k^m$ are POI view and mobility view representations for region r_k, respectively, $\parallel$ refers to the concat operation, and $\mathbf{W}$ and $\mathbf{b}$ are neural network parameters. Our model can also be regarded as transferring knowledge along bi-directions by using one view as the label for the other, which helps the intra-view module obtain information from each other to improve feature extractions.

TABLE 1: Data Description.

Dataset	Description
Census blocks	Boundaries of 270 regions split by streets in Manhattan, New York.
Taxi trips	Around 10 million taxi trip records during one month in the studied area.
POI data	Over 17,000 POIs across over 200 fine-grained categories

4.3 Multi-task Learning

Finally, we unify the different pre-tasks into a multi-task learning framework. The joint learning objective function is as follows, where α and β are parameters controlling the balance between intra and inter-view contrastive learning.

$$\mathcal{L} = \mathcal{L}_{mob} + \alpha \cdot \mathcal{L}_{poi} + \beta \cdot \mathcal{L}_{inter}. \tag{8}$$

After the model training, when given a new region r_k with mobility heat maps $\mathbf{MS}_k$, $\mathbf{MD}_k$ and POI attributes $\mathbf{f}_k$, we can obtain the final region representation as follows.

$$\mathbf{e}_k = G^p(\mathbf{f}_k) \parallel G^m(\mathbf{MS}_k, \mathbf{MD}_k). \tag{9}$$

5 Experiment

We empirically evaluate our model with two important downstream applications: land usage clustering and region popularity prediction. The land usage clustering task is to cluster regions into different groups such as central business district (CBD) and the region popularity prediction task is to infer the check-in count in each region. Both of them are commonly used in existing studies of region embedding, e.g., land usage clustering is used in [1], [5] and region popularity prediction is used in [2], [3].

5.1 Experiment Settings

Dataset. We collect several real-world datasets of New York City from NYC open data website [2]. The statistics of each dataset are reported in Table 1. The first dataset is the boundary data. We use the city boundaries designed by the US Census Bureau [3] for region embedding learning. Following the experimental settings in [1], we use the Manhattan borough as our studied area. The second dataset is the human mobility data. We use complete taxi trip records of August 2013 as our training data. Each trip record contains the pick-up and drop-off locations and time stamps. For each time stamp, we map it to an hourly-long bin in one day to construct the mobility heat maps. The last dataset is the NYC POI data provided by [22], which includes over 17,000 POIs across over 200 fine-grained categories. We regard these locations' categories as POI attributes for each region following the work in [1].

Evaluation Metrics. We use three metrics, namely Normalized Mutual Information (NMI), Adjusted Rand Index (ARI) and F-measure, for the land usage clustering task following [5]. And we use three metrics, namely Mean Absolute Error (MAE), Root Mean Square (RMSE), and the Coefficient of Determination (R^2), for the region popularity prediction task following [1]. The detailed definitions of these metrics are presented in a technical report [23].

2. https://opendata.cityofnewyork.us/
3. https://www.census.gov/data.html

TABLE 2: Performance on Two Downstream Tasks.

Method	*Land Usage Clustering*			*Popularity Prediction*		
	NMI	ARI	F-measure	MAE	RMSE	R^2
POI (TF-IDF)	0.389	0.039	0.069	281.37	417.10	0.288
HDGE [4]	0.587	0.117	0.121	216.98	339.28	0.529
ZE-Mob [5]	0.466	0.076	0.108	238.34	341.71	0.522
MV-PN [2]	0.407	0.057	0.085	276.29	419.68	0.279
CGAL [3]	0.411	0.078	0.102	275.03	415.79	0.293
MVURE [1]	0.691	0.337	0.351	222.55	336.39	0.537
M2GRL [24]	<u>0.696</u>	<u>0.361</u>	<u>0.383</u>	<u>194.01</u>	<u>303.89</u>	<u>0.625</u>
ReMVC	**0.762**	**0.474**	**0.488**	**189.92**	**296.26**	**0.643**

Baseline Methods. We compare our method with the following 7 baselines, which are existing methods of region embedding based on mobility patterns and/or POI features. **POI** merely uses the POI categories of a region with the TF-IDF method as the region embedding. **HDGE** [4] learns region representations by path sampling on a multi-layer mobility graph and spatial graph. **ZE-Mob** [5] learns region representations by employing mutual information matrix decomposition, which considers the region co-currency relations. **MV-PN** [2] learns region embedding by compressing multi-view POI networks. **CGAL** [3] extends MV-PN by employing an adversarial training method. **MVURE** [1] learns region representations by performing graph convolution on multiple homogeneous graphs. **M2GRL** [24] is a general multi-view embedding method which learns view-specific representations by graph reconstruction. Among these baselines, POI, HDGE and ZE-Mob are all single-view based methods which merely use the POI categories or human mobility patterns. MV-PN, CGAL, MVURE and M2GRL are multi-view based methods using both POI features and mobility patterns. The implementation details are presented in a technical report [23].

5.2 Land Usage Clustering

We first evaluate the learned embeddings in the land usage clustering task. As shown in the Fig. 2, the borough of Manhattan is divided into 29 districts based on land usage. We conduct k-means clustering (k=29) with the region embeddings as inputs, and the regions in the same cluster are assigned with the same type.

We show the performance results of different methods in Table 2, where the best performance is in boldface and the second best is underlined. Overall, our method performs consistently better than all the baselines in terms of three metrics, and the margin reaches near 10% for NMI and near 30% for ARI and F-measure. Specifically, POI (TF-IDF) performs the worst among the baselines. A possible reason is that the high dimensional representation based on POI categories is inefficient for the clustering task. For HDGE and ZE-Mob, the region attributes such as POIs are not considered, which may cause information loss. MVURE and M2GRL which consider multi-view region correlations perform much better than single-view based methods. For MV-PN and CGAL, they achieve comparable performance because they adopt the same idea but different training strategies. In contrast, we use the contrastive learning to highlight the discriminate features, which help our model achieve the best performance.

We also visualize the clustering results of the methods that have the top ARI scores in Fig. 2, where two regions

(a) Labels (b) ReMVC (c) M2GRL [24] (d) MVURE [1]

Fig. 2: Districts in Manhattan and region clusters.

in the same color are in the same cluster. The cluster size is used as the index of colors here. We can find that the functional regions are weakly identified by M2GRL, and some clusters are mixed with each other especially in the middle right part. In contrast, ReMVC gives more satisfying identification with clearer cluster boundaries. More visualization results can be found in a technical report [23].

5.3 Popularity Prediction

Region popularity prediction is another commonly used task [2]. Following this setting, we aggregate the check-in counts within each region as the ground truth of popularity, which can be found in [22]. To evaluate the effectiveness of different region embedding methods, we use the Lasso regression model for this downstream task. All results are obtained by 5-Fold cross-validation.

The results are also shown in Table 2. The proposed model performs consistently better than all the baselines in terms of three metrics. Among the baselines, MV-PN and CGAL achieve comparable results with the POI (TF-IDF) methods. A possible reason is that the mobility modeling module serves as an oversimplified regularization part in both methods, and the region representation is obtained by compressing POI features. These models cannot fully exploit multi-view correlations. Because human mobility is highly related to region popularity, HDGE and ZE-Mob achieve good results on this task. Compared with other multi-view learning methods MVURE and M2GRL, our method performs better with the inter-view contrastive learning module to propagate information across different views.

5.4 Module Evaluation

In this experiment, we further inspect how each module of ReMVC affects the clustering and regression performance.

Specifically, by removing the POI encoder G^p and Mobility encoder G^m from ReMVC, we can obtain two simplified models w/o POI and w/o Mob, in which we train our model only based on one of the two views. We also disable the inter-view contrastive learning module and directly concatenate two embeddings without knowledge propagation in a simplified model w/o IV. Additionally, in order to verify the necessity of exploiting contrastive learning in the intra-view feature extraction, we replace it with an AutoEncoder using MSE loss and obtain the base model ReMVC-MSE. The encoder and decoder are implemented as G_{MLP} and the size of middle layer is set as 32 for fair comparison.

From the results shown in Table 3, we have the following observations. (1) ReMVC which considers multi-view information generally has better performance than

TABLE 3: Performance Comparison of Different Modules.

Method	*Land Usage Clustering*			*Popularity Prediction*		
	NMI	ARI	F-measure	MAE	RMSE	R^2
w/o POI	0.759	0.430	0.449	253.10	385.12	0.393
w/o Mob	0.474	0.083	0.115	238.86	357.72	0.477
w/o IV	0.723	0.393	0.405	229.77	343.26	0.518
ReMVC-MSE	0.517	0.127	0.151	191.68	299.91	0.632
ReMVC	**0.762**	**0.474**	**0.488**	**189.92**	**296.26**	**0.643**

the methods considering only single-view data, i.e., w/o POI and w/o Mob. The result demonstrates that each view has its own merit to enhance the region representation. Meanwhile, the model w/o POI outperforms w/o Mob on both two tasks. One possible reason is that POIs often suffer from data sparsity problem. (2) Our full model ReMVC brings around 20% (in terms of ARI) and 17% (in terms of MAE) improvement compared with the base model w/o IV, showing that the proposed inter-view contrastive learning module can fully exploit the multi-view data and the importance of information propagation between views. (3) ReMVC outperforms ReMVC-MSE by a large margin especially in the land usage clustering task. The results show that the feature reconstruction pretext task is not as effective as contrastive learning in region representation learning. Similar to the image data, a region can be identified by its distinct structures rather than its pixel-level details.

In general, contrastive learning and multi-view cooperation are key components for our model. Due to the page limit, we present the experiment results of other ablation studies and parameters analysis in a technical report [23].

6 Conclusion

In this paper, we study the problem of learning an embedding space for regions using multi-view information. We identify and introduce two guidelines for this problem: i) comparing a region with others within each view for effective representation extraction and ii) comparing a region with itself across different views for cross-view information sharing. Based on the guidelines, we propose a novel multi-view joint learning model ReMVC for region embedding, which features in the schemes of intra-view and inter-view contrastive learning. Note that our contrastive learning structure can also be applicable to other domains. When this structure is applied to other domains, some implementation details (including the augmentation methods, loss functions, etc.) need to be adjusted to better suit the target domain.

References

[1] M. Zhang, T. Li, Y. Li, and P. Hui, "Multi-view joint graph representation learning for urban region embedding." in *IJCAI*, 2020, pp. 4431–4437.

[2] 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*, vol. 33, no. 01, 2019, pp. 906–913.

[3] 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*, 2019, pp. 1700–1708.

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

[5] 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 Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

[6] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.

[7] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 769–778.

[8] D. Paul, F. Li, and J. M. Phillips, "Semantic embedding for regions of interest," *The VLDB Journal*, vol. 30, no. 3, pp. 311–331, 2021.

[9] 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*, vol. 33, no. 01, 2019, pp. 3967–3974.

[10] T. Huang, Z. Wang, H. Sheng, A. Y. Ng, and R. Rajagopal, "M3g: Learning urban neighborhood representation from multi-modal multi-graph."

[11] P. Wang, Y. Fu, J. Zhang, X. Li, and D. Lin, "Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 6, pp. 1–28, 2018.

[12] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] P. Arsomngern, C. Long, S. Suwajanakorn, and S. Nutanong, "Towards pointsets representation learning via self-supervised learning and set augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[16] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.

[17] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "Skeletonnet: A hybrid network with a skeleton-embedding process for multi-view image representation learning," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2916–2929, 2019.

[18] Z. Zhou, Y. Hu, Y. Zhang, J. Chen, and H. Cai, "Multiview deep graph infomax to achieve unsupervised graph embedding," *IEEE Transactions on Cybernetics*, 2022.

[19] Z. Shao, Y. Xu, W. Wei, F. Wang, Z. Zhang, and F. Zhu, "Heterogeneous graph neural network with multi-view representation learning," *arXiv preprint arXiv:2108.13650*, 2021.

[20] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[21] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *arXiv preprint arXiv:2010.09709*, 2020.

[22] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, 2014.

[23] Liang Zhang, Cheng Long and Gao Cong, "Region embedding with intra and inter-view contrastive learning (technical report)," 2022, https://github.com/Liang-NTU/Techinical_Report_ReMVC/blob/main/22-TKDE-ReMVC-report.pdf.

[24] M. Wang, Y. Lin, G. Lin, K. Yang, and X.-M. Wu, "M2grl: A multi-task multi-view graph representation learning framework for web-scale recommender systems," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2349–2358.

7 APPENDIX

7.1 Implementation Details

In our experiments, we set the embedding dimension d to 32 for all the learning methods following the setting in [3] for fair comparison. We use the Adaptive Moment Estimation (Adam) optimizer to train the models with a learning rate of 0.001. The weighted parameters α and β are set to 0.001 and 1.0 respectively. For the temperature parameter τ, we set it as 0.08 following the study in [21]. The negative size for POI view, mobility view, and inter-view contrastive learning are set to 150, 10, and 5 respectively. And the augmentation parameters p and δ are set to 0.1 and 0.0001 in this study. For all the baseline methods, the learning rate is searched in {0.0005, 0.001, 0.005, 0.01}. For MV-PN and CGAL, the hidden layer of AutoEncoder is searched in {1, 2, 3}. For MVURE, we tune the number of heads in multi-head attention among {4, 8, 16}.

7.2 Evaluation Metrics

NMI, ARI and F-measure are used to evaluate the land usage clustering results in this paper.

- **Normalized Mutual Information (NMI)** measures the purity of the clustering results, which is defined as

$$\text{NMI} = \frac{I(L;C)}{[H(L)+H(C)]/2}, \tag{10}$$

  where L is the ground-truth classes and C is the predicted classes. $H(L)$ and $H(C)$ denote the entropy, and $I(L;C)$ denotes the sum of mutual information between the predicted classes and ground-truth classes. The score ranges from 0 for worst-case to 1 for perfect matching.
- **Adjusted Rand Index (ARI)** calculates the clustering accuracy by viewing the cluster assignments of all pairs of regions as different decisions. It can be defined as

$$\begin{aligned} \text{ARI} &= \frac{RI - Expected_RI}{Max_RI - Expected_RI}, \\ \text{RI} &= \frac{TP+TN}{TP+FP+TN+FN}, \end{aligned} \tag{11}$$

  where TP/FP represents true/false positive and TN/FN denotes true/false negative. For the ARI score, 1 stands for best result and 0 means random labeling.
- **F-measure** also views the clustering results of each pair as a series of binary classification decisions, then we can obtain precision, recall based on the predefined TP/FP and TN/FN. Then, F-measure is defined as follows.

$$F = \frac{(\lambda^2+1)\cdot Precision \cdot Recall}{\lambda^2 \cdot Precision + Recall}. \tag{12}$$

  We follow the setting in [5] and set $\lambda = 0.5$.

MAE, RMSE and R^2 are used to evaluate the region popularity prediction results in our study.

- **Mean Absolute Error (MAE)** measures the average absolute distance between predictions and ground-truths.
- **Root Mean Square (RMSE)** measures the average quadratic errors of all predictions.
- **Coefficient of Determination (R^2)** measures the goodness of fit for regression models, which is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{13}$$

  where $y_i, f_i, \bar{y}$ are ground truth, model prediction and the mean of ground truth respectively. The larger the R^2 is, the better the performance is.

7.3 Comparison with Other Baselines

Another work (CCDR) that uses intra and inter contrastive learning is proposed very recently [A1]. In this paper, each view corresponds user graph data and the intra-view presentation learning is based on graph structure perturbations. However, the urban region graphs do not exist naturally, and pre-defining the region graphs is non-trivial, which may suffer from the drawbacks mentioned in the Introduction section (heuristic-based measures and not learnable). On the other hand, the graph structure perturbations may affect urban function modelling significantly. For example, due to its special POI and flow patterns, the CBD region may only have very sparse connections with other regions. In this case, using structure perturbations (e.g., adding edges or dropping edges) could bring significant noise and correspondingly the augmented CBD region can not serve as a positive (similar) sample anymore. In summary, (1) our method is tailored for region data with POI view and mobility view. It is non-trivial to transform them into suitable graphs. (2) Graph structure perturbations may not be suitable for region-based contrastive learning. (3) CCDR is a supervised method aimed at solving cross-domain recommendation tasks, while our method aims to apply intra- and inter-view contrastive learning for unsupervised region embedding learning for the first time.

We compare our method and CCDR and show the results in Table 4. For constructing the region graphs, we follow a commonly used strategy named K-nearest neighbor graph (K-NNG). We tune the parameter K among {10, 20, 30, 40, 50} and select the best one. From the results, we find that our method performs consistently better than CCDR among two downstream tasks and all metrics.

TABLE 4: Comparision with CCDR (KDD 2022).

Method	*Land Usage Clustering*			*Popularity Prediction*		
	NMI	ARI	F-measure	MAE	RMSE	R^2
CCDR	0.655	0.268	0.294	248.31	379.40	0.414
ReMVC	**0.762**	**0.474**	**0.488**	**189.92**	**296.26**	**0.643**

7.4 Complete Visualization For Land Use Clustering

Fig. 3 shows the complete clustering results of five methods that have the top ARI scores, namely ReMVC, M2GRL, MVURE, HDGE and CGAL. where the same color means that the regions are in the same cluster. For HDGE, the functional regions are weakly identified since only human mobility data is used. CAGL cannot identify the correct region cluster either since it cannot effectively fuse multi-view information but mainly depends on POI modeling. M2GRL and MVURE provides better performance than other baselines and obtains more reasonable clusters. However, some

(a) Labels (b) ReMVC (c) M2GRL (d) MVURE (e) HDGE (f) CGAL

Fig. 3: Districts in Manhattan and region clusters (where the same color means that the regions are in the same cluster).

clusters are still mixed with each other especially on the middle part. In contrast, ReMVC gives the most satisfying identification with clear cluster boundaries, which demonstrates that the region functionalities are well preserved by our model.

7.5 Ablation Study for Other Modules

In this experiment, we further inspect how other modules (besides contrastive learning and multi-view cooperation which have been studied in Section 5.4) of ReMVC affect the model performance.

To further investigate the necessity of intra and inter-view contrastive learning from another perspective, we propose a cross view augmentation strategy to facilitate knowledge propagation directly in the base model ReMVC-CA instead of using the high-level inter-view contrastive learning. Specifically, taking the mobility view as an example. When constructing the positive region set for the k-th region r_k, in addition to intra-view augmentation $\mathcal{A}_m$, we define cross-view augmentation as the top-K ($K = 3$) regions, which are the most similar to r_k in the POI view (vice versa).

We also study the design of inter-view contrastive learning by replacing the "correlations" based one-to-one mapping classification task in Equation 6 with a similarity constraint-based task implemented as inner product. We call the new variant the ReMVC-Sim model.

To investigate the impacts of negative sampling methods, we replace the feature distance based sampling with Euclidean distance based sampling in model ReMVC-ES and random sampling in model ReMVC-RS respectively, where Euclidean distance based sampling means that the sampling probability for a candidate region r_n is defined as the normalization of its physical distance from the anchor region r_k and the random sampling means that the sampling probability is uniform.

We also replace the feature extractor of G^p and G^m with another commonly used structure CNN with the kernel size of 7 and rename this model as ReMVC-CNN. As for the embedding fusion operations, we consider these commonly used fusion operations, including Average, Max, Gated, Neural Networks (NN) based and Bilinear fusion, for comparison.

From experiment results in Table 5, we can have the following observations: (1) We can find that ReMVC brings huge improvement compared with ReMVC-CA, showing that our model does benefit from the hierarchical structural design of contrastive learning. This result is consistent with the finding indicated by comparing ReMVC with w/o IV in Section 5.4. (2) As for inter-view task design, we can find that our method ReMVC outperforms ReMVC-Sim by about 45% in F-measure and 28% in R^2. The underlying reason is that preserving similarity between views instead of correlation brings noise and leads to the same multi-view representations, which is detrimental to downstream tasks. (3) As for different sampling methods, we can find that ReMVC outperforms ReMVC-ES on two downstream tasks consistently, showing that the Euclidean distance sampling can hardly define effective negative regions. The possible reason is that regions far away on the map may still serve as similar urban functions. In addition, it is also unreasonable to treat regions with various functions equally. Thus, ReMVC outperforms ReMVC-RS for about 8% in land usage clustering task and 17% in popularity prediction task. (4) ReMVC achieves better results than ReMVC-CNN on both clustering and regression tasks, showing the superior ability of MLP for intra-view feature extraction. (5) Among the studied fusion operations, the max fusion performs better in the land usage clustering task while the NN based fusion performs better in the popularity prediction task. However, all these operations lead to entangled region representations. Instead, the disentangled multi-view representation by concatenation (adopted in this paper) can achieve the best results and the inter-view contrastive learning serves to propagate information across views.

7.6 Parameter Analysis

We conduct the parameter study on the weighted parameter β, α in Eq. 8, the embedding size d, the parameters p and σ for defining the augmentation operations for POI view and mobility view, and negative sizes N_p, N_m, N_i for intra/inter views respectively.

The weighted parameter β and α. As can be seen in Fig. 4a, the model performance would fluctuate with different values of β. While ReMVC remains relatively stable on clustering task, it is more sensitive to this parameter in

TABLE 5: Ablation Study for Other Modules.

Method	*Land Usage Clustering*			*Popularity Prediction*		
	NMI	ARI	F-measure	MAE	RMSE	R^2
ReMVC-CA	0.753	0.443	0.451	251.91	376.45	0.420
ReMVC-Sim	0.668	0.303	0.327	229.65	348.83	0.502
ReMVC-ES	0.751	0.442	0.457	255.81	387.65	0.385
ReMVC-RS	0.751	0.436	0.452	228.32	363.75	0.459
ReMVC-CNN	0.742	0.434	0.464	202.59	315.71	0.595
Average	0.756	0.449	0.458	231.02	353.47	0.489
Max	0.759	0.466	0.471	238.06	365.77	0.455
Gated	0.736	0.414	0.432	246.38	380.54	0.408
NN Based	0.560	0.154	0.182	236.12	347.02	0.508
Biliner	0.721	0.388	0.407	231.83	360.54	0.468
ReMVC	**0.762**	**0.474**	**0.488**	**189.92**	**296.26**	**0.643**

(a) Parameter β

(b) Parameter d

Fig. 4: Impact of hyper-parameters β and d.

popularity prediction. The performances when $\beta = 0.001$ cannot compete with that when β gets larger, which conforms to the results in ablation study and again show the necessity of introducing the inter-view contrastive learning module for cross-view information sharing. We omit the discussion for α because it obtains the similar results.

Embedding size d. Fig. 4b shows the impact of embedding size on the results. The model is quite robust under different embedding sizes ranging from 16 to 256, with the largest difference being less than 2%. It is also interesting to discover that using high-dimensional representation does not always produce the best results. For example, ARI and F-measure achieve the best when $d = 32$.

The augmentation probability p for POI view. As can be seen in Fig. 5a, the model performance would fluctuate with different values of p, where 0.1 is the best setting for both tasks. The results are reasonable because each region only includes a small amount of POIs. Meaningful augmentations are hard to generate with a small p. In fact, the augmentation is likely to be the anchor region itself so that the model cannot obtain informative positive samples. On the other hand, when p is too large, the augmentations introduce much noise and may affect the model learning.

The Gaussian variance σ for Mobility view. Fig. 5b shows the impact of various Gaussian variances on the results. We can find that the model performance decreases with σ for both tasks. The possible reason is that the injected

(a) Parameter p

(b) Parameter σ

Fig. 5: Impact of hyper-parameters p and σ.

noise with large σ will dominate the true origin/destination distributions and hence the model accepts questionable positive samples.

(a) Parameter N_p

(b) Parameter N_m

(c) Parameter N_i

Fig. 6: Impact of intra/inter view negative sizes.

The negative sizes for intra/inter views. As can be seen in Fig. 6, the model performances tend to increase firstly and then drop when negative sizes are too large for both intra and inter views. However, the largest difference between various experimental settings is less than 2% for NMI (land usage clustering task) and 4% for R^2 (popularity prediction task), showing that our model is quite robust against these hyper-parameters.

REFERENCES

[A1] R. Xie, Q. Liu, L. Wang, S. Liu, B. Zhang, and L. Lin, "Contrastive cross-domain recommendation in matching," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4226–4236.