Comprehensive Urban Region Representation Learning via Multi-view Joint Learning and Contrastive Learning

Yingde Lin[1,2], **Yuanbo Xu**[1,2*], **Lu Jiang**[3], **Pengyang Wang**[4]
[1]College of Computer Science and Technology, Jilin University
[2]MIC Lab, College of Computer Science and Technology, Jilin University
[3]College of Computer Science and Technology, Dalian Maritime University
[4]Department of Computer and Information Science, University of Macau
linyd24@mails.jlu.edu.cn, yuanbox@jlu.edu.cn, jiangl761@dlmu.edu.cn, pywang@um.edu.mo

Abstract

Urban region embedding, which learns dense vector representations for urban zones, plays a foundational role in data-driven urban intelligence. These representations are critical for downstream applications like public safety management and infrastructure development, requiring nuanced understanding of urban functionality. A core challenge remains effective fusion of multi-view data (e.g., human mobility flows and static regional attributes) into unified zone representations. To this end, we propose **MVJC**, a **M**ulti-**v**iew **J**oint Learning and **C**ontrastive Learning framework, which employs: (1) Multi-view Joint Learning (MVJL) layer to model intra-view dependencies to extract view-specific features and (2) Multi-view Contrastive Learning (MVCL) layer to perform cross-region aggregation to derive consensus representations while capturing the regional complementarity. We further introduce a structure-aware contrastive loss that mitigates false negatives by aligning representations through region topology instead of instance identity. Extensive experiments on New York City datasets demonstrate MVJC's superiority: it reduces crime prediction MAE by 9.1% (vs. 66.9 baseline) and improves land use clustering F-measure by 55.6% (vs. 0.45 baseline) over state-of-the-art method, which is attributed to MVJC's synergy of joint and contrastive learning, yielding representations that are simultaneously predictive and semantically discriminative.

Code — https://github.com/MichistaLin/MVJC

Introduction

Learning high-quality embeddings for urban regions is a critical task in urban computing. These embeddings distill complex, multi-view data to support downstream tasks like crime prediction (Yao et al. 2018) and land use clustering (Huang et al. 2018), which are essential for building smart cities.

Methods have evolved from single-view models (Wang and Li 2017; Yao et al. 2018) to multi-view (e.g., human mobility, POIs, building footprints) approaches (Zhang et al. 2021; Zheng, Yuan, and Guan 2022; Sun et al. 2024; Li et al. 2024). In practice, these heterogeneous data views are rarely independent. For instance, the functional characteristics of a geographic area (e.g., commercial zones vs. residential zones, as defined by POIs) are a key determinant of its pedestrian flow dynamics, including peak times and overall patterns. Therefore, it is imperative to develop a framework that leverages the interdependence of these views for their mutual enhancement.

Most recently, contrastive learning has emerged as a powerful self-supervised paradigm for aligning representations from different views (Zhang et al. 2023b,a; Li et al. 2023, 2024). The standard objective, however, creates a fundamental conflict when applied to clustering. By strictly treating any two different regions as a negative pair, it erroneously forces the model to push apart representations of regions that, while distinct, share the same functional role (e.g., two residential zones), as illustrated in Figure 1. This critical"false negative" problem directly undermines the goal of functional clustering by penalizing the semantic similarity that the model is designed to capture (Yan et al. 2023).

To overcome these challenges, we propose the Multi-view Joint learning and Contrastive learning (MVJC) framework, as shown in Figure 2. MVJC's hierarchical architecture directly addresses these limitations in two stages. First, a Multi-view Joint Learning (MVJL) module refines each view's features through cross-view interaction, creating enhanced view-specific embeddings. Then, a Multi-view Contrastive Learning (MVCL) module generates a consensus representation. It employs a structure-aware objective to mitigate the "false negative" problem by using learned structural relationships to align functionally similar regions, preventing their incorrect separation. This design yields robust and semantically discriminative urban region embeddings.

The main contributions of this paper are as follows:

- We propose MVJC, a novel hierarchical framework that operates in two stages: it first refines view-specific embeddings via a Multi-view Joint Learning (MVJL) module to enhance their quality, and then generates a robust consensus representation using a global, structure-aware Multi-view Contrastive Learning (MVCL) module. This design yields robust and comprehensive urban region embeddings that are resilient to noise and view-specific distortions.
- We apply the global and cross-view feature aggregation (GCFA) and structure-aware contrastive learning

*Corresponding author.

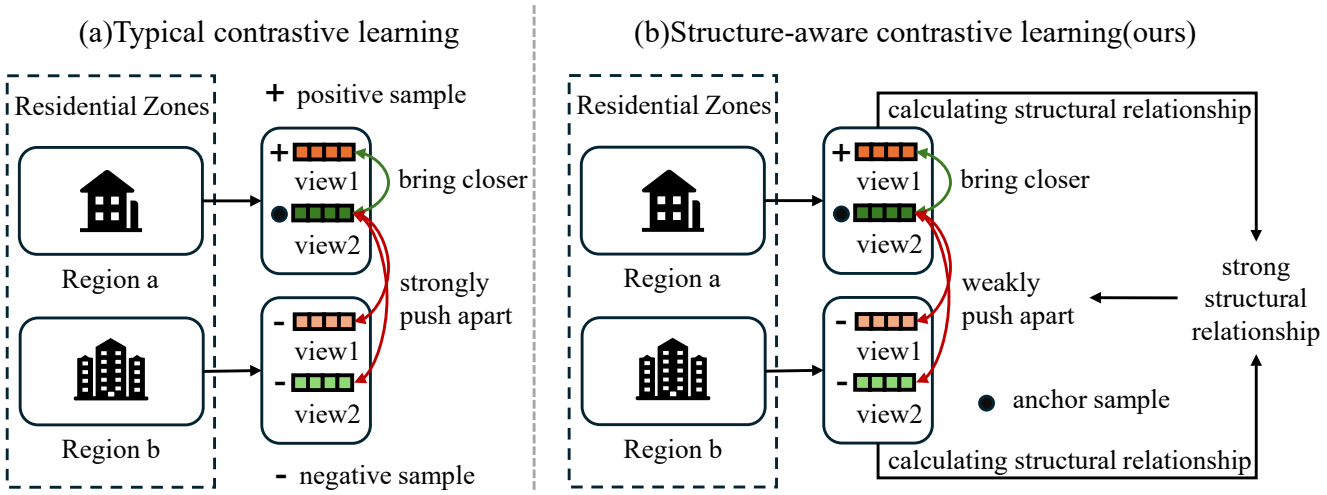


Figure 1: (a) Typical contrastive learning usually considers different views of the same region as positive sample pairs, and views of different regions as negative sample pairs. (b) If two different regions actually belong to the same functional area, their representations should also be similar.

(SACL) module to the field of urban region representation learning for the first time, and effectively integrate it with the multi-view joint learning framework to replace the traditional per-view fusion. This module utilizes the structural information of the entire dataset to learn a consistent representation for each region, thereby effectively mitigating the impact of noisy views.

- We conduct a comprehensive experimental validation on large-scale, real-world datasets from New York City. The results demonstrate that MVJC achieves new state-of-the-art performance on multiple challenging urban prediction tasks, significantly outperforming a wide range of baseline methods.

Preliminaries

Urban Region. A city can be partitioned into n disjoint urban regions by census blocks, denoted as $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$.

Human mobility. We define urban human mobility as a set of trip records that occur in urban areas. We denote a human mobility dataset as $\mathcal{M}$ and each entity in $\mathcal{M}$ is a tuple consisted of source and destination of the trip:

$$\mathcal{M} = \{m_0, m_1, \ldots, m_{|\mathcal{M}|}\}, m_i = (r_s, r_d, t), \forall m_i \in \mathcal{M},$$

where r_s is the start region, r_d is the destination region, and t is the trip start time.

Region attributes. The region attributes are the inherent social and geographic features of urban regions. A certain type of attribute of regions can be denoted as $\mathcal{A}$ as follows:

$$\mathcal{A} = \{\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}\}, \vec{a_i} \in \mathbb{R}^F, \forall \vec{a_i} \in \mathcal{A},$$

where $\vec{a_i}$ is the corresponding feature of i-th region and F is the number of dimensions of that feature. In our work, multiple region attributes, like POIs and check-ins, are considered.

Region Representation Learning. Given the human mobility $\mathcal{M}$ of a set of regions $\mathcal{R}$ and attribute features $\mathcal{A}$ of regions, we aim to learn a set of low dimensional embedding $\mathcal{E}$ to represent each region: $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}, e_i \in \mathbb{R}^d$, where $e_i \in \mathcal{E}$ is the d-dimension embedding of the region $r_i \in \mathcal{R}$ and n is the number of regions.

Methodology

Framework Overview

As shown in Figure 2, the MVJC framework comprises two modules: (1) the Multi-view Joint Learning (MVJL) module aims to refine and generate high-quality, view-specific representations; (2) the Multi-view Contrastive Learning (MVCL) module aggregates cross-region and cross-view features to learn a global consensus representation.

Multi-view Joint Learning

To learn robust representations, we jointly consider region correlations from multiple views, constructing graphs based on both human mobility and static region attributes (e.g., POIs, check-ins).

Region Correlations Based on Human Mobility

Following the formulation in MGFN (Wu et al. 2022), we model raw human mobility data as a sequence of directed, weighted graphs over time. A single mobility graph at time interval t is defined as $G_t = (V, E_t)$, where V is the set of region nodes $\{v_1, \ldots, v_n\}$ corresponding to $\mathcal{R}$. An edge $e_{ij}^t = (v_i, v_j, \omega_{ij}^t) \in E_t$ represents the volume of flow ω_{ij}^t from region r_i to region r_j during interval t. The complete human mobility dataset is a time-series multi-graph, $G = \bigcup_{t=0}^{T-1} G_t$. Then the multi-graphs are processed by the encoder Multi-Graph Fusion Networks to obtain the initial human mobility view representation $\mathcal{E}_{mob}$.

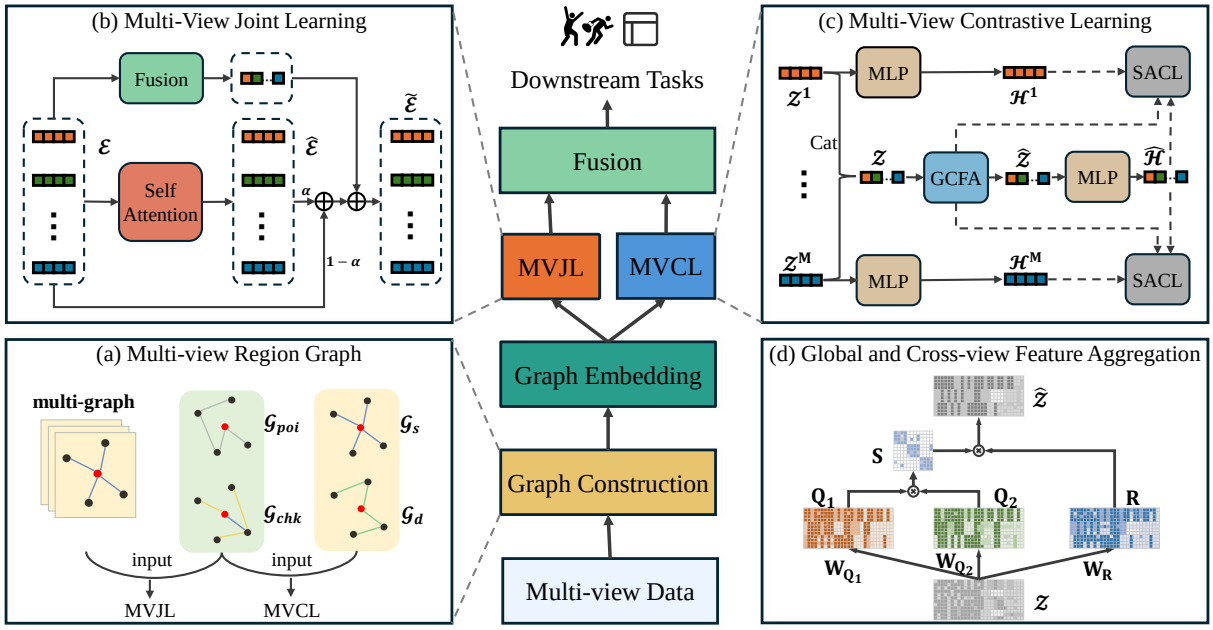


Figure 2: The framework of MVJC.

Region Correlations Based on Region Attributes

The inherent region attributes are the meta-knowledge that describes the geographic and social nature of a region. Given a type of attributes of n regions $\mathcal{A} = \{\vec{a}_i\}_{i=1}^{n}$, the corresponding region correlations are computed as

$$\mathbf{C}^{ij} = \text{sim}(\vec{a}_i, \vec{a}_j), \tag{1}$$

where $\text{sim}(\cdot)$ refers to cosine similarity. We compute region-to-region similarity matrices $\mathbf{C}_{poi}$ and $\mathbf{C}_{chk}$ based on the cosine similarity of their respective TF-IDF feature vectors. From these matrices, we construct two graphs, $\mathcal{G}_{poi}$ and $\mathcal{G}_{chk}$. In each graph, every region node is connected to its k-nearest neighbors, forming a sparse graph that captures local attribute-based similarities.

From these similarity matrices, we construct sparse k-NN graphs ($\mathcal{G}_{poi}, \mathcal{G}_{chk}$) for each attribute view. We then employ a standard Graph Attention Network (GAT) as an encoder to learn representations $\mathcal{E}_{poi}$ and $\mathcal{E}_{chk}$ from these graphs.

Multi-View Interaction

The interaction module (Figure 2(c)) achieves cross-view integration in two steps. It first employs a self-attention mechanism to propagate information among all view representations. A subsequent fusion layer then adaptively combines the resulting representations.

We employ the self-attention mechanism to propagate information across the representations of different views. Given the representations from M different views as $\{\mathcal{E}_i \in \mathbb{R}^{n\times d}\}_{i=1}^{M}$. For each representation $\mathcal{E}_i$, we associate a key matrix $\mathbf{K}_i \in \mathbb{R}^{n\times k}$ and a query matrix $\mathbf{Q}_i \in \mathbb{R}^{n\times k}$ with it as follows:

$$\mathbf{K}_i = \mathcal{E}_i\mathbf{W}_k, \quad \mathbf{Q}_i = \mathcal{E}_i\mathbf{W}_q. \tag{2}$$

For each view, we then propagate information among all views as follows:

$$[\mathbf{A}_i]_{i=1}^{M} = \text{softmax}\left(\left[\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{k}}\right]_{i=1}^{M}\right), \hat{\mathcal{E}}_i = \sum_{i=1}^{M}\mathbf{A}_i\mathcal{E}_i. \tag{3}$$

In our case, $\hat{\mathcal{E}}_i$ is considered as the relevant global information for i-th view. To incorporate this information in the learning process, we compute

$$\mathcal{E}_i' = \alpha\hat{\mathcal{E}}_i + (1-\alpha)\mathcal{E}_i, \quad 0 \le \alpha \le 1, \tag{4}$$

where $\mathcal{E}_i'$ is the representation for i-th view with global information, and α is the weight of global information.

In order to make full use of the information of other views to strengthen the representation of its own view, we employ a fusion layer that learns adaptive weights for different views as follows:

$$\mathcal{E} = \sum_{i}^{M}\mathbf{w}_i\mathcal{E}_i, \quad \mathbf{w}_i = \sigma(\mathcal{E}_i\mathbf{W}_f + \mathbf{b}_f), \tag{5}$$

where $\mathbf{w}_i$ is the weight of i-th view, which is learned by a single layer MLP network with the i-th embeddings as input. To ensure the adaptive weights in the fusion layer are properly learned, we incorporate the fused representation $\mathcal{E}$ back into each view-specific learning objective. Formally, we update the representation of each view as:

$$\tilde{\mathcal{E}}_i = (\hat{\mathcal{E}}_i + \mathcal{E})/2. \tag{6}$$

We adopt this residual-like connection to balance view-specificity and consensus, ensuring the final representation $\tilde{\mathcal{E}}_i$ incorporates shared knowledge while retaining its uniquely enhanced characteristics. In this paper, the three views output are $\tilde{\mathcal{E}}_{mob}$, $\tilde{\mathcal{E}}_{poi}$ and $\tilde{\mathcal{E}}_{chk}$, respectively.

Learning Objectives

Following HDGE (Wang and Li 2017), we use region embeddings to estimate the distribution of mobility, and learn the embedding by minimizing the difference between the true distribution and the estimated distribution. Given the source v_i, we calculate the transition probability of the destination v_j:

$$p_\omega(v_j|v_i) = \frac{\omega_{ij}}{\sum_{v_{j^*} \in N(v_i)} \omega_{ij^*}}. \tag{7}$$

Then, given the region embedding $\tilde{\mathcal{E}}^i_{mob}$, $\tilde{\mathcal{E}}^j_{mob}$ for region v_i, v_j, we estimate the transition probability:

$$\hat{p}_\omega(v_j|v_i) = \frac{\exp({\tilde{\mathcal{E}}^i_{mob}}^T \tilde{\mathcal{E}}^j_{mob})}{\sum_{j^* \in N(v_i)} \exp({\tilde{\mathcal{E}}^i_{mob}}^T \tilde{\mathcal{E}}^{j^*}_{mob})}. \tag{8}$$

Finally, the objective function of human mobility view can expressed as:

$$\mathcal{L}_{mob} = \sum_{i,j} -p_\omega(v_j|v_i) \log \hat{p}_\omega(v_j|v_i). \tag{9}$$

We design a correlation reconstruction task to ensure that the learned embeddings preserve similarities between regions across various attributes. Take POI attribute as an example, the learning objective is defined as follows:

$$\mathcal{L}_{poi} = \sum_{i,j} \left(\mathbf{C}^{ij}_{poi} - (\tilde{\mathcal{E}}^i_{poi})^T \tilde{\mathcal{E}}^j_{poi} \right)^2. \tag{10}$$

Similarly, we define the learning objective $\mathcal{L}_{chk}$ of check-in attribute. In this way, The learning objective of the multi-view joint learning part is:

$$\mathcal{L}_r = \mathcal{L}_{mob} + \mathcal{L}_{poi} + \mathcal{L}_{chk}. \tag{11}$$

Multi-view Contrastive Learning

This module is designed to discover complex correlation patterns across different data views. By doing so, it identifies clusters of urban regions that share similar comprehensive characteristics, which facilitates the learning of more discriminative and robust urban region representations.

Regions that receive human flows from the same sources or send human flows to the same targets usually play similar roles and are considered close to each other from the human mobility view (Yao et al. 2018). In this module, we define the source and destination context of a region based on inter-region interactions. Given a set of human mobility $\mathcal{M}$, the interaction weight from region r_i to region r_j is computed as: $w^{r_i}_{r_j} = |\{(r_s, r_d) \in \mathcal{M} | r_s = r_i, r_d = r_j\}|$, where $|.|$ counts the set size. Then the source and destination contexts of a region r_i are described by distributions $p_s(r|r_i)$ and $p_d(r|r_i)$ as follows:

$$p_s(r|r_i) = \frac{w^r_{r_i}}{\sum_r w^r_{r_i}}, \quad p_d(r|r_i) = \frac{w^{r_i}_r}{\sum_r w^{r_i}_r}. \tag{12}$$

Based on the source and destination context of each region, we define two types of correlations as follows,

$$\mathbf{C}^{ij}_s = \text{sim}(p_s(r|r_i), p_s(r|r_j)), \tag{13}$$

$$\mathbf{C}^{ij}_d = \text{sim}(p_d(r|r_i), p_d(r|r_j)), \tag{14}$$

where $\mathbf{C}^{ij}_s$ is the source correlation and $\mathbf{C}^{ij}_d$ represents the destination correlation. We still follow the previous method to construct graphs $\mathcal{G}_s$, $\mathcal{G}_d$, $\mathcal{G}_{poi}$ and $\mathcal{G}_{chk}$ based on region correlations $\mathbf{C}_s$, $\mathbf{C}_d$, $\mathbf{C}_{poi}$ and $\mathbf{C}_{chk}$. Then apply the GAT encoder to obtain view representation $\mathcal{Z}_s$, $\mathcal{Z}_d$, $\mathcal{Z}_{poi}$ and $\mathcal{Z}_{chk}$ and concatenate them together, denoted as $\mathcal{Z} = [\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_M]$, where $\mathcal{Z}_i \in \mathbb{R}^{n \times d}$, $\mathcal{Z} \in \mathbb{R}^{n \times Md}$.

Global and Cross-view Feature Aggregation

Conventional fusion methods generate a region's representation using only its own multi-view data, which is considered a suboptimal approach. Our approach enhances a region's consensus representation by aggregating information not just from its own views, but also from other structurally similar regions across the entire dataset. This is achieved by learning a global structure relationship matrix to guide the aggregation, as shown in Figure 2(d).

Inspired by the idea of the transformer attention mechanism (Vaswani et al. 2017), we map $\mathcal{Z}$ into different feature spaces by the $\mathbf{W}_R$ to achieve the cross-view fusion of all views, i.e.,

$$\begin{bmatrix} \mathbf{R}_1 : \\ \mathbf{R}_2 : \\ \vdots \\ \mathbf{R}_n : \end{bmatrix} = \begin{bmatrix} \mathbf{z}^1_1 & \mathbf{z}^2_1 & \cdots & \mathbf{z}^M_1 \\ \mathbf{z}^1_2 & \mathbf{z}^2_2 & \cdots & \mathbf{z}^M_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}^1_n & \mathbf{z}^2_n & \cdots & \mathbf{z}^M_n \end{bmatrix} \begin{bmatrix} \mathbf{W}_{R1} : \\ \mathbf{W}_{R2} : \\ \vdots \\ \mathbf{W}_{RM} : \end{bmatrix}, \tag{15}$$

that is $\mathbf{R}_j := \sum_{k=1}^M \mathbf{z}^k_j \mathbf{W}_{Rk}$:. Similarity, the $\mathbf{Q}_1$ and $\mathbf{Q}_2$ is obtained by $\mathbf{W}_{Q1}, \mathbf{W}_{Q2}$, i.e.,

$$\mathbf{Q}_1 = \mathcal{Z}\mathbf{W}_{Q1}, \quad \mathbf{Q}_2 = \mathcal{Z}\mathbf{W}_{Q2}, \tag{16}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{n \times d}, \mathbf{Q}_2 \in \mathbb{R}^{n \times d}$. Here, we use the matrix $\mathbf{W}_Z = \{\mathbf{W}_{Q1}, \mathbf{W}_{Q2}, \mathbf{W}_R\}$ to denote the parameters.

The structure relationship among samples is denoted as:

$$\mathbf{S} = \text{softmax}\left(\frac{\mathbf{Q}_1 \mathbf{Q}_2^T}{\sqrt{d}} \right). \tag{17}$$

The learned representation matrix $\mathbf{R}$ is enhanced by the global structure relationship matrix $\mathbf{S}$. Conceptually, the representation of each sample is updated by aggregating information from other highly correlated samples. This process pulls the representations of samples from the same cluster closer together, thereby increasing their similarity.

$$\hat{\mathcal{Z}}_i = \sum_{j=1}^n \mathbf{S}_{ij} \mathbf{R}_j, \quad \hat{\mathcal{Z}} = [\hat{\mathcal{Z}}_1; \hat{\mathcal{Z}}_2; \ldots; \hat{\mathcal{Z}}_n], \tag{18}$$

where $\mathbf{R}_j \in \mathbb{R}^{1 \times d}$ is the j-th row elements of $\mathbf{R}$, denotes the j-th sample representation, $\mathbf{S}_{ij}$ denotes the relationship between the i-th sample and the j-th sample, $\hat{\mathcal{Z}} \in \mathbb{R}^{n \times d}$. Since $\hat{\mathcal{Z}}$ is learnt from the concatenation of all views $\mathcal{Z}$, it usually contains redundancy information. Next, the output is passed through the fully connected nonlinear and linear layer to eliminate the redundancy information. The expression is described as the following equation:

$$\hat{\mathcal{H}} = \mathbf{W}_3 \left(\max(0, (\mathcal{Z} + \hat{\mathcal{Z}})\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \right) + \mathbf{b}_3. \tag{19}$$

Structure-aware Contrastive Learning

The learnt consensus representation $\hat{\mathcal{H}}$ is enhanced by global structure relationship of all samples in a batch, these data consensus representations from different views of samples in the same cluster are similar. Hence, the consensus representation $\mathcal{H}$ and view-specific representation $\mathcal{H}^v$ from the same cluster should be mapped close together. Inspired by contrastive learning methods (Chen et al. 2020), we set the consensus representation and view-specific representation from the same sample as positive pairs. However, designating all other pairs as negative can lead to inconsistencies among the representations of different samples within the same cluster, which conflicts with the clustering objective. Hence, we employ a structure-aware multi-view contrastive learning module (Yan et al. 2023). Specifically, we introduce cosine distance to compute the similarity between consensus presentation $\hat{\mathcal{H}}$ and view-specific presentation $\mathcal{H}^v$:

$$C\left(\hat{\mathcal{H}}_{i,:}, \mathcal{H}_i^v\right) = \frac{\hat{\mathcal{H}}_{i,:}^T \mathcal{H}_i^v}{\|\hat{\mathcal{H}}_{i,:}\| \|\mathcal{H}_i^v\|}. \tag{20}$$

The loss function of structure-aware multi-view contrastive learning can be defined as:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i=1}^{N} \sum_{v=1}^{V} \log \frac{e^{C(\hat{\mathcal{H}}_{i,:}, \mathcal{H}_i^v)/\tau}}{\sum_{j=1}^{N} e^{(1-S_{ij})C(\hat{\mathcal{H}}_{i,:}, \mathcal{H}_j^v)/\tau - e^{1/\tau}}}, \tag{21}$$

where τ denotes the temperature parameter, $\mathbf{S}_{ij}$ is from Eq. (17). This equation implies that a smaller value of $\mathbf{S}_{ij}$ results in a larger value of $C(\hat{\mathcal{H}}_{i,:}, \mathcal{H}_j^v)$. In other words, when the structure relationship $\mathbf{S}_{ij}$ between the i-th and j-th sample is low (not from the same cluster), their corresponding representations are inconsistent; otherwise, their corresponding representations are consistent, which solves the problem caused by other contrastive learning methods that distinguish positive and negative pairs at the sample level.

Overall Learning Objectives

In the proposed framework, the loss in our network consists of two parts:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_c. \tag{22}$$

We use simple concatenation for the final fusion to preserve the rich information from both the enhanced view-specific representations($\tilde{\mathcal{E}}$) and the global consensus representation($\hat{\mathcal{H}}$) without loss. This combined embedding is then fed to downstream task heads, allowing them to learn the optimal way to utilize these concatenated features.

$$\tilde{\mathcal{H}} = concat(\tilde{\mathcal{E}}_{mob}, \tilde{\mathcal{E}}_{poi}, \tilde{\mathcal{E}}_{chk}, \hat{\mathcal{H}}). \tag{23}$$

Experiments

Experimental Settings

Dataset We utilize a variety of real-world data from NYC Open Data specific for the Manhattan, New York area, where Taxi trips are used as human mobility. We divide the Manhattan area into 180 regions based on the community boards. The detailed description of datasets is shown in Table 1. publication.

Dataset	Description
Regions	180 regions in Manhattan.
Taxi trips	10M taxi trips during one month.
POI data	20K POIs with 13 categories.
Check-in data	100K check-in records.
Crime data	40K crime records during one year.

Table 1: Data description($\mathrm{K} = 10^3, \mathrm{M} = 10^6$).

	Crime Prediction			Land Use		
	MAE↓	RMSE↓	R^2↑	NMI↑	ARI↑	FM↑
node2vec	75.09	104.97	0.49	0.58	0.35	0.10
HDGE	72.65	96.36	0.58	0.59	0.29	0.11
ZE-Mob	101.98	132.16	0.20	0.61	0.39	0.09
MV-PN	92.30	123.96	0.30	0.38	0.16	0.07
MVURE	69.28	96.51	0.57	0.78	0.62	0.41
MGFN	70.21	89.60	0.63	0.68	0.58	0.43
HREP	67.40	86.29	0.65	0.75	0.45	0.43
ReCP	66.90	86.13	0.65	0.78	0.48	0.45
MVJC	**60.80**	**80.72**	**0.70**	**0.82**	**0.72**	**0.70**
Impr.	**9.12%**	**6.28%**	**7.69%**	**5.13%**	**16.13%**	**55.55%**

Table 2: Performance comparison of different methods on two tasks.The indicator FM stands for F-measure.

Baseline Solutions We compare the performance of MVJC with several state-of-the-art region embedding methods. Our baselines cover single-view approaches that rely on mobility data, such as HDGE (Wang and Li 2017), which learns from flow and spatial graphs, and ZE-Mob (Yao et al. 2018), which models regional co-occurrence patterns. We also include a range of multi-view methods: node2vec uses multi-view graphs of regions and concatenate the embeddings of each view (Grover and Leskovec 2016); MV-PN (Fu et al. 2019) focuses on POI and spatial structures; MVURE (Zhang et al. 2021) and MGFN (Wu et al. 2022) utilizes attention-based fusion for mobility and attributes; and HREP (Zhou et al. 2023) employs relation-aware heterogeneous graph embedding. Finally, we benchmark against ReCP (Li et al. 2024), a strong contemporary method based on multi-view contrastive learning.

Main Performance Comparison

To comprehensively evaluate the effectiveness of our proposed MVJC model, we compared it with several state-of-the-art baseline methods on two challenging downstream tasks: crime prediction and land use clustering. The experimental results are shown in Table 2.

From the results in Table 2, we can observe the following:

1. **Limitations of Single-View Methods:** Methods that use only a single data source (e.g., human mobility), such as HDGE and ZE-Mob, perform relatively poorly on both tasks. This result supports the premise that a single data source is insufficient to capture the multifaceted functions and semantics of urban regions, underscoring the

necessity of multi-view approaches.

2. **Superiority of Multi-View Methods:** Methods that integrate multiple information sources (e.g., POIs, check-ins data, and mobility), such as MVURE, MGFN, HREP, and ReCP, generally outperform single-view methods. This validates that fusing multi-dimensional data can generate more comprehensive and robust region representations. Among them, models employing more advanced fusion strategies (like attention mechanisms or contrastive learning), such as MVURE, HREP, and ReCP, typically perform better than those with simple concatenation or weighted averaging, like MV-PN.
3. **Exceptional Performance of MVJC:** Our model, MVJC, establishes a new state-of-the-art. For crime prediction, it reduces MAE by 9.12% compared to the best baseline ReCP. For land use clustering, it notably improves the F-measure by 55.6%. This substantial 55.6% F-measure improvement on a clustering-oriented task provides the strongest evidence for our central claim: the structure-aware module effectively resolves the "false negative" issue, enabling the model to correctly group regions by their underlying function rather than pushing them apart based on instance identity.

The superiority of MVJC is primarily attributed to its unique framework design. The Multi-View Joint Learning module effectively enhances the quality of each view-specific representation through inter-view interactions. Meanwhile, the structure-aware contrastive learning module learns a high-quality consensus representation that contains both view-common and view-specific characteristics through global structure aggregation and alignment. This effectively addresses the "false negative" problem in traditional contrastive learning, enabling it to achieve leading performance in both regression and clustering tasks.

Ablation Study and Parameter Analysis

Ablation Study of Modules To validate the effectiveness of the key modules in the MVJC model, we designed an ablation study. We compared the performance of the full MVJC model with two of its variants:

- **w/o JL**: This variant removes the Multi-view Joint Learning (JL) module. The encoders for each view learn independently, and their initial representations are directly fed into the subsequent structure-aware contrastive learning module.
- **w/o CL**: This variant removes the Multi-view Contrastive Learning (CL) module and uses only the output of the multi-view joint learning module for representation fusion.

The results of the ablation study are shown in Figure 3. From the analysis, we can draw the following conclusions:

1. **Effectiveness of Joint Learning (JL):** Removing JL causes a sharp decline in crime prediction (R^2 drops from 0.70 to 0.33), underscoring the importance of early cross-view interaction. While w/o JL yields a slight, coincidental gain in clustering due to aggressive contrastive focus, the substantial loss in prediction highlights the module's necessity for model generalizability.
2. **Effectiveness of Contrastive Learning (CL):** Removing CL degrades performance across tasks, most notably in land use clustering (ARI decreases by 45.8%). This confirms that structure-aware contrastive learning is crucial for generating discriminative representations by effectively grouping functionally similar regions in the embedding space.

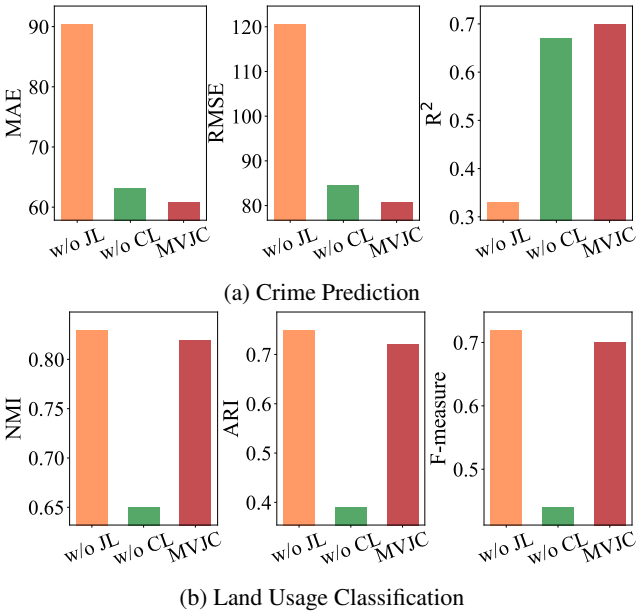


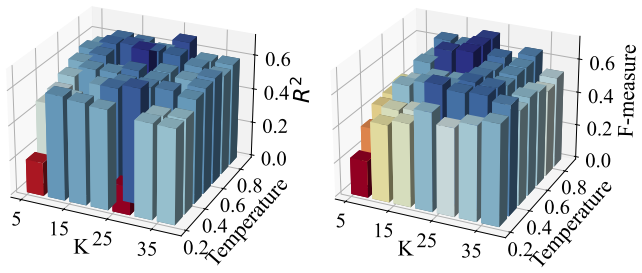
Figure 3: Performance comparison of different modules.

	Crime Prediction			Land Use		
	MAE↓	RMSE↓	R^2↑	NMI↑	ARI↑	FM↑
MVURE	69.28	96.51	0.57	0.78	0.62	0.41
ReCP	66.90	86.13	0.65	0.78	0.48	0.45
w/o-Mob	96.75	134.84	0.42	0.54	0.46	0.42
w/o-POI	78.19	95.48	0.61	0.68	0.57	0.59
w/o-Chk	64.35	85.54	0.68	0.78	0.68	0.65
MVJC	**60.80**	**80.72**	**0.70**	**0.82**	**0.72**	**0.70**

Table 3: Impact of various input views.

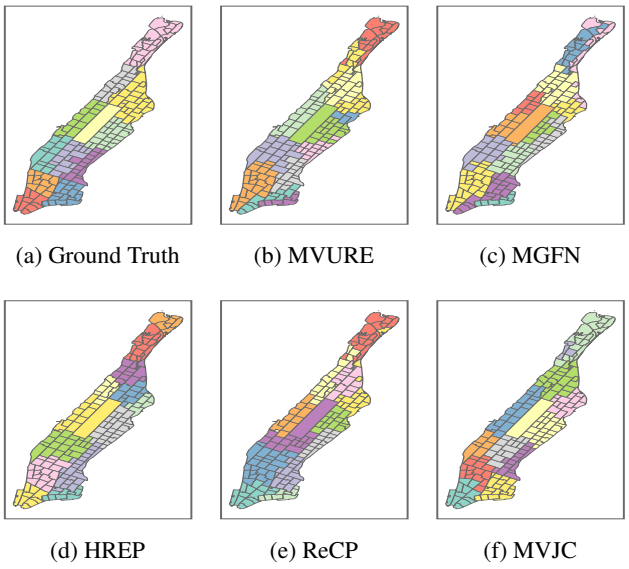
In summary, the results of the ablation study strongly demonstrate the indispensability and effectiveness of the two core modules of the MVJC model. It is the synergy of these two modules that enables MVJC to learn high-quality urban region representations.

Ablation Study of Input Views To assess the contribution of each view, we evaluated variants excluding Mobility (w/o-Mob), POI (w/o-POI), and Check-ins (w/o-Chk), comparing them against the full model and baselines (MVURE, ReCP). Results in Table 3 indicate that mobility is the most critical view, particularly for crime prediction, followed by POI. Notably, even the w/o-Chk variant outper-



(a) R^2 (b) F-measure

Figure 4: Parameter analysis of K and Temperature.



(a) Ground Truth (b) MVURE (c) MGFN

(d) HREP (e) ReCP (f) MVJC

Figure 5: Districts in Manhattan and region clusters.

forms MVURE and ReCP across all tasks (e.g., improving clustering F-measure by 44.44% over the best baseline), demonstrating MVJC's robust feature extraction capability even with reduced inputs.

Parameter Analysis K is the number of neighbors when graphs construct, and the temperature parameter describes the consistency-tolerance dilemma of contrast loss. We vary the parameter K from 5 to 40 in increments of 5, and the temperature parameter from 0.2 to 0.8 in increments of 0.1. The evaluation indicators R^2 and F-measure change accordingly as shown in Figure 4. When setting K=20 and temperature=0.6, MVJC achieves satisfactory performance.

Case Study

To intuitively evaluate our model, we visualize the land use clustering results in Figure 5. The visualization confirms that MVJC's identified clusters align significantly better with ground-truth districts than the baselines. For instance, MVJC correctly groups large, functionally coherent zones, such as commercial hubs, which other methods tend to fragment. This is primarily due to our structure-aware mechanism that mitigates the "false negative" problem. Unlike standard contrastive learning that separates all distinct instances, MVJC preserves the similarity between functionally-alike regions, preventing their incorrect separation. This capability is the key driver behind the 55.6% F-measure improvement in the land use clustering task.

Related Work

Urban Region Representation Learning

Early efforts in urban region representation focused on single-view mobility data, using flow or co-occurrence graphs (Wang and Li 2017; Yao et al. 2018). Subsequent research shifted to multi-view learning, integrating static attributes like POIs (Fu et al. 2019; Zhang et al. 2021) and semantic mobility patterns (Wu et al. 2022). Recent works have further diversified data sources by incorporating urban imagery (Li et al. 2022; Chen et al. 2024) or proposing advanced graph-based aggregation methods (Veličković et al. 2018; Huang et al. 2023; Zhao et al. 2023; Kim and Yoon 2025). Moreover, sophisticated techniques such as prompt learning (Zhou et al. 2023) and information-theoretic contrastive prediction (Li et al. 2024) have been introduced to enhance representation quality.

Multi-view Contrastive Learning

Contrastive learning has become a dominant paradigm for self-supervised representation learning (Chen et al. 2020). The core idea is to learn an embedding space where different views of the same instance are pulled together, while views of different instances are pushed apart. This has been applied to multi-view data (Lin et al. 2021; Yan et al. 2023; Sun et al. 2024; Li et al. 2024), where the different data modalities of a single region are treated as multiple views. However, as previously discussed, this standard formulation presents a "false negative" problem for clustering tasks, as it incorrectly tries to separate all distinct instances. GCFAgg (Yan et al. 2023) introduced the concept of structure-guided contrastive learning to solve this. By first learning a global similarity structure among all samples and then using this structure to downweight the repulsive force between "false negative" pairs, it aligns the contrastive objective with the clustering objective. MVJC adopts and integrates this advanced contrastive mechanism, which is a primary reason for its superior performance in the land use clustering task, as it enables the model to learn representations that are not only discriminative but also form coherent clusters.

Conclusion

In this paper, we propose MVJC, a framework that addresses key challenges in urban region representation, including sub-optimal fusion and the "false negative" problem in contrastive learning. By synergizing multi-view joint learning with a structure-aware contrastive mechanism, MVJC achieves state-of-the-art performance. The core principles are generalizable and could be extended in future work by incorporating more data modalities or adapting the embeddings for specific tasks.

Acknowledgments

This work is supported by the Natural Science Foundation of China No. 62472196, Jilin Science and Technology Research Project 20230101067JC

References

Chen, M.; Li, Z.; Huang, W.; Gong, Y.; and Yin, Y. 2024. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 319–328.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

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.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Huang, C.; Zhang, J.; Zheng, Y.; and Chawla, N. V. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 1423–1432.

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.

Kim, N.; and Yoon, Y. 2025. Effective urban region representation learning using heterogeneous urban graph attention network (HUGAT). *IEEE Access*.

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.

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

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.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

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.

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 *The Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. International Joint Conference on Artificial Intelligence (IJCAI).

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19863–19872.

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 (IJCAI-18)*.

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, Q.; Huang, C.; Xia, L.; Wang, Z.; Li, Z.; and Yiu, S. 2023a. Automated spatio-temporal graph contrastive learning. In *Proceedings of the ACM Web Conference 2023*, 295–305.

Zhang, Q.; Huang, C.; Xia, L.; Wang, Z.; Yiu, S. M.; and Han, R. 2023b. Spatial-temporal graph learning with adversarial contrastive adaptation. In *International Conference on Machine Learning*, 41151–41163. PMLR.

Zhao, Y.; Qi, J.; Trisedya, B. D.; Su, Y.; Zhang, R.; and Ren, H. 2023. Learning region similarities via graph-based deep metric learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(10): 10237–10250.

Zheng, S.; Yuan, W.; and Guan, D. 2022. Heterogeneous information network embedding with incomplete multi-view fusion. *Frontiers of Computer Science*, 16(5): 165611.

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.