Urban region representation learning via dual spatial contrasts

Quan Qin[a], Tinghua Ai[a]*, Weiming Huang[b], Shishuo Xu[c], Mingyi Du[c] and Songnian Li[d]

[a]*School of Resource and Environmental Sciences, Wuhan University, Wuhan, China*

[b]*School of Geography, University of Leeds, Leeds, UK*

[c]*School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China*

[d]*Department of Civil Engineering, Toronto Metropolitan University, Toronto, Canada*

* Correspondence: tinghuaai@whu.edu.cn

Urban region representation learning via dual spatial contrasts

Region representation learning emerges as a new research paradigm to encode urban systems and facilitate geographic mapping. Recent studies have sought to reasonably introduce inductive biases, which refer to prior assumptions that guide model learning, from a geospatial perspective to improve the quality of region representations. However, there remain challenges in incorporating the spatial effects, e.g. spatial dependency and spatial heterogeneity, into inductive biases, as they are critical to the geographic awareness of region representations. In response, we developed a novel region representation learning framework, termed Region Graph Spatial Contrastive Learning (RGSCL), by leveraging building footprints and points of interest (POIs) along with prior spatial knowledge to derive region representations. Specifically, RGSCL first constructed multi-view region graphs with POIs, building footprints, and their spatial proximity, to form a base representation space. Next, the algorithm adopted a contrastive learning mechanism with spatial effects to formulate a dual-spatial-contrast loss function to optimise the representation space. The dual-spatial-contrasts captured POI-building spatial dependency and the region's spatial heterogeneity to compose semantics in region representations. Experimental results demonstrated that RGSCL improved performance in geographic mapping. This study offers new insights into GeoAI from the perspective of inductive biases with respect to spatial effects.

Keywords: spatial representation learning, spatial effect, contrastive learning, building footprints, points of interest

1 Introduction

Representation learning aims to derive discriminative latent feature representations from multimodal data and has driven recent advances in AI for Science across various fields, such as GPT (OpenAI, 2023) and AlphaFold (Abramson et al., 2024). However, the intrinsically multimodal and complex spatial characteristics of geospatial data (Chen et al., 2025; Li et al., 2016) make it particularly difficult to transfer these successes to GeoAI, which warrants greater research attention in spatial representation learning (Gao et al., 2023; Mai et al., 2020). Urban regions are pivotal in sustaining geospatial dynamics and

functionality within urban systems, making them critical entities for spatial representation learning. Region representation learning provides an efficient and reliable solution for characterising regions by mapping comprehensive geographic information into compact and meaningful representations, thereby offering deeper insights into the intricate complexities of regional geographic information. Recently, region representation learning has been increasingly emphasised and applied to effectively address various geographic mapping tasks, such as urban functions, housing prices, traffic flow, and population mapping (Huang et al. 2023; Liu et al. 2025; Zhang et al. 2023a).

Region representation learning involves deriving region-related geospatial information from ubiquitous geospatial data to form meaningful representations. The specific knowledge and meanings embedded in these representations during learning directly influence their practical value in downstream applications, which depend on the choice of inductive bias (a.k.a. learning bias). Inductive bias is crucial in representation learning because it encourages the learning algorithm to prioritise solutions with certain properties (Battaglia et al., 2018; Goyal & Bengio, 2022). Today, most advances have benefited from diverse inductive biases, enabling models to learn region representations and their corresponding meanings from multimodal geospatial data. Despite their success, most of these studies rely on the intrinsic inductive biases of neural networks, for example, the locality of CNNs and the sequentiality of RNNs, to learn spatial relations between geospatial entities. These inductive biases, directly derived from the AI community, overemphasise fitting statistical distributions while neglecting the crucial geospatial distributions inherent in geospatial data. These studies have not sufficiently adapted inductive biases to reasonably account for geospatial characteristics and problems, resulting in region representations that lack geographic meaning and exhibit poor generality due to their limited retained information. Such limitations further indicate that

region representation learning in the literature has yet to systematically establish a theoretical framework aligned with geospatial properties.

In this context, region representation learning could fundamentally benefit from the introduction of specialised inductive biases from a geospatial perspective. In fact, spatial samples, different modalities, and their spatial distribution phenomena share intrinsic connections and patterns; that is, they are influenced by spatial effects and geographic laws (Goodchild 2004; Liu et al. 2023). Therefore, spatial effects can serve as an effective inductive bias to guide models in learning geographically meaningful solutions within an infinite representation space. To address this challenge, we developed a novel region representation learning framework called Region Graph Spatial Contrastive Learning (RGSCL), which integrates contrastive learning, spatial effects, and graph theory to form effective geospatial inductive biases for informing region representations. RGSCL leverages the semantic and morphological features of POIs and building footprints and employs graph encoding to capture spatial proximity, thereby shaping a base representation space. It further utilises a contrastive learning mechanism to sculpt an informative and discriminative representation space by synergising spatial heterogeneity and spatial dependence. Thus, RGSCL leverages the building footprint and POI information with spatial effects to effectively derive task-agnostic, general-purpose region representations, which can be further fine-tuned for various downstream geographic mapping tasks.

The remainder of this paper is organised as follows: Section 2 reviews related studies; Section 3 provides a detailed description of the RGSCL methodology; Section 4 outlines the experimental setup and presents the experimental results and analysis; Section 5 discusses the limitations and potential of RGSCL; and Section 6 concludes the study.

2 Related work

2.1 Encoder for region representation learning

The encoder serves to interpret and compress the input geospatial data, embedding region-related geospatial information into latent representations. Diverse encoder architectures with distinct inductive biases have inspired numerous studies aiming to capture the spatial effects corresponding to these biases, thereby informing region representation learning.

Earlier advances benefited from sequential models, whose inductive bias of sequentiality enabled the learning of sequentially structured spatial relations for region representation learning. The spatial relationships among geospatial entities were transformed into spatial sequences using specific spatial sampling strategies. Models such as Word2Vec (Qin et al., 2022a, 2022b; Yao et al., 2017), Doc2Vec (Niu & Silva, 2021), and Place2Vec (Yan et al., 2017) leverage the co-occurrence of geospatial entities within these sequences to capture spatial co-locations for region representation learning. However, sequential models are inherently limited by their reliance on spatial sequences, which are insufficient for modelling the complexity of real spatial relations for region representations.

Recently, graph neural networks (GNNs) have emerged as promising encoders for region representation learning. Prior to graph learning, the spatial relations among geospatial entities were appropriately transformed into spatial graphs. The message-passing mechanism was then employed to capture spatial proximity on a Delaunay Triangulation (DT) graph (Huang et al., 2023; Kong et al., 2024; Xu et al., 2022; Yan et al., 2019), road graph (Huang et al., 2025), and k-nearest neighbour (KNN) graph (Zhang et al., 2024c). The relatively arbitrary inductive bias of GNNs (Battaglia et al., 2018) provides information on region representations with complex geospatial relations beyond

Euclidean proximity, such as multiview spatial relations based on POI semantic distance, spatial distance, and mobility (Zhang et al., 2023c), dynamically adjusted traffic flow relations (Zhang et al., 2024a), and heterogeneous spatial-interaction relations between POIs (Qin et al., 2025). Despite their effectiveness in embedding spatial relations, transforming implicit geospatial relations into explicit graph structures remains challenging.

Some studies have explored leveraging of the representational capabilities of pretrained large-language models (LLMs). One common method is to interpret regional geospatial semantics from the perspective of natural language, for example, by generating embeddings for POIs (Zhang et al., 2023a, 2021) or textual descriptions of regions from street-view images (Huang et al., 2024) and aerial images (Yan et al., 2024). Another method involves empowering LLMs through instruction-tuning using geospatial data (Li et al., 2024b). However, LLMs for region representation inevitably suffer from covariance shifts and lack spatial effect inductive biases, limiting their capability to capture nuanced geospatial semantics.

2.2 Learning paradigm for region representation learning

Learning paradigms determine the training process and optimisation strategies for encoders in region representation learning. Supervised learning is a common paradigm that directly leverages task-related supervision signals to efficiently extract task-specific regional features, forming dedicated representation spaces tailored for particular geographic mapping tasks such as traffic-condition mapping (Zhang et al., 2024d). However, it overlooks task-agnostic spatial effects, as it overemphasises fitting specific downstream tasks. As a result, the learned region representation space is limited in its effectiveness for specific narrow downstream tasks. Furthermore, the sparsity of geospatial data labels (Chen et al., 2025) hinders achieving meaningful effects on scaling

laws (Bahri et al., 2024).

In contrast, self-supervised learning (Gui et al., 2024; Liu et al., 2023a) is a promising paradigm that enhances the generality of region representations derived from vast amounts of unlabelled geospatial data (Chen et al., 2020, 2025; Radford et al., 2021). One such approach is predictive learning such as context prediction (Qin et al., 2022a, 2022b; Yao et al., 2017), which enhances the spatial understanding of region representations from an NLP perspective. However, these methods remain inadequate for exploring spatial effects, which are important for urban representation. Recently, another paradigm—contrastive learning—has shed light on region–representation learning. Pulling positive samples closer and pushing negative samples away, it offers an effective solution for shaping region representations. Some studies have leveraged multimodal contrasts within regions to learn more informative representations, for example, contrasting remote sensing images and POIs (Bai et al., 2023; Xi et al., 2022), LLM-based text and remote sensing images (Yan et al., 2024), and text-location pairs from POIs (Wang et al., 2024). Additionally, a stronger inductive bias tied to spatial effects has been introduced into region representations to convey a more generic geospatial information. Li et al. (2023) compared regions and their adjacent regions to capture spatial proximity on a regional scale. Huang et al. (2023) captured hierarchical spatial relations under the spatial-scale effect by contrasting the POI-region-city hierarchy. Zhou et al. (2024) integrated the First and Third Laws of Geography as model inductive bias to define a contrast strategy and enhance spatial effect awareness.

Previous studies have explored region representation learning from the perspectives of encoder design and learning paradigms; however, two main challenges remain. First, inductive biases are often inherited from general-purpose AI models to accommodate geospatial properties only roughly, rather than being deliberately designed

from spatial effects to guide representation learning. The second challenge is an over-reliance on spatial proximity, which overlooks spatial heterogeneity. As emphasised by Goodchild (2004), spatial heterogeneity represents a more fundamental geographic principle than the First Law of Geography, serving as the primary motivation for this study.

3 Methodology

3.1 Overview

This study introduces the RGSCL framework, built upon theories of contrastive learning, geographic laws, and graph theory, which enables effective self-supervised region representation learning. Fig. 1 shows an overview of the RGSCL framework, which consists of three components: (1) Multiview region graph encoding—region graphs are defined based on the features and adjacency relations of POIs and building footprints. Graph encoding is then employed to capture the spatial proximity to derive a base region representation space; (2) Dual spatial contrastive learning—it utilises contrastive learning mechanisms to synergise spatial dependence and spatial heterogeneity to further sculpt the region representation space; and (3) Fine-tuning geographic mapping—the learned region representations are adapted to various geographic mapping tasks using few-shot learning.

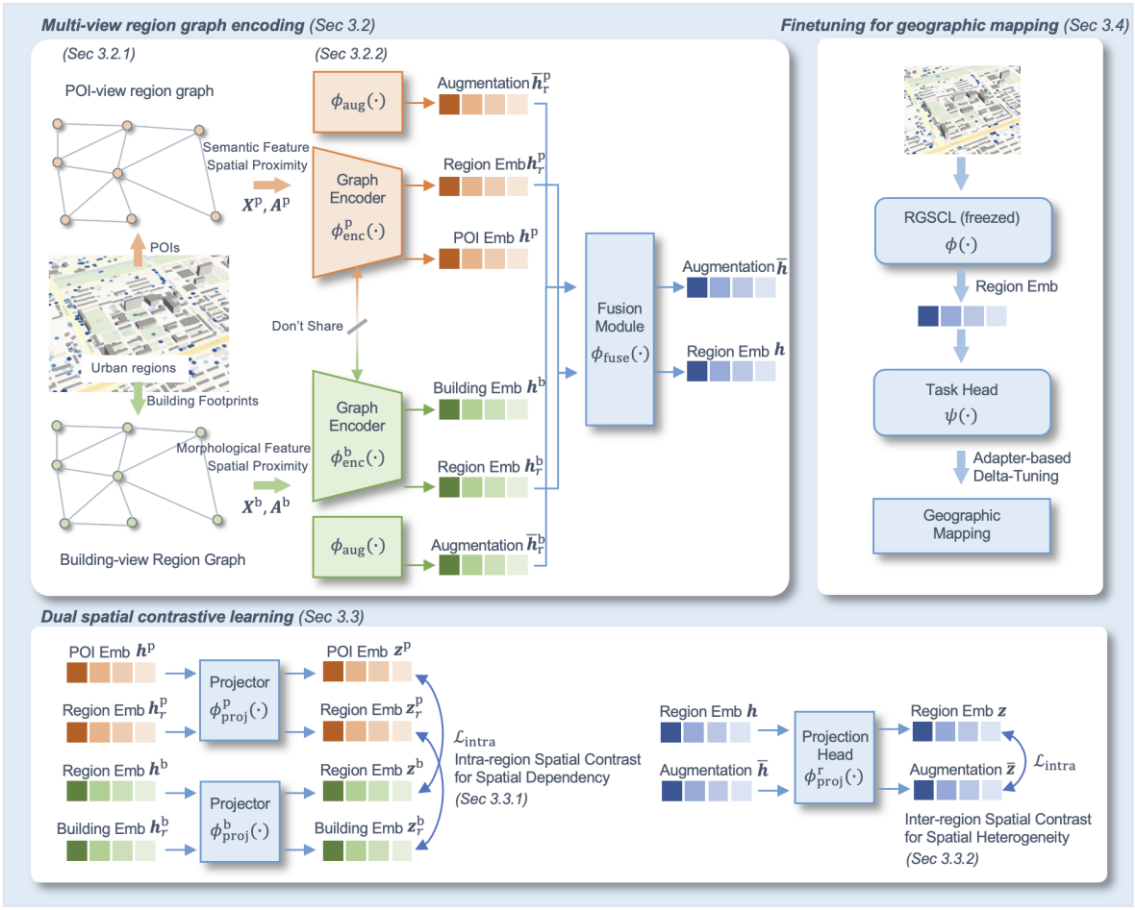

Figure 1. Framework overview.

3.2 Encoding multi-view region graph with capturing spatial proximity

3.2.1 Constructing multi-view region graph

Building footprints and POIs provide complementary regional geographic information and their spatial proximity underpins spatial autocorrelation under the First Law of Geography, which is crucial for shaping region representations. In view of this, we preserve their respective spatial proximity through graph adjacency, and propose a multi-view region graph $G^m = (\mathcal{V}^m, \mathcal{E}^m, \boldsymbol{X}^m, \boldsymbol{A}^m)$, $m \in \{\mathrm{p}, \mathrm{b}\}$ denotes the region graph in POI- and building-view, where $\mathcal{V}^m = \{v_i^m\}_{i=1}^{N_G^m}$ denotes vertexes, $\mathcal{E}^m \subseteq \mathcal{V}^m \times \mathcal{V}^m$ denotes the edges, N_G^m denotes the number of vertices. $\boldsymbol{X}^m \in \mathbb{R}^{N_G^m \times F_m}$ and $\boldsymbol{A}^m \in \{0,1\}^{N_G^m \times N_G^m}$ denotes the node feature matrix and adjacency matrix, respectively, where

$\boldsymbol{x}_i^m \in \mathbb{R}^{F_m}$ is the feature vector of v_i^m and $\boldsymbol{A}_{ij}^m = 1$ iff $\left(v_i^m, v_j^m\right) \in \mathcal{E}^m$, F_m denotes the feature dimension corresponding to the view. The construction of the multi-view region graph is illustrated in Fig. 2.

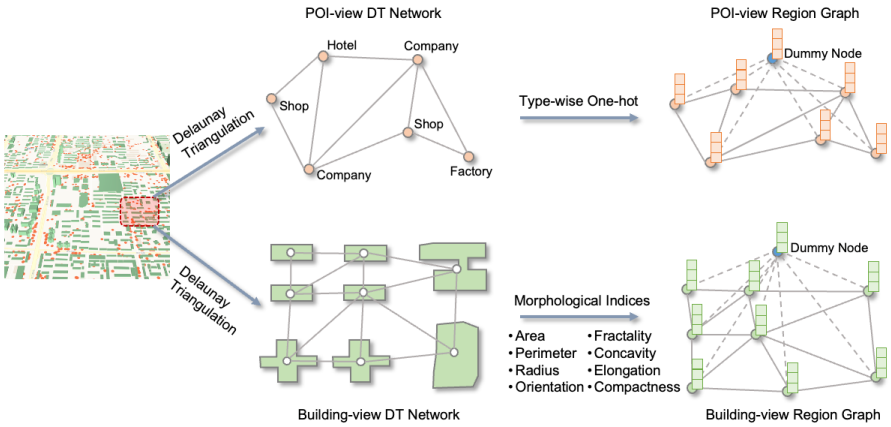

Figure 2. Construction of the multi-view region graph.

As shown in Fig. 2, the nodes in the region graph are defined by individual building footprints and POIs. To initialise the features of the footprints, several morphological indices are used to represent the two-dimensional morphological characteristics of buildings, including area, perimeter, radius, orientation, compactness, fractality, concavity, and elongation. These features are normalised using Z-score normalisation. More details on the morphological indices can be found in Appendix A1 and Yan et al. (2019). The initial features of the POIs are defined using one-hot vectors based on the POI type.

The edges in the region graph are defined based on the spatial proximity between the building footprints and POIs within the region. We employ the DT algorithm (Yan et al., 2019) to transform the respective spatial proximities into adjacency matrix $\boldsymbol{A}^m$. In a DT-based region graph, the nodes are connected to their spatially adjacent counterparts as much as possible, thus establishing valuable spatial proximity within the region graph. This enables the model to capture spatial proximity during the graph-encoding process.

To effectively derive region-level representations from node-level representations on a graph, similar to the special tokens adopted in language models such as Doc2Vec and BERT (Devlin et al. 2018; Le and Mikolov 2014) to obtain document- or sentence-level representations, we introduced a dummy node v_r^m for each region to represent region-level information. This facilitates the learning of the graph structure and mitigates the oversmoothing problem during graph readout (Liu et al. 2022). In order to stabilise model training and ensure that each view learns its own information, separate dummy nodes are assigned to the region graphs of the two views, rather than using a single dummy node to connect both, i.e., $\mathcal{V}_r^m = \{v_{r_i}^m\}_{i=1}^{N_r} \subseteq \mathcal{V}^m$. Each region dummy node is connected to all buildings and POI nodes within the region, allowing for adaptive aggregation of node information during training to update the entire region representation. In addition, each node benefits from the global information conveyed by the dummy node.

3.2.2 Encoding multi-view region graph

To facilitate the encoding of both node information and graph structure (spatial proximity) in multi-view region graphs, we adopt a graph attention network (GAT) model (Veličković et al. 2018) as the base graph encoder for the region embedding network. GAT offers a plausible solution for capturing heterogeneous spatial interactions between geospatial entities, that is, attention-based message passing among geographic nodes. We consider a dedicated encoder for each view, denoted as $\phi_{\text{enc}}^m: \mathbb{R}^{F_m} \to \mathbb{R}^F$, where $m \in \{\text{p}, \text{b}\}$, to more effectively learn the information within each view. Prior to feeding the multi-view region graph into the encoder, an affine transformation followed by batch normalisation (BN) (Ioffe and Szegedy 2015) is applied to project the multi-view features into a latent space with consistent dimensions.

$$\boldsymbol{x}_i^{\prime,m} = \mathrm{BN}(\boldsymbol{W}^f \boldsymbol{x}_i^m + \boldsymbol{b}^f) \tag{1}$$

where $\boldsymbol{W}^f \in \mathbb{R}^{F_m \times F}$ is weight matrix and $\boldsymbol{b}^f$ is bias vector. Subsequently, the encoding process for spatial proximity is designed as a one-layer GAT with a residual connection and batch normalisation and is formally defined as

$$\phi_{\mathrm{enc}}(\boldsymbol{x}_i^m) = \mathrm{BN}\left(\mathrm{GAT}\left(\boldsymbol{x}_i^{\prime,m}\right) + \boldsymbol{x}_i^{\prime,m}\right) \tag{2}$$

where the GAT adopts a multi-head attention mechanism to capture the spatial interactions between geospatial entities from different perspectives and is defined as follows:

$$\mathrm{GAT}\left(\boldsymbol{x}_i^{\prime,m}\right) = \sigma_f \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in \mathcal{C}_i} \varphi_{ij}^k W_\varphi^k \boldsymbol{x}_j^{\prime,m} \right) \tag{3}$$

where σ_f is a ReLU nonlinearity, $\mathcal{C}_i$ is a spatial context set (i.e., 1-hop adjacent nodes, including i) of node i in the graph, K denotes the number of heads for multi-head mechanism, φ_{ij}^k denotes the attention from k-th head, and φ_{ij} is defined as follows:

$$\varphi_{ij} = \sigma_\varphi \left(\sigma_a \left(\boldsymbol{a}^{\mathrm{T}} \left[\boldsymbol{W}_a \boldsymbol{x}_i^{\prime,m} \parallel \boldsymbol{W}_a \boldsymbol{x}_j^{\prime,m} \right] \right) \right) \tag{4}$$

where σ_φ is a softmax to normalise attention weights, σ_a is a LeakyReLU nonlinearity, $\boldsymbol{a} \in \mathbb{R}^{2F}$ is a weight vector for φ, $\boldsymbol{W}_\alpha \in \mathbb{R}^{F \times F}$ is a shared linear transformation that applies to all nodes, $\cdot^{\mathrm{T}}$ represents transposition, and $\parallel$ is feature-wise concatenation operation.

We encode the spatial proximity of both views using the graph encoder, diffusing node features through the graph adjacency and aggregating them to update the node features: $\boldsymbol{H}^m = \phi_{\mathrm{enc}}(\boldsymbol{X}^m)$. This yields multi-view region graph features that are then

embedded into a unified representation space. We further employ an additive attention-based fusion module (Bahdanau et al. 2015; Xi et al. 2022) $\phi_{\text{fuse}}: \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}^F$ for poi-building joint feature encoding, zipping complementary regional features to generate region embeddings $\boldsymbol{H}$. Intuitively, it learns and assigns weights to the features of both views according to their importance and relevance to the regions. Specifically, given $\boldsymbol{h}_r^{\text{b}}$ (building-view region representation) and $\boldsymbol{h}_r^{\text{p}}$ (POI-view region representation) from the same region, we can adaptively fuse the two vectors using an attention mechanism, resulting in a more comprehensive joint-view region representation that captures essential multi-view characteristics.

$$\boldsymbol{h}_r = \sum_{m \in \{\text{p,b}\}} \beta^m \boldsymbol{h}_r^m \tag{5}$$

$$\alpha^m = \boldsymbol{W}_2^\alpha \sigma_\alpha (\boldsymbol{W}_1^\alpha \boldsymbol{h}_r^m + \boldsymbol{b}^\alpha) \tag{6}$$

$$\beta^m = \frac{\exp(\alpha^m)}{\sum_{m \in \{\text{p,b}\}} \exp(\alpha^m)} \tag{7}$$

where σ_α is Tanh nonlinearity, $\boldsymbol{W}_1^\alpha$ and $\boldsymbol{W}_2^\alpha$ are weight matrices, and $\boldsymbol{b}^\alpha$ is bias vector.

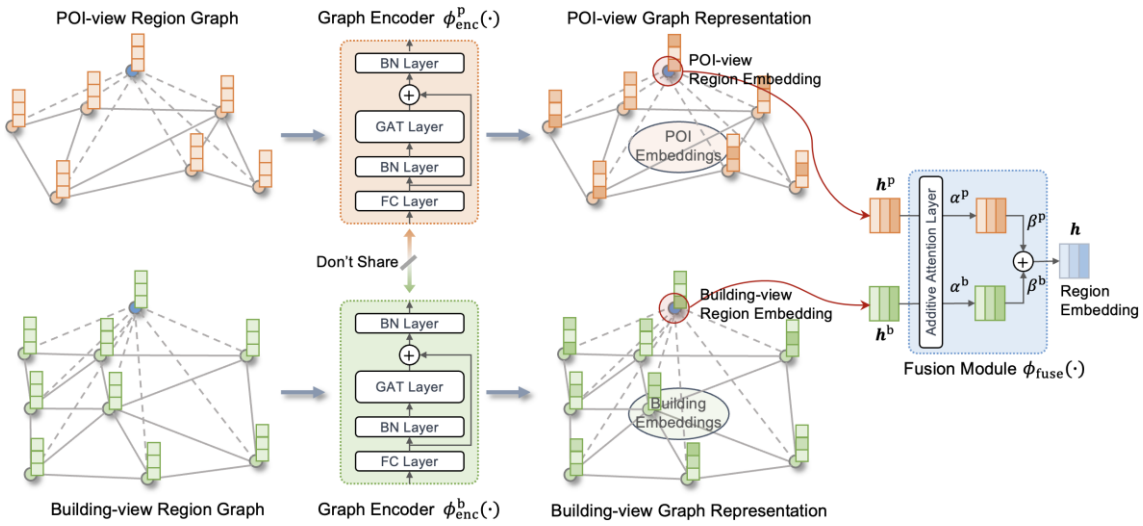

Figure 3. Multi-view joint feature encoding with region embedding network.

3.3 Model training with capturing spatial dependency and spatial heterogeneity

After defining the forward propagation of the RGSCL, an effective optimisation objective is required to guide backpropagation. The spatial distributions of building footprints and POIs emerge under the influence of various spatial effects, which can serve as a guiding principle for defining a geographically meaningful optimisation direction for RGSCL. Accordingly, we designed a dual spatial contrastive learning objective aligned with the contrastive learning paradigm to jointly capture spatial dependency and spatial heterogeneity.

3.3.1 Capturing POI-building spatial dependency via intra-region spatial contrast

POI semantics and building morphology have strong intrinsic connections, with features from both views typically spatially interdependent. For instance, factory-related POIs are generally associated with flat building footprints. Therefore, the natural correlation between POIs and building footprints serves as an effective proxy for spatial dependency. Inspired by the fact that multiview graph contrastive learning enhances representation quality by maximising mutual information (MI) across different graph views (Hassani and Khasahmadi 2020; Zhou et al. 2024), we employed intra-region spatial contrastive learning to capture the spatial dependency between views by maximising the MI of the region graph of different views within the same region, forming unified region representations that align the building morphology and POI semantics.

For negative sampling of the intra-region spatial contrast, given the region representation of one view for region r_i, the node representations from the other view are paired with it to calculate the contrastive loss. Pairs belonging to the same region are treated as positive samples, whereas pairs from different regions were treated as negative

samples; that is, (v_j, v_{r_i}) for positive samples and $(\tilde{v}_j, v_{r_i})$ for negative samples, where v denotes positive samples and $\tilde{v}$ denotes negative samples. To better achieve alignment and uniformity (Wang and Isola 2020) to benefit the representation space, we consider the two individual views (i.e. poi- and building-views) as the two contrasting aspects of a sample pair, that is, $(v_j^{\mathrm{p}}, v_{r_i}^{\mathrm{b}})$, $(\tilde{v}_j^{\mathrm{p}}, v_{r_i}^{\mathrm{b}})$ and $(v_j^{\mathrm{b}}, v_{r_i}^{\mathrm{p}})$, $(\tilde{v}_j^{\mathrm{b}}, v_{r_i}^{\mathrm{p}})$. Following this, we implemented a cross-contrast strategy to align intra-region geographic information across the two views. The intra-region spatial contrast is defined by maximising the MI between region representations (global regional information) from one view and node representations (local regional information) from the other when they are spatially consistent at the regional scale, while minimising the MI between region representations from one view and node representations from the other when they are spatially inconsistent.

Since projection head is conducive to representation learning during calculation of contrastive loss (Xue et al. 2024), we adopt a two-layer feed-forward neural network $\phi_{\mathrm{proj}}: \mathbb{R}^F \to \mathbb{R}^F$ as the projection head to map the multi-view region embeddings into another latent space to further calculate contrastive loss: $\boldsymbol{z}_i^{\mathrm{p}} = \phi_{\mathrm{proj}}^{\mathrm{p}}(\boldsymbol{h}_i^{\mathrm{p}})$, $\boldsymbol{z}_i^{\mathrm{b}} = \phi_{\mathrm{proj}}^{\mathrm{b}}(\boldsymbol{h}_i^{\mathrm{b}})$, where $\phi_{\mathrm{proj}}(\boldsymbol{h}) = \boldsymbol{W}_2^g \sigma_g(\boldsymbol{W}_1^g \boldsymbol{h} + \boldsymbol{b}^g)$, where σ_g is a ELU nonlinearity, $\boldsymbol{W}^g$ is weight matrix, and $\boldsymbol{b}^g$ is bias vector. Subsequently, the intra-region spatial contrastive loss is formally expressed as a cross-contrast as follows:

$$\mathcal{L}_{\mathrm{intra}} = \mathcal{L}_{\mathrm{poi}} + \mathcal{L}_{\mathrm{building}} \tag{8}$$

$$\mathcal{L}_{\mathrm{poi}} = -\mathbb{E}_{(v^{\mathrm{b}}, v_r^{\mathrm{p}})\sim\mathcal{M}}\left[\log \mathcal{D}_{\mathrm{intra}}(\boldsymbol{z}^{\mathrm{b}}, \boldsymbol{z}_r^{\mathrm{p}})\right] - \mathbb{E}_{(\tilde{v}^{\mathrm{b}}, v_r^{\mathrm{p}})\sim\mathcal{M}}\left[\log\left(1 - \mathcal{D}_{\mathrm{intra}}(\tilde{\boldsymbol{z}}^{\mathrm{b}}, \boldsymbol{z}_r^{\mathrm{p}})\right)\right] \tag{9}$$

$$\mathcal{L}_{\mathrm{building}} = -\mathbb{E}_{(v^{\mathrm{p}}, v_r^{\mathrm{b}})\sim\mathcal{M}}\left[\log \mathcal{D}_{\mathrm{intra}}(\boldsymbol{z}^{\mathrm{p}}, \boldsymbol{z}_r^{\mathrm{b}})\right] - \mathbb{E}_{(\tilde{v}^{\mathrm{b}}, v_r^{\mathrm{p}})\sim\mathcal{M}}\left[\log\left(1 - \mathcal{D}_{\mathrm{intra}}(\tilde{\boldsymbol{z}}^{\mathrm{p}}, \boldsymbol{z}_r^{\mathrm{b}})\right)\right] \tag{10}$$

where $\mathcal{M} = \mathcal{V}^{\mathrm{p}} \times \mathcal{V}^{\mathrm{b}}$, $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$ denote the positive and negative samples to anchor $\boldsymbol{z}_r$, respectively. Specifically, $\boldsymbol{z}$ and $\boldsymbol{z}_r$ are the embeddings of a node and dummy node located in the same region, respectively, and $\tilde{\boldsymbol{z}}$ is the embedding of a node located outside this region. Discriminator $\mathcal{D}: \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$ is employed as a proxy for maximising the MI and is calculated by a bilinear scoring function (Ishmael et al., 2018), i.e., $\mathcal{D}(\boldsymbol{u}, \boldsymbol{v}) = \sigma_{\mathcal{D}}(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{D}} \boldsymbol{v})$, where $\sigma_{\mathcal{D}}$ is a sigmoid nonlinearity and $\boldsymbol{W}_{\mathcal{D}}$ is a learnable scoring matrix.

$\mathcal{L}_{\text{intra}}$ encourages the model to pull closer the region representations of the POI and building-views that are spatially consistent at the regional scale (positive sample pairs), while pushing away the representations of the two views that are spatially inconsistent (negative sample pairs). By capturing the POI-building spatial dependency, the consistency of multi-view information in region representations is enhanced, facilitating more informative region representations.

3.3.2 Capturing spatial heterogeneity at regional scale via inter-region spatial contrast

Spatial heterogeneity emphasises the uneven distribution of geographic information, which is crucial for preserving the uniqueness of regional semantics in representations. To this end, we employ inter-region spatial contrast to capture spatial heterogeneity. For negative sampling, representations from other regions serve as negative samples to highlight the differences between regions. To understand the uniqueness of each region, we apply data augmentation to the region representations as positive samples. The inter-region spatial contrast is defined by maximising the MI between the region representation and its augmentations, while minimising the MI between representations from different regions.

Augmentations generated by appropriate data-augmentation strategies maintain similar representations, thereby enhancing the robustness of region representations

against variations and noise. Inspired by Shen et al. (2023), we employ a learnable adaptive graph augmentation module ϕ_{aug}: $\mathbb{R}^F \to \mathbb{R}^F$ to introduce subtle perturbations in node relations (edges) to simulate data noise without altering the essential characteristics of the samples, thereby creating adaptive graph augmentations (i.e., positives). We opt for a one-layer GAT to replace the original graph encoder in defining ϕ_{aug}, generating multi-view augmented features $\bar{\boldsymbol{H}}^m$, which are further processed through the feature fusion module ϕ_{fuse} to obtain the augmented region feature $\bar{\boldsymbol{H}}$. Subsequently, a projection head is utilised before calculating the contrastive loss, i.e., $\boldsymbol{z}_i = \phi_{\text{proj}}^{\text{r}}(\boldsymbol{h}_i)$. The interregion spatial contrastive loss is defined as follows:

$$\mathcal{L}_{\text{inter}} = -\mathbb{E}_{(r,\tilde{r})\sim\mathcal{R}}\left[\log \mathcal{D}_{\text{inter}}(\bar{\boldsymbol{z}}, \boldsymbol{z}) + \log\left(1 - \mathcal{D}_{\text{inter}}(\tilde{\boldsymbol{z}}, \boldsymbol{z})\right)\right] \tag{11}$$

where $\bar{\boldsymbol{z}}$ and $\tilde{\boldsymbol{z}}$ are the positive and negative samples used to anchor $\boldsymbol{z}$.

$\mathcal{L}_{\text{inter}}$ encourages the model to pull closer region representations and their augmentations (positives) into the latent space, thereby enhancing the robustness. Meanwhile, it pushes away representations from different regions, leading to the capture of spatial heterogeneity on a regional scale. The push-pull process fosters more discriminative and robust region representation.

3.3.3 Joint optimization via dual spatial contrasts

Our RGSCL aims to obtain low-dimensional, high-quality, generic region embeddings by pretraining the model with the two spatially constrained learning objectives described above. To harmonise intra- and inter-region contrasts to form dual spatial contrasts, the network is guided to synergise spatial dependency and spatial heterogeneity to shape region representations. The joint optimisation objective (loss function) for RGSCL was derived by combining the proposed dual spatial contrast losses with an ℓ_2 regularization

loss to prevent the model from overfitting on either of the two objectives. This can be formalised as follows:

$$\mathcal{L}_{\phi} = \lambda \mathcal{L}_{\text{intra}} + (1-\lambda)\mathcal{L}_{\text{inter}} + \mu \|\boldsymbol{\Theta}\|^2 \tag{12}$$

where λ is the weight parameter to balance the intra- and inter-region spatial contrastive losses, $\boldsymbol{\Theta}$ denotes all the learnable parameters of RGSCL for ℓ_2 regularisation with the strength of μ.

Fig. 4 illustrates the dual spatial contrast process of RGSCL. On the one hand, $\mathcal{L}_{\text{intra}}$ captures POI-building spatial dependency, aligning multi-view regional semantics, and enhancing the consistency of geographic information in region representations, resulting in more informative representations. On the other hand, $\mathcal{L}_{\text{inter}}$ stimulates the model to understand spatial heterogeneity at a regional scale, resulting in more discriminative representations. Finally, the model learns prior spatial knowledge (i.e. spatial dependency and spatial heterogeneity) via the joint optimisation of dual spatial contrasts, informing the region representation with more high-level and generic features. All the learnable parameters are optimised by $\mathcal{L}_{\phi}$ using stochastic gradient descend method. The RGSCL optimisation process is illustrated in Algorithm 1.

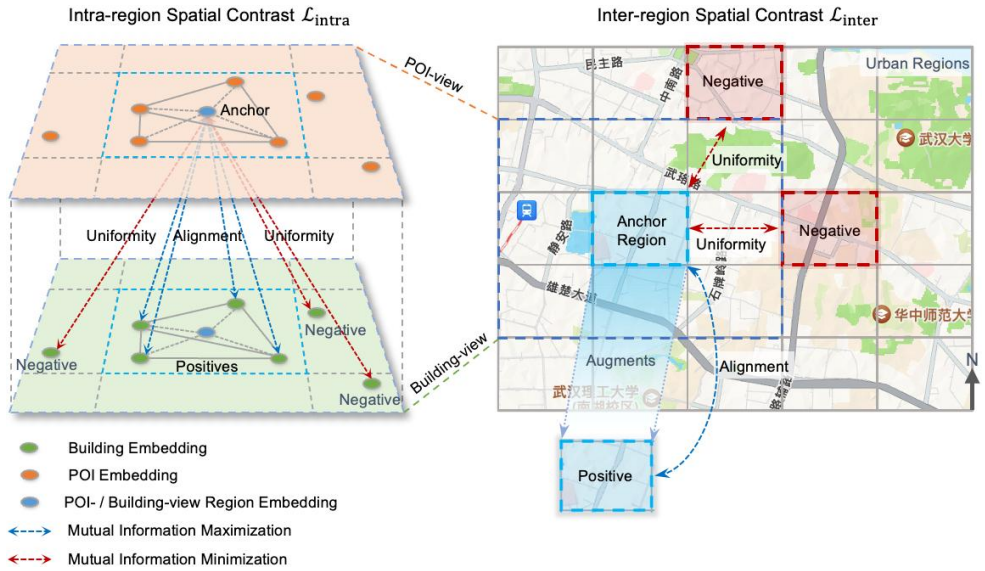

Figure 4. Illustration of dual spatial contrasts. Note that intra-region spatial contrast $\mathcal{L}_{\text{intra}}$ shown here represents only $\mathcal{L}_{\text{poi}}$ from the cross-contrast strategy while the other component $\mathcal{L}_{\text{building}}$ is inverse in form to $\mathcal{L}_{\text{poi}}$.

Algorithm 1 Learning process of RGSCL

Input: Multi-view region graph G^{p}, G^{b}
learning rate η
training epochs E
Output: Region embeddings $\boldsymbol{H}$

Initialize model parameters in each module;
Initialize POI embeddings $\boldsymbol{X}^{\text{p}}$ using one-hot;
Initialize Building embeddings $\boldsymbol{X}^{\text{b}}$ via feature engineering;
Construct POI-view Graph G^{p}, building-view Graph G^{b};
for $e \leftarrow 1$ to E **do**
    Conduct graph encoder $\phi_{\text{enc}}^{\text{p}}$ on $\boldsymbol{X}^{\text{p}}$ to get POI-view embeddings $\boldsymbol{H}^{\text{p}}$;
    Conduct graph encoder $\phi_{\text{enc}}^{\text{b}}$ on $\boldsymbol{X}^{\text{b}}$ to get building-view embeddings $\boldsymbol{H}^{\text{b}}$;
    Conduct attention-based fusion module ϕ_{fuse} to get region embeddings $\boldsymbol{H}$;
    Conduct adaptive augmentation module ϕ_{aug} to get augmented region embeddings $\bar{\boldsymbol{H}}$;
    Apply projection head $\phi_{\text{proj}}^{\text{p}}$ on $\boldsymbol{H}^{\text{p}}$ to get $\boldsymbol{Z}^{\text{p}}$;
    Apply projection head $\phi_{\text{proj}}^{\text{b}}$ on $\boldsymbol{H}^{\text{b}}$ to get $\boldsymbol{Z}^{\text{b}}$;
    Apply projection head $\phi_{\text{proj}}^{\text{r}}$ on $\boldsymbol{H}$ and $\bar{\boldsymbol{H}}$ to get $\boldsymbol{Z}$ and $\bar{\boldsymbol{Z}}$;
    Conduct intra-region contrastive learning following Eq. 8 to capture spatial dependency;
    Conduct inter-region contrastive learning following Eq. 11 to capture spatial heterogeneity;
    Calculate the dual spatial contrastive loss of RGSCL following Eq. 12;
    for each parameter $\theta \in \boldsymbol{\Theta}$ **do**
        $\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}_{\phi}}{\partial \theta}$;
    end
end
return all parameters $\boldsymbol{\Theta}$

3.4 Finetuning for geographic mapping tasks

The RGSCL infuses region representations with prior spatial knowledge via dual spatial contrasts, and its training process does not rely on supervised signals from any particular downstream task. Consequently, RGSCL learns generic region representations that are applicable to various downstream tasks. Following a pretraining-finetuning paradigm called adapter-based delta-tuning (Ding et al. 2023), we freeze the pretrained RGSCL parameters and append an additional multilayer perceptron (MLP) ψ: $\mathbb{R}^{F} \rightarrow \mathbb{R}^{F'}$ as a task head to adapt to specific geographic mapping tasks. Supervision signals from the downstream tasks were used for fine-tuning and performance evaluation.

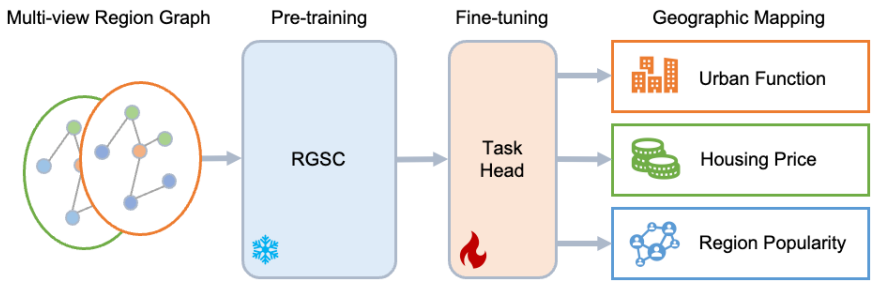

Figure 5. Fine-tuning flow of RGSCL for geographic-mapping tasks.

We evaluated the effectiveness of RGSCL by applying it to several representative geographic-mapping tasks, including urban function mapping, housing price mapping, and region popularity mapping (Huang et al., 2023; Zhang et al., 2023b, 2024b). The problem of urban function mapping is a multi-class classification task, in which each urban region is assigned to the correct land-use category based on region embeddings. Housing price mapping and region popularity mapping were formulated as regression problems in which continuous values were predicted for each region.

4 Results

4.1 Experimental setups

4.1.1 Study area and data

We conducted experiments in Beijing, China, to validate the effectiveness of the RGSCL. Five datasets were used in the study. POI and building-footprint data served as inputs to the RGSCL, whereas three other datasets were used as the ground truth for the corresponding geographic mapping tasks. After data preprocessing, 2,301 regions were retained, resulting in POI-view graphs with 157,232 nodes and 1,145,332 edges, and building-view graphs with 302,636 nodes and 2,336,228 edges. Spatial processing and geographic mapping were performed using the CGCS2000 Projected Coordinate System (EPSG: 4548).

- POI: 240 K POIs were collected from Amap (a.k.a. Gaode Map, https://ditu.amap.com/), including 100 second-level type labels and location information. Duplicate POIs and POIs with missing attributes were excluded.
- Building footprint: 300 K building footprints, including 2D shape and location information, were sourced from OSM (https://www.openstreetmap.org/). Buildings with areas smaller than 20 m² were excluded to reduce noise.
- Essential Urban Land Use Categories (EULUC) (Gong et al. 2020): The EULUC is a dataset of urban land use in China. Approximately 2.5 K regions were obtained and categorised into two hierarchical label levels. The spatial unit from the EULUC served as the regional unit in this study, with first-level labels (residential, commercial, industrial, transportation, public management, and service) used as the ground truth for urban function mapping. Regions with fewer than ten POIs or building footprints were excluded because of data sparsity.
- Housing price: 150 K housing price data points were collected from Lianjia (https://lianjia.com/). The price data for each region were averaged to serve as the ground truth for the housing-price mapping.
- Check-ins: 1M geotagged check-in records were gathered from Weibo (https://weibo.com/). Referring to (Li et al. 2024; Zhang et al. 2023b), the check-in data were aggregated within each region and used as the ground truth for region popularity mapping.

4.1.2 Parameter settings

For the pre-training stage, we instantiated an RGSCL model with tuned hyperparameters (cf. Section 3.3 for details of hyperparameter tuning). We adopted a stochastic gradient descent (SGD) optimiser with a learning rate of 2×10^{-4}, and set the training epoch to 200

without early stopping. The RGSCL was trained in minibatch mode with a batch size of 32. The gradient clipping technique and linear learning rate warm-up techniques were utilised to accelerate model convergence. For the fine-tuning stage, the data were randomly split into training, validation, and test sets at a ratio of 6:2:2 for model training and evaluation. The task head was trained in the minibatch mode with a batch size of 32 for 200 epochs. All geographic mapping experiments were repeated ten times with unfixed random seeds for reliability. For the baseline settings, we set the dimension of representation to be the same as that of RGSCL. All experiments were conducted on a server equipped with a single NVIDIA GeForce RTX 4090 GPU and 10 Xeon Platinum 8352V CPUs.

4.1.3 Baseline models

We selected several well-accepted region representation learning models as baselines for comparison with the RGSCL, each of which considers spatial effects. Unlike the joint learning conducted by RGSCL, GNN-based models (i.e. GCN and GAT) learn representations for POI, building views separately before fusion, and use a node prediction task for pretraining. All baselines employed the same architecture for the task head to ensure a consistent evaluation.

- GCN (Xu et al., 2022): This is a representative GNN architecture that performs convolution-based message passing on adjacent POIs and captures the spatial proximity between POIs. It utilises mean pooling to obtain region embeddings from node embeddings.
- GAT (Qin et al., 2025): Performs attention-based message passing to capture the spatial proximity among POIs under heterogeneous spatial interactions. It utilises mean pooling to obtain region embeddings from node embeddings.

- Doc2Vec (Niu and Silva 2021): Learns region embeddings by leveraging the spatial proximity between POIs and roads, with region embeddings obtained through synchronous training of region tokens and POI tokens.
- Place2Vec (Yan et al. 2017): This method learns POI embeddings from the spatial co-location of POI categories and subsequently aggregates POI embeddings via mean pooling to derive the region embeddings.
- RegionDCL (Li et al., 2023): This method learns region embeddings by contrasting spatially adjacent regions, thereby capturing the spatial proximity between regions.

4.1.4 Evaluation metrics

We adopted the evaluation metrics used in previous studies to assess the performance of the RGSCL in the three geographic mapping tasks. We formulated the functional zoning task as a classification problem and adopted the following evaluation metrics: Overall accuracy (OA) ↑, Cohen's kappa (Kappa) ↑, Macro-averaged F1 score (MacF1) ↑. Housing price mapping and region popularity mapping were formulated as regression tasks, and the following evaluation metrics were adopted: Mean absolute error (MAE) ↓, Root mean squared error (RMSE) ↓, Coefficient of determination (R^2) ↑. ↓ and ↑ denote the lower and higher values indicate better performance, respectively. Refer to Appendix A2 for detailed calculations of these metrics.

4.2 Performance

4.2.1 Comparison with baselines

The evaluation results ($\mu \pm \sigma$) of the geographic mapping performance of RGSCL and the baselines are listed in Table 1 and shown in Figure 6. We observed that the mapping

errors across all models did not exhibit spatial bias. Word2Vec-based models (i.e. Doc2Vec and Place2Vec) performed worse than the GNN-based models (i.e. RGSCL, GCN, and GAT). On the one hand, the input structure of Word2Vec-based models (i.e. sequential structure) is less capable of describing the complex spatial relations between geospatial entities than GNN-based models (i.e. graph structure), leading to a loss of spatial information in region representations. However, the message-passing mechanism of the GNNs appears to be more effective in capturing the spatial proximity between geospatial entities than the skip-gram used in Word2Vec-based models. Among the Word2Vec-based models, Doc2Vec performed worse than Place2Vec, indicating that the spatial proximity between POIs in Place2Vec conveyed more effective regional information than the proximity between POIs and roads in Doc2Vec. For GNN-based models, GAT outperformed GCN because the attention-based message-passing mechanism was better at learning heterogeneous spatial interactions. Compared with these models, both RGSCL and RegionDCL achieved better performance because spatial contrastive learning provided models with a stronger inductive bias. RegionDCL performed slightly better than RGSCL in the housing price task. This suggests that the spatial contrast for spatial proximity designed in RegionDCL may be more effective in modelling housing price patterns with strong spatial continuity than RGSCL, which learns spatial proximity through graph encoding. As expected, RGSCL outperformed RegionDCL in the other two tasks. A plausible explanation is that the RGSCL captures spatial heterogeneity, thereby enhancing its generality in geographic mapping tasks. This aligns with Goodchild's (2004) observation that spatial heterogeneity is a more fundamental geographic principle than the First Law of Geography. Although RGSCL, RegionDCL, and other GNN-based models capture spatial proximity, the improved performance of RGSCL reflects spatial dependency, and spatial heterogeneity knowledge

provides region representations with high-level features, boosting the task-agnostic generality of RGSCL. Therefore, we conducted a more complex experiment. The results are summarised in Table 2. For RegionDCL, the reported computational time excludes the training time for the initial building features, as the training is performed separately from the main network. RGSCL required less computational time than RegionDCL, demonstrating its efficiency advantage. Compared with other models, RGSCL sacrifices only a small and fully acceptable amount of computational time to achieve improved mapping performance, which represents a worthwhile trade-off.

Table 1. Performance of geographic mapping. The units of MAE and RMSE for housing price are CNY/m^2.

	Urban Function			Housing Price			Region Popularity		
Model	OA↑	Kappa↑	MacF1↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑
RGSCL	**0.7412±0.0075**	**0.5432±0.0092**	**0.6474±0.0085**	12604.23±135.52	16994.20±240.34	**0.433±0.014**	**213.02±9.50**	**341.63±37.46**	**0.392±0.016**
GCN	0.6992±0.0096	0.5071±0.0113	0.5988±0.0104	14337.16±142.34	18242.59±251.21	0.317±0.018	241.74±9.12	362.92±34.21	0.326±0.014
GAT	0.7054±0.0080	0.5150±0.0095	0.6042±0.0087	13933.26±138.77	18021.14±245.89	0.343±0.016	244.36±9.31	371.35±37.13	0.314±0.015
Doc2Vec	0.6602±0.0114	0.4662±0.0137	0.5679±0.0128	16638.94±146.45	20124.64±312.56	0.129±0.020	272.59±10.42	401.10±40.12	0.132±0.022
Place2Vec	0.6683±0.0114	0.4724±0.0131	0.5745±0.0124	16149.42±133.82	19564.23±234.11	0.186±0.015	260.62±9.87	394.21±38.09	0.151±0.017
RegionDCL	0.7321±0.0071	0.5232±0.0083	0.6217±0.0077	**12553.56±121.33**	**16925.22±229.78**	0.430±0.011	226.93±8.64	346.86±33.86	0.374±0.012

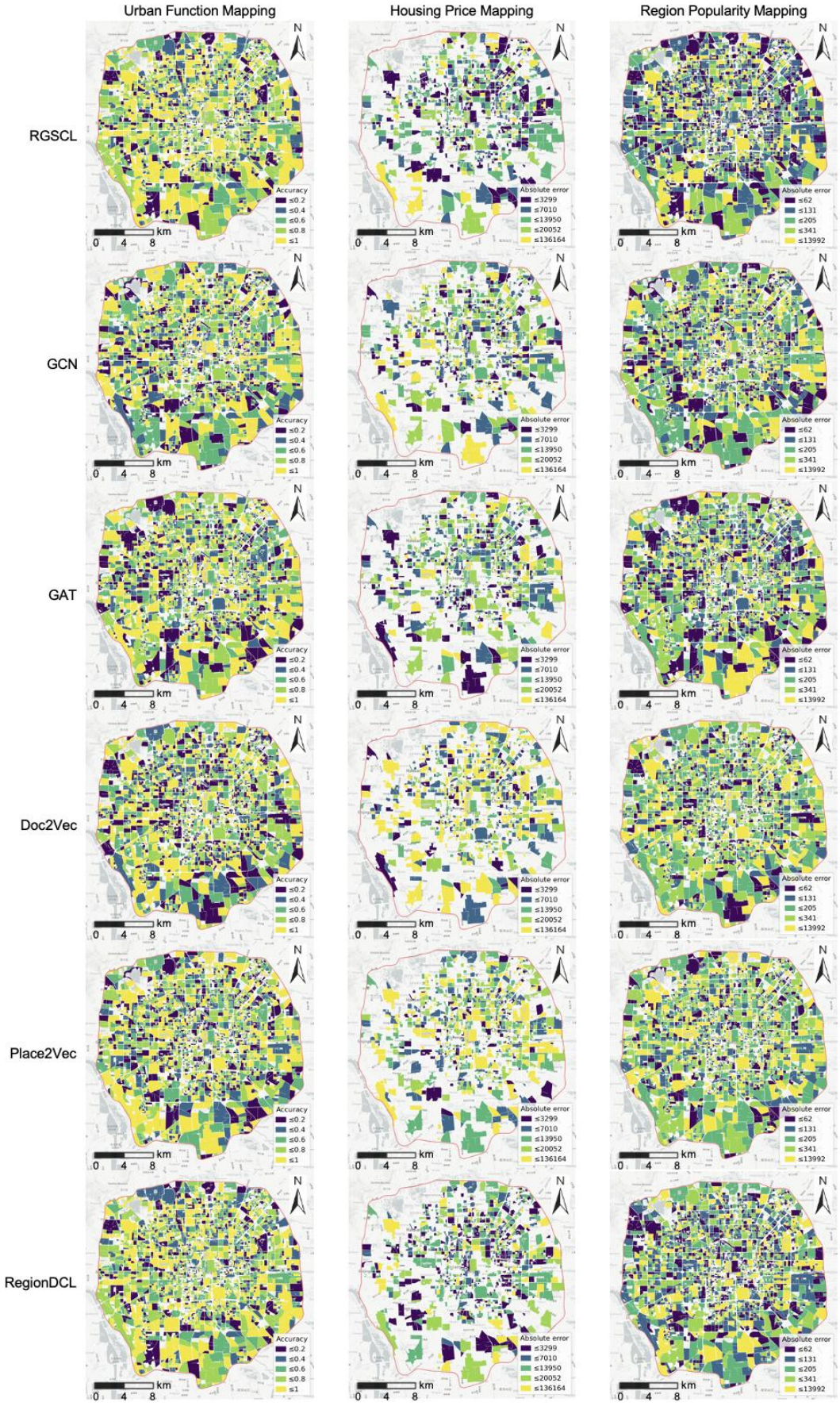

Figure 6. Spatial distribution of mapping performance.

Table 2. Computational time of model training.

Model	RGSCL	GCN	GAT	Doc2Vec	Place2Vec	RegionDCL
T_{total} (s)	5,443.76	1273.60	2045.83	750.45	560.12	9,118.74
T_{epoch} (s)	27.22	6.37	10.23	3.75	2.80	45.59

4.2.2 Representation space study

Alignment and uniformity are the two key properties of learned representations (Wang and Isola 2020). Alignment emphasises the closeness of embeddings of positive pairs, whereas uniformity emphasises the even distribution of embeddings on the unit hypersphere $\mathcal{S}^{F-1}$. Thus, we can evaluate the quality of the RGSCL representations from this perspective. To empirically verify this, we normalised the learned region representations to lie on unit hypersphere $\mathcal{S}^1$ to empirically verify this. Specifically, the more evenly the representations are distributed, the more uniform they are, and the smaller the ℓ_2 distance between the POI-view and building view region representations before feature fusion, the more aligned they are. Because Doc2Vec and Place2Vec only involve POI information and cannot evaluate alignment, they were excluded from this study.

Figure 7 presents the results for uniformity (first and second rows) and alignment (third row). The first row shows the region embedding distributions with Gaussian kernel density estimation (KDE) on $\mathcal{S}^1$. The second row shows the distributions with KDE on angles, i.e., $\arctan2(y, x)$ for each region embedding point $(x, y) \in \mathcal{S}^1$. The third row shows the ℓ_2 distances between the POI and building-view region representations. We observed that the region embeddings learned by the RGSCL were more evenly distributed on $\mathcal{S}^1$ and exhibited lower POI and building-view pair distances. These results are consistent with our expectations, demonstrating that dual spatial contrasts enhance both the uniformity and alignment of region representations, thereby improving their informativeness and discriminativeness.

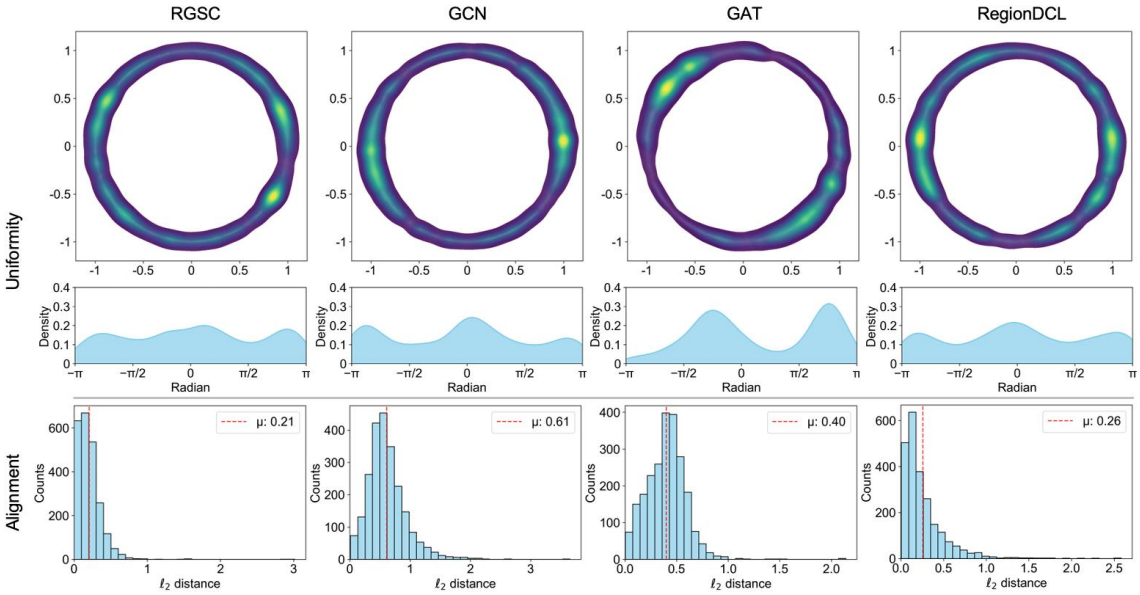

Figure 7. Region representations on $\mathcal{S}^1$.

4.3 Parameter sensitivity study

We conducted a sensitivity analysis on three key hyperparameters of RGSCL, i.e., the number of attention heads in the encoder (h), the dimension of the region embeddings (d), and the loss function weight factor (λ). h describes the multiple spatial interaction relations between geospatial entities, d refers to the dimension used to represent regions, and λ balances the contributions of inter- and intra-region spatial effects to the model. We tune h in {1, 2, 4, 8}, d in {16, 32, 64, 128, 256}, and λ in {0.1, 0.3, 0.5, 0.7, 0.9}. The results are shown in Fig. 8.

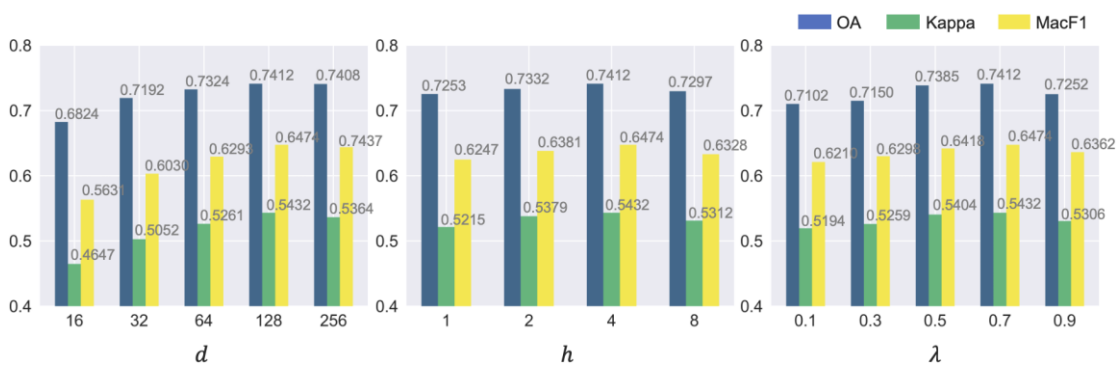

Figure 8. Parameter effect of RGSCL on urban function mapping.

We observed that the marginal performance gains of d decreased as the dimensions increased, peaking at 128. This result is intuitive because a large dimension of the representation space leads to sparse information, whereas a small dimension fails to describe complete geographic information. Regarding h, the optimal performance was achieved when $h = 4$. This may be because too few attention heads struggle to express the actual heterogeneous spatial interactions among geospatial entities, whereas too many heads may capture similar or redundant spatial interactions. Regarding λ, values between 0.5 and 0.7 yield better performance, likely because the gradient magnitudes of the losses for the two spatial effects are similar, reducing the need for subtle adjustments to prevent gradient domination issue. Finally, we find that the best performance is achieved with the hyperparameter combination of $\{h = 4, d = 128, \lambda = 0.7\}$. The experimental results indicate that RGSCL is not sensitive to the tested hyperparameters.

4.4 Ablation study

Ablation experiments were conducted on each key component of the RGSCL to validate its effectiveness. The following RGSCL variants were used:

(a) RGSCL w/o intra: It drops intra-region spatial contrast, i.e., remove $\mathcal{L}_{\text{intra}}$ from $\mathcal{L}_{\phi}$.

(b) RGSCL w/o inter: It drops inter-region spatial contrast, i.e., remove $\mathcal{L}_{\text{inter}}$ from $\mathcal{L}_{\phi}$.

(c) RGSCL w/o reg: It drops ℓ_2 regularization, i.e., remove ℓ_2 from $\mathcal{L}_{\phi}$.

(d) RGSCL-SL: RGSCL is converted from SSL into an end-to-end supervised learning model. The parameters of the RGSCL-SL were optimised from scratch using task-specific supervised signals, demonstrating the upper bound of performance for geographic mapping tasks.

(e) RGSCL-GCN replaces the GAT of the encoder with a single-layer GCN.

(f) RGSCL-mean: This replaces the dummy node with mean pooling for graph readout, which pools the building and POI embeddings within a region to derive region embeddings.

Table 3. Ablation results of RGSCL.

	Urban Function			Housing Price			Region Popularity		
Model variant	OA↑	Kappa↑	MacF1↑	MAE↓	MSE↓	R^2↑	MAE↓	MSE↓	R^2↑
RGSCL	0.7412±0.0075	0.5432±0.0092	0.6474±0.0085	12604.23±135.52	16994.20±240.34	0.433±0.014	213.02±9.50	341.63±37.46	0.392±0.016
RGSCL w/o intra	0.7225±0.0104	0.5196±0.0121	0.6219±0.0109	13494.75±157.63	17632.34±288.41	0.414±0.017	229.61±9.03	354.85±34.21	0.357±0.014
RGSCL w/o inter	0.7343±0.0065	0.5372±0.0070	0.6390±0.0072	13193.62±169.37	17328.92±305.34	0.422±0.019	230.57±10.51	356.29±38.89	0.354±0.020
RGSCL w/o reg	0.7391±0.0050	0.5403±0.0060	0.6418±0.0062	12943.64±186.85	17135.69±331.47	0.428±0.023	216.53±12.23	345.32±41.77	0.388±0.022
RGSCL-SL	**0.7654±0.0120**	**0.5573±0.0145**	**0.6539±0.0130**	**12284.49±146.22**	**16427.05±278.89**	**0.485±0.015**	**202.57±8.84**	**329.82±31.12**	**0.436±0.013**
RGSCL-GCN	0.7389±0.0053	0.5359±0.0061	0.6436±0.0058	13328.65±152.74	17621.37±279.22	0.406±0.017	220.93±10.02	346.56±38.54	0.375±0.018
RGSCL-mean	0.7252±0.0088	0.5220±0.0106	0.6277±0.0095	13582.24±121.04	17738.26±263.66	0.391±0.009	223.46±9.23	349.83±37.07	0.371±0.016

The ablation results are listed in Table 3 and the following important findings can be observed:

- Impact of spatial contrast (exp. a–d)

All ablations with respect to the loss function showed inferior performance compared with RGSCL. RGSCL without intra leads to region representations that lack knowledge of POI–building spatial dependency, whereas RGSCL without inter results in region representations that lack knowledge of spatial heterogeneity. These results indicate that each loss function is critical for model optimisation. When all losses are discarded and converted to end-to-end RGSCL-SL, the performance improves slightly compared with RGSCL. This suggests that RGSCL is very close to the upper performance bound for each task.

- Impact of graph encoder (exp. e)

It is evident that replacing the base encoder GAT with GCN (RGSCL-GCN) leads to a decline in the model performance, which is somewhat consistent with the experimental results in Tab. 1 that GAT outperforms GCN in this situation. Moreover, RGSCL-GCN surpasses GCN, further indicating the effectiveness of our dual spatial contrasts, which does not rely on a specific base graph encoder, as supported by the performance results in Table. 1 (RGSCL > GAT).

- Impact of region graph readout (exp. f)

RGSCL-mean is inferior to RGSCL, indicating that the graph readout mechanism employing dummy nodes in RGSCL is more effective in deriving region representations from node information than simple mean pooling.

5 Discussion

RGSCL has some limitations that need to be pointed out. RGSCL is as difficult as any other region representation learning method in avoiding the modifiable areal unit problem in GIS. This study tested the RGSCL at the regional scale of the EULUC, but whether different scales of region partitioning affect the RGSCL requires further investigation. In addition, both the RGSCL and the current models are static region representation learning models that ignore temporal nuances in regional information, making it difficult to capture the temporal or event dynamics of regions. Because region representations only entail information from POIs and building footprints, it may be difficult to make them applicable to all mapping tasks. Although RGSCL improves the performance of geographic mapping, interpreting the learned representations remains challenging. Opening this black box and enhancing its explainability will be an important focus for future work. In this regard, representation disentanglement holds great promise, and a plausible solution is to link the learned representations to urban design concepts such as those from *The Image of the City*, which facilitates a deeper understanding of the embeddings.

However, there is significant potential for RGSCL. First, RGSCL can easily acquire more flexible and advanced variants. One of the flexibilities of RGSCL lies in its construction of multi-view region graphs, which can be easily adapted by replacing purely geometric DT with alternative approaches such as street graphs or by incorporating geographic barriers and accessibility constraints. The spatial dependency between POI semantics and building morphology in RGSCL could benefit from adopting a more appropriate spatial scale rather than a regional scale in the intra-region spatial contrast, such as morphocompatible parcels (Fleischmann et al., 2020; Hamaina et al., 2012), which captures finer-grained spatial dependencies and yields more precise features.

RGSCL does not depend on any specific graph encoder, allowing it to be replaced with encoders, such as graph transformers, to capture more global spatial dependencies between geospatial entities.

RGSCL is an exploratory integration of spatial effects that does not fully consider other geographic or physical laws, which present significant potential space. Because spatial effects serve as an effective model inductive bias, the geospatial knowledge learned by the model can extend beyond the scope of region representations and inform broader GeoAI theories and applications.

6 Conclusions

In this study, we developed a novel region representation learning framework called RGSCL. RGSCL starts with the semantic and morphological features of POIs and building footprints, introducing a region embedding network that jointly encodes multiview regional information to derive region representations with spatial proximity. Dual spatial contrastive learning then synergises spatial dependence and spatial heterogeneity to further optimise representations and convey more geographically generic meanings. The experimental results consistently demonstrated the superiority of RGSCL, revealing the potential of spatial effects in region representation learning. This study offers new insights into region representation learning from the perspective of inductive bias with respect to spatial effects and we hope this work will encourage the GeoAI community to rethink the current training paradigms employed in region representation learning. The proposed framework is expected to be generalised to different urban environments, and in future work, we will develop a unified spatial contrast framework by incorporating additional spatial effects into the RGSCL.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Acknowledgement

The authors thank the editors and the anonymous reviewers for their insightful comments and constructive suggestions.

Funding

This work was supported in part by the Fundamental Research Funds for the Central Universities, China (Grant No. 2042022dx0001).

Notes on contributors

Quan Qin is a PhD candidate at Wuhan University. His research interests include spatial data mining and spatial representation learning.

Tinghua Ai is a Professor at Wuhan University. His research interests include multiscale representation of spatial data and cognition.

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

Shishuo Xu is an Associate Professor at Beijing University of Civil Engineering and Architecture. Her research interests include spatial data mining and spatio-temporal event detection.

Mingyi Du is a Professor at Beijing University of Civil Engineering and Architecture. His research interests include smart cities and urban remote sensing.

Songnian Li is a Professor at Toronto Metropolitan University. His research interests include geospatial data mining and knowledge discovery.

Data and codes availability statement

The codes and the data of this study are available at https://doi.org/10.6084/m9.figshare.28283762.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1409.0473

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2024). Explaining neural scaling laws. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(27), 1–8. https://doi.org/10.1073/pnas.2311878121

Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., & Liu, B. (2023). Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. *ISPRS Journal of Photogrammetry and Remote Sensing*, *201*, 193–208. https://doi.org/10.1016/j.isprsjprs.2023.05.006

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks* (pp. 1–40). http://arxiv.org/abs/1806.01261

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*, 1575–1585.

Chen, Y., Huang, W., Zhao, K., Jiang, Y., & Cong, G. (2025). Self-supervised representation learning for geospatial objects: A Survey. *Information Fusion*, *123*, 103265. https://doi.org/10.1016/j.inffus.2025.103265

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C. M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H. T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235. https://doi.org/10.1038/s42256-023-00626-4

Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2020). Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Computers, Environment and Urban Systems*, *80*, 101441. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2019.101441

Gao, S., Hu, Y., & Li, W. (2023). *Handbook of Geospatial Artificial Intelligence*.

Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., Chen, J., Chen, X., Fang, L., Feng, S., Feng, Y., Gong, Y., Gu, H., Huang, H., Huang, X., Jiao, H., Kang, Y., Lei, G., Li, A., … Xu, B. (2020). Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018. *Science Bulletin*, *65*(3), 182–187. https://doi.org/10.1016/j.scib.2019.12.007

Goodchild, M. F. (2004). The Validity and Usefulness of Laws in Geographic Information Science and Geography. *Annals of the Association of American Geographers*, *94*(2), 300–303. https://doi.org/10.1111/j.1467-8306.2004.09402008.x

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 1–49. https://doi.org/10.1098/rspa.2021.0068

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9052–9071. https://doi.org/10.1109/TPAMI.2024.3415112

Hamaina, R., Leduc, T., & Moreau, G. (2012). *Towards Urban Fabrics Characterization Based on Buildings Footprints BT - Bridging the Geographic Information Sciences: International AGILE'2012 Conference, Avignon (France), April, 24-27, 2012* (J. Gensel, D. Josselin, & D. Vandenbroucke (eds.); pp. 327–346). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29063-3_18

Hassani, K., & Khasahmadi, A. H. (2020). Contrastive multi-view representation learning on graphs. *International Conference on Machine Learning*, 4074–4084.

Huang, F., Lv, J., Li, G., & Yue, Y. (2025). SPOK : tokenizing geographic space for enhanced spatial reasoning in GeoAI SPOK : tokenizing geographic space for enhanced spatial. *International Journal of Geographical Information Science*, 1–41. https://doi.org/10.1080/13658816.2025.2497810

Huang, W., Wang, J., & Cong, G. (2024). Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science*, 1–29. https://doi.org/10.1080/13658816.2024.2347322

Huang, W., Zhang, D., Mai, G., Guo, X., & Cui, L. (2023). Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, *196*, 134–145. https://doi.org/10.1016/j.isprsjprs.2022.11.021

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, *1*, 448–456.

Ishmael, M., Aristide, B., Sai, B., Sherjil, R., & Yoshua, O. (2018). Mutual information neural estimation. *International Conference on Machine Learning*, 531–540.

Kong, B., Ai, T., Zou, X., Yan, X., & Yang, M. (2024). A graph-based neural network approach to integrate multi-source data for urban building function classification. *Computers, Environment and Urban Systems*, *110*, 102094. https://doi.org/10.1016/j.compenvurbsys.2024.102094

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, *4*, 2931–2939.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Li, Y., Huang, W., Cong, G., Wang, H., & Wang, Z. (2023). Urban Region Representation Learning with OpenStreetMap Building Footprints. *The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1363–1373. https://doi.org/10.1145/3580305.3599538

Li, Z., Huang, W., Zhao, K., Yang, M., Gong, Y., & Chen, M. (2024a). Urban Region Embedding via Multi-View Contrastive Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(8), 8724–8732. https://doi.org/10.1609/aaai.v38i8.28718

Li, Z., Xia, L., Tang, J., Xu, Y., Shi, L., Xia, L., Yin, D., & Huang, C. (2024b). UrbanGPT: Spatio-Temporal Large Language Models. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 5351–5362. https://doi.org/10.1145/3637528.3671578

Liu, J., Qin, Q., Dong, G., Wang, X., Feng, J., Zeng, Z., & Cheng, T. (2025). Beyond AlphaEarth: Toward Human-Centered Spatial Representation via POI-Guided Contrastive Learning. https://doi.org/10.48550/arXiv.2510.09894

Liu, X., Cheng, J., Song, Y., & Jiang, X. (2022). Boosting Graph Structure Learning with Dummy Nodes. *International Conference on Machine Learning*, *162*, 13704–13716.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2023a). Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 857–876. https://doi.org/10.1109/TKDE.2021.3090866

Liu, Y., Wang, K., Xing, X., Guo, H., Zhang, W., Luo, Q., Gao, S., Huang, Z., Li, H., Li, X., Wang, J., Wang, J., & Zhu, D. (2023b). On spatial effects in geographical analysis. *Dili Xuebao/Acta Geographica Sinica*, *78*(3), 517–531. https://doi.org/10.11821/dlxb202303001

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2020). Multi-scale representation learning for spatial feature distributions using grid cells. *International Conference on Learning Representations*.

Niu, H., & Silva, E. A. (2021). Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. *Computers, Environment and Urban Systems*, *88*, 101651. https://doi.org/10.1016/j.compenvurbsys.2021.101651

OpenAI. (2023). *GPT-4 Technical Report*.

Qin, Q., Ai, T., Xu, S., Zhang, Y., Huang, W., Du, M., & Li, S. (2025). Learning dual context aware POI representations for geographic mapping. *International Journal of Applied Earth Observation and Geoinformation*, *142*, 104683. https://doi.org/10.1016/j.jag.2025.104683

Qin, Q., Xu, S., Du, M., & Li, S. (2022a). Identifying urban functional zones by capturing multi-spatial distribution patterns of points of interest. *International Journal of Digital Earth*, *15*(1), 2468–2494. https://doi.org/10.1080/17538947.2022.2160841

Qin, Q., Xu, S., Du, M., & Li, S. (2022b). URBAN FUNCTIONAL ZONE IDENTIFICATION BY CONSIDERING THE HETEROGENEOUS DISTRIBUTION OF POINTS OF INTERESTS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*(4), 83–90. https://doi.org/10.5194/isprs-annals-V-4-2022-83-2022

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*, 8748–8763.

Shen, X., Sun, D., Pan, S., Zhou, X., & Yang, L. T. (2023). Neighbor Contrastive Learning on Learnable Graph Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 9782–9791. https://doi.org/10.1609/aaai.v37i8.26168

Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations*. https://doi.org/10.1007/978-3-031-01587-8_7

Wang, T., & Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning*, 9871–9881.

Wang, X., Cheng, T., Law, S., Zeng, Z., Yin, L., & Liu, J. (2024). Multimodal Contrastive Learning of Urban Space Representations from POI Data. *Computers, Environment and Urban Systems*, *120*(April), 102299. https://doi.org/10.1016/j.compenvurbsys.2025.102299

Xi, Y., Li, T., Wang, H., Li, Y., Tarkoma, S., & Hui, P. (2022). Beyond the First Law of Geography: Learning Representations of Satellite Imagery by Leveraging Point-of-Interests. *Proceedings of the ACM Web Conference 2022*, 3308–3316. https://doi.org/10.1145/3485447.3512149

Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., & He, N. (2022). A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. *Computers, Environment and Urban Systems*, *95*, 101807. https://doi.org/10.1016/j.compenvurbsys.2022.101807

Xue, Y., Gan, E., Ni, J., Joshi, S., & Mirzasoleiman, B. (2024). Investigating the Benefits of Projection Head for Representation Learning. *International Conference on Learning Representations*, 1–23. http://arxiv.org/abs/2403.11391

Yan, B., Mai, G., Janowicz, K., & Gao, S. (2017). From ITDL to Place2Vec--Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–10. https://doi.org/10.1145/3139958.3140054

Yan, X., Ai, T., Yang, M., & Yin, H. (2019). A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *150*, 259–273. https://doi.org/10.1016/j.isprsjprs.2019.02.010

Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., & Liang, Y. (2024). UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web. *Proceedings of the ACM Web Conference*, 4006–4017. https://doi.org/10.1145/3589334.3645378

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, *31*(4), 825–848. https://doi.org/10.1080/13658816.2016.1244608

Zhang, D., Wang, P., Ding, L., Wang, X., & He, J. (2024a). Spatio-Temporal Contrastive Learning-Based Adaptive Graph Augmentation for Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*, *26*(1), 1304–1318. https://doi.org/10.1109/TITS.2024.3487982

Zhang, D., Xu, R., Huang, W., Zhao, K., & Chen, M. (2023a). Towards an Integrated View of Semantic Annotation for POIs with Spatial and Textual Information. *International Joint Conference on Artificial Intelligence*, 2441–2449. https://doi.org/10.24963/ijcai.2023/271

Zhang, L., Long, C., & Cong, G. (2023b). Region Embedding with Intra and Inter-View Contrastive Learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 9031–9036. https://doi.org/10.1109/TKDE.2022.3220874

Zhang, P., Yang, M., Wang, Y., Yang, T., Yu, H., & Yan, X. (2024b). Integrating metro passenger flow data to improve the classification of urban functional regions using a heterogeneous graph neural network. *International Journal of Digital Earth*, *17*(1), 2443468. https://doi.org/10.1080/17538947.2024.2443468

Zhang, Q., Huang, C., Xia, L., Wang, Z., Yiu, S., & Han, R. (2023c). Spatial-Temporal Graph Learning with Adversarial Contrastive Adaptation. *International Conference on Machine Learning*, *202*, 41151–41163.

Zhang, Y., Chen, Z., Zheng, X., Chen, N., & Wang, Y. (2021). Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. *Journal of Hydrology*, *603*, 127053. https://doi.org/10.1016/j.jhydrol.2021.127053

Zhang, Y., Huang, W., Yao, Y., Gao, S., Cui, L., & Yan, Z. (2024c). Urban region representation learning with human trajectories: a multi-view approach incorporating transition, spatial, and temporal perspectives. *GIScience & Remote Sensing*, *61*(1). https://doi.org/10.1080/15481603.2024.2387392

Zhou, H., Huang, W., Chen, Y., He, T., Cong, G., & Ong, Y.-S. (2024). Road Network Representation Learning with the Third Law of Geography. *Advances in Neural Information Processing Systems*, 11789–11813.

Appendix

Appendix A. Description of indices and metrics involved

Appendix A1. Building morphology indices

Table A1. Description of building morphology index

Index	Notation/Equation	Description
Area	A_{b}	It denotes building area.
Perimeter	P_{b}	It denotes building perimeter.
Radius	$\frac{1}{\|\mathcal{N}\|}\sum_{i\in\mathcal{N}} R_i$	It measures average distance from each vertex of the building to its centroid.
Orientation	N/A	Orientation of the smallest bounding rectangle (SBR).
Compactness	$4\pi A_{\mathrm{b}}/P_{\mathrm{b}}^2$	It measures quadratic relation between the area and the perimeter.
Fractality	$1-\log(A_{\mathrm{b}})/2\log(P_{\mathrm{b}})$	It measures logarithmic relation between the area and the perimeter.
Concavity	$A_{\mathrm{b}}/A_{\mathrm{ch}}$	It measures area ratio of the building to its convex hull (CH).
Elongation	$L_{\mathrm{sbr}}/W_{\mathrm{sbr}}$	It measures length-width ratio of the building SBR.

Appendix A2. Evaluation metrics for geographic mapping

Table A2. Description of metrics for geographic mapping tasks.

Mapping task	Metrics	Equation	Description
Urban Function	OA	$\frac{TP+TN}{TP+TN+FP+FN}$	It measures the proportion of correct predictions relative to ground truth, where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.
	Kappa	$\frac{OA-P_e}{1-P_e}$	It measures the agreement between predictions and ground truth, where $P_e=\frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+TN+FP+FN)^2}$.
	MacF1	$\frac{1}{k}\sum_{i=1}^{k} F_1^i$	It evaluates the performance of the model by treating all classes equally, where $F_1^i=\frac{2\cdot Precision\cdot Rrecall}{Precision+Recall}$, $recision=\frac{TP}{TP+FP}$, $Recall=\frac{TP}{TP+FN}$, k is the number of function categories.
Region Popularity Housing Price	MAE	$\frac{1}{n}\sum_{i=1}^{n}\|y_i-\hat{y}_i\|$	It measures the average absolute difference between predictions and ground truth, where n is the size of testing set, y_i and $\hat{y}_i$ are prediction and ground truth of sample i.
	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	It measures the standard deviation of prediction errors.
	R^2	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$	It evaluates goodness-of-fit of the model, $\bar{y}$ is the average of the ground truth.

Appendix B. DT-based spatial proximity

Appendix B1. Natural representation of spatial proximity

DT constructs a planar network to define the structure of a region graph by connecting a set of nodes to non-overlapping triangles, thereby transforming the spatial proximity among nodes into graph adjacency. The DT-based graph offers several desirable properties for preserving spatial proximity: (1) it possesses the geometric property of maximising the minimum angle, which ensures global optimality in connecting spatially proximate nodes and effectively approximates the natural spatial proximity; (2) it provides isotropic connectivity that does not favour any specific direction (e.g. radial or

grid-like), thereby preserving the natural spatial anisotropy of geographic entities; and (3) it is a parameter-free deterministic algorithm that ensures the uniqueness and stability of the preserved spatial Overall, the DT-based graph provides a stable and natural representation of spatial proximity among geographic entities.

Appendix B2. Complexity analysis

DT construction has an average-case time complexity of $\mathcal{O}(n \cdot \log n)$ for uniform distribution, with worst-case $\mathcal{O}(n^2)$ for adversarial distributions. This complexity ensured large-scale scalability. Below, we analyse both the theoretical and empirical complexities.

We randomly sampled subsets from the POIs and building footprints to test the complexity of the DT construction. We repeated the experiments ten times and reported the mean results in Table. B1. It can be observed that the empirical time is generally consistent with the theoretical time and falls well within an acceptable range, indicating that DT enables the applicability of RGSCL framework to larger datasets.

Table B1. The empirical time T_{observe} and theoretical time T_{theory} for constructing DT-based spatial context.

DT Type	Number of Vertex	T_{observe} (s)	T_{theory}* (s)	$T_{\text{observe}}/T_{\text{theory}}$ Ratio
POI-DT	1,000 (n_{base})	0.005 (T_{base})	0.005	1.00x
	5,000	0.029	0.031	0.94x
	10,000	0.061	0.067	0.91x
	50,000	0.337	0.392	0.86x
	100,000	0.817	0.833	0.98x
	200,000	1.821	1.767	1.03x
Building-DT	1,000 (n_{base})	0.004 (T_{base})	0.004	1.00x
	5,000	0.023	0.025	0.92x
	10,000	0.053	0.053	1.00x
	50,000	0.339	0.313	1.08x
	100,000	0.757	0.667	1.13x
	200,000	1.778	1.414	1.26x

*The theoretical time of $\mathcal{O}(n \cdot \log n)$ is defined as $T_{\text{theory}} = T_{\text{base}} \cdot \frac{n \cdot \log n}{n_{\text{base}} \cdot \log n_{\text{base}}}$.