Author copy of:

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.

DOI: https://doi.org/10.1016/j.isprsjprs.2022.11.021

Learning urban region representations with POIs and hierarchical graph infomax

We present the hierarchical graph infomax (HGI) approach for learning urban region representations (vector embeddings) with points-of-interest (POIs) in a fully unsupervised manner, which can be used in various downstream tasks. Specifically, HGI comprises several key steps: (1) training category embeddings as the initial features of POIs; (2) interconnecting POIs with a graph structure and performing graph convolution to capture the uniqueness of each POI based on its spatial context; (3) aggregating POIs to the regional level using multi-head attention mechanisms, to consider the multi-faceted influence from POIs to regions; (4) performing graph convolution at the regional level to generate region representations, to incorporate the similarities between adjacent regions; (5) aggregating region representations to produce an embedding at the city level. The model is finally trained through maximizing the mutual information among the POI – region – city hierarchy, which facilitates the information from local (POIs) and global (city) scales flowing to the learned region representations, making them both locally and globally relevant. We perform extensive experiments on three downstream tasks, i.e., estimating urban functional distributions, population density, and housing price, in the study areas of Xiamen Island and Shenzhen, China. The results indicate that HGI considerably outperforms several competitive baselines in all three tasks, which proves that HGI could produce meaningful and effective region representations. In addition, the learned region representations based on POIs can potentially be used for reinforcing data representations from other modalities, e.g., remote sensing data. The implementation of HGI can be found at https://github.com/RightBank/HGI.

1 Introduction

Urban regions have been an ideal analytical scale for various studies in spatial information science, e.g., region function recognition (Zhang et al., 2017, 2018; Cao et al., 2020), population estimation (Cheng et al., 2022), and urban vitality assessment (Tu et al., 2020). Specifically, urban regions are spatially distributed neighborhoods in our cities with relatively homogenous physical and socioeconomic characteristics within each

region (Yuan et al., 2012). Today most region-level analyses have been driven by abundant urban sensory data from various sources, such as remote sensing data, human mobility data, POIs, and street view imagery. Despite the remarkable strides that have been made, most of these analyses rely on task-specific supervised learning, which leads to two major shortcomings. First, models and data representations learned for one task are not necessarily useful in other tasks. Second, in many analyses, ground truth data are sparse or even unavailable, which makes task-specific supervised learning often impractical (Wang et al., 2020).

In this context, the prominent trend of unsupervised representation learning in the machine learning community is strikingly inspirational. In fact, the success of data-driven analyses using machine learning algorithms largely depends on the choice of data representation; much of the actual effort is devoted into transforming raw data into meaningful representations (Bengio et al., 2013). Therefore, learning meaningful representations in an unsupervised fashion has been persistent endeavors from numerous communities in information sciences (Grover and Leskovec, 2016; Chen et al., 2020). For example, in natural language processing, seminal models such as Word2Vec (Mikolov et al., 2013), BERT (Devlin et al., 2019), and GPT (Radford et al., 2018) are developed to generate, among others, word representations (embeddings), which can be used for many downstream tasks, e.g., named entity recognition, text generation, and sentiment analysis. In this vein, we can naturally presume that if we manage to learn meaningful representations for urban regions to identify and disentangle the underlying explanatory factors hidden in the raw sensory data, various analyses could be fundamentally benefited (Mai et al., 2020).

In this paper, we focus on learning meaningful region representations using POIs, i.e., transforming each urban region into a low-dimensional vector embedding.

The rationale of using POIs is that they have intrinsic connections with human behaviors and the socioeconomic factors of cities (Gao et al., 2017), and generally can be readily obtained. POIs have been proven to be a competent proxy for, among others, urban function and vitality (Andrade et al., 2020; Huang et al., 2022). Although some studies have utilized multi-modal geospatial data (e.g., Zhang et al., 2017; Wang et al., 2020), we believe that mining POIs in depth, and learning meaningful representations only from such data remain a valuable scientific question, especially in view of their widespread utilization. In addition, learning meaningful POI-based region representations could also benefit the scenarios of using POIs with other types of data, e.g., remote sensing data, as a more effective solution than traditional strategies, e.g., using category frequencies or topic models (Su et al., 2021; Zhang et al., 2017).

Learning meaningful region representations from POIs is generally challenging from two perspectives: (1) how to capture and represent the rich information carried by POIs with regard to the spatial distribution of socioeconomic factors and human perceptions in our cities; (2) how to aggregate the learned POI embeddings to generate region embeddings in a meaningful manner. Previous studies mainly concentrated on encoding spatial co-occurrence patterns between POI categories, and often applied deterministic average pooling to generate region representations (Yao et al., 2017a; Yan et al., 2017; Zhai et al., 2019). Such methods could lead to the loss of some valuable information, e.g., the uniqueness of each individual POI, and the interactions between adjacent regions.

In this paper, we tackle the above challenges through interconnecting POIs into a graph structure, then each POI becomes a node, an individual region is a sub-graph, and the city constitutes an entire graph. Along this vein, we propose the HGI approach for learning region representations. This approach builds upon the recent developments

of unsupervised learning using the *infomax* principle in graph-structured data (Hjelm et al., 2019; Veličković et al., 2019), and could encode a rich set of information into region representations, including POI category semantics, POI neighborhood information, multi-faceted influence from POIs to regions, and inter-regional interactions, discrimination, as well as commonalities. HGI sophisticatedly models the hierarchical POI – region – city interactions, and generates meaningful region representations in a fully unsupervised manner.

We apply HGI for learning region representations in two study areas: Xiamen Island and Shenzhen, China, and utilize the generated region representations in three downstream tasks: estimating urban functional distributions, population density, and housing price. The comparisons with several competitive baseline models indicate that HGI could produce more meaningful and effective region representations, and thus yield considerably better results in all the downstream tasks.

Following this introduction, we review related works in Section 2. In Section 3, we elaborate the details and intuitions of the proposed HGI approach. In Section 4, we provide the results of our experiments in three downstream tasks, i.e., estimating urban functional distributions, population density, and housing price. The paper ends with a discussion in Section 5, and conclusions in Section 6. The implementation of HGI can be found at https://github.com/RightBank/HGI.

2 Related works

2.1 Learning region representations with POIs

The idea of learning POI and region representations has emerged in the last years, in order to unravel the limitations of feature engineering methods. Yao et al. (2017a) first borrowed the idea of Word2Vec to learn POI category embeddings, which mainly

captures the co-occurrence patterns between POI categories in POI strings. Yan et al. (2017) proposed the Place2Vec model to capture category co-occurrence patterns using a K-nearest-neighbors (KNN) method, and Zhai et al. (2019) subsequently utilized Place2Vec for identifying function regions. Niu & Silva (2021) applied the Doc2Vec model (a variant of Word2Vec) to embed both POIs and regions for functional land use classification. Huang et al. (2022) proposed a semantics preserved POI embedding approach that incorporates the spatial co-occurrence patterns through random walks in a POI network, and categorical semantics using a manifold learning algorithm. These works have an inherent limitation, that is, they learned representations for POI categories but not individual POIs. Imagining we have two restaurants, one of them is besides a university library and the other is nearby a railway station, in principle their semantics should be subtly differentiated in view of the largely different spatial contexts, whereas previous works generate the same category embedding for them and lose the uniqueness of each POI. A recent study in Xu et al. (2022b) modeled the uniqueness of individual POIs to generate region representations for land use classification. They considered the spatial context of each POI with a graph convolutional network (GCN; Kipf and Welling, 2017), and regarded land use classification as a supervised graph classification problem. Our study follows a similar idea in considering spatial context information of POIs, but generates region representations in a fully unsupervised manner, which can be used in a wide range of downstream tasks.

Another key component in learning region representations is the aggregation of POI embeddings to generate region embeddings. In this regard, most previous works utilize an average pooling function due to the lack of effective supervision for training a more expressive aggregation function (e.g., Yao et al., 2017a; Zhai et al., 2019); this

aggregation function omits the different importance levels of the POIs in defining a region. Niu and Silva (2021) leveraged the Doc2Vec model to obtain region embeddings as a by-product of POI embeddings. Huang et al. (2022) explicitly considered the different importance levels of POIs in the aggregation process, and employed an aggregation function coupling long short-term memory (LSTM) and attention mechanisms. This method yields better performance in identifying functional regions, but still has two limitations. (1) It requires to trained in a supervised manner, thereby cannot be used without ground truth data, e.g., in the scenario of discovering urban functional structures through clustering region representations. (2) This function can only consider one perspective of a POI's influence to its region, but in reality, the influence should be depicted from multiple perspectives. For example, assuming there is a region containing a lake and a building, from the perspective of area occupation, the lake is generally more definitive, while in terms of socioeconomic and demographic factors, the building could be more important.

Furthermore, previous studies generally neglect adjacency information at the regional level, and the generated region embeddings entirely depend on the POIs resided in the regions. Intuitively, the information propagated from adjacent regions could be informative in defining each region, as they usually have substantial similarity. Considering two distinct regions both having a company and a restaurant, one of them is surrounded by commercial regions with many office buildings, malls, etc., and the other is surrounded by several industrial areas with factories. The two regions can hardly be differentiated unless their contextual information is incorporated, by which one can naturally come to a plausible guess that the former is likely to be a commercial region, while the latter can probably be an industrial region.

There are also several studies utilizing POIs in conjunction with data from other modalities to generate region embeddings. For example, Fu et al. (2019) and Du et al. (2019a) used POIs and human mobility data for learning region representations; Wang et al. (2020) proposed the Urban2Vec model to generate region embeddings with POIs and street view imagery. Generally, such studies focused on multi-modal data fusion, instead of deeply mining the information from static POIs.

2.2 Graph representation learning for spatial data

Generalizing neural networks and learning representations for graph-structured inputs are one of the major focuses of machine learning (Bronstein et al., 2017). Several prominent strides have been made in this direction, such as the developments of GCN and graph attention networks (Veličković et al., 2018), which have fueled numerous studies on spatial data. For example, Yan et al. (2019) proposed to use a GCN for classifying building patterns based on building footprints. Xu et al. (2022a) utilized a GCN for classifying urban scenes incorporating both visual and semantic features. These studies generally follow the paradigm of supervised learning on graphs for specific tasks, and thus can hardly fit our goal, i.e., learning region representations using POIs that can benefit various downstream tasks.

In addition, graph neural networks have been widely applied to spatiotemporal data for traffic and human movement analyses. For instance, Diao et al. (2019), Zhang et al. (2021), and Dai et al. (2021) proposed different variants of graph neural networks for traffic forecasting; Luo et al. (2021) and Wang et al. (2022) utilized different architectures of GCNs for next location (POI) recommendation. Such studies designed various sophisticated graph learning models to capture both spatial and temporal patterns exhibited in their datasets (e.g., POI check-in records), in which the message passing incurred in the temporal domain is generally more pivotal than in the spatial

domain, and their training objectives depend largely on the supervision of real-world traffic flow or human movement records. As to our study, we target a different scenario with only static POIs, so the movement supervision signals are not applicable.

There have been some unsupervised graph representation learning methods relying on random walks or the reconstruction of adjacency information (e.g., Grover and Leskovec, 2016; Kipf and Welling, 2016). However, with the introduction of graph convolutions, it is unclear whether such objectives actually provide useful signals, as the graph convolutional encoders already enforce neighboring nodes to be similar (Veličković et al., 2019). Furthermore, simply applying graph neural networks to POIs still cannot fit our design principles, e.g., inter-regional interactions should be considered. The recent development of the unsupervised *deep graph infomax* (DGI; Veličković et al., 2019) sheds light on this problem, which is built upon the idea of contrastive learning between a high-level "global" representation and "local" parts of the inputs. Specifically, DGI maximizes the mutual information between node representations and the aggregated graph representation, which encourages node embeddings to embody the information from a higher level of scale (globally relevant). In this paper, we extend DGI to model the interactions between three scales, i.e., POI, region, and city, and develop the HGI model to explore hierarchical mutual information between these three analytical scales. In this way, the produced region representations (middle-level) can carry the information propagated from both (local-level) POIs, and the (global-level) city, which can thus fulfil our design principles.

3 Methodology

3.1 Overview of HGI

The overarching architecture of the proposed HGI approach is illustrated in Figure 1. The overall structure of HGI is as follows. (1) A POI category encoder ϕ_c is pretrained to generate the initial POI features, as categorical information generally plays a key role in defining the meaning of a POI. (2) A graph structure (X_p, A_p) for POIs is constructed using Delaunay triangulation (DT), and a graph convolution encoder ϕ_p is applied to the POI graph, so as to encode the spatial context and the uniqueness of each POI. (3) We utilize an aggregation function $\mathrm{AGG}_{poi-region}$ based on multi-head attention mechanisms to aggregate POI embeddings and generate region raw embeddings, in order to consider the interactions between POIs and regions from multiple perspectives. (4) A graph structure for regions (X_r, A_r) is constructed based on their adjacency relations, and then another graph convolution encoder ϕ_r is applied to this regional graph, to consider the contextual information at the regional level; this step updates region raw embeddings to the region embeddings used in downstream tasks. (5) An area-weighted summarization function $\mathrm{AGG}_{region-city}$ is applied to generate the embedding at the city level from region embeddings. (6) We perform negative sampling at both POI and regional levels, and develop a hierarchical infomax objective function to learn region embeddings in a fully unsupervised manner, i.e., the mutual information among the POI – region – city hierarchy is leveraged as the objective to train the model, including the components of ϕ_p, $\mathrm{AGG}_{poi-region}$, and ϕ_r.

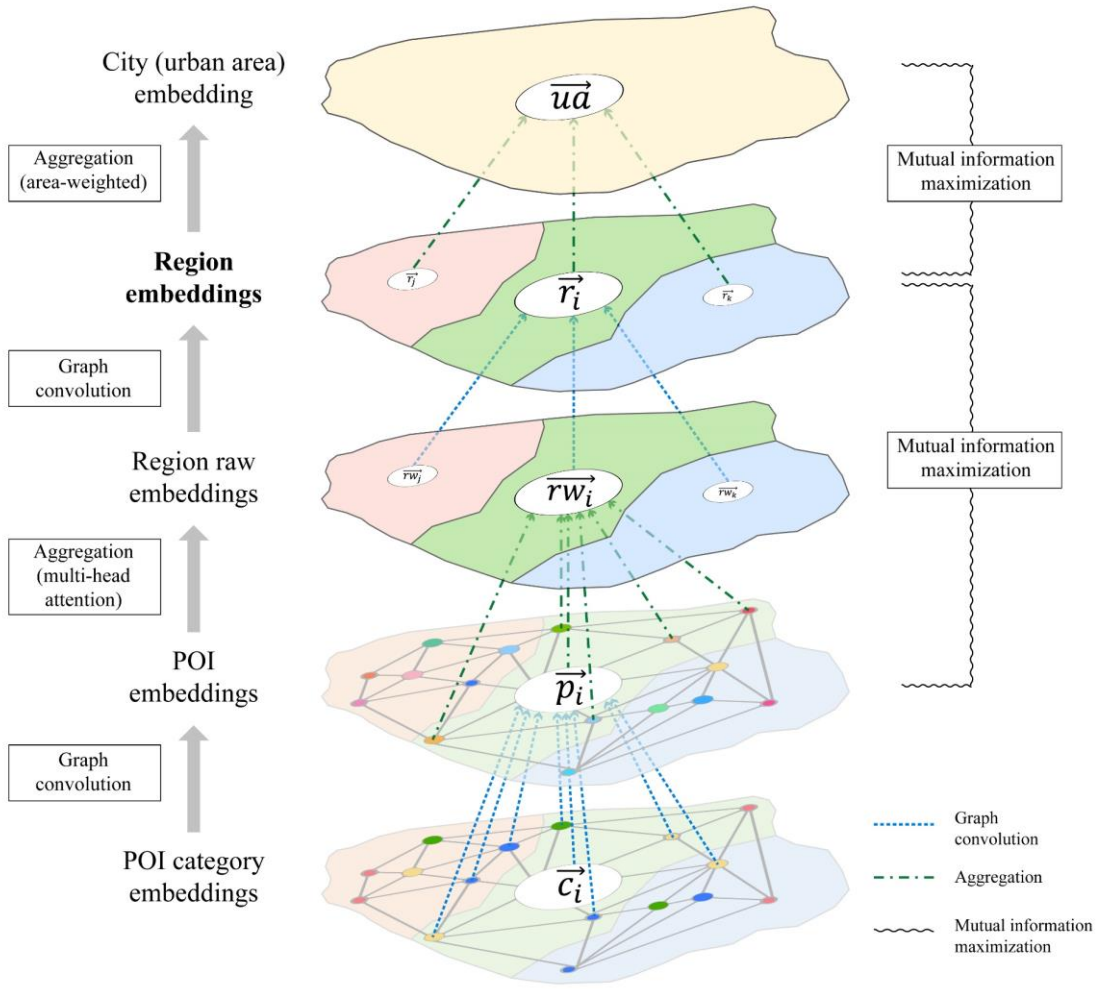

Figure 1. The overarching architecture of HGI for learning region representations.

3.2 POI category embedding

The first component in HGI is an encoder to generate POI category embeddings, which are served as initial node features in the subsequent graph learning procedures. The rationale is that category embeddings can largely reflect POI semantics, even though the uniqueness of each individual POI is neglected. There have been several studies on developing POI category encoders, such as Yao et al. (2017a), Yan et al. (2017), and Huang et al. (2022). In this paper, we employ the state-of-the-art approach in this strand, i.e., the semantic POI category embedding developed by Huang et al. (2022), which outperformed other methods in functional region inference.

Specifically, the POI category encoder ϕ_c learns embeddings for second-level POI categories, as POI categories often come with hierarchies, e.g., in Baidu Map POIs[1], a first-level category *shopping* subsumes several second-level categories such as *supermarket*, *mall*, and *grocery*. First, ϕ_c conducts spatially explicit random walks in a POI DT network to capture category co-occurrence patterns, which considers spatial distance decay, the balance between local and long-range co-occurrences, and the differentiation between intra- and cross-region co-occurrences. Second, ϕ_c incorporates categorical semantics, and employs a manifold learning algorithm, i.e., Laplacian Eigenmaps, to enforce the second-level categories that share the same first-level category to be adjacent in the embedding space. Finally, the POI category embeddings are obtained through optimizing the objective function:

$$\mathcal{L}_{\phi_c} = \sum_{c \in \mathcal{C}} \sum_{c_q \in N_{R(c)}} -\left(\log\left(\sigma_{\phi_c}\left(\vec{c}^{\mathrm{T}} \vec{c}'_{\mathrm{q}}\right)\right) - \sum_{i=1}^{k} \log\left(\sigma_{\phi_c}\left(\vec{c}^{\mathrm{T}} \vec{c}'_{n_i}\right)\right)\right) + \frac{\lambda}{2} \sum_{c_i, c_j \in \mathcal{C}} w_{ij} \left\| \vec{c}_i - \vec{c}_j \right\|_2^2 \tag{1}$$

where $\mathcal{C}$ is the set of second-level categories, $\vec{c}$ and $\vec{c}'$ denote the target and context embeddings of the second-level category c, and $N_{R(c)}$ represents the set of context categories; σ_{ϕ_c} denotes a *sigmoid* function; c_{n_i} means the categories obtained by the negative sampling process; w_{ij} = 1 if c_i and c_j belong to the same first-level category, and otherwise w_{ij} = 0; $\|\cdot\|_2^2$ denotes the operation squared ℓ_2 norm. For details, one could refer to Huang et al. (2022).

[1] https://map.baidu.com/

3.3 POI encoder that incorporates the uniqueness of each individual POI

The previous step encodes the semantics of POI categories, yet neglects the uniqueness of each single POI. In principle, POIs with the same category could manifest subtly divergent semantics due to disparate spatial contexts. For example, a *hotel* within an airport complex and another *hotel* in a university campus, despite their resemblance, have diverse semantics as the former can be largely linked to *transportation* function, while the latter could be related to *educational* purposes. Encoding the POI uniqueness could generate more meaningful POI and region embeddings.

Intuitively, the uniqueness of an individual POI can be shaped by its spatial context, i.e., its nearby POIs. To this end, a promising remedy is that modeling POIs using a graph structure and utilize the message passing mechanism in GCN, so as to generate POI embeddings through aggregating the category embeddings in a POI neighborhood. The reasons of using a graph structure are threefold: (1) graphs are more spatially compact for POIs compared to the rigid grid and sequential descriptors; (2) graphs are robust under affine and rotation transformations, i.e., its expressiveness is not compromised under varying coordinate reference systems and orientations of maps; (3) the message passing mechanism naturally models the interactions between POIs, and thus captures the spatial contexts of individual POIs. There are many different methods for constructing graphs for POIs. In particular, DT constructs compact and informative graphs from spatial vector data. Many previous studies have verified the fitness of DT graphs for modeling the interactions among spatial vector data (Yan et al., 2019; Huang et al., 2022; Xu et al., 2022b). Along this line, all the POIs in a study area are connected using DT to form a graph, then the POIs become graph nodes with the category embeddings as node features, i.e., $X_p = \{\vec{c}_1, \vec{c}_2, \dots, \vec{c}_N\}$. In the POI graph, each edge is

assigned with an unnormalized weight depending on the spatial distance and whether the two POIs are within a same region:

$$a_p(p_i, p_j) = \log\left((1 + L^{1.5})/\left(1 + l_{p_i p_j}^{1.5}\right)\right) \times w_r \quad (2)$$

where L denotes the diagonal length of the minimum bounding rectangle of all the POIs, and l represents the spatial distance between the two POIs that form the edge; w_r is a factor to differentiate intra- (w_r = 1) and cross-region (w_r = 0.4) edges. The choices of the parameters of $a_p \sim l_{p_i p_j}^{-1.5}$ and w_r = 0.4 (cross-region) are in view of the previous practices in Calafiore et al. (2021) and Huang et al. (2022). The weights are finally normalized with a linear re-scaling to [0, 1].

Upon the constructed POI graph, we apply a one-layer GCN encoder (Kipf and Welling, 2017) to generate POI embeddings with the message passing rule:

$$\phi_p(X_p, A_p) = \sigma_{\phi_p}\left(\widehat{D}_p^{-\frac{1}{2}} \hat{A}_p \widehat{D}_p^{-\frac{1}{2}} X_p \Theta_p\right) \quad (3)$$

in which $\hat{A}_p = A_p + I_N$ is the weighted POI adjacency matrix with self-loops; $\widehat{D}_p$ is the degree matrix of $\hat{A}_p$; σ_{ϕ_p} is a parametric *ReLU* (*PReLU*) function (He et al., 2015); $\Theta_p \in \mathbb{R}^{F_c \times F_p}$ is a learnable linear transformation applied to every node, where F_c is the dimension of POI category embeddings, while F_p is POI embeddings' dimension. After applying ϕ_p, the embedding of each POI is updated to be a transformed combination of the information from itself and its spatial context, which captures its uniqueness.

3.4 POI aggregation with multi-head attention mechanisms

Naturally, the next step after obtaining POI embeddings is to aggregate them and generate region embeddings in a meaningful manner. As discussed above, the aggregation should consider the differences of each POI's importance in defining its region, and the importance usually should be portrayed from multiple perspectives.

Resuming our previous example of a region having a lake and a building, the two POIs can have different importance levels to defining their region from multiple viewpoints, e.g., area occupation, economic productivity, population allocation, and cultural and leisure aspects. As we aim to learn region representations that can be used for multiple purposes, it is necessary to model such multi-faceted influence.

The multi-head attention mechanisms developed in *Transformer* (Vaswani et al., 2017; Lee et al., 2019) provides a plausible solution to tackle this problem. Specifically, an attention operation captures the relative correlations (importance levels) between two set of entities, namely how important of each POI in defining its region from a single perspective. An attention function maps query vectors $Q \in \mathbb{R}^{n_v \times d_q}$ to outputs using n_v key-value pairs $K \in \mathbb{R}^{n_v \times d_k}$, $V \in \mathbb{R}^{n_v \times d_v}$ ($d_q = d_k = d_v = d$).

$$\mathrm{Att}(Q, K, V) = \omega(QK^{\mathrm{T}})V \tag{4}$$

where ω is a scaled *softmax* function, i.e., $\omega(\cdot) = \mathrm{softmax}\left(\cdot/\sqrt{d}\right)$. The above attention operation can be extended to capture the interactions between POIs and their regions from multiple perspectives, i.e., instead of computing a single attention function, we can first project Q, K, V onto h separate vectors with the dimensions $d_q^M = d_k^M = d_v^M = d/h$ (the h projections are denoted multi-heads). Then an attention function is applied to each of these h projections, and the output is a linear transformation of the concatenation of all attention outputs:

$$\mathrm{Multihead}(Q, K, V) = \mathrm{concat}(O_1, O_2, \dots, O_h)W^O \tag{5}$$

$$\text{where } O_j = \mathrm{Att}(QW_j^Q, KW_j^K, VW_j^V) \tag{6}$$

in which $\left\{W_j^Q, W_j^K, W_j^V\right\}_{j=1}^h$ are learnable parameters.

Inspired by the *set transformer* developed by Lee et al. (2019), we can then define the aggregation function $\mathrm{AGG}_{POI-region}$ using the multi-head attention mechanisms as follows:

$$\mathrm{AGG}_{POI-region}(P_i) = H + \mathrm{rFF}(H) \tag{7}$$

$$\text{where } H = \vec{s}_i + \mathrm{Multihead}(\vec{s}_i, P_i, P_i) \tag{8}$$

in which P_i is the embeddings of the POIs in a region r_i; $\vec{s}_i$ is a randomly initialized and learnable seed vector, to which the POI embeddings in the region r_i pay multi-head attention, rFF is a linear transformation followed by a *ReLU* activation function. In this manner, the output of $\mathrm{AGG}_{POI-region}$ is a region raw embedding $\overrightarrow{rw_i}$ that aggregates the POI embeddings in the region, and $\overrightarrow{rw_i}$ reflects the different importance levels of the POIs from several perspectives (each head represents a perspective).

3.5 Region encoder that incorporates the information from adjacent regions

Thus far we have reached the regional-level representations which entirely depend on the POIs in the regions. As mentioned above, the spatial context information at the regional level could be pivotal, as nearby regions usually carry substantial similarities. Encoding adjacency information at the regional level could lift the expressiveness and discrimination power of the learned region embeddings.

In this context, we construct a regional graph to facilitate the information propagation between adjacent regions. Specifically, we conceptually view each region as a node, and build edges between the regions that share parts of borders (or each region is connected with its circumambient regions in case regions do not share borders). In this regional graph, all edges have the same importance, i.e., no specific weight is assigned. We employ a one-layer GCN to help each region sense its ambient, which is formally defined as:

$$\phi_r(X_r, A_r) = \sigma_{\phi_r}\left(\widehat{D}_r^{-\frac{1}{2}}\hat{A}_r\widehat{D}_r^{-\frac{1}{2}}X_r\Theta_r\right) \quad (9)$$

in which $\hat{A}_r = A_r + I_n$ is the adjacency matrix with self-loops; $\widehat{D}_r$ is the degree matrix of $\hat{A}_r$; σ_{ϕ_r} is a *PReLU* function; $X_r = \{\overrightarrow{rw}_1, \overrightarrow{rw}_2, \dots, \overrightarrow{rw}_n\}$ is the node features of the regional graph, with each region characterized by its region raw embedding $\overrightarrow{rw}_i$. $\Theta_r \in \mathbb{R}^{F_{rw} \times F_r}$ is a learnable linear transformation applied to every region, where F_{rw} is the dimension of region raw embeddings, while $F_r = d$ is the dimension of final region embeddings. After applying ϕ_r, each region embedding is informed by its ambient regions, and updated to $\vec{r}_i$, which serves as the region representation for downstream tasks.

3.6 Region aggregation with an area-weighted summarization

The next step is to aggregate region embeddings to generate a global representation at the city level. The rationale of generating city-level representation is to provide useful signal for training HGI, which could bring global information to region embeddings. In other words, a city-level representation could help sense general commonalities among all the regions in the study area.

In principle, we could also have a sophisticated aggregation function like $\text{AGG}_{POI-region}$ based on multi-head attention mechanisms. However, this step is to aggregate the nodes in the regional graph, which usually only has hundreds or thousands of nodes (regions). In this setting, according to the previous practice in Veličković et al. (2019), deterministic aggregation functions (e.g., average pooling) perform better than sophisticated architectures (e.g., multi-head attention mechanisms), as the latter can hardly be sufficiently trained with such a small number of nodes. In this study, instead of simply averaging all regions' representations, we leverage the prior knowledge of region areas, and believe that larger regions are generally more

definitive than small ones. Therefore, we utilize an area-weighted summarization as follows:

$$\mathrm{AGG}_{region-city}(R) = \sigma_{rc}(\sum_{i}^{n} \vec{r}_i aw_i) \tag{10}$$

where aw_i is the area proportion of region r_i in the study area (city), and σ_{rc} is a *sigmoid* function; $\mathrm{AGG}_{region-city}$ generates an embedding for the target urban area $\overrightarrow{ua}$.

3.7 Negative sampling and training objective

Till now, we have obtained all the representations at the POI, region, and city levels in the feedforward process, and the subsequent step is to define a meaningful objective to train HGI. The general idea of HGI follows the paradigm of contrastive learning, so its success depends on fulfilling the two properties of *alignment* and *uniformity* (Wang and Isola, 2020). The former can be achieved using POI – region, and region – city positive pairs, while the latter needs to be accomplished by negative sampling at both POI and region levels. The training of HGI only relies on POIs and region boundaries, and not on any ground truth data in the downstream tasks.

For a region r_i, we maximize the mutual information between its region embedding $\vec{r}_i$ and the POI embeddings P_i within the region, thereby they are positive samples. The negative samples for r_i can naturally be the POI embeddings P_j in another region r_j. In addition, we conduct hard negative sampling to prevent the model from converging to a trivial solution (Liu et al., 2018; Robinson et al., 2020). The intuition behind hard negative sampling is that if the sampling process predominately chooses POIs from very dissimilar regions as negative samples, only little useful gradient can be learned from them; it is thus crucial for the model to see some hardly distinguishable negative samples, so as to boost its discrimination power. Specifically, we represent each POI as the concatenation of its first- and second-level category one-hot

embeddings, and use an average operation to generate a rough region embedding. The hard negative sampling process selects a region r_j if the rough region embeddings of r_i and r_j have a cosine similarity between [0.6, 0.8]. In the learning process, we construct a small portion of hard negative samples among the training samples.

For the urban area (city) $\overrightarrow{ua}$, the region embeddings R are its positive samples, while the negative samples are generated through row-wise shuffling of the POI graph's feature matrix X_p to form a $\tilde{X}_p$ (replace the category embedding of each POI with a category embedding from another randomly picked POI to form a corrupted graph). Afterwards, the POI encoder ϕ_p, $\mathrm{AGG}_{POI-region}$, and the regional graph encoder ϕ_r are applied to generate corrupted region embeddings $\tilde{R}$ to be the negative samples of $\overrightarrow{ua}$.

Finally, the model can be optimized through minimizing the below objective:

$$\mathcal{L} = \alpha\mathcal{L}_{pr} + (1-\alpha)\mathcal{L}_{rc} \tag{11}$$

$$\mathcal{L}_{pr} = -\left(\frac{1}{n}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\log \mathcal{D}_{pr}(\vec{p}_i,\vec{r}_k) + \frac{1}{\tilde{n}}\sum_{k=1}^{N}\sum_{j=1}^{\tilde{n}_k}\log\left(1-\mathcal{D}_{pr}(\tilde{p}_j,\vec{r}_k)\right)\right) \tag{12}$$

$$\mathcal{L}_{rc} = -\frac{1}{N}\left(\sum_{k=1}^{N}\log \mathcal{D}_{rc}(\vec{r}_k,\overrightarrow{ua}) + \sum_{k=1}^{N}\log\left(1-\mathcal{D}_{rc}(\tilde{r}_k,\overrightarrow{ua})\right)\right) \tag{13}$$

where $n = \sum_{k=1}^{N} n_k$ and $\tilde{n} = \sum_{k=1}^{N} \tilde{n}_k$ with N being the number of regions, as well as n_k and $\tilde{n}_k$ being the number of positive and negative POIs for region k; $\mathcal{D}_{pr}(\vec{p}_i,\vec{r}_k) = \sigma_{\mathcal{L}}\left(\vec{p}_i W_{pr}\vec{r}_k\right)$ and $\mathcal{D}_{pr}(\vec{p}_i,\vec{r}_k) = \sigma_{\mathcal{L}}\left(\vec{p}_i W_{pr}\vec{r}_k\right)$; $\vec{r}_k$ is the learned embedding for the region r_k, while $\tilde{r}_k$ is its corrupted embedding; $\vec{p}_i$ is the embedding of a POI p_i located in r_k, while $\tilde{p}_j$ comes from another region; $\sigma_{\mathcal{L}}$ is a *sigmoid* function. The intuition behind the complex objective is a region-centered hierarchical mutual information maximization. $\mathcal{L}_{pr}$ maximizes the mutual information between each region embedding and the POI embeddings within the region, while pushes each region embedding away from the POI embeddings from another region (negative samples). Likewise, $\mathcal{L}_{rc}$ maximizes the

mutual information between each region embedding and the city embedding, while pushes the city embedding away from the corrupted region embeddings. The strengths of $\mathcal{L}_{pr}$ and $\mathcal{L}_{rc}$ are controlled by α, which should be tuned to find the best balance point. With this region-centered objective, the information from local-scale POIs and the global-scale city both flows to the learned region embeddings, making them both globally and locally relevant.

4 Experiment and results

4.1 Study areas and data

We generate region embeddings using HGI to demonstrate its effectiveness in two study areas: Xiamen Island and Shenzhen, China. Both two urban areas are economically prosperous, densely populated, yet extremely land-scarce, therefore sensible urban planning and management are crucial. In this context, learning meaningful and multi-purpose region representations is desirable for many real-world urban problems that can (partially) be delineated by the information entailed from POIs.

We learn region representations in the study areas using two sources of data: (1) POIs in the two areas harvested from Baidu Map in 2020, including 45,033 POIs in Xiamen Island and 303,428 POIs in Shenzhen; the POIs belong to 22 first-level categories (e.g., *food service*, and *governmental agency*), which subsume 166 second-level categories in Xiamen Island while 178 in Shenzhen (e.g., *supermarket*, and *Chinese restaurant*). (2) Traffic analysis zones (TAZs) in these two study areas, which partition each study area into many regions by the networks of major roads; we have 661 regions in Xiamen Island and 5,461 regions in Shenzhen with POIs inside. We aim to learn a representation for each region in the two study areas.

We evaluate the learned region representations in three downstream tasks. i.e., estimating urban functional distributions (the proportions of different functions that each region affords), housing price, and population density, which are pivotal tasks in understanding our cities to help effective planning. Therefore, we also acquire ground truth data for the tasks: (1) the urban function ground truth is from *Urbanscape Essential Dataset of Peking University*[2], which provides detailed spatial distributions of ten urban functions in the study areas, including *forest*, *water*, *unutilized*, *transportation*, *green space*, *industrial*, *educational & governmental*, *commercial*, *residential*, and *agricultural*; the data is produced through information extraction from remote sensing data and POIs, supplemented with human correction and modification; such data is overlapped with TAZs to obtain a ground truth functional distribution of each region. (2) Housing price data crawled from a Chinese real estate agent Lianjia[3] in 2021, and the housing prices in each region are averaged as ground truth. (3) Population density data from WorldPop 2020[4] is used; we overlap the regions with the grids in WorldPop to obtain the population density in each region.

4.2 Implementation details of HGI

We first generate POI category embeddings (64-dimensional) using the approach proposed in Huang et al. (2022) to serve as the initial POI (node) features in the POI graphs, where the Xiamen POI graph has 270,123 edges and the Shenzhen POI graph has 1,820,487 edges. Then we train the HGI model to generate region representations with the proportion of hard negative samples to be 0.25. We perform parameter analyses

[2] http://geoscape.pku.edu.cn/en.html

[3] lianjia.com

[4] https://worldpopulationreview.com/

in the study area of Xiamen Island, in which the region embedding dimension d is tuned in {32, 64, 128, 256}, the number of heads h in the aggregation function $\mathrm{AGG}_{POI-region}$ is tuned in {1, 2, 4, 8}, and the parameter α in the objective function (Equation 11) is tuned in {0.1, 0.3, 0.5, 0.7, 0.9}. We find that the parameters that lead to the best performance are d =64, h = 4, and α = 0.5. We train HGI in the two study areas separately for 2,000 epochs without a minibatch mode. During training, we set the learning rate to 0.006, and use the gradient clipping technique (maximum norm of the gradients is 0.9) as well as linear learning rate warmup for a period of 40 epochs. We find that with this training strategy, the training is both stable (not prone to collapse despite its complex loss landscapes) and efficient (with generally a large learning rate). At last, region embeddings obtained in the epoch with the lowest loss (Equation 11) are retrieved to be used in downstream tasks. Under the found best parameters, the training of HGI takes ~2.3 hours (~4.1s/epoch) and consumes 1,443MB GPU memory in Xiamen Island, and takes ~63.5 hours (~114.3s/epoch) and consumes 3,873MB GPU memory in Shenzhen using a single NVIDIA V100 GPU. Note that as an unsupervised representation learning model, HGI only entails training, so its inference efficiency is not reported.

4.3 Baseline models

We compare HGI with several baseline models that all produce 64-dimensional region representations in the study areas:

(1) Semantics preserved POI embedding (Semantic; Huang et al., 2022): This approach considers both spatial co-occurrence information and categorical semantics, and generates POI category embeddings with random walks and a manifold learning algorithm. In the urban function task (*task 1*), region

representations are obtained using either a supervised aggregation function coupling LSTM and attention mechanisms or an average pooling in an unsupervised setting; we test both the aggregation means and denote them *Semantic-s* (supervised) and *Semantic-u* (unsupervised).

(2) Place2Vec (Yan et al., 2017): This approach considers spatial co-occurrence information with a KNN sampling strategy and distance decay to learn POI category embeddings, and then generates region embeddings through averaging the POI category embeddings in each region.

(3) Doc2Vec (Niu and Silva., 2021): This model first samples POI category co-occurrences using a KNN strategy, and regards each POI as a 'word', while each region as a 'document' to train the embeddings for POI categories and regions in a conjunctive manner. This means that the aggregation from POIs to regions is performed during training the model.

(4) Latent Dirichlet Allocation (LDA): It is a generative statistical model to discover the 'topics' of a region based on its POIs, which relies on the information of POI category frequencies in each region. LDA generates a proportional distribution of topics for each region, which can be regarded as a region representation. Such topic model-based region representations have been explored in several studies, e.g., Gao et al. (2017), and Su et al. (2021).

4.4 Task 1: Estimating urban functional distributions

Our cities are composed of many regions that bear various functions for human activities, e.g., *residential*, *commercial*, and *industrial*. These functional regions often serve as the basic units of urban planning, and they greatly impact, among others, urban transportation, and resource management (Du et al., 2019, 2020; Zhang et al., 2017). In reality, the function of each region is not homogenous, but a composition of several

different functions. Therefore, understanding urban functional distributions, i.e., the proportions of different functions that each region bears, is pivotal for urban planning and management (Zhang et al., 2021).

In *task 1*, we utilize the HGI region embeddings to estimate urban functional distributions. To this end, we input the region embeddings to a shallow multilayer perceptron (MLP) with one 512-dimensional hidden layer (with *1D-batchnorm* and *tanh* activation function), and a ten-dimensional output layer with *softmax* activation function (as we estimate proportional distributions for ten functional types). The MLP is trained with the urban function ground truth separately in the two study areas (ground truth is available in all regions in Xiamen Island and 5,344 regions in Shenzhen), and a KL divergence loss function. For each study area, we randomly split the dataset into training, validation, and test datasets with the ratio of 6:2:2. We repeat the random split for ten times and experiment on them for reliability. In addition, we utilize the same MLP architecture for the baseline models; for Semantic-s, a supervised aggregation function is used (see Section 4.3).

In terms of the evaluation measures, we follow the practices from Huang et al. (2022) and treat the task as a label distribution learning problem (Geng, 2016), for which we select three representative measures (↓denotes the smaller the better, and ↑indicates the opposite):

(1) L1 distance (L1) ↓: $\sum_{k=1}^{m}\left|\hat{y}_i^{f_k} - y_i^{f_k}\right|$

(2) KL divergence (KL)↓: $\sum_{k=1}^{m} y_i^{f_k} \log\left(y_i^{f_k} / \hat{y}_i^{f_k}\right)$

(3) Cosine similarity (Cosine)↑: $\left(\sum_{k=1}^{m} \hat{y}_i^{f_k} y_i^{f_k}\right) \Big/ \left(\sqrt{\sum_{k=1}^{m} \hat{y}_i^{f_k}{}^2} \sqrt{\sum_{k=1}^{m} y_i^{f_k}{}^2}\right)$

where $\hat{y}_i^{f_k}$ is the estimated proportion of the function type f_k that region i bears, and

$y_i^{f_k}$ is the corresponding ground truth proportion.

4.4.1: Performance

We present the results of estimating functional distributions in Table 1 (mean ± standard deviation), and spatially visualize the results in terms of L1 distance in Figure 2. We can observe that the region embeddings produced by HGI (d =64, h = 4, and α = 0.5) surpasses the competitive baselines in the two study areas by all evaluation measures. This means that the functional distributions estimated using the HGI region representations have the smallest distance and relative entropy, as well as the greatest similarity to real distributions, which implies that the region representations produced by HGI are more meaningful and informative than others for sensing urban functions. Particularly, HGI can exceed the baseline Semantic-s using LSTM and attention mechanisms for aggregating POIs, which means that even using the ground truth signal to supervise the generation of region embeddings from POI embeddings can barely approach the expressiveness of HGI region embeddings. We also observe that the estimations are harder in Shenzhen than in Xiamen Island, and HGI obtains more notable gain in the more difficult study area Shenzhen (compared to Semantic-s), which is encouraging as HGI could have the potentials to be used in other complex and heterogenous cities.

As to the baselines, Semantic yields generally better results than others, and we use such embeddings as the initial POI features for HGI. Place2Vec and Doc2Vec come after Semantic with similar performance, but are both better than LDA, implying that representation learning methods (all methods but LDA) can generally outperform statistical language models such as LDA for this task, which has also been verified in a series of previous studies, e.g., Yao et al. (2017a); Zhai et al. (2019); Niu and Silva (2021).

Table 1. *Task 1* performance[5]. The best value for each evaluation measure is bolded.

Model	Xiamen Island			Shenzhen		
	L1↓	KL↓	Cosine↑	L1↓	KL↓	Cosine↑
HGI	**0.727±0.019**	**0.587±0.033**	**0.795±0.017**	**0.936±0.016**	**0.833±0.025**	**0.707±0.009**
Semantic-s	0.754±0.029	0.619±0.038	0.789±0.014	0.972±0.011	0.877±0.024	0.695±0.009
Semantic-u	0.784±0.024	0.658±0.032	0.770±0.015	0.980±0.015	0.893±0.024	0.693±0.008
Place2Vec	0.835±0.023	0.724±0.037	0.748±0.015	0.982±0.016	0.895±0.023	0.694±0.007
Doc2Vec	0.833±0.027	0.728±0.047	0.749±0.023	0.985±0.012	0.894±0.015	0.692±0.007
LDA	0.896±0.028	0.848±0.051	0.711±0.017	1.000±0.015	0.924±0.022	0.678±0.008

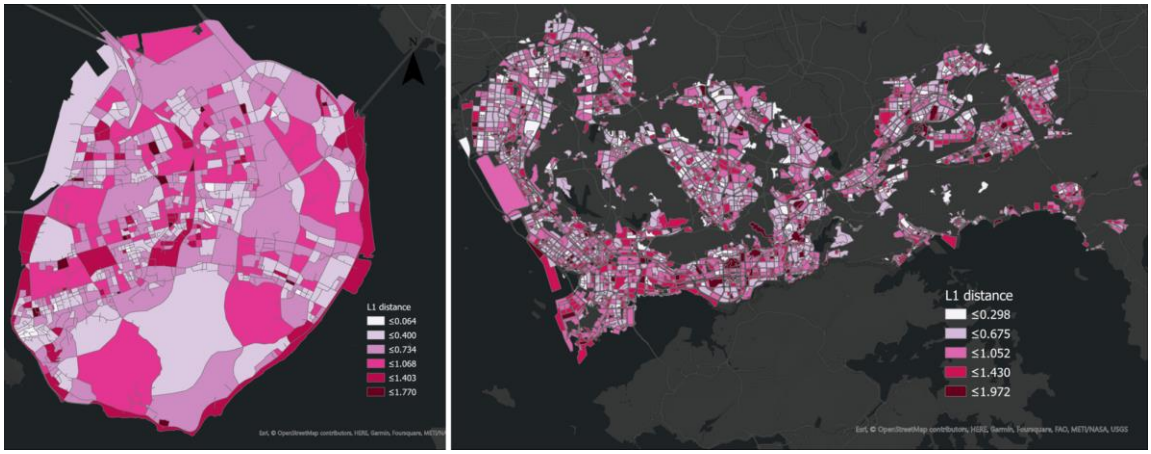

Figure 2. Spatial distribution of L1 distance errors between estimated and real functional distributions.

4.4.2: Parameter sensitivity analyses

We tune three important hyperparameters in the study area of Xiamen Island, and their results are presented in Figure 3. The first is α in the objective function (Equation 11), which balances the contributions of two mutual information maximization objectives

[5] The performances of some baseline methods are different than those reported in Huang et al., (2022), mainly because we adopt a different evaluation protocol, e.g., with fewer training samples.

among the POI – region – city hierarchy (cf. Section 3.6), we tune it in {0.1, 0.3, 0.5, 0.7, 0.9}, where larger α represents that the bonds between POIs and regions are stronger, while the correlations between regions and the city are weaker. We can observe that α = 0.5 yields the best performance, meaning that equal strengths of the information flows from POIs and the city to regions are favorable. Generally, the mutual information between POIs and regions should be no weaker than that between regions and the city, as α < 0.5 results in larger performance decline. We believe that this is reasonable as most original information is captured from POIs. The second hyperparameter is the dimension *h* of region embeddings, which we tune in {32, 64, 128, 256}. We can observe that the performance gains as *h* increases to 64, while drops at 128 and 256, which indicates that *h* = 64 leads to the best balance between informativeness and efficiency. It is also encouraging to observe that HGI outperforms all the baselines with the same embedding size (cf. Table 1). The third hyperparameter we tune is the number of heads in the POI aggregation function $\text{AGG}_{POI-region}$ using multi-head attention mechanisms, and we can observe that *h* = 4 leads to the best performance. This is probably because that smaller numbers of heads are less expressive, while larger numbers of heads would make the model harder to be trained. Overall, it is a favorable property of HGI that its performance is generally not very sensitive to different hyperparameter settings.

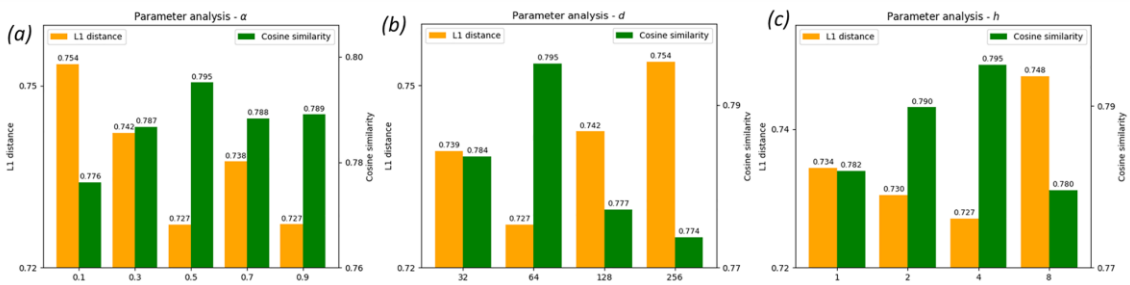

Figure 3. Parameter analyses results for (a) α in the objective function, (b) region representation dimension *d*, and (c) the number of heads *h* in the aggregation function $\text{AGG}_{POI-region}$.

4.4.3: Ablation study

We drop or replace several key components of HGI to verify their necessity in Xiamen Island, and the results are presented in Table 2. The ablations are: (1) we replace the aggregation function $\mathrm{AGG}_{poi-region}$ from POIs to regions with an average pooling, so that consider each POI equally in the aggregation process; (2) we replace the aggregation function from regions to the city embedding $\mathrm{AGG}_{region-city}$ with an average pooling; (3) we drop the graph convolution encoder ϕ_r for the regional graph, so that neglect the spatial contextual information at the regional level (note that, in this case, the infomax between region and city levels is also dropped); (4) we drop hard negative sampling strategy in the negative sampling process, so that the POI negative samples are uniformly selected; (5) we drop the mutual information maximization between regions and the city embedding, i.e., α = 1.0 in the objective function (Equation 11). We observe that all ablations lead to performance decline, particularly (3) drop region encoder and (1) average pooling for POIs result in larger performance drops than others. Overall, the full version of HGI yields the best performance.

Table 2. Ablation study results of *Task 1* in Xiamen Island. The best value for each evaluation measure is bolded.

Model	L1↓	KL↓	Cosine↑
HGI	**0.727±0.019**	**0.587±0.033**	**0.795±0.017**
(1) Average pooling for POIs	0.750±0.028	0.619±0.047	0.779±0.018
(2) Average pooling for regions	0.733±0.037	0.607±0.063	0.788±0.021
(3) Drop region encoder	0.763±0.022	0.654±0.028	0.772±0.011
(4) Drop hard sampling	0.736±0.025	0.616±0.028	0.781±0.014
(5) Drop zone-city infomax	0.738±0.024	0.603±0.041	0.787±0.019

4.5 Task 2: Population density estimation

Understanding the spatial distribution of human populations underpins various operational work, policy making, and scientific research in many related areas, e.g., disaster relief, city planning and management. Traditional population data are usually collected by census methods, which are not only labor-intensive, but also coarse in terms of granularity. Therefore, numerous studies have utilized geospatial data from various sources for population estimation at finer scales, e.g., remote sensing data and POIs (Yao et al., 2017b; Shang et al., 2020). In particular, POIs are indicative for this task, in virtue of their rich semantic information delineating human activities. Ye et al. (2019) unveiled that POIs are more useful than the brightness of night-light remote sensing data for population estimation in their study. However, most practices only used simple features of POIs, e.g., POI density, which leads to the loss of some latent information, e.g., adjacent regions are likely to have similar population densities.

In *task 2*, we use the learned region representations to estimate population density for each region against the baseline models. To this end, we input the region representations to a random forest regression model (RF; with 100 decision trees), and use 80% of the regions as training data, and the remaining 20% of the regions as testing data. We repeat the experiment for 10 times with random training test set splits, and finally report the averaged performance metrics. For evaluation, we use the classic metrics for regression tasks, i.e., root mean squared error (RMSE), mean absolute error (MAE), and R^2.

The performance of this task is presented in Table 3. We observe that HGI considerably outperforms the baselines in both study areas. HGI can obtain much less absolute errors (MAE) in both study areas. We can also observe that the HGI region embeddings yield much higher R^2 than the baselines; the variance of population density can be well captured by HGI to the extent of more than 60%. As to the baselines,

Semantic, Place2Vec, Doc2Vec generally produce comparable results, with R^2 mostly ranging between 0.1 and 0.2; LDA yields generally poorer performance than others. The results clearly demonstrate that HGI can better estimate population density in urban areas, which proves that HGI can generate more meaningful region representations for this task.

Table 3. *Task 2* performance. The best value for each evaluation measure is bolded. Units of RMSE and MAE metrics are number of people/km^2.

Model	Xiamen Island			Shenzhen		
	RMSE↓	MAE↓	R^2↑	RMSE↓	MAE↓	R^2↑
HGI	**43775.93**	**26891.21**	**0.645**	**8323.92**	**5395.82**	**0.609**
Semantic	67877.00	44941.46	0.111	11905.07	8278.01	0.191
Place2Vec	66384.16	45162.36	0.137	12017.22	8525.36	0.153
Doc2Vec	68826.79	47589.97	0.123	11832.19	8327.05	0.208
LDA	78533.49	48364.68	-0.098	12895.36	8783.99	0.060

For further analyzing the estimation errors, we spatially visualize the absolute errors in Figure 4. We notice that the estimations are generally performed well in most regions, particularly those with large and modest sizes. However, it is notable that large discrepancies usually appear in small regions, which raises the overall MAEs. We believe that this owes to that the ground truth data is from a coarser granularity, which may not well reflect the population densities in small regions; in fact, in such cases the estimations might be even more accurate. Also, the common underestimation problems in urban areas of WorldPop (Ye et al., 2019) may also be an underlying cause of the large discrepancies in small regions.

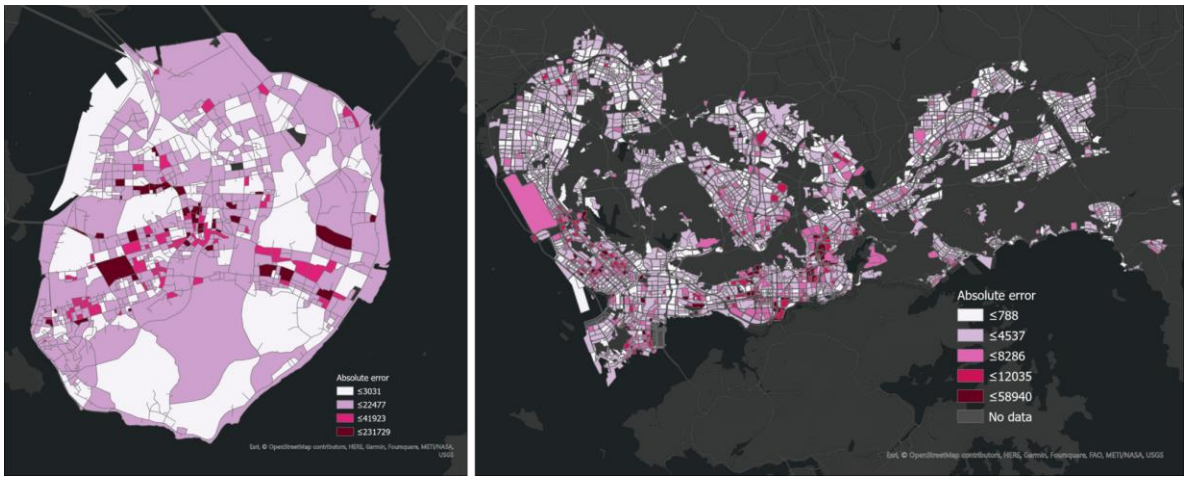

Figure 4. Spatial distribution of absolute estimation errors for population density.

4.6 Task 3: Housing price estimation

Housing prices play an important role in human habitation and economic development. Therefore, there have been tremendous amounts of studies for modeling and estimating housing prices amongst urban planners, economists, and policymakers (Cao et al., 2019). Housing prices are affected by numerous factors, and there are two types of information that have been commonly acknowledged to be pivotal, namely structural attributes and locational amenities. In this context, POIs are a powerful source of information to characterize locational amenities.

In *task 3*, we utilize the learned region representations for estimating housing prices with the same RF regression model, experimental settings, and evaluation measures as *task 2*. The housing price ground truth data is only partially available in the two study areas (424 regions in Xiamen Island, and 926 in Shenzhen), from which we select training and test datasets. The evaluation results are presented in Table 4. We find that HGI continues to prevail in this task, with the smallest errors across the study areas and highest values of R^2. The values of R^2 of HGI region embeddings indicate that they can explain the variance of housing prices to a larger extent than the baselines. The

baselines have generally similar performances, and representation learning methods continue to outperform LDA.

Table 4. *Task 3* performance. The best value for each evaluation measure is bolded. Units of RMSE and MAE metrics are CNY/m^2.

Model	Xiamen Island			Shenzhen		
	RMSE↓	MAE↓	R^2↑	RMSE↓	MAE↓	R^2↑
HGI	**11760.69**	**9075.53**	**0.346**	**19395.46**	**13682.50**	**0.416**
Semantic	13202.07	10668.92	0.132	23341.21	17507.56	0.113
Place2Vec	13591.83	10897.83	0.139	22732.39	16966.87	0.141
Doc2Vec	13293.15	10614.82	0.130	22928.67	17394.86	0.136
LDA	15525.62	12454.23	-0.127	23819.20	17625.85	0.021

We visualize the absolute errors in Figure 5 in the study areas, from which we can uncover that the estimations are performed well in city centers, while large discrepancies mostly arise in outskirts and large regions. We speculate this is due to that (1) large regions are generally more heterogenous and their inner variances are difficult to be captured; (2) the pattern of housing prices in outskirts is complicated, e.g., some outskirts can be economically undeveloped, while some can be exclusive areas with luxury residences.

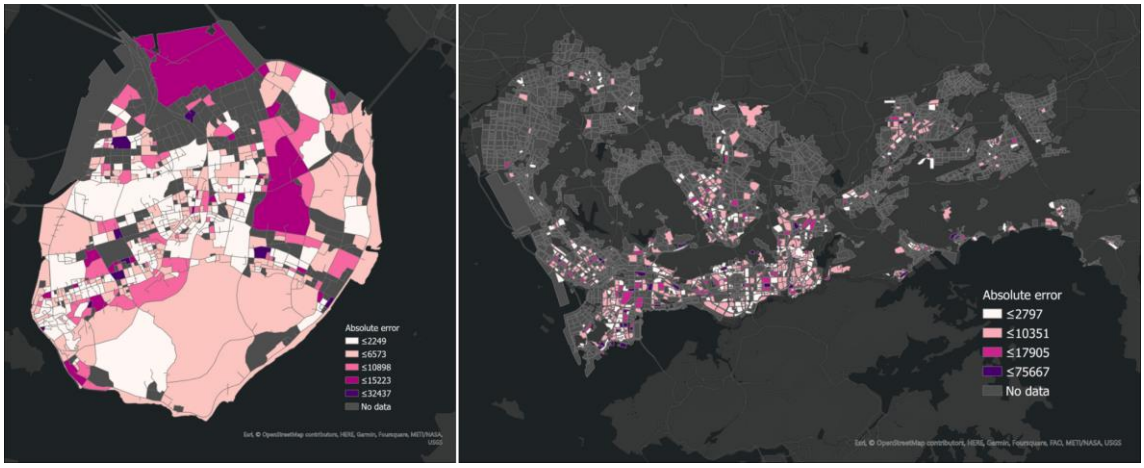

Figure 5. Spatial distribution of absolute estimation errors for housing price.

5 Discussion

Through the extensive experiments, it is encouraging to witness that HGI consistently prevails in all three downstream tasks. As the three tasks have been commonly acknowledged to be important in urban studies and spatial information science, we believe they confirm that HGI can produce more meaningful and expressive region representations than the baseline models. In fact, such results are unsurprising, as previous practices (the baselines) of representing regions with POIs mainly concentrated on modeling POI category semantics (e.g., co-occurrence patterns) and frequencies. In this paper, we take POI category embeddings as only a starting point, and further model the uniqueness of each individual POI through its spatial context, multi-faceted influence from POIs to regions, and the similarities between neighboring regions. In addition, the unsupervised objective of HGI facilitates that the information from POIs (local level) and the city (global level) flows to the regional level, so as to make region representations informative and meaningful. In addition, compared to the studies using graph neural networks for traffic or human movement analyses (cf. Section 2.2), HGI does not rely on the human behaviors exhibited in human mobility datasets, and generates meaningful region embeddings through deeply mining the information from only static POIs, which are generally easier to be collected.

The unsupervised nature of HGI unfolds many opportunities of using its region representations in various downstream tasks that have correlations with POIs. The learned region embeddings can be mainly used in two manners. First, for a particular target task such as urban vibrancy assessment, if ground truth is partially available (e.g., through field survey), one could use the pre-trained region embeddings to train a simple classifier or regression model, in order to derive the results in unknown areas. Using pre-trained region representations in different tasks can alleviate the training burden with a simple classifier or regressor. If the available ground truth is sparse (which is

often true), developing a fully supervised model with enormous parameters can be suboptimal. Second, for the tasks with no ground truth data such as similar region search (Liu et al., 2018), the HGI region embeddings can readily be used to measure similarities or uncover clusters. In principle, the regions with similar embeddings are also likely close in terms of their socioeconomic properties.

Inspired by the developments of language models, our HGI approach is able to be further strengthened by fine-tuning if some ground truth data is available for a downstream task. After training HGI in a fully unsupervised manner, we can obtain a trained encoder that takes POIs as input, and outputs region embeddings. Such an encoder is a sequential combination of the trained POI GCN ϕ_p, the aggregation function $\mathrm{AGG}_{poi-region}$, and the regional GCN ϕ_r, and all the components are equipped with already optimized parameters trained with the HGI unsupervised learning objectives. Instead of directly taking region representations, we can also fine-tune such an encoder using ground truth data in a supervised fashion, e.g., linking an MLP to the encoder to estimate urban functional distributions. In this manner, regions representations can probably be further enhanced by the supervision signal even with few ground truth data. We believe the best practices of fine-tuning HGI are worth of further investigation.

In this paper, we learn region representations based on POIs. Even though HGI can largely outperform the baseline methods, it cannot go beyond POIs. This means that our goal is to make the best of POIs and generate more meaningful representations based on such information, but we also recognize that the information in POIs is limited in nature. This also explains that the best R^2 for housing price estimation is about 0.4, as POIs can only account for part of the variations of housing price. In this regard, we believe that HGI region representations can be fused with information (representations)

from other modalities, e.g., remote sensing and human mobility data, to yield more satisfactory results. Previous studies utilizing POIs in conjunction with data from other modalities mainly used simple features from POIs or topic models, e.g., Su et al., (2021); Zhang et al., (2017). We believe that fusing HGI region representations based on POIs could considerably lift the performance for various downstream tasks, but how and when to fuse representations from different modalities also need further investigation.

6 Conclusions

In this paper, we present a novel approach HGI for learning meaningful and expressive region representations based on POIs in a fully unsupervised manner, i.e., each region is transformed to a low-dimensional vector embedding that can be used for various downstream tasks. The proposed HGI approach considers POI category semantics, the uniqueness of each single POI, multi-faceted influence from POIs to regions in the aggregation process, and similarities between adjacent regions. The model is trained with an unsupervised objective based on maximizing the mutual information among the POI – region – city hierarchy, which facilitates the information from local (POIs) and global (city) scales flowing to the learned region representations, so as to make them meaningful and effective. We conduct comprehensive experiments on three downstream tasks, i.e., estimating urban functional distributions, population density, and housing price, to verify the effectiveness of the proposed approach. The results indicate that the region representations produced by HGI considerably outperform the baseline models in all three tasks. Therefore, this confirms that HGI produces meaningful region representations that can be used for various downstream tasks in urban studies. In the future, we anticipate to further investigate fusing HGI region embeddings with information (representations) from other modalities, e.g., remote sensing data, and also

the fine-tuning techniques of HGI.

References

1. Andrade, R., Alves, A., Bento, C., 2020. POI Mining for Land Use Classification: A Case Study. ISPRS International Journal of Geo-Information 9, 493.
2. Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35, 1798-1828.
3. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine 34, 18-42.
4. Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., Singleton, A., 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. Computers, Environment and Urban Systems 85, 101539.
5. Cao, K., Diao, M., Wu, B., 2019. A big data–based geographically weighted regression model for public housing prices: A case study in Singapore. Annals of the American Association of Geographers 109, 173-186.
6. Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing 163, 82-97.
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, International conference on machine learning. PMLR, 1597-1607.
8. Cheng, Z., Wang, J., Ge, Y., 2022. Mapping monthly population distribution and variation at 1-km resolution across China. International Journal of Geographical Information Science 36, 1166-1184.
9. Dai, S., Wang, J., Huang, C., Yu, Y., Dong, J., 2021. Temporal Multi-view Graph Convolutional Networks for Citywide Traffic Volume Inference, 2021 IEEE International Conference on Data Mining (ICDM). IEEE, 1042-1047.

10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics, Minneapolis, Minnesota, 4171-4186.
11. Diao, Z., Wang, X., Zhang, D., Liu, Y., Xie, K., He, S., 2019. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting, Proceedings of the AAAI conference on artificial intelligence, 890-897.
12. Du, J., Zhang, Y., Wang, P., Leopold, J., Fu, Y., 2019a. Beyond geo-first law: Learning spatial representations via integrated autocorrelations and complementarity, 2019 IEEE International Conference on Data Mining (ICDM). IEEE, pp. 160-169.
13. Du, S., Du, S., Liu, B., Zhang, X., 2019b. Context-enabled extraction of large-scale urban functional zones from very-high-resolution images: A multiscale segmentation approach. Remote Sensing 11, 1902.
14. Du, S., Du, S., Liu, B., Zhang, X., Zheng, Z., 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. GIScience & Remote Sensing 57, 411-430.
15. Fu, Y., Wang, P., Du, J., Wu, L., Li, X., 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations, Proceedings of the AAAI Conference on Artificial Intelligence, 906-913.
16. Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Transactions in GIS 21, 446-467.
17. Geng, X., 2016. Label distribution learning. IEEE Transactions on Knowledge and Data Engineering 28, 1734-1748.
18. Grover, A., Leskovec, J., 2016. node2vec: Scalable feature learning for networks, Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 855-864.
19. He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, 1026-1034.

20. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2019. Learning deep representations by mutual information estimation and maximization, International Conference on Learning Representations.
21. Huang, W., Cui, L., Chen, M., Zhang, D., Yao, Y., 2022. Estimating urban functional distributions with semantics preserved POI embedding. International Journal of Geographical Information Science 36, 1905-1930.
22. Kipf, T.N., Welling, M., 2016. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308.
23. Kipf, T.N. and Welling, M., 2017. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations.
24. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W., 2019. Set transformer: A framework for attention-based permutation-invariant neural networks, International Conference on Machine Learning. PMLR, 3744-3753.
25. Liu, Y., Zhao, K., Cong, G., 2018. Efficient similar region search with deep metric learning, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1850-1859.
26. Luo, Y., Liu, Q., Liu, Z., 2021. Stan: Spatio-temporal attention network for next location recommendation, Proceedings of the Web Conference 2021, 2177-2185.
27. 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.
28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26.
29. 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.
30. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.

31. Robinson, J.D., Chuang, C.-Y., Sra, S., Jegelka, S., 2020. Contrastive Learning with Hard Negative Samples, International Conference on Learning Representations.
32. Shang, S., Du, S., Du, S., Zhu, S., 2020. Estimating building-scale population using multi-source spatial data. Cities, 103002.
33. Su, Y., Zhong, Y., Zhu, Q., Zhao, J., 2021. Urban scene understanding based on semantic and socioeconomic features: From high-resolution remote sensing imagery to multi-source geographic datasets. ISPRS Journal of Photogrammetry and Remote Sensing 179, 50-65.
34. Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J., Li, Q., 2020. Portraying the spatial dynamics of urban vibrancy using multisource urban big data. Computers, Environment and Urban Systems 80, 101428.
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, Advances in Neural Information Processing Systems, 5998--6008.
36. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph Attention Networks. ArXiv abs/1710.10903.
37. Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2019. Deep Graph Infomax. International Conference on Learning Representations.
38. Wang, T., Isola, P., 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, International Conference on Machine Learning. PMLR, 9929-9939.
39. Wang, Z., Li, H., Rajagopal, R., 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding, Proceedings of the AAAI Conference on Artificial Intelligence, 1013-1020.
40. Wang, Z., Zhu, Y., Liu, H., Wang, C., 2022. Learning Graph-based Disentangled Representations for Next POI Recommendation, Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1154-1163.
41. Xu, Y., Jin, S., Chen, Z., Xie, X., Hu, S., Xie, Z., 2022a. Application of a graph convolutional network with visual and semantic features to classify urban scenes. International Journal of Geographical Information Science, 1-26.

42. Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022b. 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.
43. Yan, B., Janowicz, K., Mai, G., 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.
44. 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.
45. Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017a. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. International Journal of Geographical Information Science 31, 825-848.
46. Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., Zhang, Y., 2017b. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. International Journal of Geographical Information Science 31, 1220-1244.
47. Ye, T., Zhao, N., Yang, X., Ouyang, Z., Liu, X., Chen, Q., Hu, K., Yue, W., Qi, J., Li, Z., 2019. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. Science of the total environment 658, 936-946.
48. Yuan, J., Zheng, Y. and Xie, X., 2012, August. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (186-194).
49. Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., Gu, C., 2019. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Computers, environment and urban systems 74, 1-12.
50. Zhang, J., Li, X., Yao, Y., Hong, Y., He, J., Jiang, Z., Sun, J., 2021. The Traj2Vec model to quantify residents' spatial trajectories and estimate the proportions of urban land-use types. International Journal of Geographical Information Science 35, 193-211.

51. Zhang, X., Du, S., Wang, Q., 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS Journal of Photogrammetry and Remote Sensing 132, 170-184.
52. Zhang, X., Du, S., Wang, Q., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sensing of Environment 212, 231-248.
53. Zhang, X., Huang, C., Xu, Y., Xia, L., Dai, P., Bo, L., Zhang, J., Zheng, Y., 2021. Traffic flow forecasting with spatial-temporal graph diffusion network, Proceedings of the AAAI conference on artificial intelligence, 15008-15015.