Urban Region Representation Learning with OpenStreetMap Building Footprints

Yi Li
liyi0067@e.ntu.edu.sg
Nanyang Technological University
Singapore

Weiming Huang
weiming.huang@ntu.edu.sg
Nanyang Technological University
Singapore

Gao Cong
gaocong@ntu.edu.sg
Nanyang Technological University
Singapore

Hao Wang
cshaowang@gmail.com
Nanyang Technological University
Singapore

Zheng Wang
zheng011@e.ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

The prosperity of crowdsourcing geospatial data provides increasing opportunities to understand our cities. In particular, OpenStreetMap (OSM) has become a prominent vault of geospatial data on the Web. In this context, learning urban region representations from OSM data, which is unexplored in previous work, could be profitable for various downstream tasks. In this work, we utilize OSM buildings (footprints) complemented with points of interest (POIs) to learn region representations, as buildings' shapes, spatial distributions, and properties have tight linkages to different urban functions. However, appealing as it seems, urban buildings often exhibit complex patterns to form dense or sparse areas, which brings significant challenges for unsupervised feature extraction. To address the challenges, we propose RegionDCL[1], an unsupervised framework to deeply mine urban buildings. In a nutshell, we leverage random points generated by Poisson Disk Sampling to tackle data-sparse areas and utilize triplet loss with a novel adaptive margin to preserve inter-region correlations. Furthermore, we train our model with group-level and region-level contrastive learning, making it adaptive to varying region partitions. Extensive experiments in two global cities demonstrate that RegionDCL consistently outperforms the state-of-the-art counterparts across different region partitions, and outputs effective representations for inferring urban land use and population density.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Geographic information systems*; • **Computing methodologies** → *Knowledge representation and reasoning*.

KEYWORDS

Geospatial data mining, OpenStreetMap, urban regions, representation learning

KDD '23, August 6–10, 2023, Long Beach, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599538

ACM Reference Format:
Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban Region Representation Learning with OpenStreetMap Building Footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599538

1 INTRODUCTION

Urbanization has become a defining feature of modern society, and understanding and characterizing our cities is essential for effective urban planning and management [6]. Traditional approaches for investigating the properties of cities mainly relied on field survey, which is often labor-intensive and time-consuming [22]. With the proliferation of urban big data, urban computing has emerged as a powerful tool for addressing critical urban issues and facilitating valuable urban applications [47]. By applying machine learning techniques to various types of data, such as vehicle trajectories and points of interest (POIs), we could gain new insights and develop innovative solutions to urban problems. However, while these approaches have yielded promising results, they often focus on a single task (e.g., economic growth prediction [13], air quality analysis [48]) and rely on domain expertise for supervision.

Recently, urban region representation learning emerges as a popular practice in urban computing, which transforms urban regions into vector representations (embeddings) that can be used for various downstream tasks such as identifying land use, estimating population, as well as predicting air quality and economic growth [32]. This type of approach has two key characteristics, unsupervised (or self-supervised) learning and multitasking. The unsupervised setting reduces the dependence on large amounts of labeled data, and the task-agnostic paradigm is well-suited to urban problems, as many socioeconomic factors, such as land use inference and population density estimation, share commonalities [18]. This makes it desirable to learn general-purpose region embeddings to facilitate urban sensing applications.

An important consideration in urban region representation learning is the types of input data. In this regard, several commonly-used data sources include human mobility data [9, 14, 31, 33, 38, 44, 46], POIs[9, 14, 31, 32, 38, 44], street view images [32], and geo-tagged social media [42]. Among them, human mobility data has been

[1] Code released at https://github.com/LightChaser666/RegionDCL.

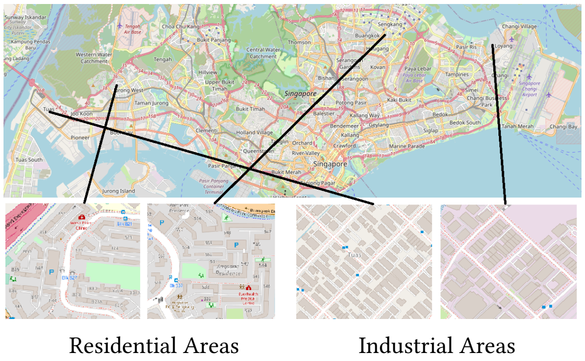


Figure 1: Typical residential and industrial areas in Singapore. Buildings in industrial areas are mainly rectangles and arranged in grids, while those in residential areas have twisted shapes and organic arrangements. The similarity in building shapes and arrangements can be observed in regions with the same land use, regardless of their spatial distances.

extensively investigated [9, 14, 31, 33, 38, 44, 46], and various techniques are developed for building mobility graphs and understanding human transitions between regions. However, human mobility data requires both devices to collect and is only feasible for limited areas and user groups. This makes it difficult to apply these methods to cities where human mobility data is not available. In addition, other data sources like street-view images entail quality, accessibility, and coverage issues for urban sensing [24]. Therefore, it is important to explore alternative data sources that have better availability in a wider range of cities.

The increasing availability of open geospatial data offers a promising solution. OpenStreetMap (OSM) is a representative crowdsourcing platform that provides increasingly comprehensive coverage of geospatial data, including building footprints, POIs, and road networks. Buildings, in particular, are readily accessible from OSM and contain valuable geometric and semantic information, especially in urban areas [1]. Utilizing such data offers two main advantages: (1) *Data Effectiveness*, as buildings are the main components of urban regions, and modern building design often follows the principle of "form follows function", meaning that building shapes and their spatial distributions are indicative of their intended use [8], e.g. commercial vs residential buildings (See Fig. 1). Furthermore, buildings are tightly linked to other factors such as population density [36, 45], and (2) *Data Availability*, as buildings are readily available from OSM. Therefore, in this study, we propose to utilize OSM building data (footprints) for urban region representation learning.

Despite its potentials, utilizing buildings for urban region representation learning poses several prominent challenges:

- **Representation Learning**. Buildings come with complex polygonal geometries, which cannot be readily encoded by previous point-oriented methods [34]. Additionally, it is crucial to capture the spatial distribution and intercorrelations of buildings, which are essential in characterizing building groups [26], but often overlooked in previous work. Furthermore, while many previous studies focus on preserving inter-region spatial proximity [14, 31, 32], we posit that maintaining building similarities across diverse regions is also crucial. As observed in Fig 1, buildings located in distant regions can exhibit similar shapes and spatial distributions. Therefore, it is important to preserve the similarities of such patterns in the learned embeddings.
- **Data-sparse Areas**. Urban buildings are often unevenly distributed in a city, resulting in many data-sparse (i.e., empty or unmapped) areas. Previous work generally treats regions as "bags of objects", neglecting the intrinsic trait of regions, which contain both developed and empty (or unmapped) parts. This lack of representation of empty areas downgrades the discrimination power of the learned embeddings, e.g, a vastly unutilized region and a small residential area both having a few buildings would likely be mistakenly regarded as similar. However, representing empty parts of regions is a non-trivial task, as these areas often possess various shapes and their characteristics are often correlated with surrounding buildings, which cannot be captured by traditional hand-crafted features (e.g., average distance between nearest neighbors [5]).

In addition, previous work has been highly dependent on one specific region partition chosen for the task at hand. However, in real-world scenarios, different downstream tasks may require varying region partitions. For example, U.S. demographics estimation is commonly done in Census Tracts [38, 44], while traffic crash estimation may require Traffic Analysis Zones [23]. Empirically, models designed for one specific region partition may not perform well when applied to a different partition, as demonstrated by the experiment results in Section 5.5.1. The inability to adapt to varying region partitions diminishes the generality of derived embeddings. Thus, it is desirable to design a training strategy that is suitable for different region partitions.

To address the aforementioned challenges, we propose RegionDCL, a **D**ual **C**ontrastive **L**earning framework that leverages OSM data to derive general urban region representations for various downstream tasks. To tackle the challenges in representation learning, we employ Transformer Encoder with distance-biased self-attention to model the complex spatial distribution and intercorrelations of buildings. In addition, we formulate a triplet loss with an adaptive margin to preserve spatial proximity and building similarities simultaneously. To handle data-sparse areas, we use Poisson Disk Sampling [2] to generate random points that explicitly mark the existence of empty spaces and interact with surrounding objects, thus significantly enhancing the embedding quality for data-sparse regions. Furthermore, we use OSM road networks to partition the city into small building groups as intermediate units for varying region partitions, and perform a dual contrastive learning strategy at both the levels of building groups and regions. This improves our model's adaptability to different partition schemes and reduces the need for re-training, ultimately enhancing the generalizability of derived embeddings.

In summary, the key contributions of this paper are as follows:

- We propose a novel framework RegionDCL that generates meaningful and effective urban region representations for various downstream tasks using OSM building footprints and POIs publicly available. This is the first work to utilize building footprints for region representation learning, to the best of our knowledge.

- We propose to annotate data-sparse (or even empty) areas with random points generated by Poisson Disk Sampling, improving the representation quality for data-sparse regions. We also propose a novel dual contrastive learning that preserves spatial proximity and building group similarity, improving the model's adaptability to multiple downstream region partitions.
- Extensive experiments on Singapore and New York City data demonstrate the superiority of the proposed framework for land use inference and population density estimation tasks. Case studies and visualization results further show RegionDCL's effectiveness in data-sparse areas and its strong adaptability to variant region partitions.

2 RELATED WORK

The field of urban region analysis has predominantly utilized task-specific supervised learning. Yuan et al. [40] utilized human mobility and POIs to identify urban functional zones. Naik et al. [19] predicted streetscape safety through street view images. Zheng et al. [48] estimated air quality through traffic and meteorological data. Yang et al. [36] encoded urban buildings for functional region classification. However, these methods are limited by their need for domain expertise and their inability to adapt to new tasks.

Several studies have attempted to overcome these limitations by focusing on learning general urban region representations in an unsupervised manner. For example, Zhai et al. [41] learn region representations by modeling the co-occurence of POI categories. Niu and Silva. [20] further take the spatial proximity of nearest neighbors into account. However, these methods focus on POI semantic preservation and are limited in modeling spatial distributions.

Recently, Wang and Li [31] proposed a mobility graph with regions as nodes and trajectories as edges, thereby uncovering the inter-region correlations. Fu et al. [9] further integrated POIs into the mobility graph with graph auto-encoders. Zhang et al. [46] integrated POIs and Check-in data into the mobility graph with collective adversarial training. Zhang et al. [43] utilize POIs and mobility data to construct different region views for contrastive learning. Porter et al. [14] combined satellite images, POIs, and the mobility graph via Graph Convolutional Networks. Zhang et al. [44] introduced an attention mechanism to capture cross-modality feature associations in the mobility graph. Wu et al. [33] fused per-hour mobility graphs in different periods to enhance urban region representations.

While these methods achieved promising results, they rely heavily on human mobility data to model inter-region associations, which can be difficult to access in practice. Additionally, they do not account for information from data-sparse areas and are confined to specific region partitions. Our proposed method, RegionDCL, addresses these limitations by leveraging widely-available OSM data and preserving building similarities and spatial distributions, which have not been previously investigated.

3 PROBLEM STATEMENT

We hereby present some definitions and give the problem statement.

Definition 3.1 (Building Footprint). A building footprint b refers to a 2-D polygonal area delineated by the exterior boundary of the building, where each vertex on the polygon has a spatial location (i.e., longitude and latitude). Each building may have a type tag (e.g., sports center). For simplicity, we use the term *building* in the rest of the paper.

In OSM data, only some of the buildings (e.g., 30% in Singapore) are annotated with type tags. For the remaining buildings without type tags, we manually assign the tag 'unknown'.

Definition 3.2 (Building Group). A building group refers to the collection of buildings in a defined spatial area. To obtain these building groups, we utilize road networks to partition the city into distinct sections, also known as Traffic Analysis Zones. The collection contains tightly connected buildings with relatively homogeneous features (e.g., shapes) [26].

Definition 3.3 (Urban Region). Urban regions $\mathcal{U}$ refer to a set of disjoint city areas, usually obtained through a certain partition approach (e.g., census tracts). Each urban region u may include multiple building groups.

Problem Statement. (*Urban Region Representation Learning*) Given a set of urban regions $\mathcal{U} = \{u_1, u_2, ...\}$, the goal of urban region representation learning is to learn a mapping function that generates a vector representation $z_i \in \mathbb{R}^k$ for each region u_i in the Euclidean space, where k is the uniform dimension for all $u_i \in \mathcal{U}$.

4 METHODOLOGY

We present our framework RegionDCL. As shown in Figure 2, it consists of three components: (1) *Feature Pre-processing*, which transforms individual geospatial objects into feature vectors and annotates empty space with random points using Poisson Disk Sampling, (2) *Building Group Encoding*, which models spatial distributions and intercorrelations of buildings with a distance-biased Transformer encoder, and (3) *Dual Contrastive Learning*, which organizes buildings into groups and preserves both building group similarity and spatial proximity to derive region representations.

4.1 Feature Pre-processing

We first present how to pre-process individual OSM buildings and POI into compact embeddings as the inputs of RegionDCL. We also demonstrate how to deal with data-sparse areas in a city.

4.1.1 Building Features. OSM buildings comprise 2-D polygonal building shapes, spatial locations, and type tags, with the 2-D shapes playing a crucial role in characterizing urban areas. For instance, Singapore's buildings in industrial areas tend to be rectangular, whereas those in residential areas are often twisted (See Fig. 1).

Despite various models for 2-D shape encoding, Convolutional Neural Network (CNN) remains one of the most widely used and effective approaches[16]. To leverage this, we employ an ImageNet pre-trained ResNet-18 [11] to encode building polygons into visual features. The visual features are made more invariant to building size and rotation by first aligning the longest edge of each building's bounding box with the horizontal axis and then resizing and rasterizing the building polygon into a 224×224 image, following the procedure outlined in [34]. In order to retain the information on the size and rotation of each building, the resulting vectors are concatenated with three scalar values (size in square meters, $cos\alpha$, $sin\alpha$) that represent the building's size and rotation angle.

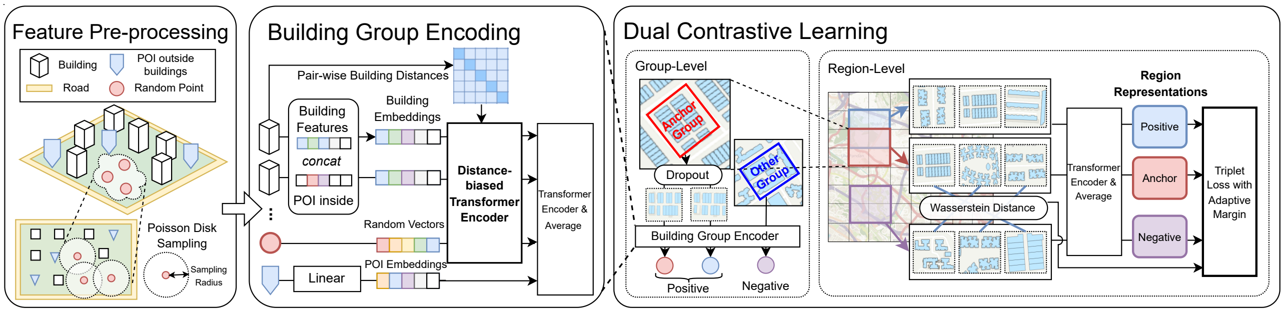


Figure 2: An overview of RegionDCL. Its backbone is dual contrastive learning with a triplet loss at both building group and region levels, where the region-level encoder is a Transformer encoder with average-pooling, and the group-level encoder is a building group encoder that consists of two Transformer encoders (i.e., Distance-biased Transformer encoder and average-pooling Transformer encoder).

While the visual appearance of a building provides some indication of its function, this information is usually not definitive. For instance, two distinct city facilities like a parking lot and a factory can both have rectangular shapes on the map. To address this limitation, we propose to supplement building features with POIs, as POIs are readily available in OSM and offer clear semantic information about the human activities performed within the building [36]. We represent each POI category as a one-hot vector and sum up the one-hot vectors of all POIs within a building. We also use one-hot vectors to represent the type tag of buildings. The extracted visual features from CNN, one-hot vector of POI category, and one-hot vector of building-type information are concatenated to form the final building features (denoted as b).

It is worth noting that certain POIs, such as small infrastructures, are often located outside buildings in the OSM data. We similarly represent them as one-hot vectors (denoted as t), but treat them separately from buildings to allow for a clearer distinction between buildings and their surrounding environment.

4.1.2 Random Points. Geospatial objects exhibit an uneven distribution across cities, resulting in the presence of data-sparse or even empty areas that have been generally neglected in previous studies. These empty areas can have distinct shapes and sizes, and be surrounded by different geospatial objects, which hold important implications for various downstream tasks. For example, a large, irregularly shaped empty area located at the edge of a city is likely to be a natural park with low population density, while a small, polygonal empty area surrounded by bus stops is likely to indicate an (unmapped) residential area with high population density. While traditional hand-crafted features, e.g. the average distance between nearest neighbors [5] and Ripley's K function [25], can provide some insights into spatial sparsity, they have limited capacity in capturing the shapes and environments of the empty areas.

To tackle the challenge, we propose to incorporate random points within these empty areas to annotate their existence. Our approach leverages the concept of Poisson Disk Sampling (PDS) [2], which allows for the uniform generation of random points within a 2-D plane space (i.e., the empty area in our case). As depicted in Fig. 2, PDS generates random points that are spatially compact but maintain a minimum user-specified distance (sampling radius r) between each pair of points. This process yields evenly distributed "sampling disks" that cover the entire plane space. The random points thus serve as annotators that provide a finer representation of empty areas, capturing their variant shapes and sizes. Furthermore, we assign all random points with a unified feature vector (denoted as s) and model their relationship with surrounding buildings with an attention mechanism (as detailed in Section 4.2). With PDS and the attention mechanism, we are able to preserve the existence as well as the contextual information of empty areas.

4.2 Building Group Encoding

After obtaining individual embeddings of buildings, POI outside buildings, and random points, we propose to encode them in groups considering their spatial distributions and intercorrelations.

4.2.1 Pair-wise building distances. The spatial distance between geospatial objects plays a crucial role in characterizing the structural information of urban regions [9]. This is especially true for buildings, where their arrangement varies depending on their intended use. For instance, industrial buildings in Singapore are often arranged in a grid-like fashion, while residential buildings are arranged more organically (as demonstrated in Fig. 1). Unlike previous work that ignores in-region spatial distances or only takes into account the nearest spatial distances [35, 36], our approach preserves all relative distances within each building group by constructing a pair-wise distance matrix D. Specifically, let $H = [b_1^T, ...b_j^T, s_1^T, ..., s_l^T]^T \in \mathbb{R}^{n \times d}$ denote the features of buildings and random points within a building group, we calculate the distance matrix D between each pair of buildings and random points via Haversine function:

$$D_{ij} = 2E \cdot \arcsin(\sqrt{\sin^2(\frac{\phi_i - \phi_j}{2}) + \cos(\phi_i)\cos(\phi_j)\sin^2(\frac{\theta_i - \theta_j}{2})}))\quad (1)$$

where D_{ij} denotes the ij-th element of matrix D, and E denotes the approximate earth's radius. ϕ_i and ϕ_j are the latitudes of the i-th and the j-th building (or random point), and θ_i and θ_j are their longitudes. The pair-wise distance matrix fully describes the spatial arrangement of geospatial objects (i.e., buildings and random points) within a building group.

4.2.2 Distance-biased Transformer encoder. We propose to use the Transformer encoder [29] to encode building groups. Unlike graph convolution layers that focus on neighbor interactions, the Transformer encoder can efficiently capture global interactions among all objects within each building group. Particularly, the self-attention mechanism can capture the contextual information of empty areas by aggregating building features to random points.

The Transformer encoder contains a multi-head self-attention module and a position-wise feed-forward network, where the multi-head self-attention is formulated as:

$$Q = HW_Q,\ K = HW_K,\ V = HW_V \tag{2}$$

$$Att_\beta(H) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \tag{3}$$

$$Mul(H) = \text{Concatenate}(Attn_1(H), ..., Attn_\beta(H)) \tag{4}$$

where W_Q, W_K, W_V are projection matrices and β is the number of attention heads. However, the above self-attention module (i.e., Eq. (3)) ignores buildings' spatial position information. Intuitively, the more distant objects should have fewer interactions (i.e., smaller attention weights) with each other. Inspired by the recent success of Graphormer [39], we add a normalized pair-wise distance matrix to self-attention as a bias term:

$$Att_\beta(H) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + \lambda\hat{D})V \tag{5}$$

$$\hat{D}_{ij} = \log(\frac{1 + \text{maxPooling}(D)^{1.5}}{1 + D_{ij}^{1.5}}) \tag{6}$$

where maxPooling(D) denotes the max value in the pair-wise distance matrix D, and λ is a trainable value. We follow Calafiore et al. [3] and Huang et al. [12] to set the exponential factor as 1.5. In such a method, the added bias term $\lambda\hat{D}$ carries the correlation of objects with respect to their relative distances to bring spatial position information to self-attention.

For those outside POIs, we feed them into a linear layer. The outputs, together with the outputs of the above distance-biased Transformer encoder, are subsequently fed into a vanilla Transformer encoder with average-pooling to generate an embedding vector for the building group.

4.3 Dual Contrastive Learning

The partitioning of urban regions is a non-trivial aspect of downstream tasks, yet it has been largely overlooked in previous studies. Existing methods have primarily focused on a single partition scheme, such as grids [14], census tracts [38], or traffic analysis zones [12]. However, this narrow focus can lead to inconsistent representation qualities, as methods that perform well under one partition scheme may not perform well under a different scheme, as shown in Section 5.5.1. To mitigate the shortcoming, we propose to use building groups as a glue of different region partitions, which is achieved through a dual contrastive learning approach at the group level and the region level.

4.3.1 Group-level Contrastive Learning. We propose encoding building groups derived from OSM's fine-grained road networks, including primary, secondary, tertiary, and footway roads. Compared to regions in downstream tasks that are typically large and diverse, building groups are more fine-grained and homogeneous in terms of urban functions, as they are the basic units in urban planning [37]. Encoding building groups into intermediate embeddings enables us to capture the finer details of urban structures. However, a novel training method is necessary to learn effective building group embeddings as the common technique of preserving cross-region human transitions in previous work [9, 31, 33, 38, 44, 46] can not be applied to our settings where human mobility data is not available.

Inspired by the recent success of contrastive learning [4], we propose to extract features with a self-supervised task, i.e. instance discrimination. The intuition is that building groups that differ in a small number of buildings should be similar, and the model is expected to discriminate the most similar building group from random sampled candidates. Specifically, given a training batch, every sample (i.e., building group) in this batch is selected as an anchor, denoted as P_i. For each anchor sample within the batch, we randomly remove/dropout a small number of buildings inside to generate a positive sample, denoted as P_i^+. The other building groups within the batch are treated as negative samples. Our model is later trained with the anchors, positive samples, and negative samples using an InfoNCE loss [28]:

$$\mathcal{L}_{InfoNCE} = -\log(\frac{e^{\text{sim}(P_i,\ P_i^+)/\tau}}{\sum_{i=0}^{n} e^{\text{sim}(P_i,\ P_j)/\tau}}) \tag{7}$$

where n is batch size, sim$(\cdot, \cdot)$ is a similarity measure function, $\tau = 0.05$ is a temperature parameter. As building group embeddings can be interpreted as a distribution of inside buildings, we apply KL-divergence as the similarity measure.

4.3.2 Region-level Contrastive Learning. Urban regions are often composed of multiple building groups, with some being more important in characterizing regions. Thus, we use a vanilla Transformer encoder with average-pooling to aggregate building group embeddings into region representations, where the self-attention weights capture the different importance of building groups.

Previous studies have shown that the effectiveness of region representations can be improved by preserving inter-region correlations [9]. Most existing work uncovers such correlations with human mobility data that directly reflect region connectivity. However, this becomes challenging with only OSM building and POI data. Leveraging the spatial proximity of the regions can be a straightforward strategy, as according to Tobler's first law of geography [27], near regions are more likely to be correlated than distant regions.

Inspired by this idea, we propose to preserve spatial proximity of regions through contrastive learning at the region level. For each region (i.e., anchor region) in a city, we select a neighboring region as the positive region and randomly choose another region as the negative region. We use a triplet Loss to preserve inter-region spatial proximity in the embedding space:

$$\mathcal{L}_{triplet} = max(||z_a - z_p|| - ||z_a - z_n|| + m,\ 0) \tag{8}$$

where $|| \cdot ||$ represents the L1 distance. z_a, z_p, and z_n denote the anchor, positive and negative region representations, respectively, and m is a hyper-parameter that controls the distance between the positive and negative embeddings in the embedding space.

While preserving the spatial proximity can be empirically effective [32], we posit that distant regions sometimes contain similar building groups. In this case, a fixed margin m would indiscriminately push all the embeddings of distant regions further apart, causing the loss of the similarities between distant regions.

To resolve this issue, we propose to dynamically adjust the margin value based on the similarity of building groups they contain. We avoid using the naïve average similarity as it can be extremum-sensitive (i.e., a group extremely different from other groups can significantly affect the average similarity.), Instead, we propose to only focus on those similar building group pairs, which is achieved by Wasserstein distance. Specifically, suppose u_a, u_b are two randomly-picked regions and $u_a \neq u_b$, we first calculate the matching cost matrix $C \in \mathbb{R}^{m \times n}$ with JS divergence:

$$C_{ij} = JS(a_i||b_j) = \frac{1}{2}KL(a_i||\frac{a_i + b_j}{2}) + \frac{1}{2}KL(b_j||\frac{a_i + b_j}{2}) \qquad (9)$$

where $a \in u_a$ and $b \in u_b$ are the building group embeddings within the two regions. We empirically use JS divergence to compute the matching cost for numerical stability. Then we utilize Wasserstein distance to obtain the minimum matching cost of the building groups between two regions:

$$W = \min_{\pi} \sum_{i=1}^{m} \sum_{i=1}^{n} \pi_{ij} C_{ij}, \quad \text{s.t.} \sum_{i=1}^{m} \pi_{ij} = 1 \text{ and } \sum_{j=1}^{n} \pi_{ij} = 1 \qquad (10)$$

which is an optimal transport problem that can be solved in polynomial time. Finally, we replace the fixed m in Eq. (8):

$$\hat{\mathcal{L}}_{triplet} = \max(||z_a - z_p|| - ||z_a - z_b|| + \lambda \cdot W,\ 0) \qquad (11)$$

where λ is a hyper-parameter to scale the Wasserstein Distance. Now, the new margin $\lambda \cdot W$ will automatically get smaller when two regions have some building groups in common and get larger when two regions are vastly different.

4.3.3 Training on Shifted Windows. We empirically find that neighboring regions may also be extremely different, which leads to unstable model performance. To address the problem, we propose to use overlapped regions as positive samples since they always share common building groups and are more likely to be similar.

Specifically, as shown in Fig. 2, we use a square window with a given size (e.g., 2km×2km) and shift the window horizontally and vertically to generate training regions. Then, the overlapped regions are selected as positive samples, while the non-overlapping windows are regarded as negative pairs. As real-world downstream region partitions are often in different sizes, we also use shifted windows with multiple sizes to further boost the model's adaptability. By training on shifted windows, our framework is decoupled from downstream region partitions compared to previous work. In the implementation, we separately train the group-level and region-level contrastive learning for better efficiency.

5 EXPERIMENTS

We conduct experiments to evaluate the effectiveness of RegionDCL from the following aspects: (1) the performance of our derived representations in downstream tasks, (2) the usefulness of our model in data-sparse regions and different region partitions, (3) the quality of building group embeddings derived by RegionDCL, and (4) the effect of important hyper-parameters.

5.1 Experimental Settings

5.1.1 Datasets. We conduct experiments with OpenStreetMap data of Singapore and New York City. New York City data has a massive population and diverse urban zone functions and is commonly used in previous work [14, 32, 33, 38, 44]. Singapore has more data-sparse regions and a much higher population density than New York City. The statistics of the used cities are shown in Table 1.

Table 1: Dataset Statistics

City	Buildings	POIs	Building Groups	Regions
Singapore	109,877	17,088	5,824	304
New York City	1,081,256	41,963	29,008	2324

For evaluation benchmark on two downstream tasks, we collect the *land use data* of Singapore and New York City from Singapore Master Plan 2019 and NYC MapPLUTO, respectively. We collect *population density data* from WorldPop for both cities. In addition, we utilize Singapore Subzones and NYC Census Tracts as region partitions in evaluation unless otherwise specified. Please refer to Appendix A for the online resources for all datasets.

5.1.2 Baselines. We compare with representative baseline models, include five models that can take the same input data as RegionDCL, three model variants and ablations, as well as two state-of-the-art region embedding techniques based on human mobility.

(1) **Baselines with the same input data as RegionDCL**.
- Place2Vec [41]: This method preserves the co-occurrence of POI categories to learn region representations. We regard building type tags as POI categories.
- Doc2Vec [20]: This method models spatial objects as words and regions as documents with their spatial co-occurrence. We take the document embeddings as region representations.
- GAE [15]: This method encodes graph nodes with GCN and optimizes node features via feature reconstruction. We average the node embeddings as region representations.
- DGI [30]: This method encodes graph nodes with GCN and then maximizes the mutual information between node embeddings and graph embeddings through contrastive learning. We take the graph embeddings as region representations.
- Urban2Vec [32]: This framework integrates two different modalities of geospatial data, and performs contrastive learning to learn region embeddings via triplet loss. We feed the same input data to Urban2Vec as to RegionDCL.

(2) **Model variants and ablations**, where we demonstrate the effectiveness of our model components.
- Transformer. We use vanilla Transformer [29] to encode buildings and POIs. We train it with group-level contrastive learning and average the derived building group embeddings as region representations to demonstrate the effectiveness of our building group encoding strategy.
- RegionDCL-no random. This is a variant of our model in which we remove the Poisson Disk Sampling part (Sec. 4.1.2).
- RegionDCL-fixed margin. This is also a variant of our model in which we remove the proposed adaptive margin (Sec. 4.3.2).

(3) **Baselines based on human mobility**. As most state-of-the-art region embedding techniques are based on human mobility, we compare RegionDCL with such methods to demonstrate that RegionDCL with readily accessible OSM data is able to outperform methods relying on human mobility data, which is often difficult to obtain.
   - MVURE [44]: This framework models the cross-modality interaction of multi-modality features (e.g., POIs and Check-ins) via graph attention networks.
   - MGFN [33]: This method learns region embeddings by clustering per-hour mobility graphs and applying self-attention.

For those graph-based methods, i.e., GAE and DGI, we construct triangle graphs with Delaunay Triangulation on buildings and POIs for each region, following [7]. Then, we use the building features and POI features wrapped in Section 4.1 as node features and initialize the edge weights with Eq. (1).

5.1.3 Implementation Details. In our experiments, all embedding dimensions are set to 64. For Place2Vec and Doc2Vec, we set window size=5, KNN sampling $k = 10$. For GAE and DGI, we use a 2-layer GCN following their suggestions. The triplet margin m in Urban2Vec and our framework are set to 10. For Transformer and our framework, the number of attention heads is set to 8. In our framework, we use a two-layer distanced-biased Transformer encoder followed by a 2-layer Transformer Encoder with average pooling. The r in Poisson Disk Sampling is set to 100 meters, and the shifted window sizes are 1000, 2000, and 3000 meters. The scaler λ in the proposed adaptive margin is set to 50.

5.2 Land Use Inference

In this task, we use the region representations derived by RegionDCL and baseline methods to infer urban functional distributions, i.e., the proportion of land use areas within each region. Following the settings in [21], we merge the fine-grained ground-truth annotations of land use into five major categories called *Residential*, *Industrial*, *Commercial*, *Open Space*, and *Others.*

5.2.1 Evaluation Metrics. The land use inference task is a label distribution learning problem [10, 12]. Evaluation metrics include L1 distance, KL-divergence, and Cosine similarities, which measure the similarity between estimated functional distributions and ground-truth labels. To infer land use, we utilize a 2-layer Multi-Layer Perceptron (MLP) with 512 hidden units and five output units. We randomly split the regions of each city into 60% training, 20% validation, and 20% test set (i.e., 5-fold cross-validation). The MLP is optimized by KL-divergence loss for 150 epochs, and the test result at the lowest validation loss is recorded. We run each algorithm 30 times and report the average value and standard deviation.

5.2.2 Results. Table 2 demonstrates the performance on land use inference in two cities, where the symbol ↓ indicates the smaller score, the better model performance; ↑ indicates opposite results. To rigorously evaluate algorithm effectiveness, we exclude the results of MVURE and MGFN in this table as they take human mobility data as input. The best results are bolded. From the table, we have the following observations:

- The performance of the model is significantly affected by the information it preserves. The GAE model that only preserves the input features with reconstruction loss is inferior to all baseline models. While methods that only preserve the spatial proximity of buildings and POIs (i.e., Place2Vec, Doc2Vec, and Urban2Vec) perform significantly better than GAE, they fail to fully capture the spatial distributions and similarities in the input features.
- Methods that extensively mine building feature similarities and spatial interactions (i.e., DGI, Transformer, and RegionDCL and its variants) exhibit significantly superiority over other baselines, which supports our design considerations of representation learning outlined in Section 4.2. Although Transformer does not explicitly account for spatial locations of buildings, it achieves competitive performance through contrastive learning on building groups, which is comparable to that of DGI that models spatial distances to neighboring objects. This highlights the effectiveness of our group-level contrastive learning strategy.
- The proposed RegionDCL outperforms all baselines by annotating empty areas, incorporating building spatial distributions, considering spatial proximity between regions, and preserving similarity among building groups. The comparison results between RegionDCL and its variant RegionDCL-no random indicate that filling empty areas with random points significantly enhances the overall embedding quality. The comparison results between RegionDCL and its variant RegionDCL-fixed margin demonstrate the effectiveness of our proposed adaptive margin in preserving building group similarities.
- RegionDCL demonstrates greater improvements in Singapore compared to New York City. Given the remarkably distinct building styles and more data-sparse areas in Singapore [26], these results further validate our design choices.

5.3 Population Density Estimation

In this task, we use the learned urban region representations to estimate the average population density in each region.

5.3.1 Evaluation Metrics. The population density estimation task is a regression problem. Evaluation metrics include absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). To infer population density, we randomly split the city into 80% training and 20% test regions and train a random forest regressor [17] on different data splits. We run each algorithm 30 times and report the average value and standard deviation.

5.3.2 Results. Experimental results are shown in Table 3. Unlike the land use inference task, methods that explicitly incorporate spatial proximity (i.e., DGI, Urban2Vec, and RegionDCL) significantly outperform the other methods. RegionDCL outperforms the best baseline in both cities (12.4% in Singapore and 5.9% in New York City in terms of R^2). The proposed random point sampling makes major contributions to the achieved improvements (6% in Singapore and 4% in New York City).

5.4 Comparison with Mobility-based Methods

We compare with two latest methods MVURE [44] and MGFN [33] to demonstrate our superiority over mobility-based methods. We collect the same official NYC Taxi Trip dataset (See Appendix A) with 10,906,859 pick-up and drop-off locations, and re-conduct

Table 2: Land Use Inference in Singapore and New York City

Models	Singapore			New York City		
	L1↓	KL↓	Cosine↑	L1↓	KL↓	Cosine↑
Urban2Vec	0.657±0.033	0.467±0.043	0.804±0.017	0.473±0.018	0.295±0.015	0.890±0.007
Place2Vec	0.645±0.039	0.451±0.047	0.812±0.018	0.518±0.016	0.308±0.012	0.878±0.005
Doc2Vec	0.679±0.050	0.469±0.058	0.789±0.027	0.506±0.015	0.299±0.016	0.885±0.008
GAE	0.759±0.040	0.547±0.051	0.765±0.022	0.589±0.011	0.365±0.011	0.855±0.007
DGI	0.598±0.029	0.372±0.032	0.846±0.012	0.433±0.009	0.237±0.012	0.907±0.005
Transformer	0.556±0.046	0.357±0.070	0.850±0.026	0.436±0.020	0.251±0.018	0.903±0.008
RegionDCL-no random	0.535±0.054	0.321±0.066	0.863±0.030	0.422±0.011	0.234±0.010	0.910±0.005
RegionDCL-fixed margin	0.515±0.042	0.303±0.040	0.872±0.020	0.426±0.011	0.248±0.018	0.905±0.008
RegionDCL	**0.498±0.038**	**0.294±0.047**	**0.879±0.021**	**0.418±0.010**	**0.229±0.008**	**0.912±0.004**

Table 3: Population Density Inference in Singapore and New York City

Models	Singapore			New York City		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Urban2Vec	6667.84±623.27	8737.27±902.41	0.303±0.119	5328.38±200.58	7410.42±261.89	0.522±0.028
Place2Vec	6952.34±713.30	9696.31±1239.65	0.171±0.121	8109.79±175.18	10228.61±261.43	0.096±0.043
Doc2Vec	6982.85±650.76	9506.81±1052.25	0.206±0.062	7734.56±247.99	9827.56±354.51	0.166±0.031
GAE	7183.24±579.82	9374.20±913.56	0.163±0.112	8010.73±290.33	10341.09±362.28	0.071±0.027
DGI	6423.44±671.25	8495.16±972.87	0.305±0.151	5330.11±261.77	7381.92±358.09	0.526±0.032
Transformer	6837.67±716.28	9042.02±1032.99	0.269±0.081	5345.17±216.30	7379.47±308.36	0.522±0.039
RegionDCL-no random	6400.50±630.35	8437.89±993.41	0.364±0.075	5228.27±210.46	7278.70±322.85	0.535±0.040
RegionDCL-fixed margin	6237.61±647.54	8387.56±948.78	0.365±0.107	5125.66±184.27	7159.65±250.12	0.551±0.033
RegionDCL	**5807.54±522.74**	**7942.74±779.44**	**0.427±0.108**	**5020.20±216.63**	**6960.51±282.35**	**0.575±0.039**
One-tailed two-sample t-test on RegionDCL and the second best method						
Test statistic	3.9651	2.4272	3.5909	4.9958	5.0616	5.2455
p-value	0.0001	0.0091	0.0003	0.0000	0.0000	0.0000

Table 4: Land Use and Population Density Inference in New York City with Different Input Data

Models	Input Data	Land Use Inference			Population Density Inference		
		L1↓	KL↓	Cosine↑	MAE↓	RMSE↓	R^2 ↑
MGFN	Mobility	0.503±0.013	0.304±0.015	0.882±0.007	5731.36±169.52	7481.89±239.00	0.516±0.036
MVURE	Mobility,POI	0.542±0.014	0.329±0.014	0.872±0.006	6557.12±265.30	8307.58±329.44	0.404±0.034
MVURE	Mobility,POI,Building	0.491±0.012	0.284±0.013	0.888±0.006	5447.44±195.38	7304.56±251.11	0.539±0.1030
RegionDCL	Building,POI	**0.418±0.010**	**0.229±0.008**	**0.912±0.004**	**5020.20±216.63**	**6960.51±282.35**	**0.575±0.039**

land use inference and population density estimation tasks for evaluation.

As shown in Table 4, when human mobility data is available, adding building data can further enhance the embedding effectiveness as demonstrated by the results of MVURE. This empirically supports our claim of data effectiveness in Section 1. While MGFN yields good results by utilizing hundreds of fine-grained per-hour mobility graphs, RegionDCL still performs the best without using mobility data, which highlights the potential of our method in those cities where human mobility data is not available.

5.5 Case Studies

We conduct case studies using Singapore data to demonstrate the capability of RegionDCL in dealing with different region partitions and data-sparse regions.

5.5.1 Adaptability to Different Region Partitions. The performance of the region representation models can be greatly influenced by the choice of region partition. To evaluate the adaptability of the models to different region partitions, we replace Singapore Subzones with 2km×2km grids that are widely used in region representation learning. As mentioned in Section 4.3.3, our model can adapt to

Table 5: Land Use Inference of Grid Regions in Singapore

Models	Land Use Inference		
	L1↓	KL↓	Cosine↑
Urban2Vec	0.726±0.024	0.527±0.028	0.764±0.014
Place2Vec	0.645±0.051	0.449±0.072	0.814±0.026
Doc2Vec	0.735±0.037	0.493±0.036	0.769±0.016
GAE	0.674±0.054	0.428±0.060	0.804±0.029
DGI	0.621±0.034	0.364±0.050	0.836±0.018
Transformer	0.541±0.044	0.326±0.053	0.860±0.020
RegionDCL	**0.485±0.020**	**0.260±0.028**	**0.890±0.012**

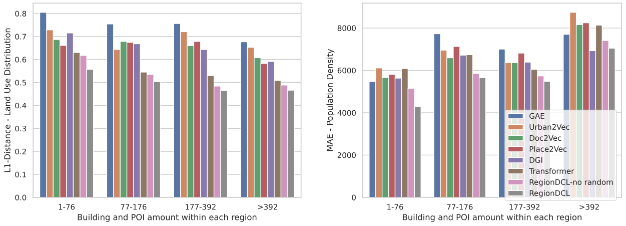


Figure 3: The Prediction Error in Regions with different numbers of Buildings and POIs in Singapore.

new region partitions without re-training, while baseline methods must be retrained on the new partition.

As shown in Table 5, methods designed for a single region partition (Urban2Vec, Doc2Vec, and DGI) suffer from a significant performance decay, while Transformer and RegionDCL still perform stably, indicating that the proposed dual contrastive learning has improved the model's adaptability to different region partitions.

5.5.2 Representation Quality in Data-sparse Regions. We perform an evaluation of the representation quality of each region on land use inference and population density estimation tasks. The results by dividing all regions into four groups of the same size (i.e., 76 regions), based on the number of POIs and buildings they contained. The average L1 distance and MAE of each group is reported in Fig. 3.

From Fig. 3, we observe that RegionDCL consistently achieves the lowest prediction error on land use inference and population density estimation tasks. This performance superiority is especially prominent in regions with fewer than 76 buildings and POIs, indicating that RegionDCL generates high-quality representations for regions even with sparse building and POI information. Furthermore, the comparison of RegionDCL with its variant RegionDCL-no random highlights that the improvement is largely attributed to the proposed random point sampling strategy.

Beyond the quantitative analysis, we also visualized the K-Means clusters of building group embeddings from RegionDCL to further demonstrate the representation quality. Please refer to Appendix B.2 for the visualization results.

5.6 Parameter Sensitivity Analysis

We show the parameter sensitivity of RegionDCL to its two key hyper-parameters (i.e., the radius r of Poisson Disk Sampling and the scaler λ in adaptive margin). We evaluate RegionDCL's performance with L1-distance for Land Use Inference and MAE for Population Density Inference in Singapore.

5.6.1 Poisson Disk Sampling Radius r. In Poisson Disk Sampling, the hyper-parameter r is used to control the radius of the sampling disk, where larger r results in more sparse random points, and vice versa. As shown in Table 6, too sparse random points (i.e., radius $\geq 150m$) lead to a performance decline, as too-few random points cannot sufficiently describe the size and shape of empty areas. On the other side, too-dense random points have a slightly negative effect as they would bring noisy data in training the model.

Table 6: The effect of Poisson Disk Sampling radius r

r	50	75	100	125	150	175
Random Points	107925	39170	18970	10749	6649	4445
L1 - Land Use↓	0.518	0.514	0.508	0.511	0.523	0.528
MAE - Population↓	6042	6128	5856	5798	6440	6540

5.6.2 The Scaler λ in Adaptive Margin. The proposed adaptive margin in the formulated triplet loss (i.e., Eq. (11)) contains a λ to scale the Wasserstein Distance, where larger λ leads to larger embedding distances between dissimilar regions. As shown in Table 7, A large or small λ would lead to significant performance decay. $50 \leq \lambda \leq 200$ is an ideal range for high model performance.

Table 7: The effect of scaler λ in Adaptive Margin

λ	1	5	10	20	50	100	200	300	500
L1 - Land Use↓	0.541	0.528	0.516	0.506	0.498	0.494	0.492	0.507	0.523
MAE - Population↓	6637	6246	6131	6048	5986	5980	6105	6298	6491

6 CONCLUSIONS

In this work, we proposed an unsupervised model RegionDCL to learn urban region representations with OpenStreetMap (OSM) data via contrastive learning in the two levels of building groups and regions. We empirically demonstrated its effectiveness on two typical downstream tasks, even for data-sparse regions or with different region partitions. We also visualized the learned building group embeddings to show its promising performance against ground truth land use. Potential future work includes: 1) integrating more OSM features (e.g., bus route) into region representations, 2) designing data augmentation techniques to boost contrastive learning for hard pattern mining, and 3) adapting RegionDCL to multiple cities for urban sensing applications.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-020[T]), the MOE Tier-2 grant MOE-T2EP20221-0015, and the Knut and Alice Wallenberg Foundation (to W.H.).

REFERENCES

[1] Anahid Basiri, Muki Haklay, Giles Foody, and Peter Mooney. 2019. Crowdsourced geospatial data quality: challenges and future directions. *International Journal of Geographical Information Science* 33, 8 (2019), 1588–1593.

[2] Robert Bridson. 2007. Fast Poisson disk sampling in arbitrary dimensions. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference.* 22–es.

[3] Alessia Calafiore, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, and Alex Singleton. 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems* 85 (2021), 101539.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. 1597–1607.

[5] Philip J. Clark and Francis C. Evans. 1954. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology* 35, 4 (1954), 445–453.

[6] UN DESA. 2019. World urbanization prospects 2018: Highlights. *UN DESA* (2019).

[7] Shihong Du, Liqun Luo, Kai Cao, and Mi Shu. 2016. Extracting building patterns with multilevel graph partition and building grouping. *ISPRS Journal of Photogrammetry and Remote Sensing* 122 (2016), 81–96.

[8] Kenneth Frampton. 2020. *Modern architecture: a critical history (world of art).* Thames & Hudson.

[9] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. 2019. Efficient Region Embedding with Multi-View Spatial Networks: A Perspective of Locality-Constrained Spatial Autocorrelations. In *Proceedings of the 32rd AAAI Conference on Artificial Intelligence.* 906–913.

[10] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[12] Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. 2022. Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science* 36, 10 (2022), 1905–1930.

[13] Bo Hui, Da Yan, Wei-Shinn Ku, and Wenlu Wang. 2020. Predicting Economic Growth by Region Embedding: A Multigraph Convolutional Network Approach. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management.* 555–564.

[14] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised Representation Learning of Spatial Data via Multimodal Embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.* 1993–2002.

[15] Thomas N. Kipf and Max Welling. 2016. Variational Graph Auto-Encoders. *CoRR* abs/1611.07308 (2016).

[16] Laksono Kurnianggoro, Wahyono, and Kang-Hyun Jo. 2018. A survey of 2D shape representation: Methods, evaluations, and future research directions. *Neurocomputing* 300 (2018), 1–16.

[17] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[18] William B. Meyer and B. L. Turner. 1992. Human Population Growth and Global Land-Use/Cover Change. *Annual Review of Ecology and Systematics* 23 (1992), 39–61.

[19] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition.* 793–799.

[20] Haifeng Niu and Elisabete A. Silva. 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 (2021), 101651.

[21] Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* 28, 9 (2014), 1988–2007.

[22] Ioannis A Pissourios. 2019. Survey methodologies of urban land uses: An oddment of the past, or a gap in contemporary planning theory? *Land Use Policy* 83 (2019), 403–411.

[23] Srinivas S. Pulugurtha, Venkata Ramana Duddu, and Yashaswi Kotagiri. 2013. Traffic analysis zone level crash estimation models based on land use characteristics. *Accident Analysis & Prevention* 50 (2013), 678–687.

[24] Zhinan Qiao and Xiaohui Yuan. 2021. Urban land-use analysis using proximate sensing imagery: a survey. *International Journal of Geographical Information Science* 35, 11 (2021), 2129–2148.

[25] B. D. Ripley. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* 13, 2 (1976), 255–266.

[26] Zhongming Shi, Jimeno A. Fonseca, and Arno Schlueter. 2021. A parametric method using vernacular urban block typologies for investigating interactions between solar energy use and urban design. *Renewable Energy* 165 (2021), 823–841.

[27] W. R. Tobler. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46, sup1 (1970), 234–240.

[28] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30.* 5998–6008.

[30] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax. In *Proceedings of the 7th International Conference on Learning Representations.*

[31] Hongjian Wang and Zhenhui Li. 2017. Region Representation Learning via Mobility Flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 237–246.

[32] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence.* 1013–1020.

[33] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-Graph Fusion Networks for Urban Region Embedding. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence.* 2312–2318.

[34] Xiongfeng Yan, Tinghua Ai, Min Yang, and Xiaohua Tong. 2021. Graph convolutional autoencoder model for the shape coding and cognition of buildings in maps. *International Journal of Geographical Information Science* 35, 3 (2021), 490–512.

[35] Xiongfeng Yan, Tinghua Ai, Min Yang, and Hongmei Yin. 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), 259–273.

[36] Min Yang, Bo Kong, Ruirong Dang, and Xiongfeng Yan. 2022. Classifying urban functional regions by integrating buildings and points-of-interest using a stacking ensemble method. *International Journal of Applied Earth Observation and Geoinformation* 108 (2022), 102753.

[37] Yao Yao, Jiaqi Zhang, Chen Qian, Yu Wang, Shuliang Ren, Zehao Yuan, and Qingfeng Guan. 2021. Delineating urban job-housing patterns at a parcel scale with street view imagery. *International Journal of Geographical Information Science* 35, 10 (2021), 1927–1950.

[38] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. 2018. Representing Urban Functions through Zone Embedding with Human Mobility Patterns. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence.* 3919–3925.

[39] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do Transformers Really Perform Badly for Graph Representation?. In *Proceedings of the 2021 Annual Conference on Neural Information Processing Systems.* 28877–28888.

[40] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 186–194.

[41] Wei Zhai, Xueyin Bai, Yu Shi, Yu Han, Zhong-Ren Peng, and Chaolin Gu. 2019. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Computers, Environment and Urban Systems* 74 (2019), 1–12.

[42] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning. In *Proceedings of the 26th International World Wide Web Conference.* 361–370.

[43] Liang Zhang, Cheng Long, and Gao Cong. 2022. Region Embedding With Intra and Inter-View Contrastive Learning. *IEEE Transactions on Knowledge and Data Engineering* (2022), 1–6.

[44] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2020. Multi-View Joint Graph Representation Learning for Urban Region Embedding. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence.* 4431–4437.

[45] Xiuyuan Zhang, Shihong Du, and Qiao Wang. 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS Journal of Photogrammetry and Remote Sensing* 132 (2017), 170–184.

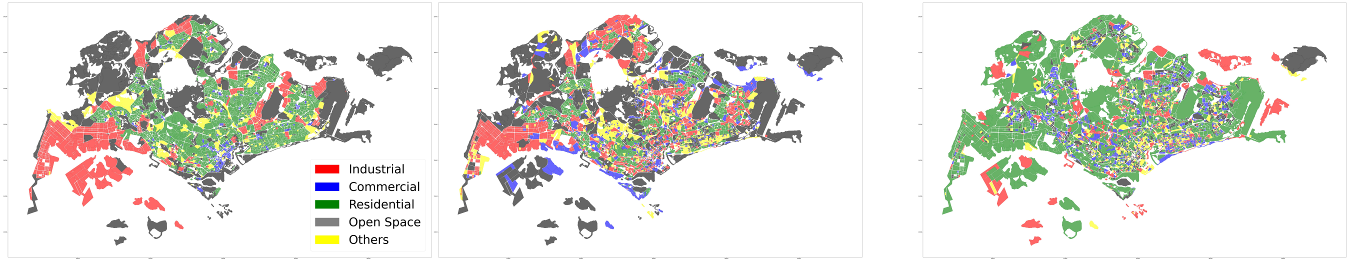
[46] Yunchao Zhang, Yanjie Fu, Pengyang Wang, Xiaolin Li, and Yu Zheng. 2019. Unifying Inter-region Autocorrelation and Intra-region Structures for Spatial Embedding via Collective Adversarial Learning. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1700–1708.

[47] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 1–55.

[48] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air: when urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1436–1444.

Table 8: The data sources and links of used datasets

Data Type	Data Source	Link
Buildings, POIs	OpenStreetMap	https://download.geofabrik.de/
Region partitions - Singapore	Singapore Public Data	https://data.gov.sg/dataset/master-plan-2019-subzone-boundary-no-sea
Land use - Singapore	Singapore Public Data	https://data.gov.sg/dataset/master-plan-2019-land-use-laye
Region partitions - New York City	NYC Planning	https://www.nyc.gov/site/planning/data-maps/open-data/census-download-metadata.page
Land use - New York City	NYC Planning	https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
Population density	WorldPop	https://hub.worldpop.org/geodata/listing?id=77
Trajectory - New York City	NYC Yellow Taxi Trip	https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t



(a) Ground truth land use

(b) RegionDCL

(c) Transformer

Figure 4: The K-Means clustering results of building group embeddings.

A DATA SOURCE

In this paper, all datasets we used are available online. We hereby provide their links in Table 8.

B ADDITIONAL EXPERIMENT RESULTS

B.1 The effect of dropout rate in the group-level contrastive learning

The proposed group-level contrastive learning contains a dropout rate to control how many objects to drop when constructing positive building group pairs. Smaller dropout rates will result in more similar positive sample pairs. As shown in Table 9, a dropout rate of 0.2 can ensure the model captures minor differences between positive samples effectively.

Table 9: The effect of dropout rate

dropout rate	0	0.1	0.2	0.3	0.4
L1 - Land Use↓	0.553	0.524	0.491	0.518	0.559
MAE - Population↓	6831	6635	6395	6542	6533

B.2 Visualization

In addition to learning region representations, RegionDCL also demonstrates impressive unsupervised capability in discovering urban functions. To showcase this capability, we visualize the derived building group embeddings by applying K-Means to group them into five categories: *Residential*, *Industrial*, *Commercial*, *Open Space*, and *Others* as described in Section 5.2.

The results, shown in Fig. 4, unveil that RegionDCL has largely captured the genuine land use. Conversely, Transformer fails to distinguish between residential, industrial, and open space areas. Our findings suggest that RegionDCL has successfully uncovered the representative urban functions from building groups, leading to high-quality region representations.