Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2023

Urban Region Representation Learning with OpenStreetMap Building Footprints

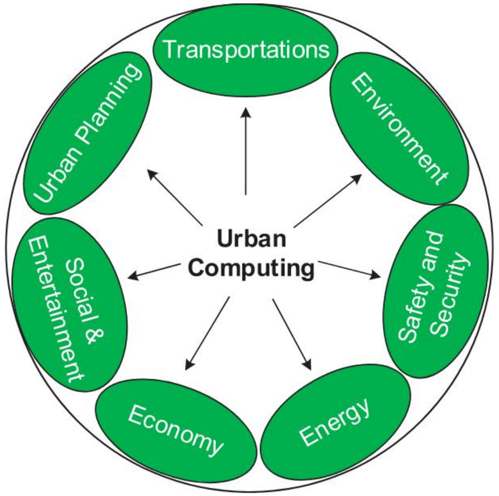
Yi Li, Weiming Huang, Gao Cong, Hao Wang, Zheng Wang

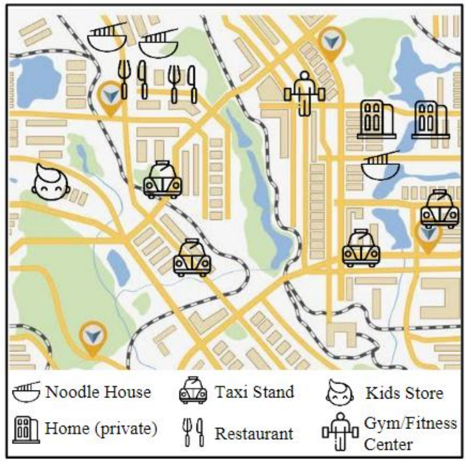
汇报人：程佳伟

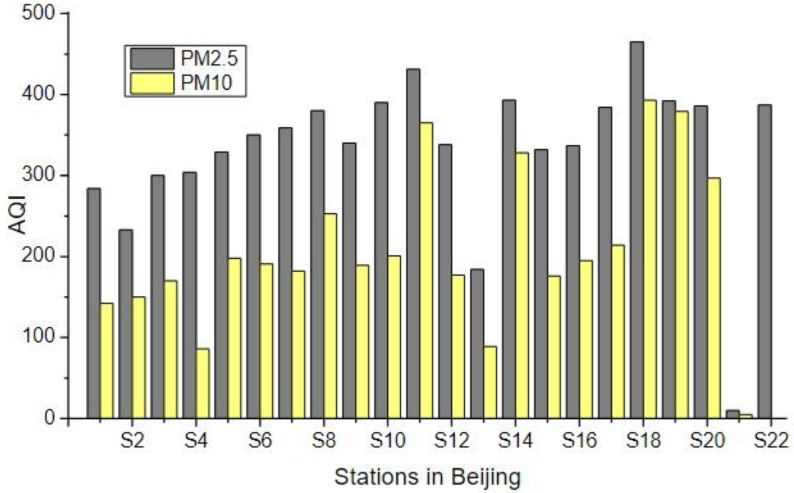
Background

- **城市化所带来的问题愈发严重**

- **随着大数据技术的普及，城市计算成为了解决城市问题的有力工具**





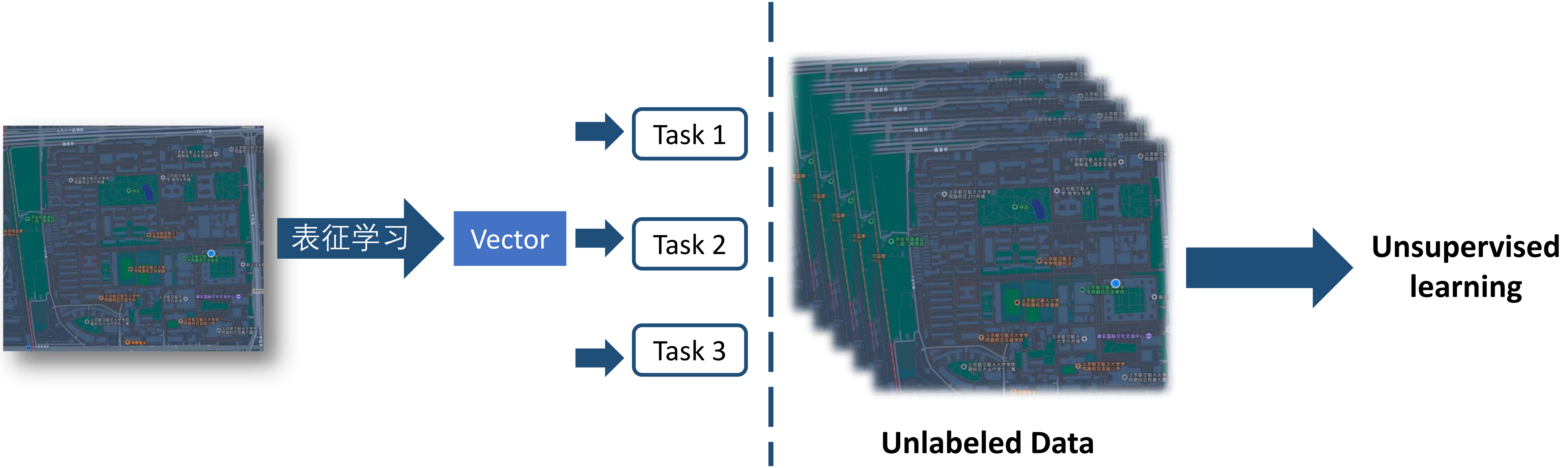


但是它们通常只专注于单个任务的解决，并且依赖于邻域专业知识进行监督

Background

- 在城市计算领域中，**区域表征学习**开始流行
- 区域表征学习具有两个优点

1. **多任务处理**，能够应用于各种下游任务，比如识别土地利用、预测空气质量等
2. **无监督学习**，减少了对大量标记数据的依赖



表征学习

Vector

Task 1

Task 2

Task 3

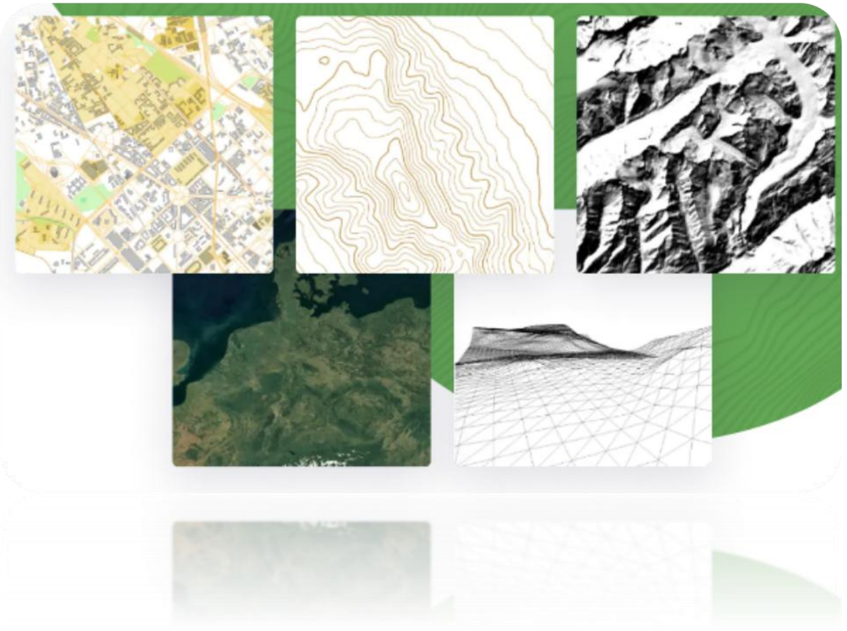
Unlabeled Data

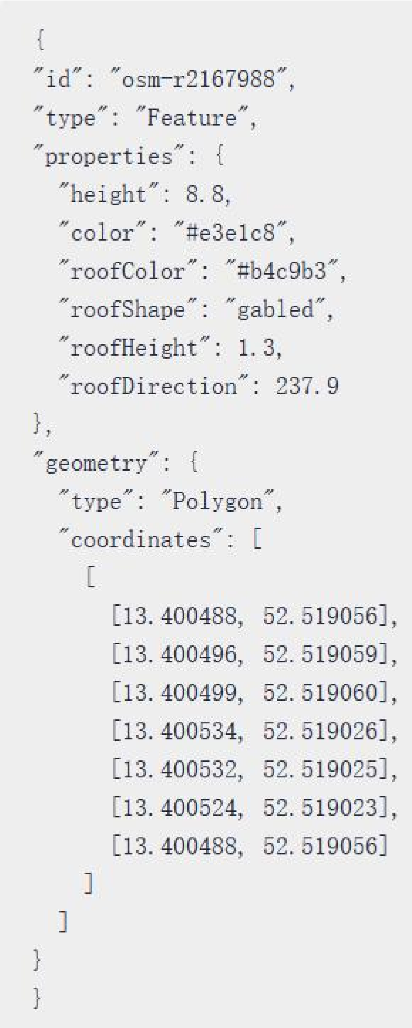
Unsupervised learning

Background

- OpenStreetMap(OSM)是一个提供地理空间数据的开源平台
- 可以提供建筑footprint、POI和道路网络的数据
- 有非常好的数据有效性和数据可用性

于是尝试用OSM的建筑footprint进行区域表征学习(first)



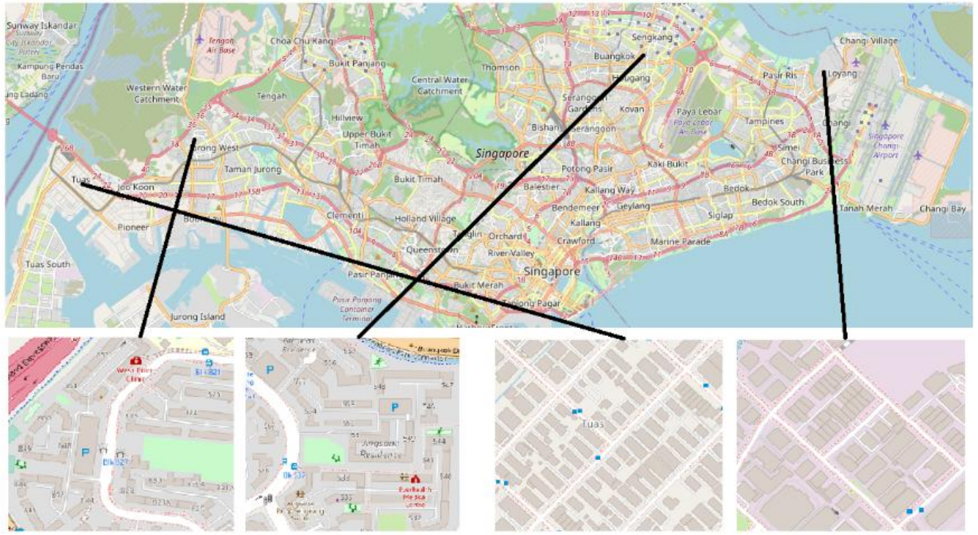


```
{
"id": "osm-r2167988",
"type": "Feature",
"properties": {
  "height": 8.8,
  "color": "#e3e1c8",
  "roofColor": "#b4c9b3",
  "roofShape": "gabled",
  "roofHeight": 1.3,
  "roofDirection": 237.9
},
"geometry": {
  "type": "Polygon",
  "coordinates": [
    [
      [13.400488, 52.519056],
      [13.400496, 52.519059],
      [13.400499, 52.519060],
      [13.400534, 52.519026],
      [13.400532, 52.519025],
      [13.400524, 52.519023],
      [13.400488, 52.519056]
    ]
  ]
}
}
```

Situation & Challenge

- Representation Learning

1. 目前的研究对于区域中的建筑物往往采用point-oriented的方式处理，而**没有考虑其几何形状**
2. 目前的研究**忽视了区域中建筑的空间分布和空间关系**
3. 目前的研究基于**距离相近则相似**的原则，但是存在反例



Residential Areas　　Industrial Areas

- Data Sparse

1. 城市区域建筑分布不均匀，有些区域是**data-sparse**的(例如empty或者unmapped)，而这些区域的特征在目前的研究中往往被忽略
2. data-sparse区域**具有不同的形状和不同的相邻环境**，容易误处理（比如大片未开发区和小片的居民区的建筑都很少，误认为相似)

- Downstream task

1. 目前研究依赖于**所有下游任务对区域的分割是一致的**，但现实中往往不是(比如人口普查区域、交通区域、行政区域)

Definition & Problem

Building Footprint

- A building footprint � 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).

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.

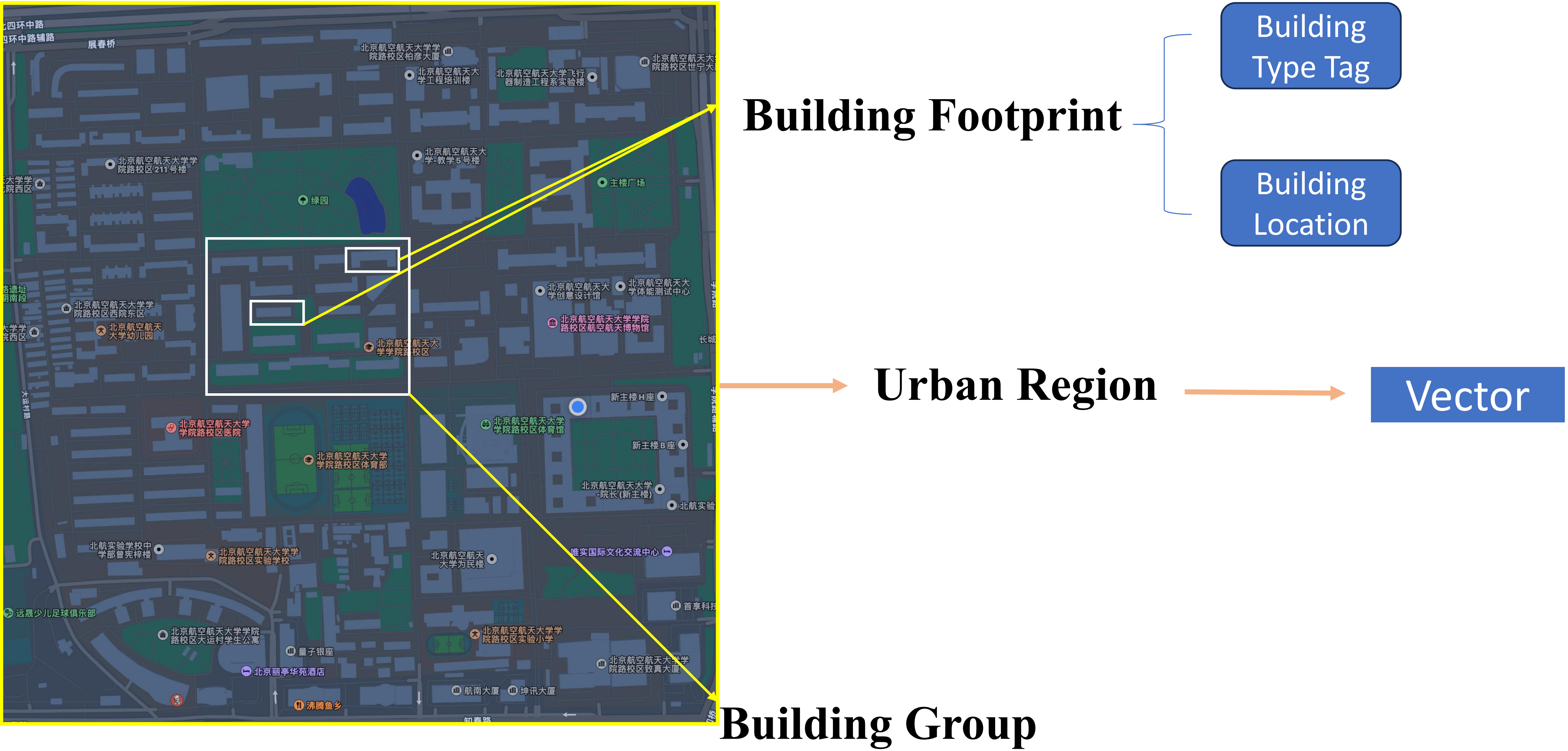
Urban Region

- Urban regions U refer to a set of disjoint city areas, usually obtained through a certain partition approach (e.g., census tracts). Each urban region � may include multiple building groups.

Problem Statement

Given a set of urban regions U = {� 1, � 2, ...}, the goal of urban region representation learning is to learn a mapping function that generates a vector representation � � for each region � � in the Euclidean space, where � is the uniform dimension for all � � ∈ U.

Definition & Problem


北四环中路
四环中路辅路
展春桥
北京航空航天大学学院路校区柏彦大厦
北京航空航天大学工程培训楼
北京航空航天大学飞行器制造工程系实验楼
北京航空航天大学院路校区世宁大
北京航空航天大学学院路校区211号楼
北京航空航天大学-教学5号楼
大学学院西区
绿园
主楼广场
路遗址朝南段
北京航空航天大学学院路校区西院东区
北京航空航天大学幼儿园
大学学院西区
北京航空航天大学创意设计馆
北京航空航天大学体能测试中心
北京航空航天大学学院路校区航空航天博物馆
北京航空航天大学学院路校区
长城
大运村路
新主楼H座
北京航空航天大学学院路校区医院
北京航空航天大学学院路校区体育馆
北京航空航天大学学院路校区体育部
新主楼B座
北京航空航天大学-院长(新主楼)
北航实验
北航实验学校中学部曾宪梓楼
北京航空航天大学学院路校区实验学校
北京航空航天大学为民楼
唯实国际文化交流中心
首享科技
远晟少儿足球俱乐部
北京航空航天大学学院路校区大运村学生公寓
量子银座
北京航空航天大学学院路校区实验小学
北京航空航天大学学院路校区致真大厦
北京丽亭华苑酒店
航南大厦
坤讯大厦
沸腾鱼乡
知春路

Building Footprint

Building Type Tag

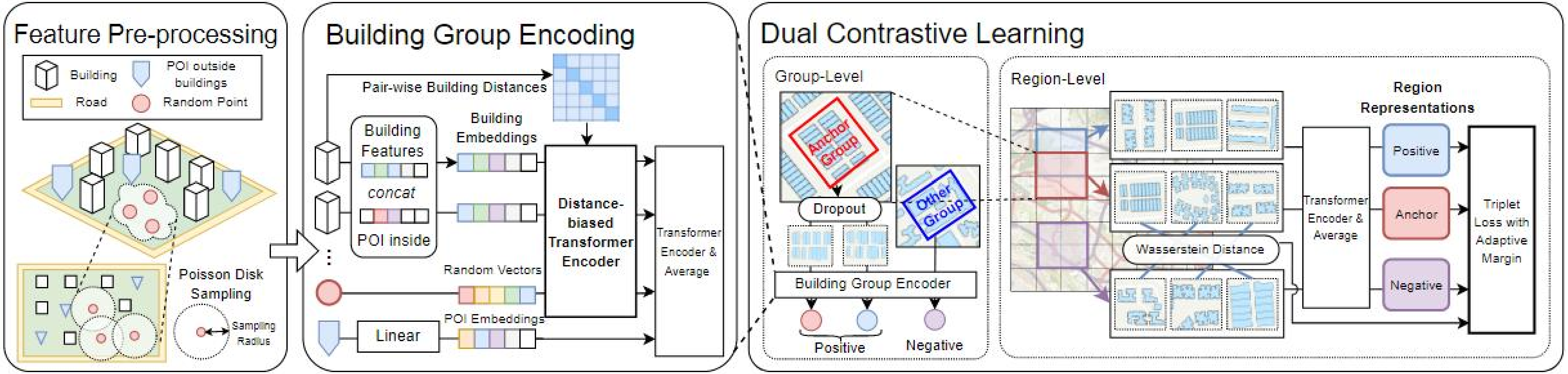
Building Location

Urban Region

Vector

Building Group

Overview: RegionDCL



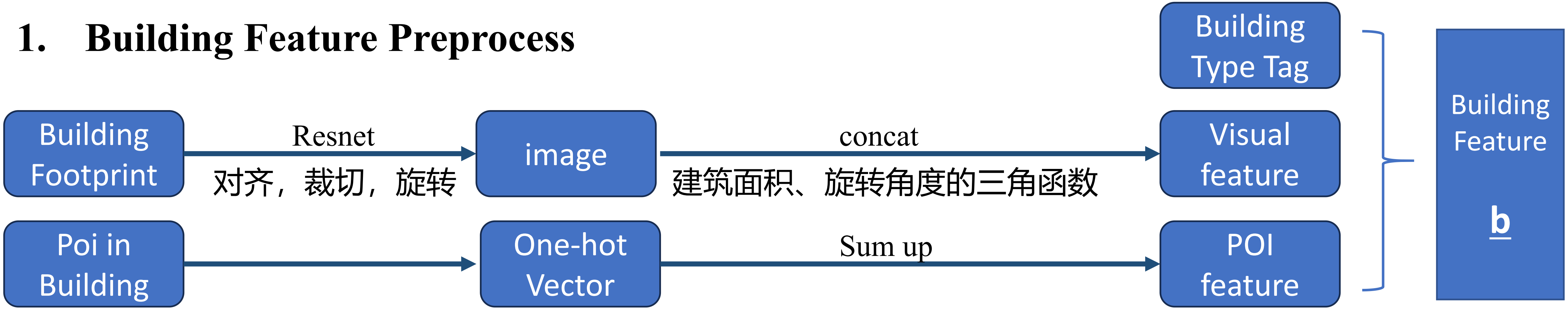
Feature Pre-process

Building Group Encoding

Dual Contrastive Learning

Feature Pre-process

1. Building Feature Preprocess



Building Footprint → Resnet (对齐，裁切，旋转) → image → concat (建筑面积、旋转角度的三角函数) → Visual feature

Poi in Building → One-hot Vector → Sum up → POI feature

Building Type Tag

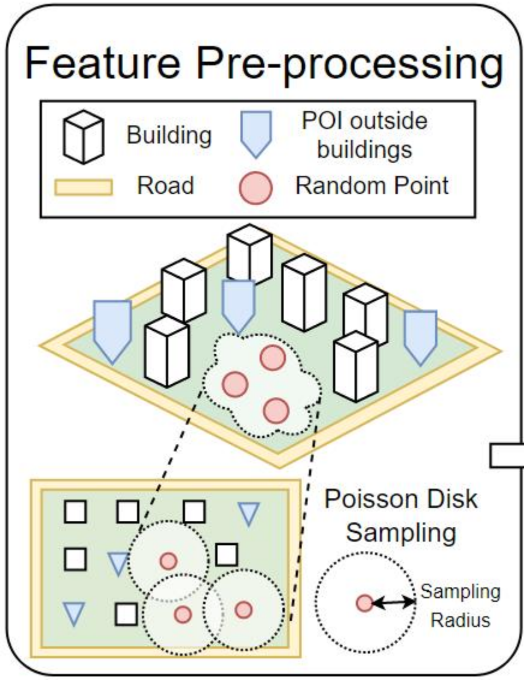
Building Type Tag, Visual feature, POI feature → Building Feature **b**

2. Random Points

泊松盘方法是一种允许在二维平面空间内均匀生成随机点，且任何两个点的距离都不会隔得太近的方法。

- 使用**泊松盘采样**方法，给data-sparse的区域来填充随机点
- 每个随机点之间保持最小的采样距离半径 r ，且携带一个统一的向量**s**
- 以此方式，我们能够保留data-sparse区域的存在和信息

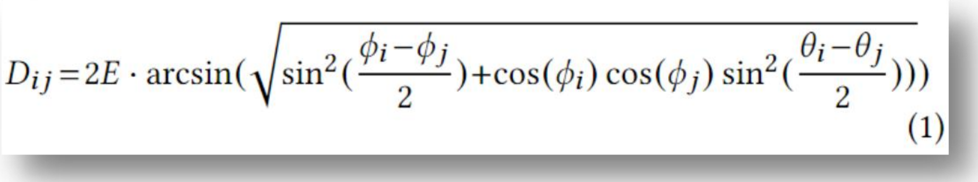
此外将building外的POI记为 **t**



Building Group Encoding

1. 计算距离矩阵

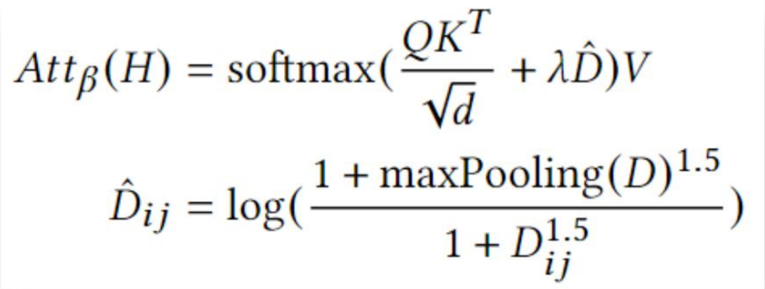
- 给定building group 的feature � $= [b_1^T, ..., b_j^T, ...s_1^T, ..., s_l^T]$
- 计算building group 中的每对建筑和随机点的距离矩阵**D**



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

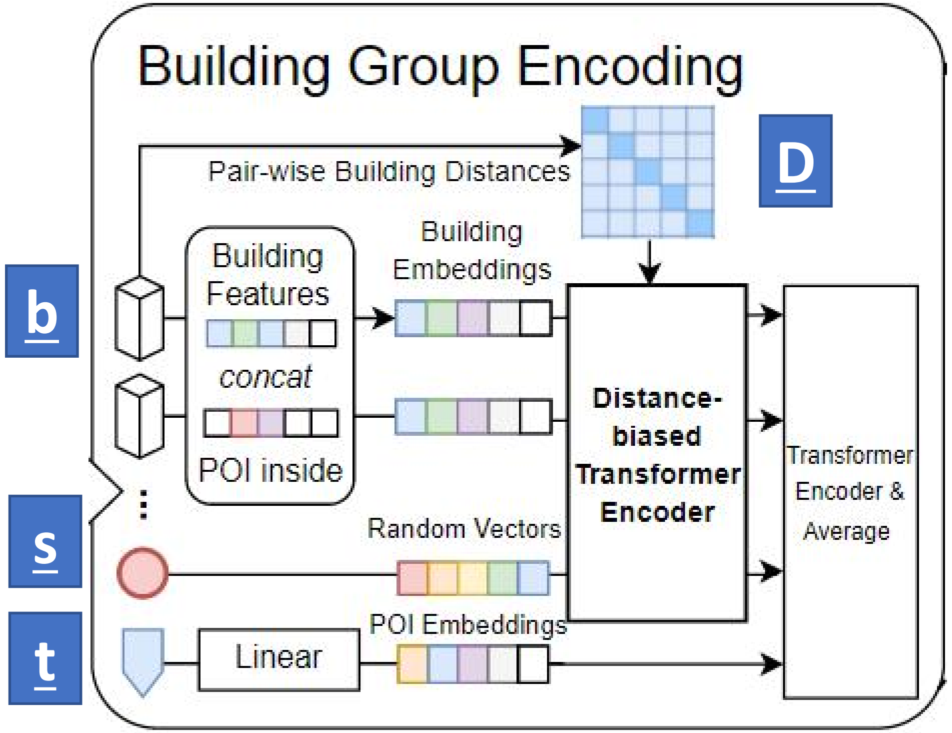
2. 使用Distance-biased Transformer Encode

$$Q = HW_Q,\ K = HW_K,\ V = HW_V$$



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

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



Dual Contrastive Learning： Group-level

Group-level Contrastive learning

- **想法：仅有一小部分不同的building group应该是相似的**

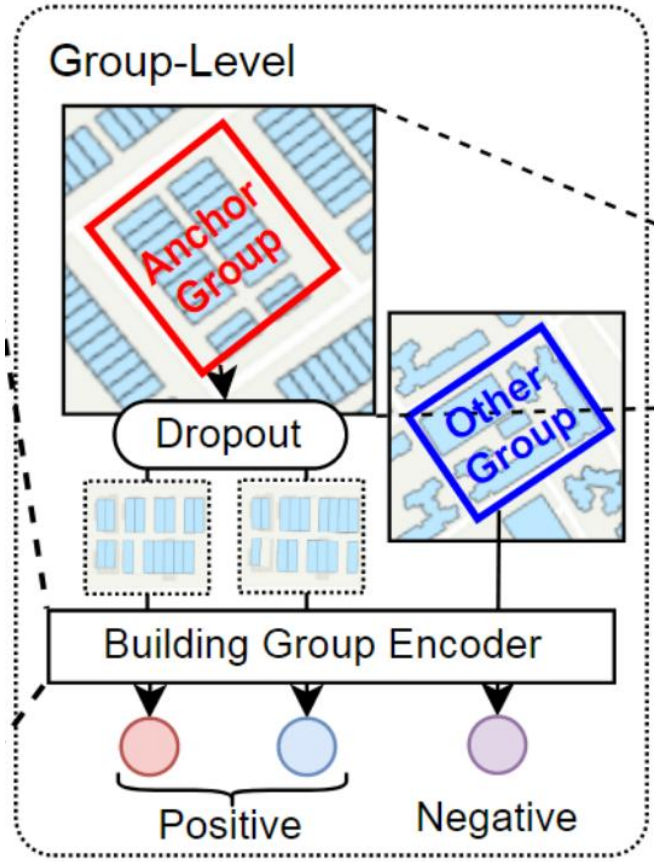
- **步骤：**

① 在一个训练batch里，依次将batch里的每个building group选择为anchor

② 随机将该anchor里的building 删除少量，删除后的anchor作为正样本

③ batch里除了anchor外的为负样本

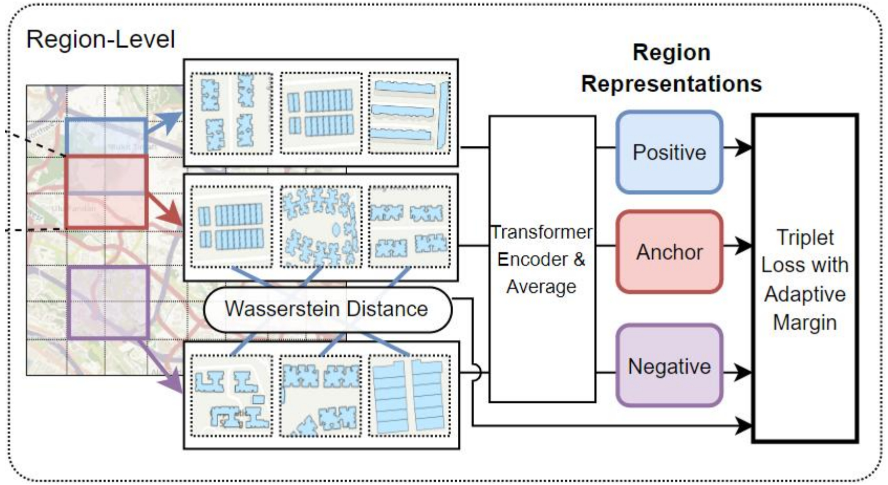
$$\mathcal{L}_{InfoNCE} = -\log(\frac{e^{\mathrm{sim}(P_i,\, P_i^{+})/\tau}}{\sum_{i=0}^{n} e^{\mathrm{sim}(P_i,\, P_j)/\tau}})$$



Dual Contrastive Learning： Region-level

Region-level Contrastive learning

- **想法：邻近的区域比远的更有可能相关，但远的区域有可能也有相似的建筑群**



1. 使用滑动窗口来选取样本区域（使用多个大小的window）
   - 使用给定大小的窗口，水平或垂直移动生成训练区域
   - 重叠的区域为正样本，随机一个非重叠的部分为负样本
2. 根据区域里的building group计算正负样本的**JS散度矩阵C**

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

a_i, b_j分别是两个区域的building group表征向量

3. 根据散度矩阵计算**Wasserstein距离**

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

   - 两个区域越相似，W越小，反之越大
4. 最后的损失函数为 $\hat{\mathcal{L}}_{triplet} = \max(||z_a - z_p|| - ||z_a - z_b|| + \lambda \cdot W,\ 0)$

其中z_a, z_p, z_n分别是anchor区，正样本，负样本的表征

Methodology Summary

- **Problem**

1. 没有考虑其几何形状
2. 忽视了区域中建筑的空间分布和空间关系
3. 存在相近相似反例
4. 忽视Data-Sparse
5. 下游任务依赖于一致的分割区域

- **Solution**

1. CNN提取Visual Feature
2. Distanced-biased Transformer
3. Region-Level Contrastive Learning
4. Poisson Disk Sampling
5. building group是基本单元，学习有效的building group representation

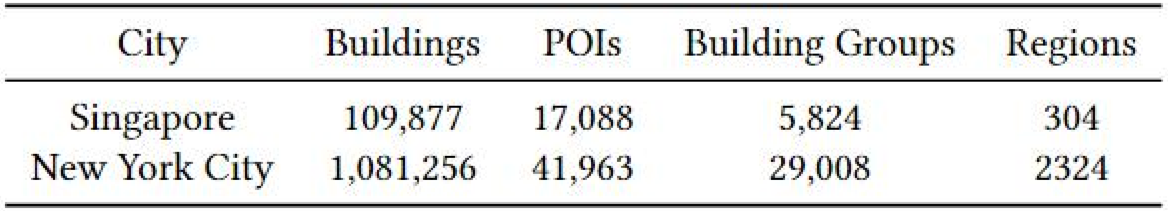
Experiment

- **Dataset**

  1. **Singapore**
  2. **New York City**

  包括OSM数据、土地使用数据、人口普查数据

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

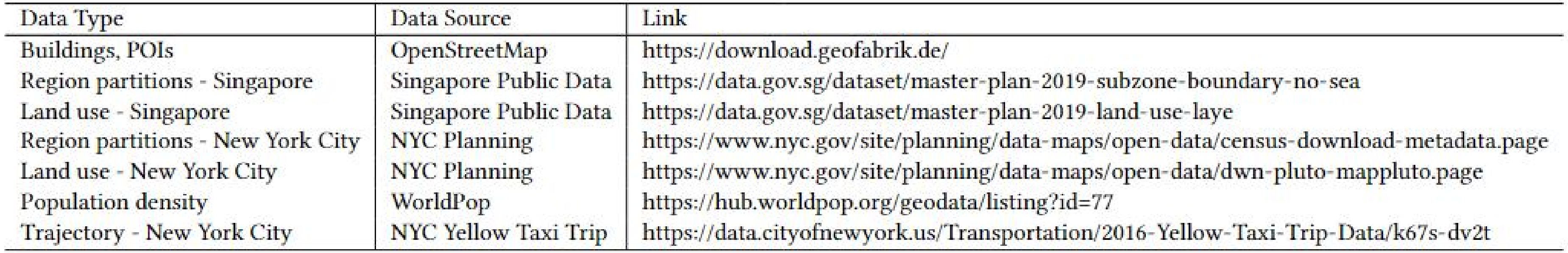
- **Downstream**

  1. **Land Use Inference (label distribution learning problem)**
  2. **Population Density Estimation (regression problem)**

- **Baseline**

  1. **Place2Vec**
  2. **Doc2Vec**
  3. **GAE**
  4. **DGI**
  5. **Urban2Vec**

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

Experiment

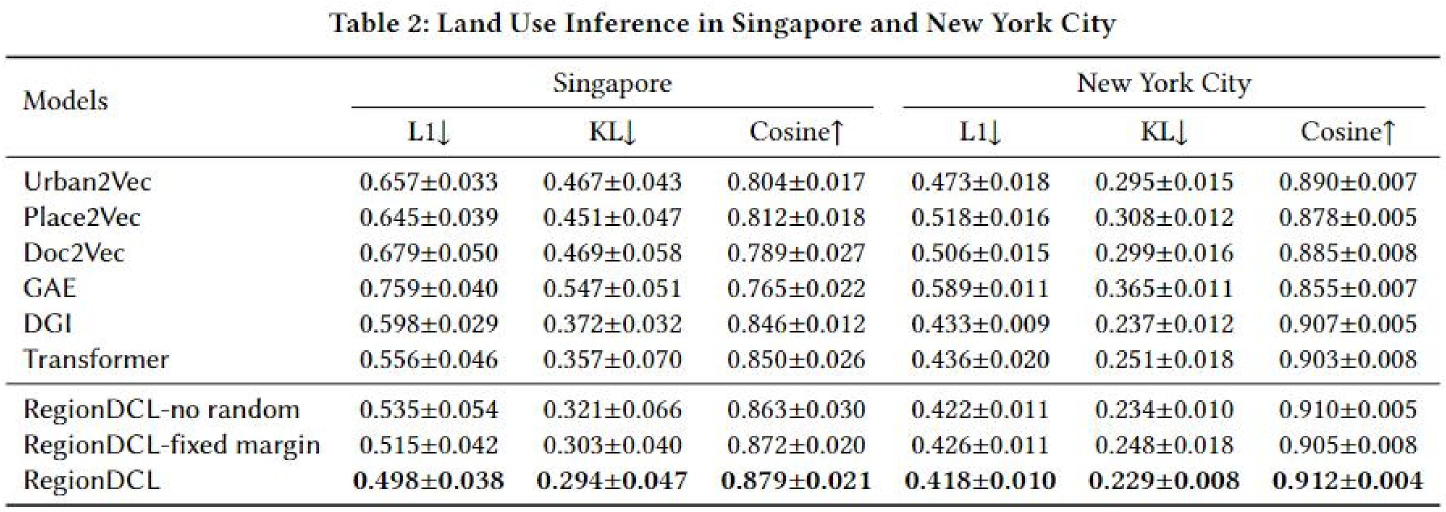
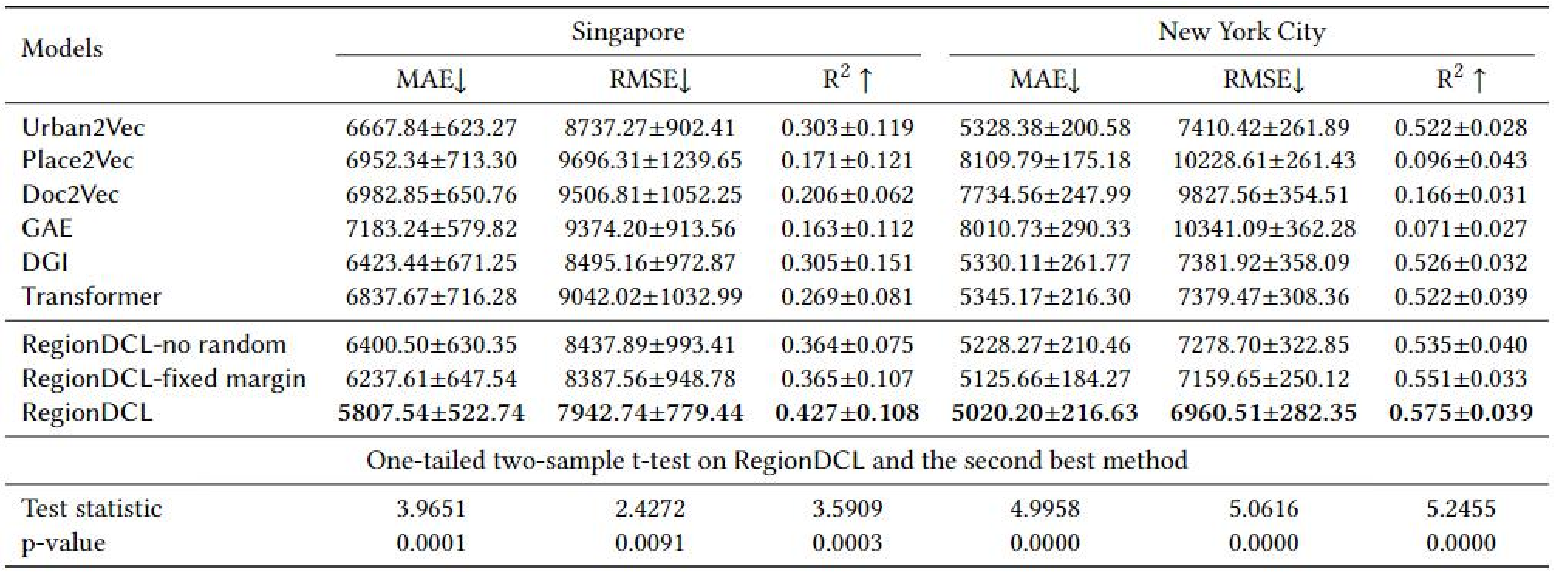


Table 2: Land Use Inference in Singapore and New York City

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**

- RegionDCL的表现优于所有baseline
- RegionDCL在新加坡表现出更大的改进。而新加坡比纽约有明显不同的建筑风格和更多的数据稀疏区域

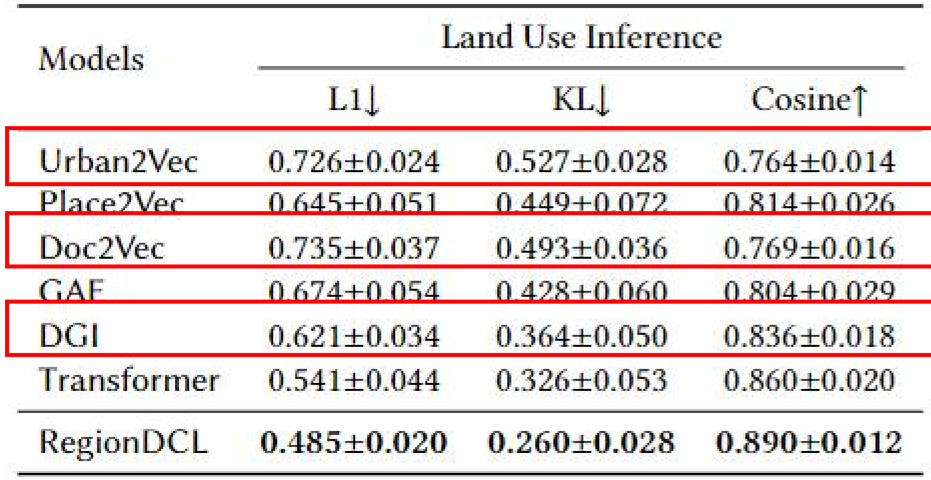


Models	Singapore			New York City		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Urban2Vec	6667.84±623.27	8737.27±902.41	0.303±0.119	5328.38±200.58	7410.42±261.89	0.522±0.028
Place2Vec	6952.34±713.30	9696.31±1239.65	0.171±0.121	8109.79±175.18	10228.61±261.43	0.096±0.043
Doc2Vec	6982.85±650.76	9506.81±1052.25	0.206±0.062	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

- RegionDCL的表现优于所有baseline

Experiment

- 将新加坡的土地分割方式替换为2*2的方格，发现有三种baseline的表现都明显下滑，RegionDCL的表现依旧良好



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**

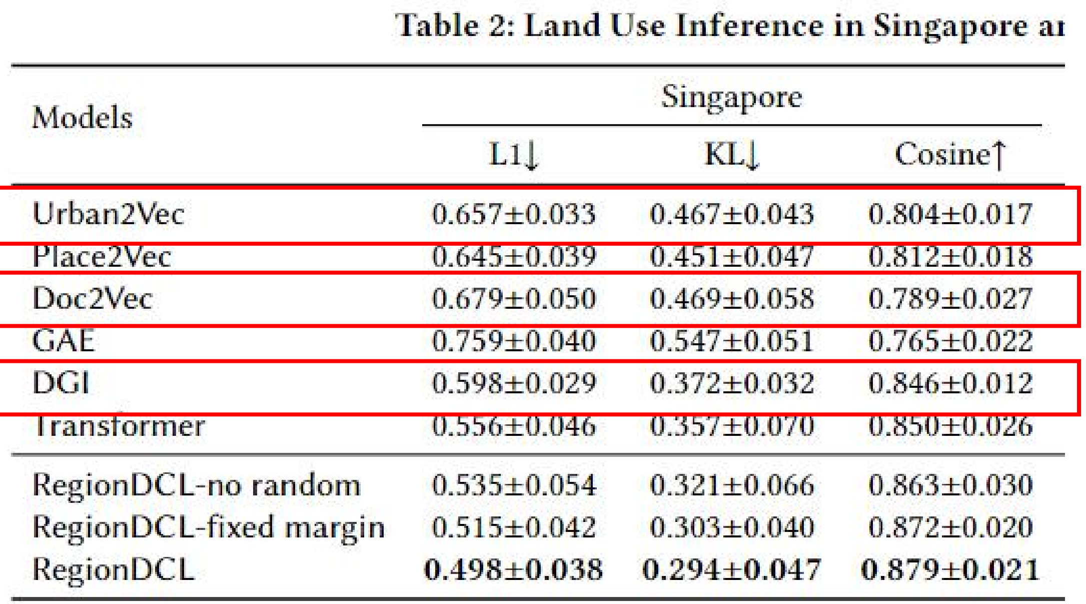
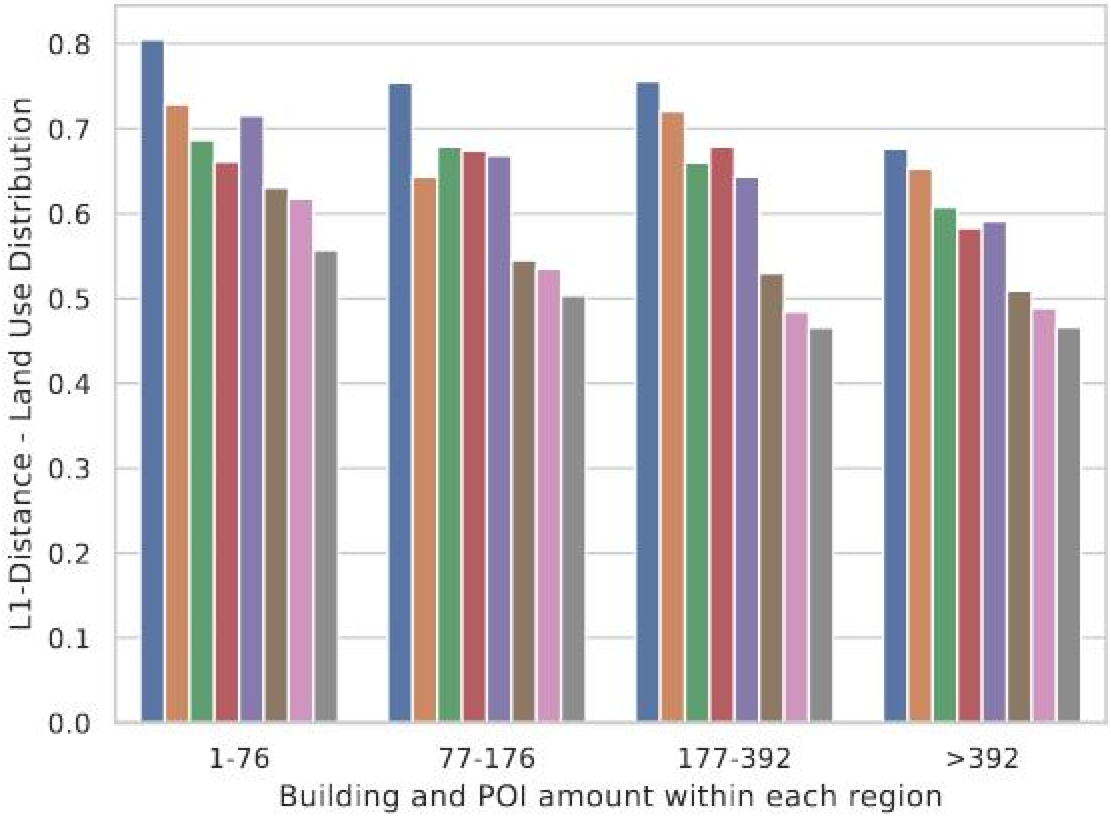


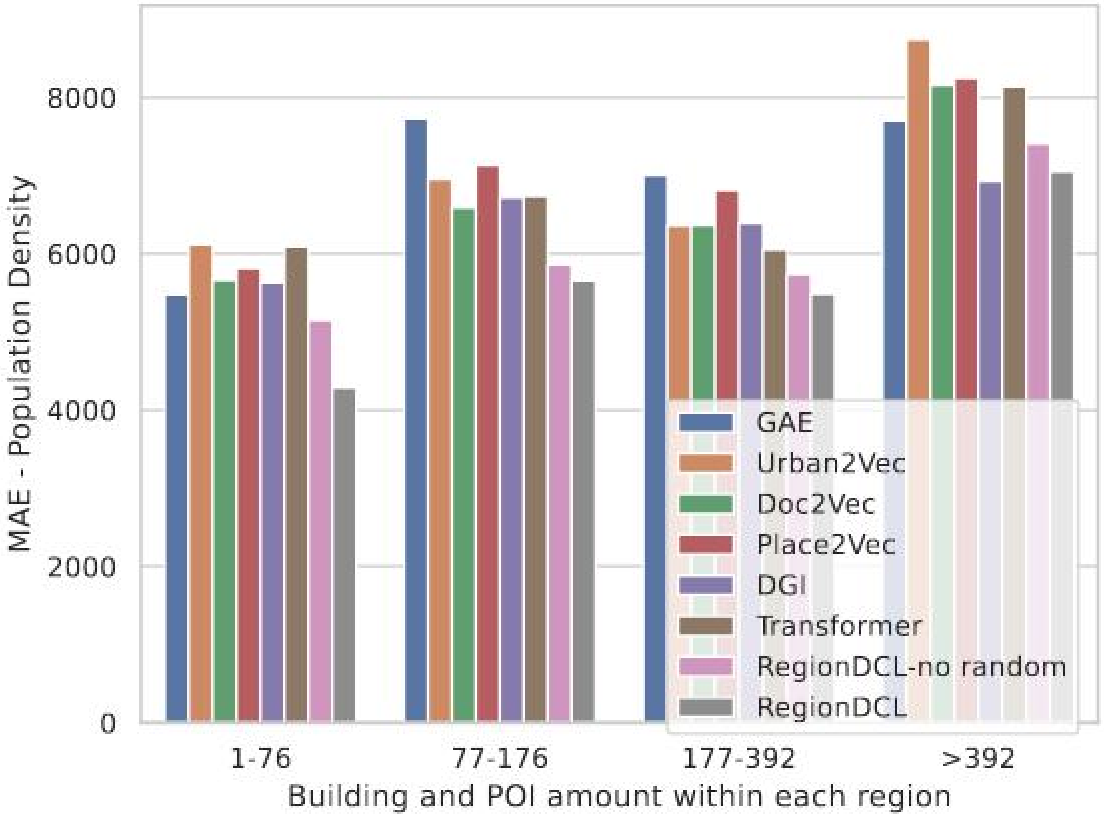
Table 2: Land Use Inference in Singapore ar

Models	Singapore		
	L1↓	KL↓	Cosine↑
Urban2Vec	0.657±0.033	0.467±0.043	0.804±0.017
Place2Vec	0.645±0.039	0.451±0.047	0.812±0.018
Doc2Vec	0.679±0.050	0.469±0.058	0.789±0.027
GAE	0.759±0.040	0.547±0.051	0.765±0.022
DGI	0.598±0.029	0.372±0.032	0.846±0.012
Transformer	0.556±0.046	0.357±0.070	0.850±0.026
RegionDCL-no random	0.535±0.054	0.321±0.066	0.863±0.030
RegionDCL-fixed margin	0.515±0.042	0.303±0.040	0.872±0.020
RegionDCL	**0.498±0.038**	**0.294±0.047**	**0.879±0.021**

Experiment

- 根据区域中的POI数目和building数目将区域分为4组
- 发现RegionDCL在土地利用推理和人口密度估计任务上始终达到最低的预测误差。在建筑物和POI少于76个的区域，这种性能优势尤为突出





谢谢大家~