



基于城市异构图对比学习的多元地图实体表征学习

杨宜凡

yfyang@buaa.edu.cn

论文链接: <https://arxiv.org/abs/2402.06135>





研究背景



城市时空数据与应用

城市信息基础设施



GPS



RFID



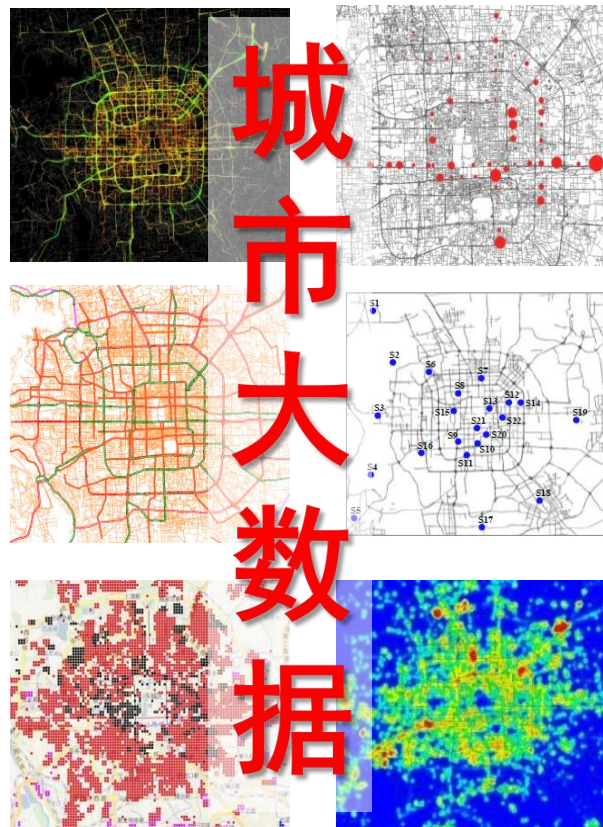
智能手机



LBS



可穿戴



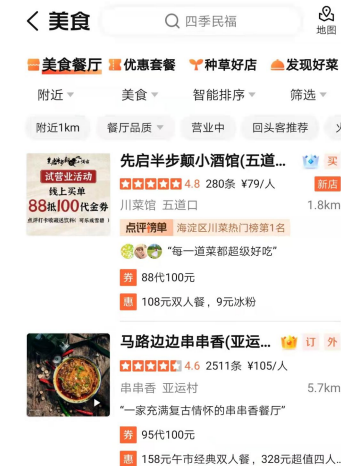
城市大数据



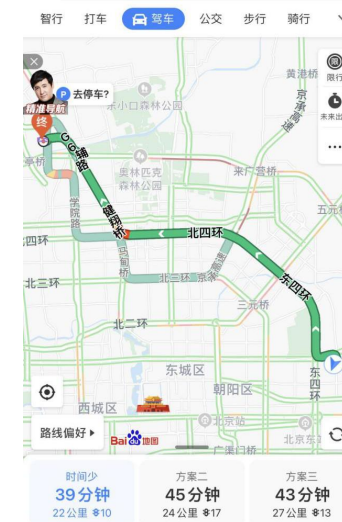
交通状况预测



需求预测与调度

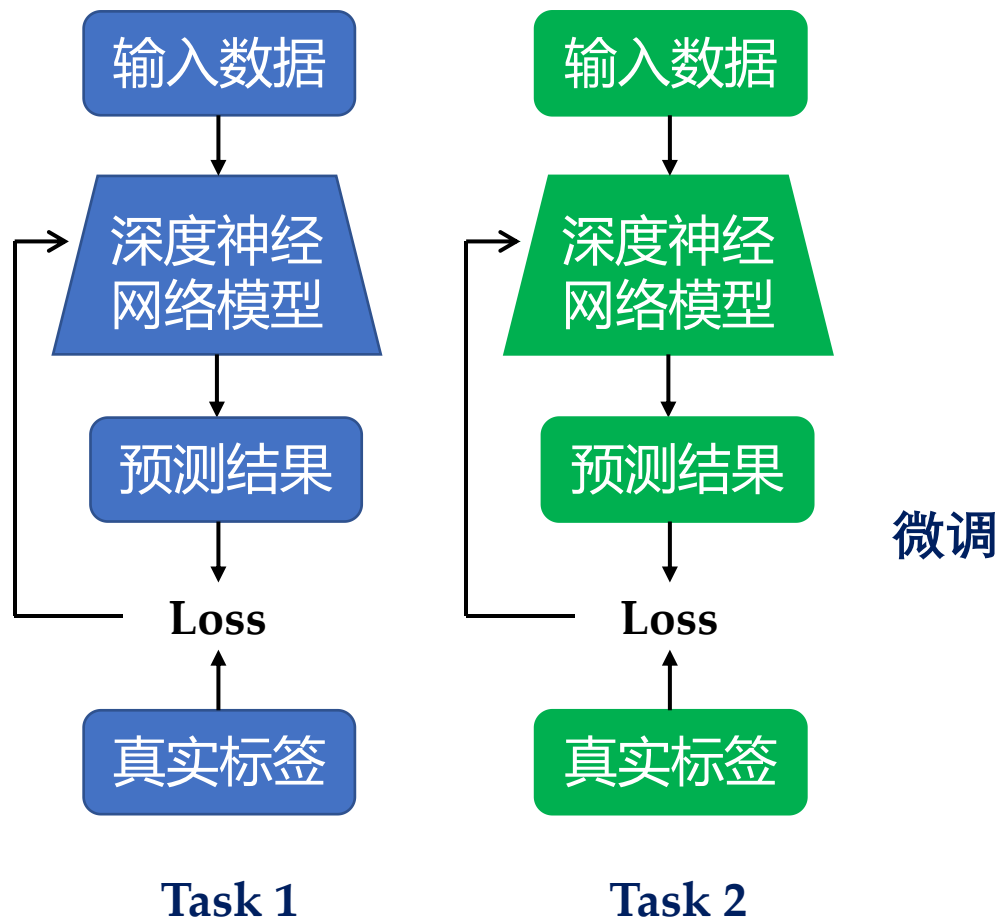


POI 推荐

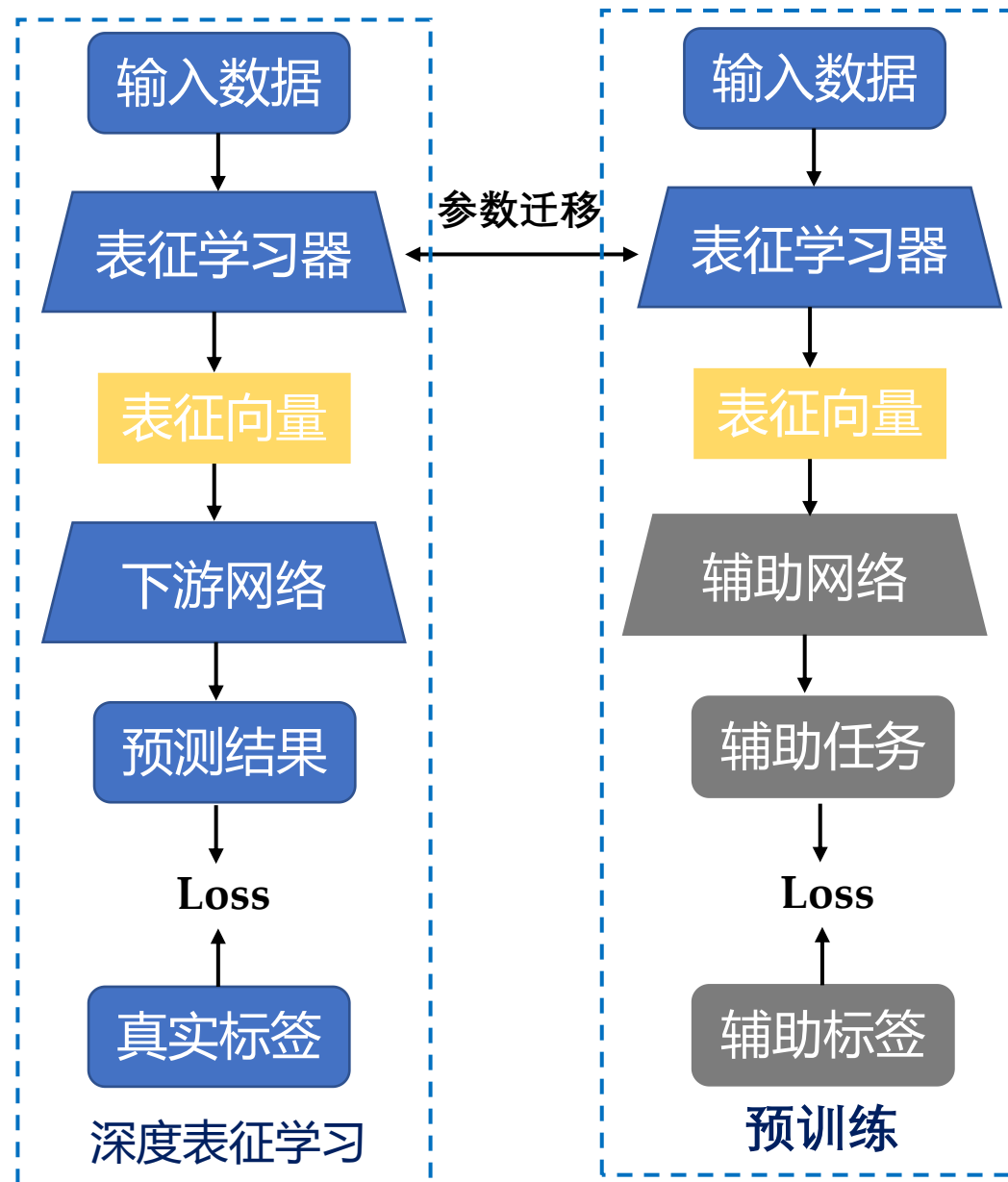


出行时间预测

深度表征学习与自监督学习



神经网络模型的计算过程 (有监督训练)



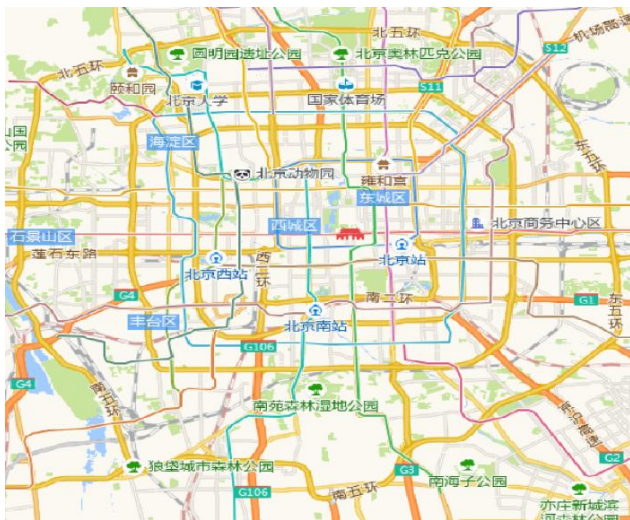
□ 自监督的方法主要包括生成式、预测式和对比式方法

- 生成式：编码器将输入 x 编码成 z ，解码器从 z 重构 x
- 预测式：可以看做是生成式，根据输入 x 构造标签，从 z 预测这个标签
- 对比式：构造正负样本，使学习到的表征保持原始数据的相似/不相似关系

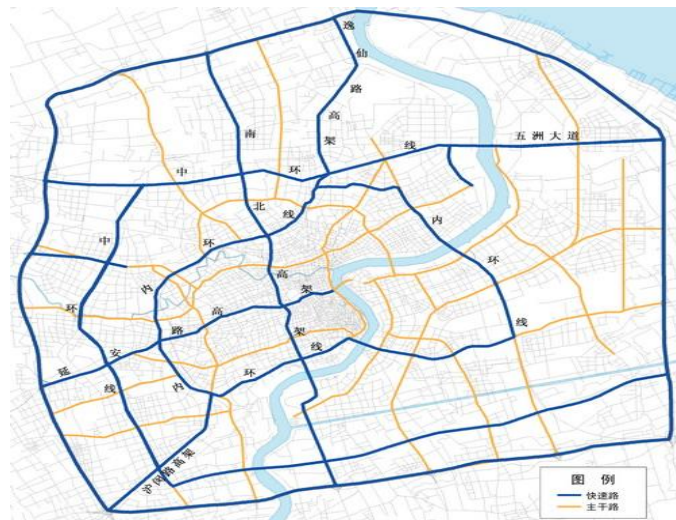
□ 不同领域的研究现状

- 计算机视觉：Deep InfoMax (ICLR 2019), SimCLR (ICML 2020), MoCo (CVPR 2020), SwAV (NIPS 2020), BYOL (NIPS 2020)...
- 自然语言处理：GPT (OpenAI 2018), BERT (NAACL 2019), TSDAE (EMNLP 2021), ConSERT (ACL 2021), SimCSE (EMNLP 2021)...
- 图数据：DGI (ICLR 2019), GraphCL (NIPS 2020), InfoGraph (ICLR 2020), GCA (WWW 2021)...
- 时间序列：TS-TCC (IJCAI 2021), TS2Vec (AAAI 2022), CoST (ICLR 2022)...

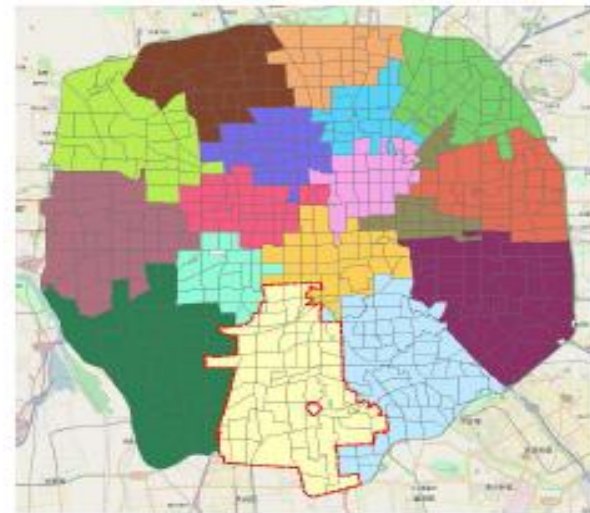
- 城市场景数据是低频、静态的城市结构数据，最直接的体现就是**城市地图数据**。
- 城市地图数据中最常见的两类就是路段和地块数据，城市中的路段构成交通路网，而交通路网又把城市地图划分成了城市地块。



城市地图



交通路网



城市地块

□ 地图实体表征学习

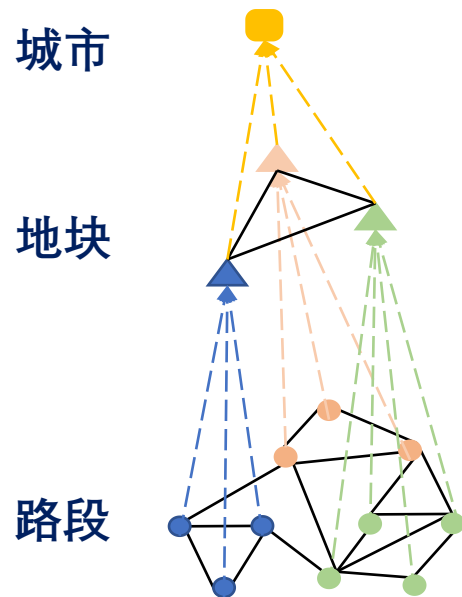
- 给定异构地图实体图，构建一个自监督模型作为嵌入函数，将保留路段和地块的异构性和不同实体之间的多视角关系，将异构图中的路段和地块均转化成表征向量。
- 自监督的方法学习到的表征向量可在各种下游任务中进行应用，包括基于路段的应用、基于地块的应用和基于轨迹的应用。

■ **单一实体**：过去的方法都是针对一类特定的地图实体进行设计的，真实的地图包含多个类别的时空实体。仅学习一种类型地图实体的表征无法满足多样化和复杂化的基于地图的应用的需求，而多个实体联合学习又存在**异构性**。

■ **关联性**：不同实体类别之间的相互关系也包含有价值的结构和语义信息，针对单个实体类别的方法必然会失去这些实体间的相互关系。



城市路段和地块

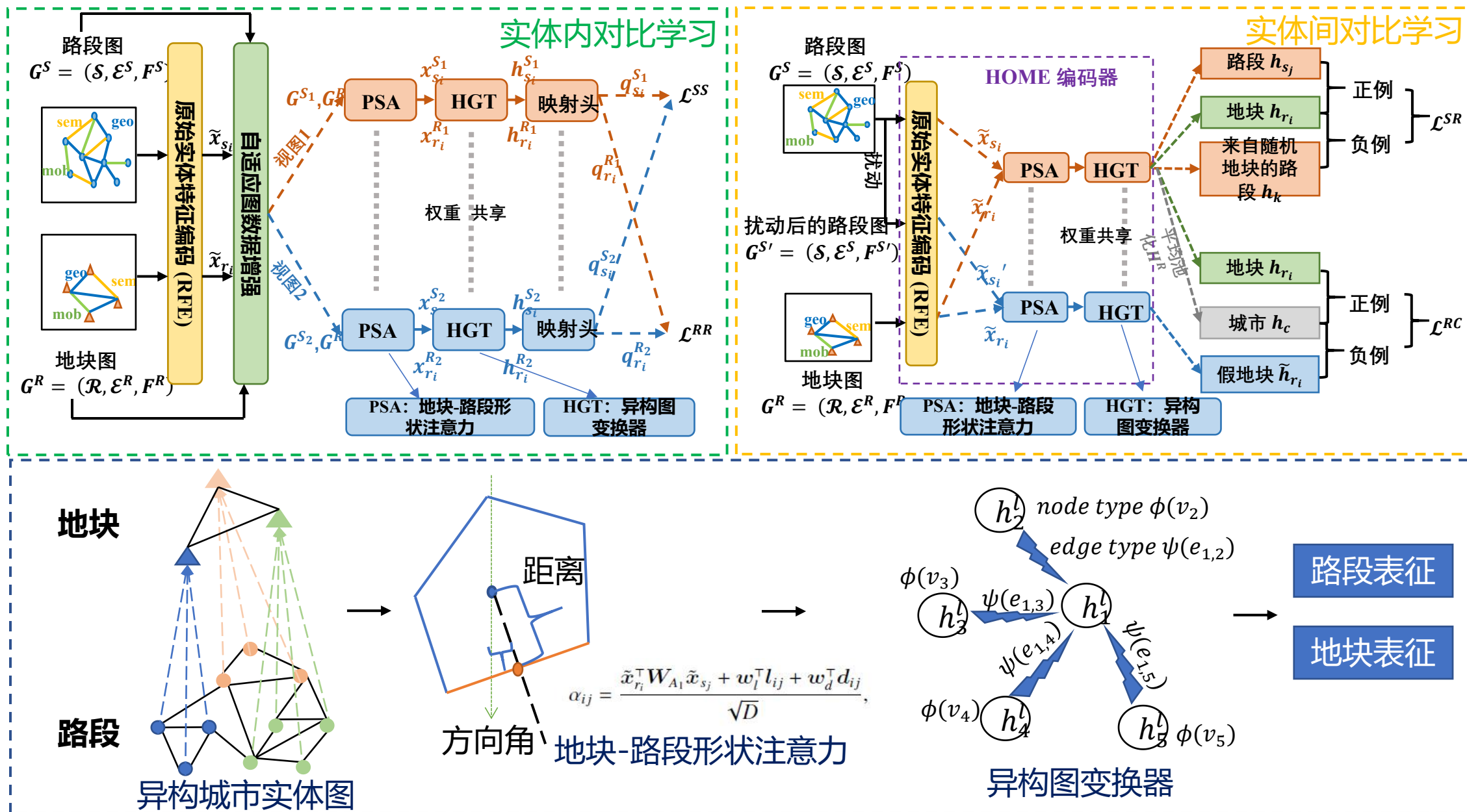




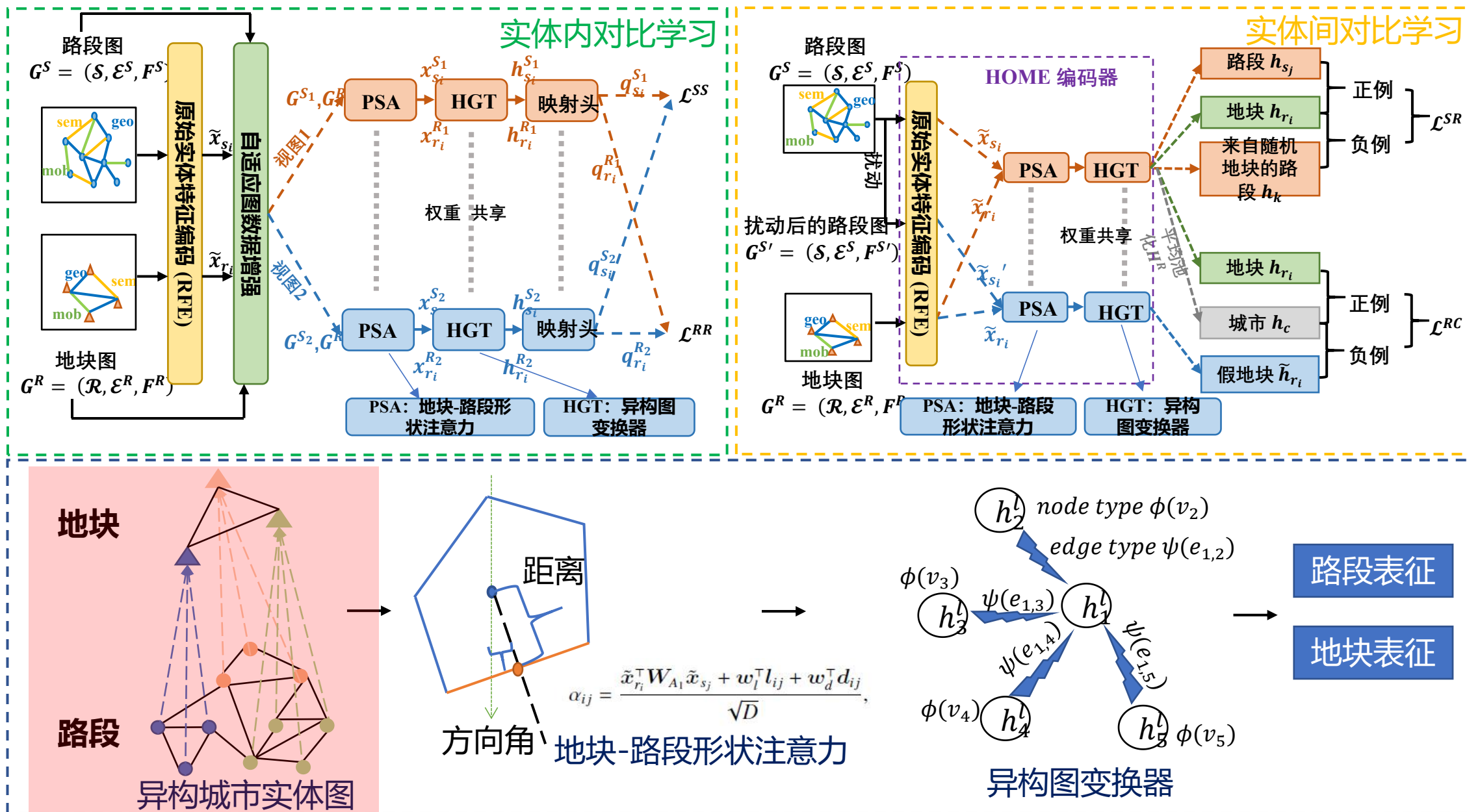
研究方法



框架结构



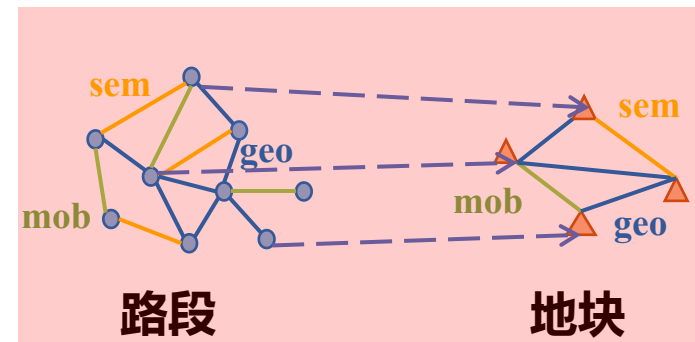
框架结构



多视角内部实体图

衡量同一类地图实体之间的关系

- 地理图：描述地图实体之间的地理关系（距离&拓扑）
- 语义图：描述地图实体之间的功能相似性（POI语义相似性）
- 移动图：描述地图实体之间的人类移动性关系（轨迹数据统计）

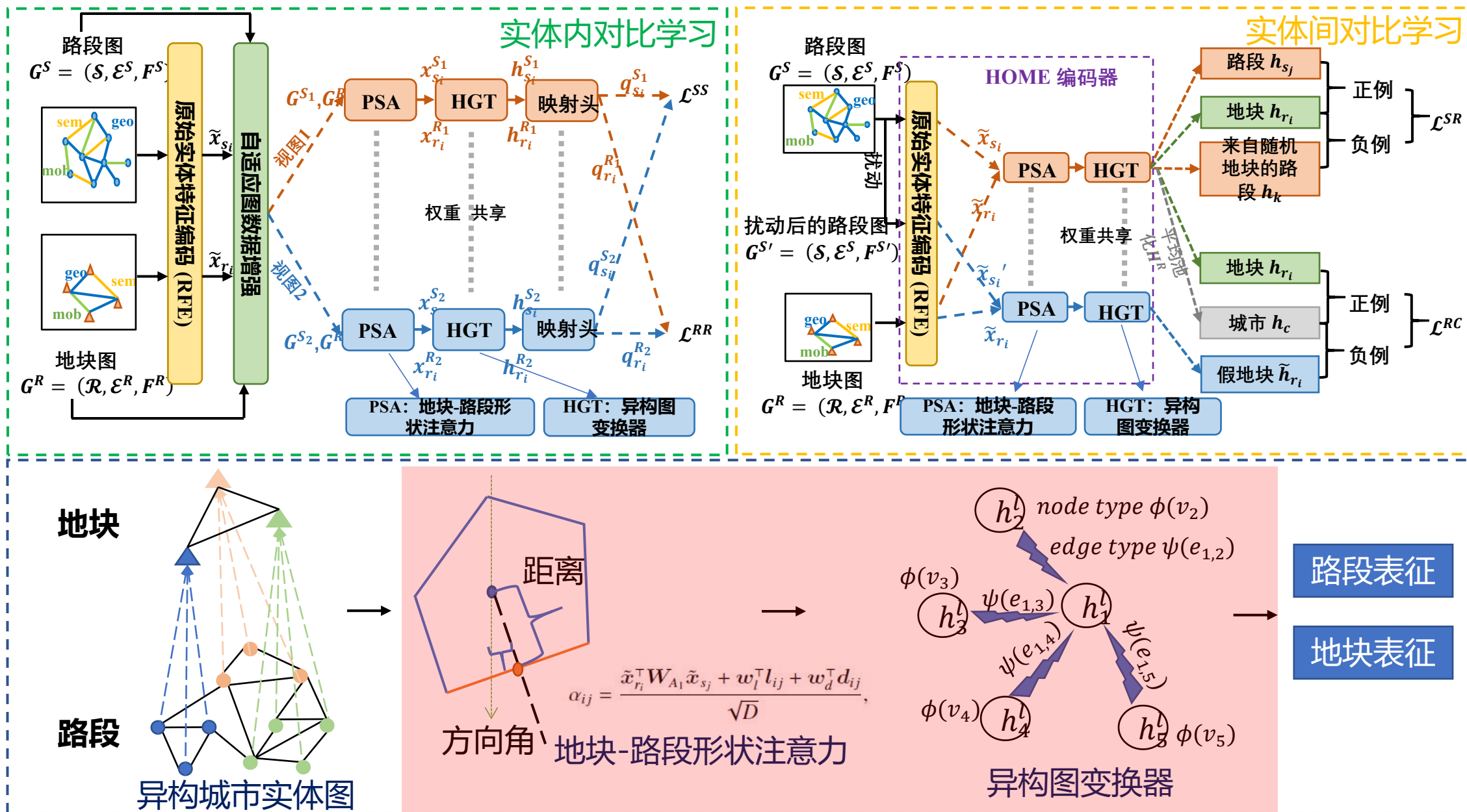


异构地图实体图

衡量不同类地图实体之间的关系：路段-地块映射关系（距离匹配）

- 异构地图实体图：节点集合为路段和地块的并集，边集合包括7类边（三类路段内边和地块内边，加上一类跨实体边）

框架结构



模型：地块-路段形状注意力

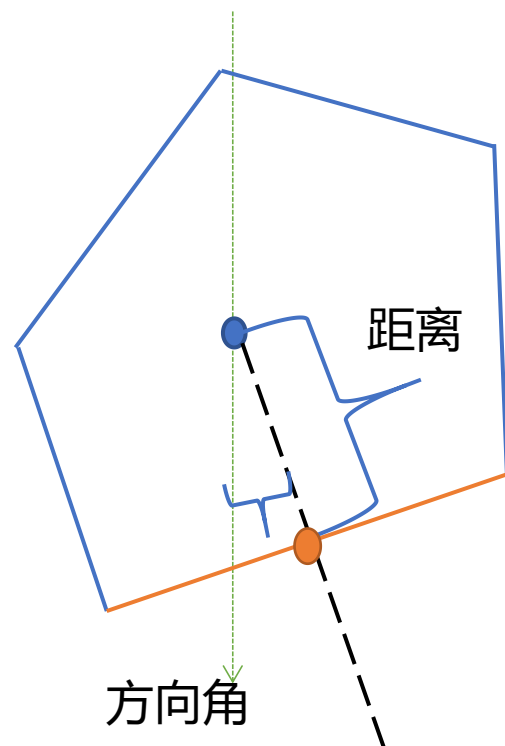
- 除了原始的特征之外，几何结构也是地图实体的重要特征。**路段环绕着地块的边界**，可以利用地块周围的路段的表征对地块的表征进行进一步的完善。
- 使用注意力机制计算地块周围路段对地块的影响，注意力的计算中考虑到地块的不规则多边形结构，引入二者质心的“**距离**”和“**方向角**”的信息。

$$ATT_{ij} = \frac{\exp(\alpha_{ij})}{\sum_{s_j \in S_{r_i}} \exp(\alpha_{ij})},$$

$$\alpha_{ij} = \frac{\tilde{x}_{r_i}^\top W_{A1} \tilde{x}_{s_j} + w_l^\top l_{ij} + w_d^\top d_{ij}}{\sqrt{D}},$$

$$a_{r_i} = \sum_{s_j \in S_{r_i}} ATT_{ij} W_{A2} \tilde{x}_{s_j},$$

$$x_{r_i} = W_{A3} a_{r_i} + \tilde{x}_{r_i},$$



不规则多边形地块及周围的路段

- HGT 模块采用节点编码器 $f_n(\cdot)$ 和边编码器 $f_e(\cdot)$ 来处理**节点和边的异构性**:

$$f_n(v_i) = \mathbf{W}_{\phi(v_i)} \mathbf{x}_{v_i},$$

$$f_e(e_{ij}) = e_{ij} \cdot \mathbf{W}_{\psi(e_{ij})} f_n(v_j),$$

上式中使用了与节点类型 $\phi(v_i)$ 和边类型 $\psi(e_{ij})$ 相关的权重参数。

- 采用注意力机制来建模异构边连接的异构节点对目标节点的影响:

$$\mathbf{q}_{v_i} = \mathbf{W}_q f_n(v_i),$$

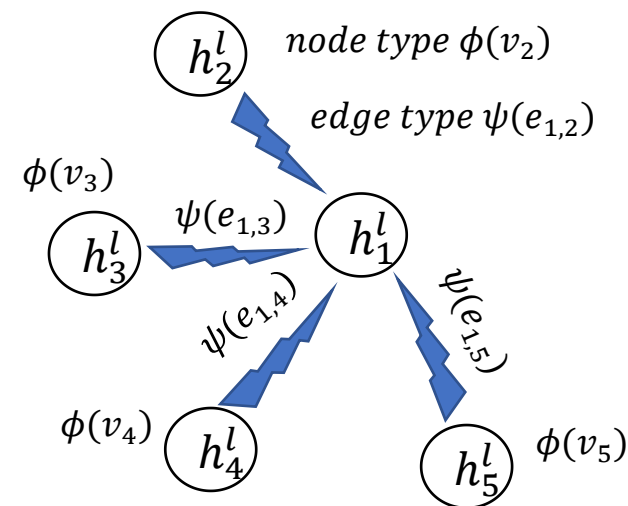
$$\mathbf{k}_{e_{ij}} = \mathbf{W}_k f_e(e_{ij}),$$

$$\mathbf{v}_{e_{ij}} = \mathbf{W}_v f_e(e_{ij}),$$

- 计算目标节点与邻域节点的注意力分数以及对邻域加权聚合后的模型的输出:

$$\alpha_{ij} = \frac{\exp(\mathbf{q}_{v_i}^\top \mathbf{k}_{e_{ij}} / \sqrt{D})}{\sum_{v_{j'} \in \mathcal{N}_i} \exp(\mathbf{q}_{v_i}^\top \mathbf{k}_{e_{ij'}} / \sqrt{D})},$$

$$\mathbf{h}_{v_i} = \sum_{v_j \in \mathcal{N}_i} \alpha_{ij} \mathbf{v}_{e_{ij}} + f_n(v_i),$$



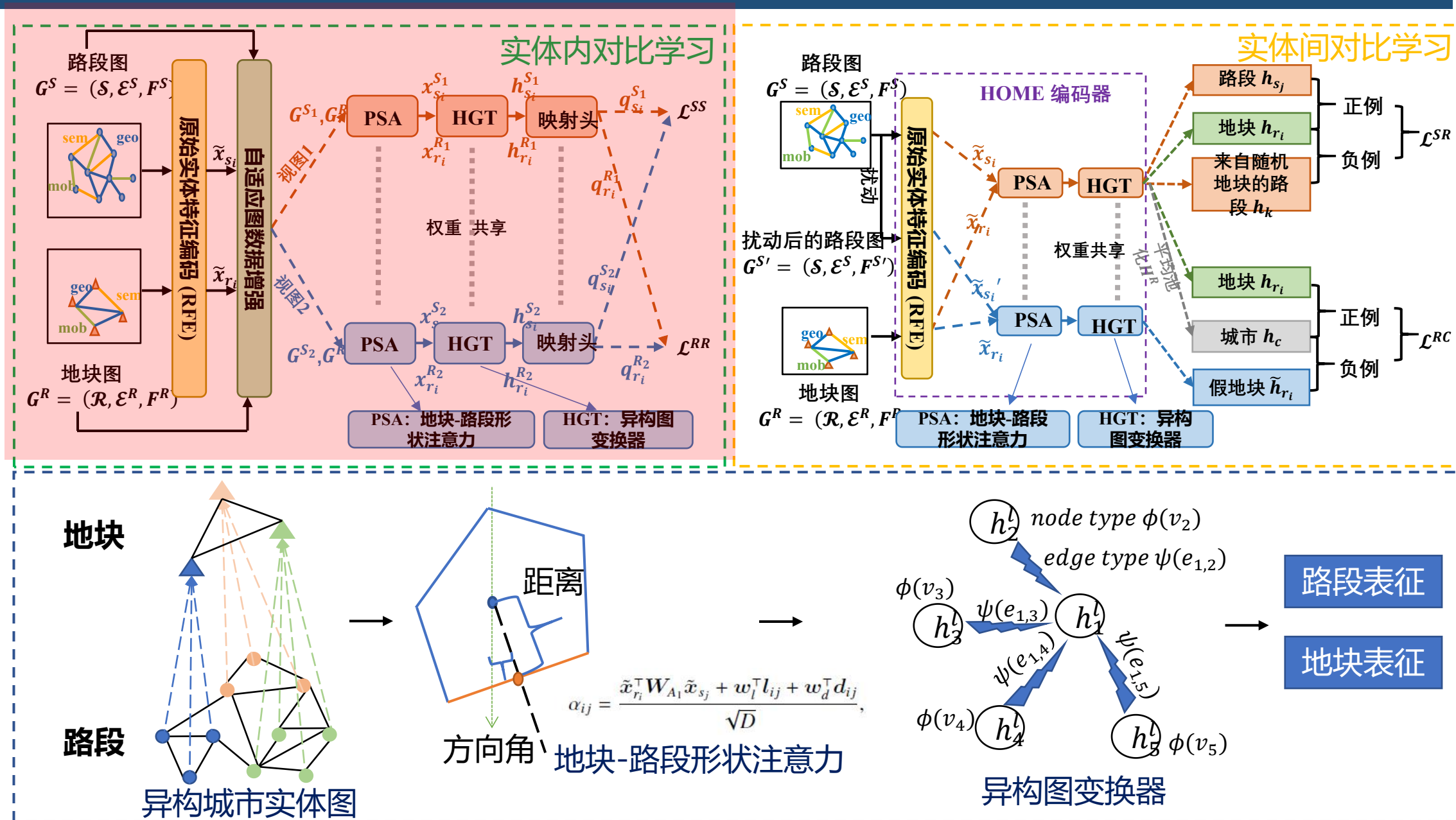
异构图变换器

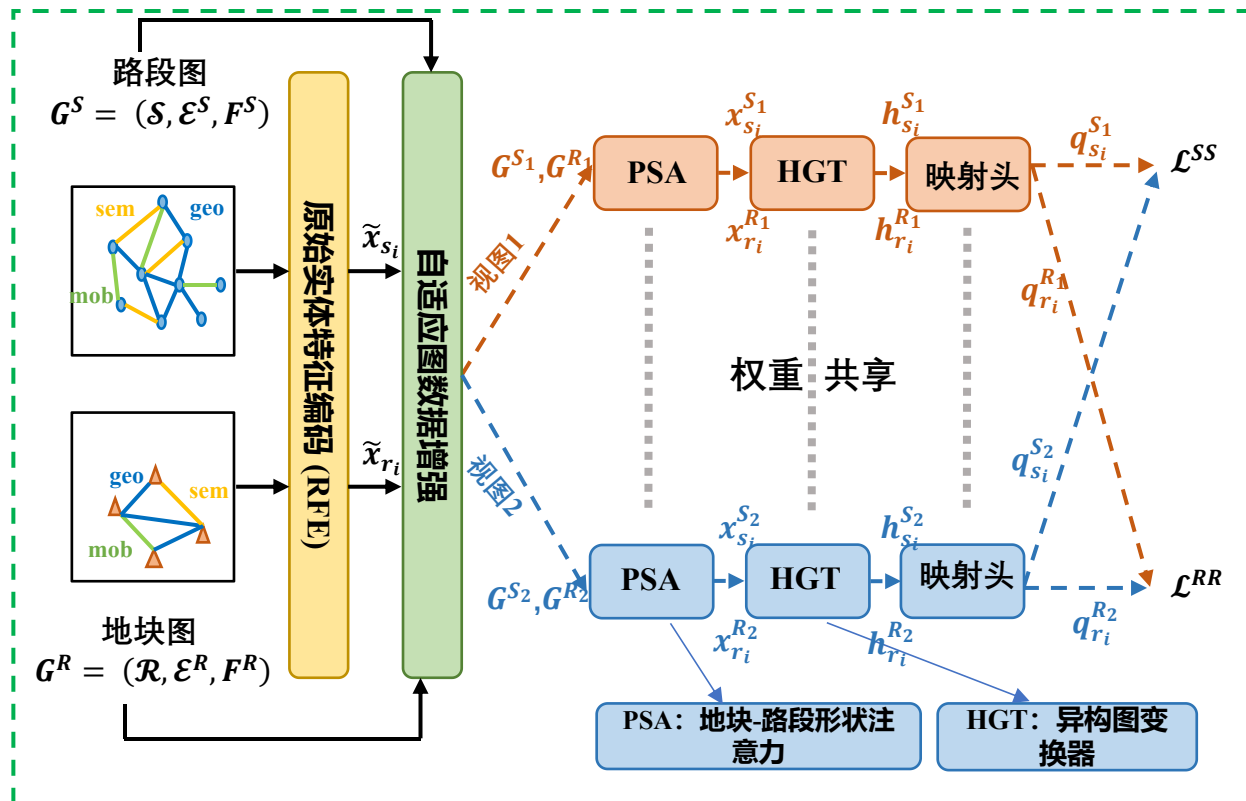


路段表征

地块表征

框架结构





最大化不同数据增强视图之间的表征一致性来学习同一级别实体（地块或路段）的表征。

自适应图数据增强

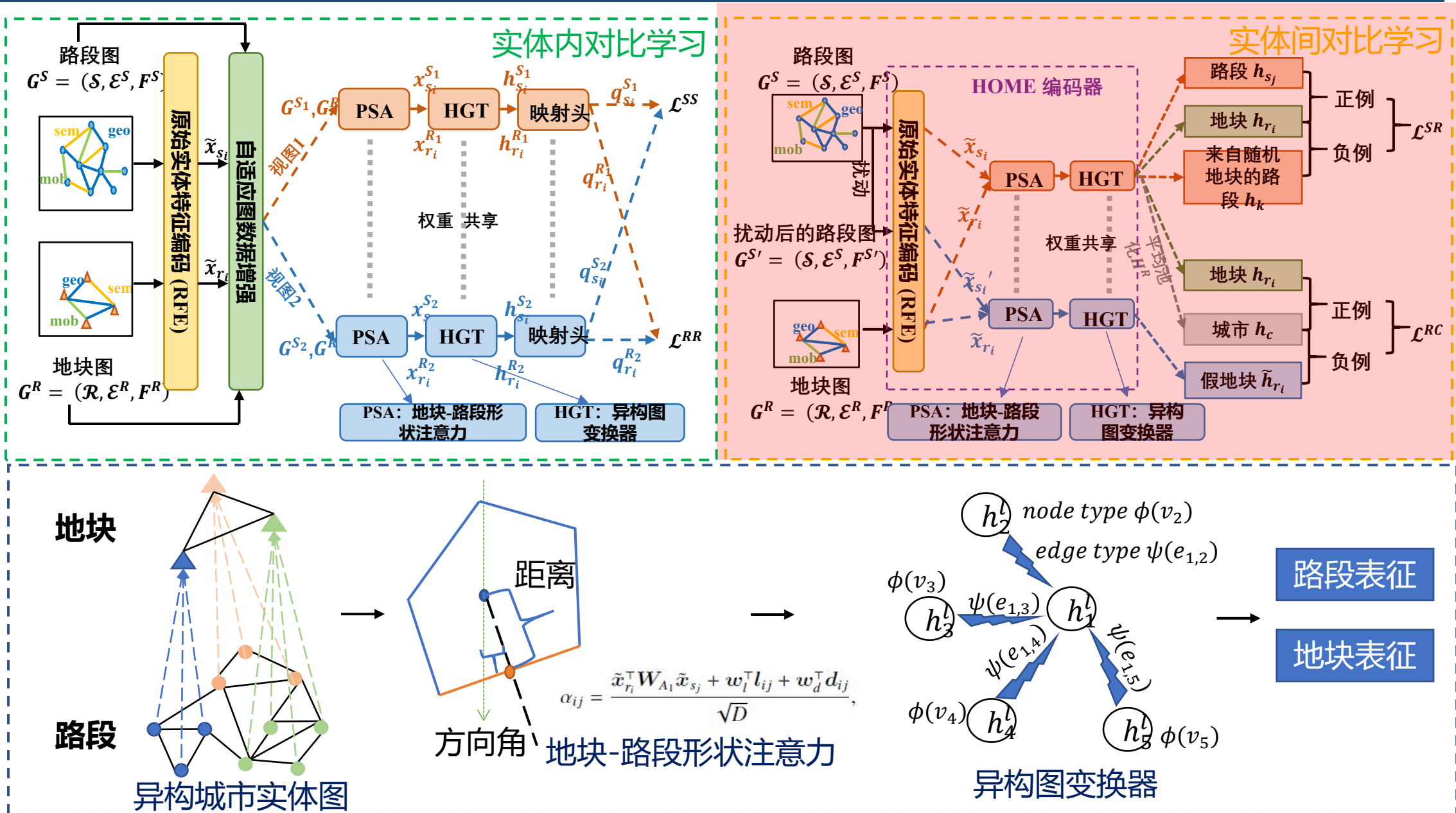
- **边增强**: 使用权重来表示边的重要性，并设置低重要性的边具有较高的删除概率。
- **节点特征增强**: 随机屏蔽节点表征向量中的元素，生成经过数据增强的图。与边增强类似，根据特征的重要性设置不同的屏蔽概率。

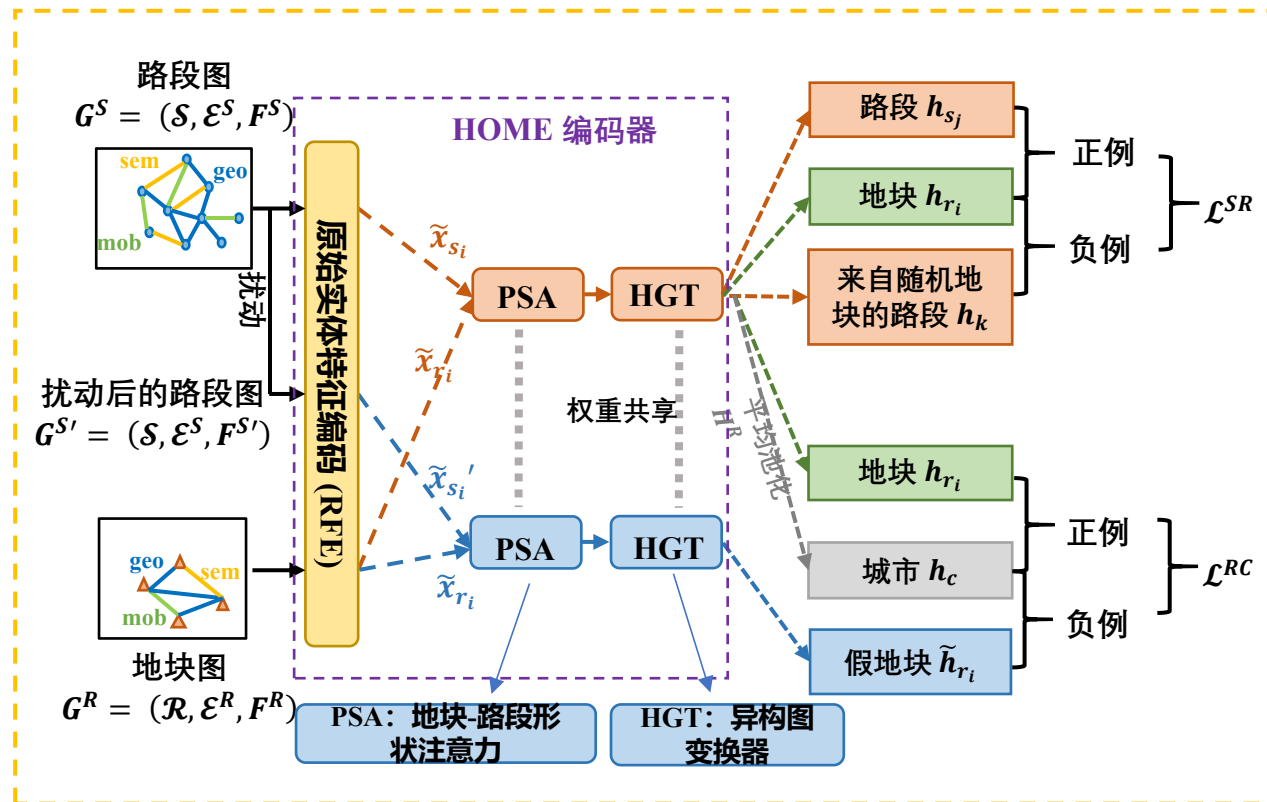
实体内对比学习损失函数

- **正样本**: 同一节点在不同增强视图下的表征
- **负样本**: 不同节点的表征

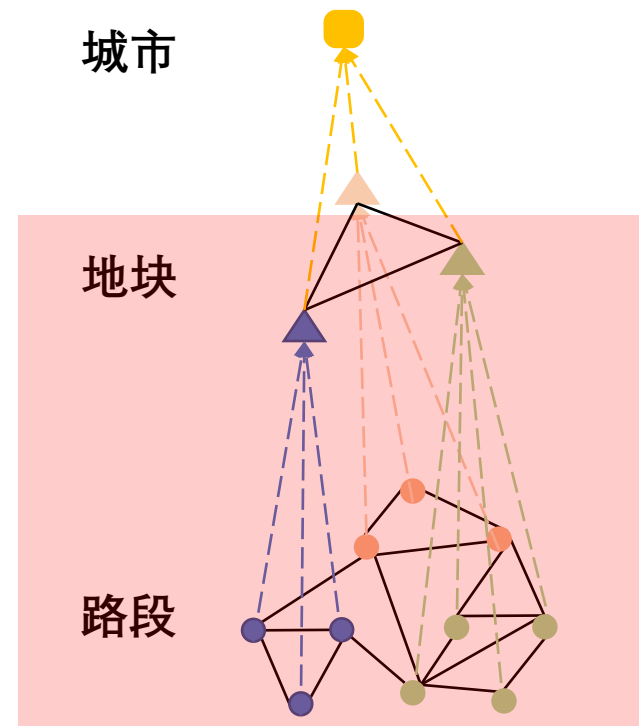
$$\mathcal{L}^{SS}(q_{s_i}^{S1}, q_{s_i}^{S2}) = -\log \frac{\exp(\text{sim}(q_{s_i}^{S1}, q_{s_i}^{S2})/\tau)}{\sum_{s_j \in S} \exp(\text{sim}(q_{s_i}^{S1}, q_{s_j}^{S2})/\tau) + \sum_{s_j \neq s_i} \exp(\text{sim}(q_{s_j}^{S1}, q_{s_j}^{S1})/\tau)}$$

框架结构





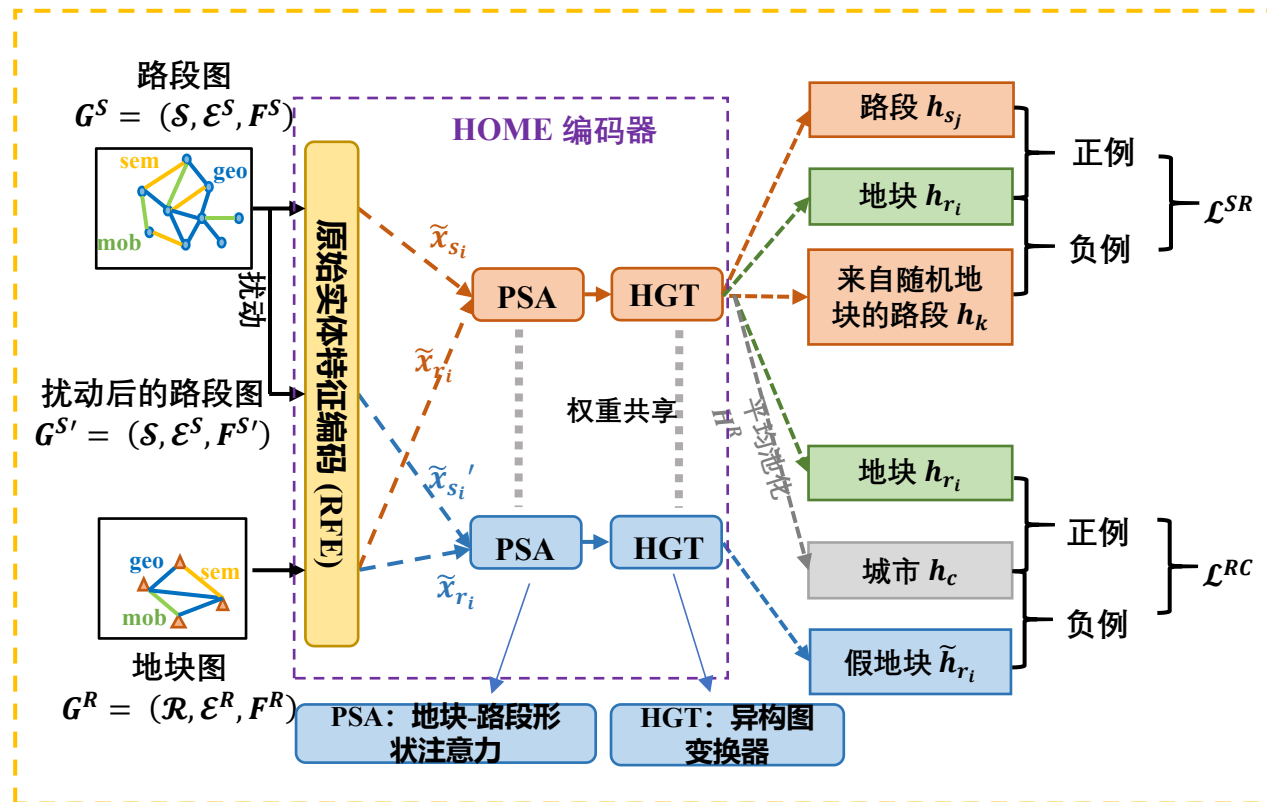
理解路段和地块之间的空间结构，从局部和全局的角度进行对比学习。



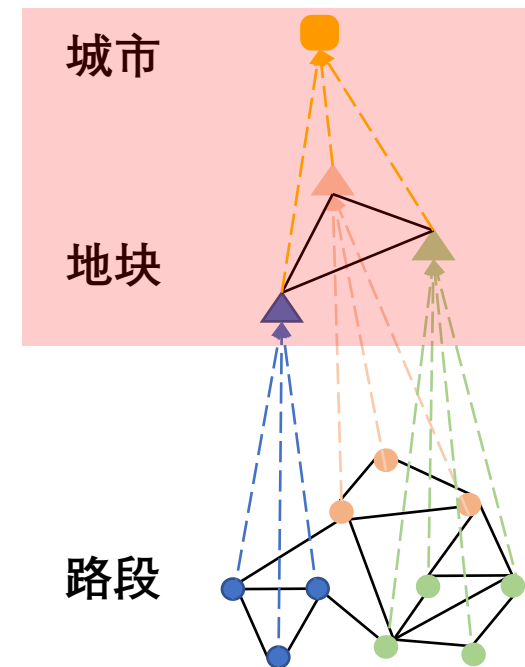
➤ 路段-地块对比学习

- **正样本**：地块以及属于该地块的路段
- **负样本**：地块以及不属于该地块的路段

$$\mathcal{L}^{SR} = - \left(\frac{1}{N_{pos}} \sum_{r_i \in \mathcal{R}} \sum_{s_j \in \mathcal{S}_{r_i}} \log \mathcal{D}(h_{r_i}, h_{s_j}) + \frac{1}{N_{neg}} \sum_{r_i \in \mathcal{R}} \sum_{s_k \in \mathcal{S}_{r'_i}} \log (1 - \mathcal{D}(h_{r_i}, h_{s_k})) \right),$$



理解路段和地块之间的空间结构，从局部和全局的角度进行对比学习。



➤ 地块-城市对比学习

- **正样本:** 城市和真实地块
- **负样本:** 城市和“假地块”，随机打乱路段得到的虚拟地块作为负样本

$$\mathcal{L}^{RC} = - \left(\frac{1}{|\mathcal{R}|} \sum_{r_i \in \mathcal{R}} \log \mathcal{D}(h_{r_i}, h_c) + \frac{1}{|\mathcal{R}|} \sum_{r_i \in \mathcal{R}} \log (1 - \mathcal{D}(\tilde{h}_{r_i}, h_c)) \right),$$



实验结果



- 来自三个城市的大规模数据集：北京（BJ）、西安（XA）和成都（CD）。
- 包含四种类型的数据，即兴趣点数据、路网数据、地块数据和轨迹数据。

表 2 城市场景数据集统计信息表

数据集	BJ	XA	CD
时间范围	2015.11	2018.11	2018.11
轨迹数量	1018312	384618	559729
用户数量	1677	26787	48295
POI 数量	81180	19107	17300
路段数量	40306	5269	6195
地块数量	11208	1056	1306
轨迹最大长度	128	64	64

本工作进行了以下三类下游任务的实验，以评估所学习到的表征：

■ 基于路段的任务

路段分类任务、交通流量预测任务、OD 流量预测任务

■ 基于地块的任务

地块分类任务、交通流量预测任务、OD 流量预测任务

■ 基于轨迹的任务

出行时间预测任务、最相似轨迹搜索任务

■ 图嵌入方法

- DeepWalk (KDD 2014)
- LINE (WWW 2015)
- Node2Vec (KDD 2016)

■ 路段表征学习方法

- RFN (SIGSPATIAL 2019)
- SRN2Vec (ACM TIST 2021)
- HRNR (KDD 2020)
- Toast (CIKM 2021)
- JCLRNT (CIKM 2022)
- HyperRoad (ACM TKDD 2023)
- SARN (EDBT 2023)

■ 地块表征学习方法

- HDGE (CIKM 2017)
- ZE-Mob (IJCAI 2018)
- MV-PN (AAAI 2019)
- MVURE (IJCAI 2020)
- Region2Vec (CIKM 2022)
- MGFN (IJCAI 2022)
- ReMVC (IEEE TKDE 2023)

■ 轨迹表征学习方法

- Trajectory2vec (IJCNN 2017)
- T2vec (ICDE 2018)
- Trembr (ACM TIST 2020)

实验：性能比较

表 3 路段级别的任务的性能对比表：第一部分

数据集	BJ			XA			CD		
	分类	流量	OD	分类	流量	OD	分类	流量	OD
模型\指标	Mi-F1	MAE	MAE	Mi-F1	MAE	MAE	Mi-F1	MAE	MAE
DeepWalk	0.364	1.333	0.227	0.441	2.103	2.642	0.339	2.168	2.809
LINE	0.345	1.334	0.213	0.411	2.115	1.912	0.323	2.470	3.015
Node2Vec	0.397	1.282	0.218	0.430	2.349	2.846	0.376	2.149	2.877
RFN	0.383	1.234	0.215	0.432	1.967	1.914	0.389	2.235	3.234
SRN2Vec	0.384	1.126	0.198	0.443	1.800	1.989	0.396	2.265	3.002
HRNR	0.399	1.123	0.219	0.441	1.838	1.702	0.409	2.223	2.825
Toast	0.462	1.166	0.201	0.457	1.822	1.692	0.439	2.193	2.735
JCLRNT	0.497	1.149	0.208	0.493	1.819	1.594	0.443	2.182	2.686
HyperRoad	0.499	1.101	0.104	0.504	1.665	1.585	0.473	2.298	2.696
SARN	0.479	1.062	0.188	0.506	1.582	1.583	0.476	2.092	2.525
HOME-GCL	0.530	1.015	0.095	0.525	1.531	1.536	0.496	1.985	2.365
提升	6.21%	4.43%	8.65%	3.75%	3.22%	2.97%	4.20%	5.11%	6.34%

表 4 路段级别的任务的性能对比表：第二部分

数据集	BJ			XA			CD		
	分类	流量	OD	分类	流量	OD	分类	流量	OD
模型\指标	Ma-F1	RMSE	RMSE	Ma-F1	RMSE	RMSE	Ma-F1	RMSE	RMSE
DeepWalk	0.146	1.942	0.407	0.318	3.246	7.810	0.278	4.156	15.106
LINE	0.123	1.946	0.400	0.281	3.411	7.502	0.261	4.632	15.102
Node2Vec	0.177	1.889	0.403	0.315	4.185	8.031	0.285	4.137	15.089
RFN	0.117	1.773	0.413	0.283	3.213	7.576	0.304	4.453	15.085
SRN2Vec	0.124	1.707	0.395	0.307	3.111	7.459	0.318	4.307	15.098
HRNR	0.181	1.702	0.410	0.302	3.125	7.799	0.322	4.318	15.127
Toast	0.201	1.674	0.394	0.319	2.958	7.505	0.378	4.224	15.086
JCLRNT	0.218	1.753	0.399	0.331	3.083	7.506	0.401	4.123	15.104
HyperRoad	0.226	1.672	0.337	0.343	2.952	7.502	0.442	4.305	15.081
SARN	0.221	1.666	0.399	0.344	2.949	7.444	0.456	4.115	15.131
HOME-GCL	0.236	1.619	0.302	0.381	2.813	7.195	0.476	3.938	14.608
提升	4.42%	2.82%	10.39%	10.76%	4.61%	3.34%	4.39%	4.30%	3.14%

表 5 地块级别的任务的性能对比表：第一部分

数据集	BJ			XA			CD		
	分类	流量	OD	分类	流量	OD	分类	流量	OD
模型\指标	Mi-F1	MAE	MAE	Mi-F1	MAE	MAE	Mi-F1	MAE	MAE
DeepWalk	0.261	2.812	0.375	0.353	4.992	4.864	0.323	6.633	5.833
LINE	0.231	3.051	0.391	0.329	5.016	5.517	0.215	7.626	6.236
Node2Vec	0.287	2.723	0.373	0.391	4.902	4.843	0.344	6.167	6.105
HDGE	0.278	2.350	0.372	0.362	4.091	4.783	0.365	6.123	6.247
ZE-Mob	0.287	3.020	0.360	0.440	3.870	4.886	0.413	5.836	6.618
MV-PN	0.296	2.791	0.358	0.387	3.907	4.912	0.334	6.124	5.916
MVURE	0.327	2.344	0.367	0.422	3.591	4.517	0.425	5.331	5.860
Region2Vec	0.311	2.291	0.391	0.431	3.623	4.551	0.423	5.213	5.931
MGFN	0.344	2.284	0.352	0.405	3.698	4.546	0.361	5.419	5.987
ReMVC	0.339	2.279	0.383	0.442	3.584	4.512	0.432	4.976	5.811
HOME-GCL	0.376	2.087	0.336	0.466	3.449	4.377	0.462	4.691	5.637
提升	9.30%	8.42%	4.55%	5.43%	3.77%	2.99%	6.94%	5.73%	2.99%

表 6 地块级别的任务的性能对比表：第二部分

数据集	BJ			XA			CD		
	分类	流量	OD	分类	流量	OD	分类	流量	OD
模型\指标	Ma-F1	RMSE	RMSE	Ma-F1	RMSE	RMSE	Ma-F1	RMSE	RMSE
DeepWalk	0.245	4.721	0.772	0.256	7.836	27.152	0.212	11.753	27.921
LINE	0.219	4.923	0.797	0.183	9.062	27.459	0.203	12.825	27.941
Node2Vec	0.256	4.623	0.753	0.279	7.441	26.203	0.273	10.032	28.087
HDGE	0.167	4.103	0.750	0.219	6.766	17.725	0.242	10.979	27.322
ZE-Mob	0.162	4.956	0.790	0.260	6.903	17.947	0.335	10.903	27.629
MV-PN	0.243	4.652	0.771	0.279	6.702	18.021	0.289	10.758	27.932
MVURE	0.221	4.402	0.723	0.283	6.481	17.921	0.326	10.240	27.523
Region2Vec	0.243	4.123	0.738	0.302	6.479	17.329	0.339	10.198	27.623
MGFN	0.309	4.097	0.724	0.297	6.478	17.812	0.309	10.267	27.998
ReMVC	0.313	4.089	0.722	0.311	6.483	17.033	0.342	9.772	27.396
HOME-GCL	0.343	3.764	0.686	0.331	6.238	16.408	0.377	9.402	26.461
提升	9.58%	7.95%	4.99%	6.43%	3.70%	3.67%	10.23%	3.79%	3.41%

在路段和地块级别的6个子任务上都取得了最优性能。

表 7 轨迹级别的任务的性能对比表：出行时间预测

数据集	BJ			XA			CD		
模型\指标	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
trajectory2vec	10.130	56.83	37.95	2.045	3.146	35.14	1.634	2.432	34.74
t2vec	10.030	56.65	36.42	2.028	3.126	33.73	1.629	2.426	34.45
Trember	9.997	55.97	34.20	2.002	3.121	32.13	1.618	2.401	34.15
Toast	10.681	47.73	52.63	2.154	3.264	33.93	1.710	2.491	37.24
JCLRNT	10.218	46.49	49.18	2.166	3.246	33.12	1.690	2.481	36.42
DeepWalk	11.737	57.67	96.72	2.338	3.365	37.53	1.862	2.554	35.12
LINE	11.843	58.65	96.43	2.462	3.455	38.37	1.999	2.795	36.77
Node2Vec	11.707	57.64	94.74	2.337	3.346	36.73	1.836	2.536	34.03
RFN	11.679	51.92	53.03	2.196	3.158	32.68	1.643	2.461	32.43
SRN2Vec	10.944	50.83	50.11	2.173	3.175	33.24	1.765	2.580	38.53
HRNR	10.418	51.10	52.13	2.154	3.185	33.51	1.671	2.413	34.90
HyperRoad	10.172	39.11	39.37	1.926	3.056	31.63	1.604	2.394	33.13
SARN	9.446	36.97	32.59	2.087	3.060	34.65	1.594	2.405	31.89
HOME-GCL	8.940	33.83	30.37	1.863	2.892	29.32	1.469	2.271	30.35
提升	5.36%	8.49%	6.81%	3.27%	5.37%	7.30%	7.84%	5.14%	4.83%

表 8 轨迹级别的任务的性能对比表：最相似轨迹搜索

数据集	BJ			XA			CD		
模型\指标	MR	HR@5	HR@20	MR	HR@5	HR@20	MR	HR@5	HR@20
trajectory2vec	40.62	0.419	0.675	20.22	0.601	0.783	13.77	0.568	0.810
t2vec	31.01	0.483	0.741	19.56	0.621	0.893	12.82	0.636	0.865
Trember	26.50	0.514	0.773	18.74	0.631	0.808	10.89	0.642	0.870
Toast	27.75	0.505	0.705	21.37	0.646	0.802	13.20	0.645	0.868
JCLRNT	23.15	0.5362	0.806	11.18	0.651	0.870	12.76	0.653	0.870
DeepWalk	41.59	0.469	0.703	19.80	0.562	0.811	21.45	0.592	0.853
LINE	77.82	0.4536	0.633	45.15	0.557	0.730	31.68	0.569	0.784
Node2Vec	32.65	0.487	0.727	18.17	0.564	0.819	19.21	0.639	0.862
RFN	75.67	0.451	0.640	22.95	0.607	0.806	10.92	0.661	0.871
SRN2Vec	39.72	0.4882	0.705	13.48	0.629	0.849	26.05	0.590	0.804
HRNR	43.13	0.4912	0.703	16.53	0.606	0.825	17.45	0.610	0.851
HyperRoad	35.39	0.445	0.697	12.36	0.624	0.860	9.79	0.654	0.865
SARN	18.04	0.557	0.826	10.77	0.664	0.879	9.54	0.663	0.877
HOME-GCL	16.83	0.610	0.871	8.59	0.707	0.911	8.53	0.691	0.907
提升	6.71%	9.52%	5.45%	20.24%	6.48%	3.64%	10.59%	4.22%	3.42%

在轨迹级别的两个子任务上取得了最优性能。

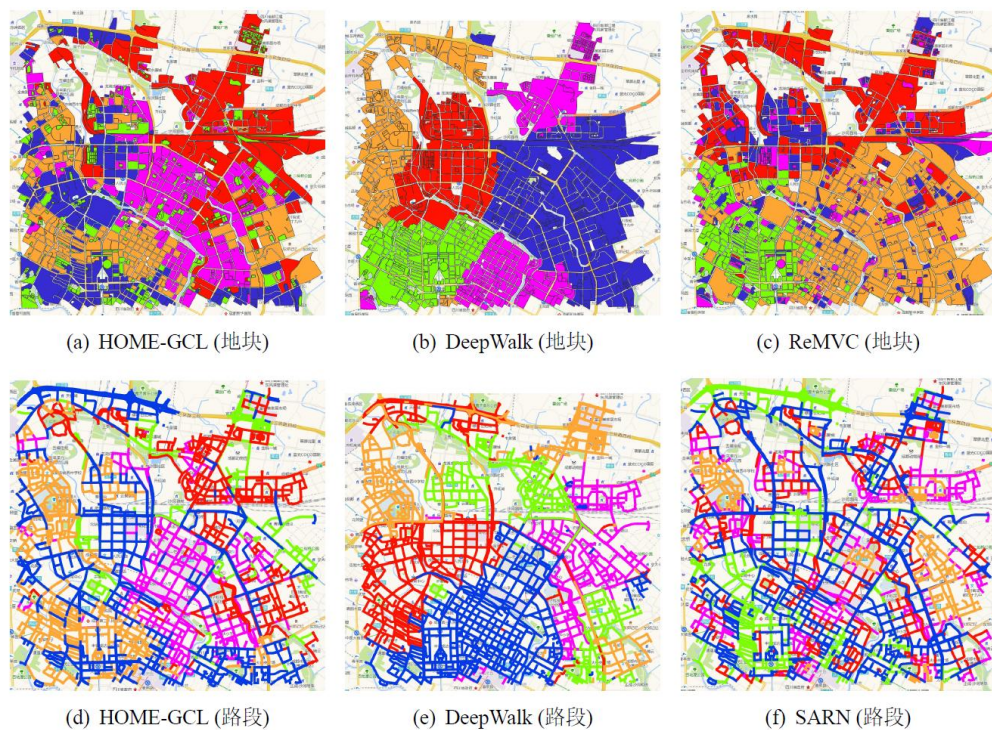


图 8 路段和地块表征聚类结果

学习表征的聚类实验：HOME-GCL 模型的聚类结果在不同类别实体的地理分布上显示出非常高的一致性，即空间上相邻的地块对应的路段也被聚类在一起。

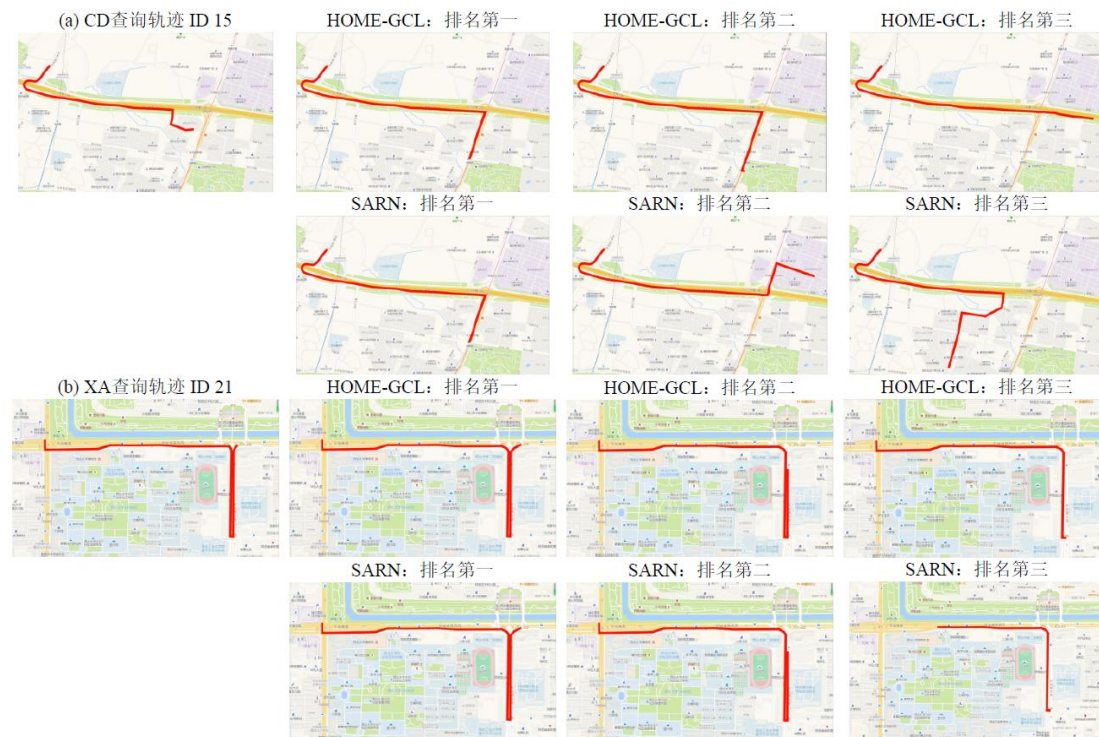


图 9 HOME-GCL 和 SARN 模型检索到的前 3 个相似轨迹的比较

前三条相似轨迹的比较：HOME-GCL有效地发现了具有相似整体特征的轨迹。相比之下，SARN检索出的轨迹与查询轨迹偏离较大

实验：消融实验&参数敏感性分析

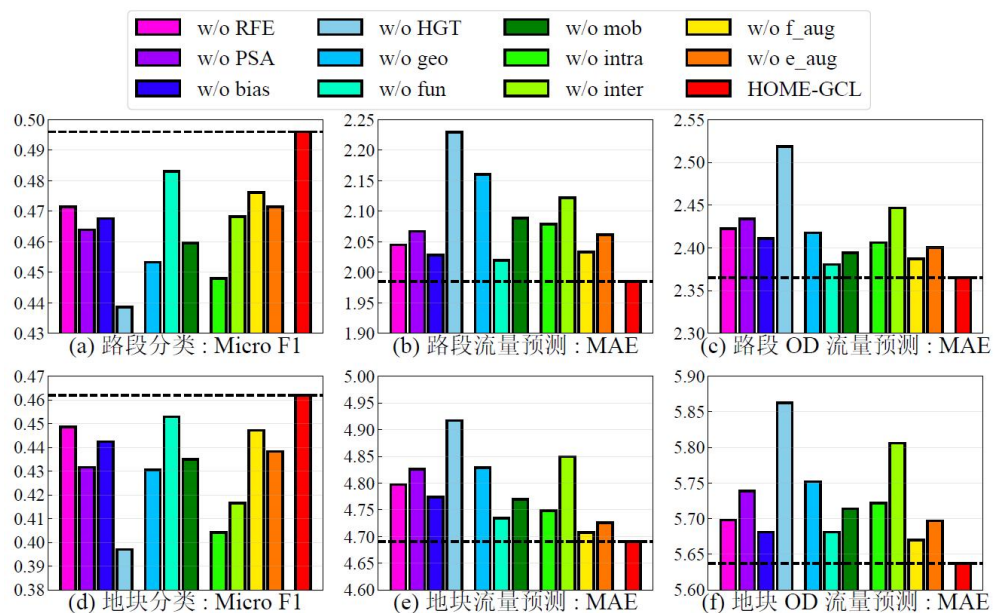


图5 CD数据集的消融实验

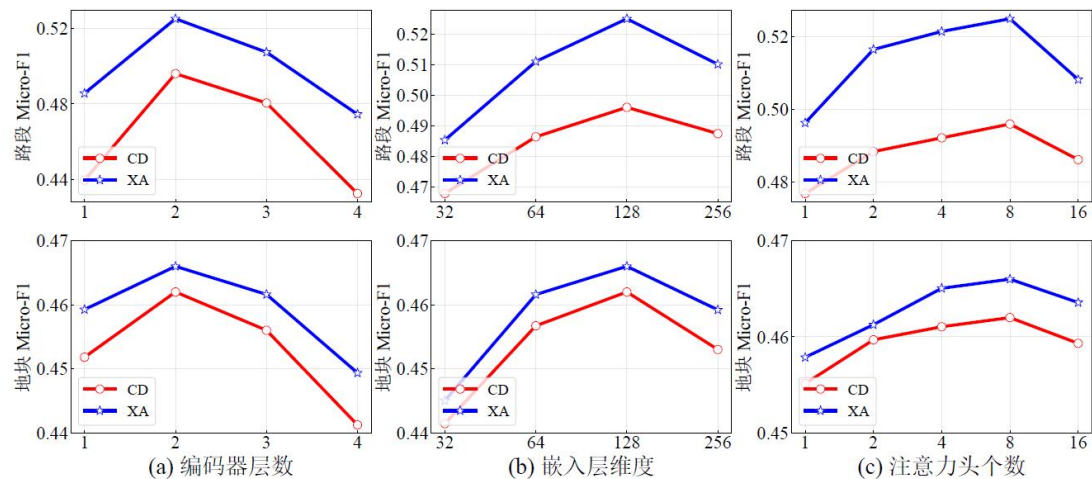


图6 CD和XA数据集的参数敏感性实验

➤ 消融实验：证明了所提出的模型的基础子模块的影响、多视角内部实体图构建的影响、对比学习任务的影响等。

➤ 敏感性分析实验：重要参数对模型效果的影响。

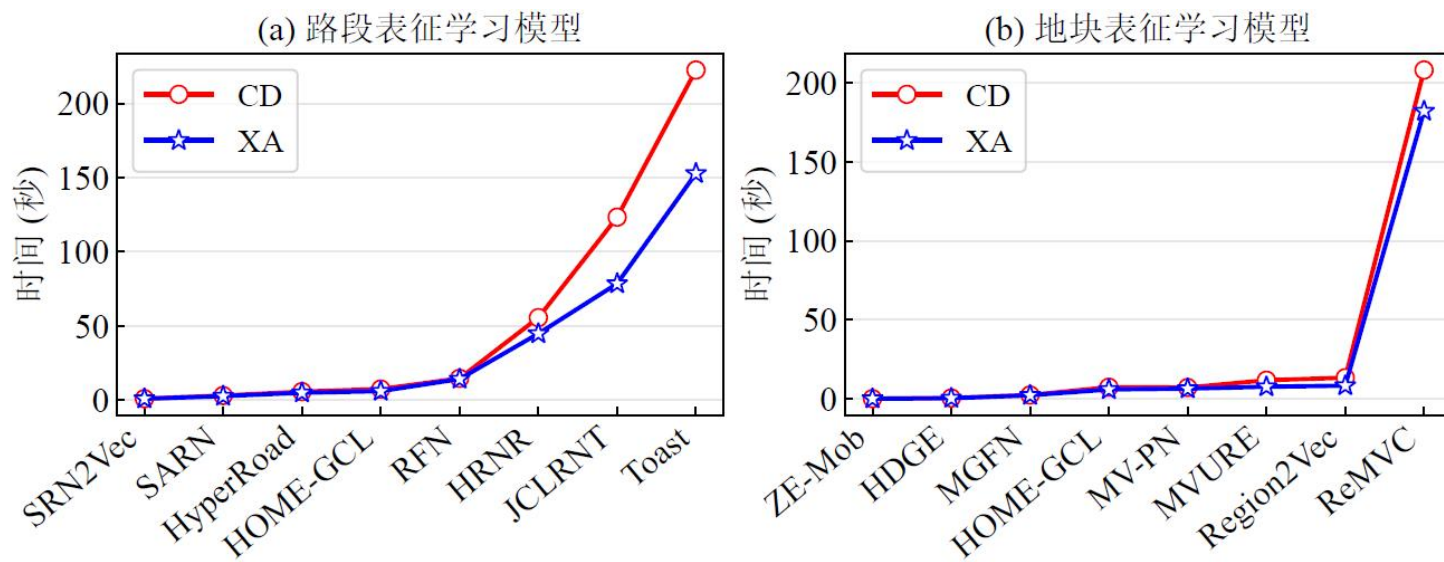


图 7 XA 和 CD 数据集上模型训练时间对比

- 一些较简单的模型，如HDGE、ZE-Mob 和SRN2Vec，训练时间较短，但它们的性能也较差。
- 尽管HOME-GCL 模型的训练时间略长于一些性能较好的基准模型，如HyperRoad、SARN 和MGFN，但HOME-GCL 模型同时获得路段和地块级别的表征。

- **论文工作总结：** 本文针对**城市地图数据**设计了一个通用且有效的表征学习方法，以自监督的方案学习的表征可以应用到多样的下游任务中，为基于城市时空数据的下游任务提供通用的解决方案。研究侧重于如何联合城市中的两个实体即路段和地块来共同学习城市场景数据的表征。
- **未来方向展望：**
 - 利用多源的数据辅助训练：POI 数据、卫星图像数据、社交媒体数据等。
 - 地图表征技术（tokenization）与向量数据库，大语言模型（Large Language Model, LLM）技术结合。



感谢聆听!

杨宜凡

yfyang@buaa.edu.cn

论文链接: <https://arxiv.org/abs/2402.06135>

