Scene-Level Vector Spatial Representation Learning for Map Understanding

Zhe Wang [1], Zhaonan Wang [2]* and Takahiro Yabe [3]

[1] Department of Computer Science and Engineering, New York University, New York, United States
[2] Division of Arts and Sciences, New York University Shanghai, Shanghai, China
[3] Center for Urban Science and Progress, New York University, New York, United States

* zhaonan.wang@nyu.edu

Keywords: graph neural networks, map understanding, spatial representation learning, vector data

Introduction

Vector spatial representation learning (VSRL) refers to transforming vector-based spatial data into machine-readable representations for spatial models to understand geographical space. As vector data are increasingly relied upon to model real-world environments (Mai et al., 2024), the quality of VSRL fundamentally determines how effectively spatial entities and spatial information are utilized in downstream tasks.

Compared to raster-based methods, VSRL directly models spatial entities in a resolution-independent manner, preserving precise geometric structures. Therefore, despite the success of many high-quality raster-based methods and data products (e.g., AlphaEarth (Brown et al., 2025), SkySense (Guo et al., 2024)), VSRL remains indispensable for tasks that require fine-grained spatial representation, such as road network modeling, parcel-level land use analysis, and trajectory pattern recognition.

In terms of representation scale, VSRL can be categorized into entity-level, pairwise-level, and scene-level (Figure 1). Existing methods have achieved strong performance at the entity and pairwise levels (Siampou et al., 2024; Yu et al., 2024; Mai et al., 2023). However, scene-level VSRL remains a key challenge, as it requires jointly modeling multiple entities and their complex spatial relationships. In this study, we aim to develop a unified scene-level VSRL framework applicable to all spatial entity types.

Literature review

Before VSRL was introduced, vector data were typically represented as raw sequences of coordinates (Figure 1) (Li et al., 2018). Although such representations explicitly describe the locations of vertices, they do not directly contain entity-level characteristics (e.g., shape and orientation) or spatial relationships (e.g., topology and orientation) among entities.

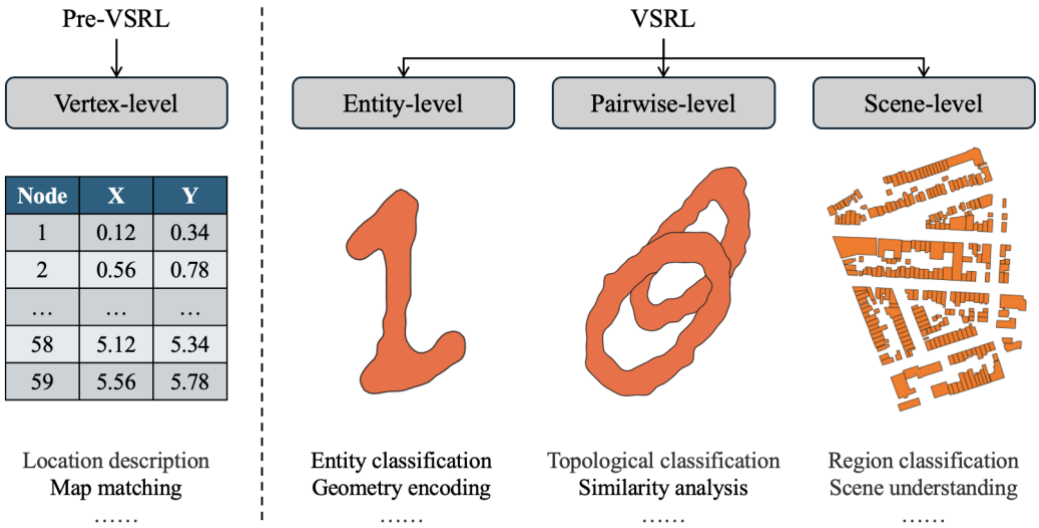

Figure 1: A taxonomy of VSRL across representation levels

Recent advances in VSRL have demonstrated promising progress in learning geometric characteristics for single entities and modeling pairwise spatial relationships (Mai et al., 2023; Siampou et al., 2024). However, two key challenges still remain for complex spatial tasks.

First, several existing methods are designed for specific entity types (e.g., points, polylines, or polygons). For example, point-based solutions focus on modeling spatial locations and distributions (Qi et al., 2017), but lack the ability to capture geometric structures; polyline-based solutions rely on sequence-based architectures such as RNNs (Li et al., 2018), which are not easily transferable to other entity types. In the meantime, limited attention has been paid to unified representation for all entity types. This limitation restricts the adaptability of VSRL methods to diverse needs.

Second, existing methods are primarily focused on single or pairwise entities. For example, the current state-of-the-art method, Geo2Vec (Chu & Shahabi, 2025), still focuses on small-scale tasks such as shape classification. However, real-world applications often focus on holistic spatial scenes and jointly analyze a large number of entities, where inter-entity relationships are as important as the entities themselves. Modeling such complex multi-entity interactions is computationally challenging and consequently remains underexplored in existing VSRL research. Entity-level VSRL is insufficient for capturing large-scale spatial patterns and supporting scene-level representation tasks. This limitation highlights the need to move beyond entity-level VSRL toward scene-level VSRL.

These limitations motivate the following research questions:

How can we move beyond entity-level VSRL toward scene-level VSRL?

How can we jointly model entities and relationships among entities within a unified representation framework?

Methodology

In this study, we propose a graph-based method for scene-level VSRL that supports all spatial entity types. In our method, the representation learning process is reformulated as a graph construction process. In this graph, nodes correspond to spatial entities and edges are created to describe the spatial relationships among entities. The overall framework (Figure 2) can be divided into two stages: node representation and edge construction.

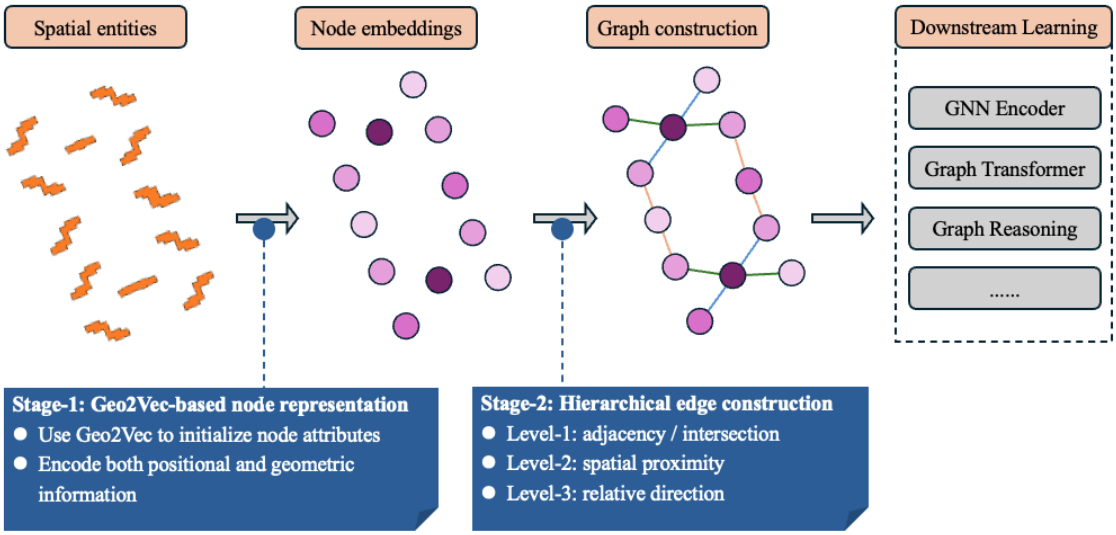

Figure 2: Framework of the proposed scene-level VSRL method

Node representation Each entity is represented as a node in the graph. We adopt the commonly used Geo2Vec to generate entity embeddings for initializing node attributes. Geo2Vec is a neural representation method for entity-level VSRL, which directly calculates both geometric and locational characteristics from the coordinate space and encodes them into a generalizable embedding space. By leveraging signed distance fields (Park et al., 2019) to convert discrete coordinates into continuous representations, Geo2Vec enables stable and expressive geometry modeling.

Edge construction Edges are constructed based on spatial relationships between entities. We consider three fundamental types of spatial relationships (Egenhofer & Franzosa, 1991; Cohn & Renz, 2008; Frank, 1991): topology, distance, and orientation, and design three hierarchical edge construction strategies to accommodate different task requirements:

Level 1 (Topological relations) Edges are created between entities that are in direct contact, including intersection and adjacency. These edges capture strong and explicit spatial relationships.

Level 2 (Proximity relations) Additional edges are introduced between entities that are spatially close but not in direct contact. These edges enable the modeling of local neighborhood structures and spatial patterns among nearby entities.

Level 3 (Directional refinement) Existing edges are further categorized based on relative orientation, enriching the representation with directional information and enabling more expressive spatial reasoning.

After graph construction, irregular spatial entity clusters are transformed into structured graph representations. This allows vector-based spatial data to be seamlessly integrated into downstream learning tasks using well-established graph-based techniques, such as GNNs. Through message passing, entity representations are refined to capture both entity-level characteristics and scene-level spatial patterns, enabling models to support multi-scale spatial tasks. Furthermore, the graph formulation naturally supports permutation invariance with respect to entity ordering, addressing a fundamental challenge in multi-entity representation.

Results

We evaluate the proposed method across three real-world datasets (Figure 3), each corresponding to one geo-entity type, in order to assess its effectiveness and generalization ability.

Offshore Wind Turbine (OWT) (Wang et al., 2025) This dataset contains 6,590 OWTs in China. The task is to classify wind farm layouts (e.g., linear layout, grid layout, irregular layout). Level 1 edge construction strategy is used in this task.

Drainage Network (DN) (Lehner & Grill, 2013) This dataset contains 1,899 river basins from Tibetan Plateau. The task is to classify basins as internal or external drainage systems. Level 1 edge construction strategy is used in this task.

Local Climate Zone (LCZ) (Demuzere et al., 2021) This dataset includes 163,441 building footprints in New York City. The task is to classify urban blocks into categories (e.g., open vs. compact zones). Level 2 edge construction strategy is used in this task.

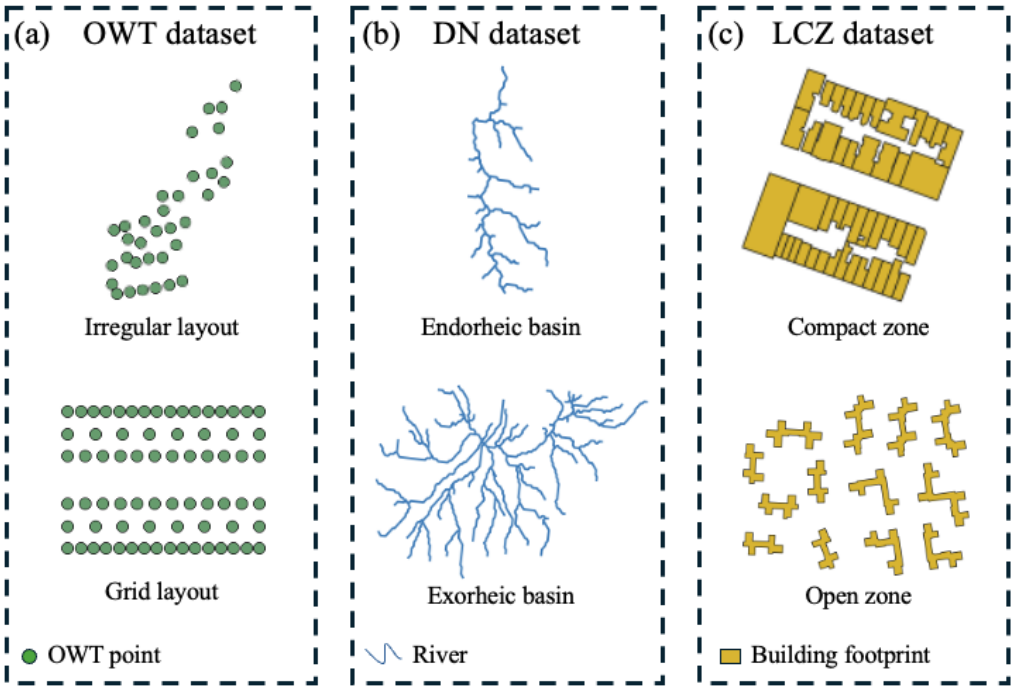

Figure 3: Examples of spatial patterns from three vector datasets; (a) OWT dataset; (b) DN dataset; (c) LCZ dataset

Three commonly used GNN models, Sage (Hamilton et al., 2017), GAT (Veličković et al., 2018), and HGT (Hu et al., 2020) are applied to process the learned representations. For baseline, we average the initial embeddings of all entities, without inter-entity message-passing. Accuracy (Acc), Balanced Accuracy (BAcc), and F1-score (F1) are used as evaluation metrics.

The results are shown in Table 1. Across all datasets, graph-based methods consistently outperform the baseline, demonstrating the importance of modeling inter-entity relationships. Among the models, Sage achieves the best performance, suggesting that neighborhood aggregation is particularly effective for capturing spatial patterns.

In addition, to analyze the impact of different graph construction strategies, we conduct an ablation experiment on the LCZ dataset, where three edge construction strategies are evaluated sequentially. The results, shown in Table 2, reveal that incorporating proximity-based edges (Level 2) significantly improves performance compared to using only topological edges (Level 1), highlighting the importance of modeling local spatial interactions. Adding directional information (Level 3) provides additional improvements in some cases, indicating that orientation-aware relationships can further enhance representation quality.

Dataset	**Model**	**Acc**	**BAcc**	**F1**
OWT	Baseline	0.43	0.25	0.16
	Sage	0.52	0.34	0.29
	GAT	0.48	0.30	0.26
	HGT	0.48	0.31	0.27
DN	Baseline	0.83	0.54	0.55
	Sage	0.87	0.65	0.61
	GAT	0.86	0.64	0.60
	HGT	0.87	0.65	0.60
LCZ	Baseline	0.64	0.39	0.39
	Sage	0.73	0.54	0.52
	GAT	0.70	0.51	0.49
	HGT	0.69	0.48	0.48

Table 1: Performance comparison of the proposed VSRL framework across different datasets.

Model	**Acc/L1**	**Acc/L2**	**Acc/L3**	**F1/L1**	**F1/L2**	**F1/L3**
Baseline	0.64	0.64	0.64	0.38	0.38	0.38
Sage	0.69	0.73	0.72	0.50	0.52	0.53
GAT	0.69	0.70	0.69	0.48	0.49	0.51
HGT	0.68	0.69	0.71	0.44	0.48	0.50

Table 2: Ablation study on different edge construction strategies.

Discussion and Conclusion

This study advances VSRL from entity-level to scene-level, where multiple spatial entities can be jointly modeled within a unified framework. By explicitly incorporating spatial relationships among entities, the proposed method enables spatial models to better capture complex spatial patterns and provides a foundation for scene-level map understanding.

To the best of our knowledge, this work is the first study to develop a scene-level VSRL method that explicitly models multi-entity spatial relationships while simultaneously supporting all spatial entity types.

References

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., & Kohli, P. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291. doi: 10.48550/arXiv.2507.22291.

Chu, C., & Shahabi, C. (2025). Geo2Vec: Shape- and distance-aware neural representation of geospatial entities. arXiv:2508.19305. doi: 10.48550/arXiv.2508.19305.

Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), Handbook of knowledge representation (pp. 551–596). Amsterdam, The Netherlands: Elsevier.

Demuzere, M., Kittner, J., & Bechtel, B. (2021). LCZ Generator: A web application to create local climate zone maps. Frontiers in Environmental Science, 9, 637455. doi: 10.3389/fenvs.2021.637455.

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. International Journal of Geographical Information Systems, 5(2), 161–174. doi: 10.1080/02693799108927841.

Frank, A. U. (1991). Qualitative spatial reasoning about cardinal directions. In D. Mark & D. White (Eds.), Proceedings of Auto-Carto 10 (pp. 148–167).

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., & Li, Y. (2024). SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 27672–27683).

Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems, 30, 1024–1034.

Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. In Proceedings of The Web Conference 2020 (pp. 2704–2710). New York, NY: ACM. doi: 10.1145/3366423.3380027.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15), 2171–2186. doi: 10.1002/hyp.9740.

Li, X., Zhao, K., Cong, G., Jensen, C. S., & Wei, W. (2018). Deep representation learning for trajectory similarity computation. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 617–628). IEEE. doi: 10.1109/ICDE.2018.00062.

Li, Y., Huang, W., Cong, G., Wang, H., & Wang, Z. (2023). Urban region representation learning with OpenStreetMap building footprints. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 1363–1373). New York, NY: ACM. doi: 10.1145/3580305.3599538.

Li, L., Xue, H., Song, Y., & Salim, F. (2024). T-JEPA: A joint-embedding predictive architecture for trajectory similarity computation. In Proceedings of the 32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 569–572). New York, NY: ACM. doi: 10.1145/3678717.3691271.

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

Mai, G., Jiang, C., Sun, W., Zhu, R., Xuan, Y., Cai, L., Janowicz, K., Ermon, S., & Lao, N. (2023). Towards general-purpose representation learning of polygonal geometries. GeoInformatica, 27(2), 289–340. doi: 10.1007/s10707-022-00481-2.

Mai, G., Yao, X., Xie, Y., Rao, J., Li, H., Zhu, Q., Li, Z., & Lao, N. (2024). SRL: Towards a general-purpose framework for spatial representation learning. In Proceedings of the 32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 465–468). New York, NY: ACM. doi: 10.1145/3678717.3691246.

Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). DeepSDF: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 165–174). doi: 10.1109/CVPR.2019.00025.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 652–660). doi: 10.1109/CVPR.2017.16.

Siampou, M. D., Li, J., Krumm, J., Shahabi, C., & Lu, H. (2024). Poly2Vec: Polymorphic Fourier-based encoding of geospatial objects for GeoAI applications. arXiv preprint arXiv:2408.14806. https://doi.org/10.48550/arXiv.2408.14806

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the International Conference on Learning Representations (ICLR).

Wang, Z., Fang, Z., Wang, X., & Wang, Z. (2025). SP-CenterNet: An accurate and lightweight deep learning approach for detecting offshore wind turbines using satellite imagery. Ocean Engineering, 334, 121433. doi: 10.1016/j.oceaneng.2025.121433.

Yu, D., Hu, Y., Li, Y., & Zhao, L. (2024). PolygonGNN: Representation learning for polygonal geometries with heterogeneous visibility graph. In Proceedings of the

30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4012–4022). New York, NY: ACM. doi: 10.1145/3637528.3671738.