GeoLink: Empowering Remote Sensing Foundation Model with OpenStreetMap Data

Lubin Bai[1]*, Xiuyuan Zhang[2], Siqi Zhang[3], Zepeng Zhang[4], Haoyu Wang[2], Wei Qin[1], Shihong Du[2]†

[1] School of Earth and Space Sciences, Peking University, Beijing, China
[2] College of Urban and Environmental Sciences, Peking University, Beijing, China
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems
Institute of Automation, CAS, Beijing, China
[4] Intelligent Maintenance and Operations Systems Lab
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Abstract

Integrating ground-level geospatial data with rich geographic context, like OpenStreetMap (OSM), into remote sensing (RS) foundation models (FMs) is essential for advancing geospatial intelligence and supporting a broad spectrum of tasks. However, modality gap between RS and OSM data, including differences in data structure, content, and spatial granularity, makes effective synergy highly challenging, and most existing RS FMs focus on imagery alone. To this end, this study presents GeoLink, a multimodal framework that leverages OSM data to enhance RS FM during both the pretraining and downstream task stages. Specifically, GeoLink enhances RS self-supervised pretraining using multi-granularity learning signals derived from OSM data, guided by cross-modal spatial correlations for information interaction and collaboration. It also introduces image mask-reconstruction to enable sparse input for efficient pretraining. For downstream tasks, GeoLink generates both unimodal and multimodal fine-grained encodings to support a wide range of applications, from common RS interpretation tasks like land cover classification to more comprehensive geographic tasks like urban function zone mapping. Extensive experiments show that incorporating OSM data during pretraining enhances the performance of the RS image encoder, while fusing RS and OSM data in downstream tasks improves the FM's adaptability to complex geographic scenarios. These results underscore the potential of multimodal synergy in advancing high-level geospatial artificial intelligence. Moreover, we find that spatial correlation plays a crucial role in enabling effective multimodal geospatial data integration. Code, checkpoints, and using examples are released at https://github.com/bailubin/GeoLink_NeurIPS2025

1 Introduction

Remote sensing (RS) serves as a powerful tool for observing and monitoring our planet. Recently, the label-free, task-agnostic nature of self-supervised learning (SSL) has enabled RS foundation models (FMs) to make significant strides [1, 2, 3, 4]. Beyond scaling up the model parameters and dataset size, many RS FMs have been specifically tailored to accommodate the unique characteristics of RS image [5, 6, 7, 8], incorporating multi-scale [2, 9], multi-temporal [4, 10, 11, 12, 13], and multi-spectral [3, 14, 15, 16] processing techniques. After pretraining, these FMs can extract

*Email: lbbai@stu.pku.edu.cn
†Corresponding author: shdu@pku.edu.cn

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

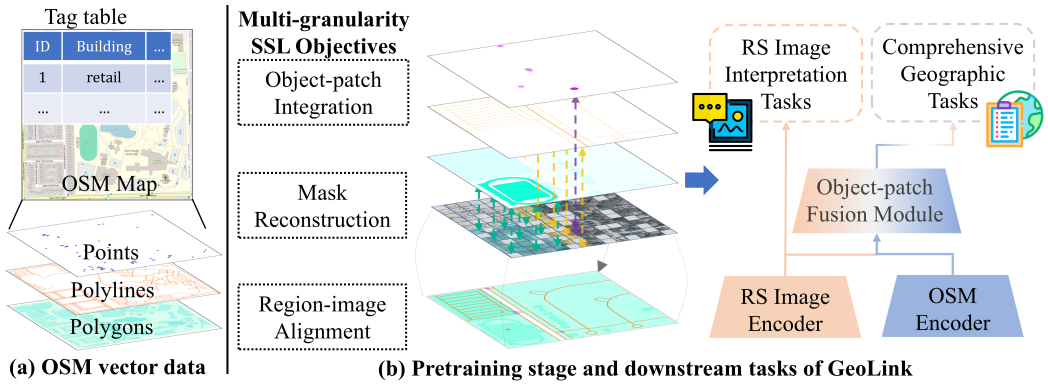


Figure 1: (a) OSM data stores the geometry information of geographic features in vector format, including points, polylines, and polygons, and leverages tags to record the semantic information. (b) GeoLink leverages multi-granularity SSL objectives to integrate RS and OSM data across multiple spatial scales, supporting both RS interpretation tasks and comprehensive geographic tasks.

meaningful, generalizable representations for RS interpretation tasks like semantic segmentation, yielding impressive performance in many domains, like environment monitoring [17] and disaster management [1, 18].

However, the integration of ground-level geospatial data remains relatively underexplored in many existing RS FMs. Ground-level geospatial data like various kinds of maps, in-situ sensor data and so on, can not only serve as RS interpretation references but also provide supplementary information for real-world applications [19, 20, 21]. Among them, OpenStreetMap (OSM) is one of the largest open-source geospatial databases of volunteered geographic information (VGI), providing rich geo-context associated with geographic locations [22]. OSM data has long been used in RS interpretation [20, 23, 24], and we believe integrating it into RS FM is essential for achieving a geo-oriented and context-aware understanding of Earth observation, as well as the advanced geospatial intelligence. First, it provides explicit location-based contextual cues that are difficult to capture from pure visual analysis, linking pixels to real-world objects and resolving ambiguities (e.g., distinguishing similar-looking buildings by location). Second, vision-language models like CLIP [25] show that structured semantics can enhance transferable representation learning, where OSM data, with its spatial hierarchies and geo-tagged attributes, plays a similar role. Finally, many geographic tasks demand a holistic understanding that RS images alone cannot provide, as they lack socioeconomic insights, while OSM data can fill this information gap.

As shown in Fig. 1(a), OSM data is originally vector-based, storing geographic objects as points, polylines, and polygons with rich tag tables, which differs significantly from RS image in data format and information content. To support RS image interpretation, most existing studies adopt indirect integration strategies. For example, converting OSM data into labels for RS images [26, 27], or constructing knowledge graphs from OSM data to provide prior knowledge for RS interpretation [28, 29]. However, such approaches tend to be manual-intensive and task-specific, confined to small-scale training datasets and experimental regions, making them misaligned with the paradigm of RS FMs. Some recent studies have explored leveraging OSM data to generate synthetic text data for RS vision-language FMs [30, 31], where relevant content descriptions associated with RS images are extracted from OSM data. They also follow an indirect way to reconcile the modality discrepancies between OSM and RS data, resulting in the loss of spatial information. To unlock OSM's potential for FM development, we aim to design a geo-spatially explicit approach that directly harnesses OSM's raw vector elements to inject geo-context into RS FMs, providing multi-perspective geographic priors while enhancing model capabilities across diverse geospatial tasks.

In this study, we introduce GeoLink, a multimodal FM that (1) enhances RS self-supervised pretraining through OSM-derived multi-granularity learning signals, (2) achieves efficient pretraining via masked input, and (3) increases the performance and diversity of downstream tasks via RS-OSM fusion. First, we design a heterogeneous graph neural network (GNN)-based OSM encoder that specifically addresses the geometric heterogeneity (points/polylines/polygons), non-Euclidean structure, and dynamic attribute tags of OSM data. Representing OSM objects as nodes and their spatial

relations as edges, the encoder performs message passing to generate both object-level (node) and region-level (graph) encodings for interaction with the RS image encoder. Second, as shown in Fig. 1(b), using location as the bridge to build multi-granularity spatial correlation, the cross-modal learning signals are derived at two levels: (1) region-image level alignment via contrastive learning with explicit spatial extent matching, and (2) object-patch level interaction through position-aware cross-attention, capturing implicit spatial associations while preserving spatial consistency for joint representation learning. Third, inspired by MAE [32] and FLIP [33], we randomly mask a large portion of image patches (e.g., 75%) during pretraining, feeding only the visible ones to accelerate training while maintaining accuracy improvements. Finally, after pretrained on 1.2 million sample pairs, we evaluate GeoLink in two task collections, as shown in Fig. 1(b), RS interpretation tasks (unimodal) and comprehensive geographic tasks (multimodal). According to the real-world application scenarios, several benchmarks are employed to assess our model across various domains, including land use/cover, agriculture, and urban planning. We find that incorporating OSM data during pretraining significantly enhances the RS image encoder's capacity, while fusing RS and OSM data in downstream tasks improves the FM's adaptability to complex geographic scenarios.

2 Related work

RS FMs. Thanks to advances in computational power and deep learning, RS FMs have rapidly evolved, trending toward multi-scale, multi-temporal, and multi-sensor designs tailored to the unique traits of RS imagery. Given the visual variance of geographic objects across resolutions, models like Scale-MAE [2] and Cross-Scale MAE [9] extend MAE with multi-scale augmentation and position embeddings to handle varying ground resolutions. As surface changes driven by seasons and human activity are common, multi-temporal RS is key for tasks like change detection and crop mapping. Approaches such as SeCo [11] use time-separated image pairs for contrastive learning, while SatMAE [4] introduces temporal embeddings to encode timestamps. Multi-sensor RS combines data from sources like multispectral and SAR to enrich downstream tasks, prompting FMs such as CROMA [3], DOFA [34], SeaMo [10], MMEarth [35], Skysense [36], and OmniSAT [37] to support multi-sensor inputs via multi-encoder or tokenizer-based designs. These works highlight the growing emphasis on multimodality. Beyond scale, time, and sensor diversity, geographic domain knowledge is also vital [38]. Geospatial vector data, such as OSM, offers rich yet underused geographic contextual information; this study explores its integration into RS FM.

Synergy of RS and geospatial vector data. A significant modality gap exists between RS and geospatial vector data, with most existing methods adopting indirect integration strategies. They can be categorized into three types based on vector data utilization: data conversion, data derivation, and knowledge graph methods. Data conversion methods utilize tools like buffering and rasterization to transform vector data into raster format, thus matching the structure of RS images for easier processing; the rasterized geospatial data may then serve as either inputs [39, 40] or training labels [22, 41, 42]. Data derivation methods generate intermediate data from vectors to assist RS tasks, such as producing image captions [30, 31], creating geospatial units from road networks [39], or constructing positive pairs for contrastive learning [43]. Knowledge graph-based methods extract geographic knowledge from vector data to build graphs that support RS image interpretation [28, 29]. While these indirect paradigms have long dominated RS-vector synergy, recent efforts explore direct integration, primarily via point data enriched with latitudinal-longitudinal priors [44, 45]. In this study, we aim to further harness the rich information contained in OSM vector data to incorporate with RS images, thereby improving performance in a wider array of geographic downstream tasks.

3 Method

Framework Overview. As shown in Fig. 2, GeoLink contains three encoders: (1) Vision Transformer (ViT)-based [46] RS image encoder f_I that encodes RS image input I into patch encodings $\varepsilon_P \in \mathbb{R}^{L_P \times D_P}$, where L_P is the number of patches and D_P is the patch feature dimension. (2) Graph Attention Convolution Network (GATConv)-based [47] OSM encoder f_O that takes the constructed OSM graph G as input, and outputs the node encodings after message passing $\varepsilon_V \in \mathbb{R}^{L_V \times D_V}$, where $\varepsilon_V = \varepsilon_{V_p} \cup \varepsilon_{V_l} \cup \varepsilon_{V_g}$. Here, L_V is the number of nodes, D_V is the node feature dimension, and p, l, and g denote the node types: point, polyline, and polygon, respectively. Details about the OSM encoder structure can be found in the Appendix A. (3) Object-patch fusion encoder f_F

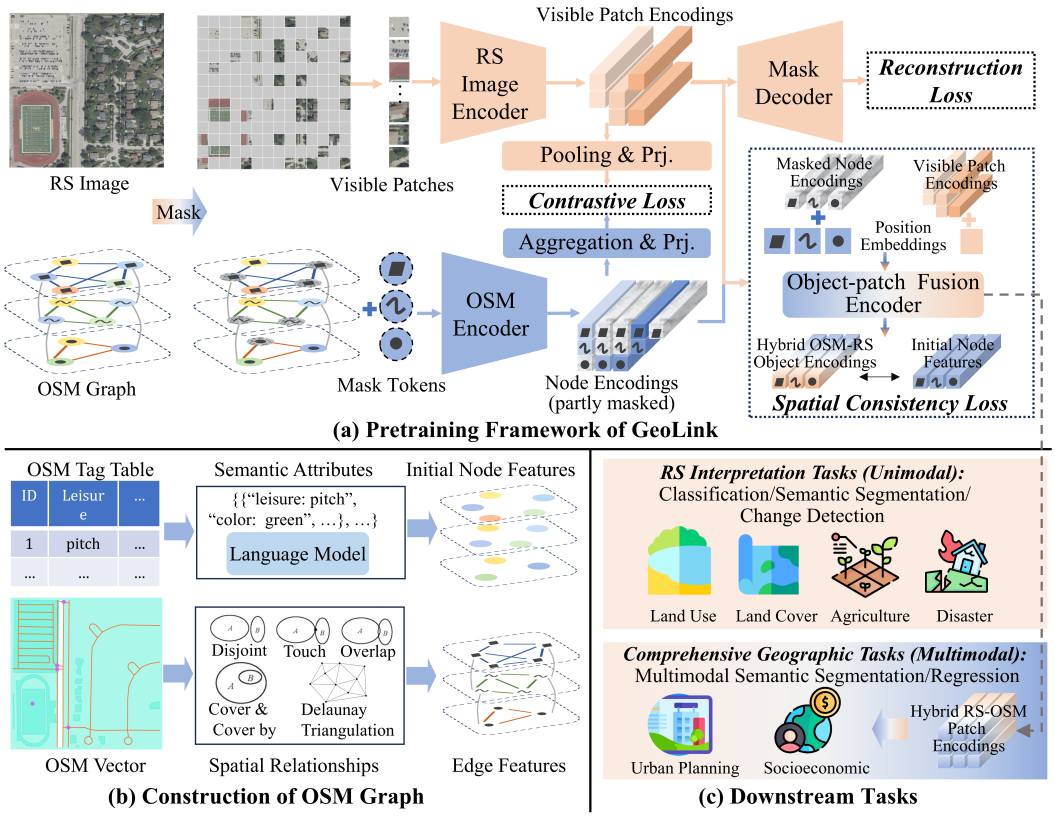


Figure 2: (a) GeoLink masks both modalities, using visible image patches and masked OSM graph as inputs. Pretraining is achieved through three SSL objectives: RS reconstruction loss, cross-modal contrastive loss, and spatial consistency loss. (b) The heterogeneous graph is employed to model OSM data, incorporating three node types and multiple spatial relationships. (c) The pretrained model can produce both unimodal and multimodal encodings, generalizing to various downstream tasks.

for fine-grained data integration, which takes the patch encodings ε_P and node encodings ε_V as input, and generates two types of multimodal encodings, including hybrid OSM-RS object encodings $\varepsilon_{OR} \in \mathbb{R}^{L_V \times D_F}$ and hybrid RS-OSM patch encodings $\varepsilon_{RO} \in \mathbb{R}^{L_P \times D_F}$, where D_F is the fusion feature dimension. During pretraining, we mask both modalities, using visible RS image patches and masked OSM graphs as inputs. Three learning objectives are leveraged to optimize the encoders, including RS reconstruction loss, cross-modal contrastive loss, and spatial consistency loss. Next, we will systematically introduce GeoLink's pretraining process.

OSM Graph Construction. As shown in Fig. 2(b), the heterogeneous graph is constructed to model the OSM vector map, where nodes represent geographic objects (points, polylines, and polygons), and edges capture various spatial relationships between them. First, we embed the OSM tags as the initial node features to represent the semantic attributes. OSM follows a free tagging system with unlimited tag categories, rendering static methods inadequate for handling unseen values. To accommodate this scenario while preserving semantics, we employ a language model to encode the OSM tags. Specifically, each OSM object's tag attributes are organized into a tag-value dictionary $D_o = \{(t_1, c_1), (t_2, c_2), \ldots, (t_n, c_n)\}$, and each tag-value pair is converted into a string $s_i = \text{concat}(t_i, ":", c_i)$, which is individually encoded using a BERT language model as $h_i = \text{BERT}(s_i)$. More frequently occurring tags (e.g., "building", "land use") better represent the general attributes of an object, while rare tags (e.g., "historic period") usually describe details. To provide a comprehensive and consistent representation, we calculate the weighted average of all tag-value encodings of an OSM object to serve as the initial node feature:$\sigma_V = \sum_{i=1}^{n} w_i h_i / \sum_{i=1}^{n} w_i$, where the weight w_i for each tag corresponds to its occurrence count in the global pretraining dataset. Second, given the scale differences between points, polylines, and polygons, we leverage spatial topological relationships rather than distances to construct edges, accounting for variations across

different vector types. For instance, point features only exhibit disjoint relationships, in which case we employ Delaunay Triangulation [48] to establish connections. A detailed exposition of the spatial relationship types among various node categories is provided in the Appendix A. Compared to spatial distance, topology embodies a definitive spatial relationship and is less susceptible to noise. The constructed OSM graph is then fed into the OSM encoder, where message passing enables each node to aggregate information from its neighbors, resulting the node encodings ε_V.

Masked Inputs. During the pretraining phase, the learning objectives are built upon masked inputs, which not only enhances representation learning but also reduces memory consumption. For RS data with a ViT encoder, the image is divided into a grid of non-overlapping patches, with a large portion randomly masked out, leaving only the visible ones as input to obtain the visible patch encodings ε_P^v. Following MAE [32], we leverage a mask decoder containing two Transformer blocks to reconstruct the masked patches. Then, we calculate the reconstruction loss between the reconstructed patches $\hat{I}^m$ and original masked patches I^m as:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_P^m}\sum_{j=1}^{L_P^m}\left(\hat{I}_{ij}^m - I_{ij}^m\right)^2,$$

where N is the batch size and L_P^m is the number of masked patches in each input image. For OSM data, we employ a node-masking strategy. For each masked node i, its initial feature σ_{V_i} is replaced by a learnable mask token while preserving its original adjacency edges with other nodes to ensure effective message passing within the OSM encoder. We adopt three different mask tokens for points, polylines, and polygons, respectively. The OSM encoder outputs encodings ε_V for all nodes, including both masked and visible ones, i.e., $\varepsilon_V = \varepsilon_V^m \cup \varepsilon_V^v$.

Region-image Level Alignment. Since each input pair of RS and OSM data covers the same geographic extent, they describe the corresponding region from different perspectives and contain correlated, complementary information. Consequently, we employ contrastive learning to align them. To obtain region-level OSM encoding, we design an aggregation module to aggregate the heterogeneous node encodings. Specifically, this module first leverages three Set2Set layers that independently aggregate nodes of each type into type-specific encodings $\varepsilon_{G_t} = \text{Set2Set}_t(\varepsilon_{V_t})$, where $t \in \{p, l, g\}$ corresponds to the node types of point, polyline, and polygon. Then a linear layer followed by a Softmax layer computes type attention to weight-sum the type-specific encodings to produce the OSM region encoding: $\varepsilon_G = \sum_{t\in\{p,l,g\}} \text{Softmax}(\text{Linear}(\varepsilon_{G_t}))\varepsilon_{G_t}$. For RS data, we employ mean pooling for the visible patch encodings ε_P^v to obtain the image-level encoding $\varepsilon_I = \text{MeanPool}(\varepsilon_P^v)$. Following [25], we project both the OSM region encoding ε_G and the RS image encoding ε_I using separate linear layers, i.e., $z_G = \text{Linear}_G(\varepsilon_G)$ and $z_I = \text{Linear}_I(\varepsilon_I)$, and contrast both modalities using the InfoNCE loss [49]:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{2N}\left(\sum_{i=1}^{N}\frac{\exp(\text{sim}(z_G^i, z_I^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_G^i, z_I^j)/\tau)} + \sum_{i=1}^{N}\frac{\exp(\text{sim}(z_I^i, z_G^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_I^i, z_G^j)/\tau)}\right),$$

where N is the batch size, τ is a temperature parameter, and $\text{sim}(\cdot,\cdot)$ is the cosine similarity function, i.e., $\text{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}$. Through contrastive learning, the inherent structured semantic information in OSM data is effectively conveyed to the image encoder, thereby guiding its pretraining process.

Object-patch Level Integration. To further facilitate interaction between the two modalities and obtain fine-grained fused representations, we design an object-patch fusion encoder. This encoder is implemented as a two-way Transformer composed of one self-attention layer and two cross-attention layers (details in Appendix A). As shown in Fig. 2(a), during pretraining, the fusion encoder accepts the masked node encodings ε_V^m and the visible patch encodings ε_P^v as inputs, and produces two types of fused encodings: hybrid OSM-RS object encodings ε_{OR}^m and hybrid RS-OSM patch encodings ε_{RO}^m. A critical challenge arises from the inherent spatial ambiguity between the two input encodings: the absence of explicit geographic correspondence makes direct cross-attention operations susceptible to erroneous feature associations, e.g., accidentally connect unrelated elements. To address this, we incorporate sinusoidal position embeddings into the fusion encoder by adding them to each input. The sinusoidal embedding handles individual coordinates, which can be directly adopted to point nodes. To capture the spatial coverage of polylines and polygons, we sample their key-points and compute the average of the sampled points' position embeddings (details see in Appendix A). These enhanced spatial signatures enable the model to progressively establish accurate cross-modal association

through attention-based learning. As shown in Fig. 2(a), the output hybrid OSM-RS object encodings integrate features from visible nodes (via message passing in the OSM encoder) and visible patches (via cross-modal interaction in the fusion encoder), which encapsulate the spatial context surrounding masked OSM objects. According to the first law of geography [50], this contextual information is strongly correlated with the intrinsic properties of the masked ones. To enforce spatial-semantic consistency, we introduce a consistency loss function that operates on the hybrid OSM-RS object encodings ε_{OR}^m and the initial features of masked nodes σ_V^m:

$$\mathcal{L}_{\text{cst}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_V^m}\sum_{j=1}^{L_V^m}\left(\varepsilon_{OR_{ij}}^m - \sigma_{V_{ij}}^m\right)^2,$$

where N is the batch size, and L_V^m is the number of masked nodes in each graph. The synergy of the position embedding and consistency constraints significantly enhances the model's capacity for cross-modal representation learning, improving the RS encoder's ability to capture fine-grained semantic features. Finally, the three objectives are combined together for pretraining: $\mathcal{L} = \alpha\mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{cont}} + \gamma\mathcal{L}_{\text{cst}}$, where the loss weights are set as $\alpha = 1$, $\beta = 0.01$, and $\gamma = 0.01$ by default (related experiments about the loss weights are provided in Appendix D).

4 Experiments

Pretraining details. To pretrain GeoLink, we construct a multimodal dataset derived from SkyScript-top30 [31]. The SkyScript-top30 dataset contains multi-source, multi-resolution RS images with RGB bands, featuring ground sample distances (GSD) ranging from 0.1 m/pixel to 30 m/pixel. For each RS image, the corresponding OSM data is downloaded from the Overpass API using its geo-coordinate and timestamp. After preprocessing like data cleaning, we obtain a final pretraining dataset of 1,271,431 matched pairs. More details are provided in the Appendix B. A ViT-L model is employed as the RS image encoder in this study. The default masking ratios for RS patch and OSM graph node is 75% and 20%. And $\tau = 0.2$ for contrastive loss. Our experiments are conducted on a Linux server equipped with 4 NVIDIA RTX6000 GPUs (48GB) using bfloat16 precision. Unlike FMs such as Scale-MAE (800 epochs) and CROMA (600 epochs) which typically require a large number of pretraining epochs, GeoLink demonstrates significantly faster convergence. We pretrain it for only 60 epochs (including 5 warmup epochs), with a batch size of 2640, a base learning rate of 1×10^{-4}, and a cosine decay schedule for learning rate cooldown. For data augmentation, we apply random cropping, horizontal flipping, and color jittering to the RS images. Notably, we perform corresponding geometric transformations on the OSM data to maintain spatial alignment with the augmented RS images. The model optimization employs AdamW with hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.95$) and a weight decay of 0.05.

Downstream Task Settings. As depicted in Figure 2(c), GeoLink supports both RS image interpretation tasks (unimodal) and comprehensive geographic tasks (multimodal). For the former, only the RS image encoder is employed, which is combined with task-specific protocols to assess the learned unimodal representations through classification, semantic segmentation, and change detection tasks. We benchmark GeoLink against six RS FMs, including GASSL [45], MMEarth [35], Scale-MAE [2], Cross-scale MAE [9], CROMA [3], and DOFA [34]. Other comparisons can be found in the Appendix D. For multimodal geographic tasks, we incorporate OSM data into the downstream tasks. We evaluate on urban function zone (UFZ) segmentation, urban village (UV) identification, population density (POP) and carbon emission (CO2) estimation tasks, which are crucial for realistic urban planning and socioeconomic analysis. These challenges typically require multi-source geographic data, making them ideal for assessing GeoLink's multimodal capabilities. The details of baseline model selection, downstream model architectures, and evaluation protocols are provided in the Appendix C.

4.1 Unimodal tasks

Classification tasks. We evaluate the RS representations learned by GeoLink using three classification protocols, including: (1) kNN ($k = 20$) evaluates representation quality by measuring instance clustering without further training; (2) linear probing assesses performance using a linear classifier on frozen RS features; (3) fine-tuning jointly updates the RS encoder and classifier for task-specific adaptation. For comprehensive evaluation, we employ seven RS benchmarks that span diverse spatial

Table 1: Comparison of different models across seven classification benchmarks under kNN, linear probing (LP), and fine-tuning (FT) evaluation protocols (Top-1 accuracy %).

Model	**Backbone**	**MLRSNet**			**EuroSAT**			**WHU-RS19**			**OPTIMAL-31**			**RESISC-45**			**AiRound**			**UCMerced**		
		kNN	LP	FT	kNN	LP	FT	kNN	LP	FT	kNN	LP	FT	kNN	LP	FT	kNN	LP	FT	kNN	LP	FT
GASSL	ResNet50	91.28	93.08	95.83	91.24	94.13	97.34	86.88	96.62	96.82	76.56	85.48	86.77	81.50	87.19	92.78	65.52	75.47	77.66	82.67	93.62	95.14
MMEarth	ConvNext V2	89.29	91.42	96.31	94.84	96.42	98.27	90.82	96.18	97.72	73.11	89.74	90.64	80.21	88.50	94.44	68.30	73.65	75.68	86.67	95.36	96.03
Scale-MAE	ViT-L	92.26	**93.56**	96.97	93.42	**97.40**	98.06	90.85	98.41	98.01	78.28	87.63	88.71	85.42	91.14	94.15	71.14	**78.20**	82.84	78.38	96.10	95.71
Cross-Scale MAE	ViT-L	92.63	93.30	96.23	93.24	95.58	97.97	89.66	97.02	97.42	79.78	87.85	88.71	85.10	90.89	93.54	72.23	74.81	80.79	84.67	95.90	96.19
CROMA	ViT-L	89.95	92.64	96.21	94.64	97.13	98.17	85.88	95.43	96.22	77.85	85.05	86.88	82.41	88.61	93.36	65.40	73.33	80.52	81.90	93.81	94.58
DOFA	ViT-L	90.73	92.40	96.36	93.93	96.79	98.20	90.04	98.12	98.31	77.28	90.89	90.54	83.04	89.85	93.85	70.27	75.48	78.35	87.41	95.16	96.50
GeoLink	ViT-L	**93.48**	93.49	**97.35**	**95.22**	97.30	**98.30**	**91.05**	**98.81**	**98.41**	**82.37**	**91.40**	**91.72**	**87.33**	**91.42**	**94.45**	**72.52**	77.59	**83.38**	**87.43**	**98.19**	**98.10**

Table 2: Comparison of encoder-freezing and fine-tuning performance across four semantic segmentation/change detection downstream datasets (mIoU %).

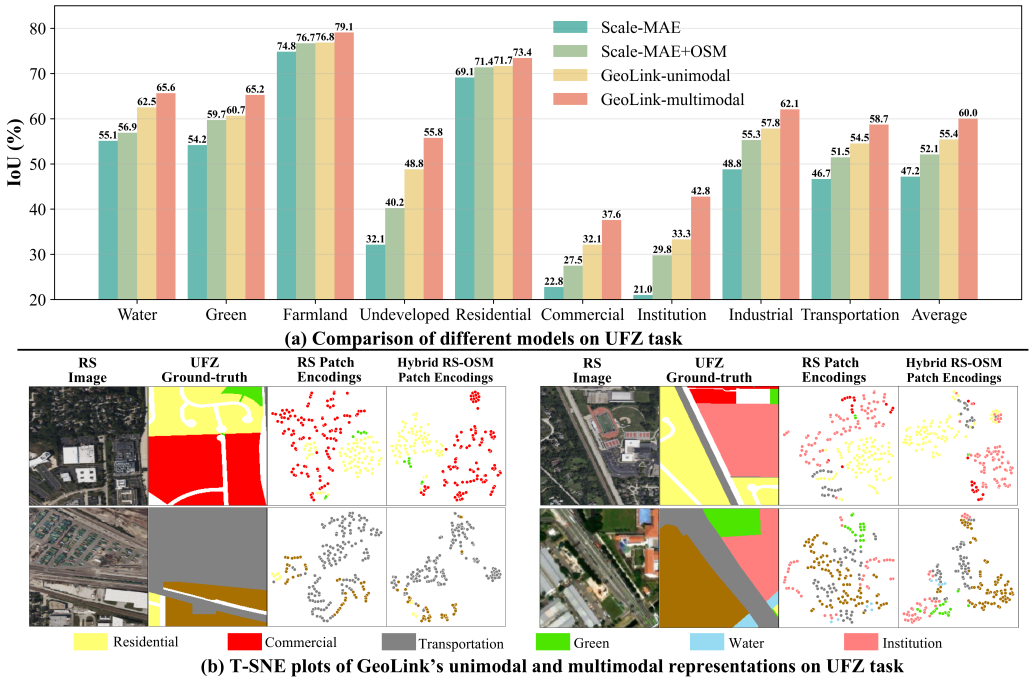
Model	Backbone	FiveBillionPixels		AI4Smallfarms		SpaceNet7		xView2	
		Freezing	Fine-tuning	Freezing	Fine-tuning	Freezing	Fine-tuning	Freezing	Fine-tuning
GASSL	ResNet50	57.47	61.37	39.65	43.29	57.63	62.09	56.27	59.87
MMEarth	ConvNext V2	56.12	62.69	37.86	43.13	62.20	62.66	56.54	59.92
Scale-MAE	ViT-L	58.94	**65.73**	41.11	45.98	62.78	63.22	58.42	60.37
Cross-Scale MAE	ViT-L	58.68	63.96	40.33	44.87	60.23	63.03	58.44	60.87
CROMA	ViT-L	58.09	63.96	41.16	45.89	58.97	61.19	57.34	59.12
DOFA	ViT-L	57.83	63.94	38.31	45.94	61.38	62.43	58.86	61.47
GeoLink	ViT-L	**60.49**	64.93	**43.26**	**47.29**	**63.22**	**64.07**	**59.94**	**61.94**

resolutions and category systems: MLRSNet [51], EuroSAT [52], WHU-RS19 [53], OPTIMAL-31 [54], RESISC-45 [55], AiRound [56], and UCMerced [57]. All benchmarks are split into 50% for training, 10% for validation, and 40% for testing. Each experiment is repeated three times under different random seeds and the average results are reported to ensure robustness. Please refer to the Appendix C for detailed downstream task settings. As shown in Table 1, GeoLink achieves state-of-the-art performance on most datasets, showcasing its superiority in learning generalizable representations. Notably, GeoLink outperforms all compared FMs by significant margins under the kNN protocol, which indicates that it has learned structured RS representations where semantically similar samples are close in the feature space. Compared to linear probing, GeoLink demonstrates even more pronounced superiority under the fine-tuning protocol. Beyond this, we also conduct data efficiency analysis, and observe that GeoLink's advantage becomes even more evident when training samples are limited. Detailed results are provided in Appendix D.

Semantic segmentation and change detection tasks. Unlike classification tasks, semantic segmentation and change detection aim to evaluate the model's ability to capture spatially detailed representations. For both tasks, we follow the protocols of the PANGAEA-bench, including using UperNet as the decoder and adopting identical training configurations. We evaluate performance on four benchmarks from PANGAEA-bench: Five-Billion-Pixels [58], AI4SmallFarms [59], xView2 [60], and SpaceNet7 [61], which cover diverse application domains such as agricultural monitoring and disaster management. The first two benchmarks correspond to semantic segmentation tasks, while the latter two are used for change detection. Table 2 presents the performance comparison of various FMs on the four benchmarks under both encoder-freezing and fine-tuning settings. GeoLink consistently achieves the best results on average in both scenarios, demonstrating the strong generalization and adaptability of its learned RS representations. The advantages of GeoLink are also evident in more challenging datasets such as AI4Smallfarms and xView2, where it consistently leads under both settings. When considered alongside the unimodal classification results, these findings further underscore the effectiveness of GeoLink's cross-modal pretraining strategy. By leveraging the supervision from OSM data, GeoLink significantly enhances the capability of the RS image encoder to learn transferable and semantically rich representations.

4.2 Multimodal tasks

UFZ segmentation. UFZs, such as residential, commercial, and institutional areas, are essential units in urban planning, reflecting complex socioeconomic and physical dynamics. However, due to the heterogeneous nature of man-made infrastructure and visual discontinuities in RS images, accurately identifying UFZs using RS data alone remains challenging. To address this, growing research efforts incorporate multi-source data [62, 63, 64, 65]. We construct a challenging real-world UFZ segmentation benchmark by refining planning maps from Chicago, Singapore, and Shenzhen, comprising 60,970 samples across nine UFZ categories. Using UperNet as the decoder, we fine-tune



(a) Comparison of different models on UFZ task

(b) T-SNE plots of GeoLink's unimodal and multimodal representations on UFZ task

Figure 3: (a) IoU (%) performances of each UFZ category. (b) T-SNE is used to visualize the learned patch encodings of GeoLink. With the incorporation of OSM data, multimodal encodings become more compact and discriminative than unimodal ones.

GeoLink and Scale-MAE under unimodal and multimodal settings. In the multimodal setup, UperNet consumes hybrid RS-OSM features produced by the object-patch fusion encoder. Detailed structures and settings are provided in the Appendix C. GeoLink outperforms Scale-MAE in both unimodal and multimodal scenarios (Fig. 3(a)). Notably, when OSM data is excluded at downstream task time, GeoLink benefits from multimodal pretraining, yielding superior performance. When OSM data is included, GeoLink's performance further improves—particularly for complex classes such as industrial, institutional, and commercial—highlighting the advantages of precise, semantically rich multimodal fusion. In addition, Scale-MAE+OSM largely outperforms Scale-MAE, further indicating the effectiveness of the proposed designed fusion encoder. Furthermore, we visualize the patch representations of GeoLink in Figure 3(b). Some categories that are easily confused in RS images exhibit scattered and overlapping distribution in unimodal scenario, while the hybrid RS-OSM patch encodings show significantly greater discriminability, highlighting GeoLink's multimodal fusion capability and the importance of multi-source information in comprehensive geospatial tasks. We also conduct a relevant while distinctive downstream task, i.e., urban village identification, and detailed results are provided in Appendix D.

UV identification. UVs—informal residential zones within or on the outskirts of cities—often suffer from deficient infrastructure and low-quality living conditions. Identifying UV is important for urban planning and sustainable development. In this study, we construct an UV semantic segmentation dataset (details in Appendix C) to evaluate the learned representations.

Table 3: Comparison of methods on the urban village identification task (IoU %).

Method	Background	UV	mIoU
Scale-MAE	90.29	58.21	74.25
Scale-MAE+OSM	91.37	68.81	80.09
GeoLink-unimodal	90.40	68.67	79.53
GeoLink-multimodal	**92.29**	**71.08**	**81.68**

The results in Table 3 demonstrate that GeoLink maintains outstanding performance in this task, showing significant improvement compared to Scale-MAE whether in unimodal or multimodal scenarios. Comprehensively analyzing both the UV and UFZ tasks, the hybrid RS-OSM patch representations generated by GeoLink effectively couple information from both data sources, making them suitable for fine-grained tasks like semantic segmentation.

POP and CO2 estimation. Fine-scale spatialized population and carbon emission data are essential for geoscience research, including climate change studies. In this work, we evaluate GeoLink on both tasks. Grid-based population density and carbon emission data for Chicago, Singapore, and Shenzhen are sourced from WorldPop and ODIAC to construct evaluation benchmarks (details in Appendix C). We use a two-layer MLP as the regression head for each task. As illustrated in Table 4, integrating OSM data significantly improves GeoLink's performance on both POP and CO2 estimation, highlighting the value of multi-source geospatial data. This supports our view that multimodal data fusion will be essential to the next generation of FMs in geography.

Table 4: Performance comparison on POP and CO2 estimation tasks (r^2 %).

Model	POP	CO2
Scale-MAE	47.18	59.12
Scale-MAE+OSM	48.29	59.97
GeoLink-Unimodal	49.76	62.37
GeoLink-Multimodal	**51.88**	**65.12**

4.3 Ablation studies

Model designing. Drawing on GeoLink's characteristics, we conduct comprehensive ablation studies to isolate the effects of key design choices, including learning objective, masking ratio, and position embedding for data fusion. Unless otherwise noted, every variant is pretrained for 60 epochs using ViT-L as the image encoder. We assess each configuration on kNN classification, linear probing, and UFZ segmentation. For kNN and linear probing, results are reported as the mean performance over seven benchmarks. The results are shown in Table 5.

(1) Learning objective. When using only $\mathcal{L}\mathrm{rec}$, the model degenerates into a standard MAE. Under this setting, performance is notably poor after 60 epochs of pretraining and only becomes competitive after 300 epochs, indicating slow convergence. In contrast, GeoLink achieves strong performance within just 60 epochs, thereby reducing computational cost. Additionally, $\mathcal{L}\mathrm{cont}$ facilitates cross-modal alignment to enhance the image encoder, while $\mathcal{L}_{\mathrm{cst}}$ primarily serves to optimize the fusion module to obtain multimodal encodings.

(2) RS masking ratio. Reducing the masking ratio to 50% for RS images leads to performance degradation and increased training cost, and we can also observe a slight performance drop at 80% and a significant collapse at 90%.

(3) OSM masking ratio. The model shows robustness to the masking ratio applied to OSM inputs, maintaining stable performance when approximately 20% of the data is masked.

(4) Position embedding for data fusion. Removing position embeddings causes a slight drop in unimodal performance but leads to a significant decline in multimodal effectiveness. This underscores the critical role of spatial correlation in the fusion of geospatial modalities. More experiments on the designing of position embedding can be found in the Appendix D.

Table 5: Ablation study on GeoLink. Performance on classification tasks (top-1 accuracy %) and UFZ segmentation (mIoU %) under varying configurations.

Ablation	Setting	Cost	kNN	Linear	UFZ-U	UFZ-M
Default	–	1.00×	87.06	92.60	55.40	60.00
(1) Learning objective	$\mathcal{L}_{\mathrm{rec}}$ (60 epochs)	0.85×	78.12	83.42	40.41	–
	$\mathcal{L}_{\mathrm{rec}}$ (300 epochs)	4.25×	85.97	90.03	49.09	–
	$\mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{cont}}$	0.92×	86.12	91.14	53.94	–
	$\mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{cst}}$	0.93×	82.34	86.77	51.06	57.39
(2) RS masking ratio	50%	2.10×	83.21	89.64	49.96	54.28
	80%	0.95×	85.21	90.17	52.01	56.04
	90%	0.90×	78.02	83.47	43.02	45.81
(3) OSM masking ratio	15%	1.00×	86.99	91.97	54.21	58.70
	25%	1.00×	86.43	91.04	53.84	58.02
(4) Fusion position embedding	Without	1.00×	86.47	91.87	53.97	56.22
(5) OSM encoder variants	GCN variant	1.00×	87.01	92.14	55.02	59.42
	Transformer variant	1.01×	86.14	91.47	53.73	58.57

(5) OSM encoder varians. We design two variant experiments to illustrate the effectiveness of the current OSM encoder. One is GCN variant which replaces the GAT in the OSM encoder with GCN while keeping all other settings unchanged, and the other is transformer variant which replaces the existing GNN-based OSM encoder with 2 standard Transformer blocks. To adapt to the Transformer's structure, each OSM node in the original GNN is treated as a token input to the Transformer. Additionally, position embedding is added to each token to preserve spatial information. Compared to the GCN variant, the attention mechanism introduced by GAT yields better performance. The Transformer variant underperforms both GCN and GAT, which we speculate is primarily due to the insufficient message-passing between OSM nodes in it, hindering the learning of OSM spatial consistency objective during pretraining.

OSM data completeness. Given the crowd-sourced nature of OSM, data completeness varies across regions, requiring models to remain robust under sparse conditions. To evaluate this, we simulate completeness by randomly removing OSM objects (Table 6). Results indicate only a minor performance drop when 20% of OSM data is removed. Even with 50% of the OSM data omitted, GeoLink maintains strong performance across most tasks, with a notable decline observed primarily on multimodal UFZ segmentation task. These findings suggest that GeoLink demonstrates substantial robustness to incomplete OSM coverage and can adapt to regions with limited OSM availability.

Table 6: Results of kNN (top-1 accuracy %) and UFZ (mIoU %) under varying OSM completeness.

Completeness	kNN	UFZ-U	UFZ-M
100%	87.06	55.40	60.00
80%	86.98	55.01	59.13
50%	86.59	54.12	57.37

5 Conclusion

In this study, we propose GeoLink, which effectively integrates geographic contextual cues from OSM data into the RS FM through semantic-spatial feature extraction and spatial-aware cross-modal interaction. This design enhances the pretraining process of RS image encoder, significantly improving its performance on image interpretation tasks. In addition, GeoLink produces fine-grained hybrid RS-OSM patch encodings tailored for comprehensive geographic tasks. Extensive evaluations across diverse benchmarks demonstrate that GeoLink outperforms previous state-of-the-art models and excels in more challenging downstream tasks such as UFZ mapping. Furthermore, our findings emphasize the pivotal role of spatial correlation in bridging and fusing multimodal geospatial data.

Despite its promising results, GeoLink has certain limitations: (1) The focus of this paper is to explore how to leverage the rich geographic information of OSM data in RS FM, and at present, it only supports RGB images and cannot process multispectral RS data. We believe the inclusion of multispectral RS images could further enhance the model, and we plan to improve the image encoder to accommodate data from various sensor types. (2) The current position embedding for OSM vectors can lead to the loss of spatial details. We argue that if position embedding can better capture the spatial characteristics, it can facilitate more accurate and deeper spatial correlations, thereby enhancing the synergy across multimodal geospatial data.

6 Acknowledgement

The work is funded by the National Key Research and Development Program of China (2023YFC3804802) and National Natural Science Foundation of China (No. 42330103). Shihong Du is the corresponding author.

References

[1] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024.

[2] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[3] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[4] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[5] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.

[6] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.

[7] Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. Fomo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27858–27868, 2025.

[8] Xuyang Li, Chenyu Li, Pedram Ghamisi, and Danfeng Hong. Fleximo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*, 2025.

[9] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023.

[10] Xuyang Li, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. Seamo: A multi-seasonal and multimodal remote sensing foundation model. *arXiv preprint arXiv:2412.19237*, 2024.

[11] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.

[12] Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, et al. Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023.

[13] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023.

[14] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. *arXiv preprint arXiv:2309.05300*, 2024.

[16] Jingtao Li, Yingyi Liu, Xinyu Wang, Yunning Peng, Chen Sun, Shaoyu Wang, Zhendong Sun, Tian Ke, Xiao Jiang, Tangwei Lu, et al. Hyperfree: A channel-adaptive and tuning-free foundation model for hyperspectral remote sensing imagery. *arXiv preprint arXiv:2503.21841*, 2025.

[17] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

[18] Chia-Yu Hsu, Wenwen Li, and Sizhe Wang. Geospatial foundation models for image analysis: Evaluating and enhancing nasa-ibm prithvi's domain adaptability. *International Journal of Geographical Information Science*, pages 1–30, 2024.

[19] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 10(2):1–46, 2024.

[20] John E Vargas-Munoz, Shivangi Srivastava, Devis Tuia, and Alexandre X Falcao. Openstreetmap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):184–199, 2020.

[21] Lubin Bai, Xiuyuan Zhang, Wei Qin, Jiang Long, Haoyu Wang, Xiaoyan Dong, and Shihong Du. From orbit to ground: A comprehensive review of multimodal self-supervised learning for remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–37, 2025.

[22] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 67–75, 2017.

[23] Munazza Usmani, Maurizio Napolitano, and Francesca Bovolo. Towards global scale segmentation with openstreetmap and remote sensing. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 8:100031, 2023.

[24] Lubin Bai, Xiuyuan Zhang, Haoyu Wang, and Shihong Du. Integrating remote sensing with openstreetmap data for comprehensive scene understanding through multi-modal self-supervised learning. *Remote Sensing of Environment*, 318:114573, 2025.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[26] Moritz Schott, Adina Zell, Sven Lautenbach, Gencer Sumbul, Michael Schultz, Alexander Zipf, and Begüm Demir. Analyzing and improving the quality and fitness for purpose of openstreetmap as labels in remote sensing applications. In *Volunteered geographic information: Interpretation, visualization and social context*, pages 21–42. Springer Nature Switzerland Cham, 2023.

[27] Taili Wan, Hongyang Lu, Qikai Lu, and Nianxue Luo. Classification of high-resolution remote-sensing image using openstreetmap information. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2305–2309, 2017.

[28] Yixiang Chen, Xu Dang, Daoyou Zhu, Yi Huang, and Kun Qin. Urban functional zone mapping by coupling domain knowledge graphs and high-resolution satellite images. *Transactions in GIS*, 28(6):1510–1535, 2024.

[29] Zhao Gun and Jianyu Chen. Novel knowledge graph-and knowledge reasoning-based classification prototype for obia using high resolution remote sensing imagery. *Remote Sensing*, 15(2):321, 2023.

[30] Yuanxin Zhao, Mi Zhang, Bingnan Yang, Zhan Zhang, Jujia Kang, and Jianya Gong. Luojiahog: A hierarchy oriented geo-aware image caption dataset for remote sensing image–text retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 222:130–151, 2025.

[31] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5805–5813, 2024.

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[33] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023.

[34] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *arXiv e-prints*, pages arXiv–2403, 2024.

[35] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024.

[36] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. 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*, pages 27672–27683, 2024.

[37] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

[38] Hao Zhang, Jin-Jian Xu, Hong-Wei Cui, Lin Li, Yaowen Yang, Chao-Sheng Tang, and Niklas Boers. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[39] Taïs Grippa, Stefanos Georganos, Soukaina Zarougui, Pauline Bognounou, Eric Diboulo, Yann Forget, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, and Eléonore Wolff. Mapping urban land use at street block level using openstreetmap, remote sensing data, and spatial metrics. *ISPRS International Journal of Geo-Information*, 7(7):246, 2018.

[40] Qiqi Zhu, Longli Ran, Yunchang Zhang, and Qingfeng Guan. Integrating geographic knowledge into deep learning for spatiotemporal local climate zone mapping derived thermal environment exploration across chinese climate zones. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217:53–75, 2024.

[41] Di Yang, Chiung-Shiuan Fu, Audrey C Smith, and Qiang Yu. Open land-use map: a regional land-use mapping strategy for incorporating openstreetmap with earth observations. *Geo-spatial information science*, 20(3):269–281, 2017.

[42] Yang Ju, Iryna Dronova, and Xavier Delclòs-Alió. A 10 m resolution urban green space map for major latin american cities from sentinel-2 remote sensing images and openstreetmap. *Scientific Data*, 9(1):586, 2022.

[43] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM web conference 2022*, pages 3308–3316, 2022.

[44] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023.

[45] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[47] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[48] Dierk Rhynsburger. Analytic delineation of thiessen polygons. *Geographical analysis*, 5(2):133–144, 1973.

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[50] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[51] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020.

[52] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[53] Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and remote sensing letters*, 8(1):173–176, 2010.

[54] Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, 2018.

[55] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[56] Gabriel Machado, Edemir Ferreira, Keiller Nogueira, Hugo Oliveira, Matheus Brito, Pedro Henrique Targino Gama, and Jefersson Alex dos Santos. Airound and cv-brct: Novel multiview datasets for scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:488–503, 2020.

[57] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.

[58] Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:178–196, 2023.

[59] Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. Ai4smallfarms: A dataset for crop field delineation in southeast asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

[60] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

[61] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[62] Lubin Bai, Weiming Huang, Xiuyuan Zhang, Shihong Du, Gao Cong, Haoyu Wang, and Bo Liu. Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. *ISPRS Journal of Photogrammetry and Remote Sensing*, 201:193–208, 2023.

[63] Wei Chen, Huiping Huang, Jinwei Dong, Yuan Zhang, Yichen Tian, and Zhiqi Yang. Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:436–452, 2018.

[64] Yifan Chen, Chaokang He, Wei Guo, Shiqi Zheng, and Bingxian Wu. Mapping urban functional areas using multisource remote sensing images and open big data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:7919–7931, 2023.

[65] Runyu Fan, Ruyi Feng, Wei Han, and Lizhe Wang. Urban functional zone mapping with a bibranch neural network via fusing remote sensing and social sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11737–11749, 2021.

[66] Stefan Leyk, Andrea E Gaughan, Susana B Adamo, Alex De Sherbinin, Deborah Balk, Sergio Freire, Amy Rose, Forrest R Stevens, Brian Blankespoor, Charlie Frye, et al. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, 11(3):1385–1409, 2019.

[67] Christopher T Lloyd, Alessandro Sorichetta, and Andrew J Tatem. High resolution global gridded data for use in population studies. *Scientific data*, 4(1):1–17, 2017.

[68] Dana R Thomson, Douglas R Leasure, Tomas Bird, Nikos Tzavidis, and Andrew J Tatem. How accurate are worldpop-global-unconstrained gridded population data at the cell-level?: A simulation analysis in urban namibia. *Plos one*, 17(7):e0271504, 2022.

[69] Tomohiro Oda, Shamil Maksyutov, and Robert J Andres. The open-source data inventory for anthropogenic co 2, version 2016 (odiac2016): a global monthly fossil fuel co 2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data*, 10(1):87–107, 2018.

[70] Tomohiro Oda and Shamil Maksyutov. A very high-resolution (1 km× 1 km) global fossil fuel co 2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*, 11(2):543–556, 2011.

[71] Ziyang Gong, Zhixiang Wei, Di Wang, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Naoto Yokoya, et al. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv preprint arXiv:2410.22629*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, our main contributions detailed in Sec. 1. Also see in Sec. 4 for more experimental evidence.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, please see Sec. 5 for limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not include any theoretical results in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we include needed experiment details in Sec. 4 and Appendix B and C. We also upload the codes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We upload the codes to recover the results. Once the blind review period is finished, we will open-source codes, instructions, constructed benchmark datasets, and model checkpoints.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see in Sec. 4 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the resource limitation, we do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the experiments compute resources in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This work focuses on a academic, publicly-available datasets. This work is not related to any private or personal data, and there's no explicit negative social impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not foresee any high risk for misuse of this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we credited them in appropriate ways.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Yes, we utilize language model Bert to encode the OSM tags, which is detailed in Sec.3

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Details of model designing

A.1 Spatial relationships contained in each edge type

Due to the different geometric structures of points, polylines, and polygons, the potential spatial relationships among these three node types also vary. Accordingly, the OSM heterogeneous graph we constructed incorporates nine edge types based on the permutations of different node types. As shown in Table 7, we summarize the spatial relationships contained by each edge type in detail. Specifically, while relationships between point nodes are represented using Delaunay triangulation, those among other node types are defined in terms of topological relations. After computing these spatial relationships, we encode them using one-hot encoding to represent the edge attributes within the heterogeneous graph.

Table 7: Topological relations between different node types

Node type	Point	Polyline	Polygon
Point	Delaunay triangulation	Touch, within	Touch, within
Polyline	Touch, contain	Touch, intersect, cover, cover by, equal	Touch, cross, cover by
Polygon	Touch, contain	Touch, cross, cover	Touch, overlap, cover, cover by, equal

A.2 The structure of OSM encoder

The OSM encoder is a lightweight heterogeneous GNN built upon GATConv, specifically designed to capture the spatial and semantic relationships embedded in OSM data. As illustrated in Fig. 4, the encoder comprises nine parallel GATConv layers, each dedicated to handling one of nine distinct edge types defined by the spatial relationships between node types. The encoder supports three node types—point, polyline, and polygon—allowing it to model the complex topology inherent in OpenStreetMap (OSM) data. Taking point-type nodes as an example, the encoder not only performs intra-type message passing (i.e., among points) but also enables inter-type interactions by propagating messages to polyline and polygon nodes through separate GATConv layers. Conversely, it also gathers contextual information from polyline and polygon neighbors using additional GATConv layers. This cross-type message exchange ensures that each node embedding is enriched by both its own type's local structure and the information from other geometric types. By explicitly modeling heterogeneous relations such as touches, within, crosses, and contains, the encoder effectively captures both the geometric connectivity and high-level geographic semantics of OSM features. The resulting node embeddings provide a rich representation of the spatial environment, which can be further integrated with RS encodings for downstream tasks.

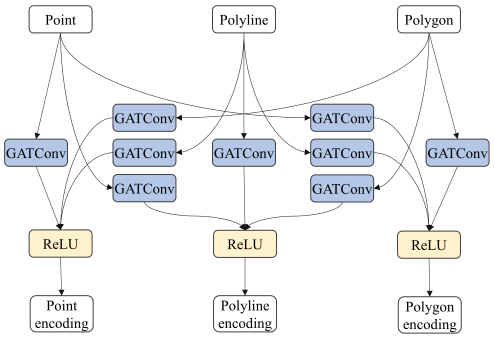


Figure 4: Detailed structure of the OSM encoder.

A.3 Position embedding of OSM objects

We use sinusoidal position embedding to uniformly represent the spatial positions of points, polylines, polygons, and image patches, thereby implicitly establishing spatial associations for fine-grained RS-OSM fusion. As mentioned in Sec. 3, we sample key-points and compute the average of the sampled points' position embeddings to capture the spatial coverage of polylines and polygons during the object-patch level integration. For a polyline vector, we sample three key-points, i.e., two endpoints and midpoint. For a polygon vector, we first compute its centroid. Then, using the centroid as the center, we sample three points inside the polygon at random radii and angles, resulting in four points to represent its position and coverage. The sinusoidal position embedding is computed for each

sampled key-point, and their average is taken as the final position embedding for the corresponding vector. In Appendix D, we evaluate the effect of the sampling number of key-points on model performance.

A.4 The structure of object-patch fusion encoder

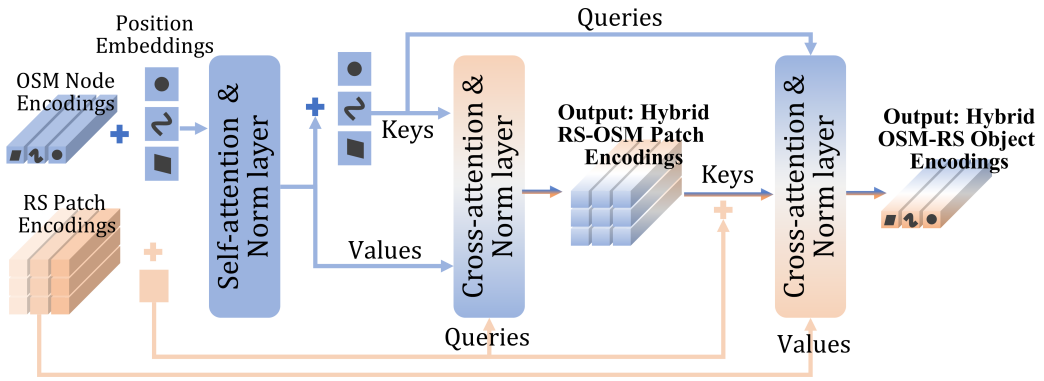


Figure 5: The detail structure of object-patch fusion encoder

As illustrated in Fig. 5, the object-patch fusion encoder is designed to effectively align and integrate RS image patches with geographic objects (node features) from OSM. It comprises a self-attention layer followed by two cross-attention layers, forming a lightweight yet expressive architecture for cross-modal fusion. The initial self-attention layer captures intra-modal relationships within the OSM object features, allowing the model to establish a coherent representation of the geographic context before engaging in cross-modal interactions.

To enhance the model's spatial sensitivity, position embeddings are added to both the keys and values prior to each attention operation. This spatial conditioning enables the encoder to better model the relative locations of RS patches and OSM objects, ensuring that attention weights are not only content-driven but also spatially aware. As a result, the model can form stronger associations between elements that are geographically close, leading to more accurate and meaningful fusion. The two cross-attention layers are asymmetrically structured to support bidirectional information exchange between modalities. The first cross-attention layer projects RS image patches as queries and OSM objects as keys/values, producing hybrid RS–OSM patch encodings that embed geographic context into visual features. Conversely, the second cross-attention layer uses OSM objects as queries to attend over RS patches, generating hybrid OSM–RS object encodings that incorporate visual cues into semantic representations of geographic objects. These two outputs are both fine-grained multimodal representations— the image patch level and the geographic object level, respectively — thereby equipping the model with the flexibility to support a broad range of downstream tasks.

B Details of the pretraining dataset

The pretraining dataset is derived from SkyScript-top30 [31], which contains 1271431 RS-OSM sample pairs. According to SkyScript, the RS images are originally downloaded from Google Earth Engine (GEE) platform, including 10 image collections like National Agriculture Imagery Program and Harmonized Sentinel-2, with a geographic coverage for all continents except Antarctic. The detailed information regarding the sources and distribution of RS images can be found in [31]. As for the OSM data, each RS image used in this study contains metadata on geographic location and timestamp, and we leverage them to retrieve corresponding OSM data via the Overpass API, which allows querying OSM data within a specified time range. After retrieving OSM data via Overpass, we apply rule-based cleaning to remove common issues—such as fixing or removing invalid polygons (e.g., self-intersections), eliminating duplicate/conflicting objects, and filtering tags with formatting or spelling errors.

To ensure diversity and reduce training burden, we remove sample pairs with overlapping spatial coverage and filter out samples with unavailable OSM data. According to the statistics shown in Fig. 6(a), each sample contains around ten OSM vector objects of points, polylines, and polygons on

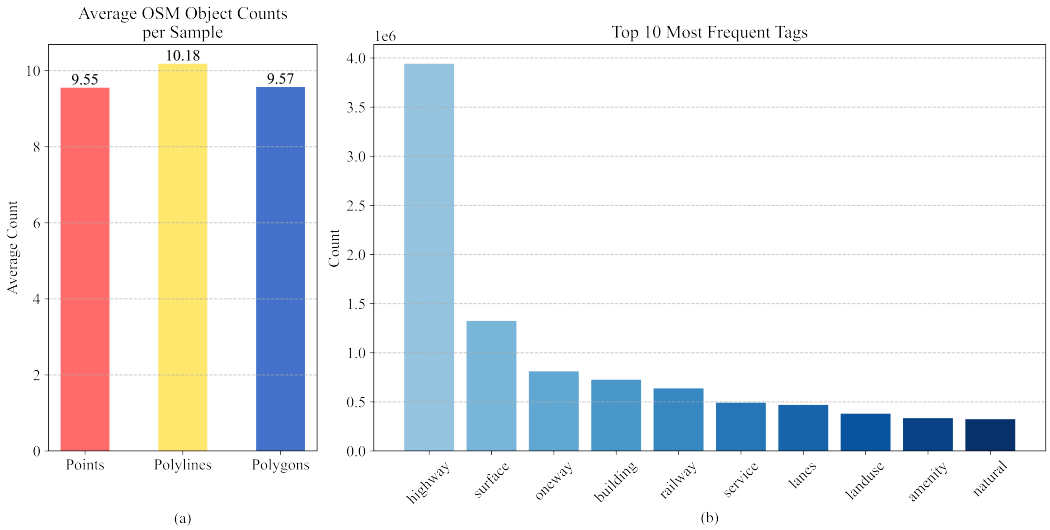


Figure 6: (a) The average counts of each kind of OSM object per sample. (b) The top 10 most frequent tags in the GeoLink pretraining dataset.

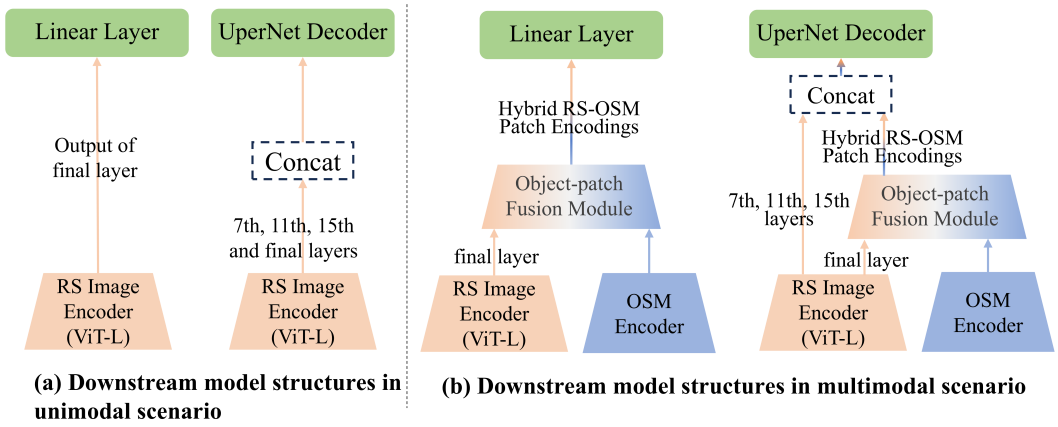


Figure 7: Model structures for diverse downstream tasks in unimodal and multimodal scenarios.

average. The pretraining dataset includes a total of 2,790 types of tags, with the top 10 most frequent ones shown in Fig. 6(b), namely: highway, surface, one way, building, railway, service, lanes, land use, amenity, and natural.

C Details of downstream task settings

C.1 The selection of baseline models

Ideally, comparisons should be made among models with identical spectral and modality settings, but in practice, this is often difficult to achieve. The diversity of spectral configurations in current RS FMs indeed makes idealized comparisons challenging, as many models are tailored to specific satellite spectral bands. For example, CROMA is built on Sentinel-1/2 data, while Prithvi-EO-2.0 [1] uses six channels shared by Sentinel-2 and Landsat: Blue, Green, Red, NIR, SWIR1, and SWIR2. Therefore, several prior works [1, 17, 18] have adopted a pragmatic approach by evaluating models with different spectral inputs on the same benchmark datasets. Although this may not ensure fully aligned comparisons, it provides meaningful insights into the models' generalization on common datasets.

To the best of our knowledge, there is currently no RS FM that uses the exact same data types as GeoLink (RS and OSM data), making it difficult to directly compare with fully modality-aligned baselines. The starting point of our experimental design is to explore how, and to what extent,

Table 8: Detailed information of benchmark datasets for multimodal downstream tasks: UFZ, UV, POP, and CO2.

Benchmark	UFZ	UV	POP	CO2
Experimental region	Chicago metropolitan area, Singapore, Shenzhen	Beijing, Shanghai	Same as UFZ	Same as UFZ
RS image source	ArcGIS World Imagery, Bing Map	ArcGIS World Imagery	ArcGIS World Imagery, Bing Map	ArcGIS World Imagery, Bing Map
GSD	1m, 3m	1m	1m, 3m	1m, 3m
Bands	RGB	RGB	RGB	RGB
Annotation source	Official statistics: Chicago Singapore Shenzhen	Expert annotation	WorldPop Dataset	ODIAC Fossil Fuel Emission Dataset
Annotation processing	Manual refinement, reclassification, spatially aligned cropping	Spatially aligned cropping	Resampling, spatially aligned cropping	Resampling, spatially aligned cropping
Image cropping size	224×224	224×224	224×224	224×224
Sample count	60,970	1,899	59,284	47,607

integrating OSM data can enhance RS FM. To fairly and comprehensively evaluate GeoLink, we consider baselines designed from different perspectives: (1) Unimodal baselines that use the same RS spectral modality as GeoLink, including GASSL, ScaleMAE, and Cross-scale MAE; (2) Multimodal baselines that incorporate additional modalities, such as MMEarth, CROMA, and DOFA. Different strategies are employed to adapt multimodal baselines to the RGB-only testing benchmark. The original DOFA model includes a dynamic projection module as a tokenizer to normalize various spectral inputs, allowing it to directly process RGB images without additional modifications. For CROMA and MMEarth, we follow PANGAEA-bench [17] employs zero-padding to fill in missing spectral bands for input.

C.2 Model structures for downstream task evaluation

As illustrated in Figure 3, we present the model architectures employed for various downstream tasks under both unimodal and multimodal settings. For linear classification and regression tasks, the unimodal approach directly utilizes the encodings output from the final layer of the RS image encoder—either the [cls] token or the mean of patch features—as input to a linear layer (left side of Fig. 7(a)). In the multimodal setting, this input is replaced by the mean of the hybrid RS-OSM patch encodings (left side of Fig. 7(a)). For semantic segmentation, we adopt the UperNet decoder. Under the unimodal setting, encodings from the 7th, 11th, 15th, and final layers of the RS encoder are fed into the decoder. In the multimodal setting, the features from the final layer are replaced with the hybrid RS-OSM patch encodings, while all other components remain consistent with the unimodal configuration. Overall, aside from minor differences in feature dimensions, the model architecture remains consistent between the unimodal and multimodal settings. The multimodal encoder can function as a plug-and-play component that integrates seamlessly into the RS FMs, and we combine it with Scale-MAE to build Scale-MAE+OSM for evaluation in Sec. 4.

C.3 Construction of the multimodal benchmark datasets

In this study, we construct three multimodal benchmark datasets to evaluate GeoLink's capability in addressing complex and comprehensive geographic tasks, including urban functional zone segmentation (UFZ), urban village identification (UV), population density estimation (POP), and carbon emission estimation (CO2). Each dataset consists of spatially aligned RS images, OSM data, and the corresponding task-specific annotations. Table 8 is a detailed overview of the construction procedures and key specifications of the four datasets. To begin with, considering factors such as geographic

Table 9: Hyperparameters for downstream tasks. LP: linear probing, Frz: freezing, FT: finetune, Cls: classification, Seg: segmentation, CD: change detection, LR: learning rate, CE: cross-entropy, MSE: mean square error

Task	Cls (LP)	Cls (FT)	Seg/CD (Frz)	Seg/CD (FT)	UFZ/UV	POP/CO2
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Batch size	64	64	64	64	32	256
LR	{1,3,5} e{-2,-3,-4}	{1,3,5} e{-2,-3,-4}	1e-4	1e-4	{1,3,5}e {-3,-4,-5}	{1,3,5}e {-1,-2,-3}
LR multiplier	–	0.01	–	0.01	0.01	0.01
Weight decay	0.05	0.05	0.05	0.05	0.05	0.05
Beta	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999
Epoch	50	50	80	80	30	30
LR scheduler	Multi-step	Multi-step	Multi-step	Multi-step	Cosine	Cosine
Default splits (train/val/test)	50/10/40 %	50/10/40 %	Same as PANGAEA-bench	Same as PANGAEA-bench	50/10/40 %	50/10/40 %
Loss function	CE	CE	CE/dice	CE/dice	CE	MSE

characteristics, level of development, and data availability, we select several representative experimental regions for the four tasks, including the Chicago metropolitan area, Singapore, Shenzhen, China and so on. High-resolution RS images for these regions are acquired from ArcGIS World Imagery and Bing Map, covering three RGB bands with GSDs of 1m or 3m. Corresponding OSM data are retrieved via the Overpass API. Subsequently, we construct high-quality annotated labels required for the four tasks through a combination of web data collection, expert annotation, and manual refinement. For the Urban Functional Zone (UFZ) task, we obtained original urban planning data from official statistics (sources listed in the table below), which are further refined using auxiliary references such as Google Maps to ensure their reliability. These refined data were then reclassified into the nine-category UFZ taxonomy: water, green space, farmland, undeveloped land, residential, commercial, institutional, industrial, and transportation. For the UV task, labels were generated entirely through manual annotation. For the POP and CO2 tasks, we leverage two well-established reanalysis datasets to obtain annotations—WorldPop and the ODIAC Fossil Fuel Emission Dataset—both of which have been extensively validated in prior research [66, 67, 68, 69, 70]. We process the RS images, OSM data, and annotations through resampling, reclassification, and other methods to obtain the final benchmark datasets.

C.4 Details of downstream evaluation protocols

To ensure the reproducibility of our experiments, we provide detailed hyperparameter settings for all downstream tasks in Table 9. All tasks are optimized using AdamW with a weight decay of 0.05 and β values set to [0.9, 0.999]. To preserve the original characteristics of the data, no data augmentation is applied in any of the tasks. For classification tasks, we perform a grid search over learning rates and report the best results for each model on each dataset. The data is split into 50% for training, 10% for validation, and 40% for testing. All results are averaged over three runs with different random seeds. For semantic segmentation and change detection tasks, all the settings of data split, learning rate, and loss function follow the default of the PANGAEA-bench [17]. Notably, at the time of paper submission, PANGAEA-bench has not yet released an officially processed version of the FiveBillionPixel dataset. Therefore, we use the dataset provided in the original FiveBillionPixel paper [58], crop the images to a size of 520×520 as input, and follow the original data split for consistency. For the UFZ, UV, POP, and CO2 tasks, we also conduct learning rate searches individually to ensure optimal performance for each task.

D Other experiments

D.1 Performance under limited annotations

Performing well under limited labeled data is one of the key metrics for evaluating FMs. To assess this, we conduct experiments using only 10% of the samples on both linear probing and UFZ tasks. Table 10 reports the average results of linear probing across seven datasets, as well as the results

for UFZ under unimodal and multimodal settings. Compared with Fig. 1 and Fig. 3, GeoLink demonstrates even more significant advantages in this scenario, indicating its stronger adaptability to downstream applications with limited training samples.

Table 10: Performance comparison on linear probing (top-1 accuracy %) and UFZ (mIoU %) tasks using only 10% samples for training.

Model	Linear probing	UFZ-unimodal	UFZ-multimodal
Scale-MAE	83.04	39.08	–
GeoLink	86.17	48.48	53.12

D.2 Impact of loss weights

GeoLink incorporates three distinct learning objectives, and their relative weighting can influence the model's performance. In this study, we always fix the weight of the image reconstruction loss to 1 and focus on adjusting the weights of the region-image contrastive loss and the spatial consistency loss. The default weight for both is set to 0.01, and we conduct experiments by scaling each of them up and down by an order of magnitude. The results are presented in Table 11. We evaluate performance on two downstream tasks: linear probing and UFZ-multimodal. The results reveal a differentiated impact of the two losses. The region-image contrastive loss primarily affects the performance of linear probing, indicating its dominant role in optimizing the image encoder. In contrast, variations in the spatial consistency loss have a greater influence on the UFZ-multimodal task, suggesting it plays a crucial role in enhancing cross-modal feature fusion. Setting both loss weights around 0.01 yields a favorable balance, facilitating effective synergy among the three objectives.

Table 11: Ablation study of contrastive and consistency loss weights. Evaluation includes linear probing (top-1 accuracy %) and UFZ-multimodal (mIoU %).

Contrastive loss	Consistency loss	Linear probing	UFZ-multimodal
0.01	0.01	92.60	60.00
0.1	0.01	91.37	58.43
0.001	0.01	91.89	58.74
0.01	0.1	91.56	57.98
0.01	0.001	92.31	56.18

D.3 Key-point number for position embedding

Table 12: Ablation study on the number of key-points for polyline and polygon representations.

Polyline key-points	Polygon key-points	Linear probing	UFZ-multimodal
Centroid + endpoints	Centroid + 3 sampling points	92.60	60.00
Centroid	Centroid	92.32	59.46
Centroid + endpoints	Centroid + 5 sampling points	91.77	58.02

Since sinusoidal position embedding is not inherently designed to represent the spatial characteristics of polyline and polygon vectors, we propose to approximate their spatial coverage by sampling a set of representative key-points. The number of sampled points can directly influence the spatial representation and, consequently, the model's performance. Therefore, in this section, we conduct experiments to explore the impact of key-point number on downstream tasks. First, we represent the positions of both polyline and polygon solely by their centroids, without sampling any key-points. The results of this setting are shown in the second row of the Table 12. Interestingly, using only the centroid as the position proxy still yields competitive performance, with only a slight drop compared to the default setting (first line). Next, we fix the polyline key-points while increasing the number of polygon key-points. Specifically, using the centroid as the center, we randomly sample five (the defalt is three) additional points within the polygon at varying radii and angles, forming a total of

six points to represent its spatial extent. The corresponding results are presented in the third row of the Table 12, where we observe a noticeable performance degradation. This can be attributed to the nature of sinusoidal position embeddings: they encode positional information through a combination of multi-frequency sine and cosine functions. Averaging multiple such embeddings tends to smooth out high-frequency variations, potentially over-smoothing the positional signal and impairing the model's ability to distinguish spatial patterns. In future work, we aim to explore more expressive and principled methods of position encoding that can seamlessly handle point, polyline, and polygon geometries within a unified framework, thereby enabling more effective spatial correlation.

D.4 Visulization of GeoLink mapping results in UFZ and UV tasks

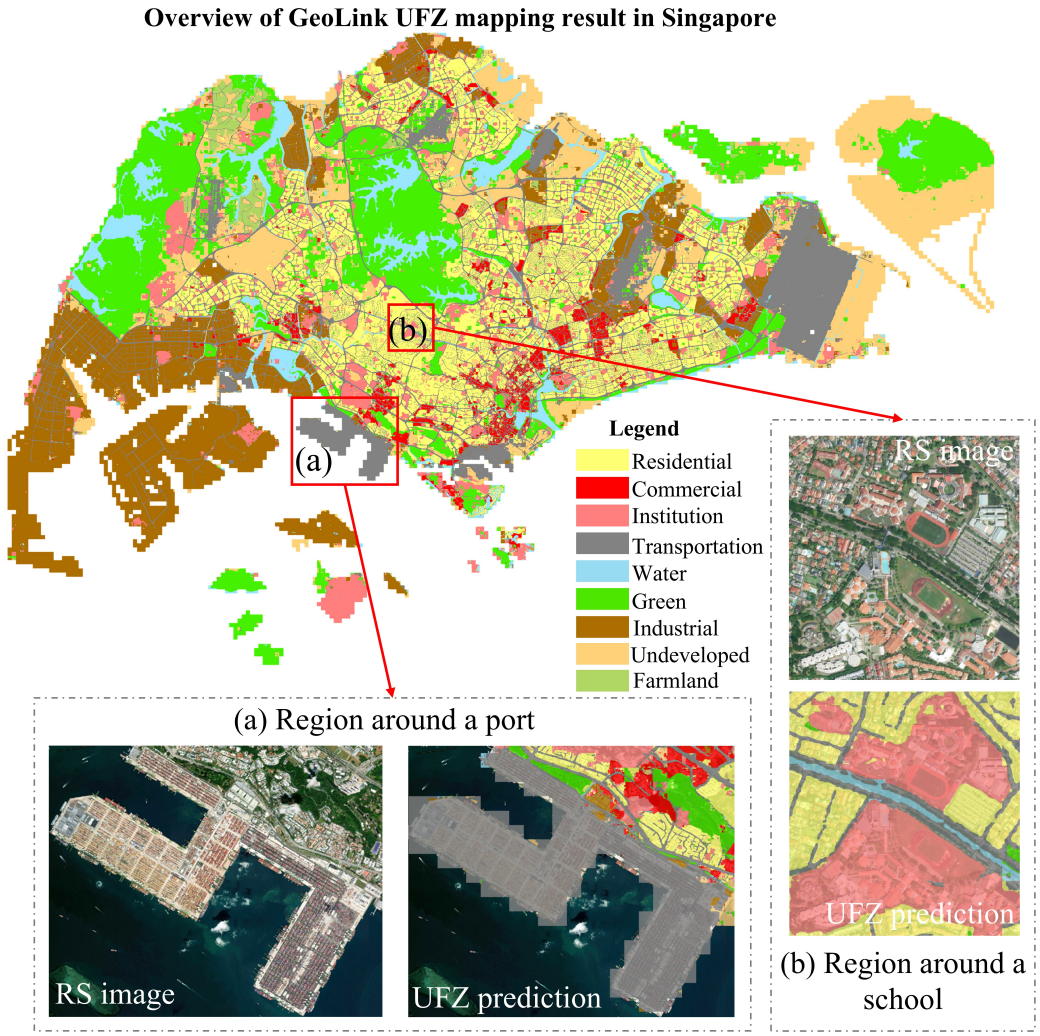


Figure 8: Overview and details of GeoLink UFZ mapping result in Singapore.

Large-scale geographic mapping is one of the most important real-world applications of RS data, and also a core capability that RS FMs should possess. In this section, we move beyond evaluating GeoLink solely through quantitative metrics, and instead demonstrate its potential for regional geographic mapping by visually comparing its predictions with real-world spatial layouts. Specifically, we leverage GeoLink to perform full-coverage predictions for Singapore and Beijing (multimodal setting) on UFZ segmentation and UV identification tasks, respectively. The predicted results are stitched together based on the geo-coordinates to generate complete UFZ map for Singapore and UV map for Beijing. As shown in Fig. 8, we first present the overall UFZ mapping results for Singapore. It is evident that the predicted spatial layout aligns well with the actual urban structure of

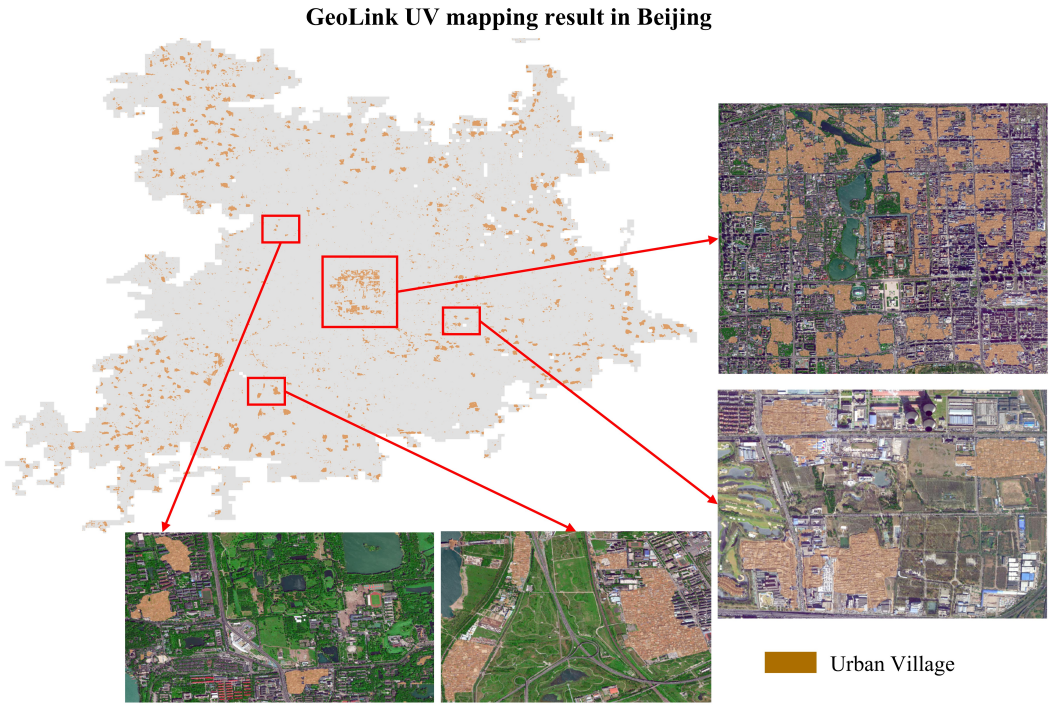


Figure 9: Overview and details of GeoLink UV mapping result in Beijing.

Singapore. For instance, the airport in the upper right and the industrial zone in the lower left are both accurately delineated. Even in the central area, where residential, commercial, and institution zones are intermingled, GeoLink is able to make reasonably precise distinctions. We further illustrate two representative local regions in detail: (a) the area surrounding the port (transport zone), where GeoLink clearly captures the spatial boundaries of the port infrastructure; (b) the area around a school (institution zone), where the model effectively leverages semantic information from OSM data and road boundary texture from RS image to delineate the functional area, despite its complexity. Fig. 9 showcases the results of urban village extraction and visualization in the Beijing region. UVs in Beijing are primarily concentrated within the second ring road, and GeoLink successfully identifies both the location and spatial extent of them (upper right), demonstrating its effectiveness in this task. Integrating RS FMs with mapping-related technologies to enable automated, high-precision geographic mapping holds significant promise across various geoscientific domains. At the same time, it represents a pressing technical challenge that remains to be fully addressed. We argue that future evaluations of RS FMs should incorporate real-world mapping performance as a critical metric, thereby continuously enhancing their practical applicability and deployment value in real geospatial scenarios.

D.5 Comparison with domain generalization model

We also compare GeoLink with CrossEarth [71], a vision FM for domain generalizable RS semantic segmentation, to evaluate the generalization ability. CrossEarth provides multiple domain generalization scenarios (e.g., urban-to-rural, different climate zones) and releases models trained on various source datasets. Based on our experimental setup and the available open-source models, we select the versions trained on ISPRS Potsdam (RGB) and CASID (Temperate Monsoon), as both are closely related to land cover tasks, similar to our two benchmarks on semantic segmentation task: FiveBillionPixels and AI4Smallfarms. This setting helps reduce the impact of domain gaps and better reflect the true performance of CrossEarth. In CrossEarth's original zero-shot setting, the source and target domains share the same label taxonomy. However, this is not the case for our benchmarks. Therefore, we modify the classification head of CrossEarth to match the number of classes in our benchmarks, while keeping all other layers and pretrained weights fixed. We finetune

the classification head on the target datasets and compare the results to GeoLink with freezing encoder. This setting helps CrossEarth adapt to our benchmarks and preserve its knowledge learned from the source domain.

Experimental results in the Table 13 show that CrossEarth exhibits strong generalization ability and performs reasonably well, although slightly below GeoLink. Notably, the model pretrained on CASID outperforms the one pretrained on Potsdam, possibly due to a closer distributional similarity to our benchmark datasets. This comparison highlights both CrossEarth's strength as a domain-generalized RS segmentation FM and GeoLink's effectiveness in leveraging OSM data to enhance the generalization of RS FM.

Table 13: Comparison GeoLink with CrossEarth on semantic segmentation tasks (mIoU %)

Model	Source domain	FiveBillionPixels	AI4Smallfarms
CrossEarth	ISPRS Potsdam	54.33	38.24
CrossEarth	CASID	56.18	41.07
GeoLink	/	60.49	43.26

D.6 Comparison with task-specific models

Currently, most geospatial mapping applications still employ task-specific models. To address the need for deciding between using task-specific models or FMs, we have designed an geographic transfer-ability experiment for discussion. We uitlize two task-specific models for comparison. (1) U-Net+OSM: using ImageNet-pretrained ResNet50 as the backbone to represent convolutional models. Since the fusion module in GeoLink is plug-and-play, we use the same OSM encoder and fusion encoder from GeoLink to ensure a fair comparison. (2) UpperNet+OSM: using ImageNet-pretrained ViT-L as the backbone, representing Transformer-based architectures.

Our UFZ benchmark includes cities with highly distinct geographic conditions: Singapore, Chicago (US), and Shenzhen (China). The original paper uses a mixed training/testing setting. To evaluate transfer-ability, we train models on one city and test on another. Specifically, we train GeoLink on Singapore and test on Shenzhen and Chicago. As shown in the Table14, although all models experience performance drops due to domain shift, GeoLink consistently outperform the task-specific models in the target cities, highlighting its superior generalization across regions. This demonstrates the effectiveness of pretrained FMs in capturing transferable geographic features.

Table 14: Comparison of geographic transfer-ability

Source Region	Target Region	U-Net+OSM	UpperNet+OSM	GeoLink (multimodal)
Singapore	Singapore	54.39	53.80	61.82
Singapore	Shenzhen	32.51	32.21	37.25
Singapore	Chicago	39.62	40.06	49.12