Knowledge Graph Enhanced for Zero-Shot Semantic Segmentation in Remote Sensing Imagery

Wubiao Huang [1], Huchen Li [1], Shuai Zhang [1], Haibing Liu [1], Zizhen Chen [1], Shihan Chen [1], Fei Deng [1, 2]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei, China - huangwubiao@whu.edu.cn, fdeng@sgg.whu.edu.cn
[2] Hubei Luojia Laboratory, Wuhan, Hubei, China - fdeng@sgg.whu.edu.cn

Keywords: Knowledge graph, Zero-shot, Semantic segmentation, Class enhanced, Remote sensing image.

Abstract

Zero-shot semantic segmentation (ZSSS) is a crucial task in remote sensing image understanding, yet existing methods still suffer from limited generalization to unseen classes. To address this issue, we propose a Knowledge Graph (KG) enhanced ZSSS framework, which introduces explicit hierarchical and relational information into class embeddings to achieve more structured and semantically consistent representations. Specifically, a KG class encoder is designed, consisting of the class enhanced query (CEQ) and class enhanced embedding (CEE) modules, which extract class-relevant subgraphs from a self-constructing Remote Sensing Semantic Class Knowledge Graph (RSSCKG) and generate knowledge-enriched embeddings through a text encoder. Experiments on three public remote sensing datasets demonstrate that the proposed method consistently improves performance across seven state-of-the-art ZSSS frameworks. The integration of KG-based embeddings yields significant gains in the evaluation metrics, with particularly strong improvements on unseen classes, while maintaining accuracy on seen classes. Compared with enhancement strategies based on large language model (LLM) generated descriptions, the proposed KG class encoder exhibit superior semantic separability and stability. These results validate the effectiveness, generalization, and scalability of the proposed framework for ZSSS in remote sensing imagery.

1. Introduction

Semantic segmentation of remote sensing imagery (RSISS) is a fundamental task in land-cover mapping, urban analysis, and environmental monitoring. Its objective is to assign a semantic label to every pixel in an image (Huang et al., 2023; Huang et al., 2024; Yang et al., 2024). In recent years, with the rapid advancement of deep learning techniques, this task has been pushed to a new level, and more research has emerged.

However, most existing RSISS models operate under closed-world assumptions, relying heavily on large quantities of manually annotated data for supervised training (Gao et al., 2025; Huang et al., 2025; Xu et al., 2024). When encountering new classes, these models typically require collecting and labeling a large number of corresponding samples and retraining, which leads to significant inefficiencies and resource waste. Moreover, the large number of classes, their complex distributions, and long-tailed, cross-domain differences make comprehensive annotation virtually impossible. This heavy dependence on manual labeling severely limits the generalization and scalability of existing models. To address these challenges, Zero-Shot Semantic Segmentation (ZSSS) has recently emerged as a promising research direction. ZSSS jointly models visual and semantic spaces, leveraging semantic reasoning to recognize classes that were unseen during training (Huang et al., 2026a; Ren et al., 2024).

Most current ZSSS methods adopt implicit knowledge representations, performing semantic embeddings of class labels by learning embedding representations from text data and mapping them to visual classifiers. However, this approach is constrained by the generalization capacity of the semantic model and the mapping function itself, making it difficult to capture structured semantic relationships. In contrast, a smaller number of studies attempt to employ explicit knowledge representations, such as knowledge bases or knowledge graphs (KGs), to model inter-class relationships and facilitate semantic generalization (Chen et al., 2023; Huang et al., 2026b). Nevertheless, due to the richness and ambiguity of remote sensing classes, there have been few studies on constructing knowledge graphs of semantic concepts in the field.

To fill this gap, this study compiles a large-scale remote sensing semantic class database by systematically reviewing existing literature, based on which we construct a Remote Sensing Semantic Class Knowledge Graph (RSSCKG). Building upon this, we propose a Knowledge Graph Enhanced Zero-Shot Semantic Segmentation (KG-ZSSS) framework, which explicitly models semantic structures and inter-class relations to enhance the model’s semantic understanding and transferability. The main contributions of this paper are summarized as follows:

1. Construction of a comprehensive RSSCKG, comprising 385 categories, 4 relationship types, and over 140000 semantic triples.
2. Design of a pluggable KG-based class encoder, which introduces class enhanced query (CEQ) and class enhanced embedding (CEE) to transfer semantic knowledge from the KG into the segmentation backbone. The CEQ module extracts semantically relevant subgraphs for target classes to filter noise and maintain semantic consistency. The CEE module aggregates node representations via a CLIP text encoder under relational constraints, producing knowledge enhanced class embeddings that align semantic and visual spaces.
3. Extensive experiments conducted on three publicly available remote sensing segmentation datasets demonstrate the effectiveness of the proposed KG class encoder and show that KG enhancement yields more distinct and interpretable semantic distributions.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the details of the proposed method. Section 4 reports experimental results and

ablation studies on multiple datasets. Section 5 concludes the paper and discusses future directions.

2. Related Work

2.1 Zero-shot Semantic Segmentation

Zero-Shot Learning (ZSL) was originally proposed for image classification tasks (Li et al., 2021b), aiming to transfer knowledge between seen and unseen classes through a shared semantic space. Early attempts extended this paradigm to semantic segmentation by integrating word embedding techniques, where class labels were projected into a high-dimensional semantic space, and each pixel was classified based on the similarity between its visual features and the nearest class embeddings. A representative work in this direction is ZS3Net (Bucher et al., 2019).

With the advent of vision-language pre-trained models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022b), ZSL has been further extended to pixel-level tasks. CLIP learns a unified multimodal embedding space through large-scale image-text contrastive learning. This makes it possible to directly calculate the semantic similarity between image regions and any text description, providing strong zero-shot transfer capability for segmentation. Subsequent studies explored various fusion and decoding strategies between textual and visual modalities. Representative examples include LSeg (Li et al., 2022a) with late feature fusion, Fusioner (Ma et al., 2022) with early fusion, and Cat-Seg (Cho et al., 2024) with mid-term fusion through cost aggregation. Several follow-up models further refined this framework, such as FGA-Seg (Li et al., 2025b) introducing pixel-text alignment loss, OVRS (Cao et al., 2025) incorporating rotation-aware perception, and domain-specific branches including GSNet (Ye et al., 2025) and RSKT-Seg (Li et al., 2025a), all achieving strong performance.

However, these methods rely solely on class names or template-based embeddings, which can lead to semantic ambiguity and incompleteness. To mitigate this issue, LMSeg (Tang et al., 2024) leverages large language models (LLMs) to generate rich class descriptions and embeds them using a text encoder. But this approach still assumes that class semantics are independent, neglecting hierarchical and relational structures. For remote sensing scenes with severe semantic overlap, this assumption can lead to the mixing of features from different classes in the embedding space, thereby reducing the model's generalization performance on unseen categories.

2.2 Knowledge Graphs in Remote Sensing

KGs represent entities and their semantic relations in an explicit, graph-structured manner, and have been widely applied in natural language processing, recommendation systems, and cross-modal retrieval (Ji et al., 2022). However, research on KGs within the remote sensing field remains limited. The first remote sensing knowledge graph (RSKG (Li et al., 2021a)) contained 117 entities, 26 relation types, and 191 triples, and was successfully applied to zero-shot classification tasks, demonstrating the potential of KG-based reasoning in this domain. Nevertheless, its small scale prevented comprehensive coverage of the diverse and fine-grained semantic classes present in remote sensing imagery. Later, Yao and Liu, (2024) proposed the mountain vegetation knowledge graph (MVKG), comprising over 4000 nodes and 10000 relations, which modeled the distribution and dynamics of mountain vegetation species.

3. Methodology

3.1 Overall Framework

As illustrated in Figure 1, we propose a KG-ZSSS framework, which differs fundamentally from conventional zero-shot segmentation architectures. The key idea is to leverage the semantic relations and hierarchical structures embedded within a KG to obtain domain-aware class embeddings, thereby improving the model’s ability to generalize to unseen classes and distinguish fine-grained semantic classes.

The proposed framework comprises four main components: a KG class encoder, an image encoder, a feature fusion and processing module, and a decoder. Specifically, a Vision Transformer (ViT)-based image encoder first extracts high-dimensional feature embeddings from the input remote sensing imagery. Meanwhile, the KG class encoder encodes each target class into a semantic embedding using KG information. Both encoders are initialized

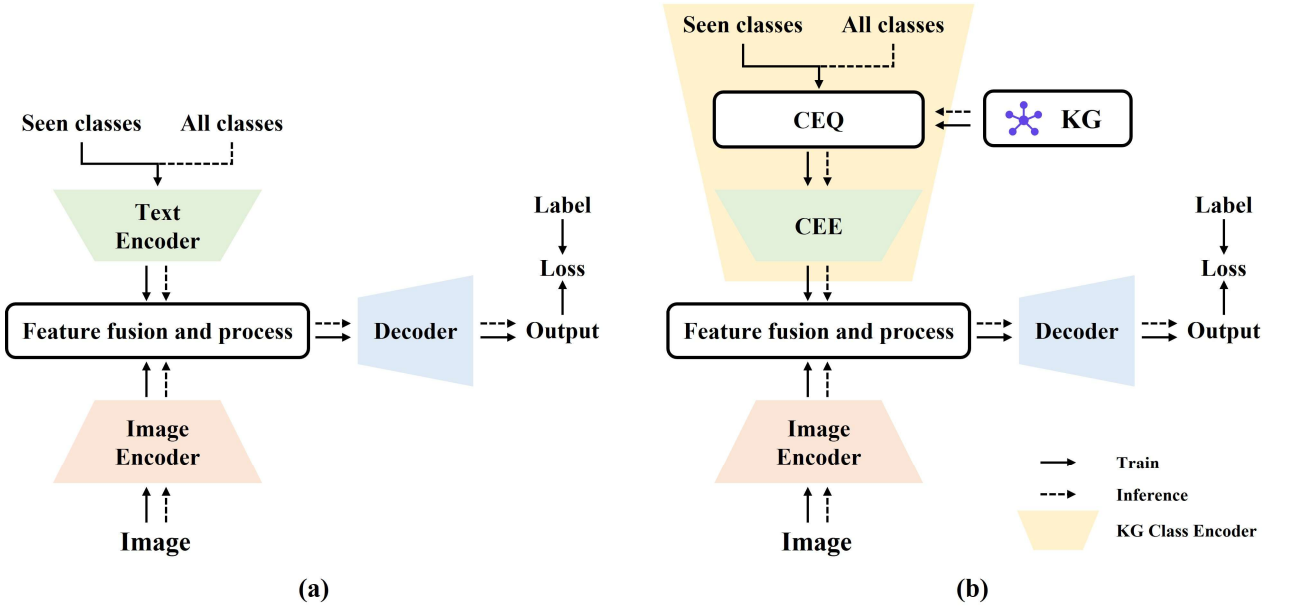

Figure 1. (a) Traditional ZSSS framework, (b) Knowledge graph enhanced ZSSS framework (Ours).

with pretrained CLIP weights to ensure strong alignment between the visual and text modalities within a shared embedding space. Next, the extracted image features and class embeddings are fed into the feature fusion module, where multimodal alignment and semantic mapping are performed to integrate the two representations. The fused features are then progressively upsampled and spatially reconstructed by the decoder to produce a pixel-wise class probability map, yielding the final segmentation result.

Within the KG class encoder, a CEQ module retrieves a semantic subgraph for each target class from the pre-constructed KG, and a CEE module encodes the subgraph into a knowledge-enchanced class representation.

During training, the model is supervised only on seen classes, with loss computed according to the corresponding ground truth labels. During inference, it can simultaneously predict both seen and unseen classes, thus achieving zero-shot segmentation. Unlike conventional approaches, our framework introduces a plug-and-play KG class encoder that can directly replace the text encoder in any existing ZSSS model without modifying other network components. Therefore, the last three modules of this paper do not provide fixed model designs, but follow the architecture of existing multiple models.

3.2 Knowledge Graph Construction

A KG is a large-scale semantic network consisting of nodes and edges, typically represented as triples of the form ⟨*head, relationship, tail*⟩. Each node denotes a semantic concept, and each edge expresses a semantic or hierarchical relation between concepts. The quality of nodes in a KG directly determines the integrity and effectiveness of its semantic database, therefore it is crucial to construct a comprehensive and structurally coherent remote sensing KG.

To this end, we first systematically collected and analyzed semantic segmentation datasets and publications in the remote sensing field, extracting all semantic classes and related node information. The extracted terms were then deduplicated and standardized to build a remote sensing semantic class database. Subsequently, domain experts manually annotated the relationships between nodes to form semantic triples. We defined four relation types: *synonym*, *contain*, *subclass*, and *different classes*. Among them, *synonym* and *different classes* are bidirectional symmetric relations, while *contain* and *subclass* are bidirectional inverse relations. The annotation process started from several commonly used high-level classes (e.g., *building*, *water*, *road*, *vegetation*, *forest*, *grass*, *brushwood*, *agricultural land*, *bareland*) and was iteratively expanded toward finer-grained subclasses or broader superclasses, thereby forming a hierarchical semantic network (Huang et al., 2026b). Finally, the constructed RSSCKG was implemented and visualized using the Neo4j platform for efficient querying and management.

As shown in Figure 2, the RSSCKG contains 385 semantic classes, 4 relation types, and 147840 triples, providing a rich and structured semantic foundation for knowledge-enhanced learning.

3.3 Class Enhanced Query

For a given set of semantic classes to be segmented, the CEQ module is used to extract semantically relevant subgraphs from the RSSCKG to support subsequent class representation enhancement. Specifically, for each semantic class, the CEQ module finds the corresponding node in the RSSCKG through precise matching. Starting from this node, it extracts all connected nodes and relations according to predefined relationship types to construct a knowledge subgraph for the target class. The corresponding Neo4j query command can be expressed as:

```
MATCH (a)-[r]->(b)
WHERE a.name = $name AND type(r) = $relationship
RETURN b.name AS Entity
```

where *$relationship* specifies the relationship type (*synonym* or *contain*), and *$name* is the name of the query class.

However, in practical applications, semantic overlap or hierarchical dependency may exist between classes. For example, *impervious surface* semantically includes *building*. Directly extracting subgraphs for both classes may result in redundant or overlapping semantics, thereby reducing discriminability. To mitigate such interference, after extracting all class-specific subgraphs, we check whether any subgraph is entirely contained within another. If such cases are detected, we perform a set-difference operation on the larger subgraph to remove the overlapping portion, ensuring that the final subgraph collection remains mutually independent and semantically disjoint. Importantly, this operation does not modify the global RSSCKG; it only constrains the class-specific subgraphs used for embedding generation. This acts as a discriminative constraint that prevents different target classes from relying on identical semantic evidence in the same segmentation episode, thereby improving class separability while preserving hierarchical knowledge in the KG.

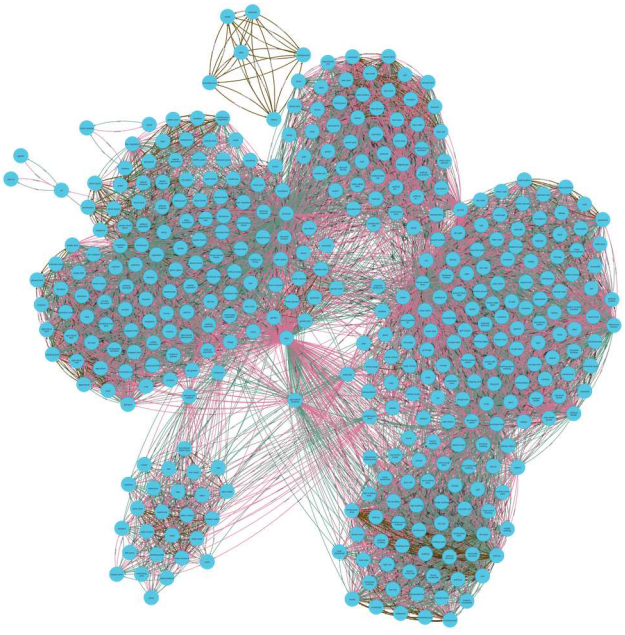

Figure 2. Visualization of the constructed RSSCKG.

3.4 Class Enhanced Embedding

Once the refined knowledge subgraphs are obtained, the CEE module converts them into high-dimensional class embeddings suitable for semantic segmentation tasks. The workflow of this module is illustrated in Figure 3.

Let the extracted subgraph $\mathcal{G}'$ consist of a node set $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$. Each node v_i is encoded using a pretrained CLIP text encoder, producing a fixed semantic embedding vector:

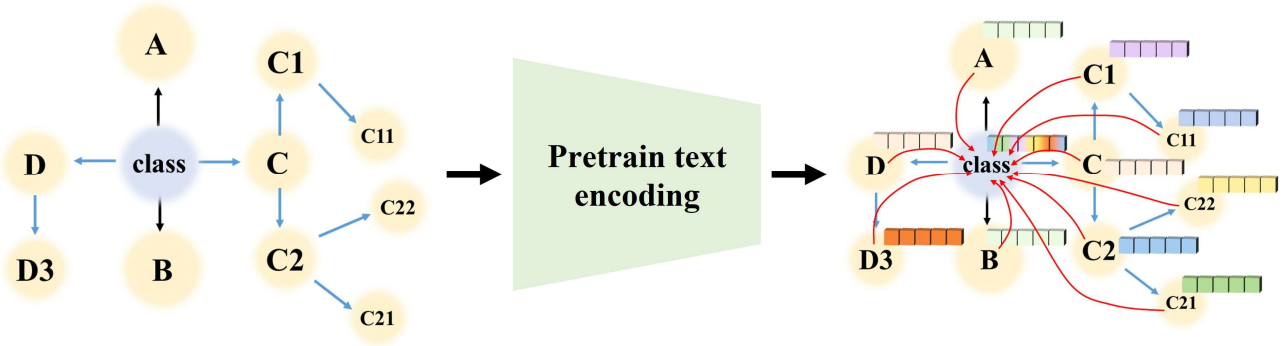

Figure 3. Workflow of the CEE module. The rectangular blocks of different colors represent different node embeddings.

$$h_i = f_{CLIP}(v_i), \quad v_i \in \mathcal{V} \tag{1}$$

Based on the edge connectivity and semantic hierarchy within the subgraph, average pooling aggregation is performed on node embeddings to obtain class enhanced representations:

$$z_c = \frac{1}{|\mathcal{V}|}\sum_{v_i \in \mathcal{V}} h_i \tag{2}$$

Finally, the set of knowledge enhanced class embeddings is formulated as:

$$Z = \{z_c | c \in C\} \tag{3}$$

where C denotes the set of target semantic classes.

Through this process, each class embedding integrates not only its intrinsic semantics but also the relational and hierarchical knowledge of its parent, child, and synonym classes, forming a richer and more structured representation in the semantic space. Consequently, the model can better capture inter-class semantic distances, leading to enhanced discrimination and improved generalization to unseen classes.

4. Experiments

4.1 Datasets and Implementation Details

4.1.1 Dataset: To comprehensively evaluate the proposed method, we conducted experiments on three publicly available remote sensing semantic segmentation datasets. Since these datasets were not originally designed for ZSSS, we manually partitioned their classes into seen and unseen classes to simulate the zero-shot scenario. The partitioning scheme is summarized in Table 1. All images were cropped into patches of size 512 × 512 for both training and inference.

EvLab-SS (Zhang et al., 2017): The dataset is originally obtained from the Chinese Geographic Condition Survey and Mapping Project, and each image is fully annotated by the Geographic Conditions Survey (NO.GDPJ 01—2013) standards. The average spatial resolution of each image is approximately 4500 × 4500 pixels. It contains 60 large-scale images captured from multiple platforms and sensors. During training, the unseen classes were defined as building, vegetation, and bareland according to task requirements.

OpenWUSU (Shi et al., 2023): The Wuhan Urban Semantic Understanding (WUSU) dataset focuses on urban structure and urbanization in Wuhan, covering key development zones such as Jiang’an and Hongshan districts, with a total geographic area of nearly 80 km². All images are acquired from GF-2 satellite imagery. During training, the unseen classes were defined as building, vegetation, water, and bare surface according to task requirements.

Potsdam
(http://www2.isprs.org/commissions/comm3/wg4/semantic-label-ing.html): This dataset represents a typical historical European city with dense buildings, narrow streets, and compact urban structures. It contains 38 images, each of size 6000 × 6000 pixels. During training, low vegetation and tree were merged into a single vegetation class.

Dataset	Seen classes	Unseen classes
EvLab-SS	building, road, water	farmland, garden, woodland, grassland, structure, digging pile, desert
OpenWUSU	road, bare surface	low building, high building, arable land, woodland, grassland, river, lake, structure, excavation
Potsdam	impervious surface, building, car	low vegetation, tree

Table 1. Seen and unseen classes used in the experiments.

4.1.2 Experimental setting: We employed the CLIP ViT-B/16 model as the backbone for both image and class encoders. All experiments were conducted on a Linux PC with two Intel (R) Xeon (R) 6133 CPUs and a NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory. During training, random scaling, random cropping, and random horizontal flipping were applied for data augmentation. We used the AdamW optimizer with a poly learning rate decay strategy, an initial learning rate of 0.0002, and a weight decay of 0.0001. The batch size was set to 4, and training was performed for a maximum of 30 k iterations.

4.1.3 Evaluation metrics: To comprehensively evaluate the performance of the proposed method on both seen and unseen classes, as well as its robustness, we calculate the Intersection over Union (IoU) metric for each class. Based on this, we calculate the seen class mean IoU ($smIoU$), unseen class mean IoU ($umIoU$), and their harmonic mean $hIoU$ to measure performance. The background class is excluded from all metric computations. The metrics are computed as follows:

$$IoU_k = \frac{TP_k}{TP_k + FP_k + FN_k} \tag{4}$$

$$smIoU = \frac{1}{|C_s|}\sum_{k \in C_s} IoU_k \tag{5}$$

$$umIoU = \frac{1}{|C_u|}\sum_{k \in C_u} IoU_k \quad (6)$$

$$hIoU = \frac{2 \times smIoU \times umIoU}{smIoU + umIoU} \quad (7)$$

where TP_k, FP_k, FN_k = the true positive, false positive and false negative pixels

k = class index

C_s, C_u = the sets of seen and unseen classes

4.2 Results

We evaluated the proposed method using seven state-of-the-art ZSSS baselines: Lseg (Li et al., 2022a), Fusioner (Ma et al., 2022), Cat-Seg (Cho et al., 2024), FGA-Seg (Li et al., 2025b), OVRS (Cao et al., 2025), GSNet (Ye et al., 2025), and RSKT-Seg (Li et al., 2025a). For each baseline, we replaced its original text encoder with our proposed KG class encoder, and evaluated both the baseline (w/o KG) and its KG-enhanced method (w KG) on all three datasets.

Table 2 presents the quantitative results. Across all seven baselines and datasets, models enhanced with the KG achieved significant performance improvements compared to the original version, with particularly notable improvements on unseen classes ($umIoU$) and the harmonic metrics ($hIoU$). These results demonstrate that KG enhancement substantially boosts generalization to unseen classes without degrading performance on seen classes, indicating effective semantic balance and robust transfer capability. Among the datasets, EvLab-SS exhibits the largest average improvement, with the mean $hIoU$ increasing by 11.64%. In OpenWUSU, the enhancement is moderate but consistently positive across all methods. For Potsdam, the average $hIoU$ rises by 7.57% after KG enhanced. Overall, RSKT-Seg achieves the best segmentation accuracy across all datasets, while traditional models such as Lseg and Fusioner perform well in simple-class settings but the performance drops sharply as class complexity increases.

Figure 4 shows the visualization results of different methods on three datasets. It can be seen that after introducing KG enhanced, each model has significantly improved in semantic boundaries, fine-grained class discrimination, and unseen class recognition. Specifically, clear improvements are observed in classes such as *digging pile* and *woodland* in EvLab-SS, *arable land*, *bare surface*, and *excavation* in OpenWUSU, and *tree* and *impervious surface* in Potsdam.

The above results confirm that by explicitly modeling hierarchical and relational semantics through the KG, the proposed framework enables the model to better capture inter-class dependencies in the semantic space, thereby achieving more robust zero-shot segmentation performance. Moreover, the proposed KG class encoder serves as a universal, easily integrable module, adaptable to diverse segmentation architectures with excellent transferability and scalability.

4.3 Comparison of Different Class Embedding Enhancement Methods

To further evaluate the semantic representation capability of the proposed method, we compared it with the class embedding enhancement strategy based on LLMs introduced in LMSeg (Tang et al., 2024). This method augments class embeddings by generating class descriptions through an LLM, thereby enriching semantic information. Figure 5 presents radar plots of the $hIoU$ metric on the EvLab-SS dataset for three variants: the baseline, the description-enhanced, and the KG-enhanced models. It is evident that in all cases, the KG-enhanced model consistently outperforms both the baseline and the description-enhanced versions, covering a significantly larger area in the radar plot. This demonstrates that KG enhancement not only achieves higher overall segmentation accuracy but also surpasses description-based augmentation in terms of stability and consistency.

While LLM-generated descriptions can indeed enrich the semantic representation of classes, they often lack explicit hierarchical and relational constraints, leading to redundancy or ambiguity in the embedding space. In contrast, the proposed KG class encoder leverages explicit hierarchical relationships and semantic links embedded in the KG, resulting in a more structured and interpretable class embedding space.

To intuitively illustrate the effect of different enhancement strategies on the embedding structure, we visualized the class embeddings produced by the baseline, description-enhanced, and KG-enhanced models using t-SNE, as shown in Figure 6. The results clearly reveal that the baseline model exhibits substantial

Backbone	Setting	EvLab-SS (%)			OpenWUSU (%)			Potsdam (%)		
		smIoU	*umIoU*	*hIoU*	*smIoU*	*umIoU*	*hIoU*	*smIoU*	*umIoU*	*hIoU*
Lseg	w/o KG	57.38	2.21	4.26	25.36	7.99	12.15	82.58	24.32	37.56
	w KG	59.23	11.66	19.48	25.25	14.90	18.74	82.95	24.28	37.56
Fusioner	w/o KG	58.08	2.56	4.90	21.54	6.55	10.05	72.41	22.95	34.85
	w KG	56.78	14.29	22.83	21.29	12.21	15.52	71.73	23.09	34.93
Cat-seg	w/o KG	47.59	6.34	11.19	8.13	8.88	8.49	30.34	24.14	26.89
	w KG	44.41	11.31	18.03	12.77	10.77	11.69	36.14	39.45	37.73
FGA-Seg	w/o KG	51.70	4.82	8.82	19.81	8.13	11.53	58.51	23.49	33.52
	w KG	49.35	14.18	22.03	19.91	12.32	15.22	70.09	36.90	48.35
OVRS	w/o KG	46.74	5.89	10.46	12.30	10.21	11.16	33.51	34.29	33.89
	w KG	41.57	11.84	18.42	14.86	11.63	13.04	34.74	43.52	38.64
GSNet	w/o KG	54.47	5.78	10.45	19.56	7.29	10.62	48.37	26.10	33.91
	w KG	47.57	14.48	22.20	18.43	13.83	15.80	51.48	39.19	44.50
RSKT-Seg	w/o KG	58.08	10.02	17.09	21.97	9.17	12.94	51.96	28.8	37.06
	w KG	52.62	16.97	25.67	22.45	17.06	19.38	60.98	40.89	48.95
Mean improve		**-3.22**	**+8.16**	**+11.64**	**+0.90**	**+4.93**	**+4.64**	**+4.35**	**+9.03**	**+7.57**

Table 2. Comparison of metrics between baselines and our method on seen and unseen classes

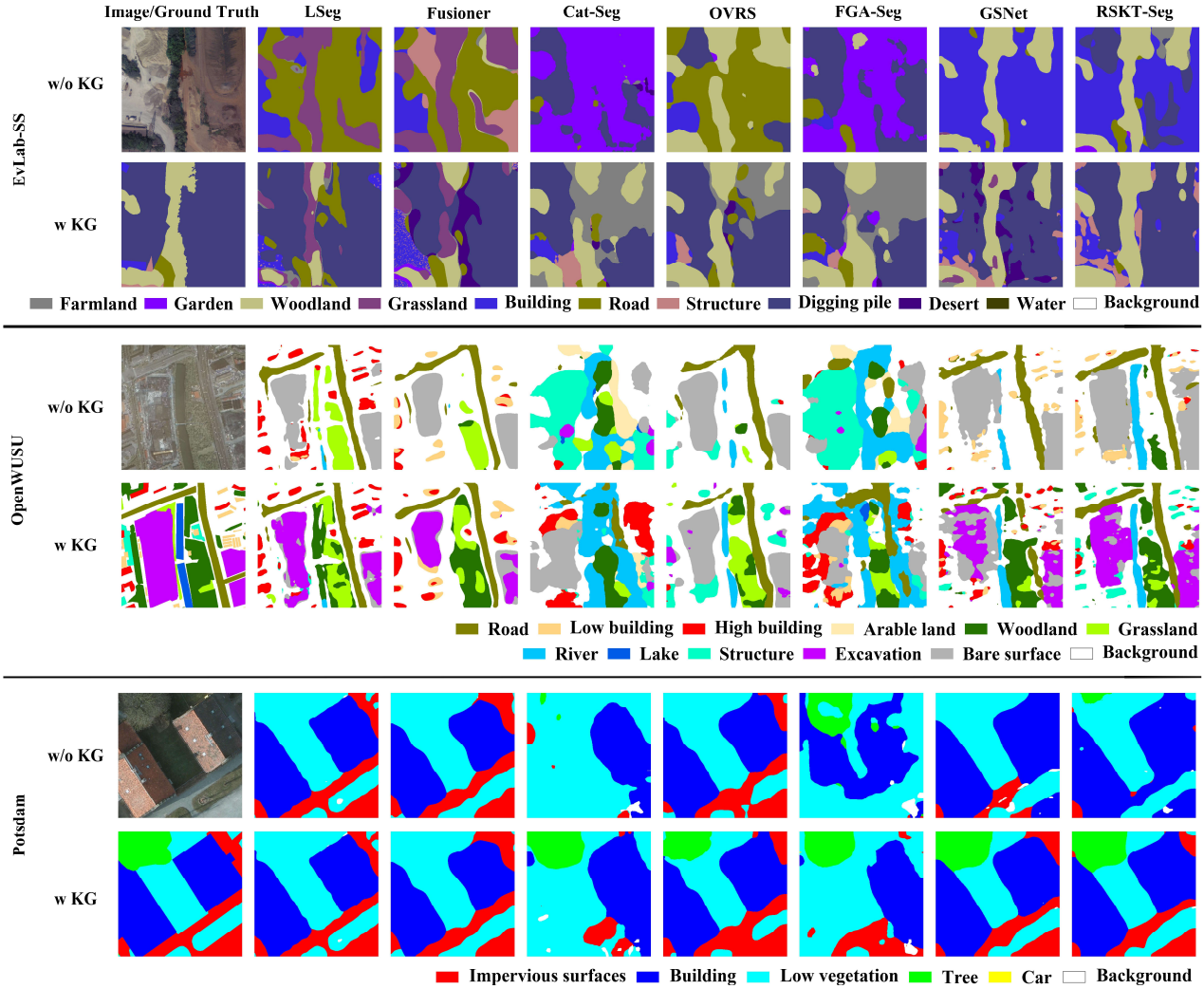

Figure 4. Qualitative results comparing GT, baseline, and our method

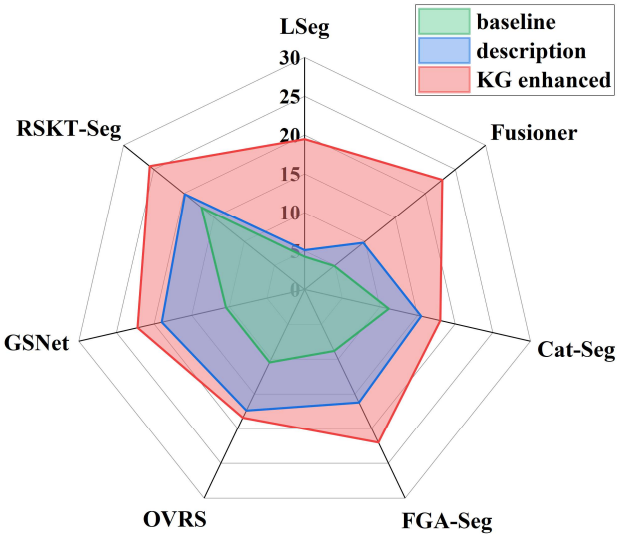

Figure 5. Radar plots of the $hIoU$ metric for different enhancement methods on the EvLab-SS dataset

class overlap, with blurred boundaries between classes, which indicating that traditional text encoders struggle to form semantically separable clusters in high-dimensional space. The description-enhanced model alleviates this issue to some extent, yielding partial class separation. However, semantic noise and fuzzy relationships inherent in LLM-generated descriptions result in the differences between semantically related classes are relatively small. In contrast, the KG-enhanced embeddings display much tighter intra-class clustering and larger inter-class separation, particularly for classes that are structurally similar but semantically distinct. This demonstrates that the proposed KG class encoder effectively constrains the embedding space through KG relations, enhancing both semantic separability and embedding consistency.

5. Conclusion

This paper presents a KG enhanced ZSSS framework for remote sensing imagery. A large-scale RSSCKG is constructed, and a KG Class Encoder based on subgraph querying and mean-pooling aggregation is designed to explicitly model hierarchical and relational dependencies among semantic classes. This enhances the richness and completeness of class representations. Extensive Experiments on EvLab-SS, OpenWUSU, and Potsdam datasets show that integrating KG-based embeddings consistently improves performance across seven state-of-the-art baselines, achieving up to 11.6% gain in $hIoU$, particularly on unseen classes. These results demonstrate that explicit semantic structure enhances generalization and class discrimination

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-125-2026 | © Author(s) 2026. CC BY 4.0 License.

compared with LLM-based or text-only embeddings. In future work, we will focus on incorporating dynamic graph reasoning and multi-source knowledge fusion to further improve model adaptability and interpretability in complex remote sensing scenes.

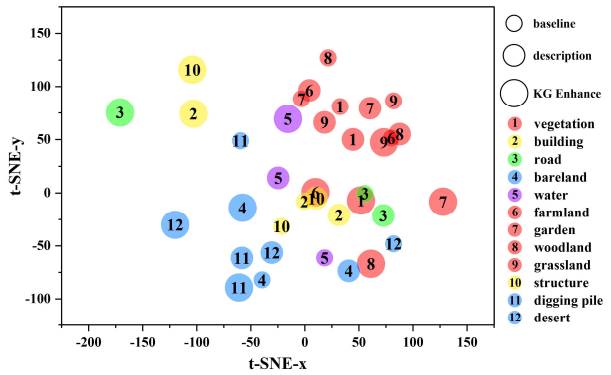

Figure 6. Visualization of t-SNE embedding of classes obtained by different enhancement methods

Acknowledgements

This work was funded by the National Key Research and Development Program of China (Grant No. 2025YFB3910300).

References

Bucher, M., Vu, T.-H., Cord, M., P\'rez, P., 2019. Zero-Shot Semantic Segmentation. *Advances in Neural Information Processing Systems*, Vancouver, Canada, 32.

Cao, Q., Chen, Y., Ma, C., Yang, X., 2025. Open-Vocabulary High-Resolution Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1-14. DOI: 10.1109/TGRS.2025.3559557.

Chen, J., Geng, Y., Chen, Z., Pan, J.Z., He, Y., Zhang, W., Horrocks, I., Chen, H., 2023. Zero-Shot and Few-Shot Learning With Knowledge Graphs: A Comprehensive Survey, *Proceedings of the IEEE*, 111(6), 653-685. DOI: 10.1109/JPROC.2023.3279374.

Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S., 2024. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 4113-4123. DOI: 10.1109/CVPR52733.2024.00394.

Gao, F., Fu, M., Cao, J., Dong, J., Du, Q., 2025. Adaptive Frequency Enhancement Network for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1-15. DOI: 10.1109/TGRS.2025.3558472.

Huang, L., Jiang, B., Lv, S., Liu, Y., Fu, Y., 2024. Deep Learning-based Semantic Segmentation of Remote Sensing Images: A Survey, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 8370-8396. DOI: 10.1109/JSTARS.2023.3335891.

Huang, W., Deng, F., Li, H., Yang, J., 2026a. HG-RSOVSSeg: Hierarchical Guidance Open-Vocabulary Semantic Segmentation Framework of High-Resolution Remote Sensing Images. *Remote Sensing*, 18, 213. DOI: 10.3390/rs18020213.

Huang, W., Deng, F., Liu, H., Ding, M., Yao, Q., 2025. Multiscale Semantic Segmentation of Remote Sensing Images Based on Edge Optimization. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1-13. DOI: 10.1109/TGRS.2025.3553524.

Huang, W., Ding, M., Deng, F., 2024. Domain-Incremental Learning for Remote Sensing Semantic Segmentation With Multifeature Constraints in Graph Space. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15. DOI: 10.1109/TGRS.2024.3481875.

Huang, W., Li, H., Zhang, S., Deng, F., 2026b. Reducing semantic ambiguity in open-vocabulary remote sensing image segmentation via knowledge graph-enhanced class representations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231, 837-853. DOI: 10.1016/j.isprsjprs.2025.11.029.

Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S., 2022. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494-514. DOI: 10.1109/TNNLS.2021.3070843.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 139, pp. 4904-4916.

Li, B., Dong, H., Zhang, D., Zhao, Z., Gao, J., Li, X., 2025a. Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing. *arXiv preprint* arXiv:2509.12040.

Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R., 2022a. Language-driven Semantic Segmentation. *International Conference on Learning Representations*.

Li, B., Zhang, D., Zhao, Z., Gao, J., Li, X., 2025b. FGAseg: Fine-Grained Pixel-Text Alignment for Open-Vocabulary Semantic Segmentation. *arXiv preprint* arXiv:2501.00877.

Li, J., Li, D., Xiong, C., Hoi, S., 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 162, pp. 12888-12900.

Li, Y., Kong, D., Zhang, Y., Tan, Y., Chen, L., 2021a. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179, 145-158. DOI: 10.1016/j.isprsjprs.2021.08.001.

Li, Y., Zhu, Z., Yu, J.-G., Zhang, Y., 2021b. Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12), 10590-10603. DOI: 10.1109/TGRS.2020.3047447.

Ma, C., Yang, Y., Wang, Y., Zhang, Y., Xie, W., 2022. Open-vocabulary semantic segmentation with frozen vision-language models. *arXiv preprint* arXiv:2210.15138.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.,

Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, PMLR, 139, pp. 8748-8763.

Ren, W., Tang, Y., Sun, Q., Zhao, C., Han, Q.L., 2024. Visual Semantic Segmentation Based on Few/Zero-Shot Learning: An Overview. *IEEE/CAA Journal of Automatica Sinica*, 11(5), 1106-1126. DOI: 10.1109/JAS.2023.123207.

Shi, S., Zhong, Y., Liu, Y., Wang, J., Wan, Y., Zhao, J., Lv, P., Zhang, L., Li, D., 2023. Multi-temporal urban semantic understanding based on GF-2 remote sensing imagery: from tri-temporal datasets to multi-task mapping. *International Journal of Digital Earth*, 16(1), 3321-3347. DOI: 10.1080/17538947.2023.2246445.

Tang, H., Zhao, Y., Huang, Y., Xu, M., Wang, J., Wu, Q., 2024. LMSeg: Unleashing the Power of Large-Scale Models for Open-Vocabulary Semantic Segmentation. *arXiv preprint* arXiv:2412.00364.

Xu, G., Jia, W., Wu, T., Chen, L., Gao, G., 2024. HAFormer: Unleashing the Power of Hierarchy-Aware Features for Lightweight Semantic Segmentation. *IEEE Transactions on Image Processing*, 33, 4202-4214. DOI: 10.1109/TIP.2024.3425048.

Yang, J., Ding, M., Huang, W., Li, Z., Zhang, Z., Wu, J., Peng, J., 2024. A Generalized Deep Learning-based Method for Rapid Co-seismic Landslide Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 16970-16983. DOI: 10.1109/JSTARS.2024.3457766.

Yao, Y., Liu, Y., 2024. The Construction of a Mountain Vegetation Knowledge Graph Incorporating With Geographical Principles, Maps, and Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15. DOI: 10.1109/TGRS.2024.3493455.

Ye, C., Zhuge, Y., Zhang, P., 2025. Towards Open-Vocabulary Remote Sensing Image Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9), 9436-9444. DOI: 10.1609/aaai.v39i9.33022.

Zhang, M., Hu, X., Zhao, L., Lv, Y., Luo, M., Pang, S., 2017. Learning Dual Multi-Scale Manifold Ranking for Semantic Segmentation of High-Resolution Images. *Remote Sensing*, 9(5), 500. DOI: 10.3390/rs9050500.