scientific reports

Check for updates

OPEN

# Large language model-driven knowledge graph reasoning for enhanced semantic segmentation

Jinhe Su[1], Xiaorong Zhang[1], Yang Luo[1], Yu Chen[1], Jinyuan Li[1], Shuting Chen[2] & Huilin Xu[1]✉

**Urban scene segmentation is essential for 3D city modeling and plays a crucial role in various remote sensing applications, including urban planning and environmental monitoring. While integrating knowledge graphs with scene segmentation has improved accuracy, existing methods often depend on dataset-specific knowledge graphs, limiting their generalizability across diverse remote sensing data. To address this, we propose a novel framework that leverages large language models (LLMs) to construct a universal knowledge graph from multi-source geospatial data and incorporate it into remote sensing semantic segmentation tasks, enhancing adaptability and robustness in urban scene understanding. Specifically, the framework comprises two key components: (1) a Graph Construction module that employs LLMs to extract cross-domain semantic relationships and build a universal knowledge graph, and (2) a Knowledge Graph Fusion module (KGFusion) that incorporates the graph into a semantic segmentation network to enhance semantic understanding. To evaluate the adaptability of our method across diverse domains, we curated a mixed dataset encompassing urban, rural, and port scenes. Experimental findings validate the efficiency and adaptability of our method, achieving 70.94% mIoU on the UAVid dataset and 63.23% on the Mixed dataset, outperforming the baseline by 0.43% and 1.04%, respectively. These results validate the robustness of our method in cross-domain scenarios and highlight its potential for broader applications in complex urban environments.**

Remote sensing image segmentation is a foundational step in integrating 3D city modeling with remote sensing data, as the quality of segmentation directly determines the efficacy of subsequent data fusion processes. Semantic segmentation is a fundamental computer vision task aimed at classifying every pixel in an image into specific categories[1–3]. In essence, it allows computers to understand and assign semantic labels to every pixel within an image. It has been extensively applied in diverse domains, including medical image segmentation[4,5], remote sensing image analysis[6], and autonomous vehicle navigation[7]. Since the introduction of fully convolutional networks (FCNs)[8] for semantic segmentation tasks, numerous advanced architectures have been developed. DeepLab[9] pioneered the use of backbone networks incorporating dilated convolutions and context extraction modules, with subsequent advancements such as DeepLabv2[10], DeepLabv3[11], PSPNet[12], and DenseASPP[13]. These methods utilize dilated convolutions and context extraction modules to capture fine-grained local contextual information, thereby significantly enhancing feature representation and pattern recognition capabilities. In remote sensing, semantic segmentation plays an essential role in urban analytics and expansive 3D city reconstruction.

As structured representations of knowledge and their interconnections, knowledge graphs have become indispensable in remote sensing image segmentation for 3D city modeling. By integrating the rich semantic information encoded in knowledge graphs, segmentation models can more effectively differentiate object categories and their spatial relationships, thereby improving segmentation performance, particularly in complex and ambiguous urban environments. For example, inter-category relationships, such as spatial adjacency and functional dependencies (e.g., connectivity within road networks), can offer valuable guidance for refining segmentation and enhancing the overall analysis of urban scenes.

Previous studies[14] have investigated the integration of domain-specific knowledge graphs to improve semantic segmentation performance. However, the dataset-specific nature of these approaches restricts their generalizability to new datasets and impedes broader adoption. Another study[15] proposed a knowledge graph for the remote sensing domain; however, its integration method for semantic segmentation is relatively complex, relying on entity-level connectivity constraints and inter-entity co-occurrence relationships. Recently, large language models (LLMs)[16–20] have rapidly advanced, with numerous models being developed. They provide

[1]The School of Computer Engineering, Jimei University, Xiamen 361021, China. [2]Chengyi College, Jimei University, Xiamen 361021, China. ✉email: xuhuilin@jmu.edu.cn

new opportunities for automating knowledge graph construction, excelling in text processing and semantic understanding, and enabling the generation of general-purpose knowledge graphs applicable to remote sensing segmentation. Leveraging these advancements holds the potential to further improve semantic segmentation and enable scalable, high-precision 3D urban reconstruction.

Traditional knowledge graphs are typically constructed for specific domains, where the entities and relationships are tightly coupled with predefined ontologies and expert-curated schemas. As a result, they exhibit limited generalization and poor adaptability when transferred to new domains, especially those with distinct visual perspectives or data distributions. This domain dependence often leads to semantic misalignment and suboptimal knowledge reuse.For example, in the UAVid dataset – a representative aerial-view remote sensing image collection – traditional ground-level knowledge graphs that focus on entities such as pedestrians or streetlights may become less effective. These objects are often visually distorted, or even irrelevant under aerial perspectives, leading to entity-relationship mismatches and degraded inference accuracy.In contrast, our approach leverages large language models (LLMs) that have been pretrained on vast and diverse corpora, enabling them to encode rich prior knowledge and semantic associations across multiple domains. This pretrained generalization allows for more flexible and robust reasoning when extracting and aligning entity relationships, even in specialized domains like remote sensing. As a result, our method offers better adaptability and transferability in constructing semantic knowledge graphs under varying data modalities and viewpoints.

To overcome the limitations of knowledge graphs in generalizability, particularly for remote sensing image segmentation in 3D city modeling, we propose a novel framework that integrates semantic prior knowledge from a knowledge graph into a semantic segmentation network. This unified approach enhances segmentation performance by leveraging structured semantic relationships. By harnessing the power of large language models (LLMs), we construct a large-scale knowledge graph specifically designed for urban environments. Additionally, we enhance DDRNet by integrating prior knowledge derived from the constructed knowledge graph.

In this knowledge graph, each node represents an urban object category, and the edges encode spatial or functional relationships, such as adjacency or connectivity. The knowledge graph, represented as a matrix, is seamlessly integrated into the segmentation network. Through graph convolution, the knowledge graph embeddings are incorporated as supplementary information into the network's feature representations. This integration significantly improves segmentation performance, particularly in complex urban scenarios such as road networks (Fig. 1(b)), as evidenced by the enhanced segmentation results compared to the baseline without knowledge graph integration (Fig. 1(a)).

Furthermore, to heighten the generalizability of our method, we introduce a Mixed dataset comprising diverse urban categories, including buildings, roads, vegetation, and other key elements. This dataset enables robust evaluation across various remote sensing scenarios. This advancement highlights the potential of knowledge graphs to enhance scene understanding in urban analysis and support precise, scalable 3D city reconstruction.

The key contributions of our work are outlined as follows:

1. We construct a knowledge graph using LLMs to guide the knowledge propagation between nodes in GCN and apply it to semantic segmentation tasks, improving the accuracy of semantic segmentation.
2. We modified traditional semantic segmentation networks by incorporating a knowledge graph fusion module, which enhances segmentation accuracy by leveraging the relationships between categories within the dataset.
3. Our method achieved improvements on the constructed multi-scene Mixed dataset compared to the UAVid dataset, demonstrating its effectiveness and generalization capability.

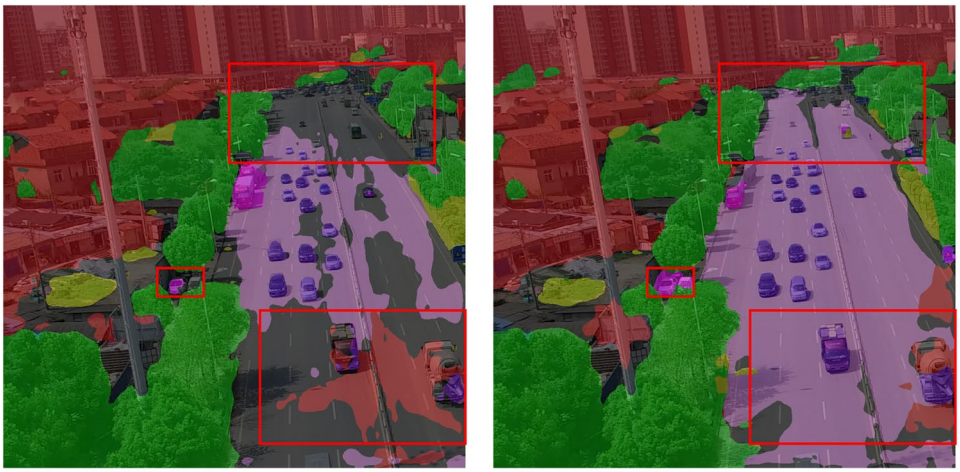


**Fig. 1**. By incorporating knowledge graph embeddings into our method(b), we can enhance semantic segmentation performance, as highlighted by the purple regions.

## Related Work

### Semantic Segmentation

Semantic segmentation, a fundamental problem in computer vision, aims to accurately classify every pixel in an image by associating it with a distinct semantic category.Since the introduction of FCNs[8], remarkable advancements have been achieved in this domain. Subsequently, the architecture evolved into SegNet[21] and U-Net[22], both of which employ the encoder-decoder architecture, a prevalent and widely utilized framework in semantic segmentation tasks. While ResNet[23] has been widely used as the encoder, more sophisticated networks like HRNet[24] and ICNet[25] have emerged to address complex segmentation scenarios. Decoders, on the other hand, typically focus on capturing global context and enhancing receptive fields. Recently, transformers[26] have been explored as decoders, leading to a series of vision transformer-based methods[27–31]. These methods aim to assist the encoder in extracting more effective pixel representations.

In addition to these standard semantic segmentation approaches, several recent studies in related domains have explored temporal, multiview, and knowledge-guided feature modeling, which provide inspiration for enhancing segmentation. Specifically, approaches like TSMGA[32] and STWANet[33] leverage temporal-spatial multiscale features and attention for change detection, while MVAFG[34] focuses on multiview fusion and advanced feature guidance to capture dependencies. Other techniques, such as the high-density iterative methods in HDIOD[35] and visual style prompt learning in VSP[36] for detail restoration, highlight the value of integrating structured knowledge and multi-scale context. These works highlight the potential of integrating structured knowledge or multi-scale context, which motivates our approach of incorporating knowledge graphs to improve segmentation accuracy.

In prior research, DDRNet[37] introduced the deep dual-resolution network and the deep aggregation pyramid pooling module.By processing high-resolution and low-resolution features concurrently, deep dual-resolution networks effectively capture both intricate details and broader contextual elements, leading to enhanced semantic segmentation performance.The Deep Aggregation Pyramid Pooling Module (DAPPM) effectively fuses multi-scale features through multi-scale pooling operations, improve the ability to recognize objects at different scales and ensure that the model can adapt to diverse scenarios. The equation of the Deep Dual-Resolution Network can be formulated as follows:

$$\begin{cases} F(x) = \mathrm{ReLU}(\mathrm{BN}(C_{3\times3}(\mathrm{ReLU}(\mathrm{BN}(C_{3\times3}(x)))))) \\ \mathrm{Output} = F(H) + (F(L) \uparrow 2) \end{cases} \tag{1}$$

where BN denotes Batch Normalization, ReLU represents the Rectified Linear Unit activation function, and $C_{3*3}$ signifies a 3x3 convolution. The DAPPM module downscales a feature map with an primary resolution of 1/64 of the input image by applying a series of strided convolutions with strides of $2^n$, thereby generating progressively lower-resolution feature maps. This process can be formalized as:

$$\begin{cases} F(X, K, s) = C_{1\times1}(C_{3\times3}(X \uparrow s)) \\ \mathrm{Output} = C_{1\times1}(F(X,5,2)||F(X,9,4)||F(X,17,8)||F(X,H\times W,1)||F(X,5,1)) + X \end{cases} \tag{2}$$

where K denotes the kernel size,s represents the upsampling factor, andF refers to a composite operation consisting of a convolution, an upsampling operation, and a subsequent $1\times1$ convolution.

### Knowledge graph

A knowledge graph[38], serves as a structured repository that delineates various entities along with their interconnections through a graphical framework. It serves as a massive semantic network, interconnecting diverse information from the world into a vast knowledge system. Knowledge graphs are versatile tools with extensive applicability across diverse domains, including large-scale integration and knowledge extraction from various data sources[39]. They can be applied to urban road scenarios, geographic sciences, and more. Constructing a knowledge graph involves information extraction, entity extraction, and relation extraction. For entity extraction, techniques like FudanNLP[40] and NLPIR[41] are commonly used. Relation extraction methods can be classified into four main types: supervised, semi-supervised, unsupervised, and open-world approaches. We observed a scarcity of projects that leverage knowledge graphs to direction semantic segmentation tasks. Therefore, we constructed a knowledge graph for our experiments to enhance segmentation accuracy.

### Graph convolutional networks (GCNs)

G raph Convolutional Networks (GCNs) are powerful neural network models designed to operate on graph-structured data. Their core principle lies in learning node representations by propagating information from neighboring nodes. Constructing a GCN typically involves data preparation, model definition, forward propagation, loss function design, backpropagation and optimization, and evaluation and tuning. The foundational framework of GCNs[42] was first introduced in 2017, demonstrating their effectiveness in semi-supervised classification. GraphSAGE[43] extended the GCN model, enabling inductive representation learning on large-scale graphs. Furthermore, the introduction of Graph Attention Networks[44] (GATs) further enhanced the performance of GCNs. APPNP[45] introduced a novel aggregation scheme that improves the performance of GCNs. Additionally, there has been research on GCNscan to enhance its performance[46,47].In this work, we aim to enhance semantic segmentation accuracy by leveraging GCNs to fuse the matrix representation of knowledge graphs with class embeddings from the dataset and previously proposed features.

## Method

This section outlines the proposed architecture specifically designed for semantic segmentation tasks. Figure 2 provides an overview of the architecture, which is composed of two key components: the Deep Aggregation Pyramid Pooling Module (DAPPM) and the Knowledge Graph Fusion Module (KGFusion). Specifically, Upon receiving an input image, the initial step involves its processing through a deep dual-resolution network to extract fine-grained details and global contextual information. For the low-resolution features, we use DAPPM to enrich the contextual information, followed by a $1\times1$ convolution for refinement before feeding the features into KGFusion. Meanwhile, we leverage large language models (LLMs) to construct a universal knowledge graph, which is converted into a matrix form and used as input to KGFusion. Additionally, category embedding vectors from the dataset are also fed into KGFusion. By leveraging the inter-class relationships in the knowledge graph, KGFusion enhances contextual understanding and improves generalization capabilities. The refined features obtained from KGFusion are subsequently integrated with those from the high-resolution branch. Ultimately, these combined features are directed through the segmentation head to produce the final prediction.

### LLM-based knowledge graph construction

Previous studies primarily focused on constructing task-specific knowledge graphs from open-source knowledge bases tailored to specific datasets, limiting their applicability to datasets from different scenarios. To address this limitation, we leveraged a large language models(LLMs)[19] to build a generalized knowledge graph. We chose ChatGLM because it provides a free API interface, enabling us to more efficiently refine the construction of our knowledge graph. Through this interface, we can conveniently extract and integrate structured semantic information, thereby enhancing the coverage and accuracy of the knowledge graph. We did not fine-tune the large language model but instead guided the model to generate data that meets our requirements through a set of designed rules and prompts. This approach not only reduces computational costs but also ensures the controllability and consistency of the data generation process, enabling it to more efficiently fulfill the demands of specific tasks. These include constraints such as:

- `The defined entities are {object1} and {object2}`
- `The output is in JSON format, unformatted, and on a single line, following the pattern: {"parent": entity1, "child": entity2, "relation_type": , "template_index": }`
- `Determine the relationship between {object1} and {object2}`
- `relation_type must be selected from the relationships defined in {object2}. There is no need for numbering; only the relationship names corresponding to the numbers are required. For example, relation_type should be a string.`

Here, `object1` and `object2` represent two categories whose relationship needs to be determined. We define a set of relations including *context*, *instance*, *precedence*, *composition*, *causality*, *input*, *output*, *similarity*, *manner*, and *association*.

The final output is a file in the following format:

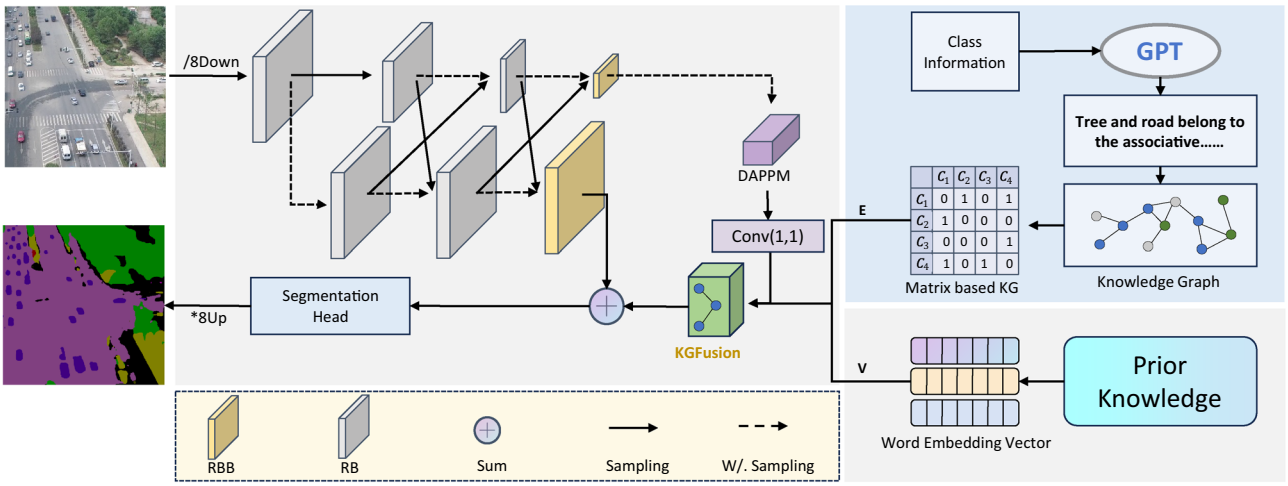


**Fig. 2**. Our network architecture consists of two primary components: a knowledge graph construction module and a knowledge graph fusion module(KGFusion). In the figure, E represents the matrix-based knowledge graph, and V denotes the class word embedding vectors. The construction module employs large language models to infer relationships between categories and definitions obtained from Baidu Baike according to predefined rules. The KGFusion module integrates the constructed knowledge graph into the network, enabling the model to leverage prior knowledge for enhanced semantic understanding and reasoning.

```
{
 "parent": entity1
 "child": entity2
 "relation_type":
 "template_index":
}
```

By extracting the relevant categories from this knowledge graph, we can construct a small-scale knowledge graph and integrate it into the semantic segmentation task.

Our knowledge graph includes a wide range of scenarios, such as urban areas, rural areas, neighborhoods, and more. Including buildings, roads, trees, moving vehicles, stationary vehicles, pedestrians, aircraft, ships, and walls. Attribute relationships, primarily spatial in nature, serve as semantic bridges between these entities, encompassing concepts like adjacency, orientation, parallelism, and enclosure. For example, a moving vehicle is typically enclosed by a road, and a tree is often adjacent to a road or building. Finally, we constructed a universal knowledge graph represented by triples. To effectively integrate this knowledge graph into the semantic segmentation network, we transformed it into a matrix form.

Furthermore, we employ GloVe 6B[48] word embeddings to obtain vector representations for each category, serving as foundational node embeddings. Each embedding is 300-dimensional, enabling the capture of rich semantic information and supporting reasoning about object relationships in semantic segmentation tasks. To facilitate integration with our subsequent KGFusion module, we converted these category embeddings into a pickle format.

We believe that the role of LLMs in constructing knowledge graphs lies in the following aspects: Firstly, large language models (LLMs) can significantly reduce the cost of manual annotation during the construction of knowledge graphs. By automatically extracting entities and their relationships from large-scale text data, LLMs accelerate the process of knowledge construction and updating, while avoiding the labor-intensive and error-prone nature of traditional manual annotation. Moreover, LLMs are capable of processing cross-domain and multimodal data, enabling the creation of more comprehensive and precise knowledge graphs.

Secondly, LLMs excel at understanding complex relationships between objects in intricate scenes. In semantic segmentation tasks, traditional methods often rely on pixel-level features, whereas LLMs, combined with knowledge reasoning, can enhance the model's ability to comprehend semantic associations among objects, resulting in segmentation outcomes that align more closely with human cognition. For instance, in agricultural scenarios, LLMs can leverage extensive prior knowledge to infer that "farmland is typically adjacent to water sources," thereby improving the recognition accuracy of targets such as irrigation facilities, farmland, and ditches in remote sensing or drone imagery, and reducing instances of missegmentation. This knowledge-driven approach not only deepens the model's understanding of complex real-world relationships but also significantly enhances its generalization capability, making it adaptable to a wider range of scenarios.

## Knowledge graph fusion module (KGFusion)

Knowledge graphs can capture semantic relationships between different categories, enabling a better understanding of the connections between objects. Our Knowledge Graph Fusion Module (KGFusion) incorporates knowledge graphs into semantic segmentation tasks, bolstering the model's capacity for generalization. In Section 3.1, we make use of large language models (LLMs) to extract a universal knowledge graph, which can be applied to scenarios with unknown classes or datasets with significant class discrepancies.To leverage the rich semantic information in knowledge graphs for semantic segmentation, we refer to[49,50] and make appropriate adjustments to adapt it to our network. We first represent the knowledge graph as a matrix, which allows us to efficiently encode the semantic relationships between different entities. Subsequently, we introduce class embedding vectors to capture the semantic characteristics of each class. By aligning these class embeddings with those in the semantic segmentation task, we can effectively narrow the divide between the elevated semantic insights within the knowledge graph and the foundational visual elements derived from the image.

As shown in Fig. 3, the KGFusion module comprises three essential elements: a graph convolutional layer, a semantic mapping layer, and a graph reasoning layer. The graph convolutional layer is instrumental in discerning the inherent connections among various entities through the dissemination of feature information across the knowledge graph's edges. This enables the model to assimilate contextual data from adjacent nodes, thereby enriching the feature depiction of every spatial area and promoting more precise semantic segmentation.

Given input feature vectors $X^l \in R^{D*H*W}$, a knowledge graph $E^{c*c}$, and class embedding vectors $v^{c*k}$, where the word embedding vectors V are derived from the GloVe library and the relational graph E is constructed from our custom knowledge graph. We use a GCN to normalize these inputs, ultimately obtaining a semantic graph enriched with contextual relationships. This semantic graph is then concatenated with the features (raw features) $X^l$ passed from the previous network layer. Subsequently, the processed features are amalgamated with the original features, culminating in the module's definitive output $X^{l+1}$. Specifically, we first input the category embedding vectors and the matrix-based knowledge graph into the graph convolutional network (GCN) for two rounds of graph convolution, enhancing the representation of local features. Furthermore, The core role of the Graph Convolutional Network (GCN) here is to deeply integrate the structured semantic information (such as entities and relationships) from the knowledge graph with visual features, thereby significantly enhancing the semantic segmentation model's ability to understand global context and complex category relationships. Since the graph convolution operation alters the feature distribution of each symbol node, it is necessary to map the symbol node representations back to the raw features. First, it is necessary to compute the compatibility $m_{v \to x_i} \in M_v$ between the local visual features and the symbol node, as shown in the following formula:

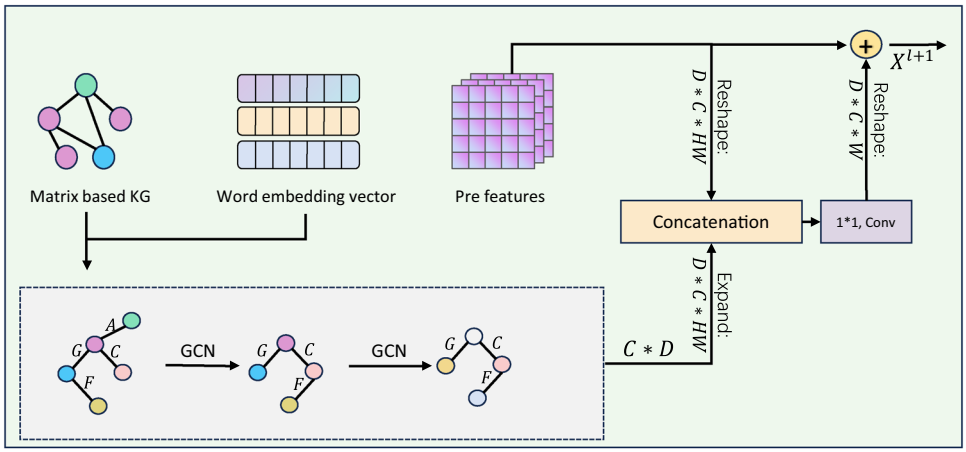


**Fig. 3**. An overview of our Knowledge Graph Fusion Module. In this module, we first apply graph convolution to both the matrix-form knowledge graph and the vector representation of categories twice to obtain a semantic graph. Subsequently, we align this semantic graph with the features derived from the network's prior operations. The aligned features are subsequently combined with the original features to generate enriched representations.

$$m_{v \to x_i} = \frac{\exp\left([v, x_i]W\right)}{\sum_{x_i} \exp\left([v, x_i]W\right)} \tag{3}$$

where W is a learnable weight matrix derived through a $1 \times 1$ convolutional operation, $x_i$ is each local feature, v is each symbol node. Next, this compatibility matrix is used to integrate semantic information into the local visual features, performing a weighted mapping of semantic features based on these relational weights. Subsequently, each local feature is updated through the weighted mapping of symbol nodes representing different semantic features. Finally, the input for the next convolutional layer is obtained as follows:

$$x^{l+1} = \sigma(M_v V^{l+1} W^a) + X^l \tag{4}$$

where $\sigma$ represents the ReLU activation function, $H_v$ denotes the compatibility matrix and $W^a$ is a learnable weight matrix for transforming the dimension.

Here, semantic mapping is primarily used to align the structured semantic information (such as entities and relationships) from the knowledge graph with visual features, thereby bridging the gap between visual features and semantic labels and further improving the performance of semantic segmentation tasks. Through semantic mapping, the model can associate pixel-level visual features in images with high-level semantic information (such as category names and entity relationships), enhancing the understanding of semantic consistency.

## Experiment

### Experimental settings

*Datasets*

**UAVid**[51]: It is designed for semantic segmentation endeavors within urban settings, with a particular emphasis on street-view scenarios. It comprises 8 semantic classes (Building, Road, Static car, Tree, Low vegetation, Human, Moving car, Background clutter) and is available in two resolutions: 3840x2160 and 4096x2160. The dataset consists of 42 sequences, partitioned into training (20 sequences), validation (7 sequences), and testing (15 sequences) sets. For our experiments, we resized each image to 1024x1024. The dataset is partitioned into training, validation, and testing subsets in an 8:1:1 ratio, containing 1728, 216, and 216 images, respectively.

**Mixed:** We constructed a multi-source heterogeneous hybrid semantic segmentation dataset by integrating several real-world remote sensing datasets (Potsdam, Vaihingen, LandCoverAI[52], UAVid[51], and UCM[53]) along with synthetic data sources including the SynthAer[54] virtual dataset and GTA-V-SID[55] game scene data. To ensure consistency across datasets, we developed a unified label mapping system that standardizes semantically equivalent but differently labeled categories (e.g., "road" vs. "lane") into common labels, while performing comprehensive format conversion and semantic alignment on all annotations. To address class imbalance issues, we implemented multi-scale data augmentation strategies during training, incorporating both geometric transformations and class-targeted augmentation techniques, which significantly improved model performance on minority classes. This hybrid dataset uniquely combines multi-perspective data from aerial, satellite, and UAV sources, effectively leveraging the complementary advantages of both real-world and virtual data. The dataset comprises 40,029 images across 19 categories, which we split into training, validation, and test sets at an 8:1:1 ratio. Tasks involving the semantic segmentation of remote sensing imagery necessitate the handling of varied and intricate image datasets, yet current collections frequently lack sufficient data volume and diversity in scenes. To overcome these constraints, we have constructed an innovative composite dataset that integrates

images from multiple remote sensing origins, sensors, and capture scenarios. This encompasses a broad spectrum of environments, including metropolitan areas, countryside locales, and natural terrains, showcasing a diverse array of object types and complex background diversity. The primary objective of this dataset is to enhance the generalization capabilities of models, enabling more robust adaptation to the challenging and heterogeneous nature of remote sensing imagery. It provides a solid benchmark for advancing algorithms in remote sensing image analysis.

*Implementation details*

Every model involved in this experiment was executed utilizing the PyTorch framework on a solitary NVIDIA GTX 3090 GPU. Specifically, we utilized the AdamW optimizer for the training of models, setting a foundational learning rate at 0.001 and a weight decay parameter at 0.01. The warmuppolylr scheduler was implemented to modulate the learning rate throughout the training process. A uniform batch size of 16 was maintained across all datasets. In the absence of pre-trained models, the training spanned 150 epochs. During the validation phase, random scaling augmentation ([0.5, 0.75, 1.0, 1.25, 1.5]) was applied. We evaluated the model using Intersection over Union (IoU) as evaluation metrics.

On the UAVid dataset, we conducted comparative experiments with three existing methods. Among them, DecoupleNet is a lightweight backbone network specifically designed for remote sensing vision tasks, suitable for various tasks such as image classification, object detection, and semantic segmentation. In this study, we compared its best-performing version, D2. On the Mixed dataset, we trained using these two methods separately, with the number of epochs, learning rate, and optimizer settings kept consistent with those of our proposed method to ensure the fairness of the experiments.

## Experiment results

The UAVid dataset, comprising diverse urban scenes captured by drones under varying illumination conditions, presents considerable difficulties for segmentation endeavors owing to the wide range of object scales. Consequently, improving segmentation performance on this dataset is highly valuable. Our experimental results, show in Table 1, illustrate that our introduced approach attains a mIoU of 70.94% on the UAVid dataset, surpassing the baseline methods DDRNet and UNetFormer by 0.43% and 3.14%, respectively. Notably, UNetFormer, with its hybrid architecture, has previously achieved a competitive mIoU on this benchmark. The experimental results on UAVid validate demonstrates the efficacy of our proposed method.

The UAVid test set was used to visually evaluate the segmentation performance of our method and DDRNet, as shown in Fig. 4. Our method excels at segmenting objects of various sizes and shapes in complex urban environments, especially in challenging UAV-based segmentation tasks. DDRNet exhibits limitations in segmenting large objects and dense, fine-grained objects accurately. Our approach addresses these shortcomings by integrating prior knowledge from knowledge graphs. This enables the model to refine predicted categories based on inter-class relationships, resulting in improved segmentation accuracy and reduced detail loss. The proposed method effectively mitigates the challenges posed by DDRNet, achieving superior visualization results.

The Mixed is a custom-curated dataset constructed by integrating multiple remote sensing image datasets, including a wide variety of environments like urban districts, countryside regions, and harbor zones. The dataset exhibits significant heterogeneity, including variations in class definitions and label inconsistencies across subsets, posing considerable challenges for semantic segmentation tasks. Notwithstanding these challenges, the Mixed dataset offers a benchmark for evaluating the generalization potential of segmentation techniques in intricate real-life situations. As demonstrated in Table 2, our method achieves competitive performance on the Mixed dataset, demonstrating relative improvements of 5.09% over UNetFormer and 1.04% over DDRNet. These results underscore the potential of our approach in effectively handling the heterogeneity and diversity of the dataset.

The qualitative visualization results on the Mixed dataset are provided as shown in Fig. 5. It is clear that our approach exhibits notable benefits compared to DDRNet. Specifically, due to the inherent complexity of the Mixed dataset, both UNetFormer and DDRNet achieve relatively low mIoU scores on this challenging benchmark. In contrast, our method effectively enhances the mIoU performance by incorporating a knowledge graph. Notably, the knowledge graph constructed in our approach exhibits superior generalizability and adaptability compared to DDRNet. This improvement can be ascribe to the fact that the Mixed dataset places a higher demand on modeling category relationships than the UAVid dataset, further highlighting the robustness of our approach.

In addition, we performed ablation experiments to evaluate the performance of various large language models (LLMs) in knowledge graph construction. For detailed comparative results across different large language models, please refer to Supplementary Table 1 and its corresponding experimental section in the Supplementary

| Method | Building | Road | Tree | Vegetation | MovingCar | StaticCar | Human | Clutter | mIoU | Params(M) | GFlops |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UnetFormer[56] | 87.40 | 81.50 | **80.20** | 63.50 | 73.60 | 56.40 | **31.00** | 68.40 | 67.80 | 11.7 | 56.9 |
| DecoupleNet D2[57] | 85.40 | 80.60 | 78.80 | 62.10 | 74.10 | 49.70 | 30.80 | 65.10 | 65.80 | 6.8 | 32.1 |
| SegFormer[58] | 86.30 | 80.10 | 79.60 | 62.30 | 72.50 | 52.50 | 28.50 | 66.60 | 66.00 | 13.7 | 63.3 |
| DDRNet (baseline)[37] | 91.64 | **84.17** | 78.82 | **71.95** | 72.09 | 67.65 | 27.14 | **70.64** | 70.51 | 5.73 | 4.91 |
| Ours | **91.79** | 83.62 | 78.56 | 71.84 | **74.15** | **69.50** | 27.97 | 70.08 | **70.94** | 5.79 | 4.93 |

**Table 1**. Comparison of results on the UAVid test set with other methods. The bold column indicates the best results.

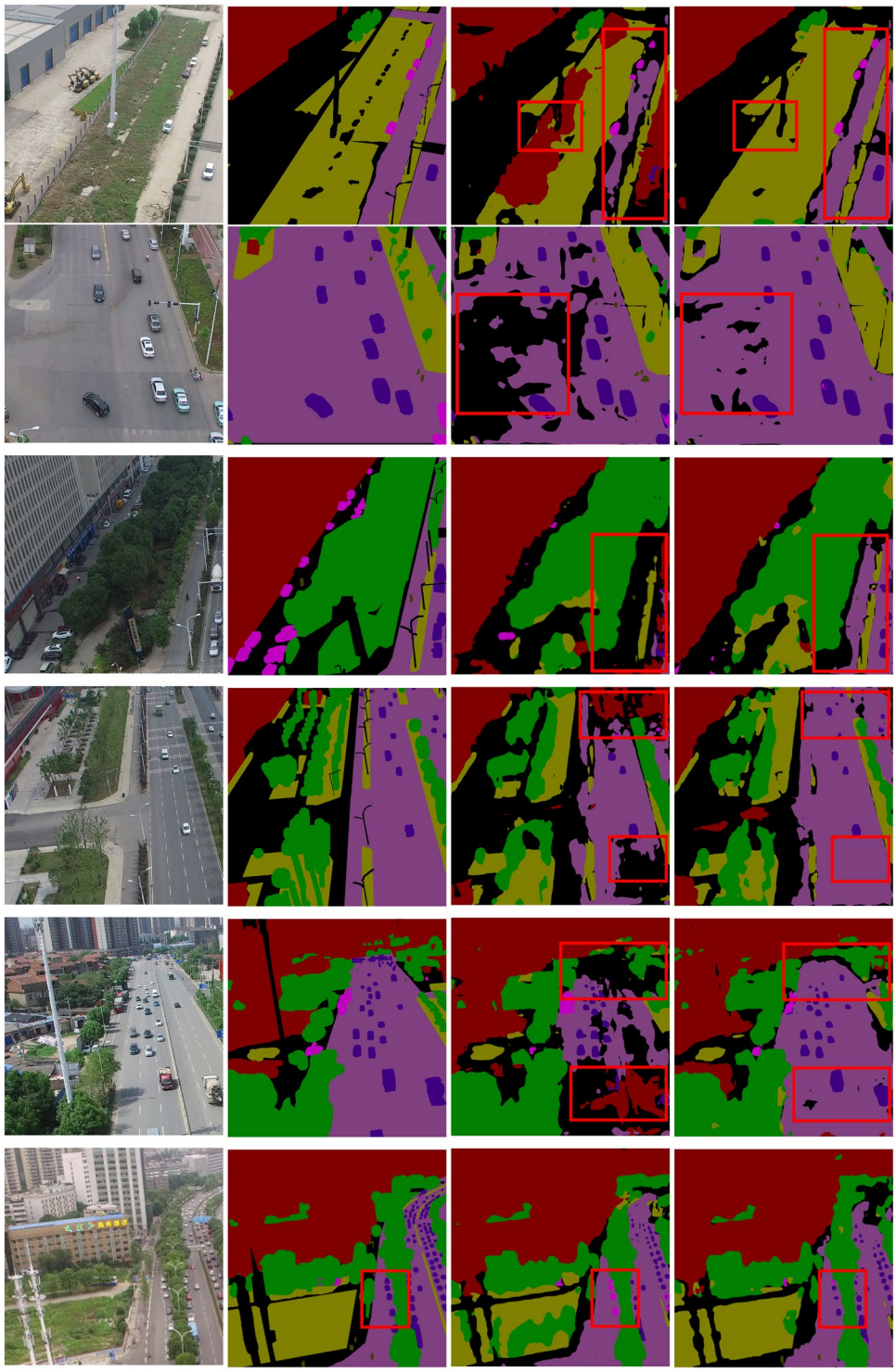


**Fig. 4**. Visualization results on the UAVid test set. Each column represents a type of image, with the rows showing different samples. The columns are input images, ground truth, baseline segmentation results, and the segmentation outcomes of our approach, respectively. Compared to DDRNet, our method offers more accurate segmentation results, especially for complex scenes and small objects, by effectively preventing detail loss and segmentation errors.

| Method | Building | LowVeg | Tree | Car | Water | Ship | Tank | Playground | Sidewalk | mIoU | Params(M) | GFlops |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| UNetformer[56] | 78.70 | 73.04 | 82.16 | 54.48 | 91.99 | 55.11 | 44.12 | 45.03 | 90.73 | 58.14 | 11.7 | 56.9 |
| DecoupleNet D2[57] | 17.87 | 61.45 | 57.02 | 25.72 | 56.48 | 43.73 | 46.52 | 28.28 | 70.77 | 43.57 | 6.8 | 32.1 |
| DDRNet (baseline)[37] | 85.75 | 76.62 | 84.99 | 60.31 | 95.23 | 48.41 | 65.61 | 67.31 | 91.26 | 62.19 | 5.73 | 4.91 |
| Ours | **86.35** | **77.88** | **86.22** | **62.3** | **95.85** | **53.69** | **71.61** | **69.76** | **91.97** | **63.23** | 5.79 | 4.93 |

**Table 2**. Comparison of results on the Mixed test set with other methods. The bold column indicates the best results. Given the extensive class information within the mixed dataset, we present a comparative analysis of accuracy for a selected subset of classes, as detailed in this table.

Materials. The experimental results demonstrate that the differences in the final accuracy improvement across models were marginal. Consequently, we adopted ChatGLM as the primary model for knowledge graph construction.

## Conclusion

In this paper, we propose a novel method for constructing a universal knowledge graph using large language models. We also argue that the relationships between categories can aid in scene-level segmentation, and thus introduce a knowledge graph fusion module to integrate the proposed prompt-based graph into the semantic segmentation network. This module combines relevant class information from the knowledge graph with the semantic segmentation task, allowing us to leverage contextual information for more accurate pixel-level segmentation. For example, knowing the 'adjacent' relationship between 'road' and 'tree' in our dataset enables our model to focus more on tree features when recognizing roads. Experiments conducted on the UAVid and Mixed datasets validate our hypothesis and the generalizability of the knowledge graph. In future work, we will explore alternative methods for capturing inter-class relationships within the dataset, beyond knowledge graphs, to further enhance their generalizability. However, due to the lack of fine-tuning, the current LLM may still exhibit limitations in capturing domain-specific nuances and may experience latency when reasoning about inter-entity relationships – a known drawback arising from the static nature of pre-trained models. While the general capabilities of pretrained LLMs are sufficient for initial relation inference, fine-tuning remains a promising direction for enhancing precision and adaptability, especially in evolving remote sensing scenarios. Additionally, although data augmentation has helped alleviate class imbalance, the diversity of samples for extremely rare categories remains limited. In future work, we plan to fine-tune the LLM using domain-specific corpora to improve knowledge adaptability, and further incorporate diffusion-based data generation to enrich minority-class representations. Moreover, although data augmentation has mitigated the class imbalance problem, the sample diversity for extremely rare categories may remain inadequate. Furthermore, we propose to employ diffusion models to generate more diverse minority-class samples.

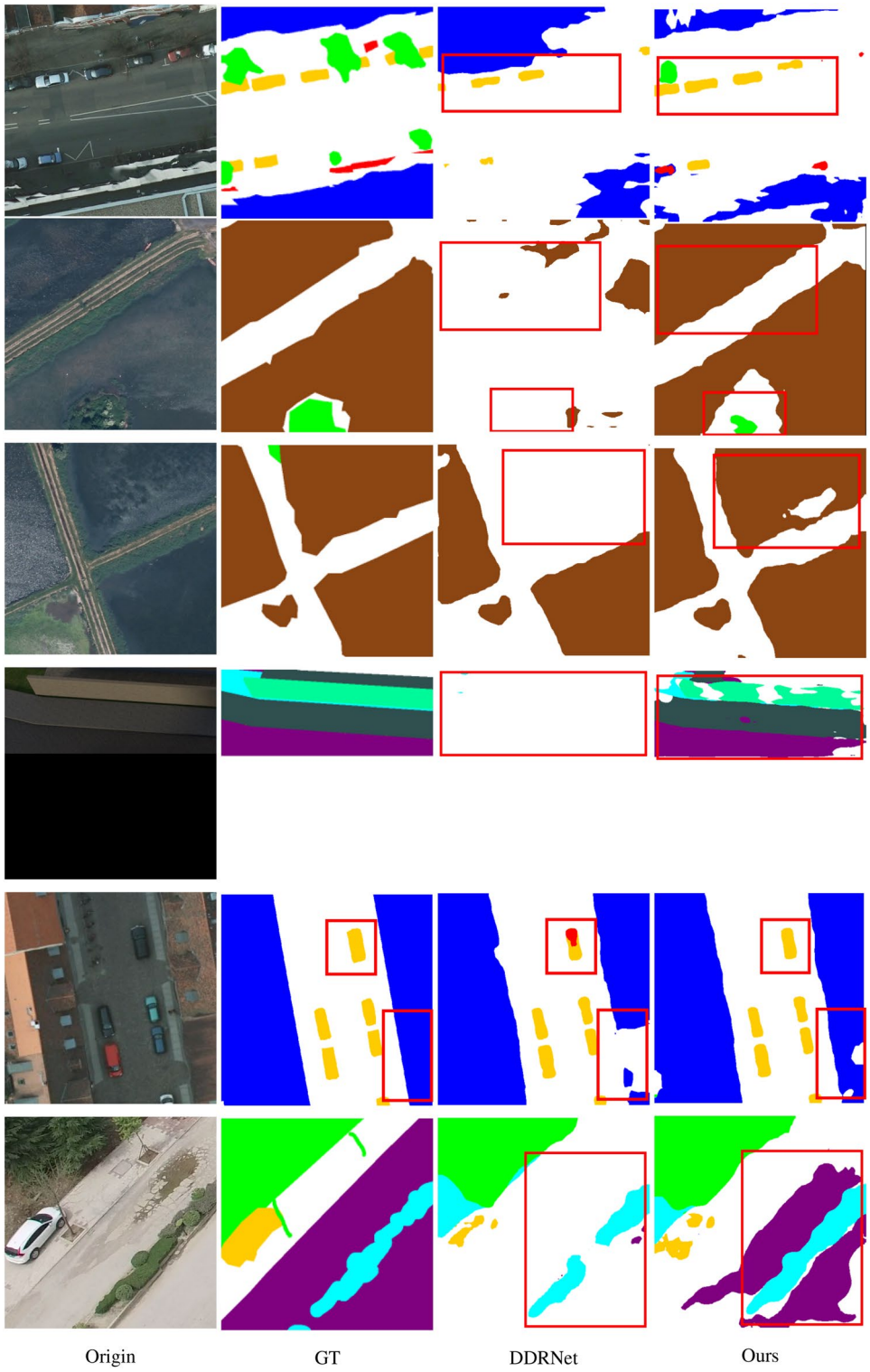


**Fig. 5**. Visualization of results on the Mixed test set. Each row shows a sample, with the columns representing the input image, ground truth, baseline segmentation result, and the segmentation outcomes of our approach, respectively. Compared to DDRNet, our method exhibits superior segmentation accuracy, particularly when segmenting objects that are similar in appearance to their background.

## Data availability

The data that support the findings of this study are available from the corresponding author, Huilin Xu, upon reasonable request.

Received: 5 March 2025; Accepted: 2 December 2025
Published online: 11 December 2025

## References


1. Liu, Z., Li, X., Luo, P., Loy, C. C. & Tang, X. Deep learning markov random field for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1814–1828 (2017).
2. Jing, L., Chen, Y. & Tian, Y. Coarse-to-fine semantic segmentation from image-level labels. *IEEE Trans. Image Process.* **29**, 225–236 (2020).
3. Ou, J., Lin, H., Qiang, Z. & Chen, Z. Survey of images semantic segmentation based on deep learning. In *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 456–463 (2022).
4. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
5. Fang, Z., Wang, Y., Xie, P., Wang, Z. & Zhang, Y. Hisynseg: Weakly-supervised histopathological image segmentation via image-mixing synthesis and consistency regularization. *IEEE Trans. Med. Imaging* **44**, 1765-1782 (2025).
6. Chen, Y. & Bruzzone, L. Toward open-world semantic segmentation of remote sensing images. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 5045–5048 (2023).
7. Yao, S. et al. Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmarks for autonomous driving on water surfaces. *IEEE Trans. Intell. Transp. Syst.* **25**, 16584-16598 (2024).
8. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
9. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. *Semantic image segmentation with deep convolutional nets and fully connected crfs* arXiv:1412.7062 (2016).
10. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
11. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. *Rethinking atrous convolution for semantic image segmentation* arXiv:1706.05587 (2017).
12. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).
13. Yang, M., Yu, K., Zhang, C., Li, Z. & Yang, K. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3684–3692 (2018).
14. Chen, S., Yang, X. & Li, Z. Improving semantic segmentation with knowledge reasoning network. *J. Vis. Commun. Image Represent.* **96**, 103923 (2023).
15. Li, Y. S. et al. Geographic knowledge graph-guided remote sensing image semantic segmentation. *Natl. Remote. Sens.* **28**, 455–469 (2024).
16. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
17. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742 (PMLR, 2023).
18. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892-34916 (2024).
19. GLM, T. et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint* arXiv:2406.12793 (2024).
20. Zhu, D., Chen, J., Shen, X., Li, X. & Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint* arXiv:2304.10592 (2023).
21. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
22. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
23. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
24. Wang, J. et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3349–3364 (2021).
25. Zhao, H., Qi, X., Shen, X., Shi, J. & Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, 405–420 (2018).
26. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
27. Guo, M.-H. et al. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **35**, 1140–1156 (2022).
28. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
29. Strudel, R., Garcia, R., Laptev, I. & Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272 (2021).
30. Wang, W. et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14408–14419 (2023).
31. Yu, W. et al. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10819–10829 (2022).
32. Zhang, X., Yuan, G., Hua, Z. & Li, J. Tsmga: Temporal-spatial multiscale graph attention network for remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **18**, 3696–3712 (2025).
33. Zhang, X., Dong, K., Cheng, D., Hua, Z. & Li, J. Stwanet: Spatio-temporal wavelet attention aggregation network for remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **18**, 8813–8830 (2025).
34. Zhang, X., Wang, Z., Li, J. & Hua, Z. Mvafg: Multiview fusion and advanced feature guidance change detection network for remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **17**, 11050–11068 (2024).
35. Zhou, Y., Xia, H., Yu, D., Cheng, J. & Li, J. Outlier detection method based on high-density iteration. *Inf. Sci.* **662**, 120286 (2024).
36. Lu, W. et al. Visual style prompt learning using diffusion models for blind face restoration. *Pattern Recognit.* **161**, 111312 (2025).
37. Hong, Y., Pan, H., Sun, W. & Jia, Y. *Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes* arXiv: 2101.06085 (2021).
38. Hogan, A. et al. Knowledge graphs. *ACM Computing Surveys (Csur)* **54**, 1–37 (2021).
39. Noy, N. et al. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue* **17**, 48–75 (2019).
40. Qiu, X., Zhang, Q. & Huang, X.-J. Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, 49–54 (2013).

41. Zhou, L. & Zhang, D. Nlpir: A theoretical framework for applying natural language processing to information retrieval. *J. Am. Soc. for Inf. Sci. Technol.* **54**, 115–123 (2003).
42. Kipf, T. N. & Welling, M. *Semi-supervised classification with graph convolutional networks* arXiv: 1609.02907 (2017).
43. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **30** (2017).
44. Veličković, P. et al. Graph attention networks. In *International Conference on Learning Representations* (2018).
45. Gasteiger, J., Bojchevski, A. & Günnemann, S. Combining neural networks with personalized pagerank for classification on graphs. In *International Conference on Learning Representations* (2019).
46. Xu, B., Xie, B. & Shen, H. Towards deeper graph neural networks via layer-adaptive. *In Companion Proceedings of the ACM Web Conference* **2023**, 103–106 (2023).
47. Chen, Z. et al. Agnn: Alternating graph-regularized neural networks to alleviate over-smoothing. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 13764-13776 (2023).
48. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543 (2014).
49. Liang, X., Hu, Z., Zhang, H., Lin, L. & Xing, E. P. Symbolic graph reasoning meets convolutions. *Adv. Neural Inf. Process. Syst.* **31** (2018).
50. Zhang, J., Peng, B., Wu, X. & Hu, J. Weakly supervised semantic segmentation by knowledge graph inference. *Eng. Appl. Artif. Intell.* **138**, 109294. https://doi.org/10.1016/j.engappai.2024.109294 (2024).
51. Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A. & Yang, M. Y. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS J. Photogramm. Remote Sens.* **165**, 108–119 (2020).
52. Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T. & Zambrzycka, A. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1102–1110 (2021).
53. Yang, Y. & Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **51**, 818–832. https://doi.org/10.1109/TGRS.2012.2205158 (2013).
54. Scanlon, M. *Semantic Annotation of Aerial Images using Deep Learning, Transfer Learning, and Synthetic Training Data*. Ph.D. thesis (2018).
55. Zou, Z., Shi, T., Li, W., Zhang, Z. & Shi, Z. Do game data generalize well for remote sensing image segmentation?. *Remote Sens.* **12**, 275 (2020).
56. Wang, L. et al. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **190**, 196–214 (2022).
57. Lu, W., Chen, S.-B., Shu, Q.-L., Tang, J. & Luo, B. Decouplenet: A lightweight backbone network with efficient feature decoupling for remote sensing visual tasks. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–13. https://doi.org/10.1109/TGRS.2024.3465496 (2024).
58. Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).


## Acknowledgements

This work was supported in part by the Natural Science Foundation of Fujian Province, China, under Grant 2023J01803; in part by the Natural Science Foundation of Xiamen, China, under Grant 3502Z202371019, 3502Z202474005.

## Author contributions

Conceptualization, J.S., X.Z. and H.X.; methodology, J.S., X.Z.; validation, J.S., H.X., X.Z. and C.S.; writing–original draft preparation, J.S., X.Z.; writing–review and editing, X.Z., J.S., H.X., C.S., Y.L., J.L. and Y.C.. All authors have read and agreed to the published version of the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-31346-x.

**Correspondence** and requests for materials should be addressed to H.X.