Multi-modal semantic segmentation for open vocabulary interactions with remote sensing images

Jinkun Dai[1], Tao Peng[1], Yuhang Xue[1], Xianping Ma[1], Yuanxin Ye[1,*]

[1] Southwest Jiaotong University, Chengdu 611756, China

Keywords: Remote Sensing, Multi-modal, Open Vocabulary, Semantic Segmentation, Vision-language Model.

Abstract

Semantic segmentation of multi-modal remote sensing imagery plays a pivotal role in land use/land cover (LULC) mapping, environmental monitoring, and precision earth observation. Current multi-modal approaches mainly focus on integrating complementary visual modalities (e.g., optical and synthetic aperture radar (SAR) imagery), yet neglect the incorporation of non-visual textual data. To address this limitation, we propose TSMNet, a text supervised multi-modal open vocabulary semantic segmentation network that synergistically integrates textual supervision with visual representation for open-vocabulary semantic segmentation. Unlike conventional multi-modal segmentation frameworks, TSMNet introduces a dual-branch text encoder to extract both scene-level semantic and object-level label information from various textual data, enabling dynamic cross-modal fusion. These text-derived features dynamically interact with visual embeddings through the proposed text-guided visual semantic fusion module, enabling domain-aware feature refinement and human-interpretable decision-making. Moreover, integrating text opens pathways for open-vocabulary semantic segmentation, enabling systems to dynamically segment targets through natural language descriptions, thereby overcoming the rigid constraints of traditional pre-defined classification heads. To verify our method, we innovatively construct two new multi-modal datasets, and carry out extensive experiments to make a comprehensive comparison between the proposed method and other state-of-the-art (SOTA) semantic segmentation models. Results demonstrate that TSMNet achieves superior segmentation accuracy while exhibiting robust generalization capabilities across diverse geographical and sensor-specific scenarios. The source code will be available at https://github.com/yeyuanxin110/TSMNet.

1. Introduction

Semantic segmentation is an advanced remote sensing technique that aims to perform pixel-wise classification of each pixel in the image, achieving pixel-level image segmentation (Xiao et al., 2025). It has become indispensable for critical Earth observation tasks such as environmental monitoring (Li et al., 2020), urban planning (Liu et al., 2018, Wu et al., 2022), disaster response and land use classification (Sundaresan and Solomon, 2025, Piramanayagam et al., 2018).

The proliferation of multi-source remote sensing data (e.g., optical, Synthetic Aperture Radar (SAR), Light Detection and Ranging (LiDAR)) driven by advances in sensor technology has revolutionized Earth observation paradigms (Zhu et al., 2017). However, the deluge of data not only drives methodological innovation but also imposes unprecedented challenges on interpretation methods (Xiong et al., 2025). Over the past decade, deep learning has become a powerful approach to addressing complex tasks (LeCun et al., 2015, Luo et al., 2023); however, most studies still concentrate on single-modal semantic segmentation (Long et al., 2015, Ronneberger et al., 2015, Chen et al., 2018).

Single-modal approaches suffer from limited discriminative power when differentiating spectrally similar classes (e.g., asphalt vs. shadow). To address this, some studies introduce additional modes to overcome the performance bottleneck (Ma et al., 2024, Zhou et al., 2025). For example, the combination of optical image and SAR image can effectively reduce the influence of bad weather and speckle noise on semantic segmentation model (Ma et al., 2025, Wei et al., 2024, Ye et al., 2025). The existing multi-modal semantic segmentation achieves by using complementary visual modes but ignores the potential of non-visual modes such as text data. Textual data offers a transformative opportunity: it embeds real-world knowledge (e.g., geographic context, material properties) that can provide critical contextual priors for model robustness (Li et al., 2023, Li et al., 2025). In addition, the integrated text opens the way for semantic segmentation of open vocabulary (Wang et al., 2024a, He et al., 2023), which enables the system to identify and classify invisible categories through natural language description, thus overcoming the strict restrictions of predefined category classification (Kawano and Aoki, 2024). Different from the traditional semantic segmentation method (Pan et al., 2025), which is limited to a fixed set of tags, open vocabulary semantic segmentation allows a wider range of concepts, making it more flexible and suitable for new scenarios in practical applications (Cheng et al., 2022).

Recently, Visual Language Models (VLMs) has attracted great attention because of its excellent open-vocabulary object recognition ability (Zhang et al., 2023, Zhang et al., 2025). This great success motivates us to explore its adaptability to multi-modal semantic segmentation tasks. VLMs have demonstrated excellent feature representation ability through cross-modal contrastive learning (Zermatten et al., 2023). However, this representation lacks pixel-level granularity, making it challenging to directly apply to dense prediction tasks. Inspired by the work of natural language processing, a series of methods are proposed in the field of vision, aiming at effectively adapting VLMs to the task of open-vocabulary semantic segmentation (Wang et al., 2024c, Lin et al., 2024). The existing main method is prompt learning, and the adapter is used to enhance the feature representation of learning. In the context of land cover mapping, eliminating the need for fixed label sets opens the door for novel map generation, in which the category information used can be directly defined by users (Cao et al., 2025). By utilizing lan-

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-3-2026-51-2026

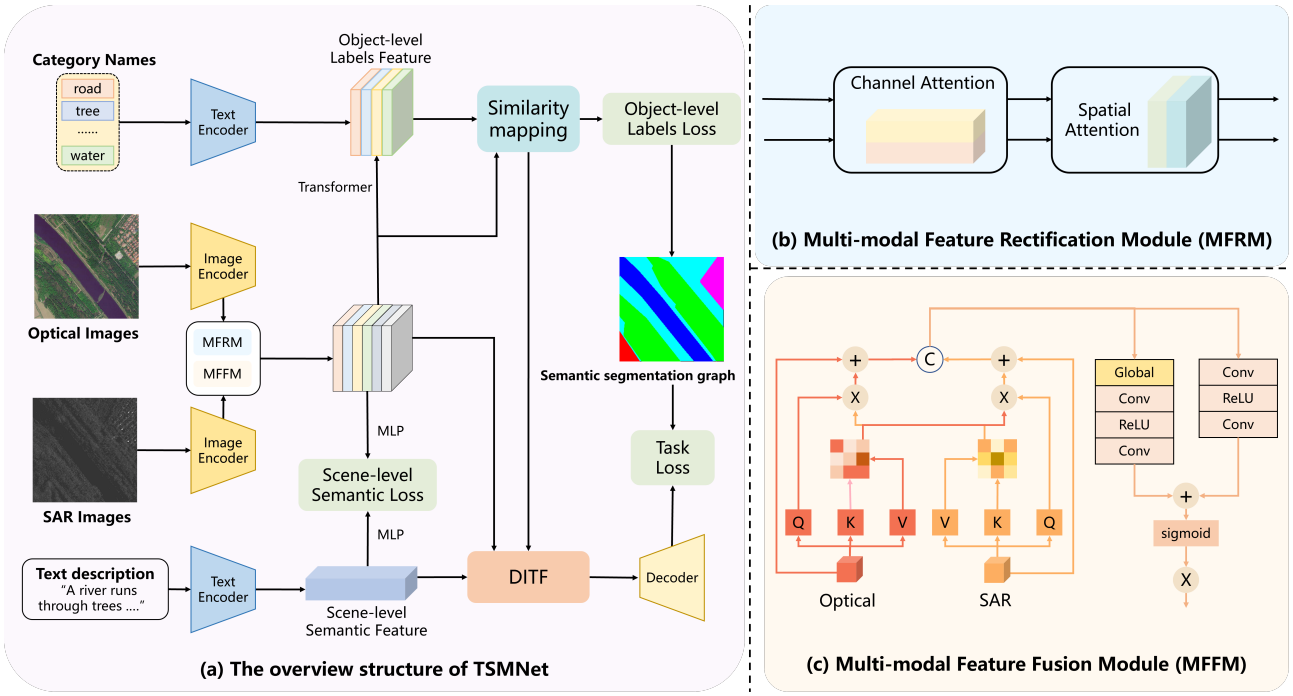

Figure 1. The overall structure of TSMNet is shown in (a). TSMNet realizes the multi-modal open vocabulary semantic segmentation based on optical images, SAR images and text data. The overall structure of MFRM is shown in (b). It dynamically adjusts the channel and spatial weights between multi-modal features to enhance the complementary information. The overall structure of MFFM is shown in (c). It introduces cross-modal attention mechanism for deep semantic interaction to realize multi-modal feature aggregation.

guage to specify the necessary classes, individuals, particularly those without expertise in remote sensing, can create customized maps tailored to their unique requirements (Zhang et al., 2024, Xu et al., 2024).

Inspired by the potential of remote sensing VLMs (Wang et al., 2024b), we propose a text-supervised multi-modal open vocabulary semantic segmentation network. Our framework aligns image-language features into a unified semantic space via contrastive learning, while enabling natural language interaction through hierarchical visual representation learning that leverages multi-layer high-level semantics. Specifically, a text supervised multi-modal open vocabulary semantic segmentation network (TSMNet) consists of the following components: multi-modal visual encoder for optical and SAR images, text encoder based on contrastive language–image pretraining (CLIP) (Radford et al., 2021), scene-level semantic and object-level label fusion module of visual language. The multi-modal visual encoder of optical and SAR images adopts a pseudo-Siamese feature extraction module, and extracts multi-level semantic information of optical and SAR images respectively through vision transformer (ViT) (Dosovitskiy et al., 2020), and constructs a multi-modal remote sensing image feature fusion network to fully integrate the detailed features of multi-modal images. For language processing, TSMNet uses the bidirectional encoder representations from transformers (BERT), which is a universal language Transformer, to design a multi-layer image language fusion module, including scene-level semantic and object-level label features. The process begins by generating two distinct types of texts using a predefined prompt template. These text features and image features are aligned by contrast learning method. Subsequently, the aligned text features are integrated with multi-modal image features through a text-guided visual semantic fusion module, ensuring rich contextual representation. TSMNet has gained strong representational capabilities of remote sensing images, enable deep exploration of the invariant features of images, and shows good generalization ability in semantic segmentation.

The main work of this paper is as follows:

1. In this paper, a TSMNet is developed, which innovatively integrates the fine-grained features of image and text modes, and realizes accurate semantic segmentation in open vocabulary scenes through multi-modal feature interaction mechanism.
2. We design a dual-branch image and text fusion module (DITF), which integrates the image features with the scene-level semantic and object-level label features of the text by optimizing the text embedding, effectively integrates the heterogeneous graphic features, enhances the dependence within and between patterns, and thus enriches the semantic information.
3. To evaluate the generalization ability and practical application value of the model, we constructed a semantic segmentation dataset of optical and SAR remote sensing images from Gaofen (GF) satellites, and describe the images manually. The dataset covers representative areas of diverse terrain, including bare ground, low vegetation, trees, houses, water, roads and other landmark categories. Each region is equipped with a high-precision ground truth label, which fills the key gap in the current multi-modal semantic segmentation dataset of integrated visual language. At the same time, we have meticulously annotated each group of images in an existing multi-modal remote sensing image dataset with detailed textual descriptions, establishing a reliable benchmark for evaluating the accuracy of our proposed model.

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-3-2026-51-2026 | © Author(s) 2026. CC BY 4.0 License.

2. Methodology

CLIP has shown great ability in open vocabulary classification, however, the gap between image-level pre-training knowledge and pixel-level segmentation task is difficult to bridge. Fine-tuning CLIP directly on the downstream segmentation dataset will inevitably damage the ability of open vocabulary recognition. Therefore, our goal is to explore a method to achieve pixel-level alignment of text features and image features, while preserving the alignment of scene features and text descriptions of images to enhance the model's understanding of scene-level semantic and object-level label text information of remote sensing images. As shown in figure 1, we propose a text-supervised multi-modal open vocabulary semantic segmentation network. In this section, we first introduce an overview of the proposed TSMNet in Section 2.1. Then, in 2.2, we discuss how to better integrate the detailed features of multi-modal images. We discuss the object-level label feature fusion of image and text in 2.3 and the scene-level semantic feature fusion of image and text in 2.4. Finally, the loss function and inference process are elaborated in Sections 2.5 and 2.6, respectively.

2.1 Overall Network Architecture

As shown in figure 1, given a set of input images $I \in R^{H \times W \times 3}$, a K-vocabulary list and a set of text descriptions, where H and W refer to the height and width of input and K refers to the number of class names, the whole network is divided into three stages, namely, the multi-modal image feature fusion stage, the object-level label feature fusion of image and text and the scene-level semantic feature fusion of image and text.

In order to inherit the abundant pre-training knowledge from CLIP, we fuse the multi-scale pixel-level features from optical and SAR images, then align the fused image features with the corresponding text features of K vocabulary at pixel level and construct the interaction between vision and language by using context-aware prompting strategy. Finally, we fuse the image features with the features described in the text to produce the semantic segmentation results of K vocabulary.

2.2 Multi-modal Image Feature Fusion Network

This paper proposes a multi-modal image feature fusion network, which combines feature rectification and cross-modal interaction to solve the multi-modal feature fusion challenge between optical and SAR images. Multi-modal data (such as optical and SAR) show significant heterogeneity, which makes it difficult for traditional linear weighted fusion methods to make full use of complementary information. Because of the complexity of multi-modal data, we use ViT as the feature extractor of optical and SAR images, and take the output of its four stages as multi-scale image features, which together with global image representation and local spatial features constitute image features. In order to better fuse the optical and SAR image features, we designed a multi-modal feature rectification module (MFRM) and a multi-modal feature fusion module (MFFM) to fuse the features of the image respectively. The formulas are shown as follows:

$$x = ViT(X), y = ViT(Y) \tag{1}$$

$$\tilde{x}, \tilde{y} = MFRM(x, y) \tag{2}$$

$$z = MFFM(\tilde{x}, \tilde{y}) \tag{3}$$

where X, Y = optical and SAR images
x, y = corresponding feature maps
$\tilde{x}, \tilde{y}$ = feature maps after multi-modal feature correction
z = multi-modal fusion feature maps of optical and SAR images

Multi-stage fusion realizes fine information integration through hierarchical processing of multi-scale features. Firstly, in the feature extraction stage, double backbone network is used to extract multi-layer features from optical and SAR images. Then, in the stage of hierarchical correction and fusion, MFRM is used to eliminate modal deviation in turn, and MFFM is used for cross-modal interaction on four scales. Finally, in the output aggregation stage, the multi-level fusion features are integrated into a multi-scale feature pyramid. This strategy dynamically assigns weights and promotes cross-modal interaction, balancing local detail enhancement and global semantic consistency, thus improving the robustness of the model in complex scenes and the generalization ability in multi-task scenes.

2.3 Object-level Label Feature Fusion of Image and Text

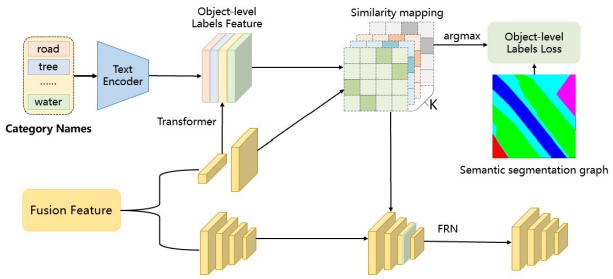

Figure 2. Object-level Label Feature Fusion Module of Image and Text.

As shown in figure 2, this paper proposes a method to align image and text features at pixel level, which effectively uses the prior knowledge of language from CLIP pre-training model. CLIP consists of two encoders, including an image encoder (ResNet and ViT) and a text encoder (Transformer). CLIP is trained through contrastive learning on large-scale image-text pairs, achieving alignment between images and text in the same embedding space. It supports cross-modal understanding and zero-shot learning, providing a powerful foundational model for multi-modal tasks. To transfer CLIP's prior knowledge to downstream semantic segmentation tasks, this paper adopts text prompts constructed based on templates. We construct text prompts by substituting the [CLS] placeholder in the template “a photo of a [CLS]” with K class names, and then encode it by the text encoder of CLIP. Previous research has demonstrated that reducing the domain gap between vision or language can significantly improve the performance of the CLIP model in downstream tasks. Therefore, we attempt to adopt a context-aware approach to enhance text features, rather than relying solely on traditional predefined templates.

Thus, the input to the text encoder becomes:

$$[p, e_k], 1 \leq k \leq K \tag{4}$$

where p = learnable text context
e_k = name embedding of the k-th class

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-3-2026-51-2026

Contextual features are composed of global image representations and local spatial features.

Including the description of visual context, can make the text more accurate. For example, ”a dog on the grass” is more accurate than ”a dog”. Usually, we can use the cross-attention mechanism in the decoder to simulate the interaction between vision and language. The context-aware strategy adopted in this paper is to refine the text features after the text encoder, that is, post-model prompt. We use template hints to generate text features, which are directly used in the query of Transformer decoder.

$$\nu_{post} = TransDecoder\,(t, [\bar{z}, z]) \tag{5}$$

where t = text feature
z = language-compatible multi-modal fusion image feature

This method encourages text features to find the most relevant visual clues, and then updates the text features through the remaining links:

$$t \leftarrow t + \gamma \nu_{post} \tag{6}$$

where γ = a learnable parameter to control the scaling of the residual

It is initialized to a very small value (e.g. 10^{-4}) to retain the linguistic prior knowledge from text features to the maximum extent.

Subsequently, we compute a pixel-text score map by measuring the compatibility between language-aligned feature maps z and text features t:

$$s = \hat{z}\hat{t}^{T} \tag{7}$$

where $\hat{z}, \hat{t}$ = the ℓ_2 normalized version of z and t along the channel dimension

The score map quantitatively represents the matching correspondence between pixel-level visual patterns and textual semantics. First, the score map can be regarded as a segmentation result, and we can use it to assist in calculating the segmentation loss. Second, we connect the score map to the final feature map to incorporate linguistic prior knowledge. Input it into a neck module based on Feature Pyramid Network (FPN) to fuse and enhance the features of different levels, so as to better support the tasks of target detection and segmentation.

2.4 Scene-level Semantic Feature Fusion of Image and Text

Compared with local category information, people often read global information from images first. Because there is a big gap between image and text description, cross-modal image-text feature alignment and fusion are the key factors in visual language representation learning. In order to solve this problem, we developed an global alignment and fusion module for image and text features, which consists of two main components, namely, image-text alignment module and image-text fusion module, as shown in figure 3.

1) Image-text alignment module: We adopt the strategy of ” alignment first and then fusion ” to deal with heterogeneous image-text features. Firstly, we apply the contrastive loss of

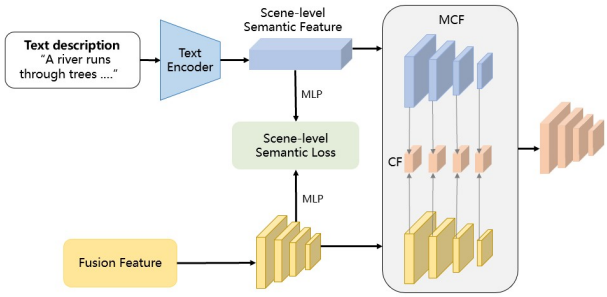

Figure 3. Scene-level Semantic Feature Fusion Module of Image and Text.

image-text to reduce the modal gap between image-text features. This process establishes the foundational relationship between the two features, facilitating enhanced image-text feature fusion in subsequent stages. Map image and text features to a shared embedded space. It is usually realized by using a projection head, which can be an MLP. Details of the image-text contrastive loss are as follows.

2) Image-text fusion module: Text features including climate information and geographical object features can be used as global priors for cross-modal feature fusion. Therefore, this paper proposes an image-text fusion module based on cross-modal attention mechanism, aiming at effectively combining image and text features and improving the feature representation ability in multi-modal tasks. The module firstly encodes the global text features and dynamically generates multi-scale text features to match the image features at different levels. Then, the cross-modal attention mechanism is used to fuse image features and text features layer by layer to capture the semantic association between them. Specifically, for each level of image features, the module interacts with the corresponding scale of text features through the cross-modal attention mechanism to generate a fused feature representation. Finally, the module outputs multi-scale fusion features to provide rich cross-modal information for downstream tasks. Experiments show that the module can significantly improve the performance of joint representation of images and texts, and provide an effective feature fusion solution for multi-modal tasks.

2.5 Loss Function

In the training process of TSMNet, three main losses are designed and calculated: object-level label loss, scene-level semantic loss and semantic segmentation task loss. Object-level tag loss is used to guide the optimization of image and tag text information, scene-level semantic loss is used to guide the optimization of image and text description information, and semantic segmentation task loss ensures that the segmentation mask generated by the model is consistent with the target area. Specifically, we use cross-entropy loss for the object-level label loss, which is represented as follows:

$$\mathcal{L}_{Object} = CrossEntropyLoss = -\sum_{i=1}^{C} y_i \log(p_i) \tag{8}$$

With:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \tag{9}$$

Where C = the total number of classes

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-3-2026-51-2026 | © Author(s) 2026. CC BY 4.0 License.

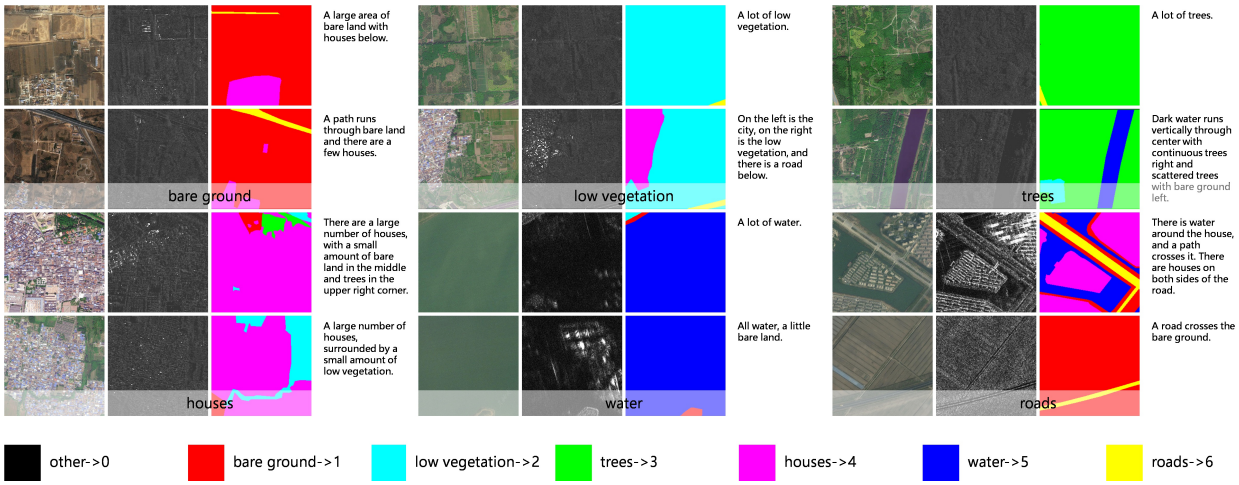

Figure 4. Some annotation examples and category percentages in SWJTU-Vision-Language dataset.

p_i = the probability of class i predicted by the model
z_i = the original output of the model to class i

Scene-level semantic loss typically employs the InfoNCE (Noise Contrastive Estimation) loss, which is widely utilized in contrastive learning frameworks. The mathematical formulation of this loss function can be expressed as follows:

$$\mathcal{L}_{Scene} = -\log \frac{\exp(sim(z_i, z_t)/\tau)}{\sum_{j=1}^{N} \exp(sim(z_i, z_{t_j})/\tau)} \tag{10}$$

where z_i, z_t = the embedding of image and text respectively
$sim(,)$ = the similarity function (such as cosine similarity)
τ = the temperature parameter
N = the number of negative samples

At the same time, we use cross-entropy loss ($\mathcal{L}_{ce}$) and dice loss ($\mathcal{L}_{dice}$) for the semantic segmentation task loss. The total loss of TSMNet is the sum of four types of losses, which can be expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{Object} + \mathcal{L}_{Scene} + \mathcal{L}_{ce} + \mathcal{L}_{dice} \tag{11}$$

2.6 Inference

As shown in figure 1, the network can generate the similarity mapping between category text and image while generating the segmentation map. In the reasoning stage, firstly, the input text categories are encoded by a text encoder, and the target embedding is obtained. Then, the cosine similarity between the semantic features of each pixel position of the target embedding and fusion features is calculated, so that the similarity mapping of each target text is generated. Finally, the final segmentation mask is obtained by selecting the label corresponding to the maximum of all similarity values at each pixel position. Compared with the traditional semantic segmentation methods that rely on a hard-coded N-way classification head (e.g., a softmax layer), this network presents an architectural advantage for open-vocabulary settings. Because the segmentation is driven by text-to-pixel cosine similarity rather than fixed output channels, the model inherently possesses the capacity to interact with arbitrary text prompts. This means that users can use various English words, synonyms, or text fragments as target queries. Although our quantitative evaluations in Section 3 are conducted on standard, fixed-class land-cover datasets to ensure fair comparisons with SOTA models, this similarity-based inference mechanism establishes a fundamental framework for dynamic, human-interpretable interactions and future zero-shot applications. It not only handles known categories effectively but also transitions the paradigm towards more flexible semantic parsing in practical scenarios.

3. Experiments and Results

3.1 Data Description

In the experiment, two multi-modal remote sensing image datasets are used to evaluate the proposed TSMNet, one is the SWJTU-Vision-Language dataset proposed in this paper, and the other is the YESeg-OPT-SAR dataset.

Dataset	All	Train	Test
SWJTU-Vision-Language	2712	800	1912
YESeg-OPT-SAR	2231	800	1431

Table 1. Sample size of two multi-modal datasets (256 × 256 pixels).

1) SWJTU-Vision-Language dataset. As shown in figure 4, this dataset is a multi-modal and high-resolution remote sensing image dataset, which is specially designed for semantic segmentation tasks driven by deep learning. By integrating optical and SAR imagery from the same area, this dataset constructs a joint semantic segmentation dataset. The dataset consists of 2,712 pairs of multi-modal images with a resolution of 3 meters, covering four major cities including Beijing, Jinan, Shanghai and Tianjin and their surrounding areas. Each pair of images is manually labeled with detailed text descriptions, which not only provide semantic information of image content, but also help the model to better understand the relationship between different features. In order to further improve the practicability of the dataset, all images are labeled at pixel level, and each pixel is accurately divided into seven different categories. This fine labeling method makes the dataset very suitable for training and evaluating deep learning models, especially in complex land use segmentation tasks. Optical images provide rich color and

Table 2. OA and mIoU (%) on the SWJTU-Vision-Language dataset.

Method	OA	mIoU	User's Accuracy						
			bare ground	low vegetation	trees	houses	water	roads	others
Deeplab v3+	65.34	37.12	63.76	52.68	51.28	72.75	80.84	58.07	38.04
CMGFNet	64.93	34.7	63.98	59.23	50.78	76.24	74.32	66.76	32.28
DDHRNet	62.63	32.66	62.83	54.65	50.93	79.59	81.42	56.33	33.71
MCANet	61.42	33.5	61.03	60.02	38.31	73.06	65.13	33.82	39.48
MSSNet	65.68	37.49	74.12	58.66	32.95	78.17	62.9	25.87	12.52
DenseCLIP	68.09	47.44	**77.17**	57.47	56.8	79.53	83.11	69.91	29.65
TACOSS	65.43	42.64	74.97	58.82	39.28	73.32	65.56	44.86	20.56
TSMNet	**69.73**	**48.83**	76.48	**61.18**	**55.2**	**80.09**	**84.82**	**69.93**	**39.49**

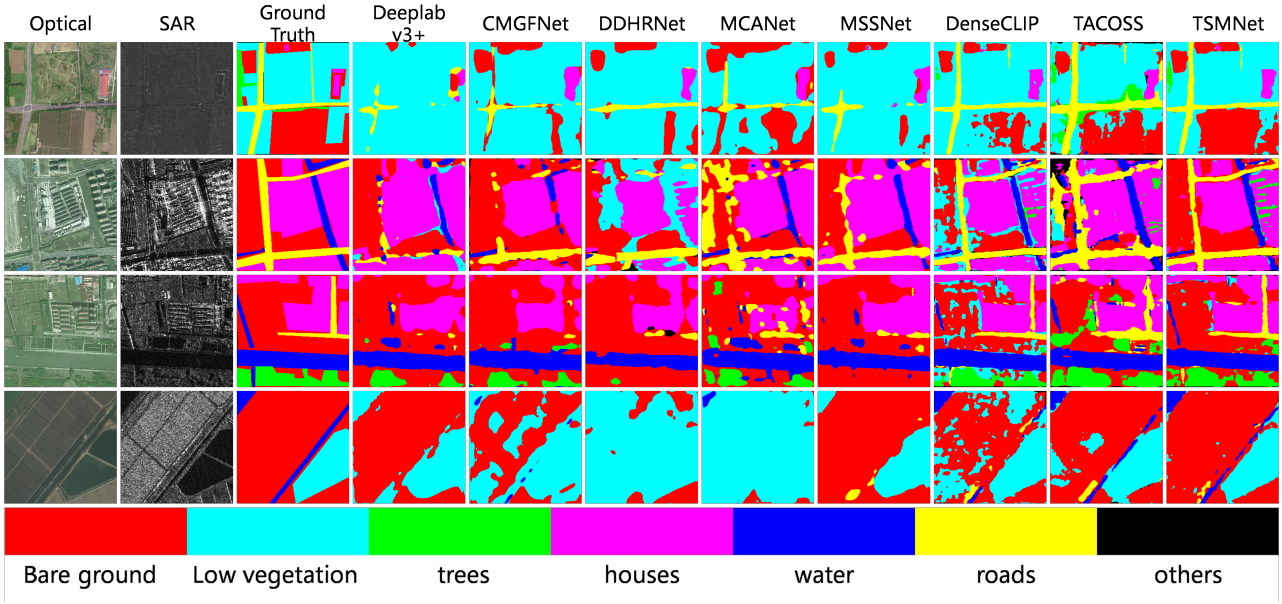

Figure 5. Examples of visualization results on the SWJTU-Vision-Language dataset on TSMNet and other multi-modal semantic segmentation networks.

texture information, while SAR images can penetrate clouds and vegetation to provide structural information of the ground. The combination of the two makes the data set maintain high segmentation accuracy even under cloudy or night conditions. The selection of Beijing, Jinan, Shanghai and Tianjin makes the dataset have extensive geographical coverage and diversified land use types, thus improving the generalization ability of the model. In addition, the construction process of dataset also takes into account the influence of time change and seasonality, and images of some areas have been collected many times in different seasons to ensure that the dataset can reflect the land use changes at different time points. This enhancement of time dimension makes the dataset not only suitable for static land use classification, but also able to support dynamic land use change monitoring and analysis. Generally speaking, the visual language dataset provides a powerful and comprehensive foundation for the semantic segmentation task driven by deep learning by combining optical and SAR images, fine pixel-level labeling, diverse geographical coverage and enhancement of time dimension. Whether used in academic research or practical application, this data set can significantly improve the performance and robustness of the model. The SWJTU-Vision-Language dataset, alongside its corresponding annotations, will be accessible at https://github.com/yeyuanxin110/SWJTU-Vision-Language.

2) YESeg-OPT-SAR dataset. The YESeg-OPT-SAR dataset has a spatial resolution of 0.5 m and includes two types of remote sensing images: high-resolution RGB optical images and SAR images. It consists of 2231 pairs of co-registered images (covering the same areas), each with a size of 256 × 256 pixels, spanning two distinct study regions. The dataset contains eight categories, namely background, bare ground, low vegetation, trees, houses, water, roads, and other. This comprehensive labeling ensures precise and detailed analysis for various applications. We have meticulously annotated each set of images in this dataset with detailed textual descriptions, establishing a reliable benchmark for evaluating the accuracy of our proposed model. Access the dataset on GitHub: https://github.com/yeyuanxin110/YESeg-OPT-SAR.

3.2 Implementation Details

Implementation Details: In order to evaluate the performance of our proposed TSMNet, we compare TSMNet with other typical semantic segmentation models (namely, Deeplab V3+, cross-modal gated fusion network (CMGFNet), multi-modal-cross attention network (MCANet), DDHRNet and MSSNet) and the latest DenseCLIP (Rao et al., 2022) and TACOSS (Zermatten et al., 2025) model. For these two datasets, we randomly select 800 to train the data, and the remaining data are used as test samples in our experiment. Table 1 shows the details about the training and test samples. In order to control the experimental variables and ensure the reliability of the experiment. In order to prevent a small number of ”background” categories from

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-3-2026-51-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 3. OA and mIoU (%) on the YESeg-OPT-SAR dataset.

Method	OA	mIoU	User's Accuracy						
			bare ground	low vegetation	trees	houses	water	roads	others
Deeplab v3+	78.42	51.89	82.07	79.32	58.6	69.92	92.37	64.92	42.52
CMGFNet	78.95	53.69	83.45	73.16	66.37	76.28	76.53	67.24	47.9
DDHRNet	79.05	53.22	81.92	68.94	35.33	67.76	87.73	38.33	46.12
MCANet	76.91	50.9	88.67	65.8	29.31	81.58	80.61	42.72	54.35
MSSNet	80.23	55.94	89.18	73.62	46.87	79.97	82.87	44.93	59.87
DenseCLIP	85.13	65.47	88.47	82.53	69.06	**83.43**	96.04	78.58	65.96
TACOSS	77.66	56.08	73.33	57.67	39.84	65.59	77.53	44.85	40.05
TSMNet	**85.33**	**66.69**	**89.85**	**84.06**	**69.56**	81.97	**96.69**	**80.89**	**69.38**

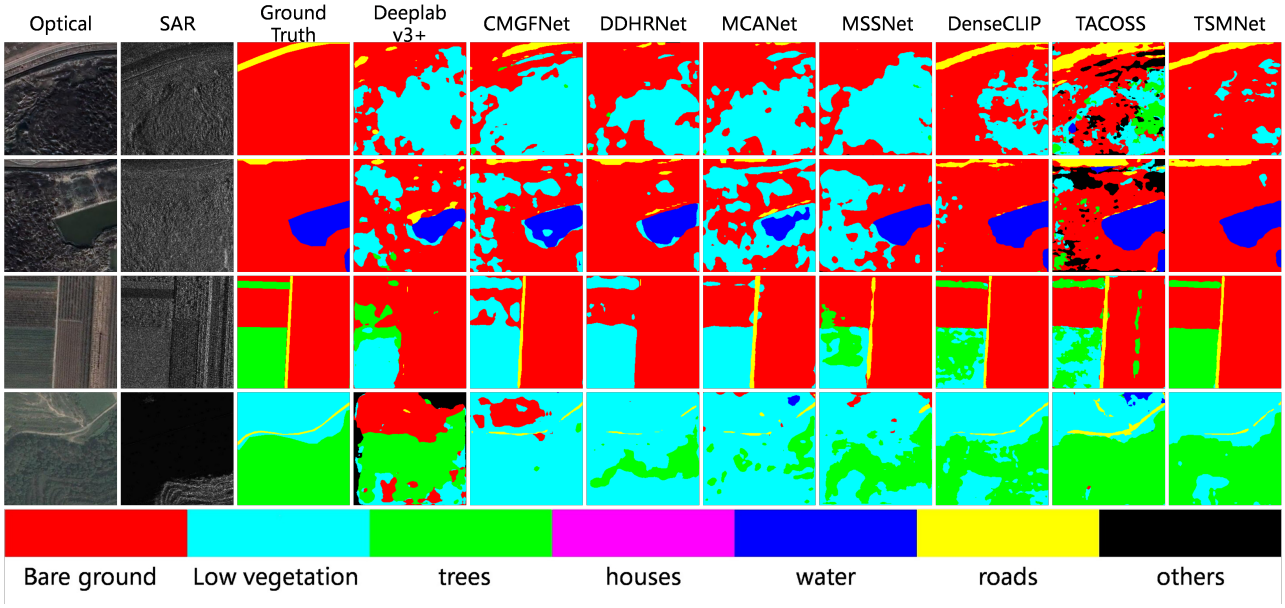

Figure 6. Examples of visualization results on the YESeg-OPT-SAR dataset on TSMNet and other multi-modal semantic segmentation networks.

affecting the experimental results, we classify them as ”other” categories for unified training.

3.3 Evaluation Metrics

Evaluation Metrics: To evaluate the proposed framework, we apply the overall accuracy (OA), coefficients, mean intersection over union (mIoU), score, precsion, and recall to determine the accuracy across the test datasets for downstream tasks.

3.4 Accuracy Analysis

While TSMNet is designed with an open-vocabulary architecture, rigorous zero-shot evaluation on dense remote sensing datasets remains challenging due to the mutually exclusive and densely annotated nature of standard LULC benchmarks. Therefore, to comprehensively evaluate the feature representation capability of our text-supervised framework, we quantitatively benchmark TSMNet against existing SOTA models using the complete predefined classes of the SWJTU-Vision-Language and YESeg-OPT-SAR datasets. We have tested eight representative semantic segmentation models and divided them into three groups. The first group includes the semantic segmentation model of monomodal remote sensing images, such as Deeplab V3+. The second group includes multi-modal remote sensing image semantic segmentation models, such as CMGFNnet, DDHRNet, MCAnet and MSSNet. The third group is multi-modal models of text supervision, such as DenseCLIP and TACOSS.

3.4.1 Performance verification on SWJTU-Vision-Language dataset: In order to evaluate the performance of TSMNet in multi-modal semantic segmentation of open vocabulary, a self-built SWJTU-Vision-Language dataset is used for experimental verification. As shown in Table 2, TSMNet achieved the highest mIoU of 48.83% under the PyTorch framework, which was 1.39 percentage points higher than the second-best model DenseCLIP. At the same time, the OA(Overall Accuracy) of TSMNet reaches 69.73%, which is 1.64% higher than that of the suboptimal model. TSMNet maintains the highest user accuracy in all categories except bare land, especially in the categories of trees (55.2%) and houses (80.09%). TSMNet also achieved the best performance in the challenging road category. As shown by the qualitative analysis results in figure 5, the specially designed multi-modal semantic segmentation network is significantly better than the single-modal semantic segmentation network, which fully proves the effectiveness of optical and SAR image feature fusion in improving semantic segmentation performance. However, the study also found that only relying on the fusion of multi-modal image features still has the limitation of insufficient understanding of the real meaning of categories, which restricts the performance of the model on multi-modal semantic segmentation datasets to some extent.

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-3-2026-51-2026 | © Author(s) 2026. CC BY 4.0 License.

It is worth noting that the multi-modal semantic segmentation model based on text and image is significantly better than the semantic segmentation model using multi-modal images only. Especially in mIoU index, TSMNet model surpasses DenseCLIP and TACOSS, which highlights the advantages of TSMNet in effectively using the image text description information to significantly improve the accuracy of land use/land cover (LULC) classification.

3.4.2 Performance analysis on YESeg-OPT-SAR dataset: In the experiment of YESeg-OPT-SAR dataset, the models specially designed for remote sensing images, such as Deeplab v3+, CMGFNet, DDHRNet, MCANet, MSSNet, show obvious lack of adaptability when dealing with complex datasets. In contrast, TSMNet achieved the highest mIoU of 66.69% with its multi-modal processing ability, which was 1.22 percentage points higher than the sub-optimal model. In addition, TSMNet also achieved the highest OA of 85.33%. As shown in table 3, TSMNet achieved the best results in all categories except houses. Compared with the multi-modal image semantic segmentation model, the multi-modal semantic segmentation model guided by natural language shows superior performance, which fully proves the important role of natural language in helping the model understand and distinguish the specific meanings of different categories.

As shown in figure 6, when identifying bare land and trees, multi-modal image semantic segmentation models tend to misjudge them as low vegetation, which shows that these models only rely on features for classification and fail to fully understand the true meaning of ground objects. In contrast, DenseCLIP, TACOSS and TSMNet, which integrate natural language and remote sensing images, can understand the meaning of categories more accurately, thus achieving more accurate semantic segmentation of open words.

3.5 Ablation Study

Dataset	MFRM	MFFM	OA	mIoU
SWJTU-Vision-Language			68.34	46.98
		✓	68.62	47.38
	✓		69.05	47.12
	✓	✓	**69.73**	**48.83**
YESeg-OPT-SAR			83.12	65.48
		✓	83.49	65.69
	✓		83.92	66.13
	✓	✓	**85.33**	**66.69**

Table 4. Effects of multi-modal image feature fusion network on accuracy (%).

3.5.1 Effectiveness of Multi-modal Image Feature Fusion Network: The section 2 introduces the multi-modal image fusion network for feature-level fusion of optical and SAR images. This module firstly corrects the multi-scale features of optical and SAR images, and then performs feature fusion through attention mechanism. In order to determine the performance of the converged network. Our ablation experiment of this network has two parts of this network. The detailed experimental results are shown in table 4. As can be seen from table 4, the addition of multi-modal image fusion network significantly improves the accuracy of semantic segmentation. This strategy effectively integrates multi-modal remote sensing data and minimizes the potential negative impact of noise information. The model can handle complex modal interactions and maintain high segmentation accuracy on different multi-modal datasets, highlighting its advantages in eliminating noise and other interference information and semantic inconsistency in the process of multi-modal fusion. Although the absolute numerical gains in table 4 may appear modest, this is primarily due to the inherent complexity of the task and the challenging nature of the high-resolution multi-modal datasets. The consistent improvements demonstrate that the MFRM and MFFM modules are not redundant; rather, they are critical for precisely mitigating SAR speckle noise and aligning highly heterogeneous visual semantics without sacrificing local spatial details.

Structure	SWJTU-Vision-Language		YESeg-OPT-SAR	
	OA	mIoU	OA	mIoU
A	68.31	47.99	84.97	64.79
B	69.21	48.36	85.23	65.39
C	68.78	48.53	85.17	65.54
D	**69.73**	**48.83**	**85.33**	**66.69**

Table 5. Effects of natural language on accuracy (%). Structure A has no natural language guidance, B has only category features, C has only text description features, and D contains two categories. Bold values indicate the highest accuracy.

3.5.2 Effectiveness of Object-level Label Feature Fusion of Image and Text Module and Scene-level Semantic Feature Fusion of Image and Text Module: The section 2 introduces the pixel level alignment module of image and text features and the global alignment and fusion module of image and text features. Natural language plays an important role in dealing with multi-modal semantic segmentation. In this paper, two natural languages, category and text description, are aligned and fused with remote sensing images respectively, and the performance of the model is significantly improved. The detailed experimental results are shown in table 5. We can see from the table that the addition of natural language, whether it is category or text description, can significantly improve the performance of semantic segmentation. Although category and text description represent the pixel-level and global alignment of remote sensing images and natural semantics, they can help the model understand reality. The combination of category and text description can further improve the performance of the model. Therefore, TSMNet adopts the method of combining and fusing remote sensing images with natural language pixel-level and global alignment.

4. Conclusion

In this study, we innovatively combine natural language processing with remote sensing image analysis, and propose a new semantic segmentation method of open vocabulary. Unlike the existing visual language model, which mainly focuses on pixel-level category alignment, our proposed TSMNet framework is more in line with human cognitive laws and emphasizes the process of image understanding from the global to the local. By aligning remote sensing image features, local category features and global description features into the semantic space, the framework realizes hierarchical visual representation learning of multi-level and high-level semantics, and effectively integrates image and text features by using cross-modal attention

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-3-2026-51-2026 | © Author(s) 2026. CC BY 4.0 License.

mechanism. In order to improve the quality of multi-modal feature fusion, we specially designed a multi-modal image feature fusion network, which can not only extract multi-modal features, but also effectively correct feature noise, thus fully retaining the detailed information of the image.

In the aspect of dataset construction, we innovatively developed two multi-modal semantic segmentation dataset for practical application scenarios, including multi-source data such as optical images, SAR images and text descriptions. At the same time, on the basis of an existing multi-modal semantic segmentation dataset lacking text information, we have made manual text annotation, which provides more semantic information for model training. The experimental results show that TSMNet has excellent performance on both test datasets. Through visual analysis, the model shows significant advantages in adaptive fusion of multi-modal information, which fully verifies the effectiveness of the proposed method. This study not only promotes the development of remote sensing image semantic segmentation technology, but also provides a new research idea for multi-modal information fusion, which has important theoretical value and practical application significance.

5. Acknowledgements

This paper is supported by the National Key Research and Development Program of China under Grant 2024YFC3015404, in part by the National Natural Science Fund of China under Grant 42271446, and in part by the Science and Technology Program of Tianjin under Grant 24YFYSHZ00080.

References

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.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, S., Ding, H., Jiang, W., 2023. Primitive generation and semantic-related alignment for universal zero-shot segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11238–11247.

Kawano, Y., Aoki, Y., 2024. MaskDiffusion: Exploiting Pre-Trained Diffusion Models for Semantic Segmentation. *IEEE Access*, 12, 127283-127293. https://api.semanticscholar.org/CorpusID:268513010.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, PMLR, 19730–19742.

Li, J., Li, Y., He, L., Plaza, A., 2020. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. *Science China Information Sciences*, 63(140301). https://doi.org/10.1007/s11432-019-2785-y.

Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z., 2025. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10545–10556.

Lin, Y., Suzuki, K., Sogo, S., 2024. Practical techniques for vision-language segmentation model in remote sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 203–210.

Liu, J., Gong, M., Qin, K., Zhang, P., 2018. A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3), 545-559.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, H., Bao, J., Wu, Y., He, X., Li, T., 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *International Conference on Machine Learning*, PMLR, 23033–23044.

Ma, X., Zhang, X., Pun, M.-O., Huang, B., 2025. A Unified Framework With Multimodal Fine-Tuning for Remote Sensing Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1-15.

Ma, X., Zhang, X., Pun, M.-O., Liu, M., 2024. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

Pan, Y., Sun, R., Wang, Y., Yang, W., Zhang, T., Zhang, Y., 2025. Purify Then Guide: A Bi-Directional Bridge Network for Open-Vocabulary Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1), 343-356.

Piramanayagam, S., Saber, E. S., Schwartzkopf, W. C., Koehler, F. W., 2018. Supervised Classification of Multisensor Remotely Sensed Images Using a Deep Learning Framework. *Remote. Sens.*, 10, 1429. https://api.semanticscholar.org/CorpusID:52932570.

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. M. Meila, T. Zhang (eds), *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 139, PMLR, 8748–8763.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Denseclip: Language-guided dense prediction with context-aware prompting. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18082–18091.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Sundaresan, A. A., Solomon, A. A., 2025. Post-disaster flooded region segmentation using DeepLabv3+ and unmanned aerial system imagery. *Natural Hazards Research*, 5(2), 363-371.

Wang, J., Ma, A., Chen, Z., Zheng, Z., Wan, Y., Zhang, L., Zhong, Y., 2024a. EarthVQANet: Multi-task visual question answering for remote sensing image understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212, 422-439.

Wang, L., Dong, S., Chen, Y., Meng, X., Fang, S., Fei, S., 2024b. MetaSegNet: Metadata-Collaborative Vision-Language Representation Learning for Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, PP, 1-1.

Wang, X., Dong, S., Zheng, X., Lu, R., Jia, J., 2024c. Explicit High-Level Semantic Network for Domain Generalization in Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-14.

Wei, K., Dai, J., Hong, D., Ye, Y., 2024. MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 135, 104241.

Wu, Y., Li, J., Yuan, Y., Qin, A. K., Miao, Q.-G., Gong, M.-G., 2022. Commonality Autoencoder: Learning Common Features for Change Detection From Heterogeneous Images. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9), 4257-4270.

Xiao, S., Wang, P., Diao, W., Fu, K., Sun, X., 2025. A Multimodal Semantic Segmentation Framework for Heterogeneous Optical and Complex SAR Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 8083-8098.

Xiong, Z., Wang, Y., Yu, W., Stewart, A. J., Zhao, J., Lehmann, N., Dujardin, T., Yuan, Z., Ghamisi, P., Zhu, X. X., 2025. DOFA-CLIP: Multimodal Vision-Language Foundation Models for Earth Observation. *arXiv preprint arXiv:2503.06312*.

Xu, W., Wang, C., Feng, X., Xu, R., Huang, L., Zhang, Z., Guo, L., Xu, S., 2024. Generalization boosted adapter for open-vocabulary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1), 520–533.

Ye, Y., Dai, J., Zhou, L., Duan, K., Tao, R., Li, W., Hong, D., 2025. Tuple Perturbation-Based Contrastive Learning Framework for Multimodal Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1-15.

Zermatten, V., Castillo-Navarro, J., Marcos, D., Tuia, D., 2025. Learning transferable land cover semantics for open vocabulary interactions with remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220, 621-636.

Zermatten, V., Navarro, J. C., Hughes, L., Kellenberger, T., Tuia, D., 2023. Text as a richer source of supervision in semantic segmentation tasks. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2219–2222.

Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L. M., Zhang, L., 2023. Mp-former: Mask-piloted transformer for image segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18074–18083.

Zhang, S., Huang, J., Wu, Y., Hu, T., Tang, W., Liu, J., 2025. Seg-diffusion: Text-to-image diffusion model for open-vocabulary semantic segmentation. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5.

Zhang, S., Zhang, B., Wu, Y., Zhou, H., Jiang, J., Ma, J., 2024. SegCLIP: Multimodal Visual-Language and Prompt Learning for High-Resolution Remote Sensing Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-16.

Zhou, L., Duan, K., Dai, J., Ye, Y., 2025. Advancing perturbation space expansion based on information fusion for semi-supervised remote sensing image semantic segmentation. *Information Fusion*, 117, 102830.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8-36.