Enrich, distill and fuse: generalized few-shot semantic segmentation in remote sensing leveraging foundation model's assistance.

GAO, T., AO, W., WANG, X.-A., ZHAO, Y., MA, P., XIE, M., FU, H., REN, J. and GAO, Z.

2024

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

Enrich, Distill and Fuse: Generalized Few-Shot Semantic Segmentation in Remote Sensing Leveraging Foundation Model's Assistance

Tianyi Gao[1] , Wei Ao[1] , Xing-ao Wang[1], Yuanhao Zhao[2], Ping Ma[2]
Mengjie Xie[1], Hang Fu[2], Jinchang Ren[2], Zhi Gao[1]
[1]Wuhan University [2]Robert Gorden University
{tianyigao, wei_ao, xingaowang}@whu.edu.cn, p.ma2@rgu.ac.uk, z17806244661@163.com,
mengjie_xie@whu.edu.cn, hangf_upc@163.com, j.ren@rgu.ac.uk, hangf_upc@163.com

Abstract

*Generalized few-shot semantic segmentation (GFSS) unifies semantic segmentation with few-shot learning, showing great potential for Earth observation tasks under data scarcity conditions, such as disaster response, urban planning, and natural resource management. GFSS requires simultaneous prediction for both base and novel classes, with the challenge lying in balancing the segmentation performance of both. Therefore, this paper introduces a novel framework named FoMA, **Fo**undation **M**odel **A**ssisted GFSS framework for remote sensing images. We aim to leverage the generic semantic knowledge inherited in foundation models. Specifically, we employ three strategies named Support Label Enrichment (SLE), Distillation of General Knowledge (DGK) and Voting Fusion of Experts (VFE). For the support images, SLE explores credible unlabeled novel categories, ensuring that each support label contains multiple novel classes. For the query images, DGK technique allows an effective transfer of generalizable knowledge of foundation models on certain categories to the GFSS learner. Additionally, VFE strategy integrates the zero-shot prediction of foundation models with the few-shot prediction of GFSS learners, achieving improved segmentation performance. Extensive experiments and ablation studies conducted on the OpenEarthMap few-shot challenge dataset demonstrate that our proposed method achieves state-of-the-art performance.*

1. Introduction

High-resolution remote sensing images (RSIs) are widely used in many fields of national economic development, such as urban infrastructure assessment [1], land use analysis [2] and environmental monitoring [3]. Semantic segmentation of RSIs is an important way to effectively utilize the image information aimed at parsing the semantic categories of ground features [4]. Although deep neural networks have made significant progress in RSI semantic segmentation, the reliance on annotated data as well as complexity and unknowns from ground targets significantly restricts their applicability [5, 6].

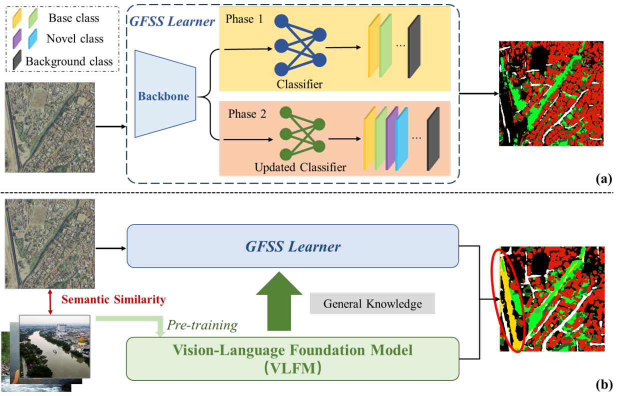

Figure 1. Comparisons between (a) traditional GFSS framework and (b) our foundation model assisted GFSS framework. Since foundation model is pre-trained on web-scale datasets, it can be regarded as a rich knowledge base of general concept, such as river (marked in red circle). It is beneficial to transfer the general knowledge from foundation model to train better GFSS learners.

Few-shot semantic segmentation (FSS) has been proposed to learn a model capable of segmenting novel classes with only a few annotated images [7]. Some researches have been conducted to improve segmentation by generating innovative supporting prototypes [8, 9] and generalized visual segmentation models (e.g., SAM [10]) have recently shown excellence in few-shot learning [11]. Nevertheless, few-shot segmentation usually requires support samples to contain classes that exist in query samples and assess only the novel classes, thus not effectively addressing the challenges of evaluation across both base and novel classes. Therefore, generalized FSS (GFSS) [12] was proposed for the recognition of both base and novel categories. It has made significant strides in processing natural imagery efficiently, allowing for the inclusion of new categories without sacrificing the accuracy of existing ones [13]. However, the objects of RSIs often exhibit more complex scales and confusing semantics. For example, when observed from satellite images, ships and the sea display significant spatial differences, whereas bridges and roads are easily confused since bridges often signify a segment of a road crossing a river. As mentioned in the literature [4, 14], current FSS models do not perform well on these categories in scenarios with limited samples, highlighting the demand of advanced knowledge to deal with various remote sensing scenes.

Recently, the surge of vision-language models such as CLIP [15] and ALIGN [16] has greatly boosted zero-shot learning. After training on large-scale image-text pairs datasets, these models exhibit astonishing recognition capabilities on unseen categories. Benefiting from their zero-shot recognition ability, the performance of few-shot segmentation models can make great progress [17–19]. Nevertheless, these models primarily learn from natural image-level supervision, raising the research question of effective transfer of this kind of knowledge to pixel-level RSI segmentation tasks.

This paper introduces a novel GFSS framework for RSIs, FoMA, **Fo**undation **M**odel **A**ssisted GFSS framework, which is shown in Figure 1. It jointly combines the foundation model and the few-shot learner, leveraging the complementary knowledge of both to enhance the segmentation performance of both the base and novel classes. We observe that a small number of samples is insufficient to train a good classifier, especially for small objects such as boats. Moreover, we also find that to simulate a more realistic GFSS setting, each image in the support set is labeled with only one novel class, while **other potential novel classes are labeled as the background**. This causes semantic ambiguity regarding the novel classes in the support labels, affecting the model's training performance under such ambiguous supervisory information. Since foundation models are pre-trained on web-scale datasets, they **gain certain general knowledge from natural images** regarding some common objects such as river and boats. Based on these findings, we attempt to leverage foundation models for providing more complementary information about support images and query images, transferring the prior knowledge of foundation models to the novel classifier in the few-shot learners by label enrichment and knowledge distillation. Finally, to enhance the segmentation performance of both base and novel classes, we combine the general knowledge of the foundation model with the domain expert knowledge of the few-shot learner, integrating the results of both through a fusion strategy. Additionally, the few-shot learner is implemented with a more robust backbone architecture, which can capture multi-scale semantic information of both base and novel categories and extract potential neighboring semantic clues.

In a nutshell, our key contributions are threefold:

1) We propose a novel foundation model assisted GFSS framework for RSIs termed FoMA, that integrates the general knowledge stored in foundational models with remote sensing domain knowledge, aiming to alleviate the lack of prior due to scarce labeled data. To our best knowledge, FoMA is the first work that adapts a foundation model pre-trained on general purpose datasets to the task of GFSS into the remote sensing field.

2) We design three strategies specially tailored for the GFSS task, in a intuition to promote the knowledge transfer from foundation models to the GFSS learner. Specifically, a Support Label Enrichment (SLE) strategy is first proposed to augment the limited support set with richer information of novel concept via the zero-shot inference of foundation model. In addition, a Distillation of General Knowledge (DGK) is designed for propagating similar semantic concept to the GFSS learner, enhancing the robustness of segmentation. Further, we propose a voting fusion module to ensemble the predictions of foundation model and our GFSS learner adaptively.

3) The proposed FoMA framework demonstrates an outstanding performance on the OpenEarthMap few-shot challenge dataset, surpassing the baseline by an improvement of 28.94%. Specifically, the performance enhancement of novel class is 31.79%, whereas the base class exhibits a 24.64% improvement in performance.

2. Related works

In this section, we list some of the studies that are most relevant to our work.

Few-shot Segmentation for Earth Observation. Few-shot segmentation [7] refers to the task of image segmentation with only a small amount of annotated data. In recent years, with the rise of the few-shot learning (FSL) [20], FSS has been widely studied in natural and medical images. Nevertheless, the extensive coverage and abundant surface information found in remote sensing imagery contribute to the complexity of applying FSS in remote sensing image segmentation, which making it a challenging research domain [21]. Inspired by the success of FSL, most current FSS research utilizes meta-learning-based models to learn how to quickly adapt to new few-shot segmentation tasks [9, 22]. Building upon the success and development of these pioneering methods, a series of FSS approaches tailored for remote sensing images have emerged. For instance, by introducing new metrics and optimization techniques to reduce intra-class variances and maximize inter-class differences, these methods aim to address the impact of complex backgrounds and the diversity of land cover on segmentation results in remote sensing imagery [14, 23, 24]. Other solutions to mitigate the impact of land cover diversity include: exploring multi-scale features of images [25] and altering image scales to reduce the resolution effects [26].

Generalized Few-shot Semantic Segmentation. Despite their considerable potential, conventional FSS methods typically struggle to simultaneously segment both base and novel classes within a query image. In response to the limitations of FSS, [12] proposed the concept of generalized few-shot semantic segmentation. In contrast to FSS, GFSS requires the model to segment all potential base and novel classes in each query image. This implies that the model needs to rapidly adapt to different categories and scenarios with a small amount of labeled data, while also retaining knowledge of the base classes. To achieve this goal, CAPL [12] utilizes two modules to learn base classes and novel classes. However, due to the richer label information available for base classes, the model is biased toward base classes in segmentation results. Recently, BAM [27] has demonstrated outstanding performance in GFSS tasks. The dual-branch structure of BAM enables the segregation of novel and base classes and their effective integration. However, the meta-learning approach poses challenges when applied to GFSS tasks with multiple novel classes. DIaM [13] enhances the consistency between novel and base classes by introducing a Kullback-Leibler term, significantly improving the segmentation results for novel classes. Nonetheless, semantic confusion of ground objects remains a current research challenge for remote sensing GFSS task .

Foundation Model for Zero-shot Semantic Segmentation. Zero-shot semantic segmentation (ZSS) aims to achieve semantic segmentation for classes without prior annotations [28]. Common cues for inferring unknown classes include shared textual attributes and visual-semantic mappings, which emphasizing the significance of aligning visual embedding with class-specific textual embedding for ZSS. Early research efforts were focused on enhancing the model's generalization capability from known to unknown classes [29–31]. Recently, as the popular of foundation models, pre-trained visual-language models such as CLIP [32] have shown astonishing performance in zero-shot classification. These models establish connections between visual features and textual features, enabling visual tasks to no longer be limited to the annotated categories in the training set. For example, MaskCLIP+ [33] introduced pseudo-labeling and self-training to apply CLIP in open vocabulary semantic segmentation. OVSeg [34] proposed a two-stage semantic segmentation framework, where the first stage generates proposal masks and the second stage performs image segmentation based on CLIP. ZegCLIP [35] directly extends CLIP's zero-shot prediction capability from images to pixel-level, aiming to maintain segmentation performance while pursuing simplicity and efficiency. CAT-Seg [36] proposes a cost aggregation-based method to optimize the paired image and text embedding of CLIP, alleviating the problem of transferring the zero-shot capabilities learned from image-level supervision to the pixel-level task. However, since these models are trained on large-scale natural image datasets, there is still significant room for improvement in their application to remote sensing images.

3. Methodology

3.1. Task definition

GFSS extends the traditional FSS task, requiring models to simultaneously segment novel classes while also considering the segmentation accuracy of base classes. In the context of GFSS, the set of classes consists of two non-overlapping parts: the base classes C^b and the novel classes C^n. For the base classes, abundant annotated samples are provided to train a base learner to obtain the ability to recognize base classes which is consistent with typical semantic segmentation pipeline. For each novel class, only K annotated support images are available to adjust the meta learner. During inference, the model needs to segment all the pixels in the query images across $1+\left|C^b\right|+\left|C^n\right|$ potential classes.

3.2. Framework of FoMA

In the GFSS task setting, the problem of segmenting novel classes has always been a challenging issue. We believe this challenge stems mainly from two reasons: First, in the few-shot setting, the number of labeled samples for novel classes is significantly insufficient compared to base classes, making it difficult to obtain a good segmentation model under the same training framework; Second, there exists an issue of semantic ambiguity in the GFSS task. Specifically, each image in the support set is labeled with only one novel class, however, the background of the image may contain pixels of other novel classes, leading to confusion between the background class and the novel classes. Under the paradigm of supervised training, this label ambiguity makes it more difficult for the model to achieve good segmentation performance on novel classes. Therefore, under this limited and incomplete supervisory information, we consider introducing other methods to enrich the label information and enhance the knowledge extraction for novel classes.

The foundation models have been pre-trained on extensive datasets, containing a wealth of general semantic knowledge and concepts, including some categories that frequently appears in remote sensing images, such as water, buildings, etc. Therefore, this general knowledge endows the foundation models with the potential for zero-shot semantic segmentation in GFSS remote sensing tasks. By leveraging the general semantic knowledge of the foundation models to guide the learning of the GFSS learner, it is possible to address the issue of insufficient supervi-

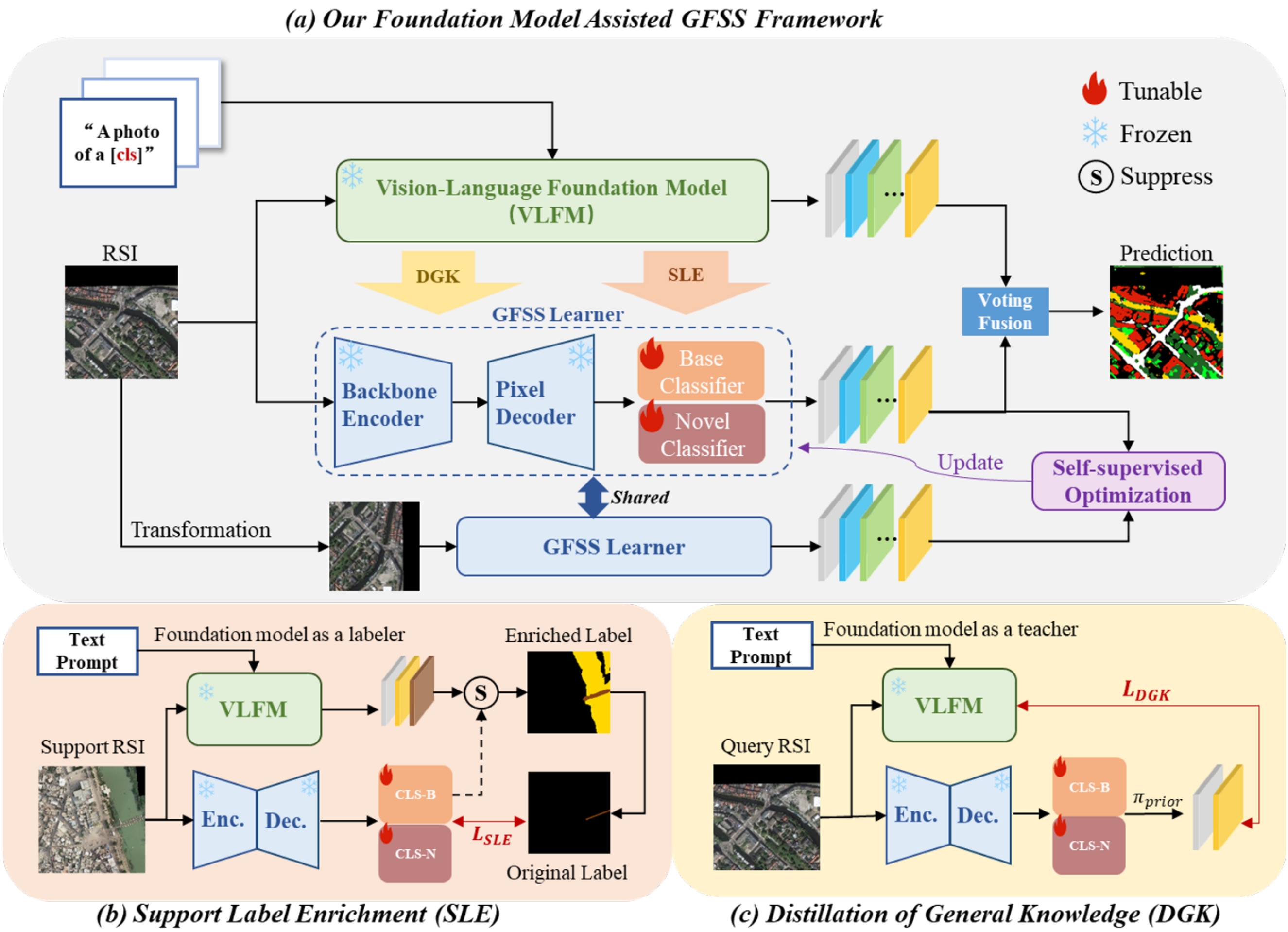

Figure 2. The overall architecture of the proposed Method. This approach introduces the general knowledge of the vision-language foundation model learned from natural images into the remote sensing image GFSS task through two modules: SLE integrates the foundation model's results on support images as pseudo-labels into the GFSS learner's training process. Concurrently, DGK distills the exceptional performance achieved by the foundation model on novel classes from query images into the GFSS learner. Furthermore, a voting fusion strategy is used to effectively merge the results of the foundation model and the GFSS learner across all classes, ensuring more accurate prediction results for the model.

sion information for novel classes. Specifically, we utilize the foundation models to annotate images in the support set with novel class labels, constructing pseudo-labels to obtain more comprehensive supervision information for the novel classes. Subsequently, we employ knowledge distillation techniques to transfer the foundation model's generic knowledge regarding novel classes in a query image to the classifier of the GFSS learner. This approach ensures that the GFSS learner's predictions for novel classes align as closely as possible with those of the foundation model, thereby enhancing its performance for these previously unseen categories. Finally, we combine the general knowledge of the foundation models with the expert knowledge of the GFSS learner, using the information of novel classes to further assist the segmentation tasks of base classes, thereby achieving improved segmentation results. The overall framework is shown in Figure 2.

3.2.1 Support Label Enrichment

Given a support image I, in our challenge setting, only one novel class is labeled. It indicates that even though some other novel classes appear in the image I, the pixels are also labeled as background, making it semantic ambiguous during the training of the novel classifier. Therefore, we aim to leverage off-the-shelf foundation models to enrich the information of current limited support labels. We name it Support Label Enrichment (SLE).

To obtain accurate pseudo labels, we utilize off-the-shelf open-vocabulary semantic segmentation models based on CLIP [37] such as CAT-Seg [36], as an automatic labeler, denoted as F_{zero}, to produce the pseudo-labels of support images. Moreover, we utilize the segmentation network F_{base}, which is trained on abundant base class data to suppress the false prediction. This process can be formulated as:

$$M^s_{zero} = \operatorname{argmax}(F_{zero}\,(I,T)) \tag{1}$$

$$M^s_{base} = \operatorname{argmax}(F_{base}\,(I)) \tag{2}$$

$$M^s_{SLE} = M^s_{gt} \cup (M^s_{zero} \cap \neg M^s_{base}) \tag{3}$$

where I and T denote the image and the text prompt describing potential classes, respectively. M^s_{zero} and M^s_{base} denote labels predicted via zero-shot segmentation models and base classifiers of GFSS learners, respectively. M^s_{SLE} denotes enriched support labels. It should be noted that F_{zero} predicts existing categories of both base classes and novel classes, but also capable of predicting more unlabeled categories, such as car and sports field, to get more accurate semantic segmentation results. Furthermore, since F_{zero} only contains general semantic knowledge and F_{base} performs well on base classes, we utilize the M^s_{base} to suppress the potential incorrect segmentation results in M^s_{zero}. Based on M^s_{SLE}, we optimize the proposed model via a cross-entropy loss:

$$L_{SLE} = CE(F_{few}\,(I)\,, M^s_{SLE}) \tag{4}$$

where F_{few} denotes the GFSS learner and CE denotes cross-entropy loss.

3.2.2 Distillation of General Knowledge

We notice that even with enriched supervision information through our SLE strategy, the performance of the novel classifier remains limited due to the constrained number of images in the support set. Therefore, we attempt to make use of the rich general priors stored in foundation models. Inspired by recent studies about incremental learning that distill knowledge from other models [38–40], we enable the foundation model to serve as a knowledgeable teacher, guiding our GFSS learner in segmenting query images. The process of Distillation of General Knowledge (DGK) can be expressed as:

$$L_{DGK} = KL\left(t \cdot \pi_{prior}(P^q_{zero}) \parallel \pi_{prior}(P^q_{few})\right) \tag{5}$$

where KL denotes Kullback-Leibler divergence and parameter t indicates the temperature which is used to control the learning efficiency of the novel classifier. P^q_{few} and P^q_{zero} is the output probability of our GFSS learner and foundation model according to query images, respectively. π_{prior} is an adjustment function based on expert prior, in order to control the range of knowledge distillation. Since some general concepts are more transferable than some remote sensing domain-specific concepts, such as agric land type 2. Based on the prior of remote sensing experts, π_{prior} moves the probabilities of less transferable categories into background class, allowing a focused distillation on more transferable categories. After iterations of distillation, the segmentation performance of our GFSS learner on the novel classes can gradually approach or even exceed the foundation model.

3.2.3 Voting Fusion of Experts

As mentioned before, the foundation model possesses generic semantic knowledge, thereby exhibiting stronger segmentation capabilities for novel classes; conversely, GFSS learners are trained on remote sensing datasets, demonstrating enhanced segmentation capabilities for base classes. Thus, organically integrating the results of different models at the decision-making level can yield superior outcomes. We find that it is beneficial to adjust the relative weights of the predictions. We refer to this operation as Voting Fusion of Experts (VFE). It is defined as:

$$P^q = argmax(w \cdot P^q_{zero} + P^q_{few}) \tag{6}$$

where P^q_{few} denotes the probability output of base and novel classes by the few-shot learner. The parameter w is the relative weight value. Most importantly, due to foundation models' varied recognition efficacy across different categories, we allocate higher weight values to the categories in which it excels at extraction, while assigning lower weight values to those where extraction performance is less proficient.

3.2.4 Self-supervised Optimization

Self-supervised optimization methods leverage unlabeled data to pre-train models, enabling more efficient utilization of limited labeled data in few-shot scenarios. The self-supervised learning paradigm we employ is grounded in the principle of consistency, leveraging inherent invariant properties of remote sensing images. Specifically, we mandate that alterations such as rotations and flips applied to the images do not alter the predicted outcome for each pixel. L_{SSL} denotes self-supervised learning based loss:

$$L_{SSL} = MSE(F_{few}(I), \Phi_{inv}(F_{few}(\Phi(I)))) \tag{7}$$

where Φ and Φ_{inv} denote the transformations and inversions applied to images, respectively, and MSE denotes mean squared error loss.

Attribute	Train Set	Val Set
Label	7 base classes (tree, rangeland, bareland, agric land type 1, road type 1 sea, lake & pond, building type 1)	4 novel classes (road type 2, river, agric land type 2, boat & ship)
Volume	258 RSIs	50 RSIs (20 for support set)

Table 1. Details of the dataset used in this work.

3.2.5 Overall Optimization

Following the general paradigm of the GFSS task [13, 41], the training of our FoMA framework also consists of two parts: base class training and novel class learning.

Base class training. In the first training phrase, the entire network of the few-shot learner is trained on the given datasets to optimize the feature representation ability and the segmentation performance. The optimization objective consists of typical segmentation loss.

Novel class learning. In the second novel class updating phrase, we keep the backbone encoder and the decoder frozen, and only fine-tune the classifier head (both base and novel). The loss function is formulated as:

$$\begin{aligned} L_{total} &= l_1 \cdot L_{SLE} + l_2 \cdot L_{DGK} + l_3 \cdot L_{SSL} \\ &\quad + l_4 \cdot L_{KD} + l_5 \cdot L_{marg-ent} + l_6 \cdot L^{q}_{cond-ent} \end{aligned} \tag{8}$$

$L_{KD}, L_{cond-ent}$ is the same as that of [13] and $L^{q}_{cond-ent}$ denotes the conditional entropy in [13] but is only calculated on query set. l_1 to l_6 denote weighting coefficients.

4. Experiments

4.1. Datasets and implementation details

Datasets and evaluation metric. The experiments utilized 308 samples of size around 1024×1024 from 11 classes in the OpenEarthMap benchmark dataset [42]. It is worth noting that these 11 classes were included in the training and validation sets in a ratio of 7:4, respectively. The detailed sample amount and class splits are shown in Table 1. For evaluation, we calculate the mean Intersection-over-Union (mIoU) on base class and novel class, as well as their weighted-sum, i.e., 0.4*base mIoU + 0.6*novel mIoU are adopted.

Implementation details. In order to obtain more accurate base classification results, we integrate the outputs of multiple segmentation models such as UNetFormer [43] and HRNet [44]. For the choice of vision-language foundation models, we employ CAT-Seg [36] built upon ViT-G/14 [45]. Except for background class and the existing categories in the training and validation sets, we utilize foundation models to extract additional unlabeled categories, including car, sports field, workshop building, pathway and parking lot, in an intuition to better delineate the outline of target classes. The implementation of our framework is carried out in an NVIDIA A100 environment using PyTorch. Since our approach incorporates the framework of DaIM [13], we select it as a baseline for comparison.

4.2. Spectral-spatial observations in datasets

Due to the complexity of datasets in many scenarios, the phenomena of similar spectra with different objects and different spectra with similar objects are prevalent. In this paper, we conducted an analysis of color confusion among different land cover types using Euclidean similarity evaluation, encompassing both *train_base_class* and *val_novel_class*. In specific, computing the Euclidean similarity on the red, green, and blue channels of pixels for different classes, and taking the average value as the final results. As shown in Figure 3, the analysis reveals significant color confusion among the following categories: (1) "Tree", "Rangeland", "Agric land type 1" and "Road type 2". (2) "Bareland", "Road type 1", "Agric land type 2", and "Road type 2". (3) "Building type 1", "Road type 1" and "Agric land type 2". (4)"Sea, lake & pond", "River" and "Boat & ship".

Despite potential confusion in coloration, many classes exhibit distinct spatial structural features. For instance, "Tree" typically manifest in clustered distributions, while "Rangeland" often display uniform land textures. Agricultural lands commonly adopt regular geometric shapes. Roads, on the other hand, present a relatively homogeneous ground texture, often appearing linear and elongated. Buildings exhibit marked height disparities compared to surrounding terrain, often showcasing distinct geometric forms such as rectangles or squares. The "River" class, in contrast to "Sea, lake & pond", display linear and flowing characteristics. To better classify different land covers, the proposed algorithm effectively utilize the spectral and spatial structural features of images, contextual semantic features, and higher-level features to enhance the discriminative ability among different land cover categories.

4.3. Segmentation results analysis

Quantitative analysis. Table 2 presents a comparison of the mIoU results of our method and the baseline across various categories. It is evident that the proposed algorithm significantly enhances the baseline's segmentation accuracy, improving the final scores by 28.94%. Specifically, the base mIoU increased by 24.64%, while the novel mIoU increased by 31.79%. This improvement stems from the incorporation of foundation model's generic knowledge, which greatly enhances the recognition of novel classes.

Regarding class-based performance, the baseline algo-

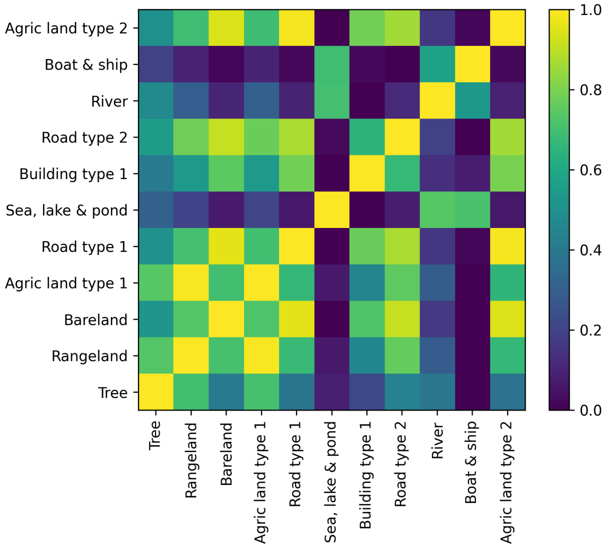

Figure 3. Color confusion analysis among different classes.

Class	Baseline	Ours
Tree	52.47	55.41
Rangeland	35.17	54.41
Bareland	6.33	23.12
Agric land type 1	37.57	68.90
Road type 1	33.78	48.04
Sea, lake, & pond	4.75	65.58
Building type 1	31.80	58.93
Base (Average)	28.84	53.48
Road type 2	0.00	17.01
River	1.45	62.67
Boat & ship	0.00	58.64
Agric land type 2	11.15	1.44
Novel (Average)	3.15	34.94
Base and Novel (Weighted)	13.42	42.36

Table 2. Performance comparison of the segmentation results between the baseline and the proposed method in terms of IoU (%). “Base and Novel (Weighted)” represents the weighted sum of base and novel mIoU, adopted as the challenge evaluation metric.

rithm fails to identify the land cover types “Road type 2” and “Boat & ship”, since these classes only have a small annotated area in the support images, making it difficult to train a good novel classifier using the common GFSS framework, whereas our algorithm successfully distinguishes these two categories leveraging the generic knowledge from the vision-language model. Particularly noteworthy is the high classification accuracy achieved for “Boat& ship”, reaching 58.64%. Furthermore, our proposed algorithm enhances the classification accuracy of the classes “River”, “Sea, lake & pond”, and “Agric land type 1” by 61.22%, 60.83% and 31.33%, respectively. This validates our framework’s ability to effectively utilize spatial structural information between scene objects and contextual semantic information to improve classification accuracy, especially for categories with significant confusion.

Qualitative analysis. In this part, we conduct a comparative analysis of the segmentation results generated by the proposed framework and the baseline approach. Figure 4 presents a visual comparison between the original image (Figure 4 (a)) and the segmentation results obtained from both methods.

It shows that in the results of the baseline, the delineation of roadways appears fragmented, failing to ensure their continuity and integrity. Moreover, a considerable portion of water is erroneously categorized as “Rangeland” or “Agric land type 1”, despite substantial dissimilarities in their features from aquatic bodies. Furthermore, the novel class, “Boat & ship”, is almost dismissed by the baseline model. In stark contrast, the algorithm proposed in this study yields segmentation results that are notably smoother and more precise. Notably, the proposed method achieves a good segmentation performance of “Boat & ship”, and an enhanced delineation of features, preserving their spatial coherence and semantic consistency. The segmentation of road networks demonstrates improved continuity, ensuring an accurate representation of their connectivity across the landscape. Furthermore, the proposed algorithm exhibits superior discriminative capacity, accurately distinguishing between land and water bodies, even in cases of subtle feature variations.

4.4. Ablation study

In order to systematically evaluate the contribution of proposed modules, we conduct a series of ablation studies, as shown in Table 3. Our experiments demonstrate that our base model notably outperformed the baseline in terms of score, primarily due to the utilization of a combination of strong backbones and the optimization strategy. Following the integration of the SLE strategy, the score improved by 1.55% compared to the base model. Furthermore, upon the implementation of the DGK strategy, the performance of model witnesses a substantial enhancement, with an improvement of 16.78% over the base model. This underscores the benefits of leveraging the foundation model’s broad semantic knowledge to enrich supervision information for novel classes or guide the classifier in learning novel concepts, ultimately resulting in improved model performance on GFSS tasks. Most importantly, although the segmentation performance of foundation models on the base classes is not as good as specialist semantic segmentation methods trained on the domain-specific dataset, through our

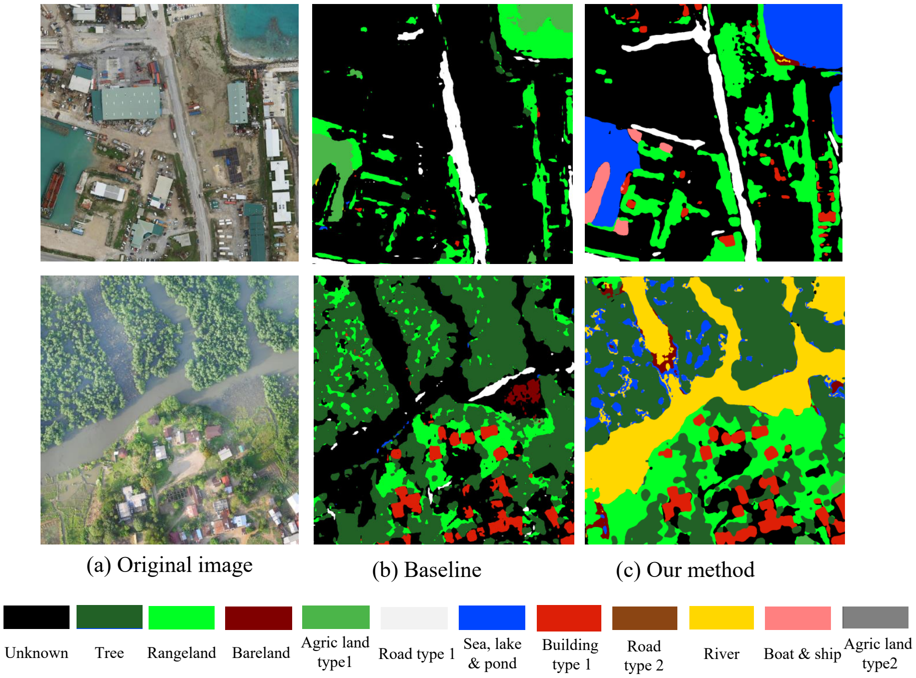

Figure 4. Qualitative comparison between the baseline and the proposed method. Our FoMA prevails in achieving more accurate and complete segmentation of rivers and boats, benefiting from a larger semantic similarity (more generalizable knowledge) between the domains of RSIs and natural images.

Method	Base	Novel	Weighted
Baseline	28.84	3.15	13.42
Our base model	50.16	3.78	22.33
+SLE	50.11	6.40	23.88
+DGK+SLE	50.49	31.53	39.11
+VFE+DGK+SLE	53.48	34.94	42.36

Table 3. Ablation studies on each component of the proposed method. “Base” and “Novel” denote the mIoU of base classes and novel classes, respectively. “Weighted” denotes the weighted sum of IoUs of base and novel classes, adopted as the server scores.

fusion mechanism, the segmentation accuracy on both base and novel classes can be further improved, resulting in a final weighted mIoU of 42.36%. This crucial finding illustrates that there are still semantic relationships in general knowledge that existing GFSS methods struggle to uncover.

5. Conclusions

In this paper, we introduced a novel GFSS framework for high-resolution RSIs, which leverages the generic knowledge from a vision-language foundation model to provide additional supervisory information. Firstly, the annotation of the support set are expanded by generating pseudo labels using the foundation model. Secondly, the promising performance of the foundation model on the query images is transferred into our novel learner through knowledge distillation. Finally, more precise segmentation results can be obtained by integrating the predictions of the foundation model with those of our GFSS learner. The experimental results on the OpenEarthMap few-shot challenge dataset demonstrated that our proposed framework achieved impressive mIoU scores for both base and novel classes. Our method can effectively address the issues of insufficient support images or poor annotation quality, and we believe that with the improvement of the foundation model’s capabilities, our proposed method will exhibit more robust recognition abilities for challenging categories.

Acknowledgements

We would like to thank Prof. Zhi Gao, Prof. Jinchang Ren, and Prof. Shunyi Zheng, for their supervision. Additionally, we extend our thanks to Dr. Hang Fu and Ms. Mengjie Xie for their advice. This work was supported in part by the National Natural Science Foundation of China under Grant 42192580 and 42192583, the Hubei Province Natural Science Foundation under Grant 2021CFA088, and the Wuhan University-Huawei Geoinformatics Innovation Laboratory. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Zhi Gao, Xuhui Zhao, Min Cao, Ziyao Li, Kangcheng Liu, and Ben M. Chen. Synergizing low rank representation and deep learning for automatic pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(10):10676–10690, 2023.

[2] Hang Fu, Genyun Sun, Li Zhang, Aizhu Zhang, Jinchang Ren, Xiuping Jia, and Feng Li. Three-dimensional singular spectrum analysis for precise land cover classification from uav-borne hyperspectral benchmark datasets. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203:115–134, 2023.

[3] Wenfan Qiao, Li Shen, Jicheng Wang, Xiaotian Yang, and Zhilin Li. A weakly supervised semantic segmentation approach for damaged building extraction from postearthquake high-resolution remote-sensing images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

[4] Linhan Wang, Shuo Lei, Jianfeng He, Shengkun Wang, Min Zhang, and Chang-Tien Lu. Self-correlation and cross-correlation learning for few-shot remote sensing image semantic segmentation. *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 2023.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] Amirreza Shaban, Shray Bansal, Z. Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *ArXiv*, abs/1709.03410, 2017.

[8] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9196–9205, 2019.

[9] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5217–5226, 2019.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[11] Jie Zhang, Xubing Yang, Rui Jiang, Wei Shao, and Li Zhang. Rsam-seg: A sam-based approach with prior knowledge integration for remote sensing image semantic segmentation. 2024.

[12] Zhuotao Tian, Xin Lai, Li Jiang, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11553–11562, 2022.

[13] Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and José Dolz. A strong baseline for generalized few-shot semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11278, 2022.

[14] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ArXiv*, abs/2102.05918, 2021.

[17] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022.

[18] Haohan Wang, Liang Liu, Wuhao Zhang, Jiangning Zhang, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Haoqian Wang. Iterative few-shot semantic segmentation from image label text. *arXiv preprint arXiv:2303.05646*, 2023.

[19] Shuai Chen, Fanman Meng, Runtong Zhang, Heqian Qiu, Hongliang Li, Qingbo Wu, and Linfeng Xu. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *IEEE Transactions on Multimedia*, 2024.

[20] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Neural Information Processing Systems*, 2016.

[21] Nicolás Catalano and Matteo Matteucci. Few shot semantic segmentation: a review of methodologies and open challenges. *ArXiv*, abs/2304.05832, 2023.

[22] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seung Wook Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. *ArXiv*, abs/2207.10866, 2022.

[23] Yuyu Jia, Junyu Gao, Wei Huang, Yuan Yuan, and Qi Wang. Holistic mutual representation enhancement for few-shot remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[24] Bing Wang, Zhirui Wang, Xian Sun, Hongqi Wang, and Kun Fu. Dmml-net: Deep metametric learning for few-shot geographic object segmentation in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

[25] Wei Ao, Shunyi Zheng, Yan Meng, and Zhi Gao. Few-shot aerial image semantic segmentation leveraging pyramid correlation fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[26] Alper Kayabasi, Gülin Tüfekci, and Ilkay Ulusoy. Elimination of non-novel segments at multi-scale for few-shot segmentation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2558–2566, 2022.

[27] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8047–8057, 2022.

[28] Max Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *ArXiv*, abs/1906.00817, 2019.

[29] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2019.

[30] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9516–9525, 2021.

[31] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123, 2022.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[33] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, 2021.

[34] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[35] Ziqi Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2022.

[36] Seokju Cho, Heeseong Shin, Sung-Jin Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seung Wook Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *ArXiv*, abs/2303.11797, 2023.

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

[38] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[39] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020.

[40] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10164–10173, 2022.

[41] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11319–11328, 2023.

[42] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6243–6253, 2022.

[43] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[44] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[45] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.