

Self-Correlation and Cross-Correlation Learning for Few-Shot Remote Sensing Image Semantic Segmentation

Linhuan Wang

linhan@vt.edu

Virginia Tech

Virginia, United States

Shuo Lei

slei@vt.edu

Virginia Tech

Virginia, United States

Jianfeng He

jianfenghe@vt.edu

Virginia Tech

Virginia, United States

Shengkun Wang

shengkun@vt.edu

Virginia Tech

Virginia, United States

Min Zhang

minzhang23@vt.edu

Virginia Tech

Virginia, United States

Chang-Tien Lu

ctlu@vt.edu

Virginia Tech

Virginia, United States

ABSTRACT

Remote sensing image semantic segmentation is an important problem for remote sensing image interpretation. Although remarkable progress has been achieved, existing deep neural network methods suffer from the reliance on massive training data. Few-shot remote sensing semantic segmentation aims at learning to segment target objects from a query image using only a few annotated support images of the target class. Most existing few-shot learning methods stem primarily from their sole focus on extracting information from support images, thereby failing to effectively address the large variance in appearance and scales of geographic objects. To tackle these challenges, we propose a **Self-Correlation and Cross-Correlation Learning Network** for the few-shot remote sensing image semantic segmentation. Our model enhances the generalization by considering both self-correlation and cross-correlation between support and query images to make segmentation predictions. To further explore the self-correlation with the query image, we propose to adopt a classical spectral method to produce a class-agnostic segmentation mask based on the basic visual information of the image. Extensive experiments on two remote sensing image datasets demonstrate the effectiveness and superiority of our model in few-shot remote sensing image semantic segmentation. Code and models will be accessed at <https://github.com/linhanwang/SCCNet>.

CCS CONCEPTS

- Computing methodologies → Neural networks.

KEYWORDS

remote sensing image semantic segmentation, few-shot learning

ACM Reference Format:

Linhuan Wang, Shuo Lei, Jianfeng He, Shengkun Wang, Min Zhang, and Chang-Tien Lu. 2018. Self-Correlation and Cross-Correlation Learning for Few-Shot Remote Sensing Image Semantic Segmentation. In *Proceedings of*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX). ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Semantic segmentation in remote sensing images has become an essential task for various applications, such as land use analysis [11], urban management [35], environmental monitoring [13], and other areas of national economic development. Although deep neural networks for semantic segmentation [2, 4, 18, 43] have achieved remarkable progress, their reliance on large-scale datasets greatly restricts their application in low-resource domains. For example, collecting an adequate amount of remote sensing data is hard, and the expense associated with hiring domain experts to annotate the data is too costly to be feasible. To reduce such burden on data annotation, few-shot semantic segmentation has been proposed [24], which aims to learn a model that can perform segmentation on novel classes with only a few annotated images.

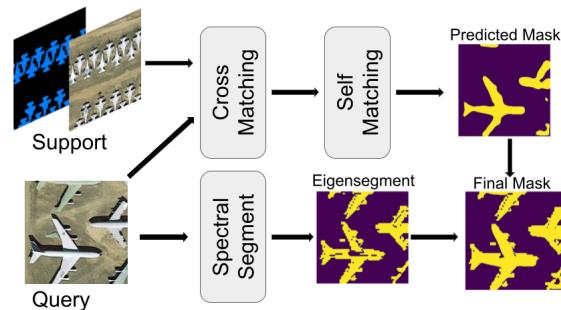


Figure 1: Overview of our proposed model (SCCNet) for few-shot remote sensing image semantic segmentation. SCCNet incorporates self-correlation information into the model and decomposes an image using the eigenvectors of a Laplacian matrix constructed from a visual feature to extract self-contained knowledge from the query image.

Recently, a group of few-shot segmentation methods adopted global average pooling [24] over the foreground region of the support features to generate class prototypes, which are then employed

to guide the segmentation process of the query image. Building upon this research direction, some studies [33, 36] strive to design more representative support prototypes to enhance segmentation performance. While significant advancements have been made for natural images, these methods encounter challenges when applied to remote sensing images, primarily due to the presence of large intra-class variances. Specifically, geographic objects of the same class can exhibit substantial variations in appearance and scales from different angles. Recently, SDM [38] proposes scale-aware focal loss to focus training on tiny hard-parsed objects and performs detailed matching with multiple prototypes for providing more accurate parsing guidance. However, SDM only considers the cross-correlation between support and query images, ignoring the self-correlation between pixels within the query image. We found that self-correlations within the query image could provide extra knowledge to help segment tiny objects, which is very important for few-shot remote sensing image semantic segmentation, particularly when there is a significant discrepancy between the support and query images.

To address the aforementioned challenges, we propose a novel model, named *SCCNet*, to leverage knowledge from query images for few-shot remote sensing image semantic segmentation. As illustrated in Fig. 1, the proposed model consists of two key components. First, we incorporate the initial query mask prediction to collect query features in high-confidence regions and then use the generated masked query features to perform self-matching with query features. Since pixels belonging to the same object are expected to exhibit higher similarity than those belonging to different objects, Self-Matching Module can provide auxiliary support information to segment the query image. Second, we propose a novel Spectral Segmentation Module to extract knowledge from query images further with classical spectral methods. Specifically, we first construct the affinity matrices using basic visual information (i.e. color and position information) and semantic information derived from the middle-layer features of the pretrained backbone. Then we decompose images using the eigenvectors of Laplacian of affinity matrices as soft segments and obtain the class-agnostic eigensegments. Since it operates solely on the query images without relying on the support annotations, it is naturally resilient to the significant discrepancies that may exist between the support and query images. The final prediction mask of the query image is obtained by fusing the optimized query mask and the eigensegment.

Our key contributions can be summarized as follows:

- We propose a **Self-Correlation and Cross-Correlation Learning Network** for the few-shot remote sensing image semantic segmentation. Our model enhances the generalization by considering both self-correlation within query images and cross-correlation between support and query images to make segmentation predictions.
- We proposed a Self-Matching Module to extract more comprehensive query information. The correlation between the initial segment and the query images is introduced to the model to tackle the large discrepancy between support and query images.

- We propose a novel Spectral Segmentation Module with spectral analysis to produce class-agnostic segmentations of query images without the supervision of any annotations.
- We evaluate the proposed model on two remote sensing image datasets for few-shot semantic segmentation tasks. Comprehensive experiments demonstrate that our *SCCNet* consistently outperforms all the baselines for both 1-shot and 5-shot settings.

2 RELATED WORK

2.1 Remote Sensing Image Semantic Segmentation

Deep learning-based methods have gained significant popularity in the remote sensing community, showcasing remarkable progress in segmenting remote sensing images. Specifically, Maggiori et al. [20] introduced a multilayer perceptron (MLP) into the segmentation network to produce better segmentation results. Yu et al. [39] introduced the pyramid pooling module as a means to address semantic segmentation in remote sensing images, while Yue et al. [40] developed TreeUNet as the first adaptive Convolutional Neural Network (CNN) specifically tailored for semantic segmentation in this domain. Zhang et al. [42] adopted the multibranch parallel convolution structure in HRNet [29] to generate multiscale feature maps and designed an adaptive spatial pooling module to aggregate more local contexts. To tackle the challenge in small-scale object segmentation, Kamppffmeyer et al. [12] assembled patch-based pixel classification and pixel-to-pixel segmentation, which introduced uncertain mapping to achieve high performance on small-scale objects. FactSeg [19] proposed a symmetrical dual-branch decoder consisting of a foreground activation branch and a semantic refinement branch. The two branches performed multiscale feature fusion through skip connection, thereby improving the accuracy of small-scale object segmentation. Furthermore, with the emergence of multiple attention mechanisms, Ding et al. [7] designed an efficient local attention embedding to enhance segmentation performance.

Although existing methods effectively demonstrate the capabilities of deep learning in remote sensing image semantic segmentation, they typically require a large number of densely-annotated images for training and have difficulties in generalizing to unseen object categories.

2.2 Few-shot Semantic Segmentation

To address the generalization issue and reduce massive training data annotation, Few-Shot Semantic Segmentation (FSS) task has been proposed, which aims to learn a model that can perform segmentation on novel classes with only a few pixel-level annotated images. Shaban et al. [24] first proposed one-shot semantic segmentation networks to address FSS. It uses global average pooling over the foreground region of the support features to generate class prototypes, which are then employed to guide the segmentation process of the query image. Building upon the concept of prototypical networks [28], utilizing prototype representations to guide mask prediction in query images has become a popular paradigm in the field of few-shot segmentation. Specifically, PANet [33] proposed a prototype alignment regularization between support and query

images to generate high-quality prototypes. PMMs [36] employ the Expectation-Maximization algorithm to generate multiple prototypes corresponding to different parts of the objects. Recently, a group of matching-based methods has been proposed to leverage dense correspondences between query images and support annotations. HSNet [22] utilizes 4D convolutions to extract precise segmentation masks by compressing the multilevel 4D correlation tensors. VAT [10] proposes a 4D Convolutional Swin Transformer to aggregate the correlation map. To fully harness the information within the support set, Yang et al. [37] employ clustering techniques to mine latent novel classes in the support set and subsequently treat them as pseudo labels during the training process.

Despite the remarkable progress achieved in natural images, Yao et al. [38] found that performance drops dramatically on unseen classes in remote sensing images. This limitation arises from the inability of these methods to effectively handle the significant variations in object appearance and scales prevalent in remote sensing images. To address this challenge, SDM [38] proposes a scaled-aware focal loss, which enhances the focus on tiny objects. DMML-Net [32] uses an affinity-based fusion mechanism to adaptively calibrate the deviation of the prototype induced by intra-class variation.

It is worth noting that all existing methods primarily focus on extracting information solely from the support set to make a segmentation. However, we argue that this approach may not be sufficient for remote sensing images, where substantial discrepancies exist between the support and query images. In this study, we aim to pioneer a novel direction by extracting the self-contained knowledge in the query images to boost the performance for few-shot remote sensing image semantic segmentation.

2.3 Spectral Methods for Segmentation

Spectral analysis originally emerged from the exploration of continuous operators on Riemannian manifolds [3]. Subsequent research efforts extended this line of research to the discrete setting of graphs, leading to numerous findings that connect global graph properties to the eigenvalues and eigenvectors of their associated Laplacian matrices. Lin et al. [17] demonstrate that the eigenvectors of graph Laplacians yield graph partitions with minimum energy. Building upon this insight, Shi et al. [26] view image segmentation as a graph partitioning problem and propose a novel global criterion called the normalized cut for image segmentation. As presented by Aksoy et al. [1], soft segmentations are automatically generated by fusing high-level and low-level image features within a graph structure. The construction of this graph facilitates the utilization of the corresponding Laplacian matrix and its eigenvectors to reveal semantic objects and capture soft transitions between them.

3 PROBLEM SETUP

Few-shot semantic segmentation aims to perform segmentation on the novel classes with only a few annotated images. Suppose we are provided with images from two non-overlapping class sets: C_{base} and C_{novel} . The training dataset \mathcal{D}_{train} is constructed from the class set C_{base} and the test dataset \mathcal{D}_{test} is constructed from the class set C_{novel} .

To mitigate the risk of overfitting caused by limited training data, we adopt a commonly used meta-learning technique known as episodic training [31]. In the K -shot setting, we employ episodic sampling to select $K+1$ annotated image pairs, denoted as $\{(I_1^s, M_1^s), (I_2^s, M_2^s), \dots, (I_K^s, M_K^s), (I^q, M^q)\}$, with the same targeted class from the training dataset \mathcal{D}_{train} . Here, $\{(I_i^s, M_i^s)\}_{i=1}^K$ represents the support samples, and (I^q, M^q) denotes the query pair. During the training phase, the segmentation model takes both the support samples $\{(I_i^s, M_i^s)\}_{i=1}^K$ and the query image I^q as inputs and generates a predicted mask \tilde{M}^q . This prediction is then supervised by the corresponding ground truth mask M^q . Similarly, during the testing phase, we employ K annotated image pairs from \mathcal{D}_{test} to infer the semantic objects present in the query images.

4 PROPOSED APPROACH

To solve the few-shot semantic segmentation problem in remote sensing images, we propose a novel model named *SCCNet* as shown in Fig 2. First, we use pre-trained CNNs (VGG [27] or Resnet [9]) as the feature extractor to generate the corresponding query and support features. In the cross-matching module, pixel-wise multi-scale correlation tensors between masked support features and query features are built and squeezed to generate the initial predicted query mask \tilde{M}_{init}^q . To tackle the high intra-class variance problem in remote sensing images, the Self-Matching Module calculates the correlations between query features masked by \tilde{M}_{init}^q and other query features. These correlations are further squeezed and merged with squeezed correlations between support and query features to generate optimized query mask \tilde{M}_{merge}^q . To further mine knowledge from the query images, in the spectral segmentation module, the classic spectral analysis method is utilized to exploit the proximity of local regions. Specifically, eigenvectors of the Laplacian of the affinity matrix are utilized as soft segments and transformed into eigensegments by Thresholding algorithms afterward. In the end, the final prediction mask of the query image is obtained by fusing the optimized query mask and the eigensegment.

4.1 Cross-Matching Module

Different from encoding an annotated support image to a feature vector to facilitate query image segmentation, we adopt the pixel-wise correlation between the support and query images to make a segmentation in our Cross-Matching Module.

Hypercorrelation pyramid construction. We extract features from query and support images and compute the correlation between them. Given a pair of query and support images, I^q and I^s , we adopt a pretrained backbone to produce a sequence of L feature maps, $\{(F_l^q, F_l^s)\}_{l=1}^L$, where F_l^q and F_l^s denote query and support feature maps at the l -th level, respectively. A support mask M^q is used to encode segmentation information and filter out the background information. We obtain a masked support feature as $\hat{F}_l^s = F_l^s \odot \zeta_l(M^s)$, where \odot denotes the Hadamard product and $\zeta_l : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{C_l \times H_l \times W_l}$ denotes a function that resizes the given tensor followed by expansion along the channel dimension of the l -th layer.

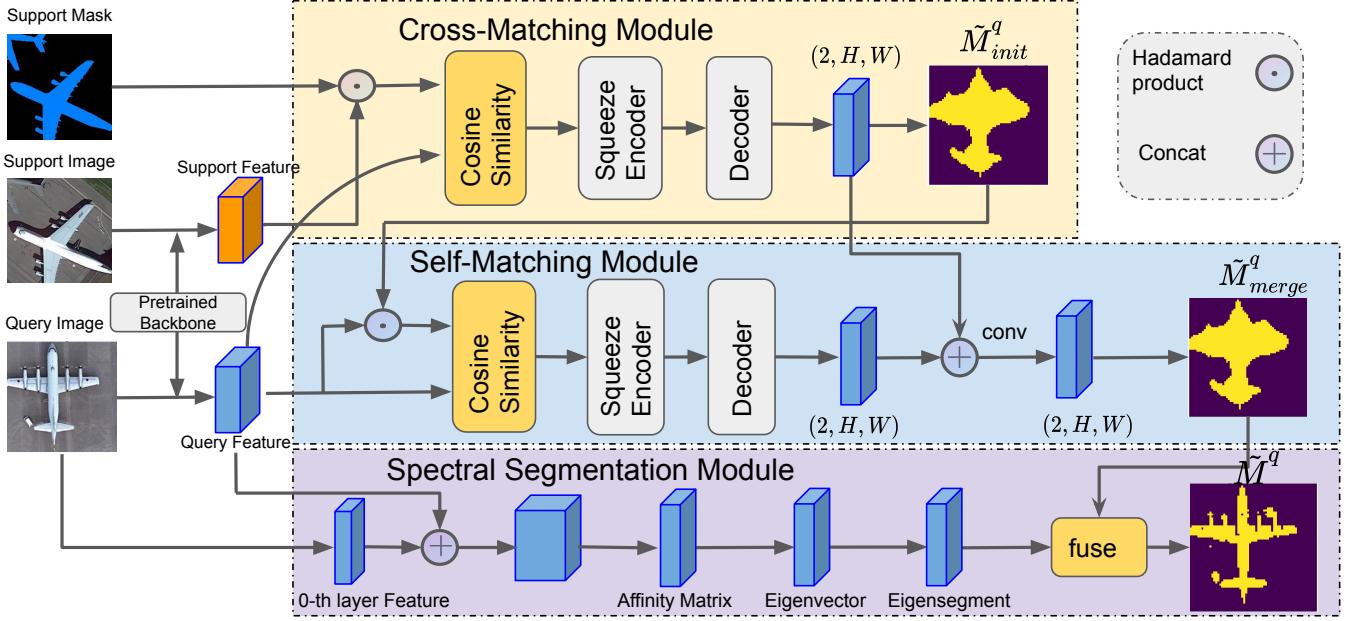


Figure 2: Overall pipeline of our proposed network. The pretrained backbone is utilized as the feature extractor to generate corresponding support and query features. In the Cross-Matching Module, correlation between masked support features and query features is built and squeezed to generate the initial query mask. In the Self-Matching Module, correlation between query image and the initial query mask is further added into the model to generate finer query mask. Finally in the Spectral Segmentation Module, the query mask is fused with eigensegments obtained by non-learning-based spectral analysis.

Given a pair of feature maps F_l^q and F_l^s , we compute a 4D correlation tensor [22] $\hat{C}_l \in \mathbb{R}^{H_l \times W_l \times H_l \times W_l}$ using cosine similarity:

$$\hat{C}_l(i, j) = \text{ReLU} \left(\frac{F_l^q(i) \cdot \hat{F}_l^s(j)}{\|F_l^q(i)\| \|F_l^s(j)\|} \right) \quad (1)$$

where i and j denote 2D spatial positions of feature maps. We collect correlation tensors computed all the intermediate features of the same spatial size and stack them to obtain a stacked correlation map $\hat{C}_p \in \mathbb{R}^{|\mathcal{L}_p| \times H_p \times W_p \times H_p \times W_p}$, where (H_p, W_p) are the height and width of the query and support feature maps, and \mathcal{L}_p is a subset of CNN layer indices $\{1, \dots, L\}$ at pyramid layer p , containing correlation maps of identical spatial size. Given P pyramidal layers, we denote the hypercorrelation pyramid as $\hat{C} = \{\hat{C}_p\}_{p=1}^P$, representing a collection of feature correlations from multiple visual aspects.

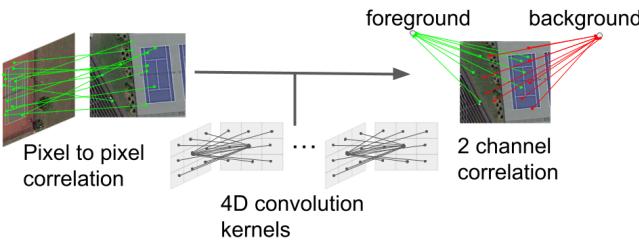


Figure 3: Simplified illustration of the effect of 4D convolution kernels that squeeze the support spatial dimensions.

Correlation Squeeze Encoder. Our encoder network takes the hypercorrelation pyramid $\hat{C} = \{\hat{C}_p\}_{p=1}^P$ to effectively squeeze it into a condensed feature map $Z \in \mathbb{R}^{128 \times H_1 \times W_1}$. As shown in Figure 3, sequences of multi-channel 4D convolution with large strides periodically squeeze the last two (support) spatial dimensions of \hat{C}_p down to (H_ϵ, W_ϵ) while the first two spatial (query) dimensions remain the same as (H_p, W_p) . Similar to FPN [16] structure, two outputs from adjacent pyramidal layers, p and $p+1$, are merged by element-wise addition after upsampling the (query) spatial dimensions of the upper layer. After merging, the output tensor of the lowest block is further compressed by average-pooling its last two (support) spatial dimensions, which in turn provides a 2-dimensional feature map $Z \in \mathbb{R}^{128 \times H_1 \times W_1}$ that signifies a condensed representation of the hypercorrelation pyramid \hat{C} .

2D-convolutional context decoder. The decoder network consists of a series of 2D convolutions, ReLU, and upsampling layers followed by softmax function. The network takes the context representation Z and predicts two-channel map $\hat{M}_{init}^q \in [0, 1]^{2 \times H \times W}$ where two channel values indicate probabilities of foreground and background. Then we take the maximum channel value at each pixel of \hat{M}_{init}^q to obtain initial query mask prediction $\tilde{M}_{init}^q \in \{0, 1\}^{H \times W}$.

4.2 Self-Matching Module

While the cross-matching module successfully captures intricate correlations between support and query images, it faces limitations when significant disparities exist between the support and query features. Consequently, the initial query mask \tilde{M}_{init}^q generated by

the cross-matching module may lack crucial details, which is a pain point for the segmentation task. To tackle this issue, Self-Matching Module (SMM) is proposed to provide auxiliary support information to segment the query image.

Suppose the query image is I^q , and the initial query mask is \tilde{M}_{init}^q . In the Self-Matching Module, different from calculating the correlation tensor between masked support features and query features, we calculate the correlation tensor between initial masked query features and query features:

$$\hat{C}_l^{self}(i, j) = \text{ReLU} \left(\frac{F_l^q(i) \cdot \hat{F}_l^q(j)}{\|F_l^q(i)\| \|F_l^q(j)\|} \right), \quad (2)$$

$$\text{where } \hat{F}_q^l = F_q^l \odot \zeta_l(\tilde{M}_{init}^q) \quad (3)$$

Following the procedure in Cross-Matching Module, we can obtain $\hat{M}_{self}^q \in [0, 1]^{2 \times H \times W}$. Then, we concatenate \hat{M}_{self}^q with M^q and utilize 1x1 conv to reduce the channel dimension to get $\hat{M}_{merge}^q \in [0, 1]^{2 \times H \times W}$.

In Self-Matching Module, the loss function \mathcal{L}_m for training the model can be computed as follows:

$$\mathcal{L}_m = \text{BCE}(\hat{M}_{merge}^q, M^q) \quad (4)$$

where $\text{BCE}(\cdot)$ is the binary cross entropy loss and M^q is the ground truth mask of the query image.

To further facilitate the Self-Matching procedure, we propose a query self-matching loss:

$$\mathcal{L}_{aux} = \text{BCE}(\hat{M}_{aux}^q, M^q) \quad (5)$$

Here, \hat{M}_{aux}^q is generated following the procedure of \hat{M}_{self}^q , but with ground truth query mask to calculate the masked query feature $\hat{F}_q^l = F_q^l \odot \zeta_l(M^q)$. The motivation is that the quality of the initial predicted query mask directly influences the auxiliary information extracted during the self-matching stage. Finally, we train the model in an end-to-end manner by jointly optimizing $\mathcal{L} = \mathcal{L}_m + \lambda \mathcal{L}_{aux}$, where λ serves as weight strength, and we set $\lambda = 1.0$ in our experiments.

4.3 Spectral Segmentation Module

Self-Matching Module incorporates the proximity between the initial query mask \tilde{M}_{init}^q and the query image within the model, effectively addressing the challenge of large intra-class variance. However, the performance of this module is influenced by the quality of \tilde{M}_{init}^q . To overcome this limitation, we employ a spectral analysis method to extract valuable knowledge from the affinity matrix, which is constructed solely based on the query image.

The derivation of the affinity matrix is the key to spectral decomposition. Inspired by Melas-Kyriazi et al. [21], we leverage the features f from the middle layer of the pretrained backbone to construct an affinity matrix. Additionally, since the features are extracted for aggregating similar features rather than anti-correlated features, we set the affinity thresholding as :

$$Z_{sem}(i, j) = f_i f_j^T \odot (f_i f_j^T > 0) \quad (6)$$

While the affinities derived from embedding features are rich in semantic information, it lacks low-level proximity including color

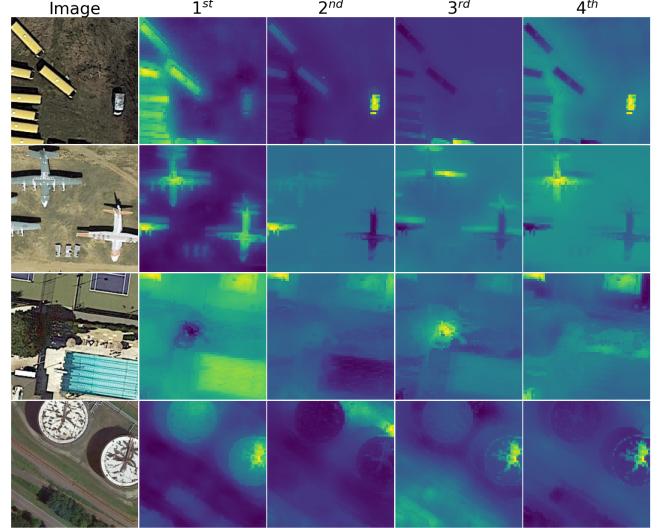


Figure 4: Visualization examples of first four Eigenvectors of our feature affinity matrix on iSAID-5^l dataset. The eigenvectors correspond to semantic regions, with the first eigenvector usually identifying the most salient object in the image.

similarity and spatial distance. To solve this problem, we adopted image matting [5, 14] to consider the basic visual information in Spectral Segmentation Module. Specifically, we first transform the input image into the HSV color space: $X(i) = (\cos(h), \sin(h), s, v, x, y)_i$, where h, s, v are the respective HSV coordinates and (x, y) denotes the spatial coordinates of pixel i . Here X contains color information and position information which can be seen as the 0-th layer feature of the network. Then, we construct a sparse affinity matrix from pixel-wise nearest neighbors based on X :

$$Z_{knn}(i, j) = \begin{cases} 1 - \|X(i) - X(j)\|, & \text{if } i \in \text{KNN}_X(j) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $\|\cdot\|$ denotes 2-norm and $i \in \text{KNN}_X(j)$ are the k-nearest neighbors of j under the distance defined by X . The overall affinity matrix is defined as the weighted sum of the two:

$$Z(i, j) = Z_{sem}(i, j) + \alpha Z_{knn}(i, j) \quad (8)$$

The residual ratio α is the hyper-parameter weighing the importance of the visual and semantic information. Empirically, we set $\alpha = 5$ in our experiments.

With the affinity matrix Z , we can compute the top N eigenvectors $\{E_i\}_{i=0}^{N-1}$ of the Laplacian L . As shown in Figure 4, after being resized to $H \times W$, the eigenvectors are soft segments with continuous values. To convert the soft segments to the hard mask predictions, we propose to introduce two thresholding algorithms into Spectral Segmentation Module. The pipeline of this combination process is illustrated in Fig. 5. Specifically, we first utilize Multi-Ostu algorithm [15] to find salient objects and adopt Adaptive Thresholding [8] algorithm to extract the sharp boundaries in the eigenvectors. Then we combine them together with Hadamard product to generate the final eigensegments $\tilde{E}_i \in \{0, 1\}^{H \times W}$:

$$\tilde{E}_i = \text{MultiOstu}(E_i) \odot \text{Adap}(E_i), i \in \{1, \dots, N-1\} \quad (9)$$

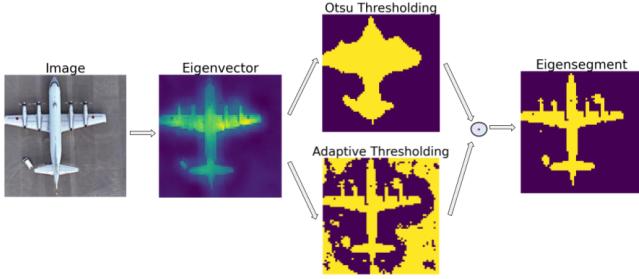


Figure 5: Pipeline of generating eigensegment from an image.

where we exclude the zero-th constant eigenvector.

4.4 Inference

Given a pair of annotated images $(I^s, M^s), (I^q, M^q)$, we first generate the predicted query mask \tilde{M}_{merge}^q through Cross-Matching and Self-Matching Modules. Meanwhile, we calculate the top $N - 1$ spectral eigensegments $\{\tilde{E}_i\}_{i=1}^{N-1}$ of each query image. Since the eigensegments are class-agnostic, we fuse the merged mask \tilde{M}_{merge}^q with the first eigenvector \tilde{E}_1 , which has the highest confidence, to obtain the final prediction. In addition, to explore the full potential of spectral segmentation, we also present the result of selecting the best \tilde{E}_{best} from $\{\tilde{E}_i\}_{i=1}^{N-1}$ ranked by IoU with ground truth query mask. The final prediction of query mask is a union of \tilde{E}_{best} and \tilde{M}_{merge}^q :

$$\tilde{M}^q = \Phi(\tilde{M}_{merge}^q, \tilde{E}_{best}) \quad (10)$$

where $\Phi(\cdot)$ is pixel-wise logical *or* function.

Our model can be easily extended to K -shot setting: Given K support image-mask pairs $\mathcal{S} = \{(I^s, M_k^s)\}_{k=1}^K$ and a query image I^q , the model performs K forward passes to provide a set of K mask predictions $\{\tilde{M}_k^q\}_{k=1}^K$. We perform voting at every pixel location by summing all the K predictions and dividing each output score by the maximum voting score. We assign foreground labels to pixels if their values are larger than some threshold τ whereas the others are classified as background. We set $\tau = 0.4$ in our experiments.

5 EXPERIMENT

To demonstrate the effectiveness of the proposed method, the experiments are organized as follows. We first describe the adopted dataset iSAID- 5^i and DLRSD- 5^i . Next, the evaluation metrics and implementation details are introduced. Then, the segmentation results and comparison with the state-of-the-art few-shot segmentation methods are presented. We finally conducted a series of ablation studies to analyze the impact of each component in our proposed method.

5.1 Datasets

iSAID- 5^i The iSAID dataset [34] contains 655,451 object instances for 15 categories across 2,806 high-resolution images, which exactly match the requirement of the few-shot segmentation task. Based on this, Yao et al. [38] create the iSAID- 5^i dataset following the setting in PASCAL- 5^i [24], and the class details are show in Table 1. Particularly, for the 15 object categories in the iSAID- 5^i dataset,

the cross-validation method is leveraged to evaluate the proposed model by using five classes in one fold as test categories \mathcal{D}_{test} and leveraging the ten classes in the left two folds as the categories of the training set \mathcal{D}_{train} . The details of the class splits are shown in Table 3, where i is the fold number. For every fold, we use the same model with the same hyperparameter setup following standard cross-validation protocol. The iSAID- 5^i dataset contains 18,076 images for training, 6,363 images for validation and the resolution of all the images is fixed to be 256×256 . Furthermore, this dataset provides sufficient size diversity for the few-shot remote sensing images' semantic segmentation task.

DLRSD- 5^i The Dense Labeling Remote Sensing Dataset (DLRSD) [25] is a publicly available dataset for evaluating multi-label remote sensing image retrieval and semantic segmentation algorithms. DLRSD contains 2,100 RGB images in total, 17 object classes and the image sizes are fixed as 256×256 pixel. To balance the number in each fold, we use 15 categories of DLRSD to build DLRSD- 5^i . The details of the class splits are shown in Table 4.

5.2 Evaluations metrics

We adopt mean intersection over union (mIoU) as our evaluation metrics. For each category, the IoU is calculated by $\text{IoU} = \frac{TP}{TP+FN+FP}$, where TP, FN, FP respectively denote the number of true positive, false negative and false positive pixels of the predicted mask. The mIoU metric averages over IoU values of all classes in a fold: $\text{mIoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c$ where C is the number of classes in the target fold and IoU_c is the intersection over union of class c .

5.3 Implementation details

For the backbone network, we employ VGG [27] and ResNet [9] families pre-trained on ImageNet [6], e.g., VGG16, ResNet50, and ResNet101. For VGG16 backbone, we extract features after every conv layer in the last two building blocks: from conv4_x to conv5_x, and after the last maxpooling layer. For ResNet backbones, we extract features at the end of each bottleneck before ReLU activation: from conv3_x to conv5_x. This feature extracting scheme results in 3 pyramidal layers ($P = 3$) for each backbones. In spectral segmentation module, we peek the layer with size 64×64 as f to construct affinity matrix Z_{sem} , which contains rich semantic information and high resolution. The image size in both iSAID- 5^i and DLRSD- 5^i is 256×256 , i.e., $H, W = 256$. This network is implemented in PyTorch [23] and optimized with SGD optimizer where the learning rate is 9e-4, the weight decay is 5e-4, and the momentum is 0.9. The learning rate is scheduled with polynomial strategy. The backbone is trained together with 10 times smaller learning rate.

5.4 Compared with SOTA

To assess the efficacy of our model, we extensively compare it with state-of-the-art (SOTA) methods [22, 30, 33, 36, 38, 41] on the iSAID- 5^i and DLRSD- 5^i dataset, employing different backbone networks and few-shot settings.

iSAID- 5^i Table 5 presents a summary of the results on iSAID- 5^i . When using \tilde{E}_1 , our method outperforms other state-of-the-art methods in almost all the experiment settings. Notable, with Resnet50 as backbone, our method achieves 4.57% and 2.33% improvement in mIoU over the state-of-the-art in the 1-shot setting

Table 1: Classes in iSAID-5ⁱ Dataset

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
ship	storage tank	baseball diamond	tennis court	basketball court	ground track field	bridge	large vehicle	small vehicle	helicopter	swimming pool	roundabout	soccer ball	plane	harbor field

Table 2: Classes in DLRSD-5ⁱ Dataset

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
airplane	bare soil	buildings	cars	chaparral	court	dock	field	grass	mobile home	pavement	sand	sea	ship	tanks

Table 3: Testing Classes for Threefold Cross Validation Test

Dataset	Test classes
iSAID-5 ⁰	ship, storage tank, baseball diamond, tennis court, basketball court
iSAID-5 ¹	ground track field, bridge, large vehicle, small vehicle, helicopter
iSAID-5 ²	swimming pool, roundabout, soccer ball field, plane, harbor

Table 4: Testing Classes for Threefold Cross Validation Test

Dataset	Test classes
DLRSD-5 ⁰	airplane, bare soil, buildings, cars, chaparral
DLRSD-5 ¹	court, dock, field, grass, mobile home
DLRSD-5 ²	pavement, sand, sea, ship, tanks

and 5-shot setting respectively. When \tilde{E}_{best} is used, the improvement is further enlarged and comes to 8.35% and 3.39%.

DLRSD-5ⁱ Table 6 presents a summary of the results on DLRSD-5ⁱ. Resnet50 is used as the backbone. When \tilde{E}_1 is used, our method achieves 2.73% and 1.26% improvement in 1-shot setting and 5-shot setting respectively. When \tilde{E}_{best} rather than \tilde{E}_1 is used, our method achieves 4.68% improvement over the state-of-the-art in the 1-shot setting.

To conduct a more thorough analysis of the performance across diverse classes in the few-shot setting, we have gathered detailed results for the one-shot scenario, utilizing the ResNet50 [9] backbone. The specific outcomes are presented in Table 7 and 8 on iSAID-5ⁱ and DLRSD-5ⁱ. On both datasets, our model demonstrates the highest performance when compared to other state-of-the-art (SOTA) methods in 10 out of 15 categories, while in the remaining classes, our model achieves the second-best performance. This substantiates the effectiveness and versatility of our approach.

Notably, we observe an intriguing trend where the improvement in the 1-shot setting is more significant than that in the 5-shot setting across all three backbones. This observation aligns with our design choice, suggesting that our method effectively mitigates intra-class variation. Conversely, in the 5-shot setting, it is more likely that some support images closely resemble the query image.

Considering the extensive analysis conducted, we can confidently conclude that our proposed method effectively tackles the

few-shot semantic segmentation task for remote sensing images. Qualitative results are shown in Fig 6.

5.5 Ablation study

Ablation study on designed modules. To further demonstrate the effectiveness of our designed modules, we conduct ablation experiments on iSAID-5ⁱ using the 1-shot setting and ResNet50 backbone. Table 9 presents the results obtained. The baseline model solely comprises the Cross-Matching Module, which is based on HSNet [22]. By introducing the Self-Matching Module, we observe a notable improvement of 3.95% in mIoU. This outcome highlights the significant benefit derived from the Self-Matching Module, which introduce proximity information between initial query mask and query image into the model.

Ablation study on fusion strategy of eigensegments. As shown in Table 9, when we fuse \tilde{E}_1 with \tilde{M}_{init}^q generated by the Cross-Matching Module, we achieve a notable improvement of 1.27% in mIoU, which proves the efficacy of the Spectral Segmentation Module. When we fuse \tilde{E}_1 with \tilde{M}_{merge}^q generated by Self-Matching Module, the total improvement comes to 4.57%, which is a large margin. In addition, our investigation reveals that the target object is not always contained within the first eigensegment, as it may not be the most salient foreground object. For instance, in the first image of Figure 4, the buses are the most salient objects and they are present in the first eigenvector, while the small vehicle is present in the second eigenvector. To fully explore the capabilities of spectral segmentations, as discussed in Section 4.4, we fuse \tilde{E}_{best} with \tilde{M}_{merge}^q . This operation yields a significant increase in improvement, with a difference of 8.13% from the baseline. This result demonstrates that the Spectral Segmentation Module, which solely mine knowledge from the query image, successfully tackles the large discrepancies between the support and query image observed in remote sensing images.

Ablation study on design of Self-Matching Module. In our model architecture, we employ a two-branch network, where the Cross-Matching Module and Self-Matching Module have separate weights. This choice doubles the number of learnable parameters in our model. To investigate the possibility of reducing memory consumption, we conduct an ablation study on a single-branch structure, where the Cross-Matching Module and Self-Matching Module share the same weights. However, as shown in Table 10, the performance of the single-branch structure is even inferior to that of HSNet [22], not to mention the two-branch network.

Table 5: Performance on iSAID- 5^i in mIoU. Some results are reported in [38]. Numbers in bold indicate the best performance and underlined ones are the second best. Superscript \dagger denotes \tilde{E}_{best} is used instead of \tilde{E}_1 .

Backbone network	Methods	1-shot				5-shot				learnable params
		fold0	fold1	fold2	mean	fold0	fold1	fold2	mean	
VGG16	PANet [33]	17.43	11.43	15.95	14.94	17.7	14.58	20.7	17.66	14.7M
	CANet [41]	19.73	17.98	30.93	22.88	23.45	20.53	30.12	24.70	26.4M
	PMMs [36]	20.87	16.07	24.65	20.53	23.31	16.61	27.43	22.45	25.8M
	PFENet [30]	16.68	15.3	27.87	19.95	18.46	18.39	28.81	21.89	10.4M
	SDM [38]	29.24	20.80	<u>34.73</u>	28.26	36.33	27.98	42.39	<u>35.57</u>	25.8M
	HSNet [22]	22.74	23.05	25.76	23.84	27.20	28.86	28.82	28.29	2.6M
	Ours	30.00	<u>27.41</u>	32.43	<u>29.94</u>	36.52	<u>31.40</u>	37.53	35.15	5.2M
Resnet50	Ours \dagger	35.71	30.33	36.68	34.24	40.40	32.56	<u>39.31</u>	37.42	5.2M
	PANet [33]	12.36	9.11	12.05	11.17	13.82	12.4	19.12	15.11	23.5M
	CANet [41]	18.8	15.62	25.79	20.07	23.86	18.54	32	24.8	20.2M
	PMMs [36]	19.02	18.51	28.42	21.98	20.89	20.87	31.23	24.33	19.6M
	PFENet [30]	18.75	17.24	22.09	19.36	19.57	18.43	26.14	21.38	10.8M
	SDM [38]	34.29	22.25	35.62	30.72	39.88	<u>30.59</u>	45.70	38.72	19.6M
	HSNet [22]	30.76	24.35	38.20	31.10	38.08	30.56	45.28	37.79	2.6M
Resnet101	Ours	<u>36.21</u>	<u>27.42</u>	<u>43.37</u>	<u>35.67</u>	<u>42.58</u>	30.30	<u>50.26</u>	<u>41.05</u>	5.2M
	Ours \dagger	40.74	31.25	46.40	39.45	44.27	31.62	50.45	42.11	5.2M
	HSNet [22]	34.91	26.51	40.84	34.09	41.71	31.08	48.54	40.44	2.6M
	Ours	<u>37.65</u>	29.19	42.99	36.13	41.87	32.12	49.63	41.20	5.2M
	Ours \dagger	40.82	31.38	45.32	39.17	42.52	32.72	<u>49.18</u>	41.47	5.2M

Table 6: Performance on DLRD- 5^i in mIoU. Resnet50 is used as the backbone

Methods	1-shot				5-shot				
	fold0	fold1	fold2	mean	fold0	fold1	fold2	mean	
SDM [38]	20.11	30.84	27.87	26.27	26.03	41.74	33.55	33.77	
HSNet [22]	22.00	47.20	34.73	34.64	27.46	<u>52.32</u>	46.23	42.00	
Ours	<u>25.34</u>	<u>48.97</u>	<u>39.73</u>	<u>37.37</u>	<u>30.22</u>	52.40	<u>47.15</u>	43.26	
Ours \dagger	26.48	49.59	41.89	39.32	30.26	51.08	47.60	42.92	

Table 7: Performance comparisons of diverse classes on the iSAID- 5^i dataset with 1-shot setting

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
SDM [38]	37.66	<u>34.37</u>	34.45	39.81	25.14	16.77	34.53	<u>30.50</u>	<u>12.42</u>	17.02	20.69	56.83	42.80	<u>40.52</u>	<u>17.26</u>
HSNet [22]	18.93	30.01	<u>37.60</u>	<u>45.33</u>	21.95	<u>25.11</u>	37.17	27.43	11.03	<u>21.01</u>	<u>32.22</u>	50.07	54.27	37.98	16.46
Ours	<u>26.76</u>	43.42	40.27	46.74	<u>22.10</u>	27.37	<u>36.75</u>	32.94	14.53	23.89	46.85	<u>55.06</u>	<u>45.77</u>	48.02	23.30

Table 8: Performance comparisons of diverse classes on the DLRD- 5^i dataset with 1-shot setting

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
SDM [38]	5.51	<u>22.74</u>	29.00	3.83	<u>39.49</u>	5.30	19.97	84.13	8.94	<u>35.90</u>	11.96	31.99	49.03	<u>38.79</u>	7.57
HSNet [22]	23.63	18.39	21.41	<u>8.55</u>	38.02	<u>63.45</u>	<u>24.56</u>	96.49	<u>18.33</u>	33.20	20.88	24.66	<u>57.13</u>	35.00	<u>35.94</u>
Ours	<u>23.58</u>	25.32	<u>26.99</u>	10.45	40.37	<u>53.27</u>	25.49	<u>96.29</u>	29.10	40.68	30.09	<u>24.80</u>	60.07	46.72	37.00

This observation suggests that the Cross-Matching Module and Self-Matching Module have subtle differences, and sharing weights actually harms the performance of the Cross-Matching Module instead of enhancing it. Nevertheless, due to the sparse design of center-pivot 4D convolution [22] we adopt, our model still has a relatively small number of learnable parameters compared to other methods [30, 33, 36, 38, 41].

Ablation study on α of spectral segmentation module. In the spectral segmentation module, α is a key hyperparameter to balance the semantic affinity matrix Z_{sem} and Z_{knn} which contains raw image information. To select the best value of α , we construct some ablation studies on iSAID- 5^i with 1-shot setting and Resnet50 backbone. As shown in Table 11, $\alpha = 5$ achieves the best performance.

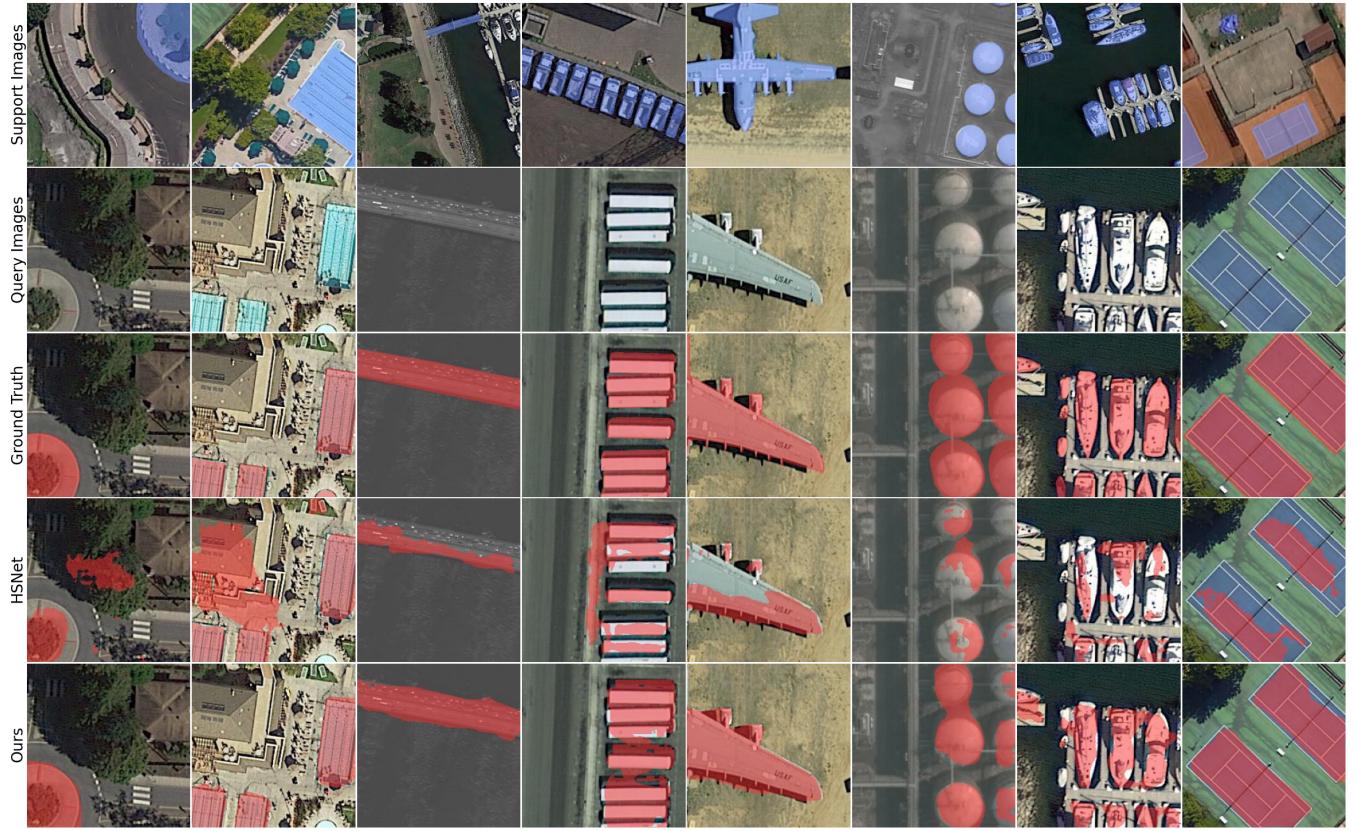
Figure 6: Qualitative results of 1-shot settings on iSAID-5ⁱ dataset.

Table 9: Ablation study on the designed components of the proposed algorithm

Self-Matching	\tilde{E}_1	\tilde{E}_{best}	fold0	fold1	fold2	mIoU
×	×	×	30.76	24.35	38.20	31.10
×	✓	×	32.65	25.34	39.12	32.37↑1.27
✓	×	×	34.64	26.85	43.36	35.05↑3.95
✓	✓	×	36.21	27.42	43.37	35.67↑4.57
✓	×	✓	39.80	29.70	48.19	39.23↑8.13

Table 10: Ablation study on design of Self-Matching Module. In single-branch design, we use same 4D conv kernels in both Cross-Matching and Self-Matching Module.

Experiments	fold0	fold1	fold2	mIoU
HSNet [22]	30.76	24.35	38.20	31.10
single-branch	26.12	25.77	38.82	30.24
two-branch	34.64	26.85	43.36	35.05

6 CONCLUSION

In this work, we propose a novel *SCCNet* for the few-shot remote sensing image semantic segmentation task. Specifically, Self-Matching Module is designed to incorporate the initial query mask

Table 11: Ablation study on the hyperparameter α in the Spectral Segmentation Module.

α	fold0	fold1	fold2	mIoU
1	36.62	27.50	42.63	35.58↑0.63
5	36.21	27.42	43.37	35.67↑0.72
10	35.91	27.41	43.41	35.58↑0.63
20	35.86	27.10	43.80	35.58↑0.63
50	35.65	26.96	43.42	35.34↑0.39

prediction to collect query features in high-confidence regions and then use the generated query prototype to perform self-matching with query features. In addition, we propose the Spectral Segmentation Module with spectral analysis methods to produce class-agnostic segmentations of query images without the supervision of any annotations. The proposed model is evaluated on two commonly adopted benchmarks for few-shot remote sensing image semantic segmentation. Without any extra knowledge or data information, our *SCCNet* outperforms previous work by a large margin.

REFERENCES

- [1] Yağız Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. 2018. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.
- [3] Jeff Cheeger. 1970. A lower bound for the smallest eigenvalue of the Laplacian. Problems in analysis (Papers dedicated to Salomon Bochner, 1969).
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.
- [5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. 2013. KNN matting. *IEEE transactions on pattern analysis and machine intelligence* 35, 9 (2013), 2175–2188.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [7] Lei Ding, Hao Tang, and Lorenzo Bruzzone. 2020. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 59, 1 (2020), 426–435.
- [8] Rafael C Gonzalez. 2009. *Digital image processing*. Pearson education india.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [10] Sungwhan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. Springer, 108–126.
- [11] Dino Ienco, Roberto Interdonato, Raffaele Gaetano, and Dinh Ho Tong Minh. 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), 11–22.
- [12] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1–9.
- [13] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing* 173 (2021), 24–49.
- [14] Anat Levin, Dani Lischinski, and Yair Weiss. 2007. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence* 30, 2 (2007), 228–242.
- [15] Ping-Sung Liao, Tse-Sheng Chen, Pau-Choo Chung, et al. 2001. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.* 17, 5 (2001), 713–727.
- [16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
- [17] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. 2020. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407* (2020).
- [18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.
- [19] Ailong Ma, Junjue Wang, Yanfei Zhong, and Zhuo Zheng. 2021. Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–16.
- [20] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. 2017. High-resolution aerial image labeling with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* 55, 12 (2017), 7092–7103.
- [21] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. 2022. Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization. In *CVPR*.
- [22] Juhong Min, Dahyun Kang, and Minsu Cho. 2021. Hypercorrelation Squeeze for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [24] Amireza Shaban, Shrav Bansal, Zhen Liu, Irfan Essa, and Byron Boots. 2017. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410* (2017).
- [25] Zhenfeng Shao, Ke Yang, and Weixun Zhou. 2018. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing* 10, 6 (2018), 964.
- [26] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.
- [27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [28] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).
- [29] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514* (2019).
- [30] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. 2020. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence* 44, 2 (2020), 1050–1065.
- [31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016).
- [32] Bing Wang, Zhirui Wang, Xian Sun, Hongqi Wang, and Kun Fu. 2021. DMML-Net: Deep metameric learning for few-shot geographic object segmentation in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–18.
- [33] Caixia Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*. 9197–9206.
- [34] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. 2019. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 28–37.
- [35] Michael Wurm, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck. 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing* 150 (2019), 59–69.
- [36] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. 2020. Prototype mixture models for few-shot semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 763–778.
- [37] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. 2021. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8721–8730.
- [38] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. 2021. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–11.
- [39] Bo Yu, Lu Yang, and Fang Chen. 2018. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 9 (2018), 3252–3261.
- [40] Kai Yue, Lei Yang, Ruirui Li, Wei Hu, Fan Zhang, and Wei Li. 2019. TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing* 156 (2019), 1–13.
- [41] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. 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*. 5217–5226.
- [42] Jing Zhang, Shaofu Lin, Lei Ding, and Lorenzo Bruzzone. 2020. Multi-scale context aggregation for semantic segmentation of remote sensing images. *Remote Sensing* 12, 4 (2020), 701.
- [43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881–2890.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009