Holistic Mutual Representation Enhancement for Few-Shot Remote Sensing Segmentation

Yuyu Jia, Junyu Gao, *Member, IEEE,* Wei Huang,
Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Few-shot segmentation endeavors to utilize a minimal amount of annotated samples (*support*) to guide the segmentation of unseen objects (*query*). Previous techniques primarily employ a *support-to-query* paradigm, neglecting to sufficiently leverage the mutual representation between query and support images, which leaves models suffering from intra-class variations and background interference in remote sensing images. This paper proposes a Holistic Mutual Representation Enhancement (HMRE) method to bridge these gaps. First, a Dual Activation (DA) module is devised to establish information symmetry between the two branches and forms the foundation for mutual representation enhancement. Subsequently, the holistic mutual enhancement is jointly constructed by the Global Semantic (GS) and Spatial Dense (SD) mutual enhancement modules. In the prediction stage for segmentation, we integrate the enhanced mutual representation into the Mutual-Fusion Decoder to activate the homologous object regions bidirectionally. To expedite the replication of investigation in this task, we further create a corresponding benchmark Flood-3i. The whole dataset is attainable at https://drive.google.com/drive/folders/1FMAKf2sszoFKjq0UrUmSLnJDbwQSpfxR. Extensive experiments on two benchmarks iSAID-5i and Flood-3i demonstrate the superiority of our proposed method, which also sets a new state-of-the-art.**

***Index Terms*—few-shot semantic segmentation, mutual representation enhancement, remote sensing images.**

I. Introduction

SEMANTIC segmentation plays a pivotal role in the domain of remote sensing image interpretation. It has been widely explored for practical applications, such as building/road extraction [1], land-cover mapping [2], [3], disaster assessment [4], and environmental protection [5]. Recently, deep learning techniques [6], [7], [8], [9] have greatly contributed to the remarkable advancements in remote sensing image segmentation algorithms. However, these successful deep-learning techniques intensely depend on extensive pixel-level annotations, which incur substantial time and labor costs for practical applications. Moreover, their generalization capability to novel tasks could be fragile [10].

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902.

Yuyu Jia, Junyu Gao, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

Wei Huang is with the Data Science in Earth Observation, Technical University of Munich, Munich 80333, Germany.

E-mail: jyy2019@mail.nwpu.edu.cn, gjy3035@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com, hw2hwei@gmail.com.

Qi Wang is the corresponding author.

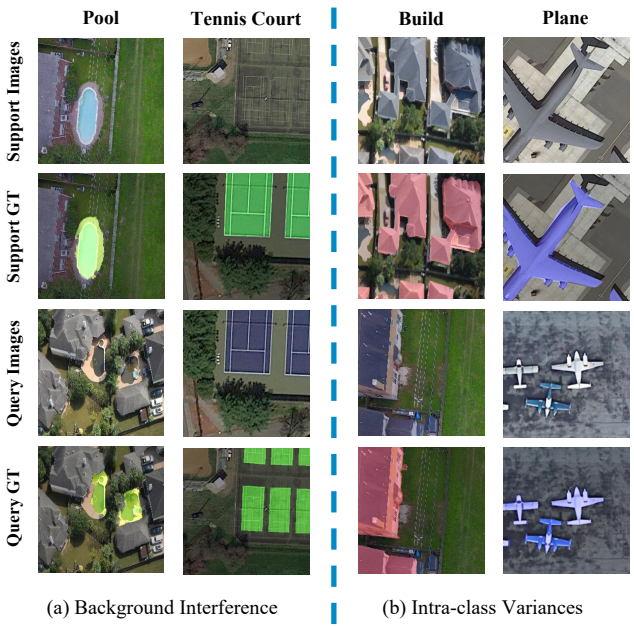


Fig. 1. Demonstration of the issues in handling the FSS task for remote sensing images. (a) Non-target regions with the same category in the background of both support and query images (*e.g.*, buildings or trees) could interfere with the segmentation of target regions (*e.g.*, pool). (b) The variations in appearance, size, and distribution of objects (*e.g.*, planes) belonging to the same category in the support and query images are substantial.

In that case, Few-Shot Learning (FSL) has been extensively investigated. It acquires transferable knowledge and subsequently generalizes it to novel (unseen) tasks with only a few labeled samples. Likewise, image segmentation inherits the paradigm of FSL, constructing a meaningful and challenging task, termed Few-Shot Segmentation (FSS). Given an image (*query*) to be segmented in the test set, the FSS model utilizes the information from several densely annotated samples (*support*) to identify target regions containing the predefined semantic category. Current FSS methods normally embrace a two-branch structure to separately extract the support and query feature, in which the support branch widely uses the Mask Average Pooling [11] operation to acquire class-specific features related to the target region. Then, the class-specific cues are embedded into the query branch to guide the segmentation through pixel-wise similarities. These techniques unidirectionally transmit category representation (*support-to-query*), formulating the FSS task as a guided segmentation task. However, under such a paradigm, the mutual information

between query and support remains underutilized in addressing intra-class variances presented by objects of the same category with diverse appearances (see Fig. 1(b)). Intuitively, the query information could, in turn, direct the segmentation of the support (*query-to-support*).

Taking these into consideration, some research in natural image processing employs symmetric structures for two-branch interaction. PANet [12] introduces a prototype alignment strategy that performs bidirectional prediction between support and query. BriNet [13] reports an information exchange module to learn the non-local spatial transforms between support and query images, which is more complex and omits the global representation in each channel. CRNet [14] proposes a cross-reference module to enhance the global mutual representation in the two branches. Despite the recent progress showcased by these investigations, they still struggle with the following points: (i) The enhancement of global mutual representation is hasty, as the incorporation of Global Average Pooling inevitably introduces a large amount of background noise. As shown in Fig. 1(a), due to the wide geographical coverage of remote sensing images, non-target regions of the same category often appear concurrently in the background of both support and query samples, which may mislead target region segmentation. (ii) The enhancement of spatial mutual representation with only mid-level features is insufficient for capturing target location. It overlooks the valuable integration of high-level features, which provide complementary information necessary for precise localization. (iii) The absence of a holistic framework that concurrently investigates both global semantic mutual enhancement and spatially dense mutual enhancement of representations.

In this work, we develop a Holistic Mutual Representation Enhancement (HMRE) method for overcoming the constraints of existing works. Fig. 2 depicts the flow of data in HMRE and concisely showcases the role of each component. Given a pair of support and query images, along with a support mask, a proposed Dual Activation (DA) module activates a pseudo-mask for the query image and a prior map for the support image, respectively. This process achieves information symmetrization of the two branches, laying the foundation for mutual representation enhancement. Subsequently in a Global Semantic (GS) mutual enhancement module, the generated pseudo-mask is utilized to extract a pure target region prototype in the query image. Then, it combines the support prototype to compute a coefficient for global semantic mutual enhancement. Intuitively, only the channels sharing analogous global target semantics in both branches will receive enhancement through the aforementioned coefficient. In addition to this, we also devise a Spatial Dense (SD) mutual enhancement module to generate the spatial dense coefficient for each branch. Compared with previous spatial interaction strategies, the proposed approach incorporates high-level spatial information (*i.e*., the support and query prior maps), which offers complementary insights essential for accurate localization. Finally, with the two types of coefficients, the Mutual Fusion Decoder achieves holistic enhancement of mutual representation and optimizes the above processes using bidirectional prediction.

Moreover, there is a shortage of remote sensing datasets specifically tailored for the FSS. To our best knowledge, only two datasets (*i.e*., iSAID-5i [15], DLRSD [16]) have been used for this purpose, where the scenes have simple backgrounds or few target objects. Considering this fact, we create a new benchmark constructed from FloodNet [17], named Flood-3i. Compared to DLRSD, the proposed dataset offers more samples, allowing for a comprehensive evaluation of algorithms. Meanwhile, each image in this dataset has a background with a wider variety of object categories than iSAID-5i, providing more challenges for the FSS task. Further details regarding this dataset will be elaborated in section IV.

our main contributions can be outlined as follows:

1) The Global Semantic mutual enhancement module is designed to mitigate the interference from non-target regions in the background while enhancing the global semantics of co-occurring target regions in the two branches.
2) The proposed Spatial Dense mutual enhancement module lightly fuses the positional information of high-level feature maps and densely enhances the spatial representation of the two branches.
3) A new benchmark, Flood-3i, is meticulously created for the FSS of remote sensing images. The Holistic Mutual Representation Enhancement method proposed in this study is evaluated on two benchmarks (*i.e*., Flood-3i and iSAID-5i), where it outperforms the prior leading techniques.

II. Related Works

A. Semantic Segmentation

Semantic segmentation assigns predefined semantic categories to each pixel, which is one of the most foundational image-processing tasks The Fully Convolutional Network (FCN) based models [18], [19] substitute fully connected layers with convolutional layers to preserve the spatial information. Furthermore, a series of optimization algorithms improve the above model structures, such as dilated convolutions [20], pyramid pooling [21], Atrous Spatial Pyramid Pooling (ASPP) [22], attention mechanism [23], encoder-decoder [24], multi-scale feature aggregation [25], *etc*.

As for the interpretation of remote sensing images, deep learning-based semantic segmentation technologies also attract widespread attention from researchers [26], [27], [28]. For instance, He *et al*. [29] proposes MANet for extracting perceptive and multi-scale representations in remote sensing images. Chen *et al*. [30] proposes AERFC, which adaptively controls convolution sampling locations and adjusts effective receptive fields to reduce training difficulty and preserve image details. HFGNet [26] places emphasis on the significance of background information and enriches the foreground saliency features.

Despite achieving high segmentation accuracy, these methods are limited by the requirement for a substantial amount of annotated data. Furthermore, they often encounter difficulties when facing unseen target objects. To effectively alleviate these, we undertake the FSS task and fully explore a class-agnostic mutual representation enhancement method to boost the model's generalization.

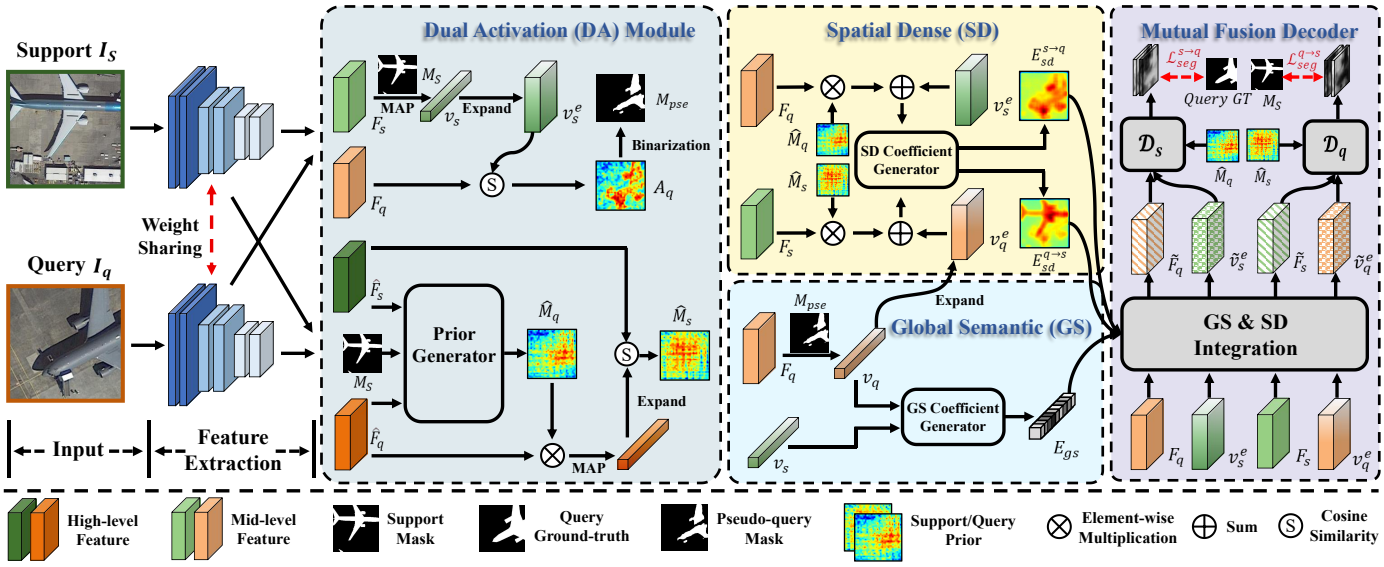


Fig. 2. Pipeline of the HMRE method for 1-shot setting. Given a query image, a support image, and a support mask, the pretrained ResNet50 [31] or VGG16 [32] with fixed weights is leveraged to extract mid-level and high-level features. To achieve symmetry of the information between the two branches, we complete a pseudo query mask and a support prior mask in the Dual Activation (DA) module. Based on this, we devise the Global Semantic (GS) and Spatial Dense (SD) mutual enhancement modules to calculate corresponding coefficients. Finally, the Mutual Fusion Decoder instantiates the holistic mutual representation enhancement by weighting the prototypes and mid-level features. Moreover, we adopt a bidirectional prediction paradigm to optimize the above process.

B. Few-Shot Learning

Few-Shot learning (FSL) aims to generalize knowledge to novel categories by leveraging a small amount of annotated data [33]. Currently, the majority of FSL algorithms are situated within the meta-learning framework. Generally speaking, these studies can be grouped into two branches: metric-based approaches [34], [35], [36], [37] and optimization-based approaches [38], [39], [40], [41]. Metric-based methods categorize a query sample according to its relative distance to each category prototype in the feature space. Optimization-based methods focus on learning optimal initialization parameters or optimizers that allow quick adaptation to novel classes.

Acquiring a vast collection of annotated remote sensing images presents a formidable challenge, rendering the FSL in this domain a task that is both significant and arduous. Researchers have dedicated extensive efforts to this undertaking. For example, Li *et al*. [42] presents DLA-MatchNet, which is designed via discriminative learning of adaptive matching. Li *et al*. [43] proposes SCL-MLNet to address the limitations of existing methods in utilizing limited annotated data and handling intra-class sample variations. Jia *et al*. [44] devises a multiview-attention mechanism to explore the information across different views of images, and boosts the model's generalization by optimizing feature distributions of hard samples. To capture intrinsic characteristics across semantic classes during the training of all base classes, Ji *et al*. [45] employs a cross-entropy loss with two auxiliary objectives.

Previous studies on FSL have primarily concentrated on accomplishing classification tasks. In contrast, our research tackles the more demanding task of few-shot semantic segmentation, which contributes to achieving a more sophisticated comprehension of remote sensing images.

C. Few-Shot Segmentation

Few-Shot Segmentation (FSS) inherits the main settings of FSL and achieves dense prediction of images, which is first proposed by Shaban *et al*. [46], termed OSLSM. Similar to OSLM, the majority of subsequent works have adopted a dual-branch network structure (*i.e*., a support branch and a query branch). Specifically, Zhang *et al*. [46] firstly utilize a mask average pooling approach to gather guidance information that exclusively considers pixels belonging to the support image. Zhang *et al*. [47] introduces CANet, a novel segmentation network that leverages a two-branch dense comparison module and an iterative optimization module for performing multi-level feature comparison and refining predicted results. Siam *et al*. [48] presents an adaptive masked proxies strategy, which leverages multiresolution average pooling on masked embeddings to construct the final segmentation layer weights for new classes. Later, PFENet [49] introduces the training-free prior mask, which significantly improves the model's generalization performance on unseen categories. On this basis, some relevant works further improve the model and achieve reliable performance, such as SD-AANet [50], MMNet [51], SCLNet [52], and PFENet++ [53].

Specific to the research of FSS of the remote sensing image, there is currently limited work in this area, therefore further investigations are needed. In current employment, Wang *et al*. [16] develops an affinity-based fusion mechanism to handle the intra-class variations. Yao *et al*. [15] mitigates the substantial variance in object appearances through a detailed matching module and a scale-aware focal loss, and more importantly, they provided a benchmark iSAID-5i for the research of FSS in remote sensing images. Both these two methods address the crucial challenge of the FSS task and propose corresponding

solutions from the view of *support to query*. Instead, this study focuses on enhancing the two-branch mutual representation and activating the homologous object regions in a bidirectional manner.

III. Methodology

A. Task Description

Few-Shot Segmentation (FSS) is the task of segmenting objects in a query image leveraging only a limited number of labeled samples. To be more precise, the model is trained on the training set $\mathcal{D}_{train}$ and its performance is evaluated on the testing set $\mathcal{D}_{test}$. Assuming that $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$ correspond to the sets of categories $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$, respectively. These two sets of categories have no overlapping elements, *i.e.*, $\mathcal{C}_{train} \cap \mathcal{C}_{test}$. We adopt the episodic learning strategy for training as prior studies in FSS [54], [55].

For a K-shot segmentation setting, each episode contains **1)** a support set $\mathcal{S} = \left\{(I_s^k, M_s^k)\right\}_{i=1}^{K}$, where I_s^k is the k-th support image and M_s^k is the corresponding binary mask; and **2)** a query set $\mathcal{Q} = \{(I_q, M_q)\}$, where I_q denotes the query image and M_q represent the corresponding binary mask. The support set $\mathcal{S}$ and the query set $\mathcal{Q}$ belong to the same category space. During the training step, both the query mask M_q and the support masks M_s are utilized, whereas only the support masks M_s are available during testing. Since the learned model is inferred without further optimization for novel categories during testing, it is crucial to train the model to find the co-occurred target regions between support images and the query image.

B. Method Overview

To alleviate intra-class variances and background interference in remote sensing images, we propose the Holistic Mutual Representation Enhancement (HMRE) strategy. To facilitate the description, we illustrate the pipeline of HMRE under the 1-shot scenario in Fig. 2. It contains three constituents (*i.e.*, Dual Activation (DA) module, Global Semantic (GS), and Spatial Dense (SD) mutual enhancement modules) and a Mutual-Fusion Decoder.

To elaborate, we adopt the pretrained backbones (ResNet50 [31] or VGG16 [32]) with shared weights to extract both mid- and high-level features. Given the support mask M_s and the query prior mask produced by the Prior Generation [49], we have the DA whose task is to further activate the support prior mask and pseudo query mask. We then decompose HMRE as GS and SD mutual enhancement modules. GS leverages the pseudo query mask and support mask to squeeze mid-level feature maps into pure prototypes for the two branches, which are used to generate a global semantic mutual enhancement coefficient. After GS, the SD follows, which fuses the high-level information (*i.e.*, prior masks) and prototypes of the opposite branch, generating the spatial dense mutual enhancement coefficients. Afterward, these two types of mutual enhancement coefficients are fed into the Mutual Fusion Decoder to implement the holistic mutual enhancement. Finally, we achieve optimization of the above process through bidirectional prediction.

C. Dual Activation (DA) Module

Following the paradigm of leading FSS techniques, we perform the mask average pooling operation on mid-level features to acquire *support prototypes*. As for high-level features, we utilize the prior generator [49] to acquire a *query prior mask* in a training-free manner to maintain generalization ability. Subsequently, these mid- and high-level features are merged and fed into the segmentation decoder, resulting in the generation of predictive masks. However, such an asymmetric dual-branch structure cannot serve as a prerequisite for mutual enhancement. That is, we need to complete the query prototype and the support prior mask.

1) Activation of the pseudo query mask: Inspired by a common ideology that features of the same object exhibit greater similarity compared to those belonging to different objects of the same category [56], we activate the pseudo mask for the query image. This process is constructed on mid-level features. As in [47], we concatenate the features extracted from *block2* and *block3* in ResNet50 or VGG16. Subsequently, a *1×1 convolution* is adopted to reduce the dimensions of the channels to *256* and formulate mid-level features. Formally, let $F_s \in \mathbb{R}^{C\times H\times W}$ and $F_q \in \mathbb{R}^{C\times H\times W}$ be the mid-level support and query features, where C, H, W are channel, height, and width, respectively.

Given F_s, F_q, and the support mask M_s, we apply the mask average pooling to calculate the support prototype as follows:

$$v_s = \mathcal{A}_{pool}(F_s \odot \mathcal{R}(M_s)) \in \mathbb{R}^C, \tag{1}$$

where $\mathcal{A}_{pool}(\cdot)$ is the average-pooling function, $\odot$ represents Hadamard Product, and $\mathcal{R}(\cdot)$ reshapes the mask M_s to fit the dimension of the support feature F_s. Afterward, we evaluate the cosine similarity between the support prototype v_s and the query feature F_q at each spatial location respectively:

$$A_q^{(x,y)} = \frac{v_s^{\mathrm{T}} \cdot F_q^{(x,y)}}{\|v_s\| \cdot \left\|F_q^{(x,y)}\right\|}, \tag{2}$$

where $A_q \in \mathbb{R}^{H\times W}$ denotes the activation map of the query feature, and (x, y) indexes it's spatial location. In essence, the higher the value of a certain position in the activation image, the more likely it is to be the target region of the query map. Based on this, we use the average value of A_q as a threshold to adaptively generate a pseudo query mask $M_{pse} \in \{0, 1\}^{H\times W}$ as:

$$M_{pse}^{(x,y)} = \mathbb{I}\left[A_q^{(x,y)} > \mathcal{A}_{avg}(A_q)\right], \tag{3}$$

where $\mathcal{A}_{avg}(\cdot)$ stands for the averaging function, and $\mathbb{I}[\cdot]$ is the indicator function.

2) Activation of the support prior mask: Following [49], we convert the pre-trained high-level features (*i.e.*, the output derived from the last layer of the backbone) into a class-agnostic prior mask that simply denotes the rough location of target regions. Formally, suppose $\hat{F}_s \in \mathbb{R}^{C\times H\times W}$ and $\hat{F}_q \in \mathbb{R}^{C\times H\times W}$ are the high-level support and query features, respectively, the prior mask of the query image can be written as:

$$\hat{M}_q = \mathcal{G}(M_s, \hat{F}_s, \hat{F}_q), \tag{4}$$

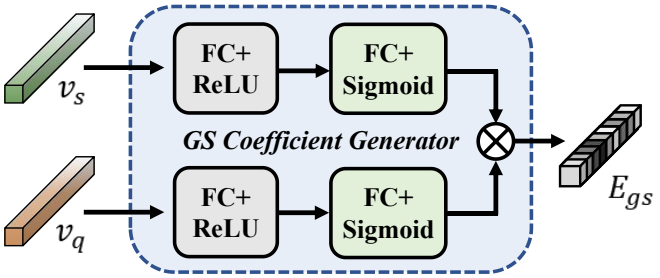


Fig. 3. Illustration of the Global Semantic (GS) Coefficient Generator.

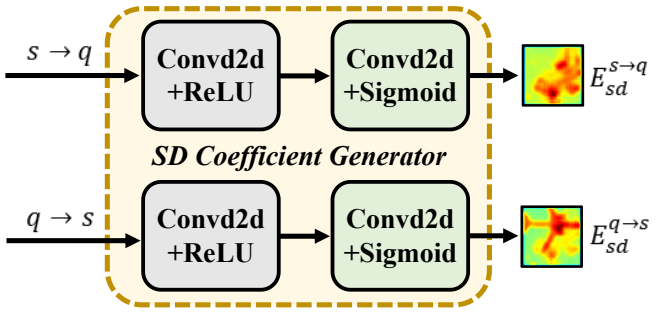


Fig. 4. Illustration of the Spatial Dense Coefficient Generator.

where $\hat{M}_q \in \mathbb{R}^{H\times W}$, and $\mathcal{G}(\cdot,\cdot,\cdot)$ is the Prior Generator. For more details please refer to [49].

To further achieve information symmetry between the two branches, we also need to fill in the support prior mask. Note that although the explicit support mask is available, directly introducing the ground truth of the support sample might not be conducive to optimizing the model from the query branch to the support branch. Therefore, we activate the support prior mask based on $\hat{M}_q$. Specifically, the mask average pooling is conducted on the high-level query feature and its corresponding prior mask:

$$\hat{v}_q = \mathcal{A}_{pool}(\hat{F}_q \odot \hat{M}_q), \tag{5}$$

where $\hat{v}_q \in \mathbb{R}^C$ encodes a proxy of the high-level query feature. Then, we compute the similarity between the high-level support feature $\hat{F}_s$ and the above proxy to produce an activation map:

$$\hat{M}_s^{(x,y)} = \frac{\hat{v}_q^{\mathrm{T}} \cdot \hat{F}_s^{(x,y)}}{\|\hat{v}_q\| \cdot \left\|\hat{F}_s^{(x,y)}\right\|}, \tag{6}$$

where $\hat{M}_s$ merely indicates the probability of pixels belonging to the target category in the support feature. Thus, we similarly regard it as a prior mask of the support image.

D. Global Semantic (GS) Mutual Enhancement Module

In the previous step, we build the information symmetry between the two branches. On this basis, the mutual enhancement of global semantic representation is subsequently performed. The study in [14] we consider most relevant to our work highlights channels with similar global semantics through a cross-reference module. Nonetheless, the utilization of global average pooling may introduce substantial background interference. In cases where common category regions exist in the backgrounds of both the support and query images, this operation might inadvertently lead the model to enhance the global semantics of the non-target channel. To counter this, we opt to use the prototype of the target region as a purer input instead of resorting to the conventional global average pooling method, which simply yet effectively focuses on co-occurrent target regions.

Similarly, we can activate the query prototype leveraging the pseudo query mask generated from the DA module as *Eq.* 1:

$$v_q = \mathcal{A}_{pool}(F_q \odot M_{pse}) \in \mathbb{R}^C. \tag{7}$$

As depicted in Fig. 3, the support and query prototypes are each directed to two-layer fully connected (FC) layers followed by a sigmoid activation function, converting the vector into the importance of each channel, confined within the interval [0,1]. Subsequently, the vectors from the two branches undergo fusion, forming the global semantic mutual enhancement coefficient E_{gs} through element-wise multiplication as:

$$E_{gs} = \phi_s(v_s) \odot \phi_q(v_q) \in \mathbb{R}^C, \tag{8}$$

where ϕ_s and ϕ_q represent the two-layer FC of two branches, respectively. Intuitively, only the channels with similar global semantics of target regions in two branches could exhibit a significant activation in the mutual enhancement coefficient E_{gs}.

E. Spatial Dense (SD) Mutual Enhancement Module

Due to the substantial differences in appearance, size, and distribution of similar target regions within remote sensing images, relying solely on global semantic enhancement is insufficient. In simpler terms, we need further take the spatial dense mutual enhancement into account. Previous related works [57], [13] predominantly center on developing spatial interaction modules for mid-level features yet disregarding complementary information in high-level features. In this study, we integrate mid-level features with the task-agnostic prior masks (*i.e.*, high-level features) generated from the DA module, constructing the spatial dense mutual enhancement between the two branches.

Given the prototypes (*i.e.*, v_s and v_q), we first expand them to fit the size of feature maps, $v_s^e \in \mathbb{R}^{C\times H\times W}$ and $v_q^e \in \mathbb{R}^{C\times H\times W}$. Regarding spatial dense enhancement of the support-to-query, the expanded query prototype v_q^e is combined with the corresponding prior mask $\hat{M}_q$. Then, the enhancement coefficient is refined leveraging a smaller network $\mathcal{F}_{sd}^*$ comprising a 2D convolution layer appended by a sigmoid activation function (see Fig. 4):

$$E_{sd}^{s\rightarrow q} = \mathcal{F}_{sd}^{s\rightarrow q}((v_q^e \odot \hat{M}_q) \oplus v_s^e) \in \mathbb{R}^{H\times W}, \tag{9}$$

where $\oplus$ denotes the element-wise sum. Similarly, the enhancement coefficient of the query-to-support can be computed as:

$$E_{sd}^{q\rightarrow s} = \mathcal{F}_{sd}^{q\rightarrow s}((v_s^e \odot \hat{M}_s) \oplus v_q^e) \in \mathbb{R}^{H\times W}. \tag{10}$$

F. Mutual Fusion Decoder

Based on the two sets of coefficients (*i.e.*, global semantic and spatial dense mutual enhancements) produced from the GS and SD modules, we establish the holistic mutual enhancement to mitigate the issue of intra-class variances and background interference in a task-agnostic manner.

Concretely, both the prototypes and mid-level features are weighted by the two types of mutual enhancement coefficients, E_{gs}, $E_{sd}^{q\to s}$, and $E_{sd}^{s\to q}$:

$$\begin{cases} \tilde{v}_s = v_s \odot E_{gs} \in \mathbb{R}^C \\ \tilde{v}_q = v_q \odot E_{gs} \in \mathbb{R}^C \\ \tilde{F}_s = F_s \odot E_{gs} \odot E_{sd}^{q\to s} \in \mathbb{R}^{C\times H\times W} \\ \tilde{F}_q = F_q \odot E_{gs} \odot E_{sd}^{s\to q} \in \mathbb{R}^{C\times H\times W} \end{cases}. \tag{11}$$

Then, the mutually enhanced mid-level features, $\tilde{F}_s$, $\tilde{F}_q$, prototypes, $\tilde{v}_s$, $\tilde{v}_q$, and prior masks, $\hat{M}_s$, $\hat{M}_q$ are all reshaped to the same spatial size and concatenated to construct support and query merged features, respectively:

$$\begin{cases} y_s^{merg} = \mathcal{C}(\tilde{F}_s, \mathcal{R}(\tilde{v}_q), \hat{M}_s) \in \mathbb{R}^{(2\times C+1)\times H\times W} \\ y_q^{merg} = \mathcal{C}(\tilde{F}_q, \mathcal{R}(\tilde{v}_s), \hat{M}_q) \in \mathbb{R}^{(2\times C+1)\times H\times W} \end{cases}, \tag{12}$$

where $\mathcal{C}(\cdot,\cdot,\cdot)$ represents concatenating the channel dimension and $\mathcal{R}(\cdot)$ is the reshape function. Finally, the merged support feature y_s^{merg} and merged query feature y_q^{merg} are separately fed into two decoders to generate segmentation prediction for two branches:

$$\begin{cases} M_s^{pre} = \mathcal{D}_s(y_s^{merg}) \in \mathbb{R}^{2\times H\times W} \\ M_q^{pre} = \mathcal{D}_q(y_q^{merg}) \in \mathbb{R}^{2\times H\times W} \end{cases}, \tag{13}$$

where the two decoders $\mathcal{D}_s(\cdot)$ and $\mathcal{D}_q(\cdot)$ possess identical structures but have different parameters, which apply a convolution block followed by a classification head as in [49] to obtain the final prediction.

To guarantee the quality of holistic mutual representation enhancement, we introduce a bidirectional optimization paradigm to train the model. Given predicted mask M_q^{pre} and ground-truth query mask M_q, we adopt the BCE loss as our main loss:

$$\mathcal{L}_{seg}^{s\to q} = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} BCE(M_q^{pre}(i,j), M_q(i,j)), \tag{14}$$

where h and w denote the height and width. Then we get the symmetric loss in the same paradigm:

$$\mathcal{L}_{seg}^{q\to s} = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} BCE(M_s^{pre}(i,j), M_s(i,j)). \tag{15}$$

In summary, the total loss is constructed as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg}^{s\to q} + \mu\mathcal{L}_{seg}^{q\to s}, \tag{16}$$

where μ balances the contribution of each branch.

G. Extension to K-shot setting

In this case, there are K support images in each episode. Each one in the K annotated images contributes to segmenting the query image. Built upon this principle, we make some adjustments to certain steps of the model to fully utilize richer support representation. Note that our model does not need to be retrained under this setting. In other words, we leverage the trained model in the 1-shot setting to perform support-to-query prediction during the inference stage.

During the Dual Activation (DA) phase, we yield K support prototypes with *Eq*. 1. Then, the cosine similarity between the query feature and the average of K support prototypes can be used to generate the activation map A_q. In addition, the K query prior masks generated from K support images are averaged to produce a more reliable query prior mask.

For the step of Global Semantic (GS) mutual enhancement, the query branch undergoes global semantic mutual enhancement with K support samples, respectively, resulting in K corresponding mutual enhancement coefficients, where the k-th one can be similar to *Eq*. 8:

$$E_{gs}^k = \phi_s^k(v_s^k) \odot \phi_q(v_q) \in \mathbb{R}^C. \tag{17}$$

In the Spatial Dense (SD) Mutual Enhancement part, the enhancement coefficient of the support-to-query can be modified as:

$$E_{sd}^{s\to q} = \mathcal{F}_{sd}((v_q^e \odot \hat{M}_q) \oplus \frac{1}{K}\sum_{k=1}^{K} v_s^{e|k}) \in \mathbb{R}^{H\times W}. \tag{18}$$

To further extend the Mutual Fusion Decoder, both the prototypes and mid-level features are enhanced as follows:

$$\begin{cases} \tilde{v}_s^k = v_s^k \odot E_{gs}^k \in \mathbb{R}^C \\ \tilde{v}_q = v_q \odot \frac{1}{K}\sum_{k=1}^{K} E_{gs}^k \in \mathbb{R}^C \\ \tilde{F}_q = F_q \odot \frac{1}{K}\sum_{k=1}^{K} E_{gs}^k \odot E_{sd}^{s\to q} \in \mathbb{R}^{C\times H\times W} \end{cases}. \tag{19}$$

Correspondingly, the merged query feature is subsequently denoted as:

$$y_q^{merg} = \mathcal{C}(\tilde{F}_q, \mathcal{R}(\frac{1}{K}\sum_{k=1}^{K}\tilde{v}_s^k), \hat{M}_q) \in \mathbb{R}^{(2\times C+1)\times H\times W}. \tag{20}$$

Finally, the merged query feature y_q^{merg} is fed into the decoder $\mathcal{D}_q$ to yield the predicted mask.

IV. EXPERIMENTS

We first present the public dataset iSAID-5i, followed by the introduction of the newly created dataset Flood-3i. Next, the implementation details and the evaluation metric are demonstrated. Then, we report the performance of the HMRE approach and other advanced methods. Finally, ablation studies are conducted to show the impact of each component.

A. Datasets

1) iSAID-5i Dataset: We do extensive experiments on a public benchmark, named iSAID-5i [15], which originates from iSAID dataset [58]. It contains 18076 samples for training and 6363 samples for testing. Each image is uniformly resized to a size of 256×256. Following the setting in [15], the distribution of class labels and quantities are shown in Table I.

Fig. 5. Illustration of samples and corresponding masks in Flood-3i.

TABLE I
NUMBER OF IMAGES PER CATEGORY FOR THE iSAID-5i DATASET.

Split ID	Category	Num of imgs in train set	Num of imgs in test set
Split0	ship	3820	1392
	storage tank	902	269
	baseball diamond	495	221
	tennis court	2485	767
	basketball court	598	160
Split1	ground track field	1405	517
	bridge	224	91
	large vehicle	2953	789
	small vehicle	2592	781
	helicopter	69	21
Split2	swimming pool	147	45
	roundabout	203	45
	soccer ball field	2000	682
	plane	1864	990
	harbor	3358	1361

2) Flood-3i Dataset: Presently, there is a scarcity of publicly available datasets for FSS research in remote sensing imagery, and they still possess certain limitations. The Dense Labeling Remote Sensing Dataset (DLRSD) [16] only involves 2100 RGB images, which diminishes the persuasiveness of algorithm evaluation. Despite the iSAID-5i dataset containing a substantial number of samples, each image has a limited variety of background classes. Instead, the FloodNet dataset [17], collected using unmanned aerial systems (UAS) for flood monitoring, offers 2343 high-resolution samples and encompasses 9 object categories, presenting a compelling choice for conducting research in FSS tasks. Drawing upon this premise, we follow the setting in iSAID-5i [15] to create the Flood-3i dataset. Specifically, we uniformly crop the raw images and masks to a size of 256×256 to extend the available data. Ultimately, the complete Flood-3i dataset is created, consisting of 13005 samples for training and 4050 samples for testing. Several examples from this dataset are illustrated in Fig. 5. For the 9 classes in the Flood-3i dataset, models are evaluated using the cross-validation method by selecting three classes as test categories denoted as $\mathcal{D}_{test}$, while utilizing the remaining six classes as training categories designated as $\mathcal{D}_{train}$. The distribution of class splits is presented in Table II.

TABLE II
NUMBER OF IMAGES PER CATEGORY FOR THE FLOOD-3i DATASET.

Split ID	Category	Num of imgs in train set	Num of imgs in test set
Split0	building flooded	869	304
	building non-flooded	1689	460
	road flooded	987	308
Split1	road non-flooded	2747	915
	water	3133	940
	tree	6457	2027
Split2	vehicle	171	46
	pool	277	76
	grass	5213	1480

B. Implement Details

1) Experimental Setting: The implementation of our model is done in PyTorch [59] and it runs on an NVIDIA GeForce RTX 3090 GPU. The ResNet50 [31] or VGG16 [32] pretrained in ImageNet [60] are adopted as the encoder to extract features with freezing parameters. Images are sized to 256 × 256 to be forwarded through the FSS model. The proposed model is trained in an episodic scheme for 80 epochs, with a batch size of 16. The stochastic gradient descent (SGD) optimizer is employed in the training phase, and the initial learning rate is set to 0.0015. Additionally, a weight decay of 0.0001 is applied, and the learning rate is adjusted using the "poly" strategy with a power value of 0.9. The balance weight μ in the total loss is empirically set to 0.8 for optimal performance.

2) Evaluation Metrics: We utilize the mean intersection over union (mIoU) as the evaluation metric. We define mIoU as the average over the set of all categories, *i.e.*, $\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\text{IoU}_i$, where N is the number of categories in each testing fold.

C. Comparison with State-of-the-Arts

We conduct a comparative analysis between HMRE and other advanced methods with the available source codes on the iSAID-5i and Flood-3i datasets. The mIoU performance of these advanced technologies including PANet [12], CANet [47], SD-AANet [50], PFENet [49], SDM [15], and SCL [52] are reported in Table III–VI.

TABLE III
COMPARISON WITH STATE-OF-THE-ARTS ON ISAID-5I DATASET IN MIOU UNDER THE ONE SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Backbone	Method	Split0						Split1						Split2						Mean
		C1	C2	C3	C4	C5	mean	C6	C7	C8	C9	C10	mean	C11	C12	C13	C14	C15	mean	
ResNet50	PANet	7.27	9.99	14.49	13.2	16.85	12.36	17.40	10.23	7.37	6.21	4.33	9.11	8.35	12.38	21.28	8.52	9.75	12.05	11.17
	CANet	7.25	27.25	43.32	6.17	9.99	18.80	6.16	17.81	28.58	8.84	16.73	15.62	17.73	56.97	13.82	27.41	13.06	25.79	20.07
	SD-AANet	25.21	39.85	43.60	21.69	22.27	30.52	35.18	47.28	23.89	25.47	14.75	29.31	9.46	27.38	20.51	12.50	22.45	18.46	26.10
	PFENet	2.98	6.30	33.29	34.49	16.71	18.75	8.68	8.67	34.62	18.19	16.08	17.24	28.58	24.47	7.37	39.76	10.31	22.09	19.36
	SDM	37.66	34.37	34.45	39.81	25.14	34.29	16.77	34.53	30.50	12.42	17.02	22.25	20.69	56.83	42.80	40.52	17.26	35.62	30.72
	HMRE(ours)	41.68	41.15	40.01	41.04	40.52	40.88	27.36	34.66	35.44	39.34	27.58	32.88	24.56	49.34	46.82	44.27	21.11	37.22	36.99
VGG16	PANet	13.85	17.78	19.78	17.8	17.94	17.43	22.41	11.75	11.75	6.36	4.88	11.43	12.96	19.11	23.49	11.87	12.33	15.95	14.94
	CANet	7.51	12.87	31.22	30.45	16.61	19.73	18.31	31.94	23.15	3.77	12.55	17.98	33.99	51.41	36.69	17.61	14.91	30.93	22.88
	SD-AANet	18.72	32.01	40.17	23.59	25.58	28.01	13.79	48.74	13.67	14.93	15.79	21.38	8.36	22.03	18.86	11.05	23.66	16.79	22.06
	PFENet	3.94	4.70	30.95	29.21	14.59	16.68	15.01	14.63	27.68	10.13	9.06	15.30	29.40	47.89	25.23	25.20	11.63	27.87	19.95
	SDM	26.55	30.57	33.01	35.06	21.03	29.24	20.54	28.89	32.45	9.62	12.59	20.80	30.79	43.52	49.47	27.51	22.37	34.73	28.26
	HMRE(ours)	30.25	36.16	41.88	43.37	30.74	36.48	31.52	36.66	35.10	18.38	19.74	28.28	29.74	45.09	49.22	30.82	21.93	35.36	33.37

TABLE IV
COMPARISON WITH STATE-OF-THE-ARTS ON ISAID-5I DATASET IN MIOU UNDER THE FIVE SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Backbone	Method	Split0						Split1						Split2						Mean
		C1	C2	C3	C4	C5	mean	C6	C7	C8	C9	C10	mean	C11	C12	C13	C14	C15	mean	
ResNet50	PANet	8.45	11.68	17.81	14.22	16.93	13.82	18.86	15.70	9.40	7.90	10.13	12.40	21.41	24.94	28.41	9.80	11.04	19.12	15.11
	CANet	32.47	6.49	39.8	25.97	14.59	23.86	25.01	25.27	32.06	1.31	9.08	18.54	26.99	73.84	33.83	20.69	4.69	32.00	24.8
	SD-AANet	40.53	37.08	46.10	29.67	37.31	38.14	38.48	49.39	26.21	30.27	16.85	32.24	9.84	28.05	21.31	19.79	22.54	20.31	30.23
	PFENet	10.13	9.48	30.71	31.23	16.31	19.57	11.53	14.15	36.07	15.58	14.82	18.43	38.52	20.90	20.41	38.66	12.22	26.14	21.38
	SDM	38.76	49.06	50.06	39.25	22.26	39.88	30.68	45.34	41.49	20.21	15.21	30.59	32.61	66.64	57.41	49.12	22.71	45.70	38.72
	HMRE(ours)	44.94	41.98	40.46	41.89	42.78	42.41	30.16	35.17	39.07	38.61	30.44	34.69	39.57	52.84	54.19	50.63	33.12	46.07	41.06
VGG16	PANet	13.40	17.79	19.04	18.11	20.15	17.70	22.97	16.18	15.65	9.37	8.73	14.58	31.65	21.63	25.36	12.70	12.15	20.70	17.66
	CANet	11.87	18.36	33.26	26.05	27.69	23.45	19.20	29.93	30.37	11.76	11.37	20.53	23.76	60.25	28.67	21.77	16.16	30.12	24.7
	SD-AANet	29.23	33.54	42.75	32.03	36.88	34.89	23.02	51.57	19.20	16.40	18.62	25.76	10.71	17.03	18.28	13.43	24.62	16.81	25.82
	PFENet	2.64	8.57	29.09	30.76	21.26	18.46	15.22	21.58	31.16	12.06	11.90	18.39	33.56	48.47	24.92	25.29	11.79	28.81	21.89
	SDM	30.81	38.59	37.34	42.58	32.35	36.33	32.32	39.71	37.23	17.73	12.93	27.98	39.20	66.78	39.06	46.17	20.72	42.39	35.57
	HMRE(ours)	31.74	36.18	42.79	42.59	33.34	37.33	33.02	37.51	35.63	20.09	20.14	29.28	42.72	47.15	47.22	45.93	33.78	43.36	36.66

1) Results of mIoU on iSAID-5i: Table III and IV show the 1-shot and 5-shot results on iSAID-5i. It is evident that our HMRE significantly surpasses other methods. Specifically, utilizing the ResNet50 backbone, the proposed method achieves a remarkable mean mIoU improvement of 6.27% in the 1-shot setting and 2.34% in the 5-shot setting compared to SDM, which is the second-best performing method. Simultaneously, with the VGG16 backbone, our 1-shot and 5-shot results respectively surpass the best competitor, *i.e.*, SDM by 5.11% and 1.09% mean mIoU. For each split, the proposed model consistently attains the best performance and demonstrates a more significant advantage in scenarios with a very restricted number of samples (*e.g.*, 1-shot setting).

For a more thorough evaluation of performance across various categories, we also provide detailed results. It is clear that our proposed method excels in performance across almost all classes. Additionally, our method significantly boosts the segmentation performance of some challenging categories. In Table III, with the ResNet50 backbone, our model gets 15.38% mIoU improvement in the basketball court category (C5). With the VGG16 backbone, an enhancement of 9.11% is achieved over the second-best approach in the category of ground track field (C6). Under the 5-shot setting (see Table IV), with the ResNet50 backbone, HMRE attains the optimal improvements, namely, 13.59%, in the helicopter category (C10). When employing the VGG16 backbone, the peak segmentation performance (progress of 9.16%) is obtained in the harbor category (C15). Meanwhile, the proposed HMRE exhibits suboptimal performance in the bridge class (C7). Upon thorough examination of the samples within this category, we have discerned that the uniform background and evenly distributed spatial elements might not be conducive to highlighting the advantages of our method.

2) Results of mIoU on Flood-3i: In Table V and VI, we conduct a comparative analysis of our approach with other methods on the Flood-3i dataset. The HMRE gives a comprehensive optimal performance. For example, our approach outperforms the second-best methods (*i.e.*, PFENet and SCL) by 5.02% and 6.22% mean mIoU under the 1-shot and 5-shot settings while employing the same backbone of ResNet50. With the VGG16 backbone, the proposed method obtains a mean mIoU 4.80% higher in the 1-shot setting, and 3.25% in the 5-shot setting. Besides, HMRE yields the best effectiveness in each split, which demonstrates its universality.

As for specific categories, our approach similarly shows its superiority. Particularly, although some methods (*e.g.*, SD-AANet in split0 and split1, SCL in split0) excel in one category, they may experience a significant decline in other categories. In the aforementioned scenario, our approach ex-

TABLE V
COMPARISON WITH STATE-OF-THE-ARTS ON FLOOD-3I DATASET IN MIOU UNDER THE ONE SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Backbone	Method	Split0				Split1				Split2				Mean
		C1	C2	C3	mean	C4	C5	C6	mean	C7	C8	C9	mean	
ResNet50	PANet	19.87	25.46	17.31	20.88	19.81	18.78	19.45	19.35	11.33	9.40	4.57	8.42	16.22
	CANet	22.89	17.54	19.82	20.08	25.29	18.62	14.05	19.32	9.22	6.86	6.28	7.45	15.62
	SD-AANet	23.70	52.93	9.34	28.66	46.79	5.32	30.09	27.4	15.87	4.62	4.83	8.44	21.5
	PFENet	44.42	47.20	30.87	40.83	44.52	38.52	29.88	37.64	6.40	14.17	2.65	7.74	28.74
	SCL	33.02	45.21	30.83	36.35	43.42	25.31	41.27	36.67	13.34	13.35	5.16	10.62	27.88
	HMRE(ours)	47.22	48.24	33.72	43.06	44.83	41.14	34.06	40.01	21.08	18.91	14.64	18.21	33.76
VGG16	PANet	16.02	27.12	12.38	18.51	17.30	14.49	21.48	17.76	12.73	6.25	4.37	7.78	14.68
	CANet	18.26	19.02	16.51	17.93	29.74	17.03	15.71	20.83	8.41	5.27	3.24	5.64	14.80
	SD-AANet	22.79	51.86	10.56	28.40	46.31	11.74	33.84	30.63	16.06	2.47	6.18	8.24	22.42
	PFENet	29.95	50.02	14.11	31.36	41.51	47.17	24.07	37.58	7.35	9.97	8.15	8.49	25.81
	SCL	31.32	57.17	10.37	32.95	46.37	27.14	34.61	36.04	11.31	12.51	4.95	9.59	26.19
	HMRE(ours)	30.40	37.80	37.05	35.08	41.86	47.94	29.21	39.67	20.73	18.32	15.59	18.21	30.99

TABLE VI
COMPARISON WITH STATE-OF-THE-ARTS ON FLOOD-3I DATASET IN MIOU UNDER THE FIVE SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.N

Backbone	Method	Split0				Split1				Split2				Mean
		C1	C2	C3	mean	C4	C5	C6	mean	C7	C8	C9	mean	
ResNet50	PANet	25.59	28.77	19.43	24.60	24.29	19.90	23.46	22.55	12.26	9.52	8.93	10.24	19.13
	CANet	21.36	19.60	24.75	21.90	33.29	19.54	15.18	22.67	8.75	6.93	7.14	7.61	17.39
	SD-AANet	27.25	51.53	6.51	28.43	56.77	6.84	34.80	32.80	17.11	3.73	4.98	8.61	23.28
	PFENet	49.21	46.82	31.54	42.52	48.03	35.71	24.52	36.09	8.21	16.78	1.34	8.78	29.13
	SCL	35.48	46.64	32.95	38.36	44.12	25.04	44.85	38.00	15.73	12.06	7.21	11.67	29.34
	HMRE(ours)	50.61	49.68	36.14	45.48	45.79	41.95	36.66	41.47	22.67	18.96	17.55	19.73	35.56
VGG16	PANet	19.07	27.44	17.23	21.25	28.22	16.89	19.8	21.64	11.86	7.42	6.50	8.59	17.16
	CANet	19.25	21.83	17.30	19.46	36.74	15.43	14.97	22.38	11.40	7.73	8.69	9.27	17.04
	SD-AANet	16.75	53.77	4.92	25.15	54.15	12.85	40.19	35.73	16.71	2.40	5.75	8.29	23.06
	PFENet	33.28	51.82	7.28	30.79	47.91	35.71	24.52	36.05	5.77	8.68	7.29	7.25	24.70
	SCL	35.24	52.83	9.23	32.43	45.51	19.66	44.77	36.65	13.22	8.00	4.31	8.51	25.86
	HMRE(ours)	30.25	36.76	35.59	34.20	39.27	45.19	31.68	38.71	16.92	14.89	11.43	14.41	29.11

hibits superior performance while possessing a smaller degree of performance variance. In addition, for the more challenging category fold split2 characterized by an imbalanced number of samples, the proposed approach reaps tremendous performance improvements. We believe that this can be attributed to the successful implementation of the holistic mutual representation enhancement, which enables the model to mine valuable information from support images.

3) Segmentation Examples: To facilitate more in-depth analysis and comprehension of the proposed model, we showcase several segmentation examples by the baseline (*i.e.*, PFENet [49]) method and our HMRE approach, as shown in Fig. 6. It can be seen that HMRE presents significantly more accurate prediction masks. For instance, when handling the minuscule masked objects in the third column (belonging to the "vehicle" category), our approach can still generate satisfactory results for the query image.

D. Ablation Studies

To examine the effect of each factor on segmentation accomplishment, a series of ablation studies are conducted with ResNet50 backbone and we set PFENet [49] as the baseline.

1) Components Analysis: The proposed HMRE contains two major components, *i.e.*, Global Semantic (GS) and Spatial Dense (SD) mutual enhancement modules. Note that we additionally incorporated a global semantic mutual enhancement strategy in CRNet [14] to probe the superiority of the devised GS module, which this paper refers to as GS-CR. Under the Flood-3i dataset, Table VII showcases the effectiveness of each component. Compared with the baseline, the utilization of GS, GS-CR, and SD modules all contribute to a substantial improvement in the segmentation consequence. By combining the GS and SD modules, HMRE achieves the best performance. Moreover, the GS module exhibits a greater performance gain for the model compared to the GS-CR strategy. This is attributed to its effective suppression of background interference in remote sensing images.

2) Effect of the Mutual Fusion Schemes: As described in Section III-F, by leveraging the two types of mutual enhancement coefficients, we propose to establish mutual fusion by weighting the prototypes and mid-level features. To study the influence of different mutual fusion schemes, we qualitatively visualize the activation maps, as illustrated in Fig. 7. From left to right, each column represents: (a) support images with annotation masks, (b) query images with the Ground Truth (GT), (c) baseline results, (d) the results of weighting

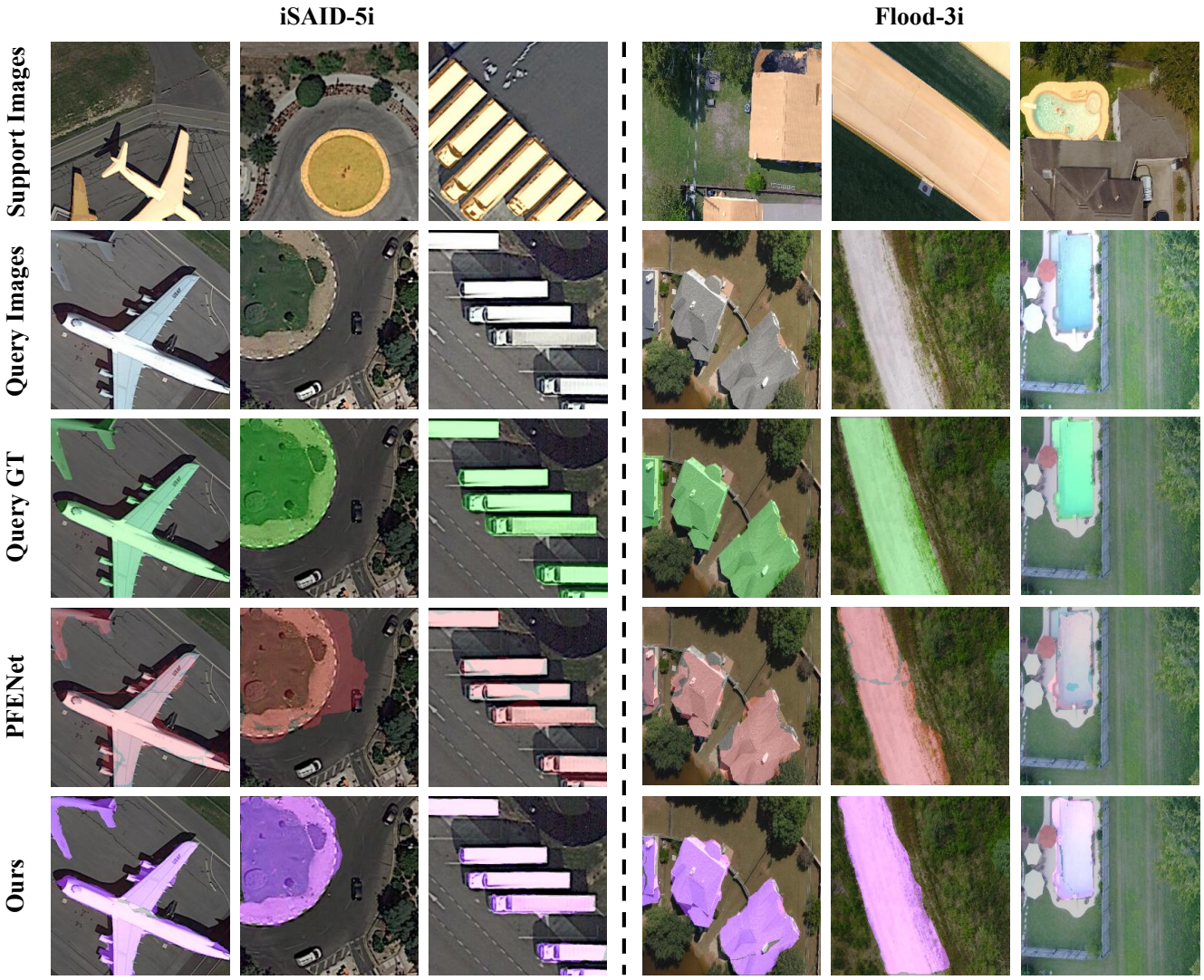


Fig. 6. Segmentation examples of HMRE and baseline method (PFENet) under the 1-shot setting. The left panel is from iSAID-5i, and the right one is from Flood-3i. Images within the same column correspond to a certain object category. The specific target regions include planes, roundabouts, vehicles, building non-flooded, road non-flooded, and pools.

prototypes, (e) the results of weighting mid-level features, (f) the results of both weighting prototypes and mid-level features (ours).

It can be observed that the baseline method erroneously activates certain non-target regions, whereas the GS module effectively suppresses non-target category interference within the background. Furthermore, the SD module is beneficial for mining more edge and detail features, particularly in cases where the spatial distribution of the targets is densely concentrated. When incorporating both the SD and GS modules, the model enables more accurate activation of the target regions within the query samples.

3) Effect of the Hyperparameter μ for Balancing the Bidirectional Optimization: We undertake an analysis to explore the influence of the hyperparameter μ, which adjusts the contribution of optimization in two directions. Concretely, the experiments are conducted on iSAID-5i and Flood-3i datasets with ResNet50 backbone. We regulate μ in the interval $[0, 2]$, as shown in Fig. 8. When μ is set as 0, it is equivalent to the model undergoing only unidirectional optimization from support to query. This leads to the loss of effective supervision for

TABLE VII
ABLATION STUDIES OF MAJOR MODEL COMPONENTS USING MIOU (%) ON FLOOD-3I DATASET.

Shot	Baseline	GS-CR	GS	SD	Split0	Split1	Split2	Mean
1	✔				40.83	37.64	7.74	28.74
1	✔	✔			41.16	38.19	10.90	30.08
1	✔		✔		42.27	39.33	14.46	32.02
1	✔			✔	41.25	38.89	15.03	31.72
1	✔	✔		✔	42.34	39.48	16.73	32.85
1	✔		✔	✔	**43.06**	**40.01**	**18.21**	**33.76**
5	✔				42.52	36.09	8.78	29.13
5	✔	✔			43.66	38.39	12.82	31.62
5	✔		✔		44.52	40.80	16.29	33.87
5	✔			✔	44.30	38.59	15.70	32.86
5	✔	✔		✔	45.32	39.87	17.84	34.34
5	✔		✔	✔	**45.48**	**41.47**	**19.73**	**35.56**

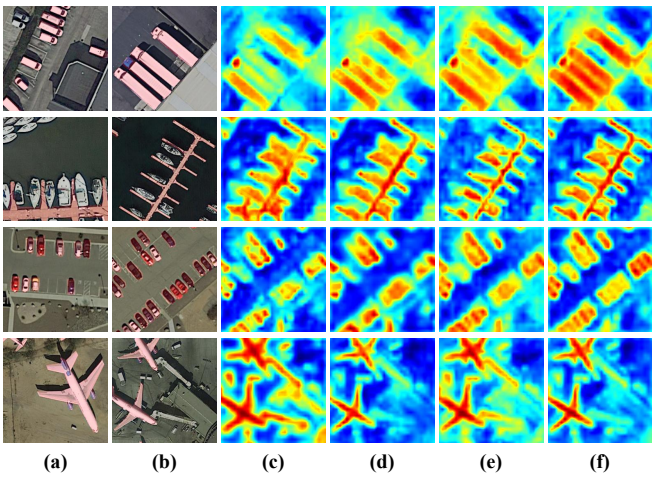


Fig. 7. Activation maps with different mutual fusion schemes under the 1-shot setting. The target classes of the top to bottom rows are "large vehicle", "harbor", "small vehicle", and "plane", respectively.

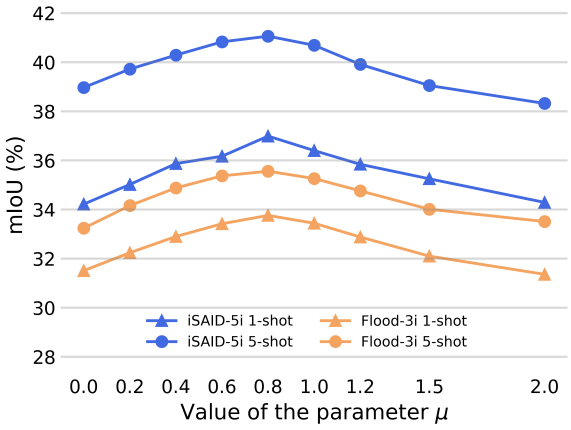


Fig. 8. Impact of the parameter μ for the few-shot segmentation mIoU.

mutual representation enhancement between the two branches, thereby resulting in a decline in model performance. As μ increases, the segmentation accuracy continuously improves due to the gradually strengthening supervision. For the two datasets and all settings, HMRE achieves the peak performance when μ is set to 0.8. Nevertheless, further increasing μ results in a decrease in the contribution of support-to-query direction optimization, reducing the segmentation accuracy of the model for query samples.

E. Extended Experiments

1) Model Discriminability: We randomly select 1000 episodes from 9 categories of the Flood-3i dataset to create a confusion matrix, as depicted in Figure 9. It is evident that the baseline method demonstrates lower discriminability among regions of similar categories (*e.g.*, c1 and c2, c3 and c4), making it susceptible to erroneous predictions. In comparison, the proposed HMRE effectively mitigates semantic confusion and demonstrates superior segmentation performance.

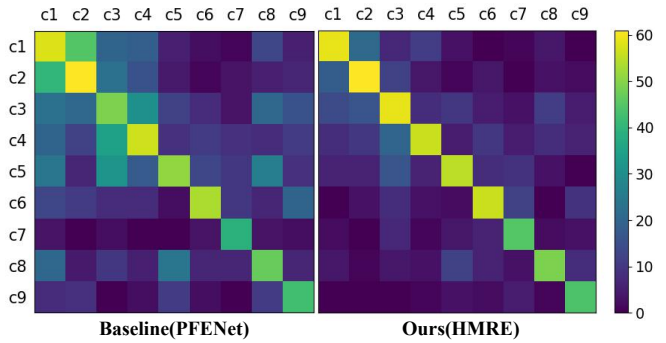


Fig. 9. Confusion matrix for the target categories on Flood-3i.

TABLE VIII
EXTENDED EXPERIMENT OF CROSS-SET FSS ON ISAID-5I → FLOOD-3I UNDER THE ONE SHOT AND FIVE SHOT SETTINGS. ALL DATA REPRESENTS THE MIOU RESULTS OBTAINED FROM 2000 EPISODES.

Backbone	Method	iSAID-5i → Flood-3i	
		1-shot	5-shot
ResNet50	Baseline (PFENet)	29.97	31.93
	Ours (HMRE)	34.63	36.29
VGG16	Baseline (PFENet)	22.77	25.71
	Ours (HMRE)	30.18	33.24

2) Cross-set FSS: Considering a scenario that is more common and challenging in practical applications, we further perform the Cross-set FSS experiment. We utilize split1 and split2 from the iSAID5-i training set as the base (seen) data, and split0 from the Flood-3i test set as the novel (unseen) data, refer to Tables I and II for the distribution of specific target categories. The results presented in Table VIII clearly demonstrate that the proposed HMRE approach outperforms the baseline method significantly in all settings. This indicates that our model exhibits superior generalization capabilities and holds promising potential for practical applications.

V. CONCLUSION

This paper proposes the HMRE approach for mitigating the intra-class variances and background interference in FSS, which constructs mutual representation enhancement between two branches. Particularly, the Dual Activation (DA) module establishes information symmetry within the dual-branch structure. Then, we decouple the holistic mutual enhancement as the Global Semantic (GS) and Spatial Dense (SD) mutual enhancement modules. Ablation experiments were organized to qualitatively and quantitatively probe the effectiveness of this strategy. To optimize the process of mutual enhancement, the devised Mutual Fusion Decoder adopts a bidirectional prediction paradigm. Moreover, we curate a new dataset called Flood-3i, which aims to facilitate the research of FSS. The comprehensive experiments conducted indicate the efficacy of our method, establishing it as the new state-of-the-art.

REFERENCES

[1] J. Hu, Q. Wang, and X. Li, "Road extraction from satellite image via auxiliary road location prediction," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2182–2185.

[2] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, p. 631–651, Sep 2012. [Online]. Available: http://dx.doi.org/10.1109/jproc.2012.2211551

[3] S. P. Abercrombie and M. A. Friedl, "Improving the consistency of multitemporal land cover maps using a hidden markov model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 2, pp. 703–713, 2016.

[4] J.-Y. Rau, L.-C. Chen, J.-K. Liu, and T.-H. Wu, "Dynamics monitoring and disaster assessment for watershed management using time-series satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1641–1649, 2007.

[5] Z. Li, Y. Xie, W. Hou, Z. Liu, Z. Bai, J. Hong, Y. Ma, H. Huang, X. Lei, X. Sun, X. Liu, B. Yang, Y. Qiao, J. Zhu, Q. Cong, Y. Zheng, M. Song, P. Zou, Z. Hu, J. Lin, and L. Fan, "In-orbit test of the polarized scanning atmospheric corrector (psac) onboard chinese environmental protection and disaster monitoring satellite constellation hj-2 a/b," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[6] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super resolution for efficient remote sensing salient object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[7] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[8] J. Wang, J. Gao, Y. Yuan, and Q. Wang, "Crowd localization from gaussian mixture scoped knowledge and scoped teacher," *IEEE Transactions on Image Processing*, vol. 32, pp. 1802–1814, 2023.

[9] Q. Wang, J. Wang, J. Gao, Y. Yuan, and X. Li, "Counting like human: Anthropoid crowd counting on modeling the similarity of objects," *arXiv preprint arXiv:2212.02248*, 2022.

[10] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[11] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[12] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Feb 2020. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00929

[13] X. Yang, B. Wang, K. Chen, X. Zhou, S. Yi, W. Ouyang, and L. Zhou, "Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation," Aug 2020.

[14] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2020.

[15] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–11, Oct 2021. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2021.3119852

[16] B. Wang, Z. Wang, X. Sun, H. Wang, and K. Fu, "Dmml-net: Deep metametric learning for few-shot geographic object segmentation in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–18, Oct 2021.

[17] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *arXiv preprint arXiv:2012.02951*, 2020.

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

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[20] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[23] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[25] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3562–3572.

[26] Z. Wang, S. Zhang, C. Zhang, and B. Wang, "Hidden feature-guided semantic segmentation network for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[27] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[28] Y. Cai, L. Fan, and Y. Fang, "Sbss: Stacking-based semantic segmentation framework for very high-resolution remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[29] P. He, L. Jiao, R. Shang, S. Wang, X. Liu, D. Quan, K. Yang, and D. Zhao, "Manet: Multi-scale aware-relation network for semantic segmentation in aerial scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[30] X. Chen, Z. Li, J. Jiang, Z. Han, S. Deng, Z. Li, T. Fang, H. Huo, Q. Li, and M. Liu, "Adaptive effective receptive field convolution for semantic segmentation of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3532–3546, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.

[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[35] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[36] G. Cheng, L. Cai, C. Lang, X. Yao, J. Chen, L. Guo, and J. Han, "Spnet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[37] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1–10.

[38] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International conference on learning representations*, 2017.

[39] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[40] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 719–11 727.

[41] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[42] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "Dla-matchnet for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7844–7853, 2020.

[43] X. Li, D. Shi, X. Diao, and H. Xu, "Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[44] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multi-view for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2023.

[45] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[46] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *CoRR*, vol. abs/1709.03410, 2017.

[47] C. Zhang, G. Lin, F. Liu, R. Yao, and C. 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*, 2019, pp. 5217–5226.

[48] M. Siam and B. N. Oreshkin, "Adaptive masked weight imprinting for few-shot segmentation," *CoRR*, vol. abs/1902.11123, 2019.

[49] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1050–1065, Aug 2020.

[50] Q. Zhao, B. Liu, S. Lyu, and H. Chen, "A self-distillation embedded supervised affinity attention model for few-shot segmentation," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[51] Z. Wu, X. Shi, G. Lin, and J. Cai, "Learning meta-class memory for few-shot semantic segmentation," Jan 2021.

[52] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov 2021.

[53] X. Luo, Z. Tian, T. Zhang, B. Yu, Y. Tang, and J. Jia, "Pfenet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask." Sep 2021.

[54] S. Moon, S. S. Sohn, H. Zhou, S. Yoon, V. Pavlovic, M. H. Khan, and M. Kapadia, "Msi: Maximize support-set information for few-shot segmentation," *arXiv preprint arXiv:2212.04673*, 2022.

[55] K. Huang, M. Cheng, Y. Wang, B. Wang, Y. Xi, F. Wang, and P. Chen, "A joint framework towards class-aware and class-agnostic alignment for few-shot segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 1471–1487.

[56] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," Jul 2022.

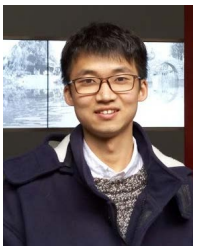
[57] G. Gao, Z. Fang, C. Han, Y. Wei, C. H. Liu, and S. Yan, "Drnet: Double recalibration network for few-shot semantic segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 6733–6746, 2022.

[58] S. Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Khan, Y. Fang, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," May 2019.

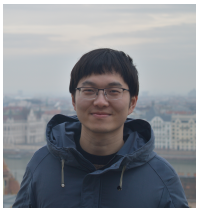
[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Yuyu Jia received the B.E. degree and the M.S. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include few-shot learning, deep learning, and remote sensing.



Junyu Gao received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021, respectively. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Wei Huang received the B.E. degree in control theory and engineering and M.E. degree in computer science and technology in the School of Artificial Intelligence, Optics and Electronics (iOPEN) from the Northwestern Polytechnical University, Xi'an, China, in 2018 and 2021, respectively. He is pursuing his Ph.D. degree at Technical University of Munich. His research interests include transfer learning, deep learning, and remote sensing.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.