Hierarchical Relation Learning for Few-shot Semantic Segmentation in Remote Sensing Images

Xin He, Yun Liu, Yong Zhou, Henghui Ding, *Member, IEEE,* Jiaqi Zhao, *Member, IEEE,*
Bing Liu, and Xudong Jiang, *Fellow, IEEE*

***Abstract*—Few-shot semantic segmentation (FSS) aims to segment specific semantic classes in a query image using only a few annotated support samples. While FSS has gained significant attention in natural image processing, it remains underexplored in the more challenging domain of remote sensing images (RSIs). Existing FSS approaches for RSIs primarily focus on enhancing feature representations of support or query images through hierarchical/multi-level feature fusion. However, unlike fully supervised segmentation that relies on feature extraction and optimization, FSS requires segmenting the query image based on its relations with annotated support images. To address this need, we propose the concept of Hierarchical Relation Learning (HRL) to explore the intrinsic support-query relations, allowing for the direct refinement of target object appearances in the query image. Specifically, we propose a Hierarchical Relation Network (HRNet), which performs single-scale relation extraction at each network hierarchy and multi-scale relation aggregation across hierarchies. In addition, we construct a Bidirectional Hierarchical Loss (BHLoss) to guide HRNet training, providing targeted supervision at each hierarchy in both top-down and bottom-up directions, thus facilitating robust multi-scale relation learning across hierarchies. Comprehensive experiments on the iSAID-5^i, DLRSD-5^i, and LoveDA-2^i datasets demonstrate the superiority of the proposed HRL. The code will be available at https://github.com/XinnHe/HRL.**

***Index Terms*—Few-shot semantic segmentation, few-shot segmentation, remote sensing, hierarchical relation learning.**

I. Introduction

SEMANTIC segmentation in remote sensing images (RSIs) aims to classify each pixel according to the semantic category of ground objects, playing a crucial role in studying environmental changes on the earth's surface and human activity patterns. In recent years, deep learning, powered by large, fine-grained labeled datasets, has significantly advanced this task, with notable success in applications such as land cover classification [1]–[3], environmental monitoring [4], and agricultural management [5], [6]. However, RSIs typically cover vast geographic areas and contain a wide variety of objects and terrains, making it impractical to manually label all semantic categories at scale. To address this challenge, few-shot segmentation (FSS) based on meta-learning was proposed [7], which leverages a few annotated examples as support data to segment novel classes in query images. FSS can significantly reduce annotation costs and enable models to adapt quickly to novel classes or unseen domains. This is particularly important for real-world RSI applications where annotated data is scarce or rapidly changing, such as post-disaster response mapping (*e.g.*, landslides, floods, collapsed buildings), rare object detection (*e.g.*, illegal open-pit mines, illegal buildings or temporary shelters), and cross-regional fine-grained crop type segmentation. In these scenarios, FSS not only enhances learning efficiency, but also improves the generalization ability and robustness of segmentation models under limited supervision.

FSS in natural images has made significant progress, particularly with prototype-based and pixel-wise methods. As shown in Fig. 1(a), the general process of prototype-based FSS methods involves first extracting prototypes from the support samples [8]–[10], which capture representative information about the target semantic classes. These prototypes are then used as prior cues to explore the relation/similarity between the support and query features, typically using Euclidean distance [8], [9], cosine distance [11]–[13], or feature concatenation [14]–[16]. Accurate support-query relation leads to reasonable predictions for the query images. Pixel-level methods (see Fig. 1(b)) take a more direct approach by performing pixel-wise support-query matching through attention mechanisms [17]–[20] or specialized convolutions [21]–[23], effectively preserving critical spatial structural information. In addition, some studies [24], [25] utilize vision transformers to enhance the representation of support prototypes by incorporating query features (see Fig. 1(c)).

Compared to FSS in natural images, research on FSS in RSIs is relatively lagging and insufficient. The primary challenge lies in the prevalent *high inter-class similarity*, *large intra-class heterogeneity*, and *significant scale variations* within RSIs (see Fig. 2). These factors seriously affect

This work was supported in part by the China Scholarship Council (No. 202306420037), in part by the National Natural Science Foundation of China (No. 62272461, 62276266, and 62172417), in part by the "Double First-Class" Project of China University of Mining and Technology for Independent Innovation and Social Service (No. 2022ZZCX06), and in part by the Six Talent Peaks Project in Jiangsu Province (No. 2015-DZXX-010 and 2018-XYDXX-044). *(Corresponding author: Yun Liu and Yong Zhou)*

X. He is with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China, and also with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. (e-mail: xhe@cumt.edu.cn)

Y. Zhou, J. Zhao, and B. Liu are with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China, and also with the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou 221116, China. (e-mail: yzhou@cumt.edu.cn; jiaqizhao@cumt.edu.cn; liubing@cumt.edu.cn)

Y. Liu is with the College of Computer Science, Nankai University, Tianjin 300350, China. (E-mail: liuyun@nankai.edu.cn)

H. Ding is with the Institute of Big Data, Fudan University, Shanghai 200433, China. (e-mail: henghui.ding@gmail.com)

X. Jiang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. (e-mail: exdjiang@ntu.edu.sg)

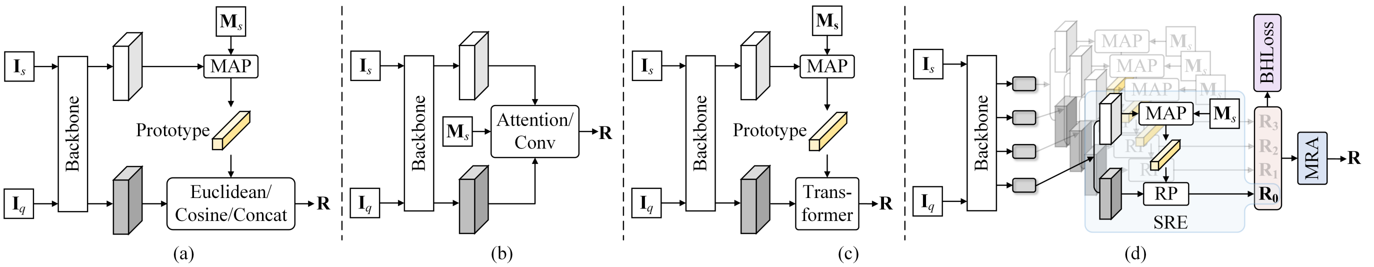

Fig. 1. Our hierarchical relation-based framework (d) in comparison with existing prototype-based (a), pixel-level based (b) and transformer-based (c) frameworks for FSS in RSIs. The inputs consist of the support image $\mathbf{I}_s$, support mask $\mathbf{M}_s$, and query image $\mathbf{I}_q$, with the output being the support-query relation $\mathbf{R}$. Here, MAP denotes masked average pooling [10]–[12]. SRE represents Single-scale Relation Extraction, which generates support-query relations $\mathbf{R}_0$, $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$ at different hierarchies. RP represents Relation Processor, which consists of a Transformer layer and a Convolutional Layer to achieve local-to-global pixel-level relation self-alignment. MRA represents Multi-scale Relation Aggregation, and BHLoss stands for Bidirectional Hierarchical Loss.

the effectiveness of traditional FSS frameworks, weakening the representativeness of prototype features and substantially increasing the difficulty of support-query feature matching under multi-scale conditions. As a result, models often fail to accurately activate the target regions in query images and even produces obvious segmentation errors.

In fully supervised semantic segmentation, to address the above characteristics of RSIs, it is common to rely on *hierarchical/multi-level feature learning* [26]–[32], where both high-level semantic information and low-level fine-grained details. The classical U-Net [33] and its variants [34]–[36] are representative approaches. They integrate semantic and fine-grained features across multiple hierarchies/levels in a bottom-up manner, significantly enhancing the feature representation capability of RSIs [37]–[46].

Inspired by this, recent studies have incorporated multi-level feature aggregation into FSS [47]–[50]. These methods typically adopt a cross-attention mechanism to enhance the perception of query features to target regions that are consistent with the objects in the support image, by modeling the similarity between support and query features and incorporating the known spatial cues from the support mask. Then, query features guided by support information at different scales are fused from bottom to top in the decoder. It is important to note that FSS differs fundamentally from fully supervised segmentation. While fully supervised methods aim to extract semantic representations from a single image, FSS relies on modeling the relation between the query and support images to identify regions corresponding to novel object classes. Therefore, the multi-level feature aggregation paradigm commonly used in fully supervised segmentation [41]–[46] may not be sufficient for FSS in RSIs. On the one hand, it is prone to the accumulation and propagation of erroneous support-query correlation clues within query features through layers, particularly when low-level features are rich in detail but lack semantic clarity. On the other hand, integrating deep (high-level) features may impair the generalization ability of the model, especially in few-shot tasks involving cross-region or cross-class segmentation [10], [15]. More importantly, these methods primarily focus on enhancing feature representations, rather than explicitly modeling or refining the support-query relations and the shape of target objects, which is particularly insufficient in complex RSI scenarios.

To address the above issues, we propose **Hierarchical Relation Learning (HRL)** for the FSS task in RSIs. HRL aims to directly refine the shape of target objects in query images by explicitly modeling and integrating hierarchical support-query relations. The framework is shown in Fig. 1(d). Specifically, we construct a **Hierarchical Relation Network (HRNet)** using the mid-layer features with generalization capability of the backbone. HRNet contains four independent Single-scale Relation Extraction (SRE) modules and a Multi-scale Relation Aggregation (MRA) module. For each network hierarchy, SRE receives prototypes, query features, and training-free priors to capture single-scale support-query relations via the Relation Processor (RP), which consists of a deformable transformer layer and a convolutional layer. Then, MRA aggregates these multi-scale relations across hierarchies in a parallel manner to avoid the accumulation of erroneous relations, and leverages RP to effectively mine discriminative relation cues to enhance relation reliability. Furthermore, we introduce a **Bidirectional Hierarchical Loss (BHLoss)** to support the training of HRNet by providing appropriate supervision at each network hierarchy in both top-down and bottom-up directions. BHLoss ensures that HRNet learns specific scale information at each hierarchy while also capturing complementary multi-scale information across multiple hierarchies, thereby enhancing the diversity and robustness of hierarchical relation representations.

In summary, our contributions are as follows:

- We propose Hierarchical Relation Learning (HRL) to explore hierarchical support-query relations for FSS in RSIs, moving beyond traditional feature optimization to directly refine target object shapes in query images.
- We construct a Hierarchical Relation Network (HRNet) for HRL, enabling single-scale support-query relation extraction and multi-scale relation integration across hierarchical levels.
- We introduce a Bidirectional Hierarchical Loss to support HRNet training by enforcing supervision on the support-query relations at each hierarchy, enhancing the diversity and robustness of hierarchical relation representations.

Extensive experiments on iSAID-5^i [51], DLRSD-5^i [52], and LoveDA-2^i [53] datasets demonstrate that the proposed HRL achieves state-of-the-art performance for FSS in RSIs, offering a new perspective for this field.

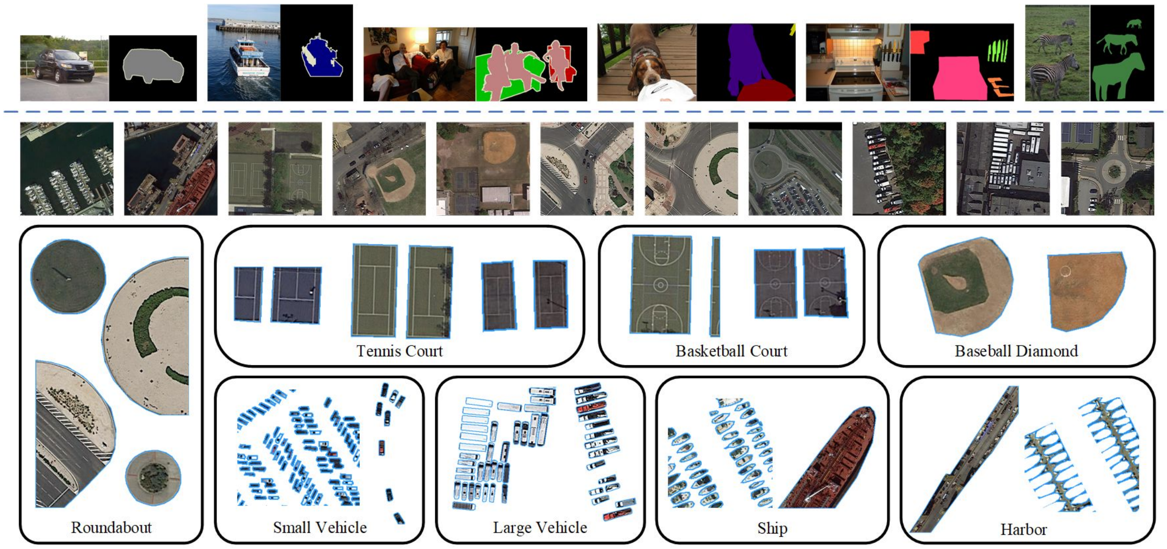

Fig. 2. Comparison between natural image segmentation and remote sensing image segmentation. Above the dashed line are natural image samples from PASCAL VOC [54] and MS-COCO [55] datasets, and below the dotted line are remote sensing image samples from the iSAID dataset [51]. Natural images are typically captured based on subjective human preferences, with each image centered around a distinct *visual subject*, such as a specific object, person, or scene that dominates the composition. This clear visual structure allows traditional FSS methods to extract prototypes and perform feature matching more directly. In contrast, RSIs are acquired indiscriminately from overhead perspectives, without a clear *visual subject*. They encompass a wide range of ground objects, resulting in more complex spatial layouts and semantic ambiguity. Within these remote sensing classes, "Tennis Court" and "Basketball Court" exhibit *high inter-class similarity* due to their nearly identical appearances, differing only in subtle line markings. Similarly, "Large Vehicle" and "Small Vehicle" often appear together and share visual characteristics, further complicating class distinction. Moreover, classes like "Roundabout", "Baseball Diamond", "Ship", and "Harbor" display substantial *large intra-class heterogeneity* in shape, structure and color. Furthermore, due to the differences in ground sampling distances, remote sensing objects of the same class often display *significant scale variations*. These factors collectively increase the difficulty of FSS.

II. Related Work

A. Few-shot Semantic Segmentation for Natural Images

We first review FSS works for natural images. As shown in Fig. 1 and discussed in §I, existing FSS methods can be briefly divided into three categories. Prototype-based FSS methods (see Fig. 1(a)) typically follow a workflow that first extracts prototypes from the support data and then computes the relationor similarity between these support prototypes and the query features. In [8], [9], prototypes are defined as the mean vectors of embedded support features. Methods such as [10]–[12], [14], [56], [57] compute a prototype for each class using the masked average pooling (MAP), ensuring that the prototypes accurately represent class features, without interference from other class pixels. However, MAP compresses the spatial features of objects, which can lead to potentially causing semantic ambiguity across different parts of an object. Some works address this by generating prototypes through methods such as the expectation maximization (EM) algorithm [16], 1D pooling operations [58], or clustering algorithms [59], effectively enhancing feature representations.

However, by compressing feature maps into prototypes, prototype-based methods do not fully resolve the loss of spatial information. To address this limitation, pixel-level methods (see Fig. 1(b)) have been proposed, gaining increased attention for their superior performance. For example, Wang *et al.* [17] introduced a democratic attention mechanism to establish robust support-query relations, thereby activating more foreground target regions. CyCTR [19] performs self-alignment of query features and cross-alignment of support-query features based on vision transformer blocks, effectively integrating relevant support information into the query while capturing its context. DCAMA [20] adopts self-attention to predict query labels as a weighted sum of the labels of all the support feature pixels, with weights based on support-query similarity. Unlike DCAMA [20], DAM [23] designs a bidirectional 3D convolution to enhance the support-query similarity matrix, followed by a query mask created by cascading this matrix with lower-level query features.

Combining the strengths of both prototype-based and pixel-level methods, ProtoFormer [24] and SCTrans [25] treat the prototype as a conditional query, with query features serving as the key and value, inputting these elements into the transformer decoder for dense calculation (see Fig. 1(c)). This approach equips the conditional query with semantic awareness of the target class in the query image, resulting in improved FSS performance.

Unlike the feature optimization approaches used in the methods above, we propose HRL to extract and aggregate hierarchical support-query relations from both network design and training loss perspectives, enabling direct refinement of target object appearances.

B. Few-shot Semantic Segmentation for RSIs

In recent years, FSS in RSIs has gained traction, focusing primarily on support-query matching strategies [13], [60]–[62] and prototype updating mechanisms [63]–[65]. Yao *et al.* [60] introduced the first FSS method for RSIs, named the Scale-aware Detailed Matching network (SDM). In particular, SDM employs Prototype Matching Modules (PMMs) [16] to generate multiple support prototypes and then calculates the similarity between these prototypes and query features. Additionally, a scale-aware loss is designed to enhance learning for small objects. PCFNet [61] combines Euclidean distance with cosine similarity. Inspired by SDM and rotational invariant, FRINet [62] performs the rotation-adaptive matching of prototypes with query features from four directions.

R^{2}Net [63] dynamically updates global prototypes to refine local prototypes, reducing errors in prototype activation, and decouples foreground and background information to minimize object confusion. Similarly, DMNet [66] maintains a meta-prototype memory for foreground and background, which stores updated semantics for current class prototypes during training and suppresses activation of previously seen classes during testing. DMML [67] strengthens the original prototypes by applying a consistency principle with query features, using the similarity between enhanced pseudo-prototypes and original prototypes as a weight to refine the original prototype. Lang *et al.* [64] progressively parsed multiple valuable sub-regions from the support features to produce more representative prototypes. MS^{2}A^{2}Net [50] leverages attention mechanisms to extract the support-query relations, and employs a stepwise U-shaped fusion strategy for enhancement. However, this fusion process can accumulate erroneous relation information from bottom to top. In contrast, our HRNet parallelly extracts and aggregates hierarchical, multi-scale relations to directly optimize multiple object scales in the query image, reducing interference from erroneous information. Furthermore, we introduce BHLoss to supervise the learning of multi-scale relations, providing a robust guarantee for capturing diverse and reliable relation representations.

C. Semantic Segmentation for RSIs

Semantic segmentation for RSIs requires precise classification of each pixel into categories such as buildings, grass, vehicles, water, vegetation, and other surface objects/stuff. To improve segmentation accuracy and efficiency in RSIs, researchers have developed novel methods aimed at enhancing global context aggregation. Traditional convolutional neural network (CNN)-based methods indirectly capture global context through techniques like dilated convolutions [68], [69], pooling operations [69], [70], superpixel clustering [71], and attention mechanisms [39], [72]–[75]. However, these methods are often limited when dealing with the complexity of remote sensing scenes [76]. Recently, vision transformers have gained prominence in computer vision due to their innate global modeling capabilities. For RSI semantic segmentation, a common approach is to use a hybrid framework combining CNNs and transformers, where CNNs supply essential spatial information and transformers provide global semantic context [42], [45], [77]–[80]. For instance, Xu *et al.* [81] embedded adaptive transformer-based fusion modules within a dense CNN structure to enhance foreground saliency and reduce background noise. MSGCNet [46] utilizes a window-based transformer to capture global information from both spatial and channel dimensions. In multi-modality RSIs, FTransUNet [27] has demonstrated that transformers effectively facilitate cross-modality information transfer and establish long-range relationships. Leveraging the strengths of both CNNs and transformers, we employ them jointly to extract support-query relations for FSS in RSIs.

III. Methodology

A. Preliminaries

FSS aims to generalize a model trained on seen/base classes to segment unseen/novel classes using only a few support samples. Under the standard FSS setup [8], which adopts the episode-based meta-learning paradigm, the training set $\mathcal{D}_{base}$ and the test set $\mathcal{D}_{novel}$ contain samples from the base classes $\mathcal{C}_{base}$ and novel classes $\mathcal{C}_{novel}$, respectively, with $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \varnothing$. The samples in $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ are independently organized into a series of episodes. Each episode $(\mathcal{S}, \mathcal{Q})_c$ for a specific class c consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. For the K-shot segmentation, the support set $\mathcal{S}$ contains K image-mask pairs, while the query set $\mathcal{Q}$ consists of a single pair. These can be expressed as $\mathcal{S} = \left\{\left(\mathbf{I}_s^k, \mathbf{M}_s^k\right)\right\}_{k=1}^{K}$ and $\mathcal{Q} = \{(\mathbf{I}_q, \mathbf{M}_q)\}$, where k indexes the image-mask pairs in $\mathcal{S}$. Notably, during meta-training, the query mask $\mathbf{M}_q$ is used to compute the segmentation loss, but it is unavailable during meta-testing. The support masks $\mathbf{M}_s^k$, however, remains accessible throughout. The goal of FSS is to learn a robust mapping function $f : (\mathcal{S}, \mathbf{I}_q) \rightarrow \mathbf{M}_q$ on $\mathcal{D}_{base}$, enabling the segmentation of novel classes in $\mathcal{D}_{novel}$. For simplicity, we describe our method in the 1-shot setting, where $\mathcal{S} = \{(\mathbf{I}_s, \mathbf{M}_s)\}$.

B. Method Overview

The importance of multi-level/hierarchical learning in fully-supervised segmentation tasks is well-established, as demonstrated by numerous studies on multi-level feature fusion strategies [26]–[32] and encoder-decoder architectures [33]–[36]. However, hierarchical learning has received limited attention in FSS. Existing approaches typically employ multi-level feature fusion, exploring the correspondence between support features/prototypes and query features at a single level [47]–[50]. Since the essence of FSS lies in extracting support-query relations to segment novel target classes in query images, this paper directly investigates Hierarchical Relation Learning (HRL), which offers a more reasonable and intuitive solution than multi-level feature fusion. Specifically, our efforts include both network design and training strategy, namely the Hierarchical Relation Network (HRNet) and Bidirectional Hierarchical Loss (BHLoss). HRNet extracts and aggregates support-query relations across multiple hierarchies (in §III-C), while BHLoss encourages the model to focus on learning relations at different scales at different hierarchies by imposing appropriate constraints at each hierarchy (in §III-D). These two components work synergistically to enhance FSS performance for RSIs. The overall pipeline of our HRL is shown in Fig. 3.

C. Hierarchical Relation Network

As the network for HRL, HRNet consists of multiple parallel relation extraction branches and a multi-scale relation aggregation branch. The former focuses on single-scale

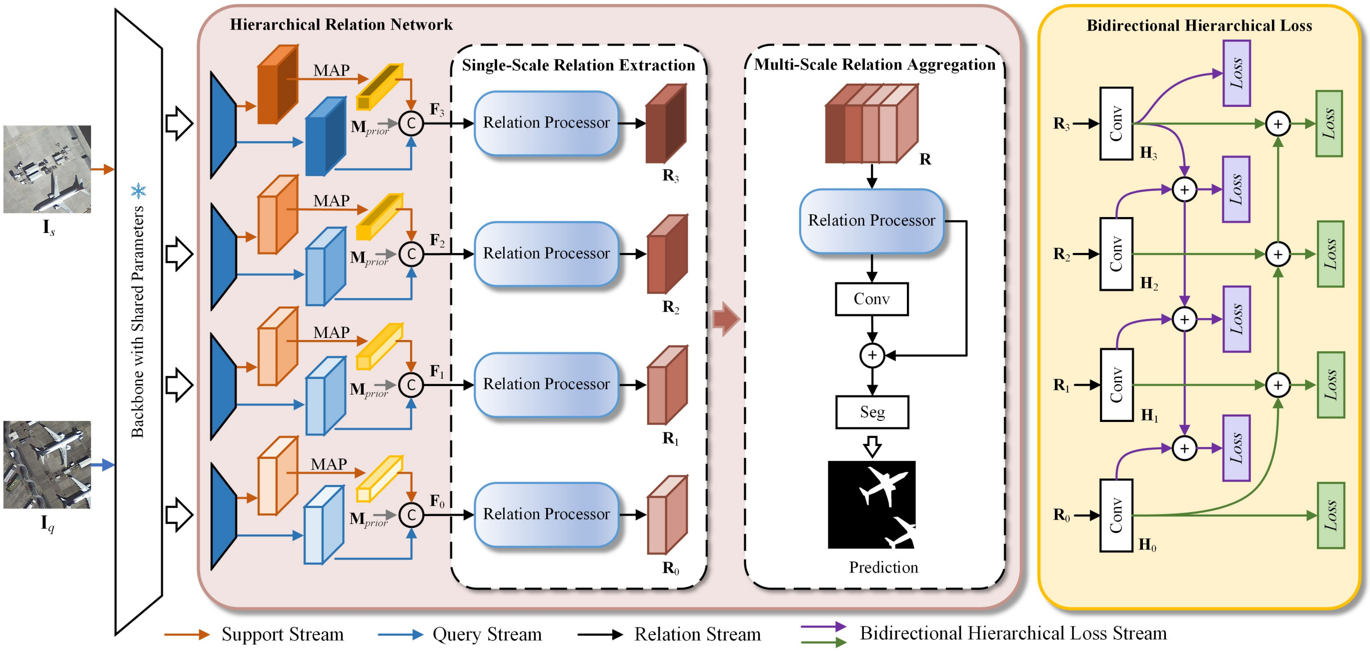

Fig. 3. Overall Pipeline of the proposed HRL. For clarity, this figure is presented in the 1-shot setting. The proposed HRNet is given on the left. We extract N hierarchies of support/query features from the frozen backbone network to preserve information generalizability and diversity. Each hierarchy of support-query feature pairs first accomplishes single-scale relation extraction to obtain a group of relations $\{\mathbf{R}_i\}_{i=1}^{N}$. Then, they are concatenated and fed into multi-scale relation aggregation to realize cross-hierarchical interactions of relation information, thereby directly segmenting novel classes in the query image. The structure of the bidirectional hierarchical loss is shown on the right. The supervision signal of the current hierarchy can be approximated by the sum of hierarchies lower than it and the sum of hierarchies higher than it. Please refer to §III-C and §III-D for more details.

support-query relation extraction at each branch, with multiple parallel branches extracting multi-scale/hierarchical support-query relations. The latter aggregates hierarchical support-query relations generated by the former, to learn a more comprehensive representation of support-query relations.

The mid-level features of the backbone (*e.g.*, Stage-3 and Stage-4 in ResNet-50 [82]) retain the ability to generalize to novel classes, providing a good basis for implementing relation matching between support and query images [15]. In previous FSS works [10], [63], it has been effective to freeze the backbone parameters and use the fused features from Stage-3 and Stage-4 as inputs to the support/query stream. This strategy ensures the generalization ability of features, thereby effective identifying target regions of novel classes. However, relying solely on features extracted by Stage-3 and Stage-4 still has certain limitations in scale expression. Considering that deeper networks have larger receptive fields [83], which is beneficial for capturing a wider range of contextual information. Thus, we add three standard residual blocks after Stage-4 to construct an extended backbone. Subsequently, the output of Stage-3 is fused with the outputs of Stage-4 and three standard residual blocks (Res-block1, Res-block2, Res-block3) to form support/query features from four hierarchies. This design not only preserves the generalization ability of mid-level features but also enriches the representation with multi-scale contextual information from subsequent layers, thus providing a more structured and discriminative feature foundation for subsequent support-query relation modeling.

Taking the 1-shot setup as an example, the support image $\mathbf{I}_s$ and query image $\mathbf{I}_q$ in an episode are fed into the extended backbone of shared parameters to obtain the feature pairs of the N hierarchies ($N = 4$), where the i-th can be represented as $(\mathbf{F}_i^s, \mathbf{F}_i^q)$ $(i \in \{1, 2, \cdots, N\})$, respectively. The size of each feature map is $H \times W \times C$. For each hierarchy, the support prototype $\mathbf{p}_i$ is obtained from the support feature $\mathbf{F}_i^s$ via masked average pooling (MAP) [10]–[12], written as $\mathbf{p}_i = \text{MAP}(\mathbf{F}_i^s, \mathbf{M}_s) \in \mathbb{R}^{1\times1\times C}$. Besides, the prior mask $\mathbf{M}_{prior}$ is derived from the cosine similarity between the high-level support and query features to assist in recognizing targets in the query image [15]. Then, the query feature $\mathbf{F}_i^q$, the prototype $\mathbf{p}_i$ corresponding to the support feature $\mathbf{F}_i^s$, and the prior mask $\mathbf{M}_{prior}$ are concatenated to obtain the feature $\mathbf{F}_i$, which can be expressed as $\mathbf{F}_i = \text{Concat}(\mathbf{F}_i^q, \mathbf{p'}_i, \mathbf{M}_{prior}) \in \mathbb{R}^{H\times W\times(2C+1)}$. Here, $\mathbf{p'}_i$ is the prototype feature of $\mathbf{p}_i$ after spatial dimension expansion.

Single-scale relation extraction. Subsequently, $\mathbf{F}_i$ is fed to the *relation processor* to extract the corresponding support-query relation $\mathbf{R}_i$ via self-aligned calculation. The relation processor contains a 3×3 convolution layer and a deformable transformer layer [84], [85] to take into account both local and global relation modeling. The convolution layer captures local relations $\mathbf{R}_i^{local} \in \mathbb{R}^{H\times W\times C}$ through weighted sum computation within the convolution kernel. Then, $\mathbf{R}_i^{local}$ is flattened into a 1-dimensional sequence as the input to the deformable transformer, which means that each spatial position is considered as a token. Unlike standard transformer that compute global self-attention across all tokens, the deformable transformer dynamically selects a small number of the most relevant spatial locations through a learnable offset and only calculates attention for these key sampling points. This strategy significantly reduces computational complexity, while adaptively accommodating scale variations of target regions between the support and query images, reducing redundant and incorrect feature matching, thereby achieving more precise global relation representation. This spread mechanism from local extraction to global pixel-level self-alignment accomplishes progressive enhancement of the relation representation.

Multi-scale relation aggregation. Based on our network design, the output of each hierarchy implies the single-scale relation between the support and query features. Fusing multi-scale features is a well-recognized and effective strategy for enhancing feature robustness and diversity in computer vision [33]–[36]. Here, we integrate multi-scale relations from different hierarchies, expecting to gain high-quality relation representations. Specifically, we concatenate support-query relations from all hierarchies to derive the multi-scale relation $\mathbf{R} = \mathrm{Concat}(\{\mathbf{R}_i\}_{i=1}^{N})$, followed by a relation processor for refinement. Then, a residual structure is accessed to prevent degradation of the relation representation, denoted as $\mathbf{R}' = \mathrm{RelationProcessor}(\mathbf{R})$ and $\mathbf{R}'' = \mathbf{R}' + \mathrm{Conv}(\mathbf{R}')$. Finally, we feed the refined relation $\mathbf{R}''$ into the segmentation head to shape the target region of the novel class in the query image, obtaining the prediction mask $\mathbf{P}$. Here, the segmentation head consists of a 3×3 convolution layer, a ReLU layer, and a 1×1 convolution layer. During training, the prediction mask $\mathbf{P}$ is supervised by $\mathbf{M}_q$ with a standard cross-entropy loss.

D. Bidirectional Hierarchical Loss

As mentioned before, we expect the model to mine the whole target regions in the query image by aggregating diverse support-query relations from different network hierarchies, thus obtaining better predictions through multi-scale relation information integration and correction. Exploring hierarchical relations allows the model to learn more discriminative representations from remote sensing objects with high intra-class variation and high inter-class similarity. To this end, we propose BHLoss to efficiently impose proper supervision at each hierarchy of HRNet to achieve the above goal.

Along this line, the i-th hierarchy should be equipped with a segmentation head $\varphi_i(\cdot)$ whose output can be expressed as $\mathbf{H}_i = \varphi_i(\mathbf{R}_i)$. Accordingly, the ground-truth $\mathbf{M}_q$ is decomposed into N binary masks corresponding to different object scales, serving as the training goals for each hierarchy, *i.e.*,

$$\mathbf{M}_q = \sum_{i=1}^{N} \mathbf{M}_i^q. \tag{1}$$

At this point, relation information mining of a specific scale is achieved by training $\varphi_i(\cdot)$ through a binary cross-entropy (BCE) loss such that $\mathbf{H}_i \sim \mathbf{M}_i^q$, expressed as,

$$\begin{aligned}\mathcal{L}_i &= \mathrm{BCE}(\mathbf{H}_i, \mathbf{M}_i^q) \\ &= -\mathbf{M}_i^q \log \mathbf{H}_i - (1 - \mathbf{M}_i^q) \log(1 - \mathbf{H}_i).\end{aligned} \tag{2}$$

TABLE I
NOVEL CLASSES FOR DIFFERENT DATASETS.

Split	Novel classes		
	iSAID-5^i	DLRSD-5^i	LoveDA-2^i
0	Ship Storage Tank Baseball Diamond Tennis Court Basketball Court	Airplane Bare Soil Building Car Chaparral	Building Road
1	Ground Track Field Bridge Large Vehicle Small Vehicle Helicopter	Court Dock Field Grass Mobile Home	Water Barren
2	Swimming pool Roundabout Soccer Ball Field Plane Harbor	Pavement Sand Sea Ship Tank	Forest Agriculture

However, it is difficult to manually decompose $\mathbf{M}_q$ into different scales, which leads to the failure of directly supervising the output of each hierarchy. The absence of an explicit $\mathbf{M}_i^q$ makes Eq. (2) unworkable. Observing Eq. (1), it is easy to notice that the decomposed binary masks $\mathbf{M}_i^q$ of one hierarchy can be approximated by $\mathbf{M}_q$ and the prediction masks of other hierarchies, denoted as

$$\mathbf{M}_i^q \sim \mathbf{M}_q - \sum_{i \neq j} \mathbf{H}_j. \tag{3}$$

Since $\mathbf{H}_i$ approximates $\mathbf{M}_i^q$, Eq. (3) can be converted to $\mathbf{M}_q \sim \sum_{i=1}^{N} \mathbf{H}_i$, notating the latter as $\hat{\mathbf{M}}_q$. The overall training loss is written as $\mathcal{L} = \mathcal{L}(\hat{\mathbf{M}}_q, \mathbf{M}_q)$, so that the gradient for the i-th hierarchy can be computed as

$$\begin{aligned}\frac{\partial(\mathcal{L})}{\partial\left(\mathbf{H}_i\right)} &= \frac{\partial(L(\hat{\mathbf{M}}_q, \mathbf{M}_q))}{\partial\left(\mathbf{H}_i\right)} \\ &= \frac{\partial(\mathcal{L}(\hat{\mathbf{M}}_q, \mathbf{M}_q))}{\partial(\hat{\mathbf{M}}_q)} \cdot \frac{\partial(\hat{\mathbf{M}}_q)}{\partial\left(\mathbf{H}_i\right)} \\ &= \frac{\partial(\mathcal{L}(\hat{\mathbf{M}}_q, \mathbf{M}_q))}{\partial(\hat{\mathbf{M}}_q)} \cdot \frac{\partial(\sum_{i=1}^{N} \mathbf{H}_i)}{\partial\left(\mathbf{H}_i\right)} \\ &= \frac{\partial(\mathcal{L}(\hat{\mathbf{M}}_q, \mathbf{M}_q))}{\partial(\hat{\mathbf{M}}_q)}.\end{aligned} \tag{4}$$

Unfortunately, the extrapolation based on Eq. (4) intuitively demonstrates that the gradient of the overall loss is independent of the output of each hierarchy, *i.e.*, any two different hierarchies i and j have equal loss gradients. That is, it is not possible to focus on different scales at different hierarchies, which is contrary to the purpose of imposing appropriate supervision signals at each network hierarchy.

To solve this problem, we propose a bidirectional approximation method to estimate the supervision signal $\mathbf{M}_i^q$ of the i-th hierarchy. $\mathbf{M}_i^q$ satisfies both conditions, approximating with the sum of the hierarchies lower than i and the sum of the hierarchies higher than i, as denoted by

$$\begin{aligned}\mathbf{M}_i^q &\sim \mathbf{M}_q - \sum_{j<i} \mathbf{H}_j, \\ \mathbf{M}_i^q &\sim \mathbf{M}_q - \sum_{j>i} \mathbf{H}_j.\end{aligned} \tag{5}$$

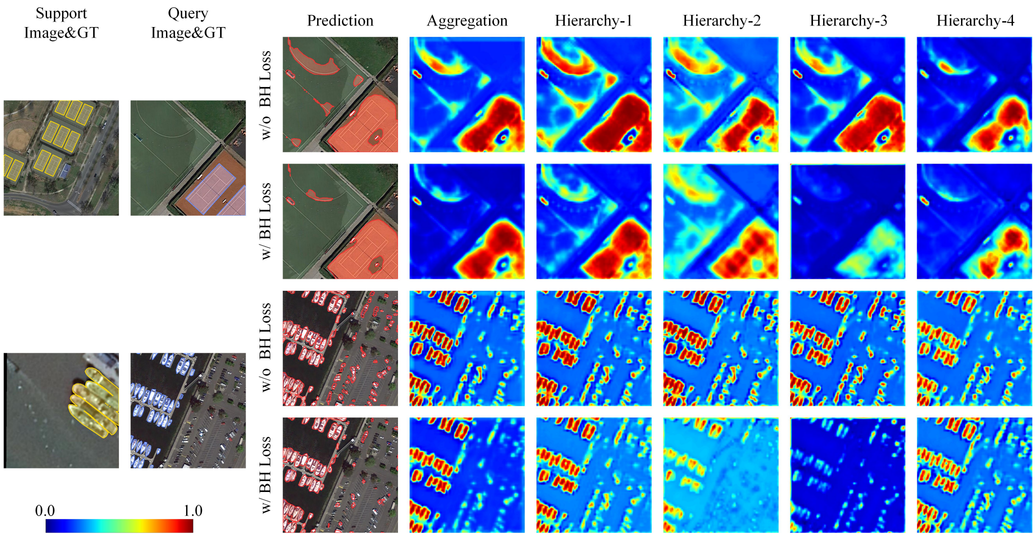

Fig. 4. Visualization of the heat maps before and after adding the BHLoss, with the 1-shot setup and ResNet-50 [82] as the backbone.

TABLE II
ABLATION STUDY OF THE RELATION PROCESSOR IN RESNET-50 [82] AS BACKBONE.

K-Shot	Baseline (Convs)	Relation Processor	mIoU			
			Split0	Split1	Split2	Mean
1-shot	✓		39.57	24.92	31.48	31.99
		✓	40.20	25.19	32.73	32.71
5-shot	✓		45.06	29.76	38.05	37.62
		✓	45.60	30.51	39.02	38.38

Such an approach allows $\mathbf{M}_i^q$ to approximate real supervision and be trainable. Taking the second hierarchy as an example, $\mathbf{M}_2^q$ needs to approximate $\mathbf{M}_q - \mathbf{H}_1$ while approximating $\mathbf{M}_q - \mathbf{H}_3 - \mathbf{H}_4$. After a simple transformation, the loss of the second hierarchy can be written as

$$\begin{cases} \mathcal{L}_2^- = \text{BCE}(\mathbf{H}_1 + \mathbf{H}_2, \mathbf{M}_q), \\ \mathcal{L}_2^+ = \text{BCE}(\mathbf{H}_2 + \mathbf{H}_3 + \mathbf{H}_4, \mathbf{M}_q). \end{cases} \tag{6}$$

The loss for other hierarchies can be derived in a similar way.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We verify the performance of HRL on iSAID-5^i [51], DLRSD-5^i [52], and LoveDA-2^i [53] datasets, with novel classes are listed in Table I.

iSAID-5^i dataset: iSAID [51] is a challenging, well-labelled, high-resolution remote sensing image dataset available for semantic and instance segmentation tasks. It contains 655,451 object instances across 15 classes. Following previous works [60]–[65], we divide the classes into 3 splits, each containing five classes. In each experiment, we use one split as novel classes for testing, while the other two splits serve as base classes for training. Additionally, we use 512×512 pixel cropped images, consistent with [63].

TABLE III
ABLATION STUDY ABOUT THE HRNET UNDER THE 5-SHOT SETTING WITH RESNET-50 [82] AS THE BACKBONE.

Split	Hierarchy Indices (SRE)				Results (MRA)
	1^{st}	2^{nd}	3^{rd}	4^{th}	
0	✓				45.60
				✓	45.66(+0.06)
	✓	✓			47.29(+1.69)
	✓	✓	✓		47.51(+1.91)
	✓	✓	✓	✓	48.10(+2.50)
1	✓				30.51
				✓	31.21(+0.70)
	✓	✓			31.22(+0.71)
	✓	✓	✓		31.53(+1.02)
	✓	✓	✓	✓	32.18(+1.67)
2	✓				39.02
				✓	39.51(+0.49)
	✓	✓			41.11(+2.09)
	✓	✓	✓		41.91(+2.89)
	✓	✓	✓	✓	42.26(+3.24)

TABLE IV
ABLATION STUDY ON THE HIERARCHICAL RELATION INTEGRATION APPROACH IN HRNET WITH RESNET-50 [82] AS THE BACKBONE.

K-Shot	Aggregation Approach	mIoU			
		Split0	Split1	Split2	Mean
1-shot	Parallel	40.96	26.95	34.03	33.98
	Bottom-up	40.25	26.31	33.39	33.32
5-shot	Parallel	48.10	32.18	42.26	40.85
	Bottom-up	46.58	29.98	41.27	39.28

DLRSD-5^i dataset: DLRSD [52] is a publicly available remote sensing image dataset for training and evaluating multi-label image retrieval and semantic segmentation models. It comprises 2,100 RGB images, each sized 256×256 pixels, spanning 17 object classes. Consistent with the setting in [86], the first 15 object classes are divided into three splits.

TABLE V
EFFECT OF THE NUMBER OF SUPPORT IMAGES WHEN USING RESNET-50 [82] AS THE BACKBONE AND THE INPUT RESOLUTION OF 512×512.

K-Shot	mIoU				FLOPs
	Split0	Split1	Split2	Mean	
1	41.50	27.84	34.51	34.62	468.8G
2	44.56	30.31	37.57	37.48	716.9G
3	45.60	31.60	39.91	39.04	964.9G
4	47.12	32.36	30.23	39.90	1213.0G
5	49.01	32.86	43.26	41.71	1461.0G

TABLE VI
ABLATION STUDY OF BIDIRECTIONAL HIERARCHICAL LOSS WITH RESNET-50 [82] AS THE BACKBONE.

K-Shot	SH Loss	BH Loss	mIoU			
			Split0	Split1	Split2	Mean
1-shot			40.96	26.95	34.03	33.98
	✓↑		40.57	27.42	34.36	34.12
	✓↓		40.64	27.38	34.09	34.01
		✓	41.50	27.84	34.51	34.62
5-shot			48.10	32.18	42.26	40.85
	✓↑		47.85	32.37	42.74	40.11
	✓↓		47.42	32.52	42.81	40.92
		✓	49.01	32.86	43.26	41.71

LoveDA-2^i dataset: LoveDA [53] contains 5,987 optical RSIs with a resolution of 1024×1024 pixels, showcasing urban and rural scenes from Wuhan, Changzhou, and Nanjing in China. These images contain 166,768 objects across 7 classes: Background, Building, Road, Water, Barren, Forest, and Agriculture. Following the setup in [66], we divide all classes except Background into three splits and employ cropped images of 473×473 pixels.

2) Implementation Details: Our model is implemented in the PyTorch framework [87] and runs on two GeForce RTX 2080 GPUs. The transformer layer in the relation processor is configured with 8 heads and an embedding dimension of 256. Its parameters are optimized using the AdamW optimizer with an initial learning rate of 10^{-4}, while the remaining parameters use the SGD optimizer with an initial learning rate of 0.001. The batch size and number of epochs are set to 8 and 12, respectively. For the iSAID-5^i [51], DLRSD-5^i [52], and LoveDA-2^i [53] datasets, we follow the training and testing configurations in [63], [86], and [66], respectively, including the image enhancement techniques, test-episode amount, test-independent run counts, the random seeding, *etc*.

3) Evaluation metrics: The *mean intersection over union* (mIoU) and foreground-background IoU (FB-IoU) are two commonly-used evaluation metrics for semantic segmentation. mIoU is calculated as the average IoU of x novel classes in each split ($x = 5$ for iSAID-5^i [51] and DLRSD-5^i [52], $x = 2$ for LoveDA-2^i [53]), while FB-IoU is the average of the foreground IoU and background IoU. We use employ metrics to comprehensively evaluate model performance.

B. Ablation Study

To clarify the benefits of each component in HRL, we conduct rigorous ablation experiments using ResNet-50 as the backbone on the iSAID-5^i dataset [51]. Notably, the baseline model only fuses prototype and query feature from a single hierarchy and then directly obtains the prediction through two standard convolution layers and a segmentation head.

1) Effects of the relation processor: We replace the two convolution layers of the baseline with a relation processor consisting of a convolution layer and a transformer layer. The results are shown in Table II. Benefiting from the local information clustering of the convolution and the global modeling capabilities of transformer, the relation processor enables an progressive extraction from local to global, allowing for more comprehensive pixel-level self-relational alignment between the support and query features. Compared to local extraction alone, it grows by an average growth of 0.72% mIoU per split in the 1-shot setting and 0.76% mIoU in the 5-shot setting.

2) Effects of the HRNet: Our HRNet consists of four independent *single-scale relation extraction* (SRE) modules and a *multi-scale relation aggregation* (MRA) module. Table III shows the experimental results for SRE from different hierarchies and the cross-hierarchy MRA. For a single hierarchy, *i.e.*, without MRA, the last SRE shows a performance increase over the first one as the depth of the network grows, but this increase is limited, rising by only 0.06%∼0.7% mIoU. In contrast, the introduction of MRA stimulates the refinement of support-query relations across hierarchies and results in a significant overall performance improvement, with gains of 2.50%, 1.67%, and 3.24% mIoU on the three splits compared to SRE, respectively. This suggests that exploring the potential relations among hierarchies effectively boosts the recognition of the target in the query image. The first and third rows of Fig. 4 visually demonstrate that MRA performs better than each single hierarchy, suggesting its superior ability to integrate complementary relation information from multiple hierarchies and thus enhance segmentation accuracy.

Further, we perform ablation experiments on the approach of aggregating hierarchical relations in HRNet. In the process of modeling support-query relations, this bottom-up aggregation strategy tends to cause erroneous relations formed at early hierarchies to be progressively propagated and accumulated in the subsequent hierarchies. In contrast, our parallel aggregation strategy can extract and integrate support-query relations across multiple levels simultaneously. This design avoids the reliance on a single accumulation path, thus reducing the risk of combining incorrect relations. Table IV shows that the parallel aggregation strategy outperforms the bottom-up strategy across various few-shot settings, demonstrating its effectiveness in preventing the accumulation and propagation of incorrect relations.

3) Effects of the number of support images: Table V lists the effect of the number of support images on FSS. It can be observed that with more support images, the model can learn more diverse and comprehensive reference information, thus establishing a more robust support-query relationship and improving the ability to recognize similar targets in the query image. However, for every increase in K, *i.e.*, one additional

TABLE VII
COMPARISON OF SEGMENTATION RESULTS WITH MIOU ON THE ISAID-5^i DATASET [51]. THE RESULTS OF THE COMPARISON METHODS ARE TRANSCRIBED FROM R^2NET [63]. THE BEST AND SECOND-BEST RESULTS ARE IN BOLD AND UNDERLINED, RESPECTIVELY.

Backbones	Methods	Pub. Years	1-shot				5-shot			
			Split0	Split1	Split2	Mean	Split0	Split1	Split2	Mean
VGG-16	*PANet* [11]	ICCV'2019	26.86	14.56	20.69	20.70	30.89	16.63	24.05	23.86
	CANet [10]	CVPR'2019	13.91	12.94	13.67	13.51	17.32	15.07	18.23	16.87
	SCL [88]	CVPR'2021	25.75	18.57	22.24	22.19	35.77	24.92	32.70	31.13
	PEFNet [15]	TPAMI'2020	28.52	17.05	18.94	21.50	37.59	23.22	30.45	30.42
	NTRENet [89]	CVPR'2022	25.78	20.01	19.88	21.89	38.43	24.21	28.99	30.54
	DCPNet [90]	CVPR'2022	28.17	16.52	22.49	22.39	39.65	22.68	29.93	30.75
	BAM [14]	CVPR'2022	33.97	16.88	21.47	24.09	38.46	22.76	28.81	30.01
	DMML [67]	TGRS'2022	24.41	18.58	19.46	20.82	28.97	21.02	22.78	24.26
	SDM [60]	TGRS'2022	24.52	16.31	21.01	20.61	26.73	19.97	26.10	24.27
	DML [13]	GRSL'2022	30.99	14.60	19.05	21.55	34.03	16.38	26.32	25.58
	TBPN [91]	Neuroc.'2023	27.86	12.32	18.16	19.45	32.79	16.28	24.27	24.45
	PCNet [64]	TGRS'2023	32.48	19.88	24.56	25.64	41.09	21.98	34.14	32.40
	R^2Net [63]	TGRS'2023	**35.27**	19.93	24.63	26.61	42.06	23.52	30.06	31.88
	MGANet [92]	TGRS'2024	30.46	20.60	20.77	23.94	32.74	23.52	25.28	27.18
	HRL (Ours)	–	33.54	**23.10**	**31.58**	**29.40**(+2.79)	**43.02**	**26.19**	**39.26**	**36.16**(+3.76)
ResNet-50	*PANet* [11]	ICCV'2019	27.56	17.23	24.60	23.13	36.54	16.05	26.22	26.27
	CANet [10]	CVPR'2019	25.51	13.50	24.45	21.15	29.32	21.85	26.91	26.03
	SCL [88]	CVPR'2021	34.78	22.77	31.20	29.58	41.29	25.73	37.70	34.91
	PEFNet [15]	TPAMI'2020	35.84	23.35	27.20	28.80	42.42	25.34	33.00	33.59
	NTRENet [89]	CVPR'2022	34.93	23.95	28.56	29.15	44.83	26.73	37.19	36.25
	DCPNet [90]	IJCAI'2022	37.83	22.86	28.92	29.87	41.52	28.18	33.43	34.38
	BAM [14]	CVPR'2022	39.43	21.69	28.64	29.92	43.29	27.92	38.62	36.61
	DMML [67]	TGRS'2022	28.45	21.02	23.46	24.31	30.61	25.23	24.08	26.18
	SDM [60]	TGRS'2022	27.96	21.99	26.27	25.92	28.50	22.05	31.07	28.27
	DML [13]	GRSL'2022	32.96	18.98	25.47	26.07	33.58	20.42	29.77	28.47
	TBPN [91]	Neuroc.'2023	29.33	16.84	25.47	23.88	30.98	20.42	28.07	26.49
	PCNet [64]	TGRS'2023	40.24	24.64	31.31	32.06	45.31	28.19	37.36	36.95
	R^2Net [63]	TGRS'2023	41.22	21.64	**35.28**	32.71	46.45	25.80	39.84	37.36
	MS^2A^2Net [50]	TGRS'2024	34.96	25.30	30.71	30.32	42.30	27.10	36.31	35.24
	MGANet [92]	TGRS'2024	35.48	22.37	27.75	28.53	38.38	26.63	35.05	33.35
	HRL (Ours)	–	**41.50**	**27.84**	34.51	**34.62**(+1.91)	**49.01**	**32.86**	**43.26**	**41.71**(+4.35)

TABLE VIII
CLASS-WISE PERFORMANCE COMPARISON IN MIOU IN 1-SHOT SETTING USING RESNET-50 [82] AS THE BACKBONE. C1-C15 CORRESPOND TO THE 15 CLASSES OF THE ISAID-5^i DATASET [51] AND DLRSD-5^i DATASET [52]. THE BEST AND SECOND-BEST RESULTS ARE IN BOLD AND UNDERLINED, RESPECTIVELY.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Mean
iSAID-5^i																
PANet [11]	21.81	36.31	23.01	42.06	14.59	12.11	17.44	22.70	12.27	21.60	30.29	24.62	26.79	25.54	15.79	23.13
CANet [10]	39.57	18.54	18.46	33.63	17.34	9.78	5.49	22.15	5.17	24.89	9.96	36.50	19.12	38.82	17.85	21.15
SCL [88]	37.61	33.63	26.68	54.75	21.22	22.60	24.40	30.22	6.71	29.93	**33.00**	44.68	18.25	44.63	15.46	29.58
PEFNet [15]	39.02	45.63	20.86	49.96	23.72	21.00	24.76	31.59	6.98	32.42	13.34	**47.64**	30.65	32.82	11.54	28.80
NERTNet [89]	33.59	42.83	22.30	49.35	21.91	21.62	28.82	25.64	9.35	34.30	23.91	38.67	25.63	40.84	13.74	28.83
DCPNet [90]	37.42	42.44	35.16	56.55	17.58	21.66	19.57	32.97	10.60	29.50	24.02	35.34	28.44	39.80	17.02	29.87
BAM [14]	36.34	39.76	**38.23**	**58.13**	24.71	18.25	12.68	35.91	11.42	30.21	28.98	40.74	29.43	33.25	10.79	29.92
TBPN [91]	25.36	41.28	30.67	32.88	16.48	13.48	9.74	27.88	12.52	20.56	11.12	34.31	23.57	40.36	17.98	23.88
DMML [67]	40.14	40.18	21.31	27.02	13.60	15.56	15.19	26.05	13.84	**34.44**	11.26	17.57	23.27	39.11	**26.12**	24.31
SDM [60]	41.77	35.50	21.41	20.81	20.29	15.60	25.60	28.66	13.29	26.79	13.61	32.35	24.59	42.79	25.75	25.92
DML [13]	35.13	42.10	30.49	41.79	15.31	13.25	16.87	24.70	14.62	25.45	10.24	35.49	25.35	41.69	18.57	26.07
R^2Net [63]	**46.87**	49.06	30.70	52.86	**26.62**	24.31	17.25	31.25	13.67	21.73	24.88	46.07	**42.29**	42.07	21.08	32.71
MGANet [92]	34.85	45.13	30.41	50.94	16.05	11.99	23.35	33.39	13.65	29.45	13.52	32.19	26.65	44.46	21.90	28.53
HRL (Ours)	43.40	**50.69**	36.01	56.99	20.54	**27.03**	**29.15**	**38.88**	**16.15**	28.01	25.23	35.94	38.72	**48.51**	24.15	**34.62**
DLRSD-5^i																
HSNet [21]	23.63	18.39	21.41	8.55	38.02	63.45	24.56	**96.49**	18.33	33.20	20.88	24.66	57.13	35.00	35.94	34.64
SDM [60]	5.51	22.74	29.00	3.83	39.49	5.30	19.97	84.13	8.94	35.90	11.96	**31.99**	49.03	38.79	7.57	26.27
SCCNet [86]	23.58	25.32	26.99	10.45	40.37	53.27	25.49	96.29	**29.10**	40.68	30.09	24.80	**60.07**	46.72	37.00	37.37
HRL (Ours)	**34.91**	**46.78**	**56.75**	**18.27**	**41.88**	**64.66**	**30.72**	71.28	26.31	**62.60**	**31.71**	27.50	36.79	**56.68**	**44.22**	**43.40**

support sample, the floating point operations (FLOPs) grow by about 248 G, which puts a higher demand on the computational resources.

4) Effects of the BHLoss: To further explore the necessity of the bidirectional hierarchical loss (BHLoss), we decompose it into two single-directional hierarchical losses (SHLoss) and conduct experiments separately for validation. The results in Table VI indicate that when only a SHLoss is applied (*i.e.*, the green or purple lines), the information flow is passed in a single direction toward either the lowest or highest layer,

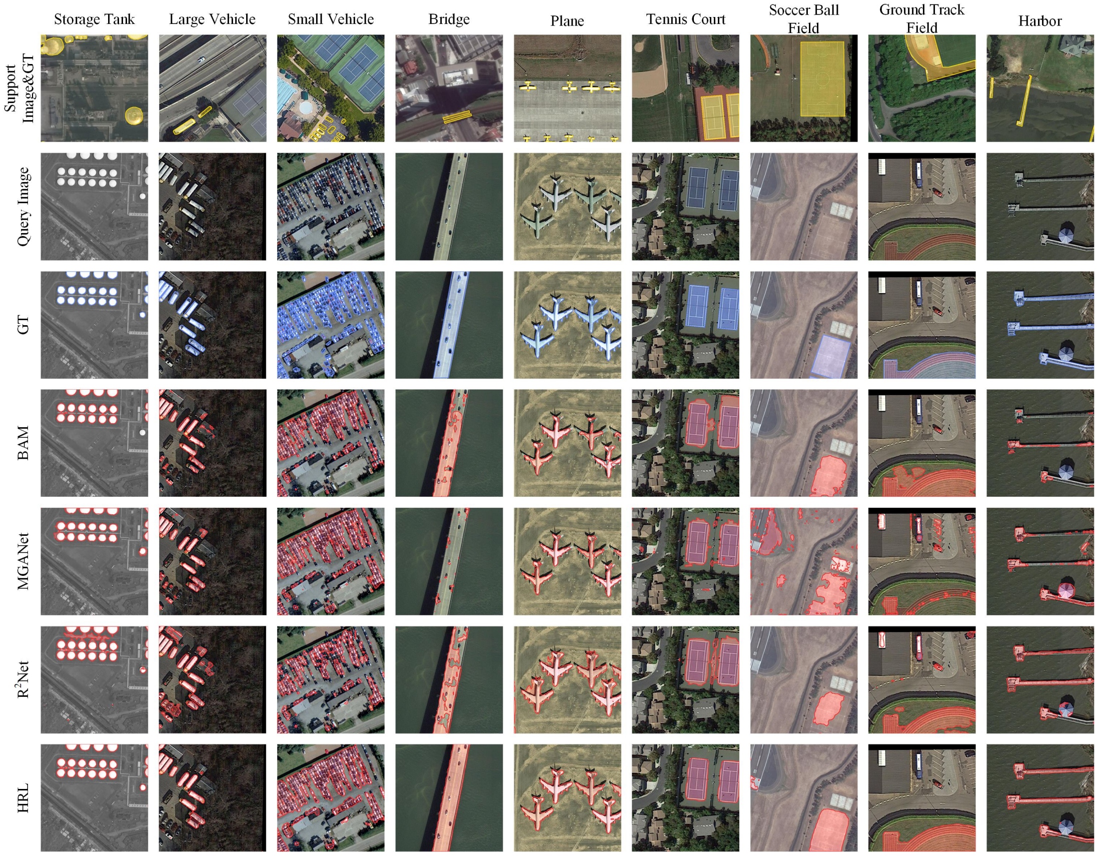

Fig. 5. Segmentation prediction of the iSAID-5^i dataset [51] in 1-shot setting.

TABLE IX
COMPARISON OF SEGMENTATION RESULTS WITH MIOU AND FB-IOU ON THE DLRSD-5^i DATASET [52]. THE RESULTS OF THE COMPARISON METHODS WITH RESNET-50 [82] AS THE BACKBONE ARE TRANSCRIBED FROM SCCNET [86]. THE BEST AND SECOND-BEST RESULTS ARE IN BOLD AND UNDERLINED, RESPECTIVELY.

Methods	Pub. Years	1-shot				5-shot			
		Split0	Split1	Split2	Mean	Split0	Split1	Split2	Mean
HSNet [21]	ICCV'2021	22.00	47.20	34.73	34.64	27.46	52.32	46.23	42.00
SDM [60]	TGRS'2022	20.11	30.84	27.87	26.27	26.03	41.74	33.55	33.77
SCCNet [86]	AGIS'2023	25.34	48.97	39.73	37.37	30.22	**52.40**	47.15	43.26
PCFNet [61]	TGRS'2023	31.09	37.52	**41.89**	36.83	36.67	42.70	**54.20**	44.52
HRL (Ours)	–	**39.72**	**51.11**	39.38	**43.40**(+6.03)	**41.37**	51.90	42.86	**45.38**(+0.86)

leading to homogeneous relation modeling. Since SHLoss unable to simultaneously constrain information extraction across different levels, it fails to deliver performance improvements. These findings validate the necessity and effectiveness of the bidirectional hierarchical loss (BHLoss) in maintaining hierarchical information balance and promoting diverse relation modeling. Specifically, after introducing BHLoss into the HRNet framework, the MIoU increased by 0.64% under the 1-shot setting and by 0.86% under the 5-shot setting. Compared with the baseline, BHLoss brings a gain of 2.63% and 4.09% MIoU in the 1-shot and 5-shot settings, respectively. Fig. 4 illustrates the heat maps for each hierarchy before and after adding BHLoss. As can be observed, our BHLoss successfully improves the diversity of hierarchical relations and suppresses the activation of non-target regions by applying appropriate supervision signals at different hierarchies.

C. Comparison with State-of-the-arts

We compare the proposed HRL with existing state-of-the-art few-shot methods and conduct both numerical and visual analyses on the iSAID-5^i [51], DLRSD-5^i [52], and LoveDA-2^i [53] datasets. Under comprehensive consideration, the comparison methods are from both natural and remote sensing image domains. In the comparison tables of segmentation results, we use italics to represent the FSS methods for natural images.

1) iSAID-5^i dataset: We first carry out experimental comparison on the iSAID-5^i dataset [51]. Table VII lists test results for state-of-the-art methods with different backbone and few-shot settings, and Table VIII shows in detail the segmentation results for each category with ResNet-50 as the backbone in the 1-shot setting. Our proposed HRL achieves the best performance for both individual splits and the total average in the 5-shot setting, improving by 3.76% mIoU over the second-ranked PCNet [64] with VGG-16 and by 4.35% mIoU over the second-ranked R^{2}Net [63] with ResNet-50. In different few-shot settings, HRL outperforms MS^{2}A^{2}Net [50], demonstrating that our HRL effectively prevents the accumulation and transmission of erroneous relations within the hierarchy. When using ResNet-50 as the backbone network, HRL slightly lags behind the advanced R^{2}Net [63] only in the Split2 of the 1-shot setting, while achieving superior results in all other configurations. According to Table VIII, this gap primarily arises from C12 (Roundabout). Due to the complex traffic patterns in “Roundabout” areas (as illustrated in Fig. 2), these regions, serving as critical nodes in road networks, exhibit significant heterogeneity in both shape and structure. They often overlap with certain road segments, resulting in considerable variance in the hierarchical relation modeling by HRL for this category, which in turn affects the final relation aggregation results. Despite the poor performance on “Roundabout”, the benefits of HRL, particularly in scale-awareness, should not be overlooked. It is worth noting that under all comparison settings, HRL consistently outperforms SDM [60], which alleviates scale variation of remote sensing objects through multi-prototype representation, as well as MGANet [92], which adopts a multi-granularity aggregation strategy. For small-scale objects, HRL surpasses the state-of-the-art R^{2}Net [63] by 11.90% mIoU on C7 (Small Vehicle) and 7.63% mIoU on C8 (Large Vehicle), respectively.

Fig. 5 visualizes the segmentation results for some episodes in the 1-shot setting using ResNet-50. We select several representative scenarios for detailed analysis. In the second column, the support image contains only “Large Vehicle”, while the query image includes both “Small Vehicle” and “Large Vehicle”. Due to the inability of BAM [14], MGANet [92], and R^{2}Net [63] to effectively distinguish inter-class relation differences during the matching process, “Small Vehicle” are incorrectly segmented. In the third column, the dense distribution and high intra-class variation of “Small Vehicle” present challenges for establishing accurate support-query correspondence. In the fourth column, “Bridge” and the road share similar construction materials, resulting in ambiguous boundaries. Moreover, the support image depicts “Bridge” within a complex urban scene, while the background of the query image is relatively simple. This significant difference in background hinders the ability of the model to locate the corresponding region in the query image. Consequently, MGANet [92] misclassifies “Bridge” as a road and only segments scattered parts of it, while BAM [14] and R^{2}Net [63] also fail to produce complete segmentation results. In the fifth column, “Plane” exhibits complex boundary structures. Benefiting from the hierarchical relation learning, HRL demonstrates superior capability in parsing fine details such as wings and tail components. In the seventh column, “Soccer Ball Field” exhibits large intra-class variation and lacks clear field markings in the query image, making it difficult to identify, even for human observers. As a result, the compared methods produce incomplete segmentations. In the ninth column, the narrow “Harbor” is closely connected to the pavilion, and both MGANet [92] and R^{2}Net [63] incorrectly classify the pavilion as part of “Harbor”. Despite these tricky scenarios, the proposed HRL demonstrates strong robustness by effectively extracting single-scale relations and aggregating multi-scale support-query relations across different hierarchies, significantly improving the stability and accuracy of FSS in RSIs.

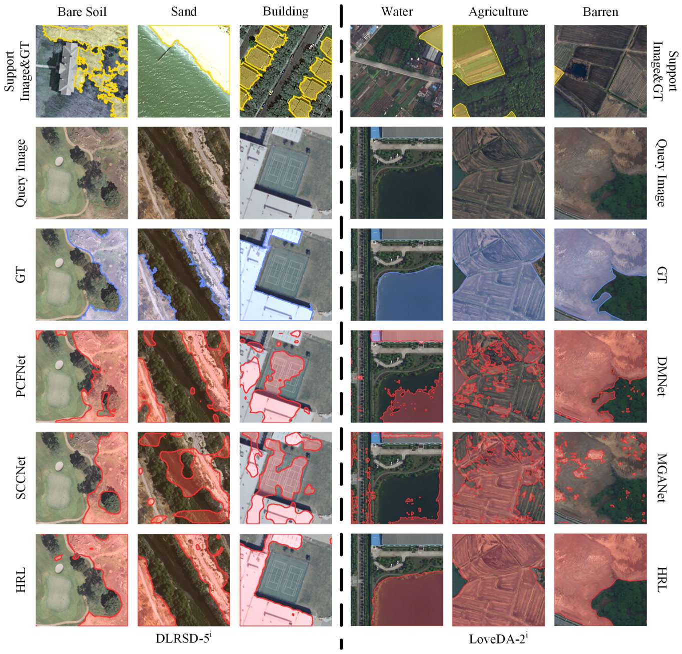

Fig. 6. Segmentation prediction of HRL with ResNet-50 [82] on DLRSD-5^i dataset [52] and LoveDA-2^i dataset [53] in the 1-shot setting.

2) DLRSD-5^i dataset: We proceed by comparing our method with existing methods on the DLRSD-5^i dataset [52]. As listed in Table IX, our HRL obtains the highest average performance across three splits on DLRSD-5^i [52], exceeding the second-ranked SCCNet [86] by 6.03% mIoU in the 1-shot setting, and surpassing the second-ranked PCFNet [61] by 0.86% mIoU in the 5-shot setting. More detailed results from Table VIII reveal that HRL is superior to the existing methods in 11 out of 15 classes. This highlights the importance and effectiveness of hierarchical relation learning in FSS for RSIs. The left side of Fig. 6 visualizes several segmentation results. The DLRSD-5^i dataset contains several challenging scenarios, such as the high inter-class similarity between “Bare Soil” and “Sand”, as well as the large intra-class variability of “Building”. In addition, both “Bare Soil” and “Sand” have ambiguous boundaries with their surrounding areas, posing a greater challenge to the robustness of the model. PCFNet [61] and SCCNet [86] both fail to accurately segment these classes due to their inability to correctly match the relationships between the support prototypes and the query images. When handling classes like “Building”, which have distinct boundaries but highly variable appearances, both methods misclassify “Court” as “Building” due to insufficient relation modeling. In contrast, HRL employs hierarchical relation extraction and cross-hierarchical relation aggregation, which effectively enhances the accuracy and diversity of support-query matching, and thus achieves better recognition performance.

TABLE X
COMPARISON OF SEGMENTATION RESULTS WITH MIOU AND FB-IOU ON THE LOVEDA-2^i DATASET [53]. THE RESULTS OF THE COMPARISON METHODS ARE TRANSCRIBED FROM MGANET [92]. THE BEST AND SECOND-BEST RESULTS ARE IN BOLD AND UNDERLINED, RESPECTIVELY.

Backbones	Methods	Pub. Years	1-shot					5-shot				
			Split0	Split1	Split2	Mean	FB-IoU	Split0	Split1	Split2	Mean	FB-IoU
VGG-16	*SCL* [88]	CVPR'2021	15.31	21.43	23.89	20.21	29.56	14.84	22.39	20.85	19.36	29.22
	PEFNet [15]	TPAMI'2020	16.14	24.35	31.57	24.02	34.72	15.08	26.82	30.18	24.03	38.40
	ASGNet [93]	CVPR'2021	15.93	21.60	29.08	22.20	39.79	17.33	26.24	35.75	26.44	38.41
	DCPNet [90]	IJCAI'2022	15.22	22.58	31.83	23.21	34.97	15.94	26.38	31.37	24.56	36.24
	NTRENet [89]	CVPR'2022	15.07	23.17	28.68	22.31	36.18	15.30	25.12	30.95	23.79	35.06
	DMNet [66]	TGRS'2023	19.71	26.23	30.43	25.46	45.78	25.02	35.51	33.62	31.38	51.70
	MGANet [92]	TGRS'2024	20.00	27.44	32.63	26.69	43.11	**25.09**	35.03	36.64	32.25	50.44
	HRL (Ours)	–	**23.76**	**33.51**	**34.48**	**30.58**(+5.12)	**47.66**(+1.88)	24.20	**41.84**	**40.27**	**35.44**(+4.06)	**52.26**(+0.56)
ResNet-50	*SCL* [88]	CVPR'2021	15.14	20.45	25.00	20.20	24.60	14.25	21.09	23.65	19.66	24.69
	PEFNet [15]	TPAMI'2020	17.13	22.20	26.49	21.94	33.48	15.83	25.73	24.74	22.10	34.77
	ASGNet [93]	CVPR'2021	15.91	20.21	22.33	19.48	36.39	18.38	26.29	36.34	27.00	39.59
	CyCTR [19]	NeurIPS'2021	13.17	23.43	21.99	19.53	38.47	13.81	27.40	26.15	22.45	42.17
	DCPNet [90]	IJCAI'2022	16.67	23.10	24.44	21.40	36.36	13.43	25.59	28.06	22.36	33.89
	NTRENet [89]	CVPR'2022	16.05	22.69	21.87	20.20	32.67	15.79	24.94	23.42	21.38	31.79
	DMNet [66]	TGRS'2023	19.29	25.52	31.53	25.45	43.40	24.62	33.80	33.12	30.51	50.93
	MGANet [92]	TGRS'2024	20.55	27.51	33.30	27.12	44.91	26.53	35.99	36.69	33.07	50.68
	HRL (Ours)	–	**24.18**	**34.69**	**37.78**	**32.22**(+6.77)	**49.97**(+6.57)	**30.76**	**40.52**	**41.75**	**37.68**(+7.17)	**55.67**(+4.74)

TABLE XI
COMPARISON OF SEGMENTATION RESULTS WITH RESNET-50 [82] ON THE ISAID-5^i DATASET [51].

Method	MIoU	FB-IoU	#Para.	FLOPs	FPS
PANet [11]	26.27	57.37	23.6M	2398.9G	2.7
CANet [10]	26.03	59.46	22.3M	1083.2G	3.1
SCL [88]	34.91	64.13	11.9M	642.1G	8.6
PEFNet [15]	33.59	63.25	10.8M	625.1G	1.3
NTRENet [89]	36.25	64.45	20.8M	641.4G	9.1
DCPNet [90]	34.38	63.77	11.3M	978.1G	6.0
BAM [14]	36.61	65.00	4.9M	696.4G	9.8
SDM [60]	28.27	59.90	29.3M	366.2G	11.8
R^{2}Net [63]	37.36	66.18	5.0M	1358.3G	4.6
MGANet [92]	33.35	62.94	10.1M	1126.4G	1.0
HRL (Ours)	**41.71**	**68.25**	10.5M	1461.0G	2.8

3) LoveDA-2^i dataset: We continue by comparing our HRL with state-of-the-art approaches on the LoveDA-2^i dataset [53]. The experimental results are presented in Table X, showing that the proposed HRL performs strongly and competitively. In the 1-shot setting, it exceeds the previous best method DMNet [66] by 5.12% MIoU with VGG-16 and by 6.77% MIoU with ResNet-50. In the 5-shot setting, HRL outperforms DMNet [66] with improvements of 4.06% and 7.17% in MIoU when using VGG-16 and ResNet-50 as backbone networks, respectively. Across all experimental settings, HRL achieves superior performance except for Split0 under the 5-shot setting with VGG-16, where its MIoU is slightly lower than that of MGANet [92] by 0.89%. This further demonstrates the robustness and stability of HRL in few-shot remote sensing image segmentation tasks. In contrast to the previous two datasets, LoveDA-2^i [53] has a larger scale and unfixed shape of remote sensing concepts (not limited to a single image, such as water and agriculture). The segmentation visualization of the various methods on the LoveDA-2i dataset is shown on the right side of Fig. 6. Although the "Water" area in the support image is relatively small and its color is similar to the surrounding environment in the query image, HRL is still able to accurately match and identify the target region, demonstrating its strong relationship modeling capability. In the second column, due to the diversity of cultivated crops and variations in uncultivated land, it is difficult to accurately segment the "Agricultural" area with only a single support image. Both DMNet [66] and MGANet [92] show fragmented segmentation results for this class. In contrast, HRL effectively improves the matching accuracy between support and query images through cross-hierarchy support-query relation correction and fusion, yielding segmentation results closer to the ground truth. In the third column, although the "Barren" in the support image occupies only a small portion and appears visually similar to the surrounding "Agricultural", HRL can still locate the corresponding region in the query image, further validating its strong robustness and generalization ability.

4) Comparison of Computing Performance: Table XI presents a comparison of the number of parameters, computational complexity (FLoating point OPerations, FLOPs), and inference speed (Frames Per Second, FPS) across several methods on 5-shot, showing that HRL strikes a good balance between performance and complexity. Despite being slower in inference speed than the state-of-the-art R^{2}Net [63], our HRL is 4.35% MIoU and 2.07% FB-IoU higher in segmentation accuracy. Compared to PEFNet, HRL has a similar parameter count but obtains a higher FB-IoU, and faster inference speed, underscoring its effectiveness. These results suggest that the proposed correlation learning approach holds promise for advancing FSS in RSIs.

V. CONCLUSION

In this paper, we propose Hierarchical Relation Learning (HRL) to enhance FSS for RSIs. Our efforts include both network architecture design (*i.e.*, HRNet) and training strategies (*i.e.*, BHLoss). HRNet is designed to perform both single-scale relation extraction at each hierarchy and multi-scale relation aggregation across hierarchies, generating enhanced relation representations for high-quality segmentation of the targets in the query image. Moreover, BHLoss is proposed to supervise the learning of single-scale relations, ensuring the diversity of single-scale relation representations for effective multi-scale relation aggregation. The proposed HRL achieves state-of-the-art performance on three popular remote sensing datasets. In future work, we plan to integrate large segmentation models into our HRL to develop a more powerful FSS model specifically adapted for RSIs.

References

[1] X. Lu, Y. Zhong, Z. Zheng, Y. Liu, J. Zhao, A. Ma, and J. Yang, “Multi-scale and multi-task deep learning framework for automatic road extraction,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 57, no. 11, pp. 9362–9377, 2019.

[2] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone, “Adversarial shape learning for building extraction in VHR remote sensing images,” *IEEE Trans. Image Process. (TIP)*, vol. 31, pp. 678–690, 2021.

[3] Q. Li, L. Mou, Y. Sun, Y. Hua, Y. Shi, and X. X. Zhu, “A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 4702315, 2024.

[4] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5607514, 2021.

[5] B. Song, H. Yang, Y. Wu, P. Zhang, B. Wang, and G. Han, “A multispectral remote sensing crop segmentation method based on segment anything model using multi-stage adaptation fine-tuning,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 4408818, 2024.

[6] H. Liu, W. Li, W. Jia, H. Sun, M. Zhang, L. Song, and Y. Gui, “Clusterformer for pine tree disease identification based on UAV remote sensing image segmentation,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 5609215, 2024.

[7] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, “One-shot learning for semantic segmentation,” in *Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–13.

[8] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Annu. Conf. Neur. Inform. Process. Syst. (NeurIPS)*, 2017, pp. 4077–4087.

[9] N. Dong and E. P. Xing, “Few-shot semantic segmentation with prototype learning,” in *Brit. Mach. Vis. Conf. (BMVC)*, 2018, p. 79.

[10] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5217–5226.

[11] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “PANet: Few-shot image semantic segmentation with prototype alignment,” in *Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9197–9206.

[12] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, “SG-One: Similarity guidance network for one-shot semantic segmentation,” *IEEE Trans. Cybernetics (TCYB)*, vol. 50, no. 9, pp. 3855–3865, 2020.

[13] X. Jiang, N. Zhou, and X. Li, “Few-shot segmentation of remote sensing images using deep metric learning,” *IEEE Geosci. Remote Sens. Lett. (GRSL)*, vol. 19, p. 6507405, 2022.

[14] C. Lang, G. Cheng, B. Tu, and J. Han, “Learning what not to segment: A new perspective on few-shot segmentation,” in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 8057–8067.

[15] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, “Prior guided feature enrichment network for few-shot segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 44, no. 2, pp. 1050–1065, 2020.

[16] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, “Prototype mixture models for few-shot semantic segmentation,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 763–778.

[17] H. Wang, X. Zhang, Y. Hu, Y. Yang, X. Cao, and X. Zhen, “Few-shot semantic segmentation with democratic attention networks,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 730–746.

[18] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, “Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation,” in *Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9587–9595.

[19] G. Zhang, G. Kang, Y. Yang, and Y. Wei, “Few-shot segmentation via cycle-consistent transformer,” in *Annu. Conf. Neur. Inform. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 21 984–21 996.

[20] X. Shi, D. Wei, Y. Zhang, D. Lu, M. Ning, J. Chen, K. Ma, and Y. Zheng, “Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 151–168.

[21] J. Min, D. Kang, and M. Cho, “Hypercorrelation squeeze for few-shot segmentation,” in *Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 6941–6952.

[22] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim, “Cost aggregation with 4D convolutional swin transformer for few-shot segmentation,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 108–126.

[23] H. Chen, Y. Dong, Z. Lu, Y. Yu, Y. Li, J. Han, and Z. Zhang, “Dense affinity matching for few-shot segmentation,” *Neurocomputing*, vol. 577, p. 127348, 2024.

[24] L. Cao, Y. Guo, Y. Yuan, and Q. Jin, “Prototype as query for few shot semantic segmentation,” *Complex & Intelligent Systems*, vol. 10, no. 5, pp. 7265–7278, 2024.

[25] J. Ding, Z. Zhang, Q. Wang, and H. Wang, “SCTrans: Self-align and cross-align transformer for few-shot segmentation,” *Image and Vision Computing*, vol. 142, p. 104893, 2024.

[26] Q. Zhao, S. Lyu, Y. Li, Y. Ma, and L. Chen, “MGML: Multigranularity multilevel feature ensemble network for remote sensing scene classification,” *IEEE Trans. Neur. Net. Learn. Syst. (TNNLS)*, vol. 34, no. 5, pp. 2308–2322, 2021.

[27] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, “A multilevel multimodal fusion transformer for remote sensing semantic segmentation,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, p. 5403215, 2024.

[28] Q. Wang, X. Luo, J. Feng, S. Li, and J. Yin, “CCENet: Cascade class-aware enhanced network for high-resolution aerial imagery semantic segmentation,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (J-STARS)*, vol. 15, pp. 6943–6956, 2022.

[29] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, “Super-resolution-based change detection network with stacked attention module for images with different resolutions,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 4403718, 2021.

[30] Y. Zhao, J. Liang, S. Huang, and P. Huang, “Hierarchical deep features progressive aggregation for remote sensing images scene classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (J-STARS)*, vol. 17, pp. 9442–9450, 2024.

[31] Z. Zhan, Z. Xiong, X. Huang, C. Yang, Y. Liu, and X. Wang, “Multi-scale feature reconstruction and inter-class attention weighting for land cover classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (J-STARS)*, vol. 17, pp. 1921–1937, 2023.

[32] X. He, Y. Zhou, B. Liu, J. Zhao, and R. Yao, “Remote sensing image semantic segmentation via class-guided structural interaction and boundary perception,” *Expert Syst. with Appl. (ESWA)*, vol. 252, p. 124019, 2024.

[33] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Int. Conf. Med. Image Comp. Comput.-Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[34] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, MICCAI*, 2018, pp. 3–11.

[35] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.

[36] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-Unet: Unet-like pure transformer for medical image segmentation,” in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 205–218.

[37] C. Peng, K. Zhang, Y. Ma, and J. Ma, “Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5601313, 2021.

[38] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, “Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 19, p. 8009205, 2021.

[39] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, “Multiattention network for semantic segmentation of fine-resolution remote sensing images,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5607713, 2021.

[40] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, “HED-UNet: Combined segmentation and edge detection for monitoring the Antarctic coastline,” *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 4300514, 2021.

[41] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[42] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 4408715, 2022.

[43] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5625711, 2022.

[44] Z. Lv, H. Huang, W. Sun, T. Lei, J. A. Benediktsson, and J. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett. (GRSL)*, vol. 20, p. 2505405, 2023.

[45] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 2004612, 2023.

[46] Q. Zeng, J. Zhou, J. Tao, L. Chen, X. Niu, and Y. Zhang, "Multiscale global context network for semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 5622913, 2024.

[47] B. Peng, Z. Tian, X. Wu, C. Wang, S. Liu, J. Su, and J. Jia, "Hierarchical dense correlation distillation for few-shot segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 23 641–23 651.

[48] F. Xiao, R. Liu, Y. Zhu, H. Zhang, J. Zhang, and S. Chen, "A dense multi-cross self-attention and adaptive gated perceptual unit method for few-shot semantic segmentation," *IEEE Trans. Artif. Intell. (TAI)*, vol. 5, no. 6, pp. 2493–2504, 2024.

[49] K. Ai, H. Hu, Q. Zhou, and Q. Guan, "SGT: Self-guided transformer for few-shot semantic segmentation," in *IEEE Int. Conf. Acoust. Speech SP (ICASSP)*, 2024, pp. 5935–5939.

[50] J. Li, M. Gong, W. Li, M. Zhang, Y. Zhang, S. Wang, and Y. Wu, "MS^{2}A^{2}Net: Multi-scale self-attention aggregation network for few-shot aerial imagery segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 4500216, 2024.

[51] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "iSAID: A large-scale dataset for instance segmentation in aerial images," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, 2019, pp. 28–37.

[52] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sensing*, vol. 10, no. 6, p. 964, 2018.

[53] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *NeurIPS Track on Datasets and Benchmarks*, 2021, pp. 1–12.

[54] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[56] X. Zhao, X. Chen, Z. Gong, W. Yao, Y. Zhang, and X. Zheng, "Contrastive enhancement using latent prototype for few-shot segmentation," *Digit. Signal Process.*, vol. 144, p. 104282, 2024.

[57] J. Wang, J. Li, C. Chen, Y. Zhang, H. Shen, and T. Zhang, "Adaptive FSS: A novel few-shot segmentation framework via prototype enhancement," in *AAAI Conf. Artif. Intell. (AAAI)*, 2024, pp. 5463–5471.

[58] J. Liu, Y. Bao, G.-S. Xie, H. Xiong, J.-J. Sonke, and E. Gavves, "Dynamic prototype convolution network for few-shot semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 11 553–11 562.

[59] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 142–158.

[60] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5611711, 2021.

[61] W. Ao, S. Zheng, Y. Meng, and Z. Gao, "Few-shot aerial image semantic segmentation leveraging pyramid correlation fusion," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5624512, 2023.

[62] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Few-shot rotation-invariant aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 62, p. 5600513, 2024.

[63] C. Lang, G. Cheng, B. Tu, and J. Han, "Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, pp. 1–11, 2023.

[64] C. Lang, J. Wang, G. Cheng, B. Tu, and J. Han, "Progressive parsing and commonality distillation for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5613610, 2023.

[65] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Holistic mutual representation enhancement for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5622613, 2023.

[66] H. Bi, Y. Feng, Z. Yan, Y. Mao, W. Diao, H. Wang, and X. Sun, "Not just learning from others but relying on yourself: A new perspective on few-shot segmentation in remote sensing," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5623621, 2023.

[67] 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 Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5611118, 2022.

[68] R. Zhang, W. Zhu, Y. Li, T. Song, Z. Li, W. Yang, L. Yang, T. Zhou, and X. Xu, "D-FusionNet: Road extraction from remote sensing images using dilated convolutional block," *GIScience & Remote Sensing*, vol. 60, no. 1, p. 2270806, 2023.

[69] Q. Liu, Y. Dong, Z. Jiang, Y. Pei, B. Zheng, L. Zheng, and Z. Fu, "Multi-pooling context network for image semantic segmentation," *Remote Sensing*, vol. 15, no. 11, p. 2800, 2023.

[70] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, p. 701, 2020.

[71] T. K. Behera, S. Bakshi, M. Nappi, and P. K. Sa, "Superpixel-based multiscale CNN approach toward multiclass object segmentation from UAV-captured aerial images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (J-STARS)*, vol. 16, pp. 1771–1784, 2023.

[72] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 57, no. 11, pp. 9277–9289, 2019.

[73] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 59, no. 1, pp. 426–435, 2021.

[74] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5609519, 2023.

[75] F. Zhou, R. Hang, H. Shuai, and Q. Liu, "Hierarchical context network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 4407612, 2021.

[76] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 58, no. 11, pp. 7557–7569, 2020.

[77] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 60, p. 5412012, 2022.

[78] P. Song, J. Li, Z. An, H. Fan, and L. Fan, "CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5900314, 2022.

[79] X. Zhou, L. Zhou, S. Gong, S. Zhong, W. Yan, and Y. Huang, "Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (J-STARS)*, vol. 17, pp. 175–189, 2024.

[80] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, vol. 61, p. 5605116, 2023.

[81] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Trans. Image Process. (TIP)*, vol. 32, pp. 1052–1064, 2023.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778.

[83] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Annu. Conf. Neur. Inform. Process. Syst. (NeurIPS)*, 2016, pp. 4898–4906.

[84] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly,

J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[85] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[86] L. Wang, S. Lei, J. He, S. Wang, M. Zhang, and C.-T. Lu, "Self-correlation and cross-correlation learning for few-shot remote sensing image semantic segmentation," in *ACM International Conference on Advances in Geographic Information Systems*, 2023, pp. 1–10.

[87] 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," in *Annu. Conf. Neur. Inform. Process. Syst. (NeurIPS)*, 2019, pp. 8024–8035.

[88] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 8312–8321.

[89] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 11 573–11 582.

[90] C. Lang, B. Tu, G. Cheng, and J. Han, "Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation," in *Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 1024–1030.

[91] G. Puthumanaillam and U. Verma, "Texture based prototypical network for few-shot semantic segmentation of forest cover: Generalizing for different geographical regions," *Neurocomputing*, vol. 538, p. 126201, 2023.

[92] S.-F. Peng, G.-S. Xie, F. Zhao, X. Shu, and Q. Liu, "Multi-granularity aggregation network for remote sensing few-shot segmentation," *IEEE Trans. Geosci. Remote Sens. (TGRS)*, pp. 1–1, 2024.

[93] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 8334–8343.