Embedding Generalized Semantic Knowledge into Few-Shot Remote Sensing Segmentation

Qi Wang, *Senior Member, IEEE,* Yuyu Jia, Junyu Gao, *Member, IEEE,* and Qiang Li, *Member, IEEE*

***Abstract*—Few-shot segmentation (FSS) for remote sensing (RS) imagery leverages supporting information from limited annotated samples to achieve query segmentation of novel classes. Previous efforts are dedicated to mining segmentation-guiding visual cues from a constrained set of support samples. However, they still struggle to address the pronounced intra-class differences in RS images, as sparse visual cues make it challenging to establish robust class-specific representations. In this paper, we propose a holistic semantic embedding (HSE) approach that effectively harnesses general semantic knowledge, *i.e.,* class description (CD) embeddings. Instead of the naive combination of CD embeddings and visual features for segmentation decoding, we investigate embedding the general semantic knowledge during the feature extraction stage. Specifically, in HSE, a spatial dense interaction module allows the interaction of visual support features with CD embeddings along the spatial dimension via self-attention. Furthermore, a global content modulation module efficiently augments the global information of the target category in both support and query features, thanks to the transformative fusion of visual features and CD embeddings. These two components holistically synergize CD embeddings and visual cues, constructing a robust class-specific representation. Through extensive experiments on the standard FSS benchmark, the proposed HSE approach demonstrates superior performance compared to peer work, setting a new state-of-the-art.**

***Index Terms*—Few-shot segmentation, remote sensing, semantic embedding, class description embeddings.**

I. Introduction

SEMANTIC segmentation, entailing the pixel-level categorization of images, is essential for deciphering remote sensing imagery and acts as a foundational method across a wide range of practical applications [1], [2], [3], [4], [5], [6]. With the development of aerial and satellite imaging devices, the collection of remote sensing data becomes feasible, further driving advancements in deep learning for fully supervised remote sensing image segmentation. However, deep learning-based methods [7], [8] still struggle with the high cost of extensive annotation data. Although researchers mitigate the stringent requirement for annotated data through semi-supervised [9], [10], [11], [12] and weakly supervised [13], [14], [15], [16] perspectives, the generalization ability of models remains a bottleneck when confronted with entirely new categories. As an effective approach to surmount these challenges, a plethora of few-shot segmentation (FSS) techniques [17], [18], [19], [20], [21], [22] have emerged. Given a few annotated samples from a novel class, *i.e., support*, FSS performs accurate segmentation on other samples within the same category, *i.e., query*.

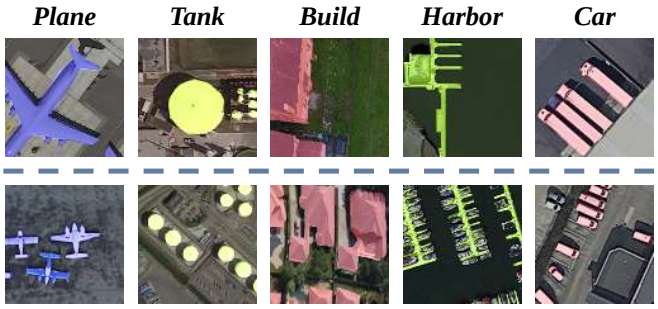


Fig. 1. Pronounced intra-class differences in remote sensing images. Even though the same object category is present in the first and second rows, their appearance, size, and distribution characteristics differ remarkably .

As Fig. 2(a) shows, FSS algorithms typically follow the segmentation-guided paradigm, exploring the extraction of class-specific visual cues from support samples and segmenting query samples through a decoder [24]. Some efforts are devoted to the segmentation decoding stage. Specifically, prototypical learning methods [25], [26], [27] employ class prototypes as stand-ins for pixel-level correlations, indirectly guiding the segmentation of query images. Affinity learning methods [28], [29], [30] directly transform pixel-level correlations into segmentation guidance information. Another school of thought mines robust and comprehensive class-specific visual representations during the feature extraction stage, serving as segmentation-guiding cues. For example, integrating multi-scale features [31], [32], [33] leveraging deep semantic correlations [34], [35], [36] and incorporating multi-level support class information [37], [38], [39], *etc*. Although showing promising results, these methods based on a single visual modality still have limitations. A minuscule number of samples often falls short in furnishing robust visual representations for segmentation guidance, especially when dealing with remote sensing images characterized by pronounced intra-class differences, as illustrated in Fig. 1.

To address the above issue, introducing general semantic knowledge, *i.e.,* class description (CD) embeddings, can supplement some missing class information in visual representations since CD embeddings are responsible only for categories and are not influenced by individual sample differences. Following this idea, AM3 [40] extracts textual prototypes

This work was supported by the National Natural Science Foundation of China under Grant 62301385, 62471394 and U21B2041.

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

E-mail: crabwq@gmail.com, jyy2019@mail.nwpu.edu.cn, gjy3035@gmail.com, liqmges@gmail.com.

Qiang Li is the corresponding author.

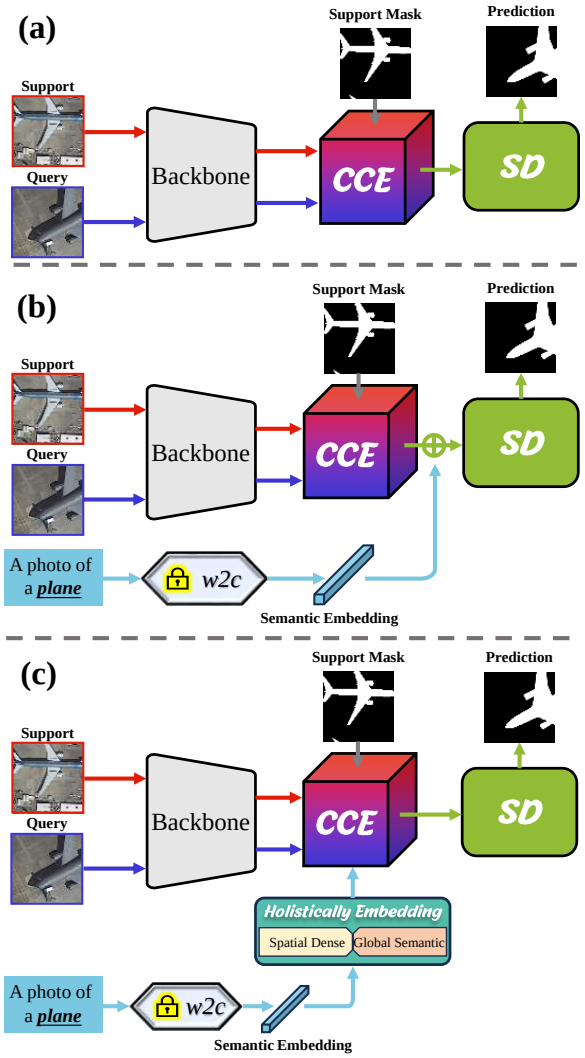


Fig. 2. Comparison between existing FSS methods and proposed HSE. (a) Many FSS algorithms adhere to a segmentation-guiding paradigm, primarily conducting research from two aspects: class-specific cues extraction (CCE) and segmentation decoder (SD). (b) Recent work, MIANet [23], introduces general semantic knowledge and combines it with visual support prototypes for segmentation guidance. (c) We further explore a holistic semantic embedding (HSE) approach that exploits the capabilities of general semantic knowledge through spatial dense interaction and global semantic modulation.

from category descriptions and combines them with visual prototypes. SP [41] modulates visual feature extraction using semantic information as prompts. However, these methods are designed for classification tasks. Recently, MIANet [23] simply supplements CD embeddings with visual representation and achieves remarkable results (Fig. 2(b)), which indicates its promising development space.

In this paper, we further explore a holistic semantic embedding (HSE) approach to harness the potential of general semantic knowledge. First, CD embeddings for each class can be obtained by a powerful pre-trained language model, such as BERT [42] and CLIP [43]. Multi-level visual features are derived through the CNN backbone, *i.e.,* ResNet50 [44] or VGG16 [45]. Subsequently, HSE is implemented as two serial complementary modules, *i.e.,* spatial dense interaction (SDI) and global content modulation (GCM), to holistically embed general semantic knowledge into the visual feature extractor (Fig. 2(c)). In particular, SDI extends the support features with general CD embeddings to facilitate the dense interaction *within the spatial dimension.* These enriched features are then processed through an interactor, which is realized using a self-attention mechanism, to guide the model to attend to spatial characteristics unique to each class. Moreover, the global content matters for comprehensively understanding the category scene. GCM converts CD embeddings and visual support prototypes into an enhancement coefficient through a modulator to enhance the support and query features *on the channel dimension*. Combining the two complementary modules, the proposed HSE approach effectively leverages the general semantic knowledge in class descriptions to mitigate intra-class individual differences in remote sensing imagery. Through extensive experiments on the standard FSS benchmarks, *i.e.,* iSAID-5^i, HSE presents significant performance improvements with different types of pre-trained language encoders and few-shot settings, realizing the efficacy of general semantic knowledge for FSS tasks.

In summary, our contributions are threefold:

1) A novel holistic semantic embedding approach is proposed to explore general semantic knowledge in class descriptions. To our best knowledge, we are the first to introduce textual modality information in few-shot remote sensing segmentation.
2) To establish robust class-specific representations for segmentation guidance, we propose two complementary modules that holistically embed semantic knowledge into the visual feature extractor from both spatial dense and global content perspectives.
3) Through comprehensive quantitative and qualitative analysis, we demonstrate the effectiveness and state-of-the-art performance of HSE on the FSS benchmark.

II. Related Works

A. Generic Semantic Segmentation

Semantic segmentation can provide pixel-level understanding for remote sensing images and has been extensively studied in recent years, driven by advancements in deep neural networks. The introduction of Fully Convolutional Networks (FCNs) [46] has pioneered the development of CNN-based methods. Following this foundation, U-Net [47] introduces an encoder-decoder structure with skip connections, enriching the feature space information. To enhance the model's multi-scale perception capability, a series of variants of atrous convolution [48], [49], [50] and adaptive pooling [51], [51], [52] techniques have emerged. Later, transformer-based methods [53], [54], [55] aim to balance local and global spatial features, effectively improving semantic segmentation performance.

Drawing from successes in natural scenes, some researchers are dedicated to the task of remote sensing semantic segmentation. Considering the broader scale variations of remote sensing objects, advanced methods predominantly employ attention mechanisms to extract global contextual information. ResU-Net [56] designs a linear attention mechanism

to approximate dot-product attention with significantly reduced computation costs, rendering the integration of attention mechanisms and CNN more flexible. Given this, MANet [57] integrates global dependencies in both channel and spatial dimensions. SLCNet [58] supervises long-range correlations through category consistency information in the ground truth segmentation map. CGGLNet [59] leverages global contextual information to address unclear boundaries and incomplete structures. However, the aforementioned methods are limited by the requirement for large-scale data training and struggle to adapt to new scenes.

B. Few-Shot Learning

Few-shot learning endeavors to discern new categories with a limited set of labeled examples. Recent investigations reveal two quintessential directions: Optimization-based methods [60], [61], [62], [63] learn a meta-learner to perform rapid adaption using a sparse set of training samples for novel categories. Metric-based methodologies [64], [65], [66], [67] cultivate a feature space wherein a suitable distance function is applied for the assessment of similarity. More recently, large-scale language models (LLMs) and vision-language models (VLMs) have propelled advancements in few-shot learning. Some superior methods ingeniously leverage textual modal knowledge to compensate for the absence of visual features [41] or to directly enhance the few-shot classifiers [40].

In the realm of remote sensing image processing, few-shot learning has likewise achieved significant progress. For example, RS-MetaNet [68] elevates the learning paradigm from individual samples to tasks through meta-training, thereby acquiring the ability to discern a metric space adept at classifying remote sensing scenes. DLA-MatchNet [69] employs the attention mechanism to discover discriminative regions. HSL-MINet [70] proposes a multiview-attention strategy designed to extract potentially shared information across various rotational perspectives of images. MVP [71] incorporates a parameter-efficient tuning method into the meta-learning framework, customizing it specifically for remote sensing images. ICSFF [72] amplifies feature representations for few-shot classification through a graph-structural feature fusion methodology. Although these methods have achieved notable success in classification tasks, they lack the capability for pixel-level understanding of remote sensing images in few-shot scenarios.

C. Few-Shot Segmentation

Few-shot semantic segmentation (FSS) aims to decode objects belonging to previously unseen categories with merely a handful of annotated samples. OSLSM [17], as a pioneering work in FSS, introduced a dual-branch segmentation paradigm. Following this, various approaches can generally be categorized into two groups according to their segmentation decoding strategies: prototypical learning methods [25], [73], [22] and affinity learning methods [28], [30], [74].

In the latest research, a state-of-the-art endeavor, MIANet [23] leverages the general semantic knowledge in category descriptions to enhance the representation of support samples, achieving promising results. However, inherent differences between remote sensing and natural images—such as pronounced intra-class variations, larger scene extents, and richer elements—impede MIANet's effective performance in the few-shot segmentation of remote sensing images.

Similarly, excellent research in few-shot remote sensing segmentation has increasingly emerged. For instance, the widely popular dataset iSAID-5^i is introduced in [75]. DMML-Net [76] constructs a deep feature pyramid comparison module and conceptualizes segmentation as a metric-based pixel classification task. PCNet [77] progressively parses the entirety of the target region into local descriptors and designs base-class memories to distillate prototypes for novel classes. HMRE [78] enhances the target information within the dual-branch network structure through global semantic and spatially dense mutual reinforcement. R2Net [79] implements dynamic global prototypes to mitigate inaccurate activations arising from intra-class differences. MS2A2Net [80] proposes an attention aggregation network to adaptively fuse multi-scale features and effectively model the foreground correlation of the dual-branch structure. However, these methods are limited to extracting visual segmentation cues from a finite number of support samples and struggle with the significant intra-class differences in remote sensing images. This paper pioneers the incorporation of textual modal information from category descriptions in this task, addressing the shortfall of support information and providing fresh insights for this research.

III. Methodology

A. Task Description

Few-shot segmentation seeks to generalize a model's segmentation capability from a training dataset $\mathcal{D}_{train}$ to a completely novel category dataset $\mathcal{D}_{novel}$, corresponding to two disjoint sets of categories, $\mathcal{C}_{train}$ and $\mathcal{C}_{novel}$. Consistent with prior studies [26], we train the model episodically, where each episode includes a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. Given the K-shot setting, each support set comprises K pairs of images I_s and their corresponding binary masks M_s, where $\mathcal{S} = \left\{I_s^k, M_s^k\right\}_{k=1}^K$. Similarly, the query set is defined as $\mathcal{Q} = \{I_q, M_q\}$, where the mask M_q of the query image is only available at the model training stage. Our goal is to mine robust segmentation-guiding information from the support set to accurately parse the query images. Note that we employ the 1-shot setting to simplify the illustration of our approach.

B. Method Overview

As depicted in Fig. 3, the proposed HSE approach encompasses an initial feature extraction stage III-C and two key complementary modules, the Spatial Dense Interaction (SDI) module III-D and the Global Content Modulation (GCM) module III-E, that explore general semantic knowledge from category descriptions, constructing robust class-specific representations for segmentation guidance. Specifically, given the support image I_s and the query image I_q, two pre-trained backbones (ResNet50 [44] or VGG16 [45]) with shared weights are adopted to extract mid- and high-level features.

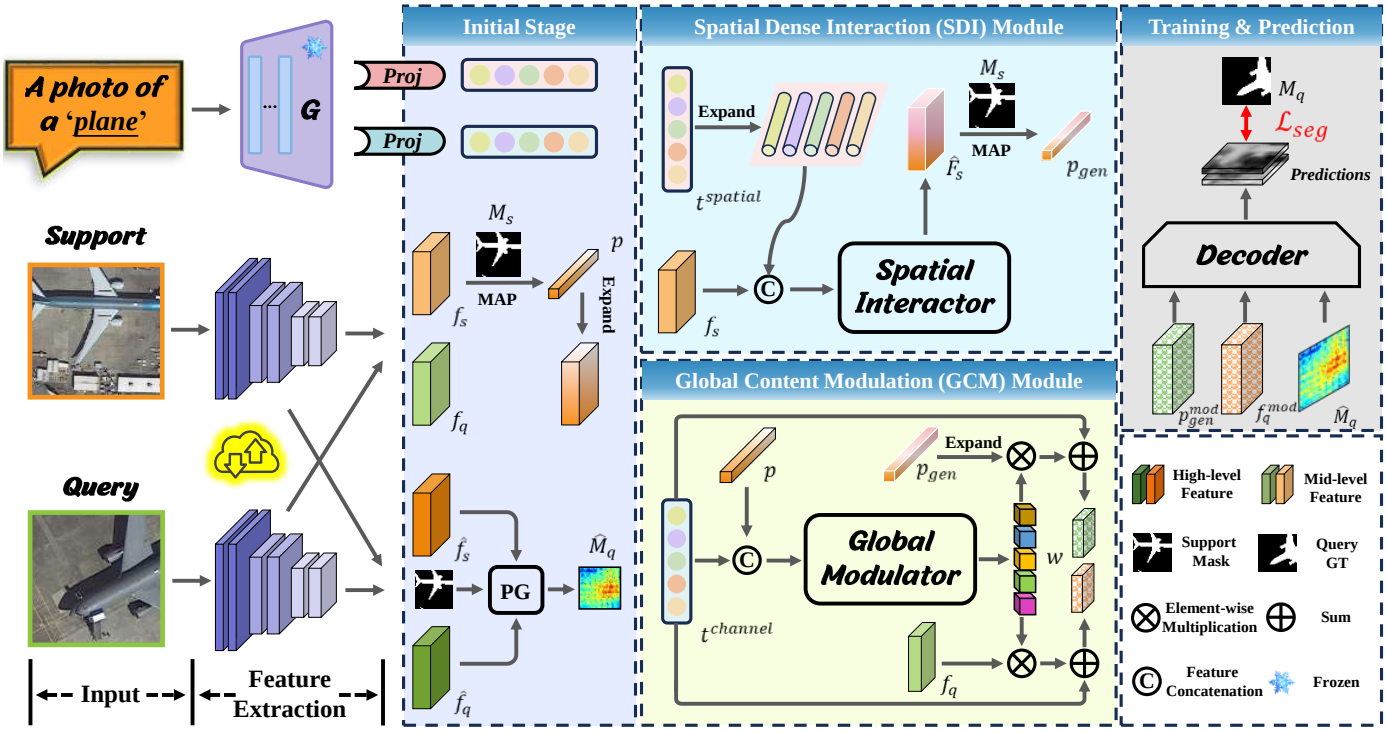


Fig. 3. Pipeline of the HSE method: it first extracts mid- and high-level support and query features, query prior masks, and CD embeddings in the initial feature extraction stage. To embed the general semantic knowledge from CD embeddings into visual cues and establish robust class-specific segmentation guidance, we design two sequential, complementary modules. The SDI module facilitates the spatial dense interaction between general semantic knowledge and individual-specific visual features. The GCM module enhances global content relevant to the target category in the support and query features through modulation coefficients. Finally, along with the modulated query feature and query prior mask, the constructed robust class-specific representation assumes the role of segmentation guidance inputted into the decoder, yielding the query prediction mask.

We then obtain the class description (CD) embedding, support prototype, and query prior mask in the initial feature extraction stage. Subsequently, the holistic embedding of general semantic knowledge is executed in two sequential modules, *i.e.,* SDI and GCM. At last, we feed the robust class-specific guidance, the modulated query feature, and the query prior mask into a segmentation decoder, obtaining the prediction query mask.

C. Initial feature Extraction Stage

The core issue in few-shot segmentation is how to transform the provided support image and its mask into class-specific segmentation guidance information. Among a plethora of studies, the predominant strategy is to compress the foreground region features of support images into prototypes, which are then used to guide the segmentation process of query images. To elaborate, background information in the middle-level support features is masked out, leaving only foreground information, which is then concentrated in the prototypes through average pooling. Formally, given the mid-level support feature $f_s \in \mathbb{R}^{C\times H\times W}$ and the corresponding support mask $M_s \in \mathbb{R}^{H\times W}$, where C, H, W represent the channel, height, and weight, the prototype $p \in \mathbb{R}^{C\times 1\times 1}$ can be expressed as:

$$p = \mathcal{A}(f_s \otimes \mathcal{R}(M_s)), \tag{1}$$

where $\mathcal{A}(\cdot)$ represents the average pooling operation, $\otimes$ denotes the Hadamard product, and $\mathcal{R}$ is a spatial scaling function.

Regarding the utilization of high-level features, the esteemed PFENet [26] transforms them into a class-agnostic prior mask to roughly indicate the target area's location in the query image. The prior mask $\hat{M}_q \in \mathbb{R}^{H\times W}$ is derived from the maximal element-wise correlation responses between the high-level features of the query and support. We abstract this process as a prior mask generator $\mathcal{PG}$, which takes support masks M_s, high-level features $\hat{f}_s$, and $\hat{f}_q$ as inputs. This can be represented as:

$$\hat{M}_q = \mathcal{PG}(\hat{f}_s, \hat{f}_q, M_s). \tag{2}$$

Moreover, with the pre-trained language model $\mathcal{G}(\cdot)$, we transform the category *name* into the CD embedding $t = \mathcal{G}(name) \in \mathbb{R}^{C_t\times 1\times 1}$, where C_t is the dimension of the embedding space.

D. Spatial Dense Interaction

Existing studies revolve around the two-level segmentation guidance cues, *i.e.,* p and $\hat{M}_q$, which are finally passed to the decoder along with the mid-level query features for segmenting the target region. However, they still struggle to break through the limitations imposed by the single visual modality, especially in the face of remote sensing images that exhibit particularly substantial intra-class differences. Essentially, the segmentation guidance cues p and $\hat{M}_q$ are both extracted from sparse, individual-specific visual features. They lack sufficient general category information to construct robust class-specific representations, rendering it arduous to efficiently encode

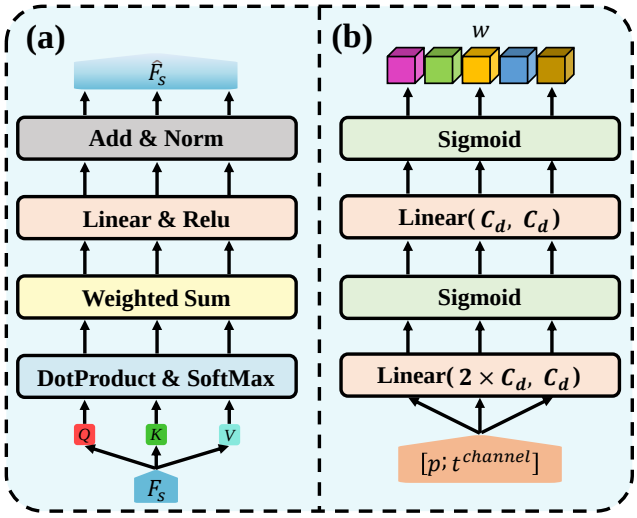


Fig. 4. Structure of the SDI (a) and GCM (b) modules.

support-query image pairs with visual disparities. In contrast, individual-agnostic CD embeddings encompass general semantic knowledge. Building on this concept, we propose a spatial dense interaction (SDI) module to embed the semantic knowledge from the spatial dimension to supplement missing general information in visual cues.

Inspired by prompt learning methods [81], [82], we concatenate CD embeddings with visual feature maps and further promote the interaction between the two through a self-attention interactor (Fig. 4 (a)), achieving semantic knowledge embedding. Given the mid-level support feature f_s of a certain category and its CD embedding $t \in \mathbb{R}^{C_t \times 1 \times 1}$, we obtain a new feature map $F_s \in \mathbb{R}^{C \times (HW+W)}$ by extending f_s with the projected CD embedding:

$$F_s = [f_s; repeat(t^{spatial})], \tag{3}$$

where $t^{spatial} = \mathcal{P}_{spatial}(t)$, $\mathcal{P}_{spatial}(\cdot)$ is a projector that maintains the channel dimension to be the same as the visual feature, $[\cdot;\cdot]$ denotes the feature concatenation operation, and *repeat*$(\cdot)$ expands the dimension through repetition operations.

Then, the extended feature map F_s is fed into the *interactor*, in which a self-attention block allows the flow of general semantic information to visual features. Formally, three input vectors $\boldsymbol{q}$, $\boldsymbol{k}$, $\boldsymbol{v}$ of the attention block have the same magnitude as F_s, the output feature map $\hat{F}_s \in \mathbb{R}^{C \times H \times W}$ can be calculated by:

$$\hat{F}_s = Resize(interactor(F_s, F_s, F_s)[:, :HW]), \tag{4}$$

where *Resize*$(\cdot)$ restores the spatial dimension $H \times W$ of the feature map.

Finally, we further compute the general support prototype p_{gen} supplemented with the semantic knowledge as:

$$p_{gen} = \mathcal{A}(\hat{F}_s \otimes \mathcal{R}(M_s)). \tag{5}$$

E. Global Content Modulation

Features from channels often contain rich global semantic content, which is closely linked to the contextual understanding of the target scene. Therefore, besides embedding general semantic knowledge, *i.e.,* CD embeddings, within the spatial dimension, it is equally crucial to modulate visual features on a channel-by-channel basis.

Given the visual support prototype $p \in \mathbb{R}^{C \times 1 \times 1 \times 1}$ of a certain category and its CD embedding $t \in \mathbb{R}^{C_t \times 1 \times 1}$, we first convert the two to an enhancement coefficient by a modulator:

$$w = modulator[p; t^{channel}], \tag{6}$$

where $w \in \mathbb{R}^{C \times 1 \times 1}$, $t^{channel} = \mathcal{P}_{channel}(t)$, $\mathcal{P}_{channel}(\cdot)$ is a projector that keeps the channel dimension consistent with the visual prototype, and the structure of the *modulator* is illustrated in Fig. 4 (b).

Then, the general prototype and query feature can be modulated by the enhancement coefficient:

$$\begin{cases} p_{gen}^{mod} = p_{gen} \otimes w + t^{channel} \\ f_q^{mod} = f_q \otimes w + t^{channel} \end{cases}. \tag{7}$$

Through such embedding of channel-level general semantic knowledge, the global information specific to the target categories contained within the support and query branches is accentuated, and the interference caused by intra-class individual differences is effectively suppressed.

F. Prediction and Training Loss

We employ the commonly used encoder-decoder structure in segmentation tasks [34], [35]. Having obtained the modulated general prototype p_{gen}^{mod} and modulated query feature f_q^{mod} from the aforementioned encoder (Section III-C-III-E), we aggregate them into the segmentation decoder $\mathcal{D}(\cdot)$ to attain the final prediction $Y \in \mathbb{R}^{H \times W}$:

$$Y = \mathcal{D}([p_{gen}^{mod}; f_q^{mod}]). \tag{8}$$

As p_{gen}^{mod} incorporates a holistic embedding of general semantic knowledge across both spatial dense and global content perspectives, it is equipped to provide robust segmentation guidance for the query image. The segmentation loss is computed utilizing the standard binary cross-entropy (BCE) function:

$$\mathcal{L} = \mathcal{D}([p_{gen}^{mod}; f_q^{mod}]). \tag{9}$$

G. Extending to K-Shot Setting

In the K-shot scenario, K support pairs $\{I_s^k, M_s^k\}_{k=1}^K$ are available to provide visual segmentation guidance. Our method requires only minimal adjustments to address the K-shot setting. In the initial feature extraction stage, the K visual prototypes and query prior masks computed by *Eq.* 1 and 2 are averaged. In addition, K support samples spatial densely interact with their respective CD embeddings, yielding $\{p_{gen}^k\}_{k=1}^K$. The final p_{gen} is obtained by averaging these values as follows:

$$p_{gen} = \frac{1}{K} \sum_{k=1}^{K} p_{gen}^k. \tag{10}$$

TABLE I
CATEGORY DISTRIBUTION AND SPLITTING OF iSAID-5^i DATASET.

Split ID	iSAID-5^0					iSAID-5^1					iSAID-5^2				
Category ID	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Category name	ship	storage tank	baseball diamond	tennis court	basketball court	ground track field	bridge	large vehicle	small vehicle	helicopter	swimming pool	roundabout	soccer ball field	plane	harbor
Num of imgs in train set	3820	902	495	2485	598	1405	224	2953	2592	69	147	203	2000	1864	3358
Num of imgs in test set	1392	269	221	767	160	517	91	789	781	21	45	45	682	990	1361

TABLE II
PERFORMANCE COMPARISON ON iSAID-5^i DATASET UNDER 1-SHOT AND 5-SHOT SETTINGS. BOLDED VALUES INDICATE THE BEST PERFORMANCE, WHILE THE SECOND-BEST PERFORMANCES ARE UNDERLINED.

Method	1-shot				5-shot			
	Split0	Split1	Split2	Mean	Split0	Split1	Split2	Mean
ResNet50								
PANet [21]	27.56	17.23	24.60	23.13	36.54	16.05	26.22	26.27
CANet [22]	25.51	13.50	24.45	21.15	29.32	21.85	26.91	26.03
SCL [38]	34.78	22.77	31.20	29.58	41.29	25.73	37.70	34.91
PFENet [26]	35.84	23.35	27.20	28.80	42.42	25.34	33.00	33.59
NERTNet [83]	34.93	23.95	28.56	29.15	44.83	26.73	37.19	36.25
DCP [27]	37.83	22.86	28.92	29.87	41.52	28.18	33.43	34.38
BAM [84]	39.43	21.69	28.64	29.92	43.29	27.92	38.62	36.61
MIANet [23]	41.43	27.26	35.70	34.80	45.75	32.34	45.21	41.10
DMML [76]	28.45	21.02	23.46	24.31	30.61	23.85	24.08	26.18
DML [85]	32.96	18.98	26.27	26.07	33.58	22.05	29.77	28.47
SDM [33]	27.96	21.99	27.82	25.92	28.50	25.23	31.07	28.27
SD-AANet [86]	30.52	29.31	18.46	26.10	28.01	21.38	16.79	22.06
TBPN [87]	29.33	16.84	25.47	23.88	30.98	20.42	28.07	26.49
HMRE [70]	40.88	32.88	37.22	36.99	42.41	34.69	**46.07**	41.06
R^2Net [79]	41.22	21.64	35.28	32.71	46.45	25.80	39.84	37.36
PCNet [77]	40.24	24.64	31.31	32.06	45.31	28.19	37.36	36.95
HSE(*ours*)	**44.03**	**35.87**	**39.26**	**39.72**	**47.03**	**36.75**	45.66	**43.15**
VGG16								
PANet [21]	26.86	14.56	20.69	20.70	30.89	16.63	24.05	23.86
CANet [22]	13.91	12.94	13.67	13.51	17.32	15.07	18.23	16.87
SCL [38]	25.75	18.57	22.24	22.19	35.77	24.92	32.70	31.13
PFENet [26]	28.52	17.05	18.94	21.50	37.59	23.22	30.45	30.42
NERTNet [83]	25.78	20.01	19.88	21.89	38.43	24.21	28.99	30.54
DCP [27]	28.17	16.52	22.49	22.39	39.65	22.68	29.93	30.75
BAM [84]	33.93	16.88	21.47	24.09	38.46	22.76	28.81	30.01
MIANet [23]	35.56	25.89	34.17	31.87	41.33	24.72	40.42	35.49
DMML [76]	24.41	18.58	19.46	20.82	28.97	21.02	22.78	24.26
DML [85]	30.99	14.60	19.05	21.55	34.03	16.38	26.32	25.58
SDM [33]	24.52	16.31	21.01	20.61	26.73	19.97	26.10	24.27
SD-AANet [86]	28.01	21.38	16.79	22.06	34.89	25.76	16.81	25.82
TBPN [87]	27.86	12.32	18.16	19.45	32.79	16.28	24.27	24.45
HMRE [70]	36.48	28.28	35.36	33.37	37.33	29.28	43.36	36.66
R^2Net [79]	35.27	19.93	24.63	26.61	**42.06**	23.52	30.06	31.88
PCNet [77]	32.48	19.88	24.56	25.64	41.09	21.98	34.14	32.40
HSE(*ours*)	**37.18**	**29.79**	**36.04**	**34.34**	40.87	**30.20**	**43.44**	**38.17**

IV. EXPERIMENTS

To ascertain the efficacy of the proposed HSE method, both 1-shot and 5-shot settings are adhered to, conducting ample quantitative and qualitative experiments on the publicly available FSS dataset. First, we delineate the FSS benchmark and the corresponding evaluation metric. Then, the implementation specifics of HSE are elucidated to facilitate realization. Following this, we present comparisons between the HSE method and other leading counterparts, analyzing the segmentation results comprehensively. Ultimately, a series of ablation studies are executed to demonstrate the impact of each component within the HSE framework.

A. Setup

1) Dataset: Despite the recent emergence of a few FSS datasets for remote sensing images, they still lack a unified partitioning standard, making it difficult to fairly compare the performance of different algorithms. In response, we adopt the iSAID-5^i dataset [75] widely used by classical algorithms. It encompasses 15 categories, with 18076 images designated for training and 6363 images for testing. Each sample is dimensioned at 3×256×256. The specific distribution of categories and their partitioning are illustrated in Table I.

2) Evaluation Metric: Building upon prior research in the domain [75], [77], we utilize the mean Intersection over Union

TABLE III
PERFORMANCE COMPARISON ACROSS SPECIFIC CATEGORIES ON iSAID-5^i DATASET UNDER THE 1-SHOT SETTING.

Category ID	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
ResNet50															
PANet [21]	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
CANet [22]	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
SCL [38]	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
PFENet [26]	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
NERTNet [83]	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
DCP [27]	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
BAM [84]	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
DMML [76]	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
DML [85]	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
SDM [33]	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
SD-AANet [86]	25.21	39.85	**43.60**	21.69	22.27	**35.18**	**47.28**	23.89	25.47	14.75	9.46	27.38	20.51	12.50	22.45
TBPN [87]	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
HMRE [70]	41.68	41.15	40.01	41.04	40.52	27.36	34.66	35.44	**39.34**	27.58	24.56	49.34	**46.82**	4.27	21.11
R^2Net [79]	**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
HSE(*ours*)	45.67	46.21	43.34	41.02	**43.91**	33.40	40.08	**36.13**	39.30	30.44	28.58	**50.61**	45.89	43.50	**27.72**
VGG16															
PANet [21]	20.05	37.71	21.18	41.22	14.15	12.17	13.82	21.05	7.89	17.88	4.36	31.68	27.55	26.88	12.97
CANet [22]	24.13	6.73	13.83	16.32	8.54	14.12	3.24	21.04	3.35	22.96	9.57	14.91	17.83	16.11	9.92
SCL [38]	28.50	32.93	19.68	29.60	18.05	22.48	7.92	31.46	8.99	22.02	14.17	16.53	19.72	**39.40**	21.37
PFENet [26]	34.32	31.81	24.20	35.43	16.86	13.98	6.01	31.68	6.76	**26.85**	8.15	17.75	20.56	33.34	14.87
NERTNet [83]	12.66	23.11	26.90	50.47	15.77	23.14	8.48	31.73	11.75	24.94	14.64	20.45	29.03	28.06	7.24
DCP [27]	27.69	38.45	25.92	33.20	15.57	17.62	12.36	26.79	8.05	17.80	22.45	18.29	18.03	37.57	16.10
BAM [84]	27.66	43.90	31.48	43.96	22.66	13.57	8.91	31.76	9.26	20.91	17.05	26.27	30.68	25.27	8.07
DMML [76]	34.75	37.36	15.15	22.85	11.94	21.41	13.85	23.92	10.24	23.50	8.17	16.32	21.08	29.63	22.09
DML [85]	27.30	42.63	19.25	**50.63**	15.13	14.16	15.94	22.40	7.74	12.74	3.79	23.73	23.47	27.40	16.88
SDM [33]	33.76	23.88	17.80	27.76	19.38	18.36	9.63	25.24	8.63	19.69	10.56	15.36	24.76	32.30	22.06
SD-AANet [86]	18.72	32.01	40.17	23.59	25.58	13.79	**48.74**	13.67	14.93	15.79	8.36	22.03	18.86	11.05	23.66
TBPN [87]	22.03	39.75	20.80	42.80	13.94	10.41	6.87	16.54	4.38	23.41	5.68	23.66	22.13	24.63	14.72
HMRE [70]	30.25	36.16	**41.88**	43.37	30.74	31.52	36.66	**35.10**	18.38	19.74	9.74	**45.09**	**49.22**	30.82	21.93
R^2Net [79]	**37.82**	**45.16**	26.27	45.30	21.81	24.11	14.38	30.92	12.21	18.03	18.66	25.02	29.64	31.95	17.87
HSE(*ours*)	37.56	40.21	40.65	36.20	**31.28**	**34.30**	38.24	33.26	**20.01**	23.14	**32.70**	43.45	47.63	30.89	**25.53**

(mIoU) as the evaluation metric. Formally, let N be the number of test (unseen) categories in each target fold, and the IoU of category c is represented by IoU_c, the $mIoU$ can be defined as:

$$mIoU = \frac{1}{N}\sum_{c=1}^{N} IoU_c. \tag{11}$$

B. Implementation Details

HSE is constructed using the PyTorch [88] framework, and all experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. We leverage ResNet50 [44] and VGG16 [45] pre-trained on ImageNet [89] as feature extractors. The baseline process involves removing the SDI and GCM modules from the proposed HSE, which directly combines the support prototype, query feature, and query prior mask [26], and then feeding them into the segmentation decoder. To preserve the generalizability of the feature extractor, we follow [26] to freeze its parameters. To derive general semantic knowledge from class name embeddings, three types of language models pre-trained on large-scale corpora are utilized, *i.e.,* CLIP [43], SBERT [90], and GloVe [91]. We augment the category name with a text template as input: *A photo of a 'category name'*.

During meta-training, we employ the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.005, a momentum of 0.9, and a weight decay of 0.0001. The batch size is set at 16, and the training epoch is fixed at 100.

During the meta-testing phase, we randomly sample 2000 episodes for model evaluation, with the final results obtained by averaging the outcomes of experiments influenced by five different random seeds.

C. Comparison With State-of-the-Art

In this section, we compare the proposed HSE method against other advanced FSS algorithms with the available source codes or public experimental results on the iSAID-5^i dataset. Under standardized experimental conditions, the quantitative and qualitative comparison results are presented to rigorously validate the superiority of HSE.

1) Quantitative Analysis: The performance comparison of our proposed HSE against other exemplary FSS algorithms is depicted in Table II, mIoU is leveraged as the evaluation metric under 1-shot and 5-shot settings. The best performance values are highlighted in bold, and the second-best results are marked with an underline. Overall, the proposed HSE method achieves the best segmentation performance across various backbones and support sample sizes. Specifically, with the ResNet50 backbone, HSE achieves the best performance in every split, enhancing the average mIoU by 2.73% under 1-shot settings. For the 5-shot setting, HSE exhibits a distinct advantage, achieving an average performance of 43.15%, representing an improvement of 2.09%. Simultaneously, employing the VGG16 backbone, HSE achieved the best performance for

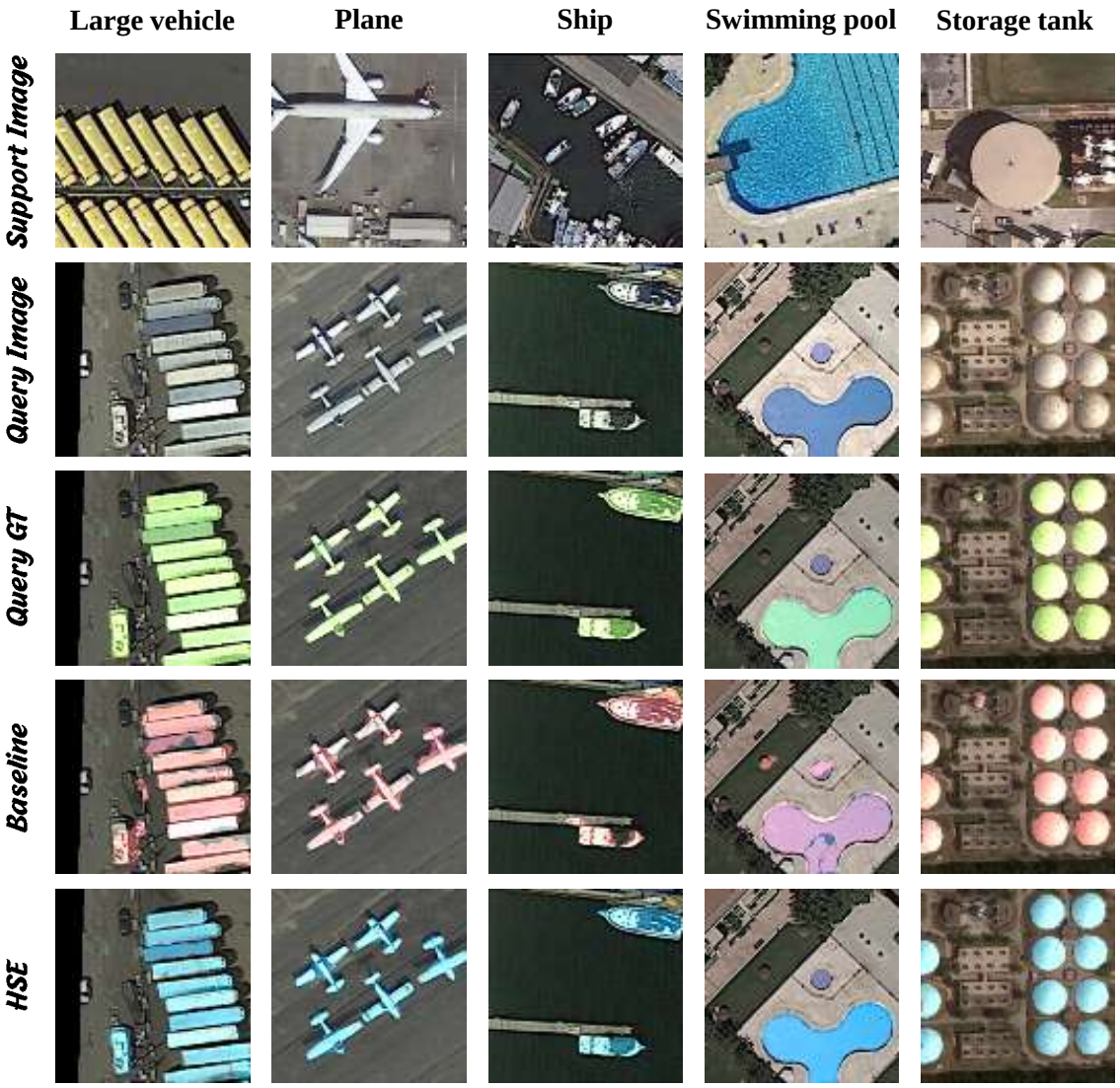


Fig. 5. Quantitative comparison of the segmentation effects of the HSE and baseline methods on iSAID-5^i dataset under the 1-shot setting.

each split under the 1-shot setting, improving the average performance by 0.97%. Like the 5-shot setting results, HSE led in both split1 and split2, surpassing the HMRE method by an average mIoU of 1.51%. In terms of comprehensive analysis, the proposed HSE method demonstrates significant advantages in FSS tasks, particularly excelling in 1-shot scenarios. This effectively validates that incorporating general semantic knowledge is beneficial for establishing robust class-specific segmentation guidance.

To specifically analyze the model's performance across various categories, we detail the segmentation results under the 1-shot setting in Table. III. The proposed HSE achieves excellent performance across the majority of categories. While some methods get better segmentation results in specific categories, their performance is not consistently stable. For example, the BAM algorithm performs well in segmenting tennis court (C4), but it struggles with smaller-sized targets such as harbor (C15). The DML method also excels at parsing tennis court (C4), yet it encounters difficulties with the small vehicle category (C9). In contrast, our method still maintains good

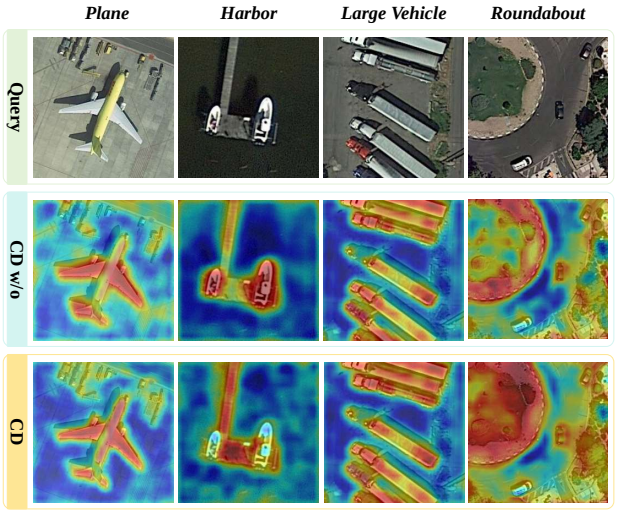


Fig. 6. Visualization of the impact of introducing CD embeddings on the attention map.

TABLE IV
ABLATION STUDY FOR EACH COMPONENT OF HSE UNDER THE 1-Shot Setting, with the Best Results Highlighted in Bold.

Methods	1-shot				Params	Inference time
	Split0	Split1	Split2	Mean		
Baseline	38.25	30.88	35.06	34.73	26.3M	34.78ms
Baseline+SDI	42.55	34.2	37.98	38.24	26.7M	35.25ms
Baseline+GSM	41.01	33.17	36.66	36.95	26.6M	35.68ms
Baseline+GSM+SDI	**44.03**	**35.87**	**39.26**	**39.72**	27.0M	36.15ms

TABLE V
IMPACT OF DIFFERENT LANGUAGE MODELS ON THE QUALITY OF GENERAL SEMANTIC KNOWLEDGE UNDER THE 1-SHOT SETTING.

Language model	1-shot			
	Split0	Split1	Split2	Mean
SBERT	43.58	35.26	38.51	39.12
GloVe	43.65	**35.90**	39.14	39.56
CLIP	**44.03**	35.87	**39.26**	**39.72**

TABLE VI
IMPACT OF DIFFERENT PROJECTOR STRUCTURES ON SEGMENTATION PERFORMANCE UNDER THE 1-SHOT SETTING.

Projector	1-shot			
	Split0	Split1	Split2	Mean
Linear	44.03	**35.87**	39.26	39.72
2-layers-MLP	**44.15**	35.73	**39.45**	**39.78**
3-layers-MLP	43.34	34.26	37.52	38.37

TABLE VII
IMPACT OF DIFFERENT TEXT DESCRIPTIONS ON SEGMENTATION PERFORMANCE UNDER THE 1-Shot Setting.

Text descriptions	1-shot			
	Split0	Split1	Split2	Mean
A photo of a {*category*}	43.52	35.82	38.78	39.37
A scene of a {*category*}	43.89	35.25	38.61	39.25
There is a {*category*} in the image	43.87	35.66	**39.40**	39.64
{*category*}	**44.03**	**35.87**	39.26	**39.72**

segmentation capabilities for challenging categories, exhibiting a more stable overall performance. This indicates that HSE has a stronger adaptability to complex scenarios, benefiting from the general CD embedding that compensates for the lack of sparse visual representation information.

2) Qualitative Analysis: To visually analyze the efficacy of the proposed HSE, Fig. 5 illustrates quantitative segmentation results in the 1-shot scenario, where the baseline approach represents HSE with the GCM and SDI components removed. Compared with the baseline method, HSE accurately activates target scene areas, generating prediction masks with more finely detailed edges. Moreover, HSE effectively reduces false activations in query image areas similar to the target category. Particularly for the "ship" and "swimming pool" categories, the baseline method misclassifies similar class regions in the background, while HSE effectively suppresses the aforementioned interference.

We visualize the impact of introducing CD embeddings on

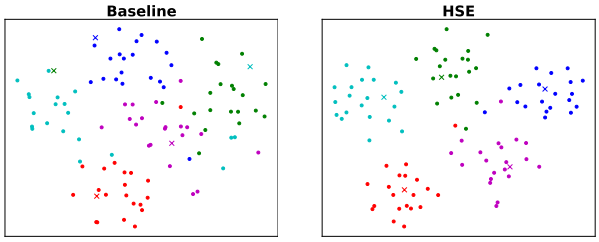


Fig. 7. t-SNE visualization of the support and query features. ● represents the query features of different categories, while × marks indicate the corresponding support features for each category.

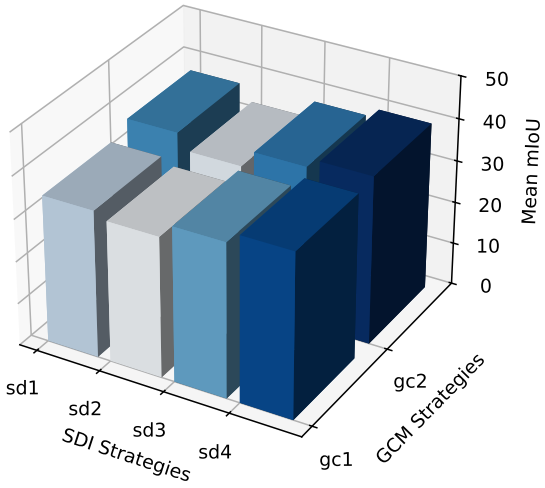


Fig. 8. Ablation studies on the different designs of SDI and GCM modules under the 1-shot setting.

the attention map in SDI. The second row in Fig. 6 represents self-attention applied solely to visual features f_s:

$$\hat{f}_s = interactor(f_s, f_s, f_s). \tag{12}$$

The second row shows the attention map $\hat{F}_s$ with the introduction of CD embeddings, as in *Eq.* 4. The introduction of CD embeddings enhances the visual information shortfall by steering the model's focus toward spatial features specific to each class, thus providing more reliable segmentation guidance.

Additionally, Fig. 7 illustrates the t-SNE [92] visualization results under the 5-way 1-shot 20-query setting. The baseline method is detailed in Section IV-B. It is evident that the proposed method significantly improves intra-class consistency. Furthermore, introducing the CD embedding enhances the distribution of the support representation (segmentation guidance), allowing it to better cover intra-class differences.

D. Ablation Studies

1) Component Analysis: To verify the effectiveness of each module within the proposed HSE framework, we organize component-level ablation experiments. As shown in Table IV, the SDI and GCM modules contribute significantly to performance, each enhancing mean mIoU by 3.51% and 1.55% respectively. This demonstrates that embedding general semantic knowledge from both spatial dense and global content

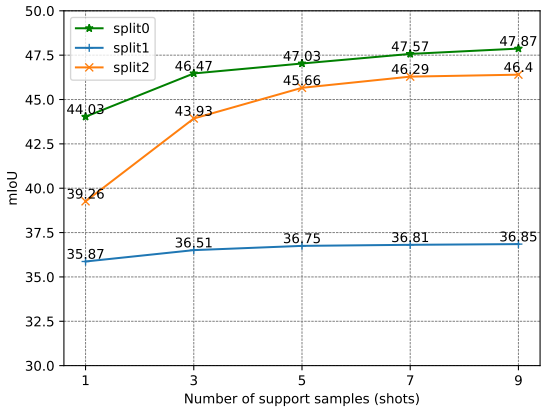


Fig. 9. Impact of the number of support samples on segmentation performance under the 1-shot setting.

perspectives is beneficial for establishing superior segmentation guidance. More importantly, the two modules complement each other, synergizing to enhance the performance of the proposed HSE further.

Regarding resource overhead, since the CLIP text encoder is only involved in a one-time computation, its parameters (151M) and time consumption (0.839s) are excluded from the overall evaluation. It can be observed that our designed SDI and GCM modules are highly efficient in computation while introducing only a minimal increase in model parameters.

2) Language Model Selection: To explore the impact of general semantic knowledge produced by different language models on segmentation performance, Table V displays the results using pre-trained SBERT, GloVe, and CLIP models respectively in the 1-shot scenario. It is evident that HSE effectively accommodates different language models and consistently achieves excellent segmentation performance. Overall, CLIP holds a slight advantage, likely because its pre-training regime better aligns visual and linguistic concepts. We select CLIP as the default language model in other experiments.

3) Various Textual Descriptions: Table VII presents the impact of embedding different textual descriptions on model performance under the 1-shot setting. Unlike natural images, where detailed template descriptions generally enhance performance, we observed that using simple category names yields better results in remote sensing imagery. We attribute this to the inherent complexity of remote sensing scenes, which encompass diverse object elements. Describing a specific category as a scene or image tends to be inaccurate, thereby diminishing the effectiveness of textual knowledge embedding.

4) Projector Structure: In the design of the SDI and GCM modules, CD embeddings t are mapped to $t^{channel}$ and $t^{spatial}$ using $\mathcal{P}_{channel}(\cdot)$ and $\mathcal{P}_{spatial}(\cdot)$ respectively. To identify the optimal projector, comparative experiments are conducted with configurations of *Linear*, *2-layers-MLP*, and *3-layers-MLP*. From Table VI, it can be observed that the structure of the projector has a minimal impact on the model's segmentation performance. Although the *2-layers-MLP* exhibits a slight advantage, considering the overhead of model parameters, we choose the *Linear* as the default projector.

5) Design of the SDI and GCM Modules: We investigate the impact of various implementation strategies for the SDI and GCM modules on segmentation performance. We evaluate three spatial dense interaction strategies (*sd1*, *sd2*, *sd3*, and *sd4*) and two global content modulation strategies (*gc1* and *gc2*). Specifically, *sd1* represents the addition of the mid-level support feature f_s and the projected CD embedding $t^{spatial}$, *sd2* denotes the element-wise multiplication of f_s and $t^{spatial}$, *sd3* stands for concatenating the visual prototype p and $t^{spatial}$ along the channel dimension, followed by two fully connected layers as in MIANet [23], and *sd4* involves concatenating f_s and $t^{spatial}$ followed by processing with self-attention, as formulated in Section III-D. *gc1* represents modulating the general prototype p_{gen} and the mid-level query feature f_q using only the enhancement coefficient w, while *gc2* employs a residual modulation approach, as described in Section III-E. Fig. 8 displays the results under different combinations of implementation strategies, where the optimal performance is obtained with the combination of *sd4* and *gc2*. This demonstrates the superiority of the methods proposed in this paper.

6) The Number of Support Samples: In Fig. 9, with the ResNet50 backbone, we illustrate the impact of varying numbers of support samples on the segmentation results for the 1-shot task. As the number of support samples increases from 1 to 9, the incorporation of richer visual representations leads to a gradual improvement in performance. However, when the number of support samples exceeds 5, the incremental improvement in segmentation performance becomes marginal. This may be due to the redundancy of target-category-relevant information contained in the numerous support samples. Additionally, as visual representations dominate the segmentation guidance, the benefits brought by the introduced generalized semantic knowledge are gradually diminished.

V. Conclusion

In this paper, we proposed the HSE method for the FSS task in remote sensing imagery. Addressing the particularly severe issue of intra-class differences in remote sensing images, we highlight that relying solely on sparse visual support representations is insufficient to establish robust segmentation guidance. Based on this observation, as the first attempt to introduce general semantic knowledge from CD embeddings into this field, we designed two complementary and sequential modules: SDI and GCM. They respectively supplement the missing target scene spatial information in sparse visual representations and modulate the global content related to target categories. Experiments conducted on the iSAID-5^i dataset validate the effectiveness of the designs proposed in this paper, while also demonstrating that the introduced HSE method sets a new benchmark for excellence.

References

[1] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2020–2030.

[2] B. Forster, "An examination of some problems and solutions in monitoring urban areas from satellite platforms," *International journal of remote sensing*, vol. 6, no. 1, pp. 139–151, 1985.

[3] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," *arXiv preprint arXiv:2312.12470*, 2023.

[4] C. Ru, S.-B. Duan, X.-G. Jiang, Z.-L. Li, Y. Jiang, H. Ren, P. Leng, and M. Gao, "Land surface temperature retrieval from landsat 8 thermal infrared data over urban areas considering geometry effect: Method and application," *IEEE Transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[5] W. Qiao, L. Shen, J. Wang, X. Yang, and Z. Li, "A weakly supervised semantic segmentation approach for damaged building extraction from postearthquake high-resolution remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[6] E. Macioszek and A. Kurek, "Extracting road traffic volume in the city before and during covid-19 through video remote sensing," *Remote Sensing*, vol. 13, no. 12, p. 2329, 2021.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2015.7298965

[8] 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*, p. 834–848, Apr 2018. [Online]. Available: http://dx.doi.org/10.1109/tpami.2017.2699184

[9] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2017.

[10] X. Zeng, Z. Wang, Y. Wang, X. Rong, P. Guo, X. Gao, and X. Sun, "Semipscn: Polarization semantic constraint network for semi-supervised segmentation in large-scale and complex-valued polsar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[11] Y. Xin, Z. Fan, X. Qi, Y. Zhang, and X. Li, "Confidence-weighted dual-teacher networks with biased contrastive learning for semi-supervised semantic segmentation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[12] W. Miao, Z. Xu, J. Geng, and W. Jiang, "Ecae: Edge-aware class activation enhancement for semisupervised remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[13] X. Zeng, T. Wang, Z. Dong, X. Zhang, and Y. Gu, "Superpixel consistency saliency map generation for weakly supervised semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[14] R. Zhou, W. Zhang, Z. Yuan, X. Rong, W. Liu, K. Fu, and X. Sun, "Weakly supervised semantic segmentation in aerial imagery via explicit pixel-level constraints," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[15] Y. Su, M. Cheng, Z. Yuan, W. Liu, W. Zeng, Z. Zhang, and C. Wang, "Spatial adaptive fusion consistency contrastive constraint: Weakly supervised building facade point cloud semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[16] Z. Li, X. Zhang, and P. Xiao, "One model is enough: Toward multiclass weakly supervised remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[17] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[18] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5249–5258.

[19] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 2020, pp. 763–778.

[20] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 622–631.

[21] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9197–9206.

[22] 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 (CVPR)*, June 2019.

[23] Y. Yang, Q. Chen, Y. Feng, and T. Huang, "Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7131–7140.

[24] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, p. 3855–3865, Sep 2020. [Online]. Available: http://dx.doi.org/10.1109/tcyb.2020.2992433

[25] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8334–8343.

[26] 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, Feb 2022. [Online]. Available: http://dx.doi.org/10.1109/tpami.2020.3013717

[27] C. Lang, B. Tu, G. Cheng, and J. Han, "Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation," *arXiv preprint arXiv:2204.09903*, 2022.

[28] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 151–168.

[29] H. Wang, X. Zhang, Y. Hu, Y. Yang, X. Cao, and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," in *Computer Vfision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 730–746.

[30] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 984–21 996, 2021.

[31] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6941–6952.

[32] Z. Wang, M. Suganuma, and T. Okatani, "Improved few-shot segmentation by redefinition of the roles of multi-level cnn features," *arXiv preprint arXiv:2109.06432*, 2021.

[33] G.-S. Xie, J. Liu, H. Xiong, and L. Shao, "Scale-aware graph neural network for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5475–5484.

[34] D. Kang and M. Cho, "Integrative few-shot learning for classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9979–9990.

[35] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim, "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 108–126.

[36] B. Liu, J. Jiao, and Q. Ye, "Harmonic feature activation for few-shot semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3142–3153, 2021.

[37] 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 266–19 276.

[38] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," *Cornell University - arXiv,Cornell University - arXiv*, Mar 2021.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 553–11 562.

[40] X. Chen, N. Rostamzadeh, B. Oreshkin, and P. Pinheiro, "Adaptive cross-modal few-shot learning," *arXiv: Learning,arXiv: Learning*, Feb 2019.

[41] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 581–23 591.

[42] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[44] 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.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations,International Conference on Learning Representations*, Jan 2015.

[46] 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.

[47] 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.

[48] 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*, p. 834–848, Apr 2018. [Online]. Available: http://dx.doi.org/10.1109/tpami.2017.2699184

[49] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Jun 2017.

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

[51] 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.

[52] X. Lian, Y. Pang, J. Han, and J. Pan, "Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation," *Pattern Recognition*, p. 107622, Feb 2021. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2020.107622

[53] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[54] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, May 2021.

[55] 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 *European conference on computer vision*. Springer, 2022, pp. 205–218.

[56] 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 Geoscience and Remote Sensing Letters*, p. 1–5, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/lgrs.2021.3063381

[57] 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 Transactions on Geoscience and Remote Sensing*, p. 1–13, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2021.3093977

[58] D. Yu and S. Ji, "Long-range correlation supervision for land-cover classification from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[59] Y. Ni, J. Liu, W. Chi, X. Wang, and D. Li, "Cgglnet: Semantic segmentation network for remote sensing images based on category-guided global–local feature interaction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[60] 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.

[61] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms." *arXiv: Learning,arXiv: Learning*, Mar 2018.

[62] L. Zintgraf, K. Shiarlis, V. Kurin, K. Hofmann, and S. Whiteson, "Fast context adaptation via meta-learning," *Cornell University - arXiv,Cornell University - arXiv*, Oct 2018.

[63] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, "Fast context adaptation via meta-learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7693–7702.

[64] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

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

[66] S. Fort, "Gaussian prototypical networks for few-shot learning on omniglot," *arXiv: Learning,arXiv: Learning*, Feb 2018.

[67] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

[68] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, p. 6983–6994, Aug 2021. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2020.3027387

[69] 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*, p. 7844–7853, Sep 2021. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2020.3033336

[70] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multiview for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[71] J. Zhu, Y. Li, K. Yang, N. Guan, Z. Fan, C. Qiu, and X. Yi, "Mvp: Meta visual prompt tuning for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[72] C. Yang, T. Liu, G. Chen, and W. Li, "Icsff: Information constraint on self-supervised feature fusion for few-shot remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[73] Y. Wang, R. Sun, Z. Zhang, and T. Zhang, "Adaptive agent transformer for few-shot segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 36–52.

[74] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9587–9595.

[75] 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*, p. 1–11, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2021.3119852

[76] 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*, p. 1–18, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2021.3116672

[77] C. Lang, J. Wang, G. Cheng, B. Tu, and J. Han, "Progressive parsing and commonality distillation for few-shot remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.

[78] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Holistic mutual representation enhancement for few-shot remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[79] C. Lang, G. Cheng, B. Tu, and J. Han, "Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, p. 1–11, Jan 2023. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2023.3301003

[80] 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 Transactions on Geoscience and Remote Sensing*, 2023.

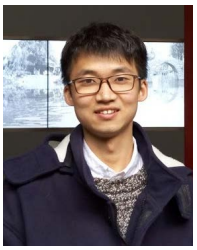
[81] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[82] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Amanda, S. Agarwal, A. Herbert-Voss, G. Krueger, H. Tom, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, C. Benjamin, J. Clark, C. Berner, M. Sam, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020.

[83] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 573–11 582.
[84] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8057–8067.
[85] X. Jiang, N. Zhou, and X. Li, "Few-shot segmentation of remote sensing images using deep metric learning," *IEEE Geoscience and Remote Sensing Letters*, p. 1–5, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/lgrs.2022.3154402
[86] 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*, vol. 16, no. 1, pp. 177–189, 2023.
[87] 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.
[88] 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.
[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
[90] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
[91] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[92] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Qi Wang (Senior Member, IEEE) 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, machine learning, pattern recognition, and remote sensing. For more information, visit the link (https://crabwq.github.io/).

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.

Qiang Li is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.