Dual-View Classifier Evolution for Generalized Remote Sensing Few-Shot Segmentation

Yuyu Jia, Wenhao Fu, Junyu Gao, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Advancements in few-shot segmentation for remote sensing images have significantly improved the ability to binarization parse novel classes using only a few supports. Generalized Few-Shot Segmentation (GFSS), a challenging and practical task, has recently attracted research attention. It involves recognizing base and novel classes while segmenting multiple categories in a query. Most GFSS methods adopt a two-stage approach: *base classifier training* and *novel classifier registering*. However, they encounter two key challenges: the data scale disparity between base and novel classes and significant intra-class variation in remote sensing images. In this paper, we present DiCE, a Dual-view Classifier Evolution method. Our approach utilizes the well-trained base classifier to allocate attention within the novel classifier, effectively addressing the disparities between the two. Simultaneously, it fosters context-driven interactions between the query and the classifier, tailoring sample-specific classifiers to mitigate intra-class variations. Furthermore, we propose a binocular hybrid training mechanism that integrates normal base training with episodic training, endowing the model with the ability to adapt to few-shot tasks. Extensive experiments on the iSAID-5^i dataset demonstrate the superior performance of DiCE.**

***Index Terms*—Generalized Few-shot Segmentation, Context Learning, Classifier Evolution, Binocular Hybrid Training.**

I. Introduction

IMAGE segmentation, a fundamental task in computer vision, focuses on pixel-level localization of target objects within specific categories. It plays a vital role in enabling intelligent interpretation of remote sensing imagery across various applications [1]. With the support of vast amounts of remote sensing data, deep learning-based segmentation techniques [2], [3] have become increasingly mature. However, these methods are constrained by the reliance on time-intensive, labor-intensive pixel-level annotations and exhibit limited generalization in specialized scenarios with scarce data availability. Few-Shot Segmentation (FSS) [4], [5], [6], [7], as an effective solution to this dilemma, has garnered significant attention from researchers. Built upon training with a sufficient amount of *base/known* data, it recognizes *novel/unseen* classes with the guidance of a limited number of labeled samples.

Contemporary FSS approaches predominantly leverage a meta-learning framework, wherein the base set is partitioned into numerous episodes to emulate few-shot scenarios. Each episode comprises a support set, annotated with corresponding mask labels, and a query set to be segmented. To extract category-specific discriminative features from the support set, techniques such as Mask Averaging Pooling (MAP) [8] are commonly employed. These extracted features are subsequently utilized to guide the segmentation process for the query set. However, as illustrated in Fig. 1, FSS inference is limited to segmenting only novel classes while ignoring base classes. In other words, FSS formulates segmentation as a binary classification problem, distinguishing only between foreground target regions and the background. This binary setup is inadequate for remote sensing image analysis, where large-scale images typically encompass diverse land cover types with multiple target regions appearing simultaneously.

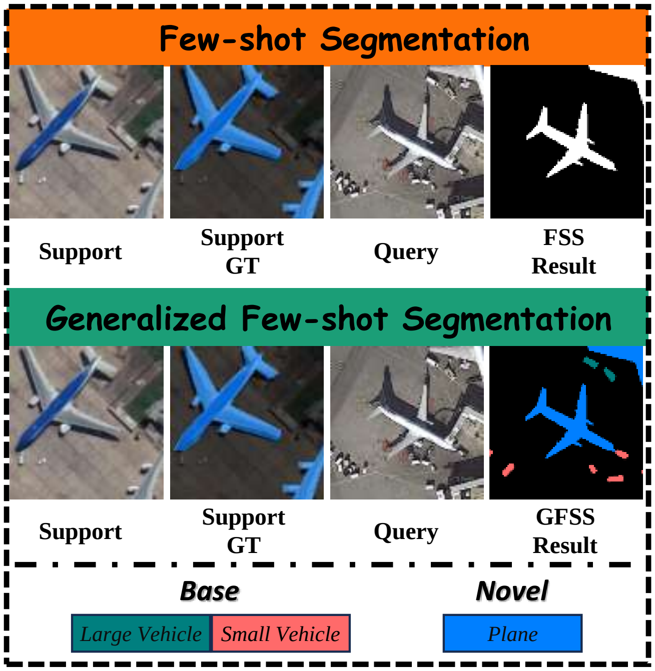


Fig. 1. In FSS, the query focuses on binary segmentation, where a single novel class is treated as the foreground and all other regions are classified as the background. In contrast, GFSS requires the segmentation of multiple classes, including both base and novel classes, within the query.

In contrast, Generalized Few-Shot Segmentation (GFSS) moves a step closer to practical applications by simultaneously segmenting both base and novel classes in queries, as depicted in Fig. 1. By incorporating the airplane sample as support for the novel class, the model not only gains the capability to identify airplanes (*novel*) in query images but also preserves its ability to recognize vehicles (*base*), a skill acquired during the

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

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

E-mail: jyy2019@mail.nwpu.edu.cn, fwh75391@163.com, gjy3035@gmail.com, crabwq@gmail.com.

Qi Wang is the corresponding author.

base training phase. While it offers practical advantages, GFSS has seen limited research, with most studies focusing primarily on natural images. Prevailing methodologies predominantly follow a two-stage learning paradigm. In the first phase, the model is trained on abundant base class data using a standard cross-entropy loss to construct a robust base classifier. Greater emphasis, however, is placed on the second phase: registering the novel class classifier. The initial idea is to extract segmentation guidance from the novel supports during inference, registering as a novel classifier and fusing it with the base one [9], [10]. Nevertheless, this fusion often results in a significant bias towards the base classes due to the severe data imbalance. To address this bias, several approaches [11], [12] fine-tune the classifier during inference with limited support supervision. Despite significant achievements, two pivotal issues warrant further in-depth exploration: **(i)** Fine-tuning the classifier risks overfitting and compromising the performance of the well-trained base classifier; **(ii)** Significant intra-class variations in remote sensing images frequently result in ambiguous classifier judgments.

In this paper, we propose a **D**ual-v**i**ew **C**lassifier **E**volution (DiCE) method. In the GFSS task setting, the base classifier, trained on extensive data, demonstrates strong discriminative capabilities, whereas the novel classifier, typically derived from limited support samples, tends to be weaker. Addressing this disparity, DiCE introduces an alternative view for classifier evolution: ***base-to-novel***. Specifically, we introduce an Attention Allocation (A^2) mechanism to transfer the extensive representations from the base classifier to the novel classifier. This mechanism further ensures that decisions concerning novel classes effectively leverage pertinent information from base classes, thereby harmonizing the biases between the two classifiers. The second view is ***query-to-clssifier***, wherein we design a Context Augmentation (CA) module leveraging transformer blocks. This module enriches the classifier with contextual information from each query, enabling it to evolve into a sample-specific representation. Consequently, it effectively addresses the challenge of intra-class variance. Additionally, we implement a Binocular Hybrid Training (BHT) strategy comprising two synergistic components: a normal base classifier and a few-shot classifier, the latter constructed by randomly sampling a limited number of instances from each batch. The joint optimization of these two enhances the model's adaptability to few-shot learning scenarios.

To summarize, the key contributions are as follows:

1) We identify two critical challenges faced by GFSS in remote sensing imagery: significant intra-class variance and base classifier bias.
2) To tackle these challenges, we propose the DiCE approach, which evolves the classifier through two views: ***query-to-clssifier*** and ***base-to-novel***, corresponding to the CA and A^2 modules, respectively.
3) We introduce a binocular hybrid training strategy to improve the model's adaptability in few-shot learning scenarios. Extensive experiments on the iSAID-5^i dataset significantly surpass state-of-the-art methods and provide new insights into GFSS.

II. Related Works

A. Few-Shot Learning

Few-Shot Learning (FSL) originated from classification tasks, pioneering a paradigm that leverages a limited number of labeled samples to enable understanding in entirely new scenarios. FSL can be categorized into three main approaches: metric-based, memory-based, and learning-based methods. Metric-based methods [13], [14], [15], [16] focus on predicting class probabilities by measuring the proximity between query and support samples in a learned metric space. Memory-based meta-learning techniques [17], [18], [19] leverage dynamic updating mechanisms to maximize the utility of historical samples, enhancing performance on the current task. Learning-based studies [20], [21], [22], [13] can be further categorized into three approaches: learning the initialization, learning the parameters, and learning the optimizer, all aimed at enhancing the model's adaptability to new tasks.

B. Few-Shot Segmentation

Few-Shot Segmentation (FSS) is based on the extensive class recognition learned during training, enabling pixel-level decoding of unseen classes with a limited number of supports. The seminal work OSLSM [4] represents the first effort to extend the principles of few-shot classification to the domain of dense prediction tasks, serving as a cornerstone for subsequent research in this field. Later literature on FSS can be divided into two main streams: prototype-based [23], [24], [25], [26], [27], [28], [29], [30], [31] and attention-based methods [32], [33], [34], [35], [36], [37], [38]. The former compresses the support information into one or more prototypes and guides the segmentation of the query by calculating similarity or concatenating them. The latter aims to establish a dense relationship between the support and the query and facilitate the transmission of support information through cross-attention mechanisms for guidance. Notably, during the inference phase, BAM [39] specifically identifies base-class objects to enhance the segmentation of novel-class regions, marking an important step toward extending the FSS task. As for FSS in remote sensing, SDM [40] pioneers a widely used dataset (iSAID-5^i) for this task and designs a detail-matching module to mitigate the intra-class variance. HMRE [41] proposes a dual-branch representation enhancement structure to address background interference in remote sensing images. HSE [42] leverages category description text embeddings in conjunction with visual representations to construct robust joint category discriminators, significantly enhancing segmentation performance.

However, methods above overlook base classes and fail to effectively identify multiple categories within the query, which renders them inadequate for practical analysis.

C. Generalized Few-Shot Segmentation

Generalized Few-Shot Semantic Segmentation (GFSS) seeks to further extend FSS by incorporating base classes during the inference phase and performing segmentation on all target regions within the query. For this task, CAPL [9]

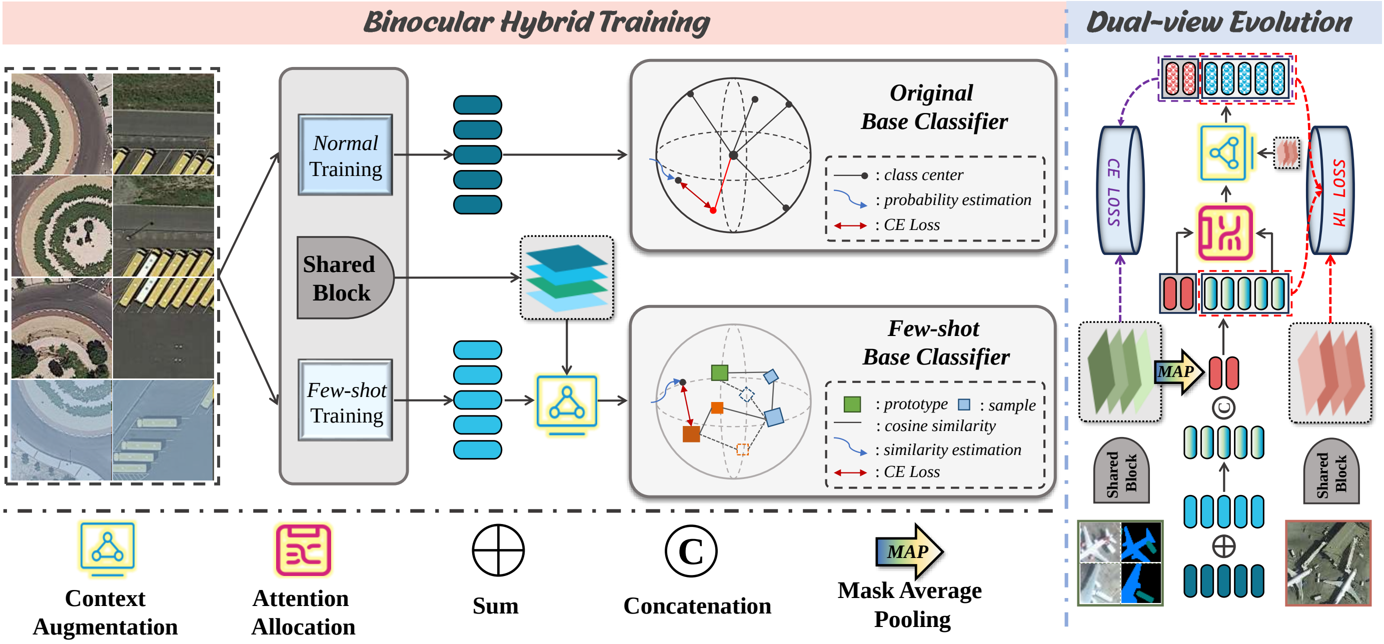


Fig. 2. The framework of DiCE. The proposed dual-view classifier evolution operates during the inference phase. The first view, *base-to-novel*, involves the design of an attention allocation module aimed at transferring the extensive representations from the well-trained base classifier to the novel classifier, thereby balancing their capability disparity. The second view, *query-to-classifier*, addresses the intra-class variance inherent in remote sensing images. Specifically, we introduce context augmentation, which leverages contextual information from the query to evolve the classifier into a sample-specific form. Furthermore, a binocular hybrid training strategy is implemented, merging normal base classifier training with optimization for few-shot scenarios randomly sampled within each batch.

sets up a baseline and mines supporting cues from within the query to guide the segmentation. POP [43] leverages the concept of orthogonal prototypes to prevent the model from degrading on base classes when generalizing to novel classes. PCN [10] addresses the base class bias issue by developing a prediction score regularization strategy and a transformer-based calibration module. A foreground context-aware module is proposed in [12] to prevent novel classes from being misclassified as background during inference. DIaM, [11] based on the InfoMax principle, fine-tunes the novel classifier during inference while employing regularization to ensure the base classifier retains its performance without degradation.

Considering the pronounced intra-class variance in remote sensing images compared to natural images, these methods are not directly applicable to remote sensing analysis. This paper, tailored to the unique characteristics of remote sensing imagery, establishes a dual-views classifier evolution framework that mitigates intra-class variance and effectively transfers the strengths of the base classifier to the novel classifier.

D. Contextual Visual Learning

Contextual learning is initially introduced in NLP and vision-language tasks, where it achieves significant success. Contextual learning for purely visual tasks was first introduced in [44], which was applied to address the image inpainting task. Painter [45] proposes the direct use of images to define visual task prompts, achieving outstanding performance across seven visual task datasets. SegGPT [46] uses the different colors of reference images as contextual prompts, successfully addressing a variety of arbitrary segmentation tasks. SINE [47] designs an in-context interaction module that transforms reference images into precise task instructions and employs a transformer-based M-Former architecture to generate segmentation masks at different granularities. Context learning has gradually emerged as a prominent approach in few-shot learning. For example, MSI [48] essentially taps into the contextual information within the background by leveraging two complementary sources of support features.

GFSS aims to decode the query by leveraging segmentation guidance from the support. The proposed DiCE provides rich contextual information from the query to the support, constructing a sample-specific classifier that mitigates the intra-class variance in remote sensing images.

III. Methodology

A. Preliminary

1) Problem formulation: In the standard GFSS setup, we are provided with a base set containing abundant samples and a novel set comprising only a limited number of samples. The two groups of data have no overlapping categories. The former comprises N_b categories $\mathcal{C}_b = \{c_1, \ldots, c_{N_b}\}$, while the latter includes N_n categories $\mathcal{C}_n = \{c_{N_b+1}, \ldots, c_{N_b+N_n}\}$. Pixels that do not belong to any predefined category are treated as the background category c_0. During training, the model can only access the base data from N_b categories. Unlike FSS, which performs binary segmentation on the query, GFSS requires the model to segment all $N_b + N_n + 1$ classes $(\mathcal{C}_b \cup \mathcal{C}_n \cup \{c_0\})$ within the query during inference, leveraging K pixel-level labeled supports, *i.e.*, K-shot for each novel class.

2) Baseline: GFSS models $\mathcal{M}$ adhere to the conventional segmentation architecture, consisting of a feature extractor $\mathcal{F}$ and a pixel-level classifier $\mathcal{G}$. Typically, the classifier $\mathcal{G} = \mathcal{G}_b^o \cup \mathcal{G}_n^o$ integrates the base $\mathcal{G}_b^o$ and novel classifiers $\mathcal{G}_n^o$ into a seamlessly unified structure.

During training, the base classifier $\mathcal{G}_b^o \in \mathbb{R}^{(N_b+1)\times d}$ is initialized with N_b categories (including the background), represented by prototypes with a channel dimension of d. Given the extracted feature map $F \in \mathbb{R}^{d\times H\times W}$, the segmentation probability distribution can be expressed as:

$$\boldsymbol{S} = \mathrm{softmax}(\mathcal{G}_b^o * F), S \in \mathbb{R}^{N_b\times H\times W}, \tag{1}$$

where H and W stand for the height and width of the extracted feature map. Subsequently, under the supervision of the sample's corresponding mask, the base classifier is optimized using the standard cross-entropy loss function.

The model is provided with support samples for N_n novel categories during the inference phase. The novel classifier $\mathcal{G}_n$ can be constructed through a mask average pooling (MAP) operation applied to the supports:

$$g_n = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{h,w} M_n^k(h,w)F_n^k(h,w)}{\sum_{h,w} M_n^k(h,w)}, g_n \in \mathbb{R}^{1\times d}, \tag{2}$$

where $\forall n \in \{N_b+1,\cdots,N_b+N_n\}$, h,w represents the spatial coordinates, F_n^k and M_n^k denote the feature map and its corresponding mask for the k-th support sample under the n-th class, and K is the number of support samples for each novel class.

B. Method Overview

The proposed DiCE framework is outlined in Fig. 2, with its core design centered on refining the classifier during the inference phase. Specifically, it employs an attention allocation module III-C1 to enable ***base-to-novel*** classifier evolution, effectively balancing biases between the two classifiers. To address the intra-class variance in remote sensing images, a context augmentation module III-C2 leverages query-specific contextual information from a ***query-to-classifier*** view, transforming the classifier into a sample-specific one. Furthermore, a binocular hybrid training mechanism III-D is introduced to optimize the feature extractor and the CA module.

C. Dual-view Classifier Evolution

1) Attention allocation: As described in the baseline methodology III-A2, the base classifier is learned from ample training data. In contrast, the novel classifier is derived solely from limited support samples, resulting in differing discriminative capabilities. Integrating the two into a final classifier can introduce a bias towards the base classes, leading to the misclassification of novel pixels with lower confidence as base categories. Existing methods [11], [12] attempt to alleviate the bias by fine-tuning the classifier with limited support supervision, but this inevitably leads to overfitting on a few novel instances and degrades the performance of the well-trained base classifier. Thus, the challenge is: *how can the novel classifier be refined while preserving the performance of the base classifier*?

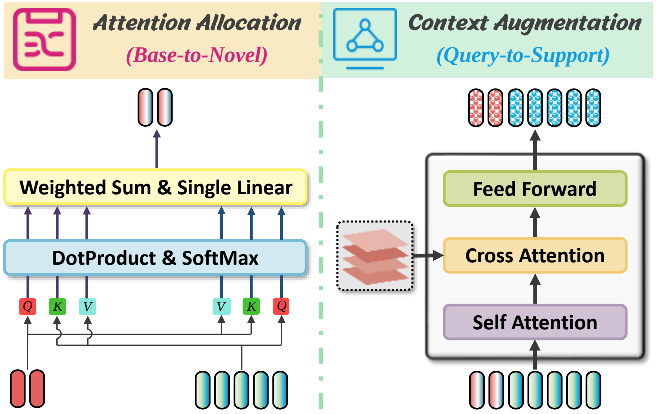


Fig. 3. Implementation of the A^2 module (left) and the CA module (right). The A^2 module takes the original base and novel classifiers as inputs and produces the refined novel classifier. The CA module is composed of standard transformer blocks, which output the final evolved classifier.

From the ***base-to-novel*** view, we propose an Attention Allocation (A^2) module. The core idea is to transfer the extensive class representations from the well-trained base classifier to the novel classifier. Additionally, we enable the novel classifier to dynamically adjust its attention weights by accessing relevant information from the base classes during decision-making.

Given the limited supervision available for finetuning during the inference phase, we implemented the A^2 module with a single linear layer to reduce the number of trainable parameters and prevent overfitting. Its detailed structure is illustrated in Fig. 3. Concretely, with the base classifier $\mathcal{G}_b$ from base training and the novel classifier $\mathcal{G}_n$ initialized from Eq. 2 as inputs, three single-linear heads project them to Q, K, and V:

$$Q = \mathcal{G}_n \cdot \theta_q; K = \mathcal{G}_b \cdot \theta_k; V = \mathcal{G}_b \cdot \theta_v. \tag{3}$$

Then, the weight allocation factor can be expressed as:

$$\Delta = \mathrm{softmax}(Q \cdot (K)^{\mathrm{T}}) \cdot V, \Delta \in \mathbb{R}^{N_n\times d}. \tag{4}$$

The refined novel classifier is obtained in a residual form:

$$\mathcal{G}_n^r = \mathcal{G}_n + \Delta. \tag{5}$$

2) Context augmentation: In remote sensing imagery, intra-class variance poses a significant challenge for few-shot tasks, as such differences hinder the classifier's ability to comprehensively capture and represent the diverse features of samples within the same category. In light of this, the proposed Context Augmentation (CA) module adopts a ***query-to-classifier*** view, driving contextual interaction between the query and the classifier. Before guiding the segmentation process, it leverages contextual information from the query to enrich the classifier, equipping it with sample-specific representational capabilities. As illustrated in Fig. 3, the CA module utilizes a transformer-based network consisting of sequentially connected self-attention, cross-attention, and feedforward blocks. It is optimized through binocular hybrid training in III-D and frozen during the inference phase to prevent overfitting.

Formally, we first concatenate the refined novel classifier $\mathcal{G}_n^r$ and the base classifier $\mathcal{G}_b$ along the channel dimension and then apply self-attention to establish relationships between the representations of different classes:

$$[\mathcal{G}_b;\mathcal{G}_n^r]_{sa}=\mathcal{SA}([\mathcal{G}_b;\mathcal{G}_n^r]), \tag{6}$$

where $[\mathcal{G}_b;\mathcal{G}_n^r]_{sa}\in\mathbb{R}^{N_b+N_n+1}$ and $\mathcal{SA}(\cdot;\cdot)$ denotes the self-attention operation. Assuming $F_q^{infer}\in\mathbb{R}^{d\times H\times W}$ represent the extracted query feature during inference, we take the segmentation of the query as an example. We flatten it to conduct the cross-attention process to acquire the augmented query classifier $\mathcal{G}_q^{aug}$:

$$\mathcal{G}_q^{aug}=\mathcal{FW}(\mathcal{CA}(R(F_q^{infer});[\mathcal{G}_b;\mathcal{G}_n^r]_{sa}), \tag{7}$$

where $\mathcal{G}_q^{aug}\in\mathbb{R}^{N_b+N_N+1}$, $R(\cdot)$ denotes the reshape operation, $\mathcal{CA}(\cdot;\cdot)$ represents cross-attention, and $\mathcal{FW}(\cdot)$ refers to the feed-forward block. Through the augmentation above, we tailor the classifier for each query, effectively addressing the challenge of intra-class variance.

3) Inference loss: During inference, the model first uses support masks as supervision signals to optimize the A^2 module. Given the extracted support feature, we derive the augmented support classifier $\mathcal{G}_s^{aug}\in\mathbb{R}^{N_b+N_N+1}$ using the approach outlined in Eq. 7:

$$\mathcal{G}_s^{aug}=\mathcal{FW}(\mathcal{CA}(R(F_s^{infer});[\mathcal{G}_b;\mathcal{G}_n^r]_{sa}). \tag{8}$$

The segmentation probability of the support can be written as:

$$\boldsymbol{S}_s=\text{softmax}(\mathcal{G}_s^{aug}*F_s^{infer}), S_s\in\mathbb{R}^{(N_b+N_n+1)\times H\times W}. \tag{9}$$

Then, the standard cross-entropy loss can be formulated as:

$$\mathcal{L}_{ce}=\sum_{i=1}^{|\mathbb{S}|}\text{CE}(\boldsymbol{Y}_i;\boldsymbol{S}_{s,i}), \tag{10}$$

where $\mathbb{S}$ represents the support set, $\boldsymbol{Y}_i$ and $\boldsymbol{S}_{s,i}$ denote the predicted probability and ground truth mask for the i-th sample in the support set, respectively. Similarly, the segmentation probability of the query can be computed by:

$$\boldsymbol{S}_q=\text{softmax}(\mathcal{G}_q^{aug}*F_q^{infer}), \tag{11}$$

where $S_q\in\mathbb{R}^{(N_b+N_n+1)\times H\times W}$.

So far, all prior meticulous designs focus on balancing the bias between the two classifiers and reducing intra-class variance. However, these objectives should not compromise the performance of the well-trained base classifier. To address this, we introduce a regularization constraint that encourages the *evolved* classifier's predictions (in Eq. 11) for base classes to closely align with those of the *old* base classifier (in Eq. 1). Specifically, we consider the prediction of the *old* base classifier as follows:

$$\boldsymbol{S}_q^{old,b}=\text{softmax}(\mathcal{G}_b*F_q^{infer}), \tag{12}$$

where $\boldsymbol{S}_q^{old,b}\in\mathbb{R}^{(N_b+1)\times H\times W}$. For the classifier evolved during inference, we consolidate the predictions of all novel classes at each pixel into a single background class, thereby obtaining pure base class predictions:

$$\boldsymbol{S}_q^b=[s_0+\sum_{i=1}^{N_n}s_{N_b+i},s_1,s_2,\cdots,s_{N_b}]. \tag{13}$$

Then, we impose the regularization constraint for the entire query set $\mathbb{Q}$ using the Kullback-Leibler divergence [49]:

$$\mathcal{L}_{kl}=\sum_{i=1}^{\mathbb{Q}}\text{KL}(\boldsymbol{S}_{q,i}^{old,b}||\boldsymbol{S}_{q,i}^b). \tag{14}$$

Finally, the overall optimization loss function for DiCE during the inference phase can be weighted as:

$$\mathcal{L}_{infer}=\mathcal{L}_{ce}+\alpha\mathcal{L}_{kl}, \tag{15}$$

where α controls the intensity of the regularization constraint.

D. Binocular Hybrid Training

Existing GFSS methods predominantly optimize the base classifier across the ***whole*** class space during the training phase, as outlined in the baseline III-A2. While this approach yields a robust base classifier, it neglects the few-shot scenarios encountered during inference. To address this limitation, we augment the training process with ***localized*** training tailored for few-shot tasks, establishing a Binocular Hybrid Training (BHT) mechanism. Specifically, assuming the class set within a batch is denoted as $\mathcal{C}_\tau$, we randomly sample one instance from each class to construct a few-shot task $\mathcal{T}$. A few-shot base classifier $\mathcal{G}_b^{\mathcal{T}}\in\mathbb{R}^{(N_b+1)\times d}$ can be obtained through the mask average pooling operation:

$$g_{b,i}^{\mathcal{T}}=\begin{cases}\frac{\sum_{h,w}M_i^{\mathcal{T}}(h,w)F_i^{\mathcal{T}}(h,w)}{\sum_{h,w}M_i^{\mathcal{T}}(h,w)}\in\mathbb{R}^{1\times d} & i\in\mathcal{C}_{\mathcal{T}}\\ g_{b,i} & \text{Otherwise}\end{cases}, \tag{16}$$

where $\forall i\in\{0,1,\cdots,N_b\}$, $\forall g_{b,i}\in\mathcal{G}_b^o$, and $\forall g_{b,i}^{\mathcal{T}}\in\mathcal{G}_b^{\mathcal{T}}$. To optimize the CA module, the few-shot base classifier is augmented through the CA module as in Eqs. 6 and 7:

$$\hat{\mathcal{G}}_b^{\mathcal{T}}=\mathcal{FW}(\mathcal{CA}(R(F^{\mathcal{T}});\mathcal{SA}(\mathcal{G}_b^{\mathcal{T}})), \tag{17}$$

where $F^{\mathcal{T}}$ is the training data in the $\mathcal{T}$. Then, the base classifier can be constructed as $\mathcal{G}_b=\mathcal{G}_b^o+\hat{\mathcal{G}}_b^{\mathcal{T}}$ and optimized through the standard cross-entropy loss, enabling the binocular hybrid training strategy.

IV. Experiments

A. Dataset

Research on GFSS tasks for remote sensing images has remained limited in recent years, and relevant datasets are scarce. Given the similarity in task settings, we utilize the classic remote sensing dataset iSAID-5^i [40] from the FSS task to conduct a series of experiments. Specifically, iSAID-5^i is derived from iSAID [50], a large-scale dataset designed for instance and semantic segmentation of remote sensing images. It further processes the images into a size of 256×256. The dataset contains 18076 training samples and 6363 testing samples, covering 15 classes. Under the GFSS setting, all classes are divided into three splits, each containing five classes. One split is designated as novel classes, while the remaining two groups serve as base classes. Fig. 4 provides an intuitive overview of the dataset.

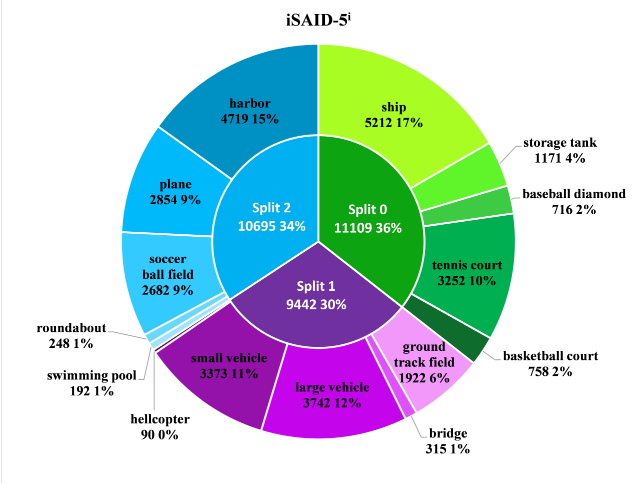


Fig. 4. Distribution overview of the iSAID-5^i dataset.

B. Implementation Details

The proposed DiCE is developed based on the PyTorch [51] framework, and all experiments are performed utilizing three NVIDIA GeForce RTX 3090 GPUs. The backbone network is a ResNet50 [52] pre-trained on ImageNet [53]. To balance computational efficiency and feature extraction, the first three blocks are frozen, reducing the number of learnable parameters while preserving the model's ability to capture essential low-level features.

During base class training, the model is trained for a total of 70 epochs using the SGD optimizer, with a batch size of 48. An initial learning rate of 0.01 is adopted, combined with a momentum value of 0.9. A weight decay coefficient of 0.0001 is applied. The learning rate decays following a polynomial schedule, where the decay power is set to 0.9.

In the novel fine-tuning phase, the backbone network and the CA module are frozen, while the classifier evolution is achieved by optimizing the A^2 module for 25 iterations using the SGD optimizer. The learning rate and weight decay are set to 2e-5 and 5e-2, respectively, with the regularization loss weighted at 0.6.

C. Evaluation Metric

This paper adopts the mean Intersection over Union (mIoU) as the evaluation metric for the segmentation model in the GFSS task. mIoU is a widely used performance indicator in semantic segmentation tasks, which evaluates the overlap between the predicted segmentation results and the ground truth. Specifically, it is calculated as the average IoU across all classes, where the IoU for each class is defined as the ratio of the intersection of the predicted and ground truth regions to their union. Mathematically, for N classes, mIoU is expressed as:

$$m\mathrm{IoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathrm{Prediction}_i \cap \mathrm{Ground\ Truth}_i|}{|\mathrm{Prediction}_i \cup \mathrm{Ground\ Truth}_i|}. \tag{18}$$

Building on this, we employ $m\mathrm{IoU}_{\mathcal{B}}$ and $m\mathrm{IoU}_{\mathcal{N}}$ to evaluate the segmentation performance of the base (including the background) and novel classes, respectively, while $m\mathrm{IoU}_{\mathcal{O}}$ representing the overall performance across all categories. However, as the base classes dominate in quantity, $m\mathrm{IoU}_{\mathcal{O}}$ is heavily influenced by their performance. To mitigate this imbalance and provide a more equitable and comprehensive evaluation of the model's capability, we propose hIoU, defined as follows:

$$h\mathrm{IoU} = \frac{2 \cdot m\mathrm{IoU}_{\mathcal{B}} \cdot m\mathrm{IoU}_{\mathcal{N}}}{m\mathrm{IoU}_{\mathcal{B}} + m\mathrm{IoU}_{\mathcal{N}}}. \tag{19}$$

TABLE I

QUANTITATIVE RESULTS ACROSS ALL SPLITS ON THE ISAID-5^i DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST RESULTS ARE MARKED WITH UNDERLINES. †INDICATES THAT THE METHOD IS SPECIFICALLY DESIGNED FOR THE GFSS TASK.

Methods	1-shot				5-shot			
	$m\mathrm{IoU}_{\mathcal{B}}$	$m\mathrm{IoU}_{\mathcal{N}}$	$m\mathrm{IoU}_{\mathcal{O}}$	hIoU	$m\mathrm{IoU}_{\mathcal{B}}$	$m\mathrm{IoU}_{\mathcal{N}}$	$m\mathrm{IoU}_{\mathcal{O}}$	hIoU
CANet [54]	46.42	9.18	34.01	15.33	46.98	10.20	34.72	16.76
PANet [30]	51.40	10.33	37.71	17.20	52.21	11.48	38.63	18.82
PFENet [28]	51.38	9.80	37.52	16.46	51.97	11.30	38.41	18.56
SCL [31]	51.67	8.12	37.15	14.03	52.11	10.19	38.14	17.05
HMRE [41]	50.38	8.40	36.39	14.40	50.90	10.38	37.39	17.24
CAPL† [9]	53.15	11.61	39.30	19.06	54.17	13.23	40.52	21.27
POP† [43]	53.67	12.10	39.81	19.75	54.55	13.85	40.98	22.09
DiCE (*ours*)	**55.98**	**13.38**	**41.78**	**21.60**	**57.71**	**14.56**	**43.33**	**23.25**

D. Compared Methods

1) Quantitative comparison: Existing GFSS research is limited. We compare the proposed DiCE framework with five state-of-the-art FSS algorithms and two GFSS techniques for natural images. Their key designs are as follows:

- CANet [54] employs a dense feature comparison module and an iterative optimization module, integrating attention mechanisms to achieve class-agnostic semantic segmentation under few-shot conditions.
- PANet [30] introduces a prototype alignment loss to align class-specific features between support and query samples, thereby improving intra-class feature consistency in FSS tasks.
- PFENet [28] employs a prior-guided mechanism and a feature enhancement module to achieve efficient alignment and fusion of features between support and query samples, thereby improving the generalization performance in FSS tasks.
- SCL [31] leverages self-guided learning to enhance the saliency of individual sample features and cross-guided learning to enable dynamic feature interaction between support and query sets, thereby improving feature representation and target region localization in FSS tasks.
- HMRE [41] addresses FSS in remote sensing imagery by introducing a holistic mutual representation enhancement module, enabling dynamic interaction and augmentation of features between the support and query sets.
- CAPL [9] introduces the first GFSS baseline by effectively utilizing co-occurrence prior knowledge from the support set and enhancing the contextual information of queries, thereby enabling balanced segmentation across both base and novel classes.

TABLE II
QUANTITATIVE PERFORMANCE ACROSS THREE SPLITS ON THE ISAID-5^i DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST RESULTS ARE MARKED WITH UNDERLINES. †DENOTES THAT THE METHOD IS SPECIFICALLY DESIGNED FOR THE GFSS TASK.

Methods	Shot	Split0				Split1				Split2			
		$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$m\text{IoU}_{\mathcal{O}}$	$h\text{IoU}$	$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$m\text{IoU}_{\mathcal{O}}$	$h\text{IoU}$	$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$m\text{IoU}_{\mathcal{O}}$	$h\text{IoU}$
CANet [54]	1	38.83	13.39	38.33	19.91	54.47	6.04	38.33	10.87	45.96	8.11	33.34	13.79
PANet [30]		44.24	14.28	34.25	21.59	61.50	6.68	43.23	12.05	48.46	10.02	35.65	16.61
PFENet [28]		45.12	14.37	34.87	21.80	60.46	6.77	42.56	12.18	48.57	8.26	35.13	14.12
SCL [31]		45.67	13.27	34.87	20.56	61.45	5.02	42.64	9.28	47.89	6.08	33.95	10.79
HMRE [41]		42.55	12.25	32.45	19.02	59.99	5.56	41.85	10.18	48.59	7.40	34.86	12.84
CAPL† [9]		47.44	14.77	36.55	22.53	61.59	6.76	43.31	12.18	50.43	<u>13.31</u>	38.06	21.06
POP† [43]		<u>48.42</u>	<u>15.96</u>	<u>37.60</u>	<u>24.01</u>	<u>61.66</u>	<u>7.05</u>	<u>43.46</u>	<u>12.65</u>	<u>50.93</u>	13.29	<u>38.38</u>	<u>21.08</u>
DiCE (*ours*)		**52.46**	**16.76**	**40.56**	**25.40**	**63.22**	**9.07**	**45.17**	**15.86**	**52.25**	**14.31**	**39.60**	**22.47**
CANet [54]	5	39.32	14.12	30.92	20.78	55.23	7.10	39.19	12.58	46.38	9.37	34.04	15.59
PANet [30]		44.98	15.14	35.03	22.65	62.39	8.01	44.26	14.20	49.27	11.30	36.61	18.38
PFENet [28]		45.88	15.46	35.74	23.13	60.94	8.21	43.36	14.47	49.08	10.22	36.13	16.92
SCL [31]		45.64	14.55	35.28	22.07	62.09	7.14	43.77	12.81	48.60	8.89	35.36	15.03
HMRE [41]		43.47	14.57	33.84	21.82	60.28	7.57	42.71	13.45	48.96	8.99	35.64	15.19
CAPL† [9]		48.51	16.59	37.87	24.72	<u>63.00</u>	<u>8.66</u>	<u>44.89</u>	<u>15.23</u>	51.01	14.43	38.82	22.50
POP† [43]		<u>49.55</u>	<u>18.23</u>	<u>39.11</u>	<u>26.65</u>	62.70	8.34	44.58	14.72	<u>51.40</u>	<u>14.98</u>	<u>39.26</u>	<u>23.20</u>
DiCE (*ours*)		**54.51**	**18.29**	**42.44**	**27.39**	**65.65**	**9.66**	**46.99**	**16.84**	**52.96**	**15.72**	**40.55**	**24.24**

TABLE III
QUANTITATIVE PERFORMANCE FOR EACH FINE-GRAINED CATEGORY ON THE ISAID-5^i DATASET. THE BACKGROUND FOR NOVEL CATEGORIES IS HIGHLIGHTED IN BLUE. †DENOTES THAT THE METHOD IS SPECIFICALLY DESIGNED FOR THE GFSS TASK.

Split	Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Split0	CANet [54]	28.37	22.67	10.72	2.86	2.32	9.31	39.01	63.11	52.97	24.65	49.41	53.45	1.07	80.10	7.47
	PANet [30]	23.06	29.73	6.72	9.52	2.37	23.46	33.27	61.40	53.34	31.65	59.29	56.36	24.76	80.77	10.15
	PFENet [28]	24.00	27.97	8.58	10.25	1.07	32.60	33.20	60.27	50.18	28.56	40.22	41.59	51.89	78.99	24.21
	SCL [31]	27.25	16.43	8.24	11.97	2.46	33.78	34.52	62.40	55.36	26.78	47.44	57.33	48.26	79.06	2.61
	CAPL† [9]	14.93	23.50	12.79	18.25	4.37	36.47	32.54	57.89	48.57	35.16	45.64	47.23	56.57	81.90	20.28
	POP† [43]	23.23	25.00	12.82	17.44	1.33	46.83	21.77	59.67	59.90	27.13	25.44	51.03	59.00	82.85	40.67
	DiCE (*ours*)	24.33	32.53	8.33	11.98	6.65	38.39	36.74	65.22	52.71	41.36	52.02	63.08	68.30	80.82	11.24
Split1	CANet [54]	55.02	74.30	73.25	82.14	60.19	11.65	0.24	10.64	7.68	0.20	43.44	58.45	66.49	16.17	6.41
	PANet [30]	55.52	71.75	72.21	83.23	59.71	10.45	0.69	11.66	10.19	0.40	49.74	66.81	63.25	75.53	6.85
	PFENet [28]	51.46	71.15	70.89	82.94	59.24	9.83	0.67	7.74	15.32	0.28	43.83	53.89	67.21	74.35	21.63
	SCL [31]	55.34	72.06	70.12	83.47	57.35	7.46	0.63	8.97	7.57	0.45	43.52	53.66	64.04	68.60	40.00
	CAPL† [9]	49.90	69.52	68.77	82.05	55.30	8.11	2.53	9.26	10.73	3.16	45.39	54.40	65.90	70.04	38.36
	POP† [43]	54.03	70.16	68.28	80.66	55.45	10.31	3.10	9.17	9.91	2.76	46.10	56.12	64.47	68.70	44.36
	DiCE (*ours*)	45.56	73.79	75.60	86.08	64.15	11.08	5.20	13.70	12.42	2.97	52.50	66.71	70.11	80.17	10.56
Split2	CANet [54]	43.25	73.10	73.88	74.24	36.15	29.74	13.24	61.35	44.60	3.03	0.13	0.38	12.92	4.37	22.73
	PANet [30]	44.05	72.67	75.58	75.69	35.30	35.14	30.00	62.32	47.12	2.03	0.28	2.09	7.30	13.37	27.04
	PFENet [28]	52.34	70.18	62.53	83.72	33.25	37.44	28.21	60.40	50.10	1.42	0.12	4.24	4.80	4.56	27.61
	SCL [31]	53.32	67.46	60.23	75.47	33.36	37.62	30.26	58.77	50.35	1.60	0.23	1.35	8.54	2.02	18.25
	CAPL† [9]	51.28	70.43	70.16	82.36	38.84	40.03	30.26	60.47	50.13	2.12	2.36	3.11	11.40	27.46	22.24
	POP† [43]	52.37	68.46	72.46	79.94	40.37	36.51	32.67	61.09	54.11	3.26	4.21	8.66	19.30	13.49	20.81
	DiCE (*ours*)	48.84	74.98	78.54	82.21	41.38	37.56	32.28	66.31	49.97	3.80	7.05	8.72	23.20	12.37	20.21

- POP [43] incorporates orthogonal constraints to learn orthogonal prototypes among classes, reducing category interference in the feature space and enhancing segmentation performance for both base and novel categories in the GFSS task.

The first five FSS algorithms employ prototype-based segmentation decoding. Building on their publicly available source codes, we generate base prototypes during training by applying average pooling to the base features. During inference, these base prototypes are combined with novel prototypes to construct the GFSS decoder.

The results in Table I reveal that the two GFSS techniques significantly outperform traditional FSS algorithms, achieving a superior balance between base and novel class segmentation. The proposed DiCE exhibits exceptional performance in both 1-shot and 5-shot scenarios, surpassing the second-best method, *i.e.,* POP [43], with improvements of 1.97% and 2.35% in $m\text{IoU}_{\mathcal{B}}$, and 1.85% and 1.16% in $h\text{IoU}$, respectively. Table II presents the segmentation performance of all methods across each split. It is observed that CAPL [9] and POP [43] significantly outperform other FSS algorithms, with each excelling in different splits. However, our proposed DiCE consistently ranks first, demonstrating superior robustness. More specifically, Table III details the segmentation performance of each category under different split settings. DiCE exhibits greater stability than other methods, which we attribute to the CA module's effectiveness in mitigating intra-class variance.

2) Qualitative comparison: Fig. 5 presents the qualitative segmentation comparison results under the split0 setting of the iSAID-5^i dataset. Overall, the visual outcomes are consistent with the performance indicated by the quantitative metrics. PFENet [28] exhibits strong recognition capabilities for base

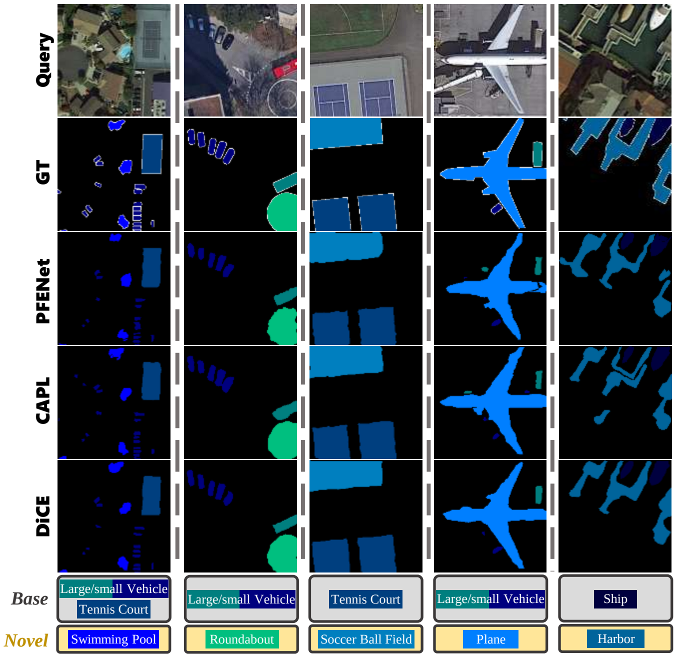


Fig. 5. Qualitative 5-shot comparison under the split0 setting of the iSAID-5^i dataset.

TABLE IV
Ablation results on the iSAID-5^i dataset of key components.

Component			1-shot			5-shot		
BHT	A^2	CA	$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$h\text{IoU}$	$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$h\text{IoU}$
✗	✗	✗	50.24	10.47	17.33	50.76	11.02	18.11
✔	✗	✗	53.04	10.21	17.12	54.69	11.17	18.55
✗	✔	✔	52.64	12.53	20.24	54.45	13.36	21.46
✔	✗	✔	54.22	11.79	19.37	56.84	12.58	20.60
✔	✔	✗	55.26	12.94	20.97	57.12	14.38	22.98
✔	✔	✔	**55.98**	**13.38**	**21.60**	**57.71**	**14.56**	**23.25**

TABLE V
Impact of different design options on performance under the 1-shot setting.

A^2	CA	$m\text{IoU}_{\mathcal{B}}$	$m\text{IoU}_{\mathcal{N}}$	$h\text{IoU}$
Random	$\mathcal{SA}$ *w/o*	53.79	11.84	19.41
Mask Average Pooling	$\mathcal{SA}$ *w/o*	54.51	12.49	20.32
Random	$\mathcal{SA}$	55.12	12.17	19.94
Mask Average Pooling	$\mathcal{SA}$	**55.98**	**13.38**	**21.60**
Optimizing parameters of the classifier		52.21	11.42	18.74
Optimizing parameters of the classifier *w/o*		**55.98**	**13.38**	**21.60**

classes but faces challenges in generalizing to novel classes. CAPL [9], specifically designed for the GFSS task, achieves commendable performance for both base and novel classes. However, it struggles to mitigate the bias toward base classes. For instance, in the first and fourth columns, it erroneously activates the small vehicle and large vehicle categories, respectively. In contrast, DiCE delivers remarkable results, achieving precise edge predictions even for small targets.

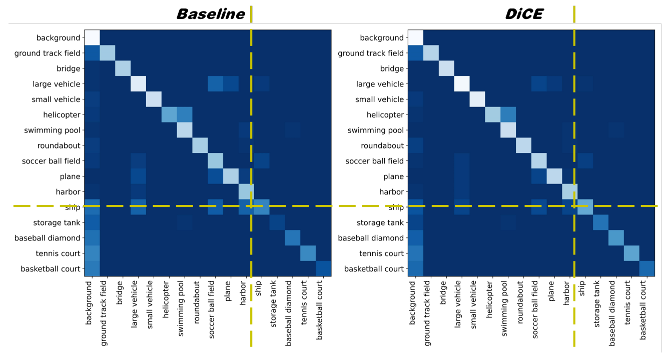


Fig. 6. Confusion matrix for the baseline (*left*) and the proposed DiCE (*right*)

TABLE VI
Gain brought by more supports.

Shot	1	3	5	7	10
$m\text{IoU}_{\mathcal{B}}$	55.98	57.19	57.71	57.98	58.26
$m\text{IoU}_{\mathcal{N}}$	13.38	14.01	14.56	14.72	15.33
$m\text{IoU}_{\mathcal{O}}$	41.78	42.80	43.33	43.56	43.95
$h\text{IoU}$	21.60	22.51	23.25	23.48	24.27

E. Further Analysis

1) Component-wise effectiveness: We perform ablation studies on three critical components of the proposed DiCE framework, as summarized in Table IV: Binocular Hybrid Training (BHT), Attention Allocation (A^2), and Context Augmentation (CA). The baseline model employs PSPNet as the backbone, with the pixel classifier constructed by directly concatenating the base prototype obtained during training and the novel prototype defined in Eq. 2. The experimental results yield the following key observations: (i) Introducing BHT alone substantially enhances performance on base classes by improving the model's adaptability to few-shot learning scenarios. (ii) The Attention Allocation and Context Augmentation modules collectively enhance the model's ability to balance base and novel classes, effectively addressing base class bias and reducing intra-class variance. The superiority of DiCE is further illustrated in Fig. 6. Compared to the baseline (*left*), DiCE (*right*) demonstrates a marked reduction in class confusion across all categories, underscoring its effectiveness in achieving robust segmentation.

2) Design options of DiCE: In Table V, we investigate the influence of different design choices within the DiCE framework on model performance. Specifically, the A^2 module is designed to refine the novel classifier by leveraging the well-trained base classifier, as the latter is constructed using limited support information.

We evaluate two initialization strategies for the novel classifier: random initialization and masked average pooling. The results consistently demonstrate that masked average pooling outperforms random initialization. This improvement can be attributed to the ability of masked pooling to generate more precise and stable class feature representations, thereby facilitating faster convergence during the inference stage.

For the CA module, we analyze the role of the self-attention ($\mathcal{SA}$) block in improving query-to-classifier interactions. The results highlight that the self-attention mechanism effectively captures the relationships among class prototypes within the classifier, leading to more accurate classification outcomes.

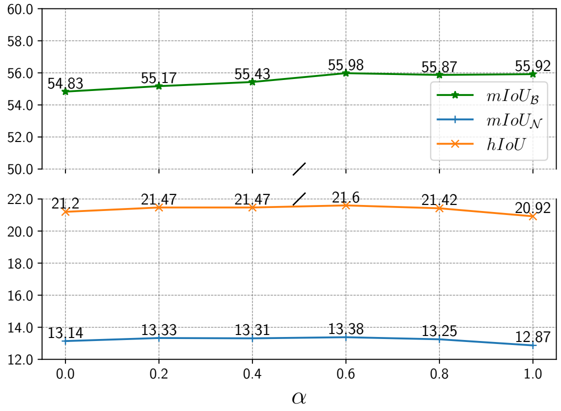


Fig. 7. Impact of hyperparameter α on performance under the 1-shot setting.

Additionally, we experimented with incorporating the classifier parameters into the optimization process during inference. However, this approach resulted in a noticeable performance degradation. A detailed analysis suggests that, given the limited number of support samples, introducing additional parameters during inference increases the risk of overfitting. To mitigate this issue, we fix the classifier parameters and instead focus on indirectly refining the classifier by optimizing the A^2 module.

3) Gain brought by more supports: Table VI illustrates the effect of increasing the number of support samples on model performance. The results reveal a consistent improvement in performance as more support information becomes available. However, the rate of improvement gradually diminishes in the later stages. This diminishing return can be attributed to the phenomenon of information saturation, where additional support samples contribute redundant information, thereby limiting further performance gains.

4) Regularization constraint loss: The regularization constraint loss in Eq. 14 is used to prevent the degradation of the well-trained base classifier during the optimization process, with the hyperparameter α controlling the weight of this loss. Fig. 7 illustrates the impact of α on model performance. As it increases, the performance of the base classifier improves significantly. However, if it becomes too large, the loss dominates the overall optimization of the model, ultimately preventing the classifier from achieving satisfactory performance. Finally, we set α to 0.6 to achieve the overall optimal performance.

V. Conclusion

In this paper, we propose a dual-view classifier evolution approach to mitigate the base bias in the GFSS task and the intra-class variation in remote sensing images. Specifically, the attention allocation module transfers the representational power of the well-trained base classifier to the novel classifier, achieving a *base-to-novel* evolution. The context augmentation module leverages contextual information from the query to enrich the classifier, making it sample-specific, thus driving a *query-to-classifier* evolution. Furthermore, we adopt a binocular hybrid training mechanism, combining the normal training of the base classifier with few-shot optimization, enabling the model to adapt to few-shot scenarios during the training phase. A series of both quantitative and qualitative experiments demonstrate the superior performance of the proposed DiCE and the effectiveness of its individual modules.

Despite its effectiveness, our approach faces practical challenges worth noting. The designs of the proposed CA and A^2 modules incorporate several attention blocks. Although the additional trainable parameters are minimal, their deployment on edge devices still presents challenges. From a societal standpoint, DiCE facilitates transformative applications in environmental monitoring. Its capability to concurrently track both base classes (*e.g.,* permanent water bodies) and novel classes (*e.g.,* seasonal flood areas) can significantly contribute to climate change adaptation efforts. In disaster response scenarios, the model has the potential to preserve the segmentation of existing urban infrastructure while identifying newly emerging landslide areas with minimal sample data, thereby expediting damage assessment and response efforts.

References

[1] C. Lang, G. Cheng, J. Wu, Z. Li, X. Xie, J. Li, and J. Han, "Toward open-world remote sensing imagery interpretation: Past, present, and future," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–38, 2024.

[2] 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

[3] 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

[4] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Procedings of the British Machine Vision Conference 2017*, Jan 2017. [Online]. Available: http://dx.doi.org/10.5244/c.31.167

[5] M. Siam, B. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00535

[6] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Computer Vision – ECCV 2020,Lecture Notes in Computer Science*, Jan 2020, p. 763–778. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-58598-3_45

[7] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00071

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

[9] Z. Tian, X. Lai, L. Jiang, S. Liu, M. Shu, H. Zhao, and J. Jia, "Generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 563–11 572.

[10] Z. Lu, S. He, D. Li, Y.-Z. Song, and T. Xiang, "Prediction calibration for generalized few-shot semantic segmentation," *IEEE transactions on image processing*, vol. 32, pp. 3311–3323, 2023.

[11] S. Hajimiri, M. Boudiaf, I. Ben Ayed, and J. Dolz, "A strong baseline for generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 269–11 278.

[12] K. Huang, F. Wang, Y. Xi, and Y. Gao, "Prototypical kernel learning and open-set foreground perception for generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 256–19 265.

[13] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," *Advances in neural information processing systems*, vol. 29, 2016.

[14] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

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

[16] S. Mai, H. Hu, and J. Xu, "Attentive matching network for few-shot learning," *Computer Vision and Image Understanding*, vol. 187, p. 102781, 2019.

[17] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[18] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," *arXiv preprint arXiv:1707.03141*, 2017.

[19] K. Tran, H. Sato, and M. Kubo, "Memory augmented matching networks for few-shot learnings," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, 2019.

[20] 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.

[21] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, "Bayesian model-agnostic meta-learning," *Advances in neural information processing systems*, vol. 31, 2018.

[22] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun, "Boil: Towards representation change for few-shot learning," *arXiv preprint arXiv:2008.08882*, 2020.

[23] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 701–719.

[24] 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.

[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*, 2021, pp. 8334–8343.

[26] 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.

[27] A. Okazawa, "Interclass prototype relation for few-shot segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 362–378.

[28] 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*, vol. 44, no. 2, pp. 1050–1065, 2020.

[29] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5463–5471.

[30] 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.

[31] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8312–8321.

[32] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8441–8448.

[33] G. Zhang, S. Navasardyan, L. Chen, Y. Zhao, Y. Wei, H. Shi *et al.*, "Mask matching transformer for few-shot segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 823–836, 2022.

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

[35] 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.

[36] Q. Xu, W. Zhao, G. Lin, and C. Long, "Self-calibrated cross attention network for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 655–665.

[37] 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.

[38] 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.

[39] 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 (CVPR)*, June 2022, pp. 8057–8067.

[40] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[41] 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*, 2023.

[42] Y. Jia, W. Huang, J. Gao, Q. Wang, and Q. Li, "Embedding generalized semantic knowledge into few-shot remote sensing segmentation," *arXiv preprint arXiv:2405.13686*, 2024.

[43] S.-A. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao, "Learning orthogonal prototypes for generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 319–11 328.

[44] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. Efros, "Visual prompting via image inpainting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 005–25 017, 2022.

[45] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6830–6839.

[46] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "Seggpt: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.

[47] Y. Liu, C. Jing, H. Li, M. Zhu, H. Chen, X. Wang, and C. Shen, "A simple image segmentation framework via in-context examples," *arXiv preprint arXiv:2410.04842*, 2024.

[48] 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.

[49] S. Kullback, "Kullback-leibler divergence," 1951.

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–37.

[51] 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.

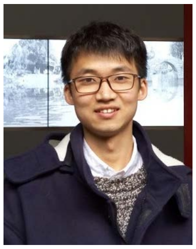
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[54] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5217–5226.

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.

Wenhao fu received the B.E. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2023. He is currently pursuing the M.S. 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.

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/).