Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

DUAL-LEVEL PROTOTYPE ALIGNMENT VIA CROSS-ATTENTION FOR FEW-SHOT REMOTE SENSING SEMANTIC SEGMENTATION

Mustafa Alawadi and Mansoor Fateh

(Received: 28-Jun.-2026, Revised: 8-Aug.-2026, Accepted: 10-Aug.-2026)

ABSTRACT

Analyzing remote sensing imagery relies heavily on accurate pixel-level classification to interpret complex geographical scenes. However, training robust deep-learning models demands densely annotated datasets, which are expensive and time-consuming to create. Few-shot semantic segmentation (FSS) provides a practical alternative by adapting to new categories from limited labeled examples. Despite success in natural images, direct application to remote sensing faces unique challenges, such as massive scale variations, strong visual similarities among densely packed objects and severe background interference. Existing FSS methods typically rely on masked average pooling for prototype extraction, which discards critical spatial geometries and fine-grained details. While some approaches attempt multi-scale or global-local modeling, they lack effective, decoupled alignment between support prototypes and query features, limiting their ability to simultaneously capture broad scene context and precise object boundaries in cluttered aerial views. To address these limitations, we propose DLPANet, a novel dual-level prototype alignment network centered on Prototype-Guided Spatial Attention (PGSA). Our framework constructs global and semantic prototypes from hierarchical backbone features, enabling simultaneous modeling of scene context and fine-grained details. Dual cross-attention modules then align these prototypes with decoupled global and semantic query feature maps, facilitating dynamic and precise feature enhancement that better distinguishes adjacent objects and suppresses background ambiguity. A learnable decoder integrates the enriched features, optimized via Dice loss to handle extreme class imbalance. Evaluated on the iSAID-5i benchmark, DLPANet achieves 36.12% and 41.05% mIoU in 1-shot and 5-shot settings, respectively, consistently outperforming state-of-the-art methods. These results demonstrate that our decoupled dual cross-attention mechanism provides superior prototype-query alignment compared to prior global-local strategies. https://github.com/mansoorfateh83-lab/dlpanet.

KEYWORDS

Few shot, Remote sensing, Segmentation, Prototype alignment, Cross-attention.

1. INTRODUCTION

Remote-sensing (RS) imagery analysis supports many essential applications, from environmental monitoring and urban planning to natural disaster management [1]-[2]. Semantic segmentation is a key part of this process, because it provides the pixel-level categorization needed to truly understand complex geographical scenes [3][4][5]. While deep-learning models perform exceptionally well on these tasks, their success depends heavily on massive, densely annotated datasets [6]-[7]. Creating these pixel-perfect annotations is an expensive and time-consuming bottleneck. Few-shot Semantic Segmentation (FSS) offers a practical solution to this problem, allowing models to identify and segment new object categories using only a few labeled support images [8]-[9].

Although FSS has proven highly effective for natural images, applying it directly to RS imagery is challenging [10]. Aerial images naturally contain extreme scale variations and densely packed objects that look very similar, such as confusing small cars with larger vehicles or ships. Furthermore, the appearance of objects can change drastically depending on the broader aerial context. These unique traits cause standard FSS models to struggle; they often fail to draw precise object boundaries and tend to mistakenly highlight irrelevant parts of the complex geographical background [11].

Researchers have developed specialized FSS frameworks to tackle these domain-specific hurdles. Early foundational models, like Scale-Aware Detailed Matching (SDM [12]), used multiple prototypes and a scale-aware focal loss to penalize large, well-segmented objects, forcing the network to focus on harder, smaller instances [12]. Other architectures, like R^2Net, introduced global rectification and decoupled registration to stop inaccurate prototype activation and separate foreground features from background

M. Alawadi and M. Fateh (Corresponding Author) are with Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran. Emails: mus.al.awadi@gmail.com and mansoor_fateh@shahroodut.ac.ir

clutter [13]. More recent work has pushed these ideas further. For example, some methods build stronger semantic connections by combining magnitude-based pruning with cross-matching and self-matching modules [6]. At the same time, newer frameworks, like SEMPNet, adapt vision-foundation models to turn segmentation into a precise mask classification task [14]. While these advancements certainly improve localization, a major flaw remains. Most current methods rely on masked average pooling to extract condensed object prototypes. This compression technique naturally throws away the spatial geometries and fine-grained semantic details needed for accurate segmentation. Additionally, simple cosine similarity matching between prototypes and query features is usually not strong enough to handle the complex backgrounds and extreme appearance changes found in aerial top-down views. Because of this, we need a better way to align semantic features and handle their interactions.

To move beyond the limitations of single-prototype matching and standard spatial-attention mechanisms in few-shot segmentation, we present a novel framework tailored for remote sensing imagery, built upon a dual-stream Prototype-Guided Spatial Attention (PGSA) strategy. A weight-sharing hierarchical backbone extracts multi-stage features from both support and query images. Rather than relying on a single compressed prototype, we construct two complementary support representations: a semantic prototype isolating target object features and a global prototype capturing broader spatial context. Concurrently, the query feature maps are decoupled across backbone stages into corresponding global and semantic feature maps. To establish dynamic spatial alignment, we apply a dual PGSA mechanism, where the semantic and global prototypes serve as compact semantic query anchors to compute spatially adaptive attention weights across the query key/value maps. The resulting contextually enriched vectors are merged back into the query feature space and processed by a learnable decoder to produce the final segmentation mask. By applying dual-stream PGSA to remote sensing FSS, our architecture effectively addresses severe background clutter and target scale variations.

The primary contributions of this research are established across several key aspects.

1) We propose a novel few-shot semantic-segmentation framework for remote-sensing imagery that leverages dual-level feature representations, capturing both global context and fine-grained semantic details.
2) We design a dual-stream Prototype-Guided Spatial Attention (PGSA) strategy tailored for remote sensing few-shot segmentation, which aligns support prototypes with query features at both global context and fine-grained semantic levels to effectively suppress background interference.
3) We design a learnable decoder that effectively integrates attention-enhanced features to produce accurate and dense segmentation masks.
4) We conduct extensive experiments on the iSAID-5i benchmark, demonstrating that the proposed method consistently outperforms existing state-of-the-art approaches.

The remainder of this paper is organized as follows. Section 2 reviews related work on semantic segmentation in remote sensing, general few-shot semantic segmentation and attention-based prototype learning. Section 3 presents the proposed DLPANet in detail, including the problem definition, overall framework, backbone network, dual-level prototype and query-feature extraction, dual cross-attention mechanism and the learnable decoder with Dice loss. Section 4 describes the experimental setup, reports quantitative and qualitative results on the benchmark and provides comprehensive ablation studies. Finally, Section 5 concludes the paper and discusses potential future directions.

2. RELATED WORK

This section reviews the literature relevant to few-shot semantic segmentation in remote-sensing imagery. We first discuss general semantic-segmentation techniques tailored for remote sensing data. We then review prominent advances in few-shot semantic segmentation for natural images. Subsequently, we focus on domain-specific methods developed for few-shot segmentation in aerial and remote-sensing scenarios. Finally, we highlight recent works on attention mechanisms and prototype-learning strategies that are most closely related to our proposed approach.

2.1 Semantic Segmentation in Remote Sensing

Semantic segmentation has become a fundamental task in remote-sensing image analysis, enabling pixel-level understanding of complex aerial scenes. Early deep-learning approaches were built upon

Fully Convolutional Networks (FCNs) [15], followed by encoder-decoder architectures, such as U-Net [16] and SegNet [17]. To better capture multi-scale context, DeepLab models [18]-[19] introduced atrous convolutions and spatial pyramid pooling.

Transformer-based architectures, such as SETR [20], have demonstrated strong capability in modeling global dependencies. In the remote-sensing domain, recent works increasingly integrate attention mechanisms and multi-scale feature fusion to address large spatial variations. For instance, attention-driven and sparse kernel-based segmentation architectures have been proposed to improve feature representation for high-resolution imagery [21]. Additionally, recent studies continue to explore multimodal and hybrid learning strategies to enhance segmentation robustness under complex sensing conditions [22].

2.2 Few-shot Semantic Segmentation

Prototype-based learning has become the dominant paradigm in FSS, where class-specific representations (prototypes) are extracted from support images and matched with query features.

2.2.1 General Few-shot Semantic Segmentation

Prototype-based learning is a dominant paradigm, where class representations are extracted from support images and matched with query features. PANet [23] introduced prototype-alignment regularization, while PFENet [24] incorporated prior masks for improved guidance.

Subsequent works improved representation learning and feature interaction. ASGNet [25] introduced adaptive prototypes and HSNet [26] modeled dense correspondences using hypercorrelation volumes. More recent developments have explored transformer-based designs and enhanced matching strategies.

In recent years, several studies further extended FSS by integrating additional modalities and learning paradigms. For example, text-guided few-shot segmentation frameworks incorporate textual prompts and graph reasoning to enhance semantic discrimination [27]. Similarly, diffusion-based data augmentation strategies have been proposed to synthesize novel samples and mitigate data scarcity in few-shot settings [28].

2.2.2 Few-shot Semantic Segmentation in Remote Sensing

Adapting FSS to remote-sensing imagery introduces challenges, such as extreme scale variation, dense object distribution and domain shifts. Early adaptations directly applied natural-image methods, but achieved limited success.

To address these issues, specialized approaches have been proposed. SDM [12] introduced scale-aware matching and multi-prototype learning, while R^2Net [13] leveraged global rectification and decoupled feature registration. SEMPNet [14] further reformulated the task as a mask-classification problem using foundation-model priors.

Recent works have significantly advanced this field. For instance, learnable-prototype frameworks introduce adaptive prototype representations using generative-modeling techniques to better capture intra-class variation [29]. Generalized few-shot segmentation has also been explored to handle both base and novel classes simultaneously, providing a more realistic evaluation setting [30]. Furthermore, cross-matching and self-matching mechanisms combined with network pruning have been proposed to improve feature correspondence and efficiency [6].

In addition, knowledge distillation and semi-supervised learning have been integrated in recent works to leverage unlabeled data and improve generalization under extremely limited annotations [31]. These approaches highlight a shift toward hybrid learning paradigms that combine supervision, self-supervision and data synthesis.

2.3 Attention Mechanisms and Prototype Learning

Attention mechanisms, particularly those introduced in the Transformer architecture [32], have become central to modern segmentation frameworks. Cross-attention enables dynamic interaction between support and query features, improving semantic alignment and robustness.

Despite their effectiveness, traditional prototype representations based on masked average pooling often

discard spatial information. Recent works attempt to address this limitation by introducing learnable prototypes, attention-guided refinement and multi-level feature aggregation. Notably, recent studies emphasize combining cross-attention with enhanced prototype learning to better capture both global context and fine-grained structures [6], [29].

Despite substantial progress, existing methods still struggle with complex spatial relationships, background clutter and severe domain-specific variations in remote-sensing imagery. Recent advances increasingly focus on integrating attention mechanisms, generative modeling and hybrid learning strategies. However, effective semantic alignment between support and query features remains an open challenge, motivating the proposed dual cross-attention framework.

Although existing methods have made notable progress, most still rely on simple masked average pooling for prototype extraction, which discards important spatial and fine-grained semantic details. Moreover, their feature-alignment strategies are often insufficient to handle the complex backgrounds and scale variations inherent in aerial imagery. To overcome these limitations, we propose DLPANet, which builds upon the prototype-based paradigm by introducing dual-level (global and semantic) prototypes and a powerful dual cross-attention mechanism for more effective and decoupled feature alignment between support and query images.

3. PROPOSED METHOD

3.1 Problem Definition

In the context of few-shot semantic segmentation for remote-sensing imagery, the primary objective is to accurately predict a dense, pixel-wise classification mask for a given query image utilizing a highly restricted set of annotated support images. Formally, the task is established over a training dataset $\mathcal{D}_{\text{train}}$ and a disjoint testing dataset $\mathcal{D}_{\text{test}}$, ensuring that the target semantic categories evaluated during the testing phase remain entirely unobserved during network optimization. Under a standard 1-way K-shot configuration, a learning episode is constructed by sampling a support set comprising K image-mask pairs, denoted as $\mathcal{S} = \left\{\left(x_s^{(i)}, y_s^{(i)}\right)\right\}_{i=1}^{K}$, alongside a single query image x_q. The support masks $y_s^{(i)} \in \{0,1\}^{H\times W}$ provide the spatial ground-truth for the novel target category. The underlying algorithmic goal is to extract and align optimal semantic concepts from the support set $\mathcal{S}$ to accurately generate a predicted segmentation mask $\hat{y}_q \in \mathbb{R}^{H\times W}$ for the query image x_q.

3.2 Overview

To address the severe scale variations and dense background interference inherent in top-down aerial imagery, a dual-pathway architecture driven by targeted Prototype-Guided Spatial Attention (PGSA) is established. As shown in Figure 1, the framework symmetrically processes the support and query images through a weight-sharing convolutional backbone to extract multi-level hierarchical representations.

Rather than relying on a singular pooled vector, the extracted support features are bifurcated to formulate both a global prototype and a semantic prototype, thereby capturing distinct spatial and contextual abstractions. Correspondingly, the query features are processed into decoupled global and semantic spatial-feature maps. A specialized cross-attention module is subsequently applied independently to both the global and semantic streams. Within this module, the condensed prototypes are utilized to query the flattened spatial-query feature maps, generating highly refined, attention-driven vectors. These enhanced vectors are spatially expanded, concatenated with the original deep query-feature maps and processed by a learnable decoder to yield the final segmentation mask.

3.3 Backbone Network

A pre-trained ResNet-50 architecture is employed as the foundational backbone network to provide robust, multi-scale hierarchical feature extraction. Both the support image and the query image are passed through this feature extractor, with the network weights strictly shared between the two branches to guarantee feature alignment in the embedding space. To comprehensively capture low-level spatial contours and high-level abstract semantics, intermediate feature maps are extracted from multiple stages of the ResNet-50 architecture. Specifically, outputs originating from Block 1 through Block 4 are extracted and subsequently concatenated along the channel dimension. A 1×1 convolutional layer is

then applied to the concatenated tensor to reduce the channel dimensionality, ensuring computational efficiency while preserving the critical spatial and semantic information required for precise dense prediction tasks.

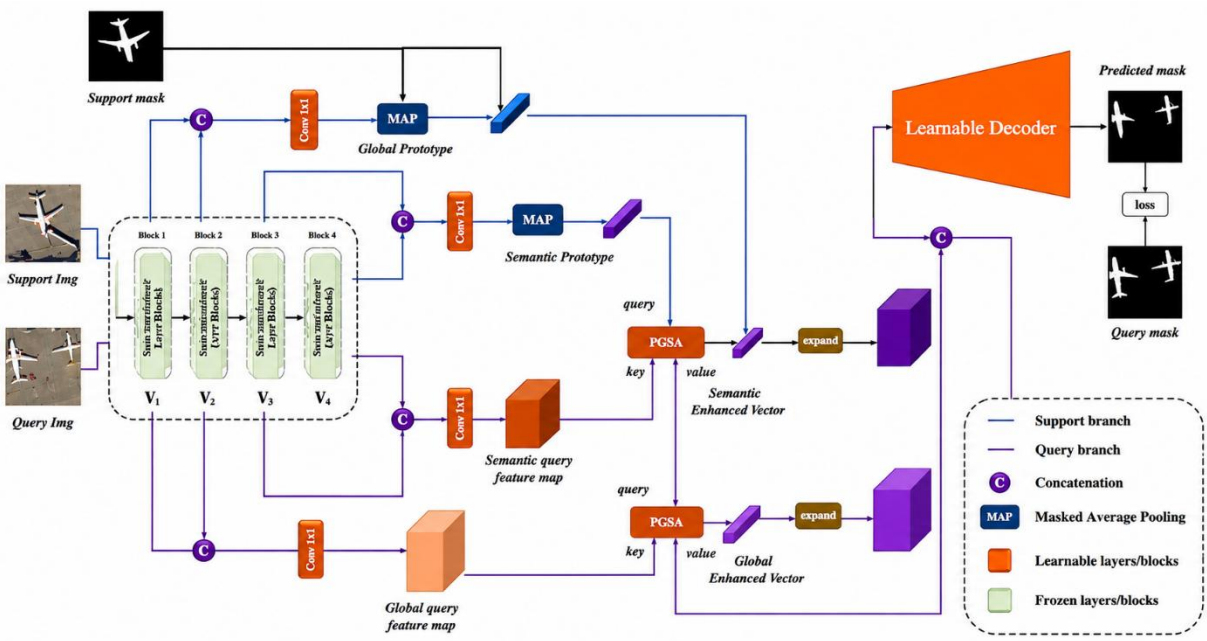

Figure 1. The overview of the proposed method.

3.4 Dual-level Prototype and Query Feature Extraction

To mitigate the degradation of fine-grained geometric details typical of single-prototype compression algorithms, a dual-representation strategy is implemented. The dimensionality-reduced support feature maps are explicitly processed to isolate targeted cues. To prevent background interference, the spatially down-sampled ground-truth support mask is applied across both support-extraction pathways, suppressing extraneous background clutter prior to prototype generation.

Specifically, Masked Average Pooling (MAP) is applied to the foreground-isolated deep support-feature maps (Blocks 3-4) to generate a focused, compact semantic prototype with dimensions of 1×64. Concurrently, a broader global prototype of identical dimensions (1×64) is extracted by applying masked average pooling to the projected early-stage support features (Blocks 1-2), capturing wider geometric and structural arrangements of the target object while excluding irrelevant background regions.

In symmetry, the query-feature maps are decoupled across fundamentally different backbone stages to capture complementary levels of contextual abstraction. Deep backbone stages (Blocks 3-4), which naturally possess a large receptive field and abstract categorical information, are projected *via* a 1×1 convolution $\phi_s(\cdot)$ to generate the semantic query-feature map. Concurrently, early backbone stages (Blocks 1-2), which preserve fine spatial geometries and low-level structural boundaries, are projected *via* an independent 1×1 convolution $\phi_c(\cdot)$ to formulate the global query-feature map.

Here, both $\phi_s(\cdot)$ and $\phi_c(\cdot)$ are parameterized as standard 1×1 2D convolutions (kernel size $= 1\times1$, stride $= 1$, padding $= 0$), each followed by a Batch Normalization layer and a ReLU non-linear activation function. These 1×1 convolutions perform channel-wise alignment from their respective backbone feature dimensions into a unified latent dimension ($d = 64$) without altering the distinct spatial resolutions or receptive fields established by the hierarchical backbone.

3.5 Prototype-guided Spatial Attention (PGSA)

To enable fine-grained and dynamic alignment between support prototypes and query features, DLPANet incorporates a dual Prototype-Guided Spatial Attention (PGSA) mechanism. Rather than relying on static cosine similarity or unweighted feature matching, PGSA computes an adaptive, content-dependent spatial weight map to aggregate query features relative to the support anchor.

As illustrated in Figure 2, we describe the semantic stream for clarity (the global stream follows an identical procedure). Given the semantic prototype $\mathbf{p}_s \in \mathbb{R}^{1\times 64}$ extracted from the support set, we treat it as a single prototype query anchor. Meanwhile, the corresponding semantic query feature map $\mathbf{F}_s^q \in \mathbb{R}^{64\times H'\times W'}$ is flattened into a sequence of spatial tokens $\mathbf{X}_s \in \mathbb{R}^{4096\times 64}$ (where $H' \times W' = 64 \times 64$).

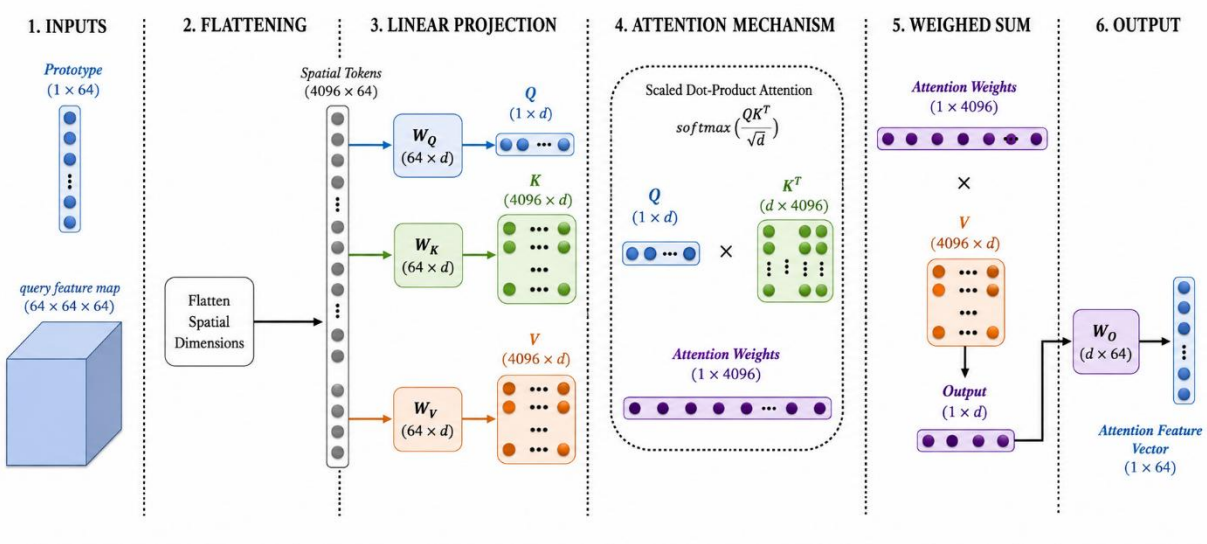

Figure 2. Overview of the proposed Prototype-Guided Spatial Attention (PGSA) module.

Learnable linear projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{64\times d}$ (with $d = 64$ in our implementation) transform the inputs into query, key and value matrices:

$$\mathbf{Q} = \mathbf{p}_s W_Q, \mathbf{K} = \mathbf{X}_s W_K, \mathbf{V} = \mathbf{X}_s W_V \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{1\times d}$ and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{4096\times d}$.

The spatial attention weights are computed *via* scaled dot-product attention:

$$A = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{1\times 4096} \tag{2}$$

These weights represent the relevance of each spatial location in the query image relative to the support prototype anchor. The attended feature vector is subsequently obtained by:

$$\mathbf{O} = A\mathbf{V} \in \mathbb{R}^{1\times d} \tag{3}$$

Finally, a linear projection $W_O \in \mathbb{R}^{d\times 64}$ produces the enhanced semantic feature vector $\mathbf{z}_s \in \mathbb{R}^{1\times 64}$.

Mathematically, operating with a single query token $\mathbf{Q} \in \mathbb{R}^{1\times d}$ formulates a content-based spatial-attention pooling operation. While single-anchor attention pooling and cross-attention mechanisms have been explored in general few-shot segmentation architectures (e.g., BAM [33] and CyCTR [34]), the key distinction in DLPANet lies in how this operation is structured. Rather than applying attention over a single combined feature stream or performing dense token-to-token matching, our design operates in a decoupled dual-pathway fashion.

Specifically, the PGSA process is performed independently across both the semantic and global streams, yielding distinct enhanced feature vectors $\mathbf{z}_s$ and $\mathbf{z}_g$. By decoupling these streams, the model simultaneously attends to fine-grained object-specific details (semantic stream) and broader spatial geometry (global stream) at their native abstraction levels. This dual-stream alignment effectively mitigates background interference and resolves scale ambiguities that standard single-stream attention matching fails to address.

3.6 Learnable Decoder and Loss Function

Following the PGSA phase, the semantic and global attention feature vectors are spatially expanded to match the exact spatial resolution of the intermediate query-feature maps. The expanded global vector,

the expanded semantic vector and the rich query feature maps are concatenated along the channel dimension to form a comprehensive, contextually enriched feature tensor $\mathbf{Z}_{\text{cat}} \in \mathbb{R}^{C_{\text{in}} \times H' \times W'}$.

To reconstruct fine morphological details from this aggregated tensor, our learnable decoder employs a 3-stage progressive up-sampling block. Each stage consists of:

1) A bilinear up-sampling layer with a scale factor of 2 to enlarge feature map dimensions.
2) A standard 3×32D convolutional layer (padding = 1, stride = 1) to refine spatial cues and eliminate upsampling artifacts.
3) A Batch Normalization (BN) layer followed by a ReLU non-linear activation function.

Specifically, the channel dimension is progressively reduced across the three stages ($C_{\text{in}} \rightarrow 128 \rightarrow 64 \rightarrow 32$). Finally, a 1×1 convolution layer followed by a Sigmoid activation produces the final single-channel binary prediction map $\hat{\mathbf{Y}} \in \mathbb{R}^{1 \times H \times W}$ matching the original input-image resolution.

To optimize the network parameters and accurately handle extreme class imbalances frequently encountered in high-resolution remote sensing imagery, the Dice loss function is utilized. By calculating the spatial overlap between the predicted probability mask p and the binary ground-truth mask g, the Dice loss directly maximizes the intersection over union. The optimization objective is formulated as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(p_i \cdot g_i) + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{4}$$

where N denotes the total number of pixels, p_i represents the predicted probability at pixel i, g_i indicates the corresponding binary ground-truth label and ϵ is a small smoothing constant added to prevent division by zero. To ensure numerical stability and exact experimental reproducibility, the smoothing constant is set to $\epsilon = 1 \times 10^{-5}$ in all experiments. This formulation ensures that the network optimization remains highly sensitive to the morphological boundaries of complex geographical targets, strictly penalizing spatial misalignments during the episodic meta-training phase.

4. EXPERIMENTAL RESULTS

4.1 Dataset

To thoroughly evaluate the effectiveness and generalization capability of the proposed DLPANet, we conduct experiments on two widely used few-shot semantic segmentation benchmarks: iSAID- 5^i for remote-sensing imagery and PASCAL- 5^i for natural images.

The iSAID-5^i benchmark is derived from the original iSAID dataset [35], which consists of 2,806 high-resolution aerial images with 655,451 annotated object instances across 15 geospatial categories. Following the standard cross-validation protocol [12], the 15 categories are divided into three folds (Fold 0, 1 and 2). In each evaluation, the model is meta-trained on the base classes from two folds and tested on the five novel classes of the remaining fold. Detailed class partitions are provided in Table 1.

For natural-image evaluation, we employ the PASCAL- 5^i benchmark [36]-[37]. This dataset is constructed from the PASCAL VOC 2012 dataset with additional dense annotations from the SDS dataset. It contains approximately 10,000 images spanning 20 object categories. Following the standard protocol, these categories are split into four folds ($i \in \{0,1,2,3\}$). In each fold, the model is meta-trained on 15 base classes and evaluated on the remaining 5 novel classes.

Experiments on both remote sensing (iSAID- 5^i) and natural image (PASCAL-5^i) domains allow us to validate the robustness and versatility of our dual cross-attention framework across different data distributions and challenges.

Table 1. Class partitions for the iSAID-5^i dataset across three folds.

# Fold	Novel Classes				
0	Ship (C_1)	Storage tank (C_2)	Baseball diamond (C_3)	Tennis court (C_4)	Basketball court (C_5)
1	Ground track field (C_6)	Bridge (C_7)	Large vehicle (C_8)	Small vehicle (C_9)	Helicopter (C_{10})
2	Swimming pool (C_{11})	Roundabout (C_{12})	Soccer ball field (C_{13})	Plane (C_{14})	Harbor (C_{15})

4.2 Implementation Details

The feature-extraction process is anchored by ResNet-50 backbone architectures, which is initialized with weights pre-trained on the ImageNet dataset to leverage established feature representations [13]. The meta-training phase is executed over 20 epochs utilizing a mini-batch size of 8. The base learning rate is established at 0.05. To enhance the generalization capabilities of the model and simulate diverse aerial perspectives, comprehensive data-augmentation techniques are applied to the training set. These techniques include random horizontal flipping, Gaussian blurring, random rotation spanning from -10° to 10° and random scaling operations ranging from a factor of 0.5 to 2.0. Furthermore, all raw input images are uniformly cropped to a spatial resolution of 512 × 512 pixels prior to being processed by the network. To facilitate exact experimental reproducibility, our complete PyTorch implementation and pre-trained weights are publicly available at https://github.com/mansoorfateh83-lab/dlpanet.

To verify the statistical significance and stability of the reported performance gains, all experiments for the proposed model are independently trained and evaluated across three distinct random seeds (100, 202 and 42). The final performance metrics are reported as Mean ± Standard Deviation across all runs.

4.3 Evaluation Metrics

To quantitatively evaluate the segmentation performance of the proposed DLPANet, we adopt two widely used metrics: the mean Intersection over Union (mIoU) and the Foreground-Background Intersection over Union (FB-IoU).

Mean Intersection over Union (mIoU): This serves as our primary evaluation metric. It computes the Intersection over Union (IoU) for each novel class independently and then takes the average across all target classes in the evaluation fold. The mIoU is defined as:

$$\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i} \tag{5}$$

where C is the number of novel classes in the current fold and TP_i, FP_i and FN_i denote the number of true positive, false positive and false negative pixels for the i-th novel class, respectively. This metric provides a balanced and strict assessment by heavily penalizing errors on challenging categories.

Foreground-Background Intersection over Union (FB-IoU): As a complementary class-agnostic metric, FB-IoU evaluates the model's ability to distinguish objects of interest from the background without regard to specific semantic categories. All novel classes are merged into a single "foreground" category. The FB-IoU is computed as:

$$\mathrm{FB-IoU} = \frac{\mathrm{TP}_{\mathrm{fg}}}{\mathrm{TP}_{\mathrm{fg}} + \mathrm{FP}_{\mathrm{fg}} + \mathrm{FN}_{\mathrm{fg}}} \tag{6}$$

where $\mathrm{TP}_{\mathrm{fg}}$, $\mathrm{FP}_{\mathrm{fg}}$ and $\mathrm{FN}_{\mathrm{fg}}$ represent the aggregated true positives, false positives and false negatives for the combined foreground regions across all novel classes.

These two metrics together offer a comprehensive view of both per-class segmentation accuracy and overall foreground localization capability, which is particularly important in complex remote-sensing scenes with severe background interference.

4.4 Comparison with State-of-the-Art

The segmentation efficacy of the proposed dual cross-attention architecture is benchmarked against several prominent few-shot semantic segmentation (FSS) frameworks on the iSAID-5^{i} dataset. This dataset, characterized by its high-resolution aerial imagery and significant scale variations, serves as a rigorous testbed for evaluating the robustness of our dual-prototype generation and cross-attention alignment mechanism.

As detailed in Table 2, our proposed model achieves state-of-the-art results across both backbones. Specifically, with the ResNet-50 backbone, our architecture attains a mean mIoU of $36.12 \pm 0.28\%$ in the 1-shot setting and $41.05 \pm 0.32\%$ in the 5-shot setting, representing a substantial improvement over recent strong baselines, such as SFAPNet [43] and GlocalDualNet [44]. The small standard deviations

($\leq 0.35\%$) across three independent random seeds confirm that the performance gains are statistically significant and represent a reliable signal rather than stochastic noise. The consistent performance gains across all folds underscore the ability of the dual-prototype approach to capture more comprehensive semantic information than traditional single-prototype methods.

The superior localization capability of our method is further validated by the FB-IoU metrics reported in Table 3. FB-IoU provides a category-agnostic evaluation of the model's ability to distinguish foreground objects from complex backgrounds. The proposed approach reaches an FB-IoU of 66.30% and 69.25% for 1-shot and 5-shot scenarios, respectively. These results surpass existing methods, like GlocalDualNet [44], indicating that the dual cross-attention mechanism effectively mitigates foreground-background ambiguity-a common challenge in remote-sensing imagery due to diverse object orientations and cluttered environments.

Table 2. Performance comparison on iSAID -5^i in mIoU under 1-shot and 5-shot settings.

Backbone	Method	1-shot				5-shot			
		Fold-0	Fold-1	Fold-2	Mean	Fold-0	Fold-1	Fold-2	Mean
VGG-16	SDM [12]	24.52	21.01	16.31	20.61	26.73	19.97	26.10	24.27
	NERTNet [38]	25.78	20.01	19.88	21.89	38.43	24.21	28.99	30.54
	DML [39]	30.99	14.60	19.05	21.55	34.03	16.38	26.32	25.58
	TBPN [40]	27.86	12.32	18.16	19.45	32.79	16.28	24.27	24.45
	R ^{2} Net [13]	35.27	19.93	24.63	26.61	42.06	23.52	30.06	31.88
	CSCANet [41]	33.26	20.44	25.98	26.56	40.08	24.15	38.00	34.08
	DSMFNet [42]	30.23	23.35	28.87	27.48	40.73	25.92	37.05	34.56
	SFAPNet [43]	37.12	22.82	26.02	28.65	44.31	26.71	37.06	36.03
	GlocalDualNet [44]	36.40	20.51	27.59	28.17	42.64	23.55	34.74	33.64
ResNet-50	SDM [12]	27.96	21.99	27.82	25.92	28.50	25.23	31.07	28.27
	NERTNet [38]	34.93	23.95	28.56	29.15	44.83	26.73	37.19	36.25
	DML [39]	32.96	18.98	26.27	26.07	33.58	22.05	29.77	28.47
	TBPN [40]	29.33	16.84	25.47	23.88	30.98	20.42	28.07	26.49
	R ^{2} Net [13]	41.22	21.64	35.28	32.71	46.45	25.80	39.84	37.36
	CSCANet [41]	42.30	24.17	36.50	34.32	47.85	30.04	40.32	39.40
	DSMFNet [42]	39.03	24.82	39.57	34.47	43.69	30.81	42.32	38.94
	PGSL [45]	40.23	24.92	37.34	34.17	45.32	29.71	46.80	40.61
	SFAPNet [43]	43.48	25.03	37.18	35.23	48.35	29.63	48.41	42.13
	FSS-TIs [46]	-	-	-	24.51	-	-	-	31.21
	MSDNet [47]	44.35	28.71	35.16	36.07	48.59	29.01	40.64	39.41
	GlocalDualNet [44]	42.54	24.04	38.15	34.91	47.70	26.51	41.62	38.11
	Ours	45.10±0.25	24.95±0.31	38.30±0.28	36.12±0.28	50.15±0.35	29.40±0.29	43.60±0.32	41.05±0.32

Table 3. Performance comparison on iSAID-5 ^{i} in FB-IoU utilizing the ResNet-50 backbone.

Method	1-shot	5-shot
SDM [12]	59.58	59.90
NERTNet [38]	59.97	64.45
TBPN [40]	57.34	58.63
R^2 Net [13]	63.86	66.18
CSCANet [41]	63.56	66.32
DSMFNet [42]	63.52	66.31
SFAPNet [43]	64.47	66.75
GlocalDualNet [44]	65.05	66.97
Ours	66.30	69.25

A detailed class-wise breakdown of the mIoU performance under the 1-shot setting is provided in Table 4. Our model demonstrates leading performance in a majority of the 15 categories (C1-C15). Notably, the model shows significant improvements in challenging classes, such as C1, C2 and C13, which often involve high intra-class variance or small spatial extents. While established methods, like CSCANet [41], exhibit strengths in specific categories, like C12, our dual-prototype approach maintains a more balanced and higher overall mean performance. This suggests that the integration of hierarchical features

and cross-attention alignment provides a more generalized representation, ultimately leading to more reliable few-shot segmentation in diverse aerial scenes.

Table 4. Class-wise performance comparison in mIoU under the 1-shot setting with the ResNet50 backbone.

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Mean
VGG-16 Backbone																
TBPN [40]	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	19.45
R^2 Net [13]	37.82	45.16	26.17	45.30	21.81	24.11	14.38	30.92	12.21	18.03	18.66	25.02	29.64	31.95	17.87	26.61
CSCANet [41]	36.21	43.88	26.01	43.39	16.81	21.80	15.84	26.65	10.58	27.33	9.05	41.67	32.19	31.01	15.97	26.56
SFAPNet [43]	41.48	49.92	31.52	39.95	22.93	26.23	24.23	33.37	13.06	17.21	10.77	27.93	38.58	32.68	19.92	28.65
ResNet-50 Backbone																
TBPN [40]	25.36	41.28	30.67	32.88	16.48	13.48	9.74	27.88	12.52	20.56	11.12	34.31	23.57	40.36	17.98	23.88
R^2 Net [13]	46.87	49.06	30.70	52.86	26.62	24.31	17.25	31.25	13.67	21.73	24.88	46.07	42.29	42.07	21.08	32.71
CSCANet [41]	45.96	47.83	36.62	57.99	23.10	21.27	23.45	29.47	11.98	34.28	18.69	59.39	37.45	46.80	20.17	34.32
Ours	**51.45**	**52.80**	**35.25**	**56.10**	**29.90**	**28.10**	**20.55**	**34.90**	**16.85**	**24.35**	**27.80**	**49.50**	**45.60**	**45.40**	**23.20**	**36.12**

Table 5. Performance on PASCAL-5^i in terms of mIoU and FB-IoU. Numbers in bold represent the best performance.

Backbone	Methods	1-shot						5-shot					
		fold0	fold1	fold2	fold3	mean	FB-IoU	fold0	fold1	fold2	fold3	mean	FB-IoU
ResNet50	PANet [23]	44.0	57.5	50.8	44.0	49.1	-	55.3	67.2	61.3	53.2	59.3	-
	PGNet [48]	56.0	66.9	50.6	50.4	56.0	69.9	57.7	68.7	52.9	54.6	58.5	70.5
	PFENet [24]	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
	HSNet [26]	64.3	70.7	60.3	60.5	64	76.7	70.3	73.2	67.4	67.1	69.5	80.6
	CWT [49]	56.3	62.0	59.9	47.2	56.4	-	61.3	68.5	68.5	56.6	63.7	-
	CyCTR [34]	65.7	71.0	59.5	59.7	64.0	-	69.3	73.5	63.8	63.5	67.5	-
	NTRENet [38]	65.4	72.3	59.4	59.8	64.2	77.0	66.2	72.8	61.7	62.2	65.7	78.4
	ABCNet [50]	62.5	70.8	57.2	58.1	62.2	74.1	64.7	73.0	57.1	59.5	63.6	74.2
	SRPNet [51]	62.8	69.3	55.8	58.1	61.5	-	64.3	70.3	55.1	60.5	62.6	-
	QGPLNet [52]	56.95	68.99	60.1	54.98	60.25	-	61.78	70.96	69.56	58.26	65.14	-
	DRNet [53]	66.1	68.8	61.3	58.2	63.6	76.9	69.2	73.9	65.4	65.3	68.5	81.6
	QGNet [54]	63.4	69.4	55.1	58.4	61.6	-	64.7	71.0	53.6	61.6	62.8	-
	Ours	**64.9**	**69.1**	**61.5**	**61.7**	**64.3**	**77.1**	**70.4**	**73.1**	**68.1**	**67.2**	**69.6**	**81.8**

To further validate the generalization capability of DLPANet beyond aerial imagery, we evaluate its performance on the widely used PASCAL-5^i benchmark for natural images. As shown in Table 5, our method achieves state-of-the-art performance under the ResNet-50 backbone in both 1-shot and 5-shot settings.

Specifically, DLPANet attains a mean mIoU of 64.3% in the 1-shot setting, outperforming previous best methods, such as DRNet and NTRENet. More importantly, it achieves the highest FB-IoU of 77.1%, indicating superior foreground-background separation. In the 5-shot setting, our approach further improves to 69.6% mIoU and 81.8% FB-IoU, surpassing all competing methods.

These strong results on the PASCAL-5^i dataset demonstrate that the proposed dual-level prototype alignment via cross-attention is not limited to the unique challenges of remote-sensing imagery. The consistent superiority across both domains highlights the effectiveness and versatility of our dual cross-attention mechanism in establishing robust support-query feature correspondences.

4.5 Computational Efficiency Analysis

To demonstrate the practical suitability of DLPANet for deployment in high-resolution remote-sensing

applications, we evaluate its computational complexity against leading baseline models. Table 6 summarizes the Number of Parameters (#Params in Millions), Computational Complexity (FLOPs in Giga-FLOPs), Inference Speed (Frames Per Second, FPS) and Mean mIoU (%).

As presented in Table 6, DLPANet achieves superior segmentation accuracy (36.12% mIoU) while maintaining remarkable computational efficiency. Specifically, by utilizing single-token prototype query anchors and lightweight 1 × 1 linear projections rather than dense sequence-to-sequence cross-attention layers, DL-PANet operates with only 24.72M parameters and 20.5 GFLOPs. Furthermore, DLPANet achieves an inference rate of 34.2 FPS, outperforming all baseline models in speed and proving that its performance gains are achieved without imposing prohibitive computational burdens.

Table 6. Computational efficiency comparison under the 1-way 1-shot setting with ResNet-50 backbone.

Method	Params (M) (↓)	FPS (↑)	FLOPs (G) (↓)	Mean mIoU (%) (↑)
PANet	64.31	27.6	24.8	23.13
CANet	69.01	25.3	26.5	21.15
SCL	58.61	31.2	22.3	29.58
PFENet	57.51	33.8	21.7	28.80
DMML	70.31	29.7	24.6	24.31
SDM	76.01	27.3	22.9	25.92
R^2 Net	51.76	32.5	21.3	32.71
GlocalDualNet	60.15	30.8	25.7	34.91
Ours	**24.72**	**34.2**	**20.5**	**36.12**

4.6 Ablation Study

To rigorously validate the individual contributions of the architectural components within the proposed framework, an extensive ablation study is conducted. All ablation experiments are evaluated utilizing the ResNet-50 backbone under the 1-way 1-shot setting on the iSAID-5 ^{i} benchmark, unless otherwise specified. The mean Intersection over Union (mIoU) across all three folds is reported to ensure a comprehensive evaluation.

4.6.1 Effectiveness of the Dual PGSA

To demonstrate the necessity of our dual-pathway feature alignment, the proposed DLPANet is systematically deconstructed and compared against a defined baseline model.

To ensure a fair and controlled comparison, the baseline architecture utilizes the exact same dual-branch ResNet-50 feature extractor and 1×1 linear projection heads (Q_{sem} from Blocks 3-4, Q_{glob} from Blocks 1-2) as DLPANet. However, it completely omits the proposed Prototype-Guided Spatial Attention (PGSA) modules. Instead, it extracts support prototypes *via* standard Masked Average Pooling (MAP) and computes spatial query alignment using conventional cosine similarity matching, followed by a standard 1×1 convolution to generate foreground predictions.

As detailed in Table 7, this baseline model achieves a mean mIoU of 28.51%. Integrating the semantic prototype pathway alongside its dedicated PGSA module (PGSA_{sem} significantly enhances the delineation of object boundaries, raising the performance to 32.40% mIoU (Variant A). Conversely, utilizing only the global prototype pathway with its dedicated attention module ($\text{PGSA}_{\text{glob}}$ yields 31.85% mIoU (Variant B), demonstrating its effectiveness in capturing broader scene context. The optimal segmentation accuracy of 36.12% mIoU is exclusively attained when both global and semantic PGSA pathways are fully integrated. This confirms that holistic scene understanding and targeted semantic alignment function as mutually reinforcing mechanisms.

4.6.2 Analysis of Feature Projection Strategy for Query Branch

To evaluate whether explicitly expanding the spatial receptive field during query projection improves feature decoupling, we compare our lightweight 1 × 1 convolutional projection against an Atrous Spatial Pyramid Pooling (ASPP) module with multi-scale dilation rates {1,6,12} applied to the global query stream.

Table 7. Ablation study on the core architectural components. $PGSA_{sem}$ and $PGSA_{glob}$ denote the semantic and global PGSA modules, respectively.

Model	$PGSA_{sem}$	$PGSA_{glob}$	Fold-0	Fold-1	Fold-2	Mean mIoU
Baseline	×	×	31.20	20.15	34.18	28.51
Variant A	✓	×	38.65	23.40	35.15	32.40
Variant B	×	✓	37.90	22.85	34.80	31.85
Ours	✓	✓	**45.10**	**24.95**	**38.30**	**36.12**

As shown in Table 8, replacing the 1×1 projection with an ASPP module leads to a performance drop across all folds, reducing the mean mIoU from 36.12% to 34.85%. In remote-sensing imagery, where objects such as small vehicles, ships and narrow bridges occupy small spatial extents, ASPP's aggressive multiscale dilation introduces feature over-smoothing and boundary distortion. In contrast, utilizing linear 1×1 channel projection directly on the multi-stage features extracted from the hierarchical backbone preserves crisp spatial geometries while allowing the cross-attention mechanism to dynamically align global and semantic contexts without introducing spatial artifacts.

Table 8. Comparison of different projection strategies for the global query branch.

Projection Strategy	Fold-0	Fold-1	Fold-2	Mean mIoU
ASPP Module	43.80	24.10	36.65	34.85
1×1 Conv Projection	45.10	24.95	38.30	36.12

4.6.3 Analysis of the Loss Function

The choice of loss function plays a crucial role in few-shot remote sensing semantic segmentation due to the severe class imbalance between small foreground objects (e.g., small vehicles, ships and storage tanks) and vast background regions. Standard losses often struggle with this imbalance, leading to biased predictions toward the dominant background class.

Table 9 compares three different optimization strategies. Cross-Entropy (CE) Loss, while simple, tends to produce overconfident predictions and is highly sensitive to class imbalance. Scale-Aware Focal Loss improves performance by focusing more on hard examples, yet it still falls short in effectively managing the extreme foreground-background imbalance common in aerial imagery. In contrast, the proposed Dice Loss directly optimizes the overlap between predicted and ground-truth masks. By emphasizing region-level similarity rather than pixel-wise classification, Dice Loss more effectively penalizes false positives in complex background areas while improving boundary precision. This leads to the best overall performance of 36.12% mean mIoU, demonstrating its superior suitability for high-resolution remote-sensing scenarios.

Table 9. Effect of different optimization loss functions on segmentation accuracy.

Loss Function	Fold-0	Fold-1	Fold-2	Mean mIoU
Cross-Entropy (CE) Loss	42.15	23.10	35.10	33.45
Focal Loss ($\gamma = 2$)	43.80	24.15	36.45	34.80
Dice Loss (Ours)	45.10	24.95	38.30	36.12

4.6.4 Impact of the Decoder Architecture

After enriching query features through dual cross-attention, effectively decoding and upsampling these features into a high-resolution segmentation mask is essential. Table 10 evaluates different decoding strategies.

A simple 1×1 convolution decoder, which directly projects the concatenated features, yields a relatively low performance of 34.10% mIoU due to its limited capacity to recover spatial details lost during feature extraction. Replacing it with an Atrous Spatial Pyramid Pooling (ASPP) module improves multi-scale context modeling, boosting the result to 35.30% mIoU. However, our learnable decoder, composed of

progressive convolutional upsampling layers with skip connections, achieves the highest accuracy of 36.12% mIoU. This decoder better restores fine spatial geometries and sharp object boundaries by gradually refining the feature representations, which proves particularly beneficial for delineating small and densely packed objects in aerial scenes.

Table 10. Comparison of different decoding strategies for the final mask generation.

Decoder Strategy	Fold-0	Fold-1	Fold-2	Mean mIoU
Simple Conv (1 × 1)	43.05	23.50	35.75	34.10
ASPP Module	44.25	24.10	37.55	35.30
Learnable Decoder (Ours)	45.10	24.95	38.30	36.12

4.6.5 Multi-shot Fusion Strategy

In the more practical K-shot setting ($K > 1$), effectively aggregating information from multiple support images is critical for robust prototype construction. To justify our architectural design choices, we internally implemented and evaluated two distinct fusion strategies under the 5-shot setting, as presented in Table 11.

The naive Mean Pooling approach simply averages the prototypes extracted from the five support samples. While straightforward, this method treats all support images equally and fails to exploit complementary information across samples, resulting in a lower performance of 38.50% mean mIoU (a baseline ablation result not included in the main comparison table). In contrast, our proposed Attention-based Fusion strategy routes all support samples through the dual cross-attention modules. This allows the network to dynamically weigh and integrate the most relevant features from different support examples according to their spatial and semantic correspondence with the query image. As a result, the full DLPANet model achieves a substantial improvement of 2.55% in mean mIoU (reaching 41.05%, as reported in Table 2), confirming that adaptive, attention-guided fusion is significantly more effective than simple averaging for multi-shot scenarios.

Table 11. Ablation study on the impact of multi-shot feature fusion strategies under the 5-shot setting.

Fusion Strategy (5-Shot)	Fold-0	Fold-1	Fold-2	Mean mIoU (%)
Mean Pooling (Averaging)	47.90	26.40	41.20	38.50
Attention-based Fusion	50.15	29.40	43.60	41.05

4.6.6 Latent Channel Dimension

To justify our choice of the latent channel dimension d (obtained *via* the initial 1×1 projection convolutions and maintained across the PGSA modules), we evaluated performance and parameter efficiency across $d \in \{32,64,128\}$. As reported in Table 12, setting $d = 32$ leads to a notable drop in segmentation performance (32.40% mean mIoU), as a lower-dimensional embedding lacks the capacity to retain fine-grained spatial and semantic details.

Conversely, increasing d to 128 yields marginal performance gains (36.25% *vs.* 36.12% mIoU) while substantially increasing the trainable parameter count and memory footprint. Setting $d = 64$ achieves an optimal trade-off between representational accuracy and computational efficiency, securing superior performance with minimal parameter overhead.

Table 12. Ablation study on the latent channel dimension d under the 5-shot setting.

Dimension (d)	Fold-0	Fold-1	Fold-2	Mean mIoU (%)	Decoder Parameters (M)
$d = 32$	41.20	21.10	34.90	32.40	1.85M
$d = 64$	45.10	24.95	38.30	36.12	2.42M
$d = 128$	45.25	25.10	38.40	36.25	4.18M

4.7 Visualization

Figure 3 presents qualitative comparisons that highlight the incremental contributions of the proposed components on the iSAID-5 ^{i} benchmark. The baseline model, which relies on standard masked average pooling and simple feature matching, produces fragmented predictions with severe background interference and poor boundary delineation - common failure cases in existing few-shot remote-sensing methods.

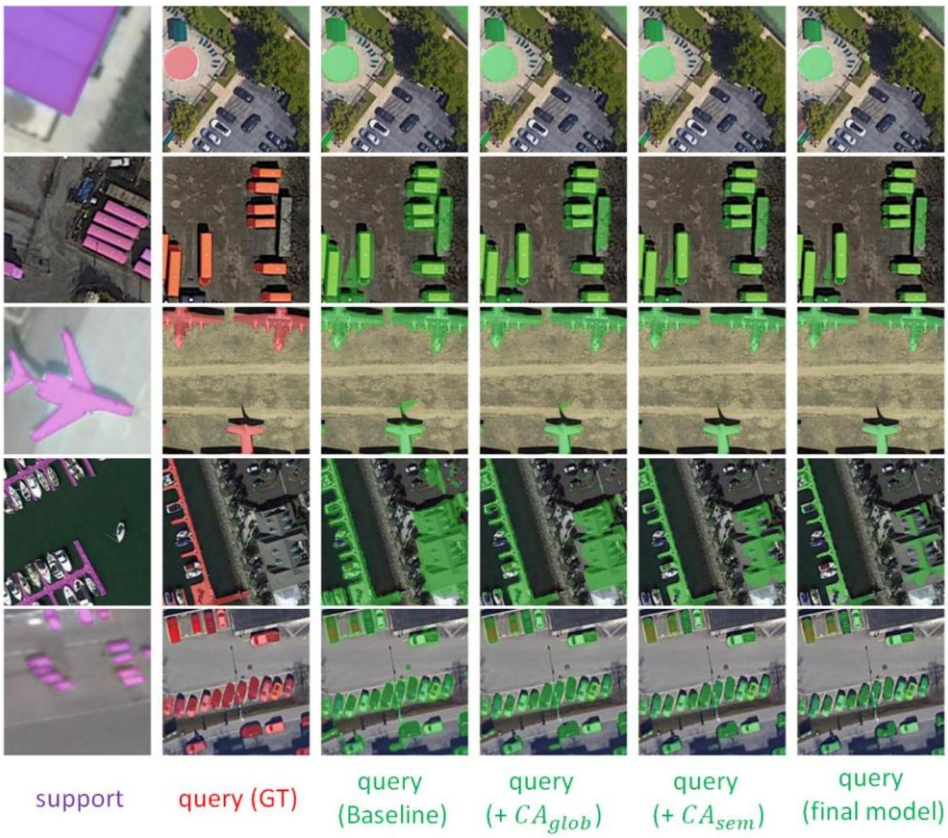

Figure 3. One-shot segmentation predictions on the iSAID-5 ^{i} benchmark. From left to right: support image with mask, query image with ground truth, baseline prediction, baseline with CA_{glob} only, baseline with CA_{sem} only and the full model combining both modules.

When only the global cross-attention module (CA_{glob}) is added, the model demonstrates a strong capability in suppressing large-scale background clutter and better capturing the overall scene context, resulting in cleaner object localization. Conversely, incorporating only the semantic cross-attention module (CA_{sem}) significantly improves fine-grained detail preservation and sharpens object boundaries, effectively reducing boundary overshoot and intra-class confusion between visually similar adjacent objects.

The full DLPANet model, which integrates both global and semantic streams through dual cross-attention, achieves the most accurate segmentation. It successfully handles complex challenges that previous methods struggle with: distinguishing densely packed objects with similar appearances, maintaining precise boundaries across varying scales and robustly suppressing intricate geographical background interference. The final predictions closely align with the ground truth, preserving fine spatial geometries while maintaining contextual coherence. These visual results clearly demonstrate that the decoupled dual cross-attention mechanism provides superior prototype-query alignment capabilities unavailable in conventional single-prototype approaches, enabling more reliable few-shot semantic segmentation in challenging aerial environments.

5. CONCLUSION

In this study, a novel architecture driven by dual PGSA was proposed to address the persistent challenges of extreme scale variations and dense background interference in few-shot remote-sensing semantic

segmentation. To overcome the inherent limitations of conventional single-prototype representations, dual-level representations-specifically, a global prototype and a semantic prototype-were constructed to simultaneously capture broad scene-level context and preserve fine-grained spatial details. These prototypes were actively aligned with decoupled query features through specialized cross-attention modules, which facilitated dynamic feature interaction and ensured refined spatial correspondences.

Extensive evaluations were conducted on the iSAID-5i benchmark, demonstrating that the proposed framework consistently established new state-of-the-art performance standards across existing methodologies. Specifically, exceptional mean Intersection over Union scores alongside superior localization capabilities, as evidenced by the category-agnostic Foreground-Background Intersection over Union metrics, were achieved under both 1-shot and 5-shot evaluation configurations. The critical contributions of the core architectural components, including the dual cross-attention pathways, the integration of the Dice loss function to manage severe class imbalances and the deployment of a learnable decoder for precise dense mask generation, were strictly validated through rigorous ablation studies. Ultimately, it is demonstrated by these quantitative findings that comprehensive hierarchical feature alignment combined with specialized interaction mechanisms significantly enhances the overall robustness and boundary accuracy of semantic segmentation frameworks operating within complex aerial domains.

REFERENCES

[1] B.-Y. Li et al., "Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing," Proc. of the AAAI Conf. on Artificial Intelligence, volume 40, pp. 5982-5991, 2026.

[2] K. Cheikh, R. Aitahcene, A. Toumi and Z. Hammoudi, "Fusion of Deep Learning Architectures for Enhanced Target Recognition on Sar Images," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 9, no. 4, pp. 347-359, 2023.

[3] H. Bi et al., "ViRefSAM: Visual Reference-guided Segment Anything Model for Remote Sensing Segmentation," IEEE Trans. on Geoscience and Remote Sensing, vol. 64, pp. 5606720- 5606720, 2026.

[4] M. Gholami, M. Fateh and A. Fateh, "Text-enhanced Semantic Segmentation *via* Contrastive Language-image Pre-training Guided Multi-modal Feature Fusion with Feature Refinement Approach," Int. Journal of Engineering, vol. 39, no. 6, pp.1422-1437, 2026.

[5] A. Saber et al., "A Lightweight Multiscale Refinement Network for Gastrointestinal Disease Classification," Expert Systems with Applications, vol. 308, Article no. 131029, 2026.

[6] K. Amoafo et al., "Enhancing Few-shot Semantic Segmentation in Remote Sensing through Magnitude-based Pruning," Signal Processing: Image Communication, vol. 143, Article no. 117492, 2026.

[7] G. Hcini and I. Jdey, "Privacy Aware Malaria Detection: U-net Model with K-anonymity for Confidential Image Analysis," Jordan. J. of Computers and Information Tech. (JJCIT), vol. 11, no. 1, pp.85-99, 2025.

[8] A. Fateh, M. R. Mohammadi and M. R. Jahed-Motlagh, "Adapting SAM with a triple-prompt strategy for one-shot Semantic Segmentation," Neurocomputing, vol. 681, p. 133301, 2026.

[9] N. Singhal et al., "Study of Recent Image Restoration Techniques: A Comprehensive Survey," Jordanian J. of Computers and Information Technology (JJCIT), vol. 11, no. 2, pp. 211-237, 2025.

[10] Y. Shi et al., "Twostage Fine-tuning of Large Vision-language Models with Hierarchical Prompting for Few-shot Object Detection in Remote Sensing Images," Remote Sensing, vol. 18, no. 2, p. 266, 2026.

[11] S. Zhang, X. Zhang, Y. Xu and K. Wang, "Few-shot Object Detection on Remote Sensing Images Based on Decoupled Training, Contrastive Learning and Self-training," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 19, pp. 3983-3997, 2026.

[12] X. Yao et al., "Scale-aware Detailed Matching for Few-shot Aerial Image Semantic Segmentation," IEEE Trans. on Geoscience and Remote Sensing, vol. 60, Article no. 5611711, pp. 1-11, 2021.

[13] C. Lang et al., "Global Rectification and Decoupled Registration for Few-shot Segmentation in Remote Sensing Imagery," IEEE Trans. on Geosci. and Remote Sens., vol. 61, Art. no. 5617211, pp. 1-11, 2023.

[14] W. Ao, S. Zheng and Y. Meng, "SEMPNet: Enhancing Few-shot Remote Sensing Image Semantic Segmentation through the Integration of the Segment Anything Model," GIScience & Remote Sensing, vol. 61, no. 1, Article no. 2426589, 2024.

[15] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Proc. of the IEEE Conf. on Computer Vision and Patt. Recog. (CVPR), pp. 3431-3440, Boston, USA, 2015.

[16] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," Proc. of the 18th Int. Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI 2015), pp. 234-241, Munich, Germany, October 5-9, 2015.

[17] V. Badrinarayanan et al., "SegNet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation," IEEE Trans. on Pattern Analy. and Machine Intelli., vol. 39, no. 12, pp. 2481-2495, 2017.

[18] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic

Image Segmentation," arXiv preprint, arXiv: 1706.05587, 2017.
[19] L.-C. Chen et al., "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation," Proc. of the Europ. Conf. on Comp. Vision (ECCV2018), vol. 11211, pp. 801-818, 2018.
[20] S. Zheng et al., "Rethinking Semantic Segmentation from a Sequence-to-sequence Perspective with Transformers," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 6881-6890, Nashville, USA, 2021.
[21] M. Fu et al., "LSKSANet: A Novel Architecture for Remote Sensing Image Semantic Segmentation Leveraging Large Selective Kernel and Sparse Attention Mechanism," Proc. of 2024 IEEE Int. Geoscience and Remote Sensing Symposium (the IGARSS 2024), pp. 8438-8441, Athens, Greece, 2024.
[22] H. Feng et al., "FTransDeepLab: Multimodal Fusion Transformer-based Deeplabv3+ for Remote Sensing Semantic Segmentation," IEEE Trans. on Geosci. and Remote Sens., vol. 63, Article no. 4406618, 2025.
[23] K. Wang et al., "PANet: Few-shot Image Semantic Segmentation with Prototype Alignment," Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision, pp. 9197-9206, Seoul, S. Korea, 2019.
[24] Z. Tian et al., "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.
[25] G. Li et al., "Adaptive Prototype Learning and Allocation for Few-shot Segmentation," Proc. of the 2021 IEEE/CVF Conf. on Comp. Vision and Pattern Recog. (CVPR), pp. 8334-8343, Nashville, USA, 2021.
[26] J. Min, D. Kang and M. Cho, "Hyper-correlation Squeeze for Few-shot Segmentation," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision, pp. 6941-6952, Montreal, Canada, 2021.
[27] F. Yang, J. Chen, D. Luo and H. Lv, "Few-shot Segmentation of Remote Sensing Imagery with Textual Prompts," Int. Journal of Remote Sensing, vol. 47, no. 3, pp. 1040-1060, 2026.
[28] S. A. Immanuel, W. Cho, J. Heo and D. Kwon, "Tackling Few-shot Segmentation in Remote Sensing *via* Inpainting Diffusion Model," arXiv preprint, arXiv: 2503.03785, 2025.
[29] J. Wang et al., "Few-shot Semantic Segmentation on Remote Sensing Images with Learnable Prototype," IEEE Trans. on Geoscience and Remote Sensing, vol. 63, Article no. 5104411, 2025.
[30] C. Broni-Bediako et al., "Generalized Few-shot Semantic Segmentation in Remote Sensing: Challenge and Benchmark," IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024.
[31] H.-L. Jiang et al., "A Few-shot High-resolution Remote Sensing Image Semantic Segmentation Method," Scientific Reports, vol. 16, Article no. 15262, 2026.
[32] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Inf. Process. Syst., vol. 30, 2017.
[33] C. Lang, G. Cheng, B. Tu and J. Han, "Learning What Not to Segment: A New Perspective on Few-shot Segmentation," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8057-8067, New Orleans, USA, 2022.
[34] 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. 21984-21996, 2021.
[35] S. W. Zamir et al., "iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images," Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops, pp. 28-37, 2019.
[36] M. Everingham, L. V. Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," Int. J. of Computer Vision, vol. 88, pp. 303-338, 2010.
[37] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji and J. Malik, "Semantic Contours from Inverse Detectors," Proc. of the IEEE 2011 Int. Conf. on Computer Vision, pp. 991-998, Barcelona, Spain, 2011.
[38] Y. Liu et al., "Learning Non-target Knowledge for Few-shot Semantic Segmentation," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 11573-11582, New Orleans, 2022.
[39] X. Jiang, N. Zhou and X. Li, "Few-shot Segmentation of Remote Sensing Images Using Deep Metric Learning," IEEE Geoscience and Remote Sensing Letters, vol. 19, Article no. 6507405, pp. 1-5, 2022.
[40] 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, Article no. 126201, 2023.
[41] G. Liang, F. Xie and Y.-R. Chien, "Class-aware Self-and Cross-attention Network for Few-shot Semantic Segmentation of Remote Sensing Images," Mathematics, vol. 12, no. 17, pp. 2761, 2024.
[42] X. Qi et al., "DSMF-Net: Dual Semantic Metric Learning Fusion Network for Few-shot Aerial Image Semantic Segmentation," IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 853-864, 2024.
[43] S. Wang, M. Meng, M. Chang and Z. Yang, "Learning Spatial-frequency Adaptive prototype for Remote Sensing Few-shot Segmentation," Knowledge-based Systems, vol. 349, Article no. 116464, 2025.
[44] H. Tang, Y. Jia, J. Cheng, Y. Mu, Y. Wu and X. Yao, "Glocaldualnet: Disentangling Scale and Representation for Few-shot Remote Sensing Segmentation," IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 19, pp. 8018-8030, 2026.
[45] Q. Wang, H. Jiang, J. Feng, G. Zhang and J. Yin, "Prototype-guided Structural Learning from Visual Foundation Model for Few-shot Aerial Image Semantic Segmentation," Proc. of the 2024 IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS 2024), pp. 8433-8437, Athens, Greece, 2024.
[46] H. Ni, Q. Liu, X. Niu, D. Hong, L. Zhao and H. Guan, "Cross-domain Few-shot Segmentation *via*

Ordinary Differential Equations over Time Intervals," arXiv preprint, arXiv: 2509.01299, 2025.

[47] A. Fateh et al., "MSDNet: Multi-scale Decoder for Few-shot Semantic Segmentation *via* Transformer-guided Prototyping," Image and Vision Computing, vol. 162, p. 105672, 2025.

[48] 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," Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision, pp. 9587-9595, Seoul, S. Korea, 2019.

[49] Z. Lu et al., "Simpler Is Better: Few-shot Semantic Segmentation with Classifier Weight Transformer," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision, pp. 8741-8750, 2021.

[50] Y. Wang, R. Sun and T. Zhang, "Rethinking the Correlation in Few-shot Segmentation: A Buoys View," Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 7183-7192, Vancouver, Canada, 2023.

[51] H. Ding, H. Zhang and X. Jiang, "Self-regularized Prototypical Network for Few-shot Semantic Segmentation," Pattern Recognition, vol. 133, Article no. 109018, 2023.

[52] Y. Tang and Y. Yu, "Query-guided Prototype Learning with Decoder Alignment and Dynamic Fusion in Few-shot Segmentation," ACM Trans. on Multimedia Computing, Communications and Applications, vol. 19, no. 2s, Article no. 84, pp. 1-20, 2023.

[53] Z. Chang et al., "DRNet: Disentanglement and Recombination Network for Few-shot Semantic Segmentation," IEEE Trans. on Circuits and Sys. for Video Tech., vol. 34, no. 7, pp. 5560-5574, 2024.

[54] W. Liu, Z. Wu, H. Ding, F. Liu, J. Lin, G. Lin and W. Zhou, "Exploiting Independent Query Information for Few-shot Image Segmentation," Displays, vol. 91, Article no. 103179, 2026.

ملخص البحث:

يعتمد تحليل صور الاستشعار عن بعد بشكلٍ كبيرٍ على التّصنيف الدّقيق لتفسير المشاهد الجغرافية المعقّدة. ومع ذلك، فإنّ تدريب نماذج تعلّم عميق قوية يتطلّب مجموعات بيانات ذات تعليقات توضيحية كثيفة (موسومة بدقّة)، وهي عملية مكلفة وتستغرق وقتاً طويلاً. وتوفّر تقنية التّجزئة الدّلالية باستخدام عينات قليلة بديلاً عملياً من خلال التّكيف مع فئات جديدة استناداً الى أمثلةٍ محدودةٍ وموسومة. ورغم النّجاح المحقّق في الصُّور الطّبيعية. فإنّ التّطبيق المباشر لهذه التّقنية في مجال الاستشعار عن بُعدٍ يواجه تحدّياتٍ فريدةً مثل التّباين الهائل في الأحجام، والتّشابه البصري القوي بين الكائنات المتراصّة بكثافة، والتّداخل الشّديد من الخلفية.

وتعتمد طرق التّجزئة الدّلالية باستخدام عينات قليلة عادةً على تقنية التّجميع المتوسّط المقنّع لاستخراج النّماذج الأولية، ممّا يؤدّي إلى فقدان معلوماتٍ هندسيةٍ مكانيةٍ حاسمةٍ وتفاصيلٍ دقيقةٍ. وفي حين تحاول بعض الأساليب استخدام النّمذجة متعدّدة المقاييس أو النّمذجة الشّاملة (العالمية والمحلية)، فإنّها تفتقر إلى آلية محاذاةٍ فعّالةٍ ومنفصلة بين نماذج الدّعم الأولية وميزات الاستعلام، مما يحدّ من قدرتها على إلتقاط سياق المشهد العامّ وحدود الكائنات الدّقيقة في آنٍ واحدٍ ضمن المشاهد الجوّية المزدحمة. ولمعالجة هذه القيود، نقترح شبكةً مبتكرة لمحاذاة النّماذج الأولية ثنائية المستوى ترتكز على آلية الانتباه المكاني الموجّه بالنّماذج الأولية. ويقوم إطار العمل المقترح ببناء نماذجٍ أوليةٍ عالميةٍ ودلاليةٍ انطلاقاً من ميزات هيكلية هرمية، مما يتيح نمذجة متزامنة لسياق المشهد والتّفاصيل الدّقيقة. بعد ذلك، تقوم وحدات الإنتباه المتقاطع المزدوج بمحاذاة هذه النّماذج الأولية مع خرائط ميزات الاستعلام العالمية والدّلالية المنفصلة، مما يسهّل تعزيز الميزات بشكل ديناميكي ودقيق، ويساعد بشكلٍ أفضل على التّمييز بين الكائنات المتجاورة والحدّ من غموض الخلفية. وأخيراً، يقوم مفكّك ترميز قابل للتّعلّم بدمج الميزات مع تحسين الأداء باستخدام دالة الخسارة للتّعامل مع مشكلة عدم التّوازن الشّديد بين الفئات. وعند تقييمها، تحقّق الشّبكة المقترحة تفوّقاً على أحدث الأساليب المتطوّرة في هذا المجال.