Few-Shot Remote Sensing Scene Classification via Parameter-free Attention and Region Matching

Yuyu Jia[a,1], Chenchen Sun[a,1], Junyu Gao[a] and Qi Wang[a,*]

[a] *School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China*

ARTICLE INFO

Keywords:
Remote Sensing
Few-shot Learning
Region Attention

ABSTRACT

Few-shot remote sensing scene classification, a pivotal task in geospatial scene understanding, has drawn considerable attention as a means to address annotation scarcity in Earth observation. While recent advancements exploit metric-based learning, conventional methods that rely on global feature aggregation, *e.g.,* prototype networks, often entangle target objects with cluttered backgrounds—an inherent limitation given the heterogeneous land-cover elements in remote sensing imagery. Although parametric attention mechanisms partially alleviate this issue, they tend to overfit base-class patterns, limiting adaptability to novel categories with diverse intra-class variations. To tackle these challenges, we propose the Parameter-free Attention with Selective Region Matching (PA-SRM) framework, which integrates two cascaded components: a parameter-free region attention module and a local description classifier. The former dynamically emphasizes discriminative regions by jointly assessing semantic similarity and spatial coherence. At the same time, the latter explicitly employs entropy-aware multi-region voting to suppress residual background interference in queries. Extensive experiments on NWPU-RESISC45, WHU-RS19, UCM, and AID datasets validate the superiority of PRA-SRM and the effectiveness of its components.

1. Introduction

Remote sensing image classification (Dou et al., 2024; Liu et al., 2024; Xu et al., 2023), a cornerstone in geospatial imagery analysis, endeavors to categorize remote scenes through their unique visual signatures. Expansive remote sensing datasets have catalyzed remarkable progress in fully supervised deep learning frameworks for image interpretation. Nevertheless, these methodologies remain inherently reliant on large-scale annotated training corpora. This constraint poses significant challenges when deploying in unexplored geographical regions or identifying emerging land-use patterns. Inspired by human cognition, which swiftly adapts to novel visual concepts through sparse exemplars, Few-Shot Learning (FSL) emerges as a transformative solution, enabling robust identification of unseen categories with minimal supervisory cues (Qiu et al., 2024; Zhang et al., 2023b; Jia et al., 2023a,c). This approach serves as a crucial bridge between theoretical advancements in machine perception and the practical demands of global environmental monitoring ecosystems.

Unlike conventional classification tasks, the training and testing datasets in FSL contain entirely disjoint classes. Given a limited set of labeled images from novel categories (termed the support set), FSL aims to accurately classify unseen images from the same categories (termed the query set). It predominantly employs a meta-learning framework to enable rapid model adaptation to novel scenarios with limited supervision. Current state-of-the-art techniques, such as prototype-based methods (Snell et al., 2017; Jia et al., 2024; Wang et al., 2024; Jia et al., 2025) and relation networks (Sung et al., 2018; Wu et al., 2019; He et al., 2020; Ding et al., 2019) emphasize *global feature metric learning*. This is often achieved through spatial aggregation of image representations, typically via global average pooling. It is then followed by a similarity assessment using static metric functions or learnable metric modules. Nonetheless, these methods face considerable challenges when applied to remote sensing imagery due to inherent domain disparities in visual characteristics.

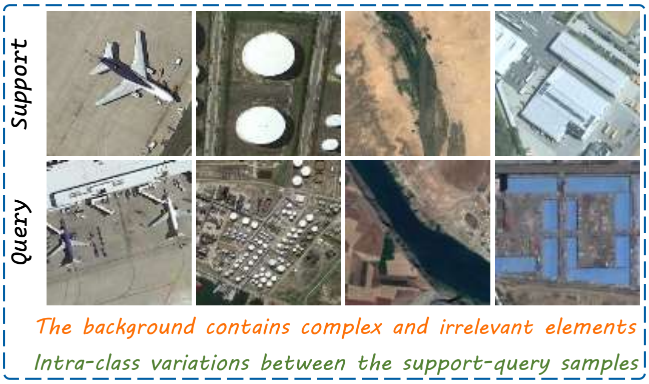

Support
Query
The background contains complex and irrelevant elements
Intra-class variations between the support-query samples

Figure 1: Two inherent attributes of remote sensing images.

The first issue lies in the ***intricate background elements*** present in remote sensing images. As illustrated in Fig. 1, a query image of an airplane often includes irrelevant elements such as ground vehicles, aprons, and terminal buildings. Global feature metrics, which aggregate spatial information indiscriminately, inherently capture these background elements, resulting in biased similarity measurements (Yao et al., 2021; Cheng et al., 2021). To mitigate this issue, recent studies have introduced parametric attention mechanisms to emphasize discriminative regions (Xu et al., 2024a; Zeng & Geng, 2022). Unfortunately, this strategy faces another fun-

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

jyy2019@mail.nwpu.edu.cn (Y. Jia); sunchenchen@mail.nwpu.edu.cn (C. Sun); gjy3035@gmail.com (J. Gao); crabwq@gmail.com (Q. Wang)

[1] The two authors contribute equally to this work.

damental hurdle in remote sensing: ***pronounced intra-class variations***. As shown in Fig. 1, storage tanks exhibit substantial appearance changes under different viewing angles and lighting conditions. Parametric attention modules, trained on a limited set of base classes during meta-training, often overfit to specific local patterns (e.g., circular roofs) and fail to adapt their focus to novel classes with distinct visual characteristics. This dilemma raises a critical question: *Can we develop a parameter-free attention mechanism capable of reducing background interference while dynamically localizing discriminative regions between support-query pairs?*

Our work responds to this question by proposing a Parameter-free Attention with Selective Region Matching (PA-SRM) framework for few-shot remote sensing classification. As depicted in Fig. 2, PA-SRM operates through two synergistic stages: (i) ***Parameter-free Region Attention*** (PRA): Instead of introducing trainable parameters, this module partitions the image into fragmented regions and computes attention weights based on semantic similarity and spatial consistency. By leveraging these criteria, PRA effectively suppresses background noise and adaptively highlights discriminative regions—such as aircraft fuselages obscured by shadows or partially occluded storage tanks. Unlike pixel-level attention methods that rely on fine-grained correlations, PRA captures region-level dependencies, thereby preserving structural integrity and enhancing the model's capacity to interpret both spatial and semantic contexts. Operating at the region level while avoiding parameterized learning constraints, PRA achieves a balance between computational efficiency and contextual awareness, enabling robust generalization to novel classes with substantial intra-class variability. (ii) ***Local Description Classifier*** (LDC): During classification, LDC selects query regions with high discriminative power relative to support prototypes and integrates multi-region similarity scores through entropy-aware voting. This strategy strengthens consensus patterns while reducing the influence of outlier responses caused by residual background interference.

To summarize, the key contributions are as follows:

1) We identify two primary challenges in FSL for remote sensing imagery: intricate background elements and pronounced intra-class variations. To address these, we propose the PA-SRM framework, which performs better than competing algorithms.
2) We introduce a parameter-free region attention module that jointly accounts for semantic similarity and spatial consistency across regions. Avoiding additional trainable parameters efficiently captures discriminative regional features while suppressing irrelevant background interference.
3) We propose a local descriptor-based classifier for the metric decision stage to further mitigate the influence of residual background noise on classification.

2. Related Work

2.1. Remote Sensing Image Classification

Remote sensing image classification has evolved significantly with the advent of machine learning and deep learning. Conventional methods primarily relied on handcrafted spectral and spatial features, such as texture descriptors (e.g., Gray-Level Co-occurrence Matrix) and spectral indices (e.g., NDVI), alongside classifiers like Support Vector Machines (SVMs) and Random Forests (Mountrakis et al., 2011). While these approaches are effective for simple landscapes, they often falter in complex scenes due to their limited feature representation and susceptibility to intra-class variability.

The advent of deep learning revolutionized the field. Convolutional Neural Networks (CNNs), such as ResNet (He et al., 2016) and VGG (Simonyan, 2014), demonstrate superior performance by automatically learning hierarchical features from raw pixels (Zhang et al., 2016). Transfer learning emerges as a key strategy to address the challenges of limited labeled data in remote sensing, where models pre-trained on large-scale natural image datasets (e.g., ImageNet) were fine-tuned on smaller remote sensing benchmarks (Penatti et al., 2015). Recent works further incorporate attention mechanisms (Fu et al., 2019; Zhao et al., 2024b,a; Shen et al., 2022) and multi-scale fusion techniques (Li et al., 2021c; Liu et al., 2025; Zhang et al., 2023a; Zheng et al., 2024) to enhance discriminative feature extraction, particularly for objects with varying sizes and orientations.

Despite progress, conventional deep learning approaches remain data-hungry and often underperform in low-data regimes, such as rare land-cover categories or geographically restricted regions. To mitigate this, researchers explore data augmentation (Wang et al., 2021; Yu et al., 2017; Stivaktakis et al., 2019) and semi-supervised learning (Han et al., 2018; Miao et al., 2022; Huang et al., 2023). However, these methods still necessitate considerable annotation efforts and overlook the generalization to novel scenes.

2.2. Deep Metric Learning

Deep Metric Learning (DML) maps input samples to a feature space where pairwise distances reflect their semantic similarity. Specifically, DML seeks to minimize the distance between features of the same class while maximizing the separation between features of different classes. Through carefully designed sampling strategies and loss functions, DML constructs a discriminative embedding space, effectively addressing challenges of low inter-class variance and high intra-class variability.

Early applications of DML in remote sensing utilized Siamese networks (Bertinetto et al., 2016), which optimize contrastive loss to distinguish image pairs. This approach demonstrates exceptional performance in fine-grained land-cover classification. Building upon this, triplet networks (Hoffer & Ailon, 2015) enhance feature representation by incorporating relative similarities within image triplets (anchor, positive, negative), thus enabling a more precise delineation of class boundaries. Recent innovations integrate attention mechanisms and multi-scale feature extraction into DML frameworks, boosting their robustness and adaptability (Wang et al., 2019). These techniques skillfully capture spatial patterns and spectral variations inherent in remote sensing imagery. Notably, attention-based DML models have led to

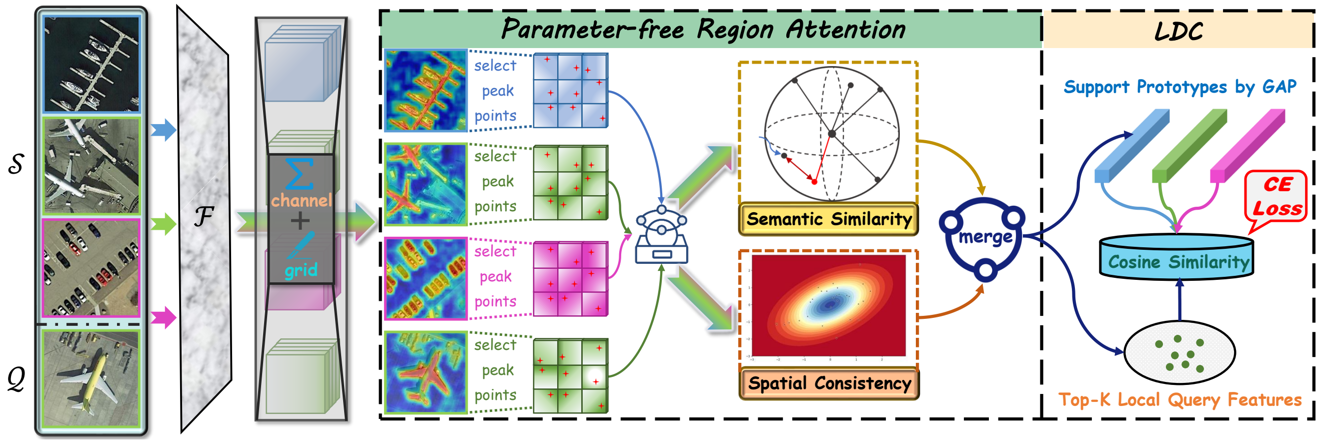


Figure 2: The overall pipeline of the proposed PA-SRM. For clarity, we use a 3-way 1-shot setting as an example. The framework consists of two main modules: **(a) Parameter-free Region Attention (PRA)**, which partitions input images into fragmented regions and computes attention weights based on semantic similarity and spatial consistency, adaptively suppressing background noise and highlighting discriminative regions without introducing trainable parameters. **(b) Local Description Classifier (LDC)**, which leverages local prototype construction and entropy-aware similarity voting to achieve robust classification by aggregating multi-region similarity scores while mitigating residual background interference. The synergy of these two modules effectively addresses the challenges of intricate backgrounds and pronounced intra-class variations in remote sensing imagery.

considerable improvements in handling complex scenes with objects of diverse sizes and orientations (Kim et al., 2018).

Despite these advancements, DML in remote sensing continues to confront enduring challenges, particularly the scarcity of labeled data and the high dimensionality of imagery. To overcome these limitations, researchers have delved into few-shot learning (Jia et al., 2023b) and domain adaptation (Tuia et al., 2016), which allow DML to generalize more effectively to novel and unseen environments. Furthermore, self-supervised learning (Wang et al., 2022b) has emerged as a promising avenue, harnessing large-scale unlabeled datasets to minimize the dependence on extensive annotation while still achieving competitive performance.

2.3. Few-shot Learning

Few-Shot Learning (FSL) seeks to generalize to novel tasks using limited labeled samples, which primarily fall into pre-training-based methods (Zeng et al., 2021; Wang et al., 2022a; Xiong et al., 2021; Gong et al., 2022) and meta-learning-based paradigms (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Zhang et al., 2020). Pre-training-based methods utilize large-scale datasets to train backbone networks for extracting transferable features, whereas meta-learning approaches focus on optimizing models to generalize effectively to unseen classes by training on a diverse set of tasks with limited samples. Moreover, several studies show that incorporating class-descriptive information encoded by CLIP into the construction of visual prototypes can significantly enhance the robustness of metric-based classification (Chen et al., 2023; Zhang et al., 2024; Shen et al., 2024).

Recent advancements have extended FSL methodologies to the remote sensing community. DUSN (Qin et al., 2024) uses class subspaces as a metric benchmark, reducing classification confusion through inter-class and intra-class constraints. HSL-MINet (Jia et al., 2023b) improves model decision boundaries through hard sample learning and multi-view integration, addressing rotational insensitivity and uncertainty in difficult sample distributions. Given the rich distribution of land-cover types and background elements in remote sensing imagery, recent work has employed attention mechanisms to capture discriminative regions. For example, DLA-MatchNet (Li et al., 2021b) employs an attention mechanism to explore the semantic relationships both between channels and across spatial dimensions, thereby capturing discriminative regions. ACL-Net (Xu et al., 2024b) introduces an attention-based contrastive learning network to align and enhance target image features, emphasizing the target regions and augmenting the network's feature extraction capability. Although highly effective, attention mechanisms involving many parameters inevitably introduce base-class biases, diminishing the model's ability to generalize to novel classes. Our work focuses on a parameter-free attention design that explicitly selects key regions within the query image for class decision-making.

3. Method

3.1. Preliminaries

3.1.1. Problem Definition

The goal of FSL is to extend the model's classification ability from the training set $\mathcal{D}_{\text{train}}$ to the test set $\mathcal{D}_{\text{test}}$, where the categories in these sets are completely distinct. Both datasets contain multiple sample-label pairs, $\boldsymbol{I}_*$ and $\boldsymbol{y}_*$. Following the episodic learning mechanism (Snell et al., 2017) that has been widely adopted previously, we treat an episode as a basic task. Each episode involves classifying samples in the *query set* $\mathcal{Q}$ based on a few given labeled samples in the *support set* $\mathcal{S}$. The latter typically consists of N classes, each providing K samples, commonly referred to as an N-way K-shot setting. The model is trained on $\mathcal{D}_{\text{train}}$ and tested on

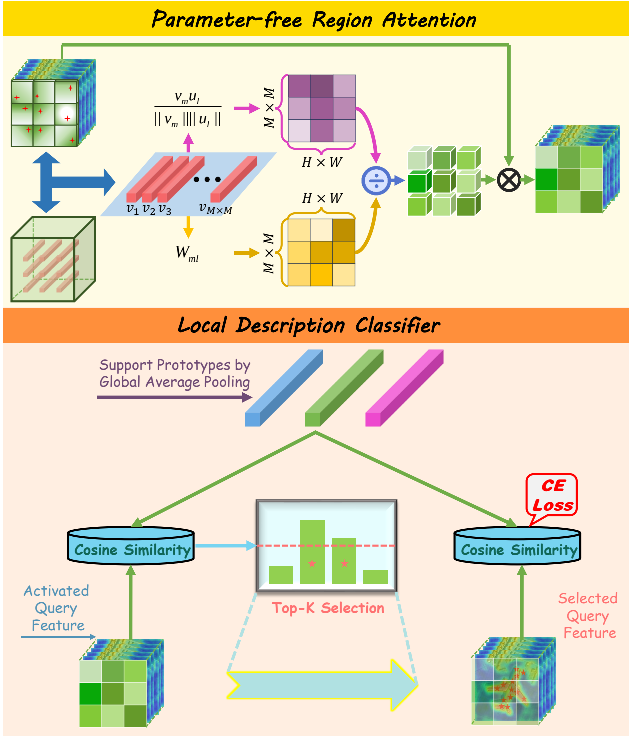


Figure 3: Configuration diagram of PRA and LDC modules. Since the PRA module processes all images independently, we use a query feature map as an example to illustrate the workflow.

$\mathcal{D}_{\text{test}}$ to evaluate its ability to generalize to new categories.

3.1.2. Baseline

We adopt ProtoNet (Snell et al., 2017) as our baseline model. Its core lies in the fact that the prototype of each class is obtained by averaging the global features of the K samples belonging to that class in the support set. Specifically, $N \times K$ support images and $|Q|$ query images are fed into the feature extractor$\mathcal{F}(\cdot)$, resulting in support features and query features $\{F^s, F^q\} \in \mathbb{R}^{D\times H\times W}$, where D, H, and W represent the number of channels, height, and width of the feature map, respectively. The support features of clacategoryss c are compressed through a global average pooling operation $GAP(\cdot)$ and averaged to obtain the category prototype:

$$p_c = \frac{1}{K} \sum_{(F_i^s, y_i)\in S} GAP(F_i^s) \cdot \mathbb{I}(y_i = c), \tag{1}$$

where $p_c \in \mathbb{R}^D$ and $\mathbb{I}(\cdot)$ is the indicator function. Next, the probability of each query sample belonging to a class c can be obtained based on the distance between the query sample and the N prototype vectors:

$$\mathbf{P}(y_i = c | I_i) = \frac{\exp(-d(GAP(F_i^s), p_c))}{\sum_{\acute{c}} \exp(-d(GAP(F_i^s), p_{\acute{c}}))}, \tag{2}$$

where $d(\cdot)$ denotes the distance calculation function, such as Euclidean distance or cosine similarity.

Finally, the classification loss function in the form of cross-entropy can be constructed as:

$$J_{base} = -\frac{1}{|Q|} \sum_{(I_i, y_i)\in Q} \log(\mathbf{P}(y_i = c | I_i)). \tag{3}$$

3.2. Method Overview

As shown in Fig. 2, the proposed Parameter-free Attention with Selective Region Matching (PA-SRM) framework, based on ProtoNet, consists of two core cascaded modules: the Parameter-free Region Attention (PRA) and Local Description Classifier (LDC) modules, designed for few-shot remote sensing scene classification. Specifically, the support and query images are fed into a shared backbone network, *e.g.*, ResNet12 (Oreshkin et al., 2018), to obtain feature maps. The PRA module 3.3 processes support and query feature maps independently, computing region-level attention weights in a parameter-free manner based on semantic similarity and spatial consistency. This approach emphasizes discriminative regions while mitigating base-class bias. Subsequently, the LDC module 3.4 follows, selecting discriminative local regions in the query image for classification decisions, further eliminating residual background noise interference.

3.3. Parameter-free Region Attention

In the literature of FSL, the GAP operation in Eq. 1 is widely used for global semantic similarity matching in the metric space. However, in remote sensing images rich in land cover details, the background often includes elements unrelated to the scene category. Direct global compression of feature maps introduces background interference, degrading classification performance. To mitigate this issue, various approaches have been proposed, including attention mechanisms (Zeng & Geng, 2022) to identify discriminative regions and multi-prototype strategies (Li et al., 2021a) to represent regional semantics better. Despite their effectiveness, these approaches introduce significant training parameters, increasing base-class bias, while pixel-level correlation computations disrupt the structural integrity of target regions. For remote sensing data with high intra-class variability, such limitations hinder the model's generalization capability.

To strike a balance between mitigating background interference and enhancing the model's generalization, we designed the PRA mechanism. As illustrated in Fig. 3, it operates through a hierarchical process: (i) The feature map is partitioned into local sub-regions, with representative points identified within each region (3.3.1); (ii) jointly evaluating semantic similarity and spatial consistency to compute attention coefficients (3.3.2); (iii) aggregating region-based features using the computed attention coefficients to produce an enhanced feature map (3.3.3).

3.3.1. Region Representative Points Localization

Starting from the feature map output by the backbone network, all channels are aggregated, retaining only the spatial dimensions, thus visibly highlighting regions with high-density information:

$$\mathbf{A} = \sum_{d=1}^{D} F[d, :, :], \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{H\times W}$, $F[d, :, :]$ represents the spatial value of the feature map F in the d-th channel. Since each feature map is processed independently, the index labels are ignored. Then, we evenly divide $\mathbf{A}$ into $M \times M$ grid sub-regions and select the peak points of each sub-region as representative points to explore the local land cover information. The coordinates of the representative points can be expressed as $\{G_m(x, y) | m = 1, \cdots, M \times M\}$.

3.3.2. *Attention Coefficient Computation*

After identifying representative points of sub-regions based on the peaks of the activation map, the question arises: *how can these points truly become representative*? We propose explaining this from two criteria.

Each representative point is expected to aggregate features with high semantic relevance, consistent with the criterion of ***semantic similarity***. First, the semantic vector v_m of the representative point $G_m(x, y)$, mapped onto the original feature map F, is represented as:

$$v_m = F[:, G_m(x), G_m(y)], \tag{5}$$

where $v_m \in \mathbb{R}^D$, $G_m(x)$ and $G_m(y)$ represent the coordinates on the x and y axes, respectively. The semantic weight coefficient T_{ml} is obtained by computing the similarity between v_m and the semantic vectors at other locations in the feature map F:

$$T_{ml} = \frac{v_m u_l}{\parallel v_m \parallel \parallel u_l \parallel}, \tag{6}$$

where $l = 1, \dots, H \times W$, and u_l is the channel feature at the l-th coordinate index in F.

New issues arise here: rigidly grid-based partitioning may fragment the complete target body, disrupting its spatial structural information. Additionally, an exclusive focus on semantic similarity may misactivate background elements that exhibit local similarity to the target area yet lack spatial coherence. For example, this could lead to the erroneous identification of dispersed runway segments as an aircraft's fuselage. Based on this, we propose another criterion—***spatial consistency***. Specifically, we introduce a neighborhood spatial weight based on Euclidean distance, guiding the model to prioritize regions within the activation map $\mathbf{A}$ that are proximal to the representative points:

$$W_{ml} = \frac{1}{1 + \left\| G_m(x, y) - G_l(x, y) \right\|_2}, \tag{7}$$

where $\|\cdot\|_2$ denotes the Euclidean distance and $G_l(x, y)$ represents the x and y axis coordinates of the l-th index position in the activation map $\mathbf{A}$.

3.3.3. *Attention-Guided Feature Aggregation*

To balance both semantic similarity and spatial consistency criteria, we define the parameter-free region attention coefficient as:

$$R_{ml} = T_{ml} \times W_{ml}. \tag{8}$$

Intuitively, a larger R_{ml} indicates stronger semantic correlation between the representative point $G_m(x, y)$ and the channel feature at the l-th index in the feature map F, along with tighter spatial aggregation. Finally, under the guidance of attention, we aggregate the local region features corresponding to the representative points:

$$\hat{v}_m = \sum_{l=1}^{H\times W} \frac{\exp(R_{ml})}{\sum_{\hat{l}=1}^{H\times W} \exp(R_{m\hat{l}})} \cdot v_l. \tag{9}$$

Thus, all support and query feature maps are aggregated into $M \times M$ local region features through parameter-free attention.

3.4. Local Description Classifier

Although the PRA module has aggregated discriminative region features through parameter-free attention, residual background interference still exists in certain local descriptors (*e.g.,* vegetation patches near buildings), which could potentially mislead the classification decision. To address this issue, we introduce an entropy-aware similarity voting mechanism within the LDC module, which identifies the top-K most discriminative query regions.

For a given class c and a query sample, we first calculate the similarity between the local region features $\hat{v}_m^q$ and the global class prototype p_c:

$$s_{m,c} = \frac{\hat{v}_m^q p_c}{\parallel \hat{v}_m^q \parallel \parallel p_c \parallel}, \tag{10}$$

where p_c is obtained by averaging the local region features of all supports under class c:

$$p_c = \frac{1}{K} \sum_{(\hat{v}^s_{m|i}, y_i) \in S} \left(\frac{1}{M \times M} \sum_{M=1}^{M\times M} \hat{v}^s_{m|i}\right) \cdot \mathbb{I}(y_i = c). \tag{11}$$

Then, the discriminative power of each local region feature is evaluated by calculating its information entropy based on the inter-class similarity distribution:

$$H_m = -\sum_{c \in C_{task}} \tilde{s}_{m,c} \log \tilde{s}_{m,c}, \tag{12}$$

where C_{task} stands for the class set in an episode. $\tilde{s}_{m,c}$ is the inter-class similarity distribution, which can be defined as:

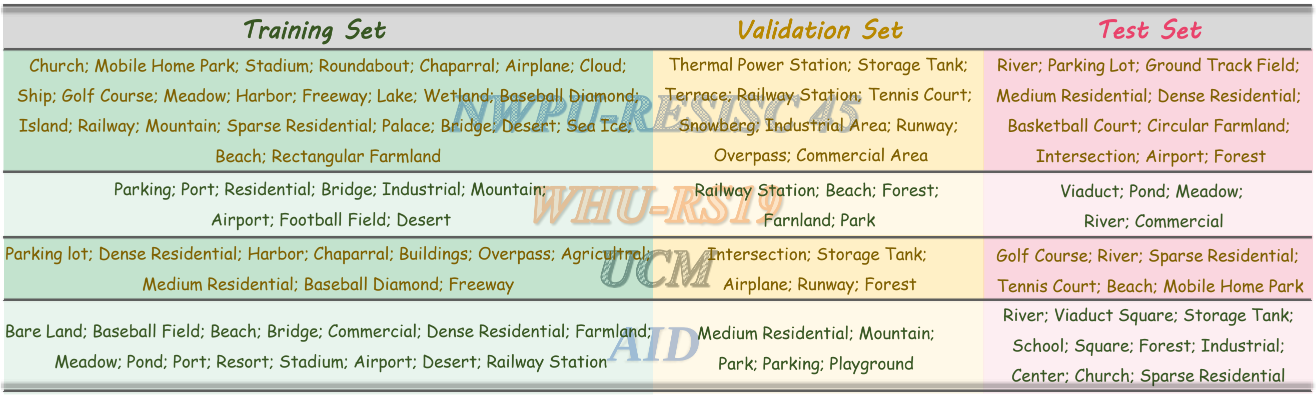
$$\tilde{s}_{m,c} = \frac{\exp(s_{m,c})}{\sum_{\acute{c} \in C_{task}} \exp(s_{m,\acute{c}})}. \tag{13}$$

A low entropy value H_m indicates that the region is highly aligned with a specific class prototype, while a high-entropy region corresponds to semantically ambiguous areas (*e.g.,* land that may appear in multiple categories). Subsequently, we select the top-K regions with the lowest entropy, $\mathcal{R}_K$, to calculate the final classification score:

$$y^{pred|c} = \sum_{\hat{v}_k^q \in \mathcal{R}_K} s_{k,c}. \tag{14}$$

Further, we can obtain the probability distribution for the j-th query sample in an episode as:

$$\mathbf{P}(y_j^{pred|c} | I_j) = \frac{\exp(y_j^{pred|c})}{\sum_{\acute{c} \in C_{task}} \exp(y_j^{pred|\acute{c}})}. \tag{15}$$



	Training Set	Validation Set	Test Set
NWPU-RESISC45	Church; Mobile Home Park; Stadium; Roundabout; Chaparral; Airplane; Cloud; Ship; Golf Course; Meadow; Harbor; Freeway; Lake; Wetland; Baseball Diamond; Island; Railway; Mountain; Sparse Residential; Palace; Bridge; Desert; Sea Ice; Beach; Rectangular Farmland	Thermal Power Station; Storage Tank; Terrace; Railway Station; Tennis Court; Snowberg; Industrial Area; Runway; Overpass; Commercial Area	River; Parking Lot; Ground Track Field; Medium Residential; Dense Residential; Basketball Court; Circular Farmland; Intersection; Airport; Forest
WHU-RS19	Parking; Port; Residential; Bridge; Industrial; Mountain; Airport; Football Field; Desert	Railway Station; Beach; Forest; Farnland; Park	Viaduct; Pond; Meadow; River; Commercial
UCM	Parking lot; Dense Residential; Harbor; Chaparral; Buildings; Overpass; Agricultral; Medium Residential; Baseball Diamond; Freeway	Intersection; Storage Tank; Airplane; Runway; Forest	Golf Course; River; Sparse Residential; Tennis Court; Beach; Mobile Home Park
AID	Bare Land; Baseball Field; Beach; Bridge; Commercial; Dense Residential; Farmland; Meadow; Pond; Port; Resort; Stadium; Airport; Desert; Railway Station	Medium Residential; Mountain; Park; Parking; Playground	River; Viaduct Square; Storage Tank; School; Square; Forest; Industrial; Center; Church; Sparse Residential

Figure 4: Standard split representation of four few-shot remote sensing scene classification datasets.

3.5. Loss Function

Our core PRA and LDC modules are designed parameter-free, making them plug-and-play for any metric learning-based FSL framework. Additionally, they employ the most basic cross-entropy training paradigm, consistent with that of ProtoNet (Snell et al., 2017):

$$\mathcal{L} = -\frac{1}{|Q|} \sum_{(\boldsymbol{I}_j, y_j) \in Q} \log(\mathbf{P}(y_j^{pred|c} | \boldsymbol{I}_j)). \tag{16}$$

4. Experiments

4.1. Datasets

4.1.1. NWPU-RESISC45

The NWPU-RESISC45 dataset (Cheng et al., 2017), developed by Northwestern Polytechnical University, is a widely used benchmark for remote sensing scene classification. It comprises 31,500 RGB images (256×256 pixels) evenly distributed across 45 categories, each containing 700 images. These categories span diverse natural landscapes (e.g., forest, desert), human-made structures (e.g., airport, harbor), and mixed environments (e.g., farmland, urban areas). Its high intra-class variability and inter-class similarity make it a robust benchmark for evaluating scene classification algorithms, particularly regarding generalization and robustness.

4.1.2. WHU-RS19

The WHU-RS19 dataset (Guofeng Sheng & Sun, 2012), created by Wuhan University, is a high-resolution benchmark for remote sensing scene classification. It comprises 1,005 images, each measuring 600×600 pixels, evenly distributed across 19 categories with 50 images per class. The dataset comprises various scenes, including airports, bridges, industrial areas, forests, rivers, and residential areas.

4.1.3. UCM

The UCMerced Land-Use Dataset (UCM) (Yang & Newsam, 2010), developed by the University of California, Merced, comprises 2,100 aerial images with a spatial resolution of approximately 0.3 meters per pixel and a fixed size of 256×256 pixels. Organized into 21 evenly distributed categories, it includes both natural scenes (e.g., forest, river, beach) and human-made structures (e.g., buildings, parking lots, runways), with 100 images per class. The imagery, sourced from the USGS National Map urban areas, offers high-quality and diverse scene representations.

4.1.4. AID

The Aerial Image Dataset (AID) (Xia et al., 2017) is a large-scale benchmark for scene classification in remote sensing, consisting of 10,000 high-resolution aerial images sourced from Google Earth. Each image has a fixed resolution of 600×600 pixels and is categorized into 30 diverse scene types, including both natural (e.g., farmland, park) and human-made environments (e.g., airport, residential areas). Class sizes range from 220 to 420 images, resulting in an imbalanced distribution. The dataset captures varying imaging conditions, such as differences in resolution, lighting, and seasonal changes, making it well-suited for assessing the robustness of classification algorithms.

To ensure a fair comparison, we follow the experimental setup described by (Xu et al., 2024b), which entails partitioning these benchmarks into training, validation, and testing sets. The specific partitioning is detailed in Fig. 4.

4.2. Implementation Details

We implement PA-SRM using the PyTorch framework (Paszke et al., 2019). All experiments are conducted on NVIDIA GTX 1080Ti GPUs. Following the standard FSL task protocol, we adopt the ResNet12 (Oreshkin et al., 2018) architecture as the backbone for PA-SRM without pre-training on ImageNet. Input images are uniformly resized to 84×84 pixels and then passed through the backbone network to generate feature maps with a dimensionality of $640 \times 5 \times 5$. For optimization, we employ stochastic gradient descent (SGD) with a momentum of 0.95 and a weight decay of 1×10^{-4}. The model is trained for 30 epochs, and each episode contains 15 query samples per category.

We conduct FSL experiments under two widely used scenarios: 5-way 1-shot and 5-way 5-shot. During the testing phase, we perform 2000 random samplings on the test set and report the average classification accuracy as the final result.

Table 1
Comparison of different methods on NWPU-RESISC45 and WHU-RS19 datasets

Method	Type	Params	NWPU-RESISC45 5-way		WHU-RS19 5-way	
			1-shot	5-shot	1-shot	5-shot
MAML (Finn et al., 2017)	Optimization-based	9.21M	48.04±0.21	62.98±0.47	50.87±0.23	64.26±0.32
Meta-SGD (Li et al., 2017)	Optimization-based	18.97M	40.96±0.68	47.46±0.37	51.78±1.05	65.47±0.65
LLSR (Zhai et al., 2019)	Optimization-based	-	51.43	72.90	51.10	70.65
MatchingNet (Vinyals et al., 2016)	Metric-based	10.72M	40.31±0.13	47.27±0.38	51.25±0.61	54.36±0.38
ProtoNet (Snell et al., 2017)	Metric-based	11.19M	41.38±0.26	62.77±0.14	58.17±0.56	80.54±0.42
RelationNet (Sung et al., 2018)	Metric-based	27.07M	66.21±0.28	78.37±0.28	61.74±0.51	79.15±0.35
DLA-MatchNet (Li et al., 2021b)	Metric-based	50.91M	68.80±0.70	81.63±0.46	68.27±1.83	79.89±0.33
SPNet (Cheng et al., 2021)	Metric-based	-	67.84±0.87	83.94±0.50	81.06±0.60	88.04±0.28
SCL-MLNet (Li et al., 2021d)	Metric-based	191.59M	62.21±1.12	80.86±0.76	63.36±0.88	77.62±0.81
TAE-Net (Huang et al., 2021)	Metric-based	-	69.13±0.83	82.37±0.52	73.67±0.74	88.95±0.52
GES-Net (Yuan et al., 2022)	Metric-based	-	70.83±0.085	82.27±0.55	75.84±0.78	82.37±0.38
MKN (Cui et al., 2022)	Metric-based	-	65.84±0.89	82.67±0.55	-	-
MPCL-Net (Ma et al., 2023)	Metric-based	45.01M	55.94±0.04	76.24±0.12	61.84±0.12	80.34±0.54
TDNET (Wang et al., 2023)	Metric-based	8.33M	65.85±0.53	82.16±0.32	64.24±0.51	84.15±0.32
HiReNet (Tian et al., 2024)	Metric-based	13.94M	70.43±0.90	81.24±0.58	-	-
PA-SRM(*ours*)	Metric-based	**14.53M**	**72.65±0.43**	**83.64±0.61**	**77.49±0.71**	**88.86±0.58**

Table 2
Comparison of different methods on UCM and AID datasets

Method	Type	Params	UCM 5-way		AID 5-way	
			1-shot	5-shot	1-shot	5-shot
MAML (Finn et al., 2017)	Optimization-based	9.21M	48.94±0.31	60.61±0.29	48.16±0.37	61.13±0.94
Meta-SGD (Li et al., 2017)	Optimization-based	18.97M	51.13±0.95	63.68±0.59	50.01±0.93	65.31±0.38
LLSR (Zhai et al., 2019)	Optimization-based	-	39.47	57.40	-	-
MatchingNet (Vinyals et al., 2016)	Metric-based	10.72M	34.68±0.91	53.34±0.17	42.17±0.78	52.34±0.89
ProtoNet (Snell et al., 2017)	Metric-based	11.19M	52.34±0.19	69.28±0.67	49.91±0.47	70.48±0.21
RelationNet (Sung et al., 2018)	Metric-based	27.07M	48.48±0.75	62.17±0.33	53.51±0.68	68.65±0.95
DLA-MatchNet (Li et al., 2021b)	Metric-based	50.91M	53.76±0.62	63.01±0.51	53.41±0.99	70.61±0.36
SPNet (Cheng et al., 2021)	Metric-based	-	57.64±0.73	73.52±0.51	-	-
SCL-MLNet (Li et al., 2021d)	Metric-based	191.59M	51.37±0.79	68.09±0.92	59.46±0.96	76.31±0.68
TAE-Net (Huang et al., 2021)	Metric-based	-	60.21±0.72	77.44±0.51	-	-
GES-Net (Yuan et al., 2022)	Metric-based	-	-	-	-	-
MKN (Cui et al., 2022)	Metric-based	-	58.45±0.54	77.92±0.23	57.29±0.59	75.42±0.31
MPCL-Net (Ma et al., 2023)	Metric-based	45.01M	56.41±0.21	76.57±0.07	60.61±0.43	76.78±0.08
TDNET (Wang et al., 2023)	Metric-based	8.33M	-	-	67.48±0.51	80.56±0.36
HiReNet (Tian et al., 2024)	Metric-based	13.94M	58.60±0.80	76.84±0.56	59.43±0.66	74.12±0.43
PA-SRM(*ours*)	Metric-based	**14.53M**	**60.79±0.28**	**78.64±0.42**	**68.75±0.36**	**81.47±0.65**

4.3. Experimental Results and Comparisons

To demonstrate the advancement of the proposed PA-SRM, a comparison is made with various state-of-the-art FSL methods on four publicly available datasets, including MatchingNet (Vinyals et al., 2016), ProtoNet (Snell et al., 2017), MAML (Finn et al., 2017), Meta-SGD (Li et al., 2017), LLSR (Zhai et al., 2019), RelationNet (Sung et al., 2018), DLA-MatchNet (Li et al., 2021b), SPNet (Cheng et al., 2021), SCL-MLNet (Li et al., 2021d), TAE-Net (Huang et al., 2021), GES-Net (Yuan et al., 2022), MKN (Cui et al., 2022), MPCL-Net (Ma et al., 2023), TDNET (Wang et al., 2023), and HiReNet (Tian et al., 2024). Tables 1 and 2 clearly illustrate the paradigms of the methods above, their parameter counts, and the classification accuracies under both 1-shot and 5-shot settings.

PA-SRM performs superiorly in few-shot remote sensing classification, as evidenced by three critical aspects: (i) It achieves a significant performance improvement over the second-best method, with an average increase of 1.33 in classification accuracy under the 1-shot setting across four datasets. (ii) In addition to its overall performance advantage, PA-SRM benefits from the parameter-free design of its core module, which results in a lighter parameter footprint. (iii) The most competitive recent methods, such as HiReNet and TDNET, excel on different datasets. However, our method ranks top across all four datasets, demonstrating the model's

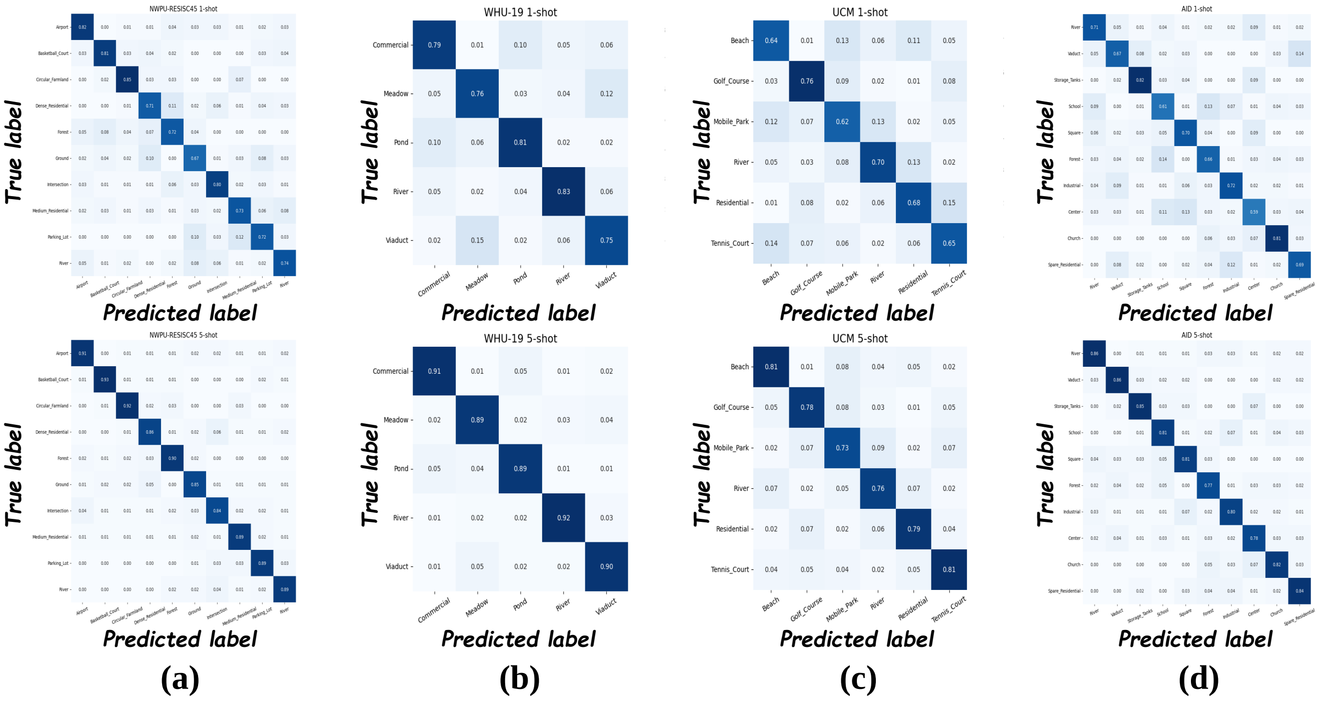


Figure 5: Confusion matrix visualization across four datasets. (a) NWPU-RESISC45; (b) WHU-RS19; (c) UCM; (d) AID.

Table 3
Ablation study of core model components on the NWPU-RESISC45 dataset.

	PRA	LDC	Baseline	1-shot	5-shot	FLOPs
I			✓	67.84±0.36	78.43±0.52	22.25G
II	✓		✓	70.64±0.88	82.61±0.32	22.25G
III		✓	✓	70.32±0.64	80.54±0.13	22.25G
IV	✓	✓	✓	**72.65±0.43**	**84.64±0.61**	22.25G

Table 4
Ablation study of core model components on the WHU-RS19 dataset.

	PRA	LDC	Baseline	1-shot	5-shot	FLOPs
I			✓	65.44±0.62	79.72±0.52	22.25G
II	✓		✓	73.68±0.29	85.11±0.72	22.25G
III		✓	✓	71.47±0.82	82.39±0.26	22.25G
IV	✓	✓	✓	**77.49±0.71**	**88.86±0.58**	22.25G

generalization ability and robustness.

To provide a more intuitive demonstration of the model's discriminative ability for each category, Fig. 5 presents the confusion matrix results for the model across four datasets under two experimental setups. It is evident that the model effectively suppresses inter-class ambiguity for all categories.

4.4. Ablation Study

4.4.1. Module Ablation of PA-SRM

We conduct ablation experiments on four datasets under two experimental settings to validate the effectiveness of our two core designs, PRA 3.3 and LDC 3.4:

Table 5
Ablation study of core model components on the UCM dataset.

	PRA	LDC	Baseline	1-shot	5-shot	FLOPs
I			✓	54.21±0.87	70.85±0.36	22.25G
II	✓		✓	58.28±0.27	76.54±0.80	22.25G
III		✓	✓	57.45±0.78	72.84±0.61	22.25G
IV	✓	✓	✓	**60.79±0.28**	**78.64±0.42**	22.25G

Table 6
Ablation study of core model components on the AID dataset.

	PRA	LDC	Baseline	1-shot	5-shot	FLOPs
I			✓	61.56±0.73	76.54±0.67	22.25G
II	✓		✓	66.69±0.93	80.95±0.28	22.25G
III		✓	✓	63.27±0.46	78.32±0.98	22.25G
IV	✓	✓	✓	**68.75±0.36**	**82.47±0.65**	22.25G

I-*Baseline:* We use the ResNet12-configured ProtoNet as the baseline for PA-SRM.

II-*Baseline + PRA:* Building upon the baseline, we integrate the PRA module to concentrate on the discriminative features within images while employing a parameter-free paradigm to mitigate base-class bias.

III-*Baseline + LDC:* Directly split the feature maps outputted by the baseline and use the LDC module to filter region features that resemble category prototypes for classification decision-making.

IV-*Baseline + PRA + LDC (PA-SRM):* The overall framework of the proposed PA-SRM.

As shown in Tables 3-6, the PRA and LDC modules we

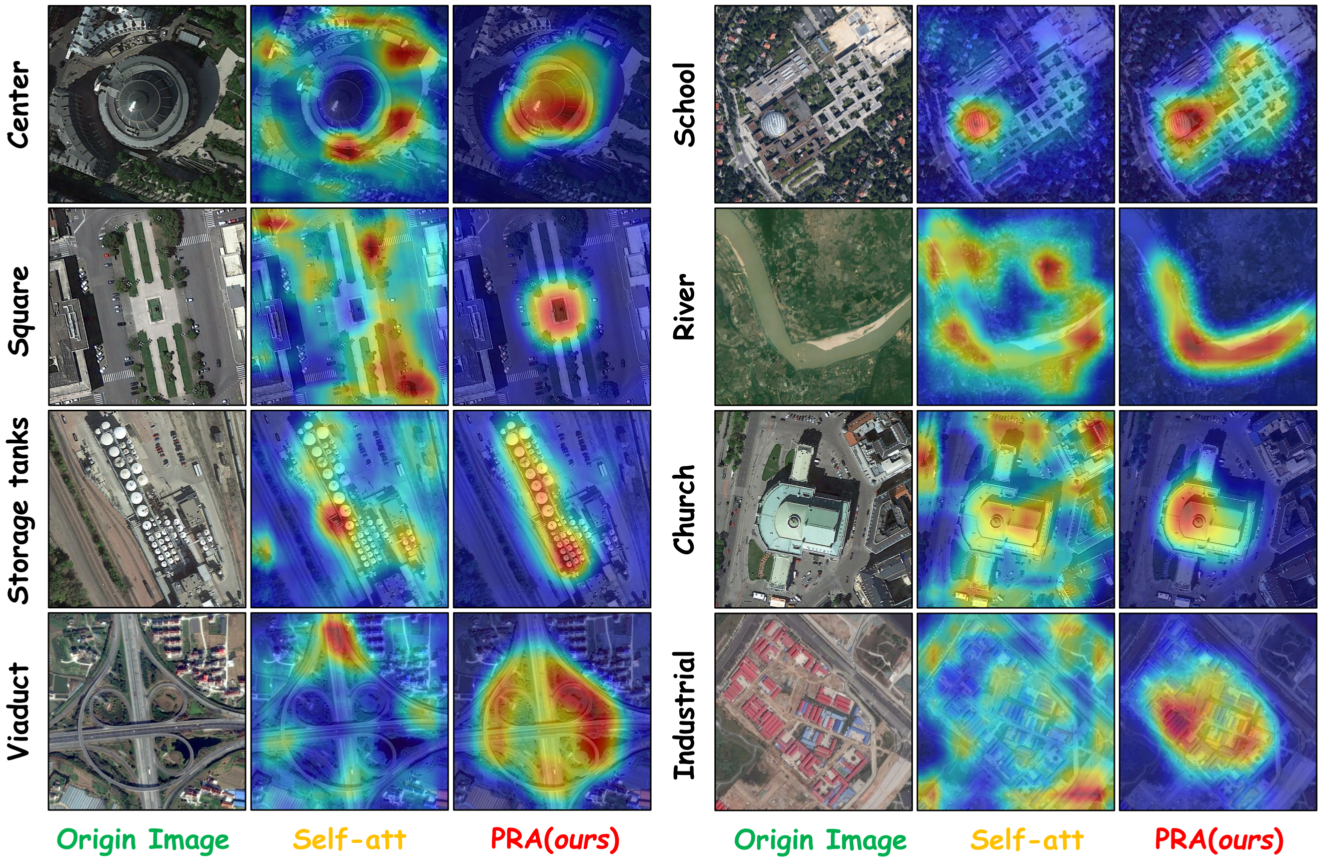


Figure 6: Visual comparison of the ability of self-attention and the PRA module to capture critical regional features.

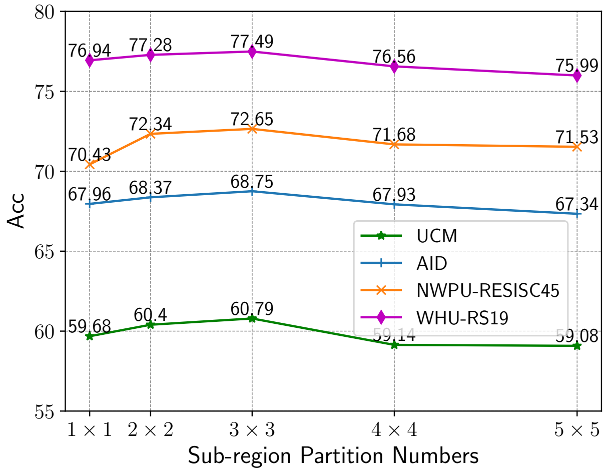


Figure 7: Impact of sub-region partition numbers.

propose contribute to FSL tasks, aligning with their original design intent. FLOPs analysis indicates that these two modules introduce no additional computational overhead compared to the baseline while effectively collaborating at feature extraction and classification stages.

Table 7
Impact of the attention coefficient strategy.

Dataset	Semantic	Spatial	1-shot	5-shot
NWPU-RESISC45		✓	71.05	81.80
	✓		71.77	83.02
	✓	✓	72.65	84.64
WHU-RS19		✓	73.89	84.55
	✓		76.29	86.92
	✓	✓	77.49	88.86
AID		✓	65.23	80.29
	✓		67.10	80.84
	✓	✓	68.75	82.47
UCM		✓	58.64	75.60
	✓		59.58	74.26
	✓	✓	60.79	78.64

4.4.2. Design of the PRA Module

From Section 3.3, the PRA module comprises two pivotal steps: region representative points localization (Section 3.3.1) and attention coefficient computation (Section 3.3.2). We conduct ablation experiments to evaluate its design and utilize gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) to visualize its focus on critical regions.

Firstly, we investigate the impact of varying sub-region partition numbers on 1-shot performance across four datasets,

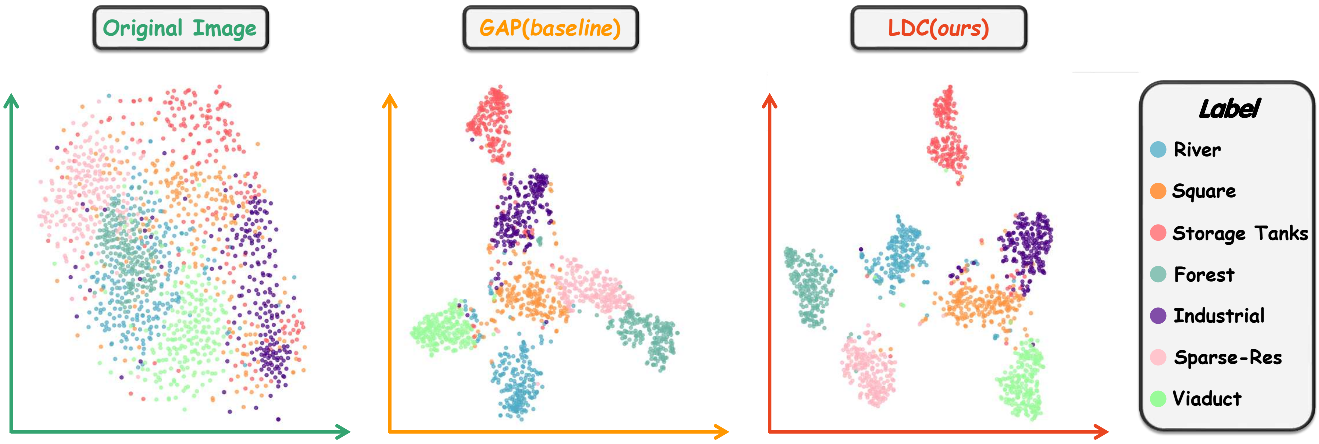


Figure 8: t-SNE visualization of feature distributions processed by GAP operation and the LDC module.

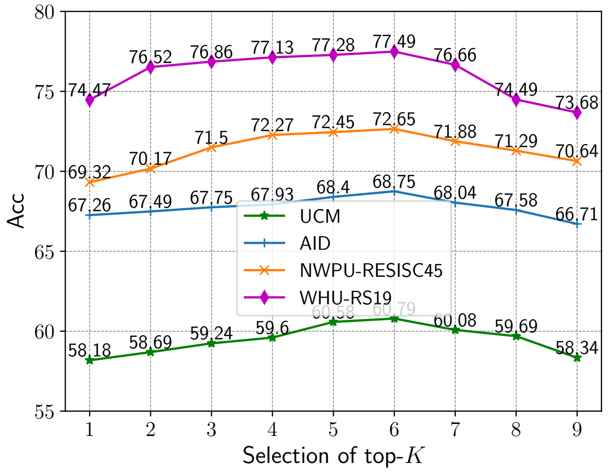


Figure 9: Impact of the number of selected sub-regions.

as illustrated in Fig. 7. When the number of sub-regions is too few (*e.g.,* 2×2), grid-based partitioning becomes overly coarse, failing to capture fine-grained local details and leading to the loss of critical semantic information, such as subtle land cover variations essential for classification. In contrast, excessively fine partitioning (*e.g.,* 4×4 or 5×5) risks over-fragmentation, disrupting the structural coherence of target objects and impairing the model's ability to extract meaningful regional features. A 3×3 partition achieves a balance, offering sufficient granularity to capture local discriminative features while maintaining spatial and semantic consistency across regions.

Next, considering that the parameter-free region attention coefficient R_{ml} (Eq. 8) is derived by multiplying the semantic similarity factor T_{ml} (Eq. 6) and spatial consistency W_{ml} (Eq. 7) factor, we investigate the respective contributions of these two factors to the PRA module. As evidenced by the results in Table 7, both components contribute significantly to the overall performance. The model performs optimally when aggregating semantically similar features while preserving the regions' spatially structured information.

Finally, Fig. 6 provides a comprehensive visual comparison of the Grad-CAM outputs generated by our proposed PRA module and a conventional self-attention mechanism. Note that all the samples displayed are drawn from the test set. The PRA module demonstrates superior capability in capturing critical regional features. In contrast, constrained by many trainable parameters and base-class bias, self-attention performs poorly on unseen classes. This limitation is particularly problematic in remote sensing scene classification, where significant intra-class variation is prevalent.

4.4.3. *Design of the LDC Module*

In the LDC module, we adopt an entropy-aware similarity voting mechanism to select the top-K discriminative regions from the query samples for classification decisions. Fig. 9 illustrates the impact of varying K values on classification performance. The model achieves optimal performance when selecting the top-9 low-entropy regions, which can be attributed to a balance between discriminative coverage and noise suppression:

1) The limitation of a small K: A small number of regions emphasize highly discriminative local features but fail to capture multi-part dependencies, rendering the classification vulnerable to local occlusions or viewpoint changes.

2) The degradation of a large K: Too many regions introduce high-entropy background noise, leading to diluted classification confidence

Moreover, in Fig. 8, we present a t-SNE (Van der Maaten & Hinton, 2008) visualization that contrasts the results of applying traditional Global Average Pooling (GAP) to multiple query features with those processed through the LDC module. It can be observed that, after being filtered through the proposed LDC module, the query feature distribution used for classification decisions becomes more discriminative.

5. Conclusion

This paper identifies the limitations of traditional metric-based approaches in few-shot learning (FSL) for remote sensing images, where global feature aggregation often introduces complex background interference. While self-attention mechanisms have been employed to address this issue, their reliance on extensive training parameters leads to base-class bias, hindering generalization to unseen scenes with significant intra-class variation. To tackle these challenges, we propose a Parameter-Free Region Attention (PRA) module that emphasizes discriminative regions while avoiding base-class bias. Moreover, we introduce an entropy-aware similarity voting mechanism within the LDC module to enhance classification by selecting highly discriminative query regions, further suppressing background interference. Extensive comparative and ablation experiments validate the superior performance of the proposed PA-SRM framework and the efficacy of its individual components.

References

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14* (pp. 850–865). Springer.

Chen, W., Si, C., Zhang, Z., Wang, L., Wang, Z., & Tan, T. (2023). Semantic prompt for few-shot image recognition. *arXiv preprint arXiv:2303.14123*, .

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

Cheng, G., Li, Z., Yao, X., Guo, L., & Wei, Z. (2017). Remote sensing image scene classification using bag of convolutional features. *IEEE Geoscience and Remote Sensing Letters*, *14*, 1735–1739.

Cui, Z., Yang, W., Chen, L., & Li, H. (2022). Mkn: Metakernel networks for few shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11.

Ding, Y., Tian, X., Yin, L., Chen, X., Liu, S., Yang, B., & Zheng, W. (2019). Multi-scale relation network for few-shot learning based on meta-learning. In *International Conference on Computer Vision Systems* (pp. 343–352). Springer.

Dou, P., Huang, C., Han, W., Hou, J., Zhang, Y., & Gu, J. (2024). Remote sensing image classification using an ensemble framework without multiple classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, *208*, 190–209.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1126–1135). volume 70.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3146–3154).

Gong, M., Li, J., Zhang, Y., Wu, Y., & Zhang, M. (2022). Two-path aggregation attention network with quad-patch data augmentation for few-shot scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–16.

Guofeng Sheng, T. X., Wen Yang, & Sun, H. (2012). High-resolution satellite scene classification using a sparse coding based multiple feature combination. *International Journal of Remote Sensing*, *33*, 2395–2412.

Han, W., Feng, R., Wang, L., & Cheng, Y. (2018). A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*, 23–43.

He, J., Hong, R., Liu, X., Xu, M., Zha, Z.-J., & Wang, M. (2020). Memory-augmented relation network for few-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1236–1244).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3* (pp. 84–92). Springer.

Huang, W., Shi, Y., Xiong, Z., Wang, Q., & Zhu, X. X. (2023). Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *195*, 192–203.

Huang, W., Yuan, Z., Yang, A., Tang, C., & Luo, X. (2021). Tae-net: Task-adaptive embedding network for few-shot remote sensing scene classification. *Remote Sensing*, *14*, 111.

Jia, Y., Gao, J., Huang, W., Yuan, Y., & Wang, Q. (2023a). Exploring hard samples in multi-view for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, .

Jia, Y., Gao, J., Huang, W., Yuan, Y., & Wang, Q. (2023b). Exploring hard samples in multiview for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–14.

Jia, Y., Gao, J., Huang, W., Yuan, Y., & Wang, Q. (2023c). Holistic mutual representation enhancement for few-shot remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, .

Jia, Y., Li, J., & Wang, Q. (2025). Generalized few-shot semantic segmentation for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, (pp. 1–1). doi:10.1109/TGRS.2025.3531874.

Jia, Y., Zhou, Q., Huang, W., Gao, J., & Wang, Q. (2024). Like humans to few-shot learning through knowledge permeation of vision and text. *arXiv preprint arXiv:2405.12543*, .

Kim, W., Goyal, B., Chawla, K., Lee, J., & Kwon, K. (2018). Attention-based ensemble for deep metric learning. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 736–751).

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

Li, L., Han, J., Yao, X., Cheng, G., & Guo, L. (2021b). Dla-matchnet for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *59*, 7844–7853.

Li, M., Lei, L., Li, X., & Sun, Y. (2021c). A multi-scale feature aggregation network based on channel-spatial attention for remote sensing scene classification. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* (pp. 4916–4919).

Li, X., Shi, D., Diao, X., & Xu, H. (2021d). Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–12.

Li, Z., Zhou, F., Chen, F., & Li, H. (2017). Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, .

Liu, Y., Li, H., Hu, C., Luo, S., Luo, Y., & Chen, C. W. (2025). Learning to aggregate multi-scale context for instance segmentation in remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, *36*, 595–609.

Liu, Y., Zhong, Y., Shi, S., & Zhang, L. (2024). Scale-aware deep reinforcement learning for high resolution remote sensing imagery classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *209*, 296–311.

Ma, J., Lin, W., Tang, X., Zhang, X., Liu, F., & Jiao, L. (2023). Multi-pretext-task prototypes guided dynamic contrastive learning network for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, .

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*.

Miao, W., Geng, J., & Jiang, W. (2022). Semi-supervised remote-sensing image scene classification using representation consistency siamese network. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS journal of photogrammetry and remote sensing*, *66*, 247–259.

Oreshkin, B., Rodríguez López, P., & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, *31*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. volume 32.

Penatti, O. A., Nogueira, K., & Dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44–51).

Qin, A., Chen, F., Li, Q., Tang, L., Yang, F., Zhao, Y., & Gao, C. (2024). Deep updated subspace networks for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–14.

Qiu, C., Zhang, X., Tong, X., Guan, N., Yi, X., Yang, K., Zhu, J., & Yu, A. (2024). Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, *209*, 368–382.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Shen, J., Cao, B., Zhang, C., Wang, R., & Wang, Q. (2022). Remote sensing scene classification based on attention-enabled progressively searching. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13.

Shen, W., Zhong, Y., & Ma, A. (2024). Self-supporting adaptive prototype learning for remote sensing few-shot semantic segmentation. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 9701–9706).

Simonyan, K. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, .

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, *30*.

Stivaktakis, R., Tsagkatakis, G., & Tsakalides, P. (2019). Deep learning for multilabel land cover scene categorization using data augmentation. *IEEE Geoscience and Remote Sensing Letters*, *16*, 1031–1035.

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

Tian, F., Lei, S., Zhou, Y., Cheng, J., Liang, G., Zou, Z., Li, H.-C., & Shi, Z. (2024). Hirenet: Hierarchical-relation network for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–10.

Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, *4*, 41–57.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, *29*.

Wang, B., Wang, Z., Sun, X., He, Q., Wang, H., & Fu, K. (2023). Tdnet: A novel transductive learning framework with conditional metric embedding for few-shot remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 4591–4606.

Wang, J., Wang, X., Xing, L., Liu, B.-D., & Li, Z. (2022a). Class-shared sparsepca for few-shot remote sensing scene classification. *Remote Sensing*, *14*, 2304.

Wang, Q., Jia, Y., Gao, J., & Li, Q. (2024). Embedding generalized semantic knowledge into few-shot remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, .

Wang, W., Liu, X., & Mou, X. (2021). Data augmentation and spectral structure features for limited samples hyperspectral classification. *Remote Sensing*, *13*, 547.

Wang, X., Han, X., Huang, W., Dong, D., & Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5022–5030).

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022b). Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*, 213–247.

Wu, Z., Li, Y., Guo, L., & Jia, K. (2019). Parn: Position-aware relation networks for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6659–6667).

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (2017). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*, 3965–3981.

Xiong, Y., Xu, K., Dou, Y., Zhao, Y., & Gao, Z. (2021). Wrmatch: Improving fixmatch with weighted nuclear-norm regularization for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

Xu, W., Wang, J., Wei, Z., Peng, M., & Wu, Y. (2023). Deep semantic-visual alignment for zero-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *198*, 140–152.

Xu, Y., Bi, H., Yu, H., Lu, W., Li, P., Li, X., & Sun, X. (2024a). Attention-based contrastive learning for few-shot remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, .

Xu, Y., Bi, H., Yu, H., Lu, W., Li, P., Li, X., & Sun, X. (2024b). Attention-based contrastive learning for few-shot remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–17.

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 270–279).

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

Yu, X., Wu, X., Luo, C., & Ren, P. (2017). Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, *54*, 741–758.

Yuan, Z., Huang, W., Tang, C., Yang, A., & Luo, X. (2022). Graph-based embedding smoothing network for few-shot scene classification of remote sensing images. *Remote Sensing*, *14*, 1161.

Zeng, Q., & Geng, J. (2022). Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *191*, 143–154.

Zeng, Q., Geng, J., Huang, K., Jiang, W., & Guo, J. (2021). Prototype calibration with feature generation for few-shot remote sensing image scene classification. *Remote Sensing*, *13*, 2728.

Zhai, M., Liu, H., & Sun, F. (2019). Lifelong learning for scene recognition in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *16*, 1472–1476.

Zhang, C., Cai, Y., Lin, G., & Shen, C. (2020). Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12203–12213).

Zhang, D., Yang, Y., Liu, X., Ma, W., & Jiao, L. (2023a). Dense cross-scale transformer with channel learning for remote sensing scene classification. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6101–6104).

Zhang, H., Xu, J., Jiang, S., & He, Z. (2024). Simple semantic-aided few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 28588–28597).

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and remote sensing magazine*, *4*, 22–40.

Zhang, T., Zhang, X., Zhu, P., Jia, X., Tang, X., & Jiao, L. (2023b). Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *195*, 353–364.

Zhao, Y., Chen, Y., Rong, Y., Xiong, S., & Lu, X. (2024a). Global-group attention network with focal attention loss for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–14.

Zhao, Y., Gong, M., Qin, A. K., Zhang, M., Hu, Z., Gao, T., & Pu, Y.

(2024b). Gradient-guided multiscale focal attention network for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–18.

Zheng, J., Li, Z., Duan, P., Kang, X., & Fu, W. (2024). Hyperspectral remote sensing scene classification with spectral-spatial convolutional network. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 9464–9467).