

Article

Few-Shot Remote Sensing Scene Classification via Fusion of Zigzag Scanning Feature Sequence and Riemannian Geometric Barycenter Network

Xiliang Chen ^{1,2,3,4,5,*} , Longwei Li ^{2,6,7} , Yufeng Chen ^{1,4}, Lei Liu ^{1,4}, Zhenyu Wang ^{1,4}, Mingqing Liu ⁵ , Xiaojie Liu ⁵ and Guobin Zhu ⁵ 

¹ School of Intelligent Connected Vehicle, Hubei University of Automotive Technology, Shiyan 442000, China

² Anhui Province Key Laboratory of Realistic Geographic Environment, Chuzhou University, Chuzhou 239000, China; lilw@chzu.edu.cn

³ Key Laboratory of Automotive Power Train and Electronic Control, Hubei University of Automotive Technology, Shiyan 442000, China

⁴ Shiyan Key Laboratory of Air-Ground Crowd Cooperation Technology and Application, Hubei University of Automotive Technology, Shiyan 442000, China

⁵ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

⁶ Anhui Engineering Research Center of Remote Sensing and Geo-informatics, Chuzhou 239000, China

⁷ Anhui Center for Collaborative Innovation in Geographical Information Integration and Application, Chuzhou 239000, China

* Correspondence: 20240087@huat.edu.cn

Highlights

What are the main findings?

- A Riemannian manifold feature embedding layer is proposed by combining zigzag scanning, ARMA sequence modeling, and SPD manifold embedding to encode higher-order statistical dependencies in remote sensing features.
- A Riemannian feature extraction and barycenter learning framework is developed to capture non-Euclidean geometric properties and estimate more stable class prototypes under few-shot conditions.

What is the implication of the main finding?

- The proposed method addresses the limitations of conventional Euclidean few-shot learning by exploiting the intrinsic manifold structure of remote sensing image features and preserving spatial-structural information that is often lost during pooling or flattening.
- The proposed method improves classification robustness in label-scarce scenarios by reducing prototype instability caused by large intra-class variations and high inter-class similarities.



Academic Editor: Salah Bourennane

Received: 23 April 2026

Revised: 24 June 2026

Accepted: 2 July 2026

Published: 7 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

Few-shot remote sensing scene classification aims to accurately recognize unseen scene categories using only a scarce number of labeled samples, which has emerged as a research hotspot in the field of remote sensing image interpretation. However, remote sensing images intrinsically suffer from large intra-class variations, high inter-class similarities, and complex background interferences. Traditional few-shot learning methods typically perform feature metric learning in Euclidean space, making it difficult to capture the non-Euclidean geometric distribution characteristics of remote sensing features, and they often neglect the spatial structural information embedded in feature maps. To address these issues, this paper proposes a novel few-shot remote sensing scene classification method, termed ZSFS-RGBN, which integrates a Zigzag Scanning Feature Sequence with a

Riemannian Geometric Barycenter Network. Specifically, ResNet12 is first employed as the backbone to extract deep convolutional feature maps from both the support and query sets. Second, a Zigzag scanning strategy is introduced to reorganize the two-dimensional feature maps into one-dimensional feature sequences, thereby effectively preserving the spatial locality and structural continuity of the features. Third, an autoregressive moving average (ARMA) model is constructed to characterize the spatial dependencies of the feature sequences, and its state parameters are mapped onto a symmetric positive definite (SPD) matrix manifold, enabling the deep semantic representations of remote sensing scenes in a non-Euclidean geometric space. Finally, a Riemannian geometric barycenter network is designed to learn the Riemannian barycenter of each category on the SPD manifold, where a joint loss function is introduced to simultaneously optimize intra-class compactness and inter-class separability. Comprehensive experiments are conducted on three public remote sensing scene datasets: NWPU-RESISC45, UC Merced Land-Use, and WHU-RS19. Experimental results demonstrate that the proposed method consistently outperforms several representative state-of-the-art approaches under both 5-way 1-shot and 5-way 5-shot settings. Furthermore, ablation studies verify the effectiveness of each component within the proposed framework.

Keywords: few-shot remote sensing scene classification; zigzag scanning feature sequence; symmetric positive definite (SPD) matrix; Riemannian geometric barycenter network.

1. Introduction

As a fundamental research task in the field of remote sensing (RS) image understanding and intelligent interpretation, RS image scene classification aims to automatically categorize input images into corresponding scene classes based on the spatial configurations of ground objects, texture distribution patterns, semantic contextual relations, and overall regional structures. Driven by the rapid advancements in high-resolution satellites, aerial photography platforms, and unmanned aerial vehicle (UAV) systems, modern RS data exhibit a clear trend toward higher spatial resolution, wider coverage, denser observation frequencies, and increasingly complex scene content. Consequently, the automatic, accurate, and efficient extraction of scene semantic information from massive RS imagery has become an indispensable technical foundation for numerous applications, such as land-use and land-cover (LULC) mapping, urban functional zone identification, ecological environmental assessment, disaster risk analysis, and military situational awareness [1–3].

Early remote sensing scene classification methods primarily relied on hand-crafted visual features, such as color histograms, texture descriptors, Scale-Invariant Feature Transform (SIFT) [4], Local Binary Patterns (LBP) [5], histogram of oriented gradients (HOG) [6], and bag-of-visual-words (BoVW) models [7]. While these approaches demonstrate certain effectiveness on specific datasets and simple scenes, their feature representations are typically limited to low-level visual attributes. Consequently, they struggle to adequately characterize the complex multi-object compositional relationships, scale variations, and high-level semantic structures inherent in high-resolution remote sensing scenes. In recent years, the rapid development of deep convolutional neural networks (CNNs) has significantly propelled research in remote sensing image scene classification. Deep networks represented by AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet can automatically learn hierarchical representations—ranging from edges and textures to local objects and global scene semantics—through multilayer nonlinear mappings [8–11]. These models have achieved performance significantly superior to traditional methods across multiple

benchmark datasets. Building upon this foundation, researchers have further incorporated techniques such as multi-scale feature fusion, attention mechanisms, context modeling, local region matching, graph structure reasoning, knowledge distillation strategy and Vision Transformers [12–16]. These advancements aim to enhance the models' representational capability for the complex spatial structures and long-range dependencies within remote sensing images. However, deep learning models typically rely heavily on large-scale annotated samples for training. Annotating remote sensing images is not only highly costly but also strictly requires specialized geospatial knowledge and expert interpretation experience. For new regions, novel categories, or time-sensitive emergency tasks, obtaining sufficient annotated samples is often impractical. Therefore, achieving reliable recognition of remote sensing scene categories under the condition of limited annotated samples has emerged as one of the most urgent core challenges in the intelligent interpretation of remote sensing data.

Few-shot learning aims to achieve rapid generalization to novel categories using only a few labeled samples. Its core idea is to construct a task distribution from a large set of base categories, enabling the model to learn transferable discriminative knowledge across tasks during training, and then perform classification on new tasks with only a few support samples during testing [17–20]. In recent years, few-shot learning has been widely introduced into the field of remote sensing scene classification, giving rise to the important research direction of few-shot remote sensing scene classification. Compared with natural images, remote sensing scenes exhibit more prominent issues of intra-class variation and inter-class similarity. On one hand, the same category can present significant differences under different geographic regions, seasons, imaging angles, spatial resolutions, and sensor conditions. For example, farmland shows markedly different texture patterns depending on the season and crop type, while man-made scenes such as airports, harbors, and industrial areas also exhibit large visual variations due to differences in spatial scale and functional layout. On the other hand, different categories often share similar land cover components and texture patterns, for instance, highways versus runways, commercial areas versus dense residential areas, rivers versus coastlines, and forests versus wetlands—all exhibit strong visual similarities. Under few-shot conditions, the limited support samples cannot adequately cover the intra-class distribution, leading to problems such as biased class prototype estimation, unstable embedding space distribution, and blurred classification boundaries for query samples.

Existing few-shot remote sensing scene classification methods can generally be categorized into metric learning-based, meta-learning-based, feature enhancement-based, and graph relational modeling-based approaches. Metric learning methods typically learn an embedding space and perform classification based on the distance or similarity between support and query samples, with representative examples including Siamese Networks, Matching Networks, Prototypical Networks, and Relation Networks [17–20]. Meta-learning methods, by contrast, aim to optimize model initialization or learn task adaptation strategies so that the model can adapt to novel categories after only a few gradient updates [21]. Methods tailored for remote sensing scenarios further incorporate local region descriptors, attention mechanisms, class prototype rectification, graph relational inference, and hierarchical semantic modeling to alleviate the classification difficulties caused by the complexity of remote sensing imagery [22–27]. Although these methods have substantially advanced the performance of few-shot remote sensing scene classification, most of them still conduct embedding learning and similarity measurement primarily in a Euclidean feature space. In general, sample representations are obtained through global average pooling, vector concatenation, or linear projection, and classification is then performed using Euclidean distance, cosine distance, or learnable similarity functions. To some extent, this modeling

paradigm overlooks the underlying higher-order statistical structures and non-Euclidean geometric properties inherent in remote sensing scene features.

In fact, remote sensing scenes are typically composed of multiple land-cover objects, texture patterns, and spatial structures [28]. Their deep convolutional features not only encode channel-wise response intensities but also contain higher-order correlations across different spatial locations, feature channels, and local regions. If convolutional feature maps are simply compressed into Euclidean vectors, critical spatial arrangement information and second-order statistical dependencies may be lost. In recent years, symmetric positive definite (SPD) matrices and their associated Riemannian manifolds have attracted extensive attention in fields such as pattern recognition, computer vision, electroencephalogram signal analysis, and image classification [29,30]. SPD matrices can naturally characterize the covariance structure, second-order statistical information, and multivariate correlations of sample features. However, the space of SPD matrices is not a linear Euclidean space but a Riemannian manifold with intrinsic curvature. Ignoring its geometric structure and directly applying Euclidean operations may destroy the positive definiteness of the matrices as well as their intrinsic distance relationships. In contrast, performing metric learning, feature transformation, and mean estimation on the Riemannian manifold enables a more faithful characterization of the geometric relationships among SPD samples [31,32]. Therefore, integrating deep convolutional features with Riemannian manifold learning provides a new research avenue for addressing higher-order correlation modeling and stable class prototype estimation in few-shot remote sensing scene classification.

Based on the above analysis, this paper proposes a few-shot remote sensing scene classification method that integrates a zigzag scanning feature sequence with a Riemannian geometric barycenter network. The core idea of this method is as follows: First, the deep convolutional feature maps of remote sensing images are extracted using ResNet12, retaining Euclidean feature representations with local spatial semantics. Second, the two-dimensional feature maps are converted into one-dimensional feature sequences with spatial order constraints via zigzag scanning so that the feature sequences preserve as much two-dimensional local adjacency and structural continuity as possible during the flattening process. Third, an autoregressive moving average (ARMA) model is employed to model the dynamic dependencies of the serialized features, and an SPD matrix is constructed based on its state-dependent structure, thereby embedding the sample representation from Euclidean space into a Riemannian manifold space. Fourth, manifold feature extraction is performed through SPD network layers such as bilinear mapping and eigenvalue rectification. Finally, a Riemannian geometric barycenter network is designed to estimate the class geometric barycenter in the SPD manifold, and the class prototypes are optimized through intra-class compactness and inter-class separability constraints, thereby enhancing discriminability and generalization capability under few-shot conditions. The main contributions of this paper are summarized as follows.

(1) A Riemannian manifold feature embedding (RMFE_{mb}) layer is proposed by combining zigzag scanning, ARMA sequence modeling, and SPD manifold embedding. It encodes higher-order statistical dependencies into symmetric positive definite matrices and embeds them onto the Riemannian manifold while preserving geometric structure.

(2) A Riemannian manifold feature extraction (RMFE_{xt}) layer is developed, which performs a Bilinear Mapping (BiMap) to preserve geodesic distances, followed by a Rectified Eigenvalues (ReEig) operation to extract principal components directly on the Riemannian manifold. This layer captures non-Euclidean geometric properties beyond conventional Euclidean methods.

(3) A Riemannian geometric barycenter (RieGB) layer is designed to compute task-level and class-level barycenters on the SPD manifold. It jointly optimizes intra-class compact-

ness and inter-class margins, alleviating unstable prototypes and ambiguous decision boundaries under few-shot conditions.

The rest of the paper is organized as follows: the related work is briefly introduced in Section 2, which includes remote sensing image scene classification, image classification based on Riemannian manifold space, and few-Shot remote sensing image scene classification. In Section 3, we introduce the proposed method. Section 4 presents the experiment results. We present a discussion in Section 5, and we draw a conclusion in Section 6.

2. Related Work

2.1. Remote Sensing Image Scene Classification

Remote sensing image scene classification is one of the core tasks in remote sensing image interpretation and aims to automatically recognize and annotate the ground scene categories presented in remote sensing images. With the widespread application of high-resolution remote sensing satellites, the massive amount of remote sensing data acquired provides abundant data sources for scene classification, while also imposing higher demands on the accuracy and efficiency of classification algorithms. Early remote sensing scene classification methods primarily relied on handcrafted low-level features, such as color histograms, gray-level co-occurrence matrix (GLCM), Gabor, histogram of oriented gradients (HOG), and scale-invariant feature transform (SIFT) [4]. Researchers typically combined these features with traditional machine learning classifiers for scene recognition. For example, Yang and Newsam proposed a land use classification method based on the bag-of-visual-words (BoVW) model [7], which clustered local features to construct a visual vocabulary and represented images as word frequency vectors. Cheng et al. extracted mid-level visual elements from remote sensing images using the probabilistic latent semantic analysis (pLSA) model, achieving effective scene classification [33]. However, these handcrafted feature-based methods have obvious limitations: on the one hand, feature design relies heavily on domain knowledge and engineering experience, resulting in limited generalization ability; on the other hand, low-level features are insufficient for representing the complex high-level semantic information in remote sensing images, leading to a sharp decline in classification performance in complex scenes.

Benefiting from the development of deep learning, convolutional neural networks (CNNs) have become the mainstream paradigm for remote sensing scene classification due to their powerful feature learning capability [34]. Penatti et al. first verified that pre-trained CNN features can generalize well from natural images to remote sensing scenes [35]. Subsequently, numerous CNN-based methods have been developed. Chaib et al. fused multi-layer CNN features using discriminant correlation analysis to improve accuracy [36]. Liu et al. designed a multi-scale CNN to handle scale variations in remote sensing objects [37]. Lu et al. proposed a feature aggregation network that enhances representation by aggregating middle-layer features [38]. Wei et al. introduced a memory-efficient feature-replay class-incremental learning framework for dynamic real-world scene classification with streaming data and emerging classes [39]. He et al. proposed a covariance pooling network to capture second-order statistical information with fewer parameters [40]. Furthermore, the integration of attention mechanisms has significantly strengthened feature discrimination [13,41].

In recent years, transformer-based models have been introduced to capture long-range dependencies. Wu et al. proposed the spectral-spatial sparse selection transformer (S4Former), a Vision Transformer-based network integrating sparse representation and iterative optimization, which achieves high accuracy on multispectral remote sensing scene classification with limited training samples. [42]. Xu et al. proposed Hy-MSDA, a hybrid CNN-Transformer framework with consistency learning and dynamic weighting,

which effectively exploits multi-source information for unsupervised domain adaptation in remote sensing scene classification and outperforms state-of-the-art methods [43]. Transformer-based and graph-based models exhibit strong capability in relational modeling; such methods generally involve larger parameter scales and show a stronger dependence on substantial training data. In contrast, zigzag scanning provides a deterministic and parameter-free feature organization strategy that effectively preserves the local-to-global spatial continuity within feature maps, while ARMA modeling is able to capture sequential dependencies with relatively low parameter complexity. Therefore, this design is better suited to few-shot remote sensing scenarios under limited-sample conditions.

Despite great progress made, deep learning-based methods generally require massive amounts of annotated data. In practical applications, obtaining sufficient annotations is expensive and time-consuming, especially for new regions and rare classes. Moreover, most methods can only recognize seen classes and lack the ability to identify novel categories [44]. Therefore, few-shot remote sensing scene classification under limited labeled samples has become an urgent and critical research direction.

2.2. Image Classification Based on Riemannian Manifold Space

Riemannian geometry provides a powerful mathematical tool for handling nonlinear data distributions. In remote sensing image processing, feature covariance matrices naturally lie on symmetric positive definite (SPD) manifolds, and the geometric metrics on Riemannian manifolds can more accurately characterize similarity relationships among samples [45]. The core idea of such methods is to map image features into non-Euclidean geometric spaces and perform similarity measurement using geodesic distances or kernel functions on the manifold, thereby overcoming the limitations of the linearity assumption in Euclidean space. Harandi et al. proposed a Riemannian sparse representation method based on covariance matrices, achieving superior performance over traditional methods on multiple datasets [46]. Huang and Van Gool pioneered the SPD neural network (SPDNet), which takes SPD matrices as input and preserves the manifold structure through bilinear mapping layers (BiMap), eigenvalue rectification layers (ReEig), and logarithmic mapping layers (LogEig), enabling end-to-end Riemannian deep learning [31]. SPDNet and its variants have demonstrated excellent performance in tasks such as image classification and action recognition. Liu et al. proposed a bag-of-matrix-summarization method based on SPDNet, which obtains fixed-length feature representations by vectorized summarization of SPD features, and they designed a margin-based loss function [47]. Brooks et al. proposed a batch-normalized version of SPDNet, which normalizes SPD features using Riemannian barycenters and parallel transport, significantly improving training stability and classification performance [48]. Chakraborty et al. designed ManifoldNet, which performs feature fusion on the SPD manifold via a weighted Fréchet mean layer, enhancing the model's representational capacity [49]. In specific applications of remote sensing scene classification, Zhang et al. proposed a Lie group model integrating multi-scale features and hybrid attention, balancing classification accuracy and model lightness [45]. Xu et al. designed intrinsic means and kernel functions based on Lie group theory, enhancing model robustness [50,51]. Cheng et al. reduced the interference of marginal samples through a marginal sample constraint strategy and enhanced feature discriminability by combining dense connections with global covariance pooling [52]. Li et al. proposed a method for jointly learning first-order and second-order statistical information, constructing a comprehensive loss function to enhance feature discriminability [53]. Wu et al. proposed a deep covariance alignment method to achieve domain-invariant representations of category features, enhancing cross-region adaptability [54].

Although Riemannian manifold-based methods have achieved substantial progress in remote sensing scene classification, SPDNet and ManifoldNet primarily focus on learning discriminative representations on the SPD manifold, typically based on covariance-form inputs. Traditional covariance pooling and global covariance pooling methods generally construct second-order statistics directly from feature maps without explicitly accounting for the organizational patterns of spatial structures and the sequential dependencies inherent in remote sensing scenes. In contrast, the proposed ZSFS-RGBN method does not merely apply SPD learning to Riemannian manifold-based feature extraction; rather, it establishes an integrated framework that spans from spatial feature organization to the estimation of geometric prototypes on the Riemannian manifold.

2.3. Few-Shot Remote Sensing Image Scene Classification

Few-shot remote sensing image scene classification aims to recognize novel scene categories using extremely few labeled samples and has become a research hotspot in the field of remote sensing image processing. Existing methods can be mainly divided into three categories: transfer learning-based methods, meta-learning-based methods, and metric learning-based methods [55]. Transfer learning-based methods typically pre-train CNN models on large-scale natural image datasets and then fine-tune them on target remote sensing datasets. Although such methods can leverage the generalization ability of pre-trained models, their performance degrades significantly when there is a large distribution shift between the source and target domains. Meta-learning-based methods train models by simulating few-shot learning scenarios (episode training) to enable the model to "learn how to learn". Model-agnostic meta-learning (MAML) is one of the most representative meta-learning approaches, which learns a set of good initial parameters so that the model can quickly adapt to new tasks with only a few gradient updates [21]. Zhai et al. first introduced MAML into remote sensing scene recognition and proposed a lifelong few-shot learning model (LLSR) that learns classifiers from limited samples and transfers knowledge from one dataset to another [44]. Li et al. proposed RS-MetaNet, which enhances model generalization by learning a distance metric and a balancing loss to maximize distances between different categories [56]. Metric learning-based methods learn an embedding space in which samples of the same class are close to each other while those of different classes are far apart. Prototypical Networks (ProtoNet) are a classic metric learning approach that computes class prototypes as the mean of support set samples and then classifies query samples based on their distances to the prototypes [19]. Li et al. proposed DLA-MatchNet, which introduces channel and spatial attention to automatically discover discriminative regions and employs a nonlinear network as a similarity measurer [57]. Cheng et al. designed SPNet, which obtains more accurate prototype representations through prototype self-calibration and mutual calibration [22]. Xu et al. proposed an attention-based contrastive learning method (ACL-Net), which highlights representative features through an attention optimization module and designs a dictionary-based contrastive loss to enhance inter-class separability [27].

Few-shot learning methods have achieved good performance in remote sensing scene classification. However, these methods still face several key issues. First, most of them perform feature metric learning and prototype learning in Euclidean space, ignoring the non-Euclidean geometric structure of remote sensing image features. Second, the spatial structure information of feature maps is lost during global pooling or flattening, leading to insufficient modeling of local details and spatial relationships. Third, with extremely few labeled samples, the problems of large intra-class variance and high inter-class similarity are further amplified. Therefore, how to integrate Euclidean-space deep feature extraction with Riemannian manifold geometric modeling while effectively exploiting the spatial

structural information in feature maps represents an important direction for breaking through the performance bottleneck of few-shot remote sensing scene classification.

3. Methodology

3.1. Overall Architecture of the Proposed Method

The overall architecture of the ZSFS-RGBN method proposed in this paper is illustrated in Figure 1. The method mainly consists of four core components: a Euclidean space feature extraction layer, a Riemannian manifold feature embedding layer, a Riemannian manifold feature extraction layer, and a Riemannian geometric barycenter network layer. Specifically, remote sensing images are first fed into a pre-trained ResNet12 backbone network for deep feature extraction, outputting a feature tensor of size $H \times W \times C$, where H and W denote the height and width of the feature map, respectively, and C denotes the number of channels. Subsequently, the feature map of each channel is reorganized into a one-dimensional sequence via zigzag scanning, effectively preserving the spatial correlation and energy concentration characteristics among pixels. On this basis, an autoregressive moving average (ARMA) model is constructed to capture the spatiotemporal dependencies in the sequence. After obtaining the model parameters, they are mapped onto a Grassmann manifold to achieve feature embedding in a non-Euclidean geometric space. Furthermore, the points on the Grassmann manifold are mapped onto a symmetric positive definite (SPD) manifold, and the SPDNet network is employed for deep Riemannian manifold feature extraction. Finally, a Riemannian geometric barycenter network is designed to compute the Riemannian barycenter of samples from each class, and a joint supervised loss function is constructed to achieve joint optimization of intra-class compactness and inter-class separability. The entire model is trained in an end-to-end manner, enabling the collaborative optimization of feature learning and metric learning through the joint loss function.

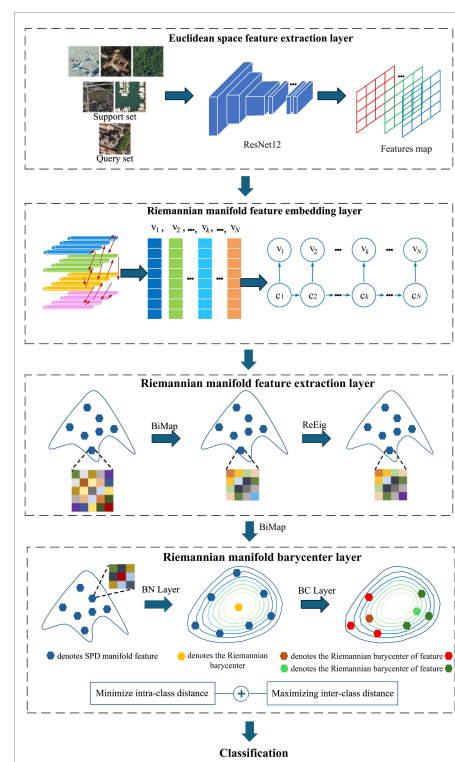


Figure 1. The overall architecture of ZSFS-RGBN.

3.2. Euclidean Space Feature Extraction

In the Euclidean space feature extraction stage, this paper adopts ResNet12 as the backbone network for deep feature extraction from input remote sensing images. ResNet12 is a lightweight variant of the ResNet family, specifically designed for few-shot learning tasks, maintaining strong feature representation capability while controlling network parameters and computational complexity [10]. The core design concepts of ResNet12 include a residual learning mechanism, batch normalization technology, and a global average pooling strategy. The residual learning mechanism solves the problems of gradient vanishing and network degradation in deep network training by introducing skip connections. In traditional CNNs, as the number of layers increases, theoretically stronger feature representation capability should be achieved, but due to gradient attenuation during backpropagation, the training error of deep networks may actually be higher than that of shallow networks. Residual networks form an identity mapping path by directly bypassing one or more convolutional layers and adding the input to the output, transforming the learning objective from directly learning the desired mapping $H(x)$ to learning the residual mapping $F(x) = H(x) - x$. When the optimal mapping is close to the identity mapping, the residual mapping is optimized to be near zero, which is easier to optimize than directly learning the identity mapping. The specific structure of a residual block can be expressed as:

$$y = \mathcal{F}(x, \{W_i\}) + x \quad (1)$$

where x and y denote the input and output tensors of the residual block, respectively, $\mathcal{F}(\cdot)$ represents the residual mapping to be learned, typically consisting of two or three convolutional layers followed by batch normalization and ReLU activation functions. The residual connection allows gradients to propagate directly along the identity path to shallow layers, effectively alleviating the gradient vanishing problem and enabling stable training of deeper networks.

The specific network structure of ResNet12 is as follows: first, a 3×3 convolutional layer with 64 output channels and stride 1, followed by batch normalization and ReLU activation. Subsequently, four residual blocks (ResBlocks) are used, each containing two 3×3 convolutional layers, each followed by batch normalization and ReLU activation. The output channels of the four residual blocks are 64, 128, 256, and 512, respectively. After each residual block, if the number of channels changes, a 1×1 convolution is used for skip connection dimension matching. After the fourth residual block, global average pooling compresses the spatial dimension of each feature map to 1×1 , obtaining a 512-dimensional feature vector. For few-shot learning tasks, ResNet12 typically does not connect a fully connected classification layer after global average pooling but directly outputs feature vectors for use in downstream metric learning modules. Compared with deeper networks such as ResNet34 or ResNet50, ResNet12 has fewer parameters and faster inference speed, making it more suitable for rapid adaptation in few-shot scenarios. The feature extraction process of ResNet12 can be formalized as: given an input image $I \in \mathbb{R}^{224 \times 224 \times 3}$, after passing through the feature extraction network $\Phi(\cdot)$, a feature map $F \in \mathbb{R}^{H \times W \times C}$ is obtained, where H and W depend on the downsampling factor in the network. In the ResNet12 configuration adopted in this paper, after four residual blocks with stride-2 downsampling, $H = W = 14$ and $C = 512$. This feature map retains sufficient spatial resolution for subsequent zigzag scanning serialization operations while containing rich semantic information.

3.3. Riemannian Manifold Feature Embedding

3.3.1. Zigzag Feature Sequence

After obtaining the 2D feature map $F \in \mathbb{R}^{H \times W \times C}$ from ResNet12, how to effectively organize and represent these features becomes a key issue in few-shot remote sensing

scene classification. Traditional global average pooling averages features across all spatial positions to obtain a C -dimensional global feature vector, which significantly reduces dimensionality but completely loses spatial structural information. Traditional row-wise or column-wise flattening preserves information from all spatial positions but destroys local adjacency relationships among features and is sensitive to image rotation and flipping. To address these issues, this paper introduces a zigzag scanning strategy to serialize the 2D feature maps into 1D feature sequences.

Zigzag scanning was originally applied in the field of image compression, particularly in encoding discrete cosine transform coefficients in the JPEG standard. Its basic idea is to traverse matrix elements along diagonal directions in an alternating pattern, starting from the top-left corner, moving in a diagonal direction, changing direction upon reaching boundaries, until all elements are traversed. This scanning method has two advantages: first, zigzag scanning maintains spatial locality, with adjacent elements in the original feature map remaining adjacent or close in the scanned sequence; second, zigzag scanning is relatively robust to geometric transformations such as rotation and flipping, because the geometric properties of the scanning path align with the statistical characteristics of natural images.

For the feature map F , zigzag scanning is performed independently on each channel. Let the feature map of the c -th channel be $F_c \in \mathbb{R}^{H \times W}$. Figure 2 shows the zigzag scanning strategies of feature maps with different sizes. Taking the 4×4 feature map (Figure 2b) as an example, the zigzag scanning path follows the order:

$(0,0) \rightarrow (0,1) \rightarrow (1,0) \rightarrow (2,0) \rightarrow (1,1) \rightarrow (0,2) \rightarrow (0,3) \rightarrow (1,2) \rightarrow (2,1) \rightarrow (3,0) \rightarrow (3,1) \rightarrow (2,2) \rightarrow (1,3) \rightarrow (2,3) \rightarrow (3,2) \rightarrow (3,3)$.

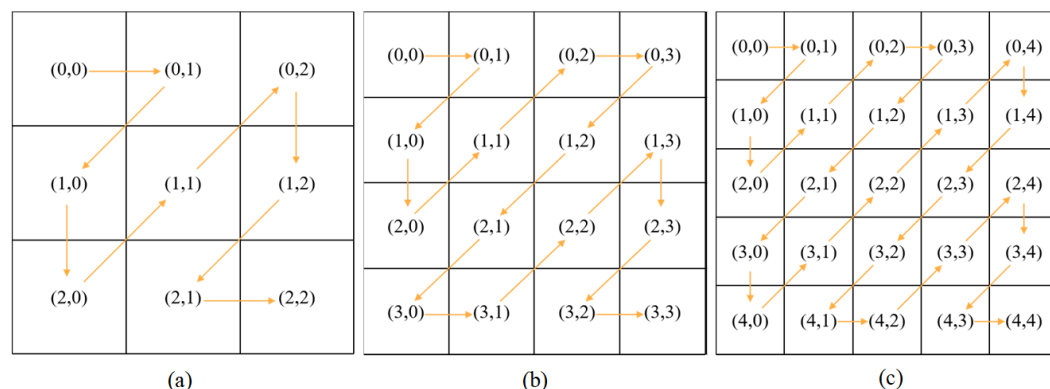


Figure 2. The zigzag scanning strategies of feature maps with different sizes: (a) 3×3 feature map; (b) 4×4 feature map; (c) 5×5 feature map.

The general traversal rules are defined as follows, starting from the top-left corner $(0,0)$ with an initial horizontal direction, and alternating diagonal traversal upon hitting matrix boundaries. For a square matrix with $H = W$, the zigzag scanning path can be formally described as:

Initialization: Start at $(r=0, c=0)$, first move right to $(0,1)$.
When moving in the down-left diagonal direction: If $r = H-1$ (bottom boundary), move right to $(r, c+1)$, switch to up-right direction; If $c = 0$ (left boundary), move down to $(r+1, c)$, switch to up-right direction; Otherwise, move to $(r+1, c-1)$, keep down-left direction.
When moving in the up-right diagonal direction: If $c = W-1$ (right boundary), move down to $(r+1, c)$, switch to down-left direction; If $r = 0$ (top boundary), move right to $(r, c+1)$, switch to down-left direction; Otherwise, move to $(r-1, c+1)$, keep up-right direction.

Following these rules, the traversal alternates between down-left and up-right diagonal passes, moving right or down at the matrix boundaries to start the next diagonal pass. After

traversing all $H \times W$ positions, a one-dimensional sequence is obtained. Concatenating the zigzag sequences of all C channels along the channel dimension yields the final feature sequence $V = [v_1, v_2, \dots, v_T]$, where $T = H \times W$ is the sequence length, and $v_t \in \mathbb{R}^C$ is the vector of feature values at the t -th position in zigzag order across all channels. This feature sequence V serves as the input data for subsequent ARMA modeling. The advantage of zigzag scanning over row-wise and column-wise scanning is: row-wise scanning first scans the first row from left to right, then the second row, etc., which maintains locality within rows but causes jumps when switching between rows; column-wise scanning similarly suffers from jumps between columns; while zigzag scanning traverses along diagonal directions, keeping spatially adjacent points adjacent in the sequence, better matching the spatial correlation structure of natural images.

3.3.2. Autoregressive Moving Average Model

After obtaining the zigzag feature sequence $V = [v_1, v_2, \dots, v_T]$, it is necessary to further model the dependency relationships among feature vectors in the sequence to capture the spatial structural patterns of remote sensing images. Traditional sequence modeling methods such as recurrent neural networks or Transformers are powerful but require large amounts of training data and have high computational complexity, making them unsuitable for few-shot scenarios. This paper employs the autoregressive moving average (ARMA) model to model the feature sequence, which has low parameterization and good statistical interpretability, and its parameters can be naturally mapped to Riemannian manifolds.

The ARMA model is a classical time series analysis method that expresses the current observation as a linear combination of historical observations plus a linear combination of historical noises. In this paper, the zigzag sequence V is treated as a multivariate time series, where each time step t corresponds to a spatial position on the zigzag path, and the observation vector $v_t \in \mathbb{R}^C$ is the feature value of all channels at that position. To capture spatial dependencies in the sequence, this paper adopts a state-space form of the ARMA model, whose core idea is to introduce a low-dimensional state variable to capture the core dynamic characteristics of the sequence. The state-space ARMA model consists of an observation equation and a state transition equation, expressed as:

$$\begin{cases} y_t = A \cdot x_t + \varepsilon_t & \text{(Observation equation)} \\ x_{t+1} = B \cdot x_t + \eta_t & \text{(State transition equation)} \end{cases} \quad (2)$$

Where $x_t \in \mathbb{R}^d$ is the state vector at time t , d is the state dimension, typically much smaller than the observation dimension C , used to extract the core dependencies of the sequence. $A \in \mathbb{R}^{C \times d}$ is the observation matrix, mapping the low-dimensional state vector to the high-dimensional observation space. $B \in \mathbb{R}^{d \times d}$ is the state transition matrix, characterizing the dynamic evolution of the state over time and reflecting the spatial dependency relationships between adjacent feature vectors in the zigzag sequence. $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$ is the zero-mean Gaussian observation noise, with $\Sigma_\varepsilon \in \mathbb{R}^{C \times C}$ as the covariance matrix, corresponding to noise interference in the feature extraction process. $\eta_t \sim \mathcal{N}(0, \Sigma_\eta)$ is the zero-mean Gaussian state noise, with $\Sigma_\eta \in \mathbb{R}^{d \times d}$ as the covariance matrix, corresponding to uncertainty in state changes.

Estimating the ARMA model parameters $(A, B, \Sigma_\varepsilon, \Sigma_\eta)$ is a critical issue under few-shot conditions. Since the state variable x_t is an unobservable latent variable, direct maximum likelihood estimation is complex. This paper adopts an estimation method based on singular value decomposition (SVD), which has high computational efficiency and numerical stability. First, to avoid redundant representation of the state space, an orthogonality constraint $A^T A = I_d$ is imposed on the observation matrix, where I_d is the d -order identity matrix. This constraint ensures the uniqueness of the state space basis

and simplifies parameter solving. Then, singular value decomposition is performed on the zigzag feature sequence matrix $V \in \mathbb{R}^{C \times T}$:

$$V = U \cdot \Lambda \cdot W^T \quad (3)$$

Where $U \in \mathbb{R}^{C \times r}$ is the matrix of left singular vectors (columns orthogonal), $\Lambda \in \mathbb{R}^{r \times r}$ is a diagonal matrix with singular values on the diagonal, $W \in \mathbb{R}^{T \times r}$ is the matrix of right singular vectors (columns orthogonal), and $r = \min(C, T)$ is the rank of the matrix. According to the orthogonality constraint, the estimate of the observation matrix takes the first d left singular vectors, i.e., $\hat{A} = U_{:,1:d}$, because the left singular vectors already contain the core semantic information of the feature sequence and satisfy the orthogonality condition. The estimate of the state vector sequence is $\hat{X} = \Lambda_{1:d,1:d} \cdot W_{:,1:d}^T \in \mathbb{R}^{d \times T}$, where each column of $\hat{X}$ corresponds to the state vector $\hat{x}_t$ at a given time.

The state transition matrix B must satisfy $\hat{x}_{t+1} \approx B \cdot \hat{x}_t$ and is solved using the least squares method with the pseudoinverse:

$$\hat{B} = \hat{X}_{2:T} \cdot (\hat{X}_{1:T-1})^\dagger \quad (4)$$

Where $(\cdot)^\dagger$ denotes the matrix pseudoinverse. When $T - 1 \geq d$, the pseudoinverse exists and is unique; when the sequence length T is small, the pseudoinverse still provides a stable minimum-norm solution. The state noise is $\eta_t = \hat{x}_{t+1} - \hat{B} \cdot \hat{x}_t$, and its covariance matrix is estimated as:

$$\hat{\Sigma}_\eta = \frac{1}{T-1} \sum_{t=1}^{T-1} \eta_t \cdot \eta_t^T \quad (5)$$

The observation noise is $\varepsilon_t = y_t - \hat{A} \cdot \hat{x}_t$, and its covariance matrix is estimated as:

$$\hat{\Sigma}_\varepsilon = \frac{1}{T} \sum_{t=1}^T \varepsilon_t \cdot \varepsilon_t^T \quad (6)$$

Thus, the zigzag feature sequence of each sample is represented by the ARMA model parameters $(\hat{A}, \hat{B}, \hat{\Sigma}_\eta)$. These parameters capture the dynamic evolution patterns of the feature sequence, reflecting the spatial dependency characteristics of land cover structures in remote sensing images. However, directly using these parameters as feature vectors ignores the non-Euclidean topological structure of the parameter space, so they need to be further mapped to a Riemannian manifold.

3.4. Riemannian Manifold Feature Extraction

The ARMA model parameters $(\hat{A}, \hat{B}, \hat{\Sigma}_\eta)$ constitute a dynamic description of the zigzag feature sequence, but these parameters lie in spaces with specific geometric structures, and directly applying Euclidean methods would lose important geometric information. Specifically, the observation matrix $\hat{A}$ satisfies the orthogonality constraint $\hat{A}^T \hat{A} = I_d$, and the d -dimensional subspace $\text{span}(\hat{A})$ spanned by its column vectors corresponds to a point on the Grassmann manifold $\mathcal{G}_{d,C-d}$. The state transition matrix $\hat{B}$ lies in the matrix space $\mathbb{R}^{d \times d}$, and the state noise covariance $\hat{\Sigma}_\eta$ lies on the symmetric positive definite matrix manifold $\mathcal{P}_d$. To fully exploit these geometric structures, this paper first maps the ARMA parameters to the Grassmann manifold and then adopts the SPDNet architecture for deeper Riemannian manifold feature extraction.

The core idea of mapping ARMA parameters to the Grassmann manifold is to construct an extended observation matrix that incorporates the dynamic information of state transitions into the subspace representation. Specifically, the extended observation matrix is defined as:

$$O = [\hat{A}^T, (\hat{A} \cdot \hat{B})^T, (\hat{A} \cdot \hat{B}^2)^T, \dots, (\hat{A} \cdot \hat{B}^{L-1})^T]^T \in \mathbb{R}^{L \times C} \quad (7)$$

Where L is the length of the zigzag sequence, and $\hat{A} \cdot \hat{B}^l$ denotes the observation matrix after l steps of state transition. Each row of the extended observation matrix O corresponds to an observation after a specific order of state transition, and the subspace $\text{span}(O^T)$ spanned by its row vectors fuses dynamic information from multiple time scales. QR decomposition is performed on O to obtain an orthonormal basis matrix $P \in \mathbb{R}^{C \times L}$ (satisfying $P^T P = I_L$), and the subspace corresponding to this basis matrix is the point g on the Grassmann manifold $\mathcal{G}_{L, C-L}$. This mapping process encodes the dynamic characteristics of the ARMA model as geometric points on the Grassmann manifold, enabling similarity between different samples to be measured by geodesic distances on the manifold.

After obtaining the Grassmann manifold representation, this paper uses a Riemannian manifold feature extraction network based on the SPDNet architecture to further process these representations. SPDNet is a deep neural network specifically designed for handling symmetric positive definite matrix data and is capable of performing nonlinear transformations and feature learning while preserving SPD structure. First, the Grassmann manifold points need to be converted into SPD matrix representations. Given the orthonormal basis matrix $P \in \mathbb{R}^{C \times L}$, construct the SPD matrix $G = PP^T \in \mathbb{R}^{C \times C}$, which is a rank- L positive semidefinite matrix. By adding a regularization term ϵI , it becomes strictly positive definite: $G_\epsilon = G + \epsilon I$, where ϵ is a small positive number. Thus, each sample is represented as an SPD matrix $G_\epsilon \in \mathcal{P}_C$ on the symmetric positive definite matrix manifold.

The core structure of SPDNet includes BiMap layers and ReEig layers. Let the input of the l -th layer be the SPD matrix $M^{(l-1)} \in \mathcal{P}_{d_{l-1}}$. The operation of a BiMap layer is defined as:

$$M^{(l)} = T_l M^{(l-1)} T_l^T \quad (8)$$

Where $T_l \in \mathbb{R}^{d_l \times d_{l-1}}$ is a column full-rank transformation matrix ($d_l < d_{l-1}$), and its row number d_l determines the dimension of the output SPD matrix. To ensure that the output remains an SPD matrix, T_l must satisfy the full-rank condition. The BiMap layer functions similarly to a fully connected or convolutional layer in a CNN, projecting high-dimensional SPD matrices to low-dimensional SPD spaces through bilinear transformation while preserving the manifold structure. This operation can be understood as a coordinate transformation on the manifold, recombining original features into more discriminative representations. The transformation matrix T_l is a learnable parameter of the network, optimized through backpropagation.

The ReEig layer (rectified eigenvalues layer) is analogous to the ReLU activation function in CNNs and performs a nonlinear transformation on the eigenvalues of SPD matrices to enhance feature discriminability. The ReEig layer is defined as:

$$M^{(l)} = U^{(l-1)} \max(\tau I, \Sigma^{(l-1)}) (U^{(l-1)})^T \quad (9)$$

Where $M^{(l-1)} = U^{(l-1)} \Sigma^{(l-1)} (U^{(l-1)})^T$ is the eigenvalue decomposition of the input SPD matrix, $\Sigma^{(l-1)} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_{d_{l-1}})$ is a diagonal matrix with eigenvalues on the diagonal, τ is a preset threshold parameter, and $\max(\tau I, \Sigma^{(l-1)})$ takes the element-wise maximum, i.e., $\max(\tau, \sigma_i)$. This operation raises all eigenvalues smaller than τ to τ , preventing numerical instability due to excessively small eigenvalues while introducing nonlinear activation capability. Through the ReEig layer, the network can learn more discriminative feature representations, suppressing noise and redundant information. After stacking multiple BiMap and ReEig layers, the features are input into the Riemannian geometric barycenter network for classification.

3.5. Riemannian Geometric Barycenter Network

To further enhance the discriminative power of deep features, this paper designs a Riemannian geometric barycenter network (RGBN). The core idea of this network is to cluster features of the same class on the Riemannian manifold while separating features of different classes. The RGBN consists of two submodules: the barycentric normalization layer and the barycentric classification layer.

The barycentric normalization layer aims to align and stabilize the SPD features within a batch. Since Riemannian distances follow a right-skewed distribution, direct computation of the barycenter can be unstable due to outliers, and feature distributions may vary across different batches. Let the SPD features in a mini-batch be $\{G_i\}_{i=1}^{N_b}$, where N_b is the batch size. First, compute the Riemannian barycenter (geometric mean) $\bar{G}$ of these features, defined as the point minimizing the weighted sum of squared distances:

$$\bar{G} = \arg \min_{G \in \mathcal{P}_c} \sum_{i=1}^{N_b} \delta_R^2(G_i, G) \quad (10)$$

Where $\delta_R(\cdot, \cdot)$ is the affine-invariant Riemannian distance, defined as:

$$\delta_R(G_i, G_j) = \|\log(G_i^{-1/2} G_j G_i^{-1/2})\|_F = \left(\sum_{c=1}^C \log^2 \lambda_c \right)^{1/2} \quad (11)$$

λ_c is the eigenvalues of the matrix $G_i^{-1/2} G_j G_i^{-1/2}$. The Riemannian barycenter can be computed iteratively using the Karcher flow algorithm: first project all SPD features to the tangent space at $\bar{G}$ via the logarithmic map, compute the arithmetic mean in the tangent space, and then map back to the manifold via the exponential map, repeating until convergence.

The process of aligning each feature G_i to the barycenter $\bar{G}$ involves three steps. First, project G_i to the tangent space at the barycenter using the logarithmic map:

$$S_i = \log_{\bar{G}}(G_i) = \bar{G}^{1/2} \log(\bar{G}^{-1/2} G_i \bar{G}^{-1/2}) \bar{G}^{1/2} \quad (12)$$

Second, perform parallel transport of S_i in the tangent space, moving the tangent vector from the tangent space at $\bar{G}$ to the tangent space at a reference point G_0 . Parallel transport is defined as:

$$\Gamma_{\bar{G} \rightarrow G_0}(S_i) = (G_0 \bar{G}^{-1})^{1/2} S_i (G_0 \bar{G}^{-1})^{1/2} \quad (13)$$

Finally, map the tangent vector back to the manifold using the exponential map, obtaining the normalized SPD feature:

$$\tilde{G}_i = \exp_{G_0}(\Gamma_{\bar{G} \rightarrow G_0}(S_i)) = G_0^{1/2} \exp(G_0^{-1/2} \Gamma_{\bar{G} \rightarrow G_0}(S_i) G_0^{-1/2}) G_0^{1/2} \quad (14)$$

After barycentric normalization, the SPD features within the batch are aligned to a common reference point, reducing distributional differences and the influence of outliers.

The goal of the barycentric classification layer is to simultaneously minimize intra-class distances and maximize inter-class distances. Let the normalized features be $\{\tilde{G}_i\}$ with corresponding labels $y_i \in \{1, 2, \dots, N_c\}$, where N_c is the number of classes. For each class c , a learnable Riemannian barycenter B_c is maintained, initialized as the Riemannian barycenter of the support samples of that class and updated during training via gradient descent. The intra-class compactness loss encourages features of the same class to approach their class barycenter:

$$\mathcal{L}_{\text{intra}} = \frac{1}{2N_b} \sum_{i=1}^{N_b} \delta_R^2(\tilde{G}_i, B_{y_i}) \quad (15)$$

The inter-class separation loss increases the distance differences between features of different classes and their barycenters. Inspired by the Z-score concept in statistical testing, a loss term based on distance distribution normalization is designed:

$$\mathcal{L}_{\text{inter}} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{j \neq y_i}^{N_c} \frac{\log(d_{i,j}/\mu_i)^2}{\log(\sigma_i)} + \text{CrossEntropy}(z_i, y_i) \quad (16)$$

Where $d_{i,j} = \delta_R(\tilde{G}_i, B_j)$ is the Riemannian distance between sample $\tilde{G}_i$ and the barycenter B_j of class j , μ_i and σ_i are the running mean and standard deviation of $d_{i,j} (j \neq y_i)$ in the current batch, and z_i is the classification logits obtained by passing $\tilde{G}_i$ through the LogEig layer and a fully connected layer. The first term measures the deviation of samples from non-class barycenters using standardized distances, increasing the penalty when a sample falls within the distance distribution of other classes; the second term is the standard cross-entropy loss, ensuring basic classification capability.

To ensure the numerical stability of the logarithmic term $\log(\sigma_i)$ in $\mathcal{L}_{\text{inter}}$, we use an exponential moving average to update σ_i during training:

$$\sigma_i^{(t)} = \beta \sigma_i^{(t-1)} + (1 - \beta) \hat{\sigma}_i^{(t)}, \quad (17)$$

Where $\hat{\sigma}_i^{(t)}$ is the batch-wise estimate at the t -th iteration and $\beta = 0.9$ is the momentum coefficient. Before evaluating the logarithm, σ_i is clamped into a valid numerical range:

$$\tilde{\sigma}_i^{(t)} = \text{clamp}(\sigma_i^{(t)}, \epsilon, 1 - \epsilon), \quad (18)$$

Where $\epsilon = 10^{-6}$. The logarithmic term is therefore computed as $\log(\tilde{\sigma}_i^{(t)})$. This prevents the input of the logarithm from becoming zero or extremely small. The derivative of this term is:

$$\frac{\partial \log(\tilde{\sigma}_i)}{\partial \tilde{\sigma}_i} = \frac{1}{\tilde{\sigma}_i}. \quad (19)$$

Since $\tilde{\sigma}_i \in [\epsilon, 1 - \epsilon]$, the gradient is bounded by $1/\epsilon$, avoiding infinite gradients. Moreover, exponential moving average smoothing and gradient clipping further suppress abrupt gradient oscillations, ensuring stable end-to-end optimization.

The proposed ZSFS-RGBN method adopts an end-to-end joint training strategy, with the total loss function composed of several complementary loss terms that jointly optimize the quality of feature representations and classification performance. The design of the total loss function follows three principles: feature discriminability, classification accuracy, and geometric constraint integrity. Feature discriminability requires compact intra-class features and separated inter-class features; classification accuracy requires the model to correctly predict sample categories; geometric constraint integrity requires that operations on the manifold maintain the closure of the SPD structure.

The complete form of the total loss function is:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{intra}} + (1 - \lambda) \mathcal{L}_{\text{inter}}$$

Where $\mathcal{L}_{\text{intra}}$ is the intra-class compactness loss, $\mathcal{L}_{\text{inter}}$ is the inter-class separation loss, λ is a hyperparameter balancing the loss terms. Through this end-to-end joint optimization, ZSFS-RGBN learns deep feature representations with intra-class compactness and

inter-class separability, significantly improving the performance of remote sensing scene classification under few-shot conditions.

4. Experiments and Analysis

4.1. Datasets Description

To evaluate the performance and robustness of the proposed framework, we conducted experiments on three widely recognized remote sensing image scene classification benchmarks: NWPU-RESISC45 (NWPU) [2], UC-Merced (UC) [7], and WHU-RS19 (WHU) [58]. In order to maintain a rigorous evaluation with prior research, we adopted a category-based splitting strategy for the training (S_{train}), validation (S_{val}), and testing (S_{test}) sets. The detailed distribution of scene categories across these three subsets is documented in Table 1.

Table 1. The splits for the training dataset, validation dataset, and testing dataset.

Name	S_{train}	S_{val}	S_{test}
NWPU dataset	Chaparral, Railway, Ship, Lake, Sparse residential, Wetland, Harbor, Mountain, Tennis court, Cloud, Stadium, Airplane, Meadow, Desert, Bridge, Church, Freeway, Palace, Roundabout, Island, Mobile home park, Rectangular farmland, Sea ice, Golf course, Beach	Runway, Park, Thermal power station, Industrial area, Snowberg, Railway station, Commercial area, Overpass, Storage tank, Terrace	Airport, Forest, Ground track field, Parking lot, Dense residential, Intersection, River, Circular farmland, Basketball court, Medium residential
UC dataset	Freeway, Overpass, Agricultural, Harbor, Parking lot, Chaparral, Buildings, Dense residential, Baseball diamond, Medium residential	Forest, Airplane, Runway, Intersection, Storage tanks	Beach, Mobile home park, Sparse residential, River, Golf course, Tennis court
WHU dataset	Airport, Mountain, Bridge, Parking, Industrial, Desert, Residential, Football field, Port	Park, Forest, Beach, Railway station, Farmland	Pond, River, Commercial, Viaduct, Meadow

The NWPU dataset, released by Northwestern Polytechnical University, stands as one of the most comprehensive and challenging benchmarks in the field. It comprises 31,500 remote sensing images categorized into 45 distinct scene classes, with 700 images per class. Each image features a spatial resolution of 256×256 pixels. Due to its large scale and high intraclass variation, it provides a rigorous test for the generalization capabilities of our model. In our experiments, 25 categories were allocated for training, 10 for validation, and the remaining 10 for testing.

The UC dataset is a classic benchmark for land-use scene recognition, which contains 2100 high-resolution images across 21 land-use categories. Each category consists of 100 images with a pixel size of 256×256 . The images were originally derived from the USGS National Map Urban Area Imagery collection. For this study, the 21 categories were split into 10 for training, 5 for validation, and 6 for testing, following the distribution specified in Table 1.

The WHU dataset is provided by Wuhan University; this dataset is composed of 950 high-resolution images (600×600 pixels) extracted from Google Earth. It encompasses 19 scene categories, with 50 images per category. Despite its relatively small total volume, the dataset is characterized by significant variations in scale, illumination, and orientation,

making it an ideal choice for testing model sensitivity to environmental changes. We divided the 19 categories into 9 for training, 5 for validation, and 5 for testing.

4.2. Experiment Setup

All experiments in this paper are implemented under the PyTorch 2.2.2 deep learning framework and run on a hardware platform configured with an NVIDIA RTX 2080Ti GPU to ensure efficient training and inference. To maintain input consistency and reduce the impact of resolution differences on feature extraction, all remote sensing images from the three datasets are uniformly resized to a fixed resolution of 224×224 pixels before being fed into the network. During the training phase, a series of effective data augmentation strategies are applied to improve the generalization ability of the model and alleviate overfitting under limited samples, which include random horizontal flipping, random center cropping, and random grayscale transformation. In accordance with the standard evaluation paradigm of few-shot learning, this paper adopts an episode-based meta-training mechanism for model optimization. Specifically, 2000 episodes are randomly constructed for the training stage, 500 episodes for the validation stage, and another 2000 independent episodes for the final testing stage to ensure statistical reliability. Each episode strictly follows the 5-way K -shot setting ($K = 1$ for 1-shot and $K = 5$ for 5-shot). In each episode, 5 distinct categories are first sampled from the corresponding dataset split, then K labeled images per category are selected as the support set for prototype learning, and 15 unlabeled images per category are used as the query set for performance evaluation. To quantitatively evaluate the classification performance of the proposed method and all comparison approaches, the average classification accuracy (CA) over 2000 testing episodes with a 95% confidence interval is utilized as the primary evaluation metric. For model optimization, the Stochastic Gradient Descent (SGD) optimizer is employed to update all network parameters, and the initial learning rate is set to 1×10^{-3} to ensure stable convergence and effective gradient propagation.

4.3. Analysis of Experiment Results

To comprehensively evaluate the effectiveness of our proposed ZSFS-RGBN method, we compared it with various state-of-the-art (SOTA) few-shot learning methods across three public remote sensing datasets: NWPU-RESISC45 (NWPU), UC-Merced (UC), and WHU-RS19 (WHU). The evaluation was conducted under the standard 5-way 1-shot and 5-way 5-shot classification settings. The quantitative results are presented in Table 2, Table 3 and Table 4.

As shown in Table 2, the NWPU dataset is highly challenging due to its large scale and severe intraclass variations. Despite these challenges, our proposed ZSFS-RGBN method achieves the highest classification accuracies of 84.59% and 92.19% in the 5-way 1-shot and 5-way 5-shot settings, respectively. Compared to the second-best performing method, DiffPR-Net, our approach yields significant performance gains of 2.74% in the 1-shot scenario and 2.78% in the 5-shot scenario. Furthermore, compared to classic meta-learning approaches such as MAML and RelationNet (RN), the improvements are remarkably substantial (e.g., over 35% higher than MAML in 1-shot). This demonstrates that the ZSFS-RGBN framework is highly capable of extracting discriminative and robust features even when data is extremely scarce and scene complexity is high.

Table 2. Classification accuracies (%) of different methods on the nwpu dataset.

Method	Five-Way	
	One-Shot	Five-Shot
MN [18]	42.81 ± 0.12	47.71 ± 0.26
S2M2 [59]	63.24 ± 0.47	83.23 ± 0.28
Meta-SGD [60]	60.56 ± 0.28	74.33 ± 0.65
TPN [61]	66.51 ± 0.87	78.50 ± 0.56
DeepEMD [62]	64.39 ± 0.84	78.01 ± 0.56
RN [20]	66.05 ± 0.32	78.24 ± 0.47
IDLN [63]	75.25 ± 0.75	84.67 ± 0.23
PN [19]	41.57 ± 0.42	63.09 ± 0.08
DLA-MatchNet [57]	68.80 ± 0.70	81.63 ± 0.46
GES-Net [64]	70.83 ± 0.85	82.27 ± 0.55
MAML [21]	47.32 ± 0.10	63.03 ± 0.55
MKN [54]	65.84 ± 0.89	82.67 ± 0.55
Baseline (Conv-4) [65]	62.21 ± 0.48	79.13 ± 0.32
SCL-MLNet [66]	62.21 ± 1.12	80.86 ± 0.76
LLSR [44]	51.43	72.90
TAE-Net [67]	69.13 ± 0.83	82.37 ± 0.52
RS-MetaNet [56]	46.21 ± 0.58	68.75 ± 0.70
CovaMNet [68]	43.62 ± 0.67	67.23 ± 0.15
SIB [69]	67.34 ± 0.81	78.28 ± 0.48
TSC [70]	73.26 ± 0.15	84.62 ± 0.35
CAN+T [71]	69.89 ± 0.58	81.04 ± 0.33
SPNet [22]	67.84 ± 0.87	83.94 ± 0.50
Qin et al. [72]	69.48 ± 0.85	84.00 ± 0.46
PCFGNet [73]	72.05 ± 0.75	85.07 ± 0.45
DN4AM [74]	70.15 ± 0.82	85.69 ± 0.69
TeAW [75]	70.23 ± 0.45	85.77 ± 0.25
ICSFF [76]	70.93 ± 0.53	86.83 ± 0.77
LDRNet [25]	71.91 ± 0.65	84.50 ± 0.40
PA-SRM [26]	72.65 ± 0.43	83.64 ± 0.61
ACL-Net [27]	76.13 ± 0.24	86.54 ± 0.23
DBA-RMCL [77]	77.15 ± 0.34	86.64 ± 0.20
DiffPR-Net [78]	81.85 ± 0.32	89.41 ± 0.20
ZSFS-RGBN (ours)	84.59 ± 0.39	92.19 ± 0.47

Table 3 reports the comparative results on the UC dataset. While this dataset has fewer categories and images, it remains a standard benchmark for land-use classification. The proposed ZSFS-RGBN method once again consistently outperforms all other compared methods, achieving 63.22% for 5-way 1-shot and 82.61% for 5-way 5-shot tasks. Notably, in the 5-shot setting, our method surpasses the highly competitive DBA-RMCL model by a notable margin of 3.88%. This indicates that our model can effectively leverage slightly increased support samples to construct a much more accurate class representation, proving its strong adaptability to standard land-use scenarios.

The WHU dataset is characterized by variations in illumination, scale, and shooting angles. Table 4 illustrates that our ZSFS-RGBN model exhibits exceptional robustness to these environmental changes. It achieves a superior accuracy of 89.14% in the 1-shot setting, exceeding the second-best DBA-RMCL (87.33%) by 1.81%. In the 5-shot setting, our method reaches an impressive accuracy of 98.69%, outperforming the strong baseline ICSFF (97.56%) by 1.13%. The ability to achieve near-perfect classification in the 5-shot scenario confirms that the proposed method effectively mitigates the negative impacts of multi-scale and illumination variations inherent in high-resolution remote sensing imagery.

Across all three datasets, a consistent trend can be observed: the performance in the 5-way 5-shot setting is uniformly better than that in the 5-way 1-shot setting for all methods. This is intuitive as providing five support samples per class rather than one significantly enriches the reference information, thereby reducing the risk of overfitting and improving the reliability of the prototype or feature matching. Most importantly, the proposed ZSFS-RGBN method maintains its leading position across datasets of varying scales and complexities. Whether dealing with the massive 45-category NWPU dataset or the highly variable WHU dataset, our model demonstrates superior generalization capability and stability, verifying the effectiveness of the core algorithmic design in ZSFS-RGBN for few-shot remote sensing scene classification.

Table 3. Classification accuracies (%) of different methods on the uc dataset.

Method	Five-Way	
	One-Shot	Five-Shot
MN [18]	54.35 ± 0.27	68.03 ± 0.33
PN [19]	51.46 ± 0.16	68.07 ± 0.23
Meta-SGD [60]	48.00 ± 0.14	63.44 ± 0.26
MAML [21]	47.27 ± 0.17	61.92 ± 0.18
RN [20]	48.15 ± 0.28	62.01 ± 0.06
DLA-MatchNet [57]	53.76 ± 0.62	63.01 ± 0.51
SCL-MLNet [66]	51.37 ± 0.79	68.09 ± 0.92
MKN [54]	58.45 ± 0.54	77.92 ± 0.32
TeAW [75]	56.94 ± 0.39	77.50 ± 0.27
MPCL-Net [79]	56.46 ± 0.21	76.57 ± 0.07
ACL-Net [27]	59.74 ± 0.39	74.98 ± 0.29
DBA-RMCL [77]	61.50 ± 0.21	78.73 ± 0.25
ZSFS-RGBN (ours)	63.22 ± 0.72	82.61 ± 0.50

Table 4. Classification accuracies (%) of different methods on the whu dataset.

Method	Five-Way	
	One-Shot	Five-Shot
Meta-SGD [60]	51.78 ± 1.05	65.47 ± 0.65
GES-Net [64]	75.87 ± 0.78	82.37 ± 0.38
MN [18]	51.36 ± 0.68	54.27 ± 0.41
S2M2 [59]	69.00 ± 0.41	82.14 ± 0.21
MAML [21]	51.06 ± 0.21	65.83 ± 0.17
RN [20]	62.70 ± 0.43	78.25 ± 0.34
PCFGNet [73]	72.41 ± 0.91	85.26 ± 0.66
PN [19]	58.62 ± 0.41	80.63 ± 0.39
DeepEMD [62]	63.76 ± 0.74	78.25 ± 0.43
DLA-MatchNet [57]	68.27 ± 1.83	79.89 ± 0.33
CovaMNet [68]	59.06 ± 0.73	70.29 ± 0.62
Qin et al. [72]	61.62 ± 0.75	81.32 ± 0.62
SCL-MLNet [66]	63.36 ± 0.88	77.62 ± 0.81
LLSR [44]	57.1	70.65
Baseline (Conv-4) [65]	71.42 ± 0.33	86.12 ± 0.16
SIB [69]	55.84 ± 0.65	65.96 ± 0.48
TSC [70]	70.99 ± 0.74	82.18 ± 0.32
CAN+T [71]	69.79 ± 0.56	79.71 ± 0.22
DN4AM [74]	84.98 ± 0.63	96.72 ± 0.20
TeAW [75]	85.32 ± 0.52	96.97 ± 0.16
ICSFF [76]	87.31 ± 0.25	97.56 ± 0.08

Table 4. Cont.

Method	Five-Way	
	One-Shot	Five-Shot
PA-SRM [26]	77.49 ± 0.71	88.86 ± 0.58
ACL-Net [27]	78.30 ± 0.32	90.43 ± 0.15
DiffPR-Net [78]	85.43 ± 0.21	95.69 ± 0.10
DBA-RMCL [77]	87.33 ± 0.22	93.88 ± 0.10
ZSFS-RGBN (ours)	89.14 ± 0.47	98.69 ± 0.37

4.4. Ablation Studies

To systematically evaluate the contribution of each core component in the proposed ZSFS-RGBN framework, we conduct comprehensive ablation experiments on the NWPU, UC, and WHU datasets under both 5-way 1-shot and 5-way 5-shot settings. The three key components under investigation are RMFEmb, RMFExt, and RieGB. RMFEmb includes zigzag scanning, ARMA modeling, and mapping to the Grassmann/SPD manifold. Removing it replaces the entire manifold embedding with simple channel-wise flattening and direct Euclidean feature vectors. RMFExt consists of the BiMap and ReEig layers that perform structure-preserving transformations on the SPD manifold. Removing RMFExt bypasses these layers, feeding the embedded SPD matrices directly into the barycenter layer. RieGB includes the barycentric normalization and barycentric classification layers with intra-class compactness and inter-class separation losses. Removing it replaces the RieGB layer with a standard SPDNet classification head (LogEig layer followed by a fully connected layer with cross-entropy loss).

Based on these, we construct four model variants:

Baseline (w/o all): Removes all three components. The ResNet12 feature maps are globally average-pooled into Euclidean vectors, and classification is performed using standard Prototypical Networks with Euclidean distance.

+ RMFEmb: Adds the RMFEmb layer to the baseline, but without RMFExt and RieGB. Classification uses a single BiMap layer followed by LogEig and fully connected layers.

+ RMFEmb + RMFExt: Further adds the RMFExt component, but still without the RieGB layer.

ZSFS-RGBN (Full): Includes all three components (RMFEmb, RMFExt, RieGB) as described in Section 3.

All methods use the same ResNet12 backbone, training episodes, and hyperparameters for a fair comparison. The experimental results are reported in Table 5.

Table 5. Ablation study results of ZSFS-RGBN on three datasets (%).

Methods	NWPU		UC Merced		WHU-RS19	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
Baseline (w/o all)	76.13 ± 0.45	85.86 ± 0.38	59.21 ± 0.38	73.54 ± 0.41	83.35 ± 0.40	94.62 ± 0.28
+RMFEmb	78.52 ± 0.41	87.43 ± 0.35	61.08 ± 0.45	77.56 ± 0.36	85.21 ± 0.44	95.43 ± 0.35
+RMFEmb+RMFExt	81.09 ± 0.38	89.02 ± 0.32	61.95 ± 0.52	79.13 ± 0.44	87.46 ± 0.39	96.01 ± 0.31
ZSFS-RGBN (Full)	84.59 ± 0.39	92.19 ± 0.47	63.22 ± 0.72	82.61 ± 0.50	89.14 ± 0.47	98.69 ± 0.37

Based on the results in Table 5, the progressive integration of our proposed modules demonstrates a clear and consistent performance boost over the Euclidean baseline. Introducing the RMFEmb module brings steady accuracy gains across all datasets. Under the 1-shot setting, it improves accuracy by 2.39% on NWPU, 1.87% on UC Merced, and 1.86% on WHU-RS19. Under the 5-shot setting, the gains reach 1.57% on NWPU, 4.02% on

UC Merced, and 0.81% on WHU-RS19. This indicates that encoding feature maps into Riemannian manifolds successfully captures higher-order statistical and spatial information typically lost in standard global average pooling. Building upon RMFEmb, the addition of the RMFExt module further refines these geometric representations. On the NWPU dataset, accuracy increases by 2.57% under the 1-shot setting and 1.59% under the 5-shot setting. On UC Merced, the improvements are 0.87% under the 1-shot setting and 1.57% under the 5-shot setting. On WHU-RS19, gains reach 2.25% (1-shot) and 0.58% (5-shot). These results prove that performing structure-preserving transformations directly on the manifold effectively extracts more discriminative non-Euclidean features. Finally, integrating the RieGB layer completes the architecture and provides substantial performance leaps. Under the 1-shot setting, further improvements are 3.50% on NWPU, 1.27% on UC Merced, and 1.68% on WHU-RS19. Under the 5-shot setting, the gains are 3.17% on NWPU, 3.48% on UC Merced, and 2.68% on WHU-RS19. This highlights that computing the Riemannian barycenter accurately estimates stable class prototypes, sharpening decision boundaries and reducing prototype bias when multiple support samples are available.

The ablation study confirms that the RMFEmb, RMFExt, and RieGB components are indispensable and mutually reinforcing. Working in synergy, the complete ZSFS-RGBN framework achieves total absolute improvements over the baseline ranging from 4.01% (UC Merced, 1-shot) to 9.07% (UC Merced, 5-shot), with other settings improving by 5.79%–8.46%. These significant gains thoroughly validate the superiority and necessity of our proposed Riemannian geometric framework for few-shot remote sensing image classification.

4.5. Hyperparameters Analysis

(1) Sensitivity analysis of balance factor

The balance factor λ in the joint loss function is a critical hyperparameter that regulates the relative importance between intra-class compactness loss and inter-class separability loss. To determine the optimal value of λ for few-shot remote sensing scene classification, we conducted exhaustive ablation experiments by varying λ from 0 to 1.0 at intervals of 0.1, covering all typical weighting ratios. All experiments were implemented on three benchmark datasets, namely NWPU, UC, and WHU, under two standard few-shot settings: 5-way 1-shot and 5-way 5-shot. All other experimental configurations and network parameters were kept strictly consistent to ensure fairness.

The quantitative classification results under different λ values are visualized in Figure 3. As observed from the experimental curves, the classification accuracy exhibits obvious fluctuations with the continuous adjustment of λ , which is mainly attributed to the different sensitivities of each remote sensing dataset to the loss balance ratio under limited sample conditions. For the 5-way 1-shot setting, the optimal λ values vary across datasets: the proposed model achieves the best performance when $\lambda = 0.6$ on the NWPU dataset, $\lambda = 0.3$ on the UC dataset, and $\lambda = 0.5$ on the WHU dataset. For the 5-way 5-shot setting, the optimal values shift to 0.9 for NWPU, 0.6 for UC, and 0.9 for WHU. The overall results demonstrate that the classification performance is highly sensitive to the selection of λ , and inappropriate values will weaken the feature discriminability by breaking the balance between intra-class aggregation and inter-class separation. The optimal λ values obtained from comparative experiments are adopted in all subsequent formal experiments of the ZSFS-RGBN model, ensuring the stability and superiority of classification performance on diverse remote sensing datasets.

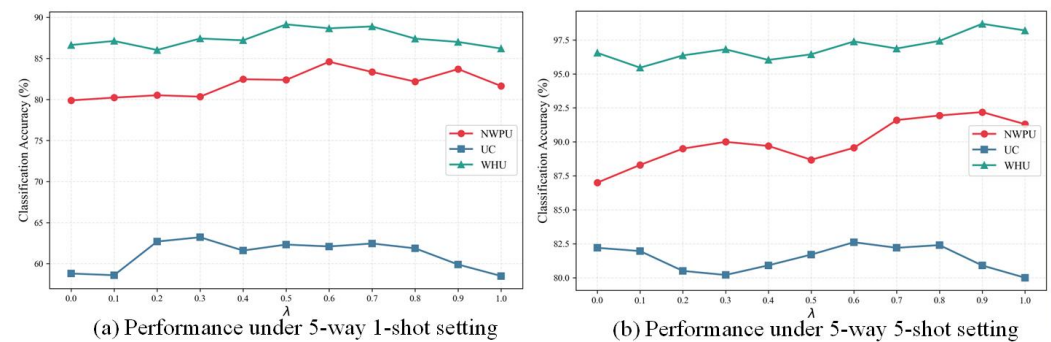


Figure 3. Classification accuracy of the proposed ZSFS-RGBN with different values of λ on three datasets under 5-way 1-shot and 5-way 5-shot settings.

(2) Sensitivity analysis of zigzag sequence length

Table 6 reports the classification accuracies of the proposed ZSFS-RGBN with different zigzag sequence lengths L on the three datasets under 5-way 1-shot and 5-way 5-shot settings. It can be observed that the model performance is sensitive to the choice of L . As L increases from 4 to 16, the accuracy generally improves on all three datasets, indicating that an appropriate zigzag sequence length helps preserve more effective spatial structure information for feature representation and classification. However, when L is further increased to 25, the performance drops consistently, suggesting that excessively long sequences may introduce redundant information and weaken feature discriminability. Among all settings, $L = 16$ achieves the best results on NWPU, UC Merced, and WHU-RS19 under both 1-shot and 5-shot tasks, demonstrating that it provides a better balance between local detail preservation and global structural modeling. Therefore, $L = 16$ is adopted in all formal experiments.

Table 6. Classification accuracy (%) of the proposed ZSFS-RGBN with different zigzag sequence lengths (L) on the three datasets under 5-way 1-shot and 5-way 5-shot settings.

L	NWPU		UC Merced		WHU-RS19	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
4	82.38 $\pm$ 0.51	90.73 $\pm$ 0.43	60.91 $\pm$ 0.77	79.84 $\pm$ 0.58	87.56 $\pm$ 0.49	97.24 $\pm$ 0.36
9	84.07 $\pm$ 0.37	91.52 $\pm$ 0.48	61.88 $\pm$ 0.69	81.17 $\pm$ 0.46	88.47 $\pm$ 0.55	98.03 $\pm$ 0.42
16 (ours)	84.59 $\pm$ 0.39	92.19 $\pm$ 0.47	63.22 $\pm$ 0.72	82.61 $\pm$ 0.50	89.14 $\pm$ 0.47	98.69 $\pm$ 0.37
25	83.66 $\pm$ 0.45	91.31 $\pm$ 0.52	62.04 $\pm$ 0.74	80.96 $\pm$ 0.63	88.29 $\pm$ 0.51	97.81 $\pm$ 0.40

4.6. Computation Efficiency Analysis

To evaluate the computational cost introduced by the proposed method while improving classification performance, we compare the proposed ZSFS-RGBN with several representative few-shot classification methods on the NWPU dataset under the 5-way 1-shot setting. Specifically, the number of parameters, FLOPs, training time, and inference time are reported, as shown in Table 7. It can be observed that MN and PN have relatively simple architectures, with the same parameter size of 12.42 M and FLOPs of 516.2 G and 514.8 G, respectively, resulting in relatively low training and inference costs. Although RN enhances feature interaction capability, its parameter size increases substantially to 23.49 M, accompanied by higher computational complexity and runtime overhead. In contrast, ZSFS-RGBN contains only 12.60 M parameters, which is still much lower than that of RN, while its FLOPs, training time, and inference time reach 742.6 G, 3.06 s, and 0.548 s, respectively. This indicates that the additional cost of ZSFS-RGBN mainly arises from its more sophisticated feature representation and geometric modeling process, rather

than from a significant increase in model size. According to Table 2, under the same setting, ZSFS-RGBN achieves a classification accuracy of 84.59%, which is significantly higher than those of MN, PN, and RN. Overall, although ZSFS-RGBN introduces higher computational complexity and runtime cost, it yields substantial performance gains with a relatively well-controlled parameter size, demonstrating a favorable trade-off between accuracy and complexity.

Table 7. Computational cost and complexity of different methods on NWPU dataset under 5-way 1-shot setting.

Methods	Parameters (M)	FLOPs (G)	Training Times (s/episode)	Inference Times (s/episode)
MN	12.42	516.2	2.04	0.368
PN	12.42	514.8	1.98	0.356
RN	23.49	688.4	2.74	0.492
ZSFS-RGBN (ours)	12.60	742.6	3.06	0.548

4.7. Visualization Analysis

To intuitively demonstrate the discriminative capability of the proposed ZSFS-RGBN method in the few-shot remote sensing scene classification task, we employed t-SNE to project the deep features into a two-dimensional space for visualization. The experiments were conducted on the NWPU-RESISC45 dataset under the 5-way 5-shot setting. Specifically, we selected all 10 categories from the testing set and randomly sampled 100 images from each category for the comparative experiments by feature visualization. The visualization results are shown in Figure 4. Figure 4a shows the original image feature visualization, and Figure 4b shows the features extracted from the baseline. Figure 4c shows the features extracted from ZSFS-RGBN. From Figure 4a, it can be observed that the original image features of different categories are severely mixed; this confirms that low-level pixel information alone is insufficient for distinguishing complex remote sensing scenes. In contrast, Figure 4b shows that the baseline method, which relies on Euclidean space feature learning and global average pooling, achieves a certain degree of clustering. Remarkably, Figure 4c demonstrates that the proposed ZSFS-RGBN method significantly improves feature discriminability. Features belonging to the same category are tightly clustered into compact clusters, and the inter-class margins are substantially enlarged. The t-SNE visualization thus provides strong evidence that ZSFS-RGBN is capable of learning highly discriminative feature representations under few-shot conditions.

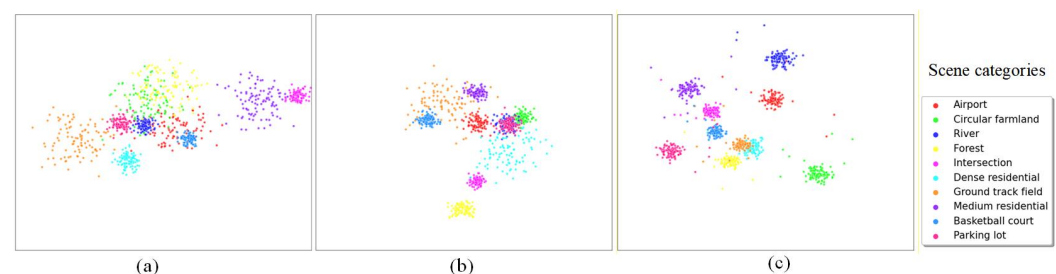


Figure 4. t-SNE visualization of features on the NWPU dataset under 5-way 5-shot settings: (a) original image feature visualization; (b) features extracted from the baseline; (c) features extracted from ZSFS-RGBN.

5. Discussion

The experiment results presented in Section 4 demonstrate that the proposed ZSFS-RGBN method consistently achieves superior performance over a wide range of state-of-

the-art few-shot learning methods on three benchmark remote sensing scene datasets. In this section, we discuss the underlying reasons for these improvements, the implications of our findings, the limitations of the current work, and several promising directions for future research.

The success of ZSFS-RGBN can be attributed to three synergistic innovations that directly address the inherent challenges of few-shot remote sensing scene classification: large intra-class variation, high inter-class similarity, and loss of spatial structure in conventional Euclidean embeddings. First, the zigzag scanning strategy effectively preserves spatial locality and energy concentration. Unlike global average pooling, which discards all spatial information, or row-wise/column-wise flattening, which breaks diagonal continuity, zigzag scanning traverses feature maps along alternating diagonal directions. This maintains the adjacency of spatially close features in the resulting 1D sequence, which is crucial for capturing the anisotropic patterns frequently observed in remote sensing scenes. The t-SNE visualization in Figure 4 confirms that the baseline method, which lacks this strategy, produces overlapping clusters, whereas ZSFS-RGBN yields well-separated and compact clusters. Second, the ARMA-based SPD embedding maps the feature sequence onto a Grassmann manifold, capturing second-order statistics and dynamic dependencies. Many remote sensing scenes are characterized not only by individual objects but also by the spatial arrangement and co-occurrence of textures. The ARMA model, through its state-space formulation, extracts latent spatial dynamics that reflect these structural relationships. By representing each sample as a point on a Riemannian manifold, the method naturally respects the non-Euclidean geometry features. This is a key departure from Euclidean metric learning, which treats all feature dimensions independently and linearly. The ablation study in Table 5 shows that adding RMFEb alone already brings clear gains, confirming the value of manifold embedding. Third, the RieGB jointly optimizes intra-class compactness and inter-class separation directly on the manifold. In few-shot scenarios, class prototypes estimated by simple averaging in Euclidean space are often biased due to the limited number of support samples. The RieGB layer computes the true Riemannian barycenter on the SPD manifold, which is a more robust and geometrically meaningful central tendency. The custom inter-class loss further pushes different class barycenters apart while pulling query samples toward their correct prototype. The ablation results indicate that RieGB provides the most substantial improvement in the 5-shot setting, where more support samples allow the barycenter to be estimated more reliably.

Our method outperforms recent strong baselines such as DiffPR-Net, ICSFF, DBA-RMCL, and ACL-Net across all three datasets. Notably, DiffPR-Net uses diffusion-based data augmentation to generate additional training samples, yet it still lags behind ZSFS-RGBN. This suggests that simply increasing the number of samples is not sufficient. Similarly, ACL-Net employs attention-based contrastive learning in Euclidean space, but it fails to capture the higher-order geometric structure that our Riemannian approach inherently provides.

Despite the strong performance of ZSFS-RGBN, two main limitations remain. First, the method assumes that the feature distributions of base and novel classes are similar, but in real applications, a large domain shift often exists (e.g., different sensors, geographic regions, or imaging conditions). The current Riemannian embedding does not explicitly model such domain discrepancies, which may degrade cross-domain few-shot transferability. Second, the method only processes single-modality optical images, whereas modern remote sensing platforms provide complementary data such as SAR, hyperspectral, and LiDAR. Fusing these heterogeneous modalities could offer richer information, but the current framework lacks a mechanism to integrate multi-modal SPD representations.

To overcome the above limitations, future research will focus on two directions. For cross-domain transferability, we plan to incorporate domain adaptation techniques directly

into the Riemannian manifold, for instance by aligning class-specific SPD barycenters across domains via parallel transport or learning domain-invariant ARMA parameters through adversarial training. For multi-modal fusion, we will extend the framework to handle multiple modalities by constructing a joint SPD representation on a product manifold or learning a shared Riemannian space via bilinear fusion, enabling the model to exploit complementary information from different sensors and further boost few-shot classification accuracy in challenging real-world scenarios.

6. Conclusions

This paper proposes a novel few-shot remote sensing scene classification method called ZSFS-RGBN, which combines zigzag scanning feature serialization and Riemannian geometric barycenter learning to effectively preserve spatial structural information and model the non-Euclidean geometric properties of remote sensing features. By optimizing intra-class compactness and inter-class separability on SPD manifolds via a Riemannian barycenter network, the proposed approach alleviates the instability of prototype estimation caused by large intra-class variations and high inter-class similarity in remote sensing scenes. Extensive experiments on three public datasets demonstrate that ZSFS-RGBN consistently outperforms state-of-the-art methods under both 5-way 1-shot and 5-way 5-shot settings, and ablation and visualization results verify the effectiveness of each key component. This work provides a new and effective paradigm for few-shot remote sensing scene classification with strong practical value for real-world applications with limited annotation resources.

Author Contributions: Conceptualization, L.L. (Longwei Li) and X.C.; methodology, X.C.; software, Z.W.; validation, M.L., X.L. and X.C.; formal analysis, X.C.; investigation, X.C.; resources, Y.C.; data curation, X.C.; writing—original draft preparation, X.C.; writing—review and editing, G.Z.; visualization, L.L. (Lei Liu); supervision, Y.C.; project administration, X.C.; funding acquisition, X.C. and L.L. (Lei Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by Open Foundation of Shiyan Key Laboratory of Air-Ground Crowd Cooperation Technology and Application grant number SYZDK32025A02, in part by the Natural Science Foundation of Hubei Province of China under grant number 2026AFC0344, in part by Educational Commission of Hubei Province of China grant number Q20241805, in part by the National Natural Science Foundation of China grant number 12403083, in part by the Open Foundation of Anhui Province Key Laboratory of Realistic Geographic Environment grant number 2024PGE004, in part by the Open Foundation of Key Laboratory of Automotive Power Train and Electronic Control(Hubei University of Automotive Technology) grant number ZDK12024B02, in part by the Open Foundation of Startup Foundation of Doctor scientific research from Hubei University of Automotive Technology grant number BK202487.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to express our gratitude to the editor and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756.

2. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883.
3. Zhu, Q.; Zhong, Y.; Zhang, L.; Li, D. Scene classification based on the fully sparse semantic topic model. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5525–5538.
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
5. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987.
6. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
7. Yang, Y.; Newsam, S. 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, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
12. Yang, B.; Chen, Y.; Ghamisi, P. LVM-StARS: Large vision model soft adaption for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
13. Wu, X.; Ma, Z.; Cui, F.Q.; He, H.; Hou, Y.; Yang, J. Dual-branch CNN with enhanced bidirectional multiscale sparse cross-attention for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5617014.
14. Zhao, B.; Zhong, Y.; Zhang, L. A spectral-spatial attention network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1990–2001.
15. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087.
16. Xue, W.; Ai, J.; Zhu, Y.; Sun, X.; Zhang, Y.; Gao, G. LMCNet: Lightweight Modality Compensation Network Via Knowledge Distillation for Salient Ship Detection Under Missing-Modality Conditions. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 6547–6560.
17. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the International Conference on Machine Learning Deep Learning Workshop, Lille, France, 10–11 July 2015.
18. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; Kavukcuoglu, K. Matching networks for one shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3630–3638.
19. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
20. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.S.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
21. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
22. Cheng, G.; Cai, L.; Lang, C.; Yao, X.; Chen, J.; Guo, L.; Han, J. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5608011.
23. Tian, F.; Lei, S.; Zhou, Y.; Cheng, J.; Liang, G.; Zou, Z.; Li, H.C.; Shi, Z. HiReNet: Hierarchical-relation network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5603710.
24. Regan, J.; Khodayar, M. FGRMNet: Fully graph relational matching network for few-shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12.
25. Qin, A.; Luo, B.; Li, Q.; Zou, C.; Zhao, Y.; Song, T.; Gao, C. Local-descriptors-based rectification network for few-shot remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9566–9581.
26. Jia, Y.; Sun, C.; Gao, J.; Wang, Q. Few-shot remote sensing scene classification via parameter-free attention and region matching. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 265–275.
27. Xu, Y.; Bi, H.; Yu, H.; Lu, W.; Li, P.; Li, X.; Sun, X. Attention-based contrastive learning for few-shot remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620317.
28. Ai, J.; Mao, Y.; Luo, Q.; Jia, L.; Xing, M. SAR Target Classification Using the Multikernel-Size Feature Fusion-Based Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13.

29. Tuzel, O.; Porikli, F.; Meer, P. Region covariance: A fast descriptor for detection and classification. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 589–600.
30. Moakher, M.; Batchelor, P.G. Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*; Springer: Berlin, Germany, 2006; pp. 285–298.
31. Huang, Z.; Van Gool, L. A Riemannian network for SPD matrix learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 2036–2042.
32. Gao, Z.; Wu, Y.; Bu, X.; Yu, T.; Yuan, J.; Jia, Y. Learning a robust representation via a deep network on symmetric positive definite manifolds. *Pattern Recognit.* **2019**, *92*, 1–12.
33. Cheng, G.; Han, J.; Zhou, P. Learning low-level and mid-level features for remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1813–1817.
34. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325.
35. Penatti, O.A.B.; Nogueira, K.; dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 44–51.
36. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4775–4784.
37. Liu, L.; Yu, Y.; Yang, X.; Gao, L. Multi-scale CNN for remote sensing image scene classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018.
38. Lu, X.; Gong, T.; Zheng, X. Multisource compensation network for cross-domain scene classification. *IEEE Trans. Geosci. Remote Sens.* **2020**.
39. Wei, Y.; Liu, Y.; Niu, B.; Xiang, X.; Lin, J.; Hu, Y.; Wu, Y. A memory-efficient class-incremental learning framework for remote sensing scene classification via feature replay. *Remote Sens.* **2026**, *18*, 896.
40. He, N.; Fang, L.; Li, S.; Plaza, A.; Plaza, J. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6899–6910.
41. Wei, J.; Li, E.; Zhang, C. Cross-layer feature fusion and attention-based class feature alignment network for unsupervised cross-domain remote sensing scene classification. *Remote Sens.* **2026**, *18*, 859.
42. Wu, N.; Lv, J.; Jin, W. S4Former: A spectral–spatial sparse selection transformer for multispectral remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5001605.
43. Xu, K.; Zhu, Z.; Wang, W.; Fan, C.; Wu, B.; Jia, Z. Enhancing remote sensing scene classification with Hy-MSDA: A hybrid CNN–transformer for multisource domain adaptation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5600715.
44. Zhai, M.; Liu, H.; Sun, F. Lifelong learning for scene recognition in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1472–1476.
45. Zhang, Z.Y.; Xu, C.J. A Lie group-based model for remote scene classification with multi-scale feature fusion and mixed attention mechanisms. *Int. J. Remote Sens.* **2025**, *46*, 3800–3830.
46. Harandi, M.T.; Sanderson, C.; Hartley, R.; Lovell, B.C. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 216–229.
47. Liu, H.; Li, J.; Wu, Y.; Ji, R. Learning neural bag-of-matrix-summarization with Riemannian network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
48. Brooks, D.A.; Schwander, O.; Barbaresco, F.; Schneider, J.Y.; Cord, M. Riemannian batch normalization for SPD neural networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
49. Chakraborty, R.; Bouza, J.; Manton, J.H.; Vemuri, B.C. ManifoldNet: A deep neural network for manifold-valued data with applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 799–810.
50. Xu, C.; Zhu, G.; Shu, J. A lightweight intrinsic mean for remote sensing classification with Lie group kernel function. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1741–1745.
51. Xu, C.; Zhu, G.; Shu, J. A lightweight and robust Lie group-convolutional neural networks joint representation for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5501415.
52. Cheng, D.; Dong, Z.; Wang, J.; Yang, X. A dual-scale siamese densely connected network with MRF for SAR image classification. *Remote Sens. Lett.* **2022**, *13*, 247–257.
53. Li, E.; Samat, A.; Zhang, C.; Du, P.; Liu, W. First and second-order information fusion networks for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8014405.
54. Wu, L.; Lu, M.; Fang, L. Deep covariance alignment for domain adaptive remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620811.

55. Cui, Z.; Yang, W.; Chen, L.; Li, H. MKN: Metakernel networks for few shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4705611.
56. Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; Tao, C. RS-MetaNet: Deep metametric learning for few-shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6983–6994.
57. Li, L.; Han, J.; Yao, X.; Cheng, G.; Guo, L. DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7844–7853.
58. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412.
59. Mangla, P.; Singh, M.; Sinha, A.; Kumari, N.; Balasubramanian, V.N.; Krishnamurthy, B. Charting the Right Manifold: Manifold Mixup for Few-shot Learning. In Proceedings of the Workshop on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020.
60. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv* **2017**, arXiv:1707.09835.
61. Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S.J.; Yang, Y. Learning to propagate labels: Transductive propagation network for few-shot learning. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019, pp. 1–14.
62. Zhang, C.; Cai, Y.; Lin, G.; Shen, C. DeepEMD: Differentiable Earth Mover’s Distance for Few-Shot Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5632–5648.
63. Zeng, Q.; Geng, J.; Jiang, W.; Huang, K.; Wang, Z. IDLN: Iterative Distribution Learning Network for Few-Shot Remote Sensing Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8020505.
64. Yuan, Z.; Huang, W.; Tang, C.; Yang, A.; Luo, X. Graph-Based Embedding Smoothing Network for Few-Shot Scene Classification of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1161.
65. Chen, W.Y.; Liu, Y.C.; Kira, Z.; Wang, Y.; Huang, J.B. A Closer Look at Few-shot Classification. *arXiv* **2019**, arXiv:1904.04232.
66. Li, X.; Shi, D.; Diao, X.; Xu, H. SCL-MLNet: Boosting Few-Shot Remote Sensing Scene Classification via Self-Supervised Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5801112.
67. Huang, W.; Yuan, Z.; Yang, A.; Tang, C.; Luo, X. TAE-Net: Task-Adaptive Embedding Network for Few-Shot Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 111.
68. Li, W.; Xu, J.; Huo, J.; Wang, L.; Gao, Y.; Luo, J. Distribution Consistency Based Covariance Metric Networks for Few-Shot Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8642–8649.
69. Hu, S.X.; Moreno, P.G.; Xiao, Y.; Shen, X.; Obozinski, G.; Lawrence, N.D.; Damianou, A. Empirical Bayes Transductive Meta-Learning with Synthetic Gradients. *arXiv* **2020**, arXiv:2004.12696.
70. Zeng, Q.; Geng, J. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 143–154.
71. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Cross Attention Network for Few-shot Classification. *arXiv* **2019**, arXiv:1910.07677.
72. Qin, A.; Chen, F.; Li, Q.; Song, T.; Zhao, Y.; Gao, C. Few-Shot Remote Sensing Scene Classification via Subspace Based on Multiscale Feature Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13292–13307.
73. Zeng, Q.; Geng, J.; Huang, K.; Jiang, W.; Guo, J. Prototype Calibration with Feature Generation for Few-Shot Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 2728.
74. Chen, Y.; Li, Y.; Mao, H.; Chai, X.; Jiao, L. A Novel Deep Nearest Neighbor Neural Network for Few-Shot Remote Sensing Image Scene Classification. *Remote Sens.* **2023**, *15*, 666.
75. Cheng, K.; Yang, C.; Fan, Z.; Wu, D.; Guan, N. TeAw: Text-Aware Few-Shot Remote Sensing Image Scene Classification. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
76. Yang, C.; Liu, T.; Chen, G.; Li, W. ICSFF: Information Constraint on Self-Supervised Feature Fusion for Few-Shot Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5800312.
77. Dong, R.; Xia, J.; Jiao, L.; Zhao, J.; Hong, L.; Yun, Z. DBA-RMCL: Refined Metric Contrastive Learning With Dual-Branch Attention for Few-Shot Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5623016.
78. Zhu, Y.; Han, J.; Pan, B.; Shi, Z. DiffPR-Net: Few-Shot Remote Sensing Scene Classification Based on Generative Diffusion and Prototype Rectified Model. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5606213.
79. Ma, J.; Lin, W.; Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Multipretext-task prototypes guided dynamic contrastive learning network for few-shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5614216.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.