IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Few-Shot Remote Sensing Scene Recognition via Feature Enhancing Learning

YE RONG[1,2], QING-YI KONG[2,3], (IEEE Member), AND GUANG-LONG WANG[1]
[1]Army Engineering University of PLA, Shijiazhuang 050005, China
[2]Hebei jiaotong vocational and technical college,Shijiazhuang 050035, China
[3]Hebei Kingston Technology Co., Ltd.,Xinji 052360, China

Corresponding authors: Guang-Long Wang (glwang2005@163.com) and Qing-Yi Kong (402574328@qq.com)

This work was supported by the Hebei Light Electric Special Vehicle Technology Innovation Center for providing technical support and the Hebei Kingston Technology Co., Ltd. Also, this research was supported by the Xinji Science and Technology Plan Project (No. 202104), the Science and Technology Research Project of Colleges and Universities in Hebei Province (No. ZC2021261), and the Science and Technology Research Project of Hebei Universities (No. ZC2022012).

ABSTRACT Inspired by the human brain's efficient knowledge assimilation, few-shot learning presents significant promise in machine learning and artificial intelligence (AI). The field aims to create models that classify and recognize new categories using a few labeled samples. However, existing models' capacity to utilize knowledge from benchmark datasets requires enhancement. In response, our research introduces an innovative model for few-shot scene recognition in optical remote-sensing imagery. This model incorporates a cross-attention mechanism to improve feature extraction with limited data. Additionally, it employs triplet angular loss to introduce geometric constraints in the feature space. Our method also combines non-linear projection transformations with fine-grained, instance-level classification. This enhances separation between classes and improves prediction accuracy. We have evaluated our model using three renowned datasets for few-shot remote sensing imagery: Northwestern Polytechnical University (NWPU), the Aerial Image Dataset (AID), and the University of California, Merced (UCM). The outcomes affirm our method's capability to feature across diverse remote sensing scene categories. Our approach represents advancing precision and efficacy of detecting category-specific features in few-shot learning contexts.

INDEX TERMS Few-shot learning, remote sensing imagery, feature enhancement learning, contrastive learning.

I. INTRODUCTION

Remote sensing is a technology that gathers information about objects without direct contact, a key component of which is optical remote sensing imagery. These images, captured through remote sensing techniques, visually depict the Earth's surface and environmental data. They are known for their extensive coverage, quick acquisition, real-time solid capabilities, and comprehensive informational content. This makes them extremely useful in resource exploration [1], [2], environmental monitoring [3], meteorological forecasting [4], and military surveillance [5]. Nevertheless, accurately classifying and recognizing remote-sensing images presents complex challenges, especially when dealing with ecological variations. Challenges such as "spectral confusion," where different materials generate similar spectral signatures, and "spectral variability," where the same material produces different spectral responses under various conditions, arise [6], [7]. Coupled with the limited availability of labeled training samples, these factors pose significant challenges to data collection and constrain cost-effectiveness. Given their reliance on abundant training samples, traditional methods often underperform when faced with such complexities.

Inspired by the rapid learning ability of the human brain, the development of few-shot learning technologies has advanced quickly. These technologies may solve the classification challenges of remote sensing imagery. Few-shot learning [8] involves classifying and recognizing new categories using a small number of labeled samples by effectively applying knowledge from a more extensive set of labeled samples. It emulates the human ability to generalize from a small number of examples. Applying few-shot learning to remote sensing offers a promising approach to addressing the scarcity of labeled training data and enhancing classification for underrepresented or entirely new classes.

However, the few-shot classification of remote sensing images faces a series of challenges and limitations [9]. First and foremost, remote sensing data is inherently complex, with attributes such as high dimensionality, spectral variability, and class imbalance. Moreover, because of the scarcity of samples, models are susceptible to overfitting, which makes it challenging to generalize to new, unseen examples. Furthermore, the restricted availability of training data may impede the model's capacity to capture intricate terrestrial features. Overcoming these obstacles is essential for effectively harnessing the potential of few-shot learning. Furthermore, transfer learning, semi-supervised learning, and other advanced techniques also offer effective strategies to address these issues.

This paper presents a few-shot remote-sensing scene recognition method using feature enhancement learning (FEL). The method tackles the challenges presented by a scarcity of labeled training data and intricate remote sensing imagery. By integrating feature enhancement techniques with the few-shot learning paradigm, we aim to explore further and develop new methods designed explicitly for few-shot remote sensing scene recognition. This approach seeks to improve feature extraction's quality and discriminative power, even with limited samples, enhancing image recognition's performance and generalization ability. This will contribute to a more accurate interpretation and utilization of remote sensing image data, resulting in more reliable classification results for practical resource management and environmental monitoring applications. Furthermore, we conducted comprehensive experiments on benchmark remote sensing datasets to evaluate the effectiveness of our proposed method, demonstrating its potential to push the boundaries of few-shot remote sensing scene recognition.

II. RELATED WORK

Here, we offer background information on the forthcoming study topic.

A. META-LEARNING

Meta-learning, or "learning to learn," refers to a technique that enables machines to leverage previously acquired knowledge or models to generalize new knowledge and tackle a wide range of novel tasks that display significant differences. We typically classify approaches in this field into metric-based, model-based, and optimization-based methods.

Metric-based meta-learning shares similarities with the k-nearest neighbors algorithm, with its primary goal being to calculate distances between samples efficiently. In addition to classic network models, Koch et al. proposed Siamese networks [10]; Vinyals et al. introduced Matching Networks (MatchingNet) [11]; Snell et al. developed Prototypical Networks (ProtoNet) [12]; and Sung et al. suggested Relation Networks (RelationNet) [13]. The field also includes several notable developments. Induction Networks (InductionNet) [14] represent a novel advancement in the field of metric-based meta-learning. These networks focus on identifying the optimal matching image regions to enhance meta-learning performance. The FD-DAML networks [15] are designed to address challenges related to variations in domain distribution, label mismatches, and inadequately labeled samples to enhance the model's ability to generalize. The DCMLN model [16] utilizes a meta-testing approach that combines gradients and metrics to improve the performance of meta-learning. These recent advancements enhance the field of metric-based meta-learning, offering additional opportunities to tackle complex problems. They provide new insights and techniques in distance measurement and sample matching, essential for enhancing meta-learning's performance and adaptability.

Model-based meta-learning is a critical technology that aims to achieve rapid learning by adapting model architecture and parameters. New models and methods have emerged beyond the traditional convolutional neural networks (CNNs) [17], recurrent neural networks (RNNs) [18], and extended short-term memory networks (LSTMs) [19], leading to significant advancements in meta-learning. These innovative models and methods encompass memory-augmented neural networks (MANNs) [20], which incorporate the neural Turing machine concept to enhance the model's adaptability across diverse tasks, facilitating quicker learning and improved generalization. Metanets (MetaNet) [21] primarily focus on cross-task meta-learning, allowing models to adapt to various tasks rapidly. The Simple Neural Attentive Meta-Learner (SNAIL) [22] utilizes temporal convolutions and soft attention mechanisms to provide an efficient learning approach for meta-learning. Moreover, notable emerging achievements include XDNet [20], MedOptNet [23], FedMeta-FFD [24], STDP-PNN [25], and DIFF-WRN [26]. These contributions introduce new possibilities for advancing model-based meta-learning, broadening the scope of its applications, and enhancing model performance in multi-task learning and rapid adaptability.

Optimization-based meta-learning is a crucial approach that utilizes meta-learning frameworks to determine the optimal gradient descent directions and achieve parameter optimization using limited samples. Several classic models and methods have succeeded remarkably, including model-agnostic meta-learning (MAML) [27]. This quintessential approach enables rapid optimization of models for few-shot tasks and quick adaptation to new tasks. It stands as a prototypical example of a meta-learning architecture. LSTM-Based Meta-Optimizer [28]: Ravi and his colleagues proposed a meta-optimizer based on Long Short-Term Memory Networks (LSTM). This approach aims to learn the initial state of optimizers to facilitate rapid optimization on new tasks.

Furthermore, there have been recent research breakthroughs, including MAML-SR [29], Proto-MAML [30] (which integrates the principles of prototypical networks), and MAML-PFL [31], among others. These advancements have expanded the range of applications for optimization-based meta-learning methods, offering more solutions for addressing the challenges of few-shot parameter optimization and making significant contributions to the development of the meta-learning field. These methods are crucial for improving parameter tuning and optimization with limited data.

Few-shot learning is a subfield of meta-learning within the supervised learning domain that focuses on addressing the challenge of learning and generalizing effectively when training samples are minimal. This article focuses on few-shot learning as the core technology, aiming to address the common issue of limited sample sizes encountered in g imagery.

B. FEW-SHOT REMOTE SENSING IMAGE CLASSIFICATION

Remote sensing image classification encounters challenges such as limited raw data, difficulties annotating datasets, and various sensor characteristic limitations, all contributing to a shortage of training sample data. In recent years, numerous researchers have dedicated their efforts to studying few-shot remote sensing image classification and have proposed several effective methods. Data augmentation, which involves distortions like random cropping, rotation, flipping, and noise, is an accessible method for increasing the number of samples and improving models' generalization ability and classification performance [32], [33]. However, this approach does not address the fundamental issue of limited samples. Modeling and simulating remote sensing imaging for data augmentation [33] can rapidly produce images, yet these synthetic images often show notable differences from natural images. Deep generative models employ sample synthesis, sample transfer, and multimodal techniques for data augmentation [34], [35]. They automatically generate samples without manual design, learn richer intrinsic features, and improve the similarity of the distribution to the original data. However, improving the fidelity of images with complex background information is still necessary.

Transfer learning techniques [9] leverage knowledge from a source domain to improve few-shot remote sensing image classification tasks. These techniques involve transferring models and features. This approach can reduce the problem of insufficient samples, but it requires careful attention to acquiring source data. Metric learning [49] techniques that aim to find an optimal metric space are relatively straightforward and do not require fine-tuning. However, their ability to generalize to the diversity within remote sensing images and scale variations needs further research, as demonstrated by SCL-MLNet [37], [38]. Meta-learning approaches involve learning from multiple few-shot examples and accumulating general knowledge to enhance generalization across different tasks. They exhibit rapid adaptability and can train on various assignments simultaneously. However, a diverse dataset with multiple classes is a prerequisite. Designing lightweight models [24] can decrease the number of model parameters and the risk of overfitting. Self-learning and self-training techniques [42] enable the exploration of unlabeled data, enhancing learning from limited datasets. We can use these methods individually or combine them to improve the few-shot remote sensing image classification performance. The specific choice depends on the nature of the problem and the available data.

In summary, few-shot remote sensing image classification is a challenging task that requires various methods to achieve adequate resolution. Although researchers have proposed many practical approaches in recent years, there remains significant room for improvement in current models when tested on public benchmark datasets, and they still fall short of practical, real-world applications. Further enhancements and refinements are necessary. Moving forward, we need to explore further the issues and solutions related to few-shot remote sensing image classification to support the application and advancement of remote sensing imagery.

C. ATTENTION MECHANISM

Experts first introduced attention mechanisms in visual imaging in the 1990s. Later, these mechanisms merged closely with deep learning. Image processing extensively employs attention mechanisms. The Recurrent Neural Network Model RNN [43] was the first to incorporate spatial attention, using RNNs and reinforcement learning for its implementation. CNNs [44] combine explicit translational invariance and implicit rotational invariance within neural networks. Spatial Transformer Networks (STN) [45] can dynamically perform spatial transformations and align data.

The field subsequently entered an era that actively utilized channel attention, adaptively recalibrating channels through attention weights. The prominent methods include squeeze-and-excitation networks (SENet) [46] and the Convolutional Block Attention Module (CBAM) [47]. SENet models the relationships between channels to determine their weights. CBAM combines spatial and channel attention. This method segments the feature map and assigns weights to each region, enhancing the influence of individual channels. Average Pooling (GAP) [48] combines global spatial details with channel data to assign weights to feature maps across spatial and channel dimensions. Meanwhile, SPP-Net [49] applied Spatial Pyramid Pooling (SPP) or global average pooling to extract spatial information, increasing the focus on crucial regions and features of the feature map.

Finally, in 2017, incorporating attention mechanisms in Natural Language Processing (NLP) [50] marked the beginning of the self-attention era. Proposers Multi-Head Self-Attention (MHSA) [50] to integrate multi-scale features, enhancing the model's generalization capabilities. In 2020, the introduction of the Vision Transformer (ViT) [51] to the field of computer vision marked a significant milestone. Following this development, scholars delved deeply into transfer learning between different tasks, scaling of model sizes, and systemic analysis, leading to further substantial breakthroughs.

This article leverages the characteristics of few-shot learning and introduces an innovative attention mechanism architecture. The goal is to reveal the potential relationships between support and query sets in few-shot tasks. This approach enhances the quality of feature extraction in scenarios with limited samples.

D. CONTRASTIVE LEARNING

In recent years, deep learning-based contrastive learning techniques have undergone rapid development. Methods such as SimCLR [52] (self-supervised contrastive learning), MoCo [53] (momentum contrast), BYOL [54] (bootstrap your own latent), SwAV [55] (unsupervised clustering), and SimSiam [56] (simple Siamese networks) have emerged as exemplary contrastive learning algorithms. These methods typically use twin-like neural network architectures and, during training, compare either positive pairs (different augmentations of the same image) or antagonistic pairs (augmentations of other photos). Neural network models use deep contrastive learning to automatically improve feature representations by comparing numerous positive and negative pairs. CPC [57] (Contrastive Predictive Coding) has established the baseline for deep contrastive learning. Contrastive Predictive Coding (CPC) aims to maximize the alignment between predicted and actual outcomes in sequential data. It enhances the feature extraction network and incorporates the InfoNCE loss, which has become a standard in contrastive learning research.

Khosla et al. [58] expanded the concept of contrastive learning to supervised learning and introduced supervised contrastive learning loss (SCL loss). The goal was to leverage labeled data to improve the model's feature representation capabilities. Chen et al. [52] developed a semi-supervised contrastive learning algorithm that first undergoes pretraining on all data through contrastive learning. The pre-trained model's knowledge is then transferred to a new model using labeled data through distillation learning. Contrastive learning actively improves the quality of data feature extraction in unsupervised learning scenarios. It creates proxy tasks for backbone networks to handle unlabeled data.

Numerous scholars have conducted in-depth studies combining self-supervised learning with contrastive learning to devise innovative strategies that enhance the model's ability to learn image representations. These approaches have achieved outstanding performance across multiple benchmark datasets, exemplified by MoCo [53], SimCLR [52], BYOL [54], SwAV [55], and SimSiam [56] methodologies. By integrating contrastive learning with knowledge distillation, they have managed to match the performance of models trained on much larger datasets, with DINO [59] being a typical example of such research.

Furthermore, Cui introduced a versatile parametric method that allows the model to learn visual representations and task-specific parameters within a unified framework, thereby improving the model's generalizability across various tasks [60]. Yin proposed carefully selecting sample pairs for clustering objectives, which, when combined with contrastive learning, has enhanced clustering performance [61].

This paper builds on these advancements by presenting a triplet angular loss constraint. This constraint aim to shrink distances between positive samples and expand distances between negative ones. As a result, this constraint enhances the discriminative ability of the features extracted by the backbone network from the feature space.

III. MATERIALS AND METHODS

In this section, we outline our proposed method, which includes subsections on introducing notation, the complete objective, the rotation prediction task, the contrastive prediction task, and network regularization.

A. NOTATION AND PRELIMINARY

In this section, we define and explain the few-shot classification problem using relevant notation. In few-shot classification, we often work with a large amount of labeled data D_{Base} and their respective classes C_{Base}. We aim to train a model to generalize to new, unlabeled instances D_{Novel} from entirely new classes C_{Novel}. It is important to note that the sets of base classes C_{Base} and novel classes C_{Novel} are disjoint, i.e., $C_{Base} \cap C_{Novel} = \phi$.

Few-shot learning addresses the challenge of having limited samples for the novel classes. Due to the need for large amounts of data and labels in deep learning models, training directly on new class categories is not feasible. Therefore, we leverage the data from the base classes to facilitate learning. Given the scarcity of samples, it is necessary to construct numerous N-way K-shot Q-query sets for training using the base class data. An N-way K-shot Q-query set comprises N distinct classes, and each class contains K support samples and Q query samples for training and evaluation, respectively.

These NK support samples comprise a support set $S = \{X_i, Y_i\}_{i=1}^{NK}$, while the query set $Q = \{X_i\}_{i=1}^{NQ}$ consists of NQ unlabeled query samples. The few-shot classification task aims to utilize the NK support samples to recognize the unlabeled NQ data within the query set.

B. OVERVIEW OF THE FRAMEWORK

We propose a feature-augmentation learning-based framework for few-shot remote sensing scene recognition, shown in Fig. 1. The model primarily consists of three components.

1) FEATRUE EMBEDDING MODULE

The backbone network (such as Conv-4 or ResNet-12) processes the original remote sensing image data to extract initial, shallow representations. Then, a series of stacked cross-attention modules process the features of the support and query sets, refining the features for both sets.

2) CONTRASTIVE EMBEDDING MODULE

This module further refines the features of the support and query sets within the feature space by clustering similar features closer together and pushing dissimilar features apart. It employs a triplet angular loss to implement geometric spatial constraints.

3) INSTANCE-LEVEL CLASSIFICATION MODULE

This module first applies a nonlinear projection transformation to the refined features of the support and query sets. Then, it conducts fine-grained classification at the instance level. This process enhances the model's capacity to distinguish

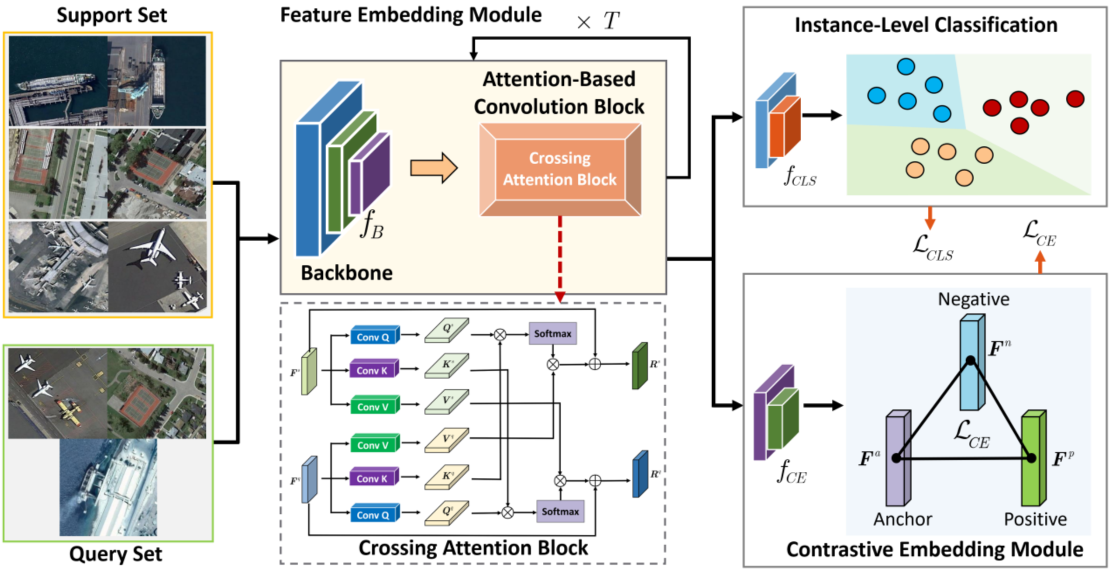


FIGURE 1. **The Proposed Model Framework Diagram.**

between different classes in various episodic tasks. During testing, one can directly utilize the instance-level classification module to infer and predict the samples in the query set.

C. FEATURE EMBEDDING MODULE

We input the remote sensing image data into a backbone network, denoted as $f_B(\cdot)$, to obtain shallow representations of both the support and query set images in Equation (1).

$$\begin{aligned} F^s &= f_B(X^s), \\ F^q &= f_B(X^q), \end{aligned} \tag{1}$$

Here, X^s and X^q represent the original remote sensing image data from the support and query sets.

In the t-th cross-attention module, the input is the output of the $(t-1)$-th cross-attention module. We can express this module mathematically in Equations (2) and (3).

$$\begin{aligned} R^s[t] = \text{softmax}&\left(\frac{f_Q(R^s[t-1]) \otimes f_K(R^q[t-1])}{\sqrt{d_s}}\right) \\ &\otimes f_v(R^q[t-1]) \oplus R^s[t-1] \end{aligned} \tag{2}$$

$$\begin{aligned} R^q[t] = \text{softmax}&\left(\frac{f_Q(R^q[t-1]) \otimes f_K(R^s[t-1])}{\sqrt{d_q}}\right) \\ &\otimes f_v(R^s[t-1]) \oplus R^q[t-1] \end{aligned} \tag{3}$$

In this context, d_s and d_q represent the sizes of the last-order dimensions of the feature sets for the support and query sets, respectively. Notably, when $t = 1$, $R^s[0] = F^s$, $R^q[0] = F^q$. As can be seen from the above formulation, during the process of refining the support set features $R^s[t]$, guidance is explicitly drawn from the previous moment's query set features $R^q[t-1]$. Similarly, the query set features $R^q[t]$ explicitly incorporate information from the support set features $R^s[t-1]$. The cross-attention module we propose explicitly allows for interaction between the support and query set features, enabling the unlabeled query set samples to move closer in the feature space to their similar classes.

Furthermore, the module utilizes residual connections to preserve the features from the previous moment, effectively retaining valuable features and eliminating redundant ones. This approach also tackles the vanishing gradient problem that may arise when stacking multiple layers of cross-attention modules.

D. COMPARE EMBEDDED MODULES

The support and query set samples undergo significant feature refinement within the feature embedding module. Since we treat each episodic task as wholly independent and identically distributed, we anticipate that our model will learn more abstract and advanced general features. Drawing inspiration from the concept of contrastive learning in self-supervised learning, which has achieved notable success in unsupervised learning, we propose a contrastive embedding module. This module imposes additional constraints on the feature embedding module to enhance the expressivity of features across different classes. Firstly, the refined features $R^s[T]$ of the support set undergo a nonlinear transformation through $f_{CE}(\cdot)$, expressed in Equation (4).

$$Z^s = f_{CE}(R^s[T]) \tag{4}$$

Traditional contrastive loss establishes fixed boundaries between classes. The loss function generates negative gradients that may not always aim in the optimal direction. This means they don't guarantee pushing negative samples away from the center of the positive samples' class. To address these issues, we employ the geometric space constraint of angular loss, which incorporates angular constraints to achieve this objective. For any given episodic task, we organize the support set features Z^s into several triplets $\langle F^a, F^p, F^n \rangle$ using label information. The anchor feature F^a and the positive sample feature F^p have the same label in each triplet. However, the negative sample feature F^n belongs to a different category. You can calculate the angular loss from the support set features Z^s with the formula provided in Equation (5).

$$\mathcal{L}_{CE} = \frac{1}{N \times (N-1)} \sum_{n=1}^{N} log \left\{ 1 + \sum_{m=1, m \neq n}^{N} e^{d\left\langle F_n^a, F_n^p, F_m^n \right\rangle} \right\} \tag{5}$$

Where $d \left\langle F_n^a, F_n^p, F_m^n \right\rangle$ represents the loss for an individual triplet, calculated in Equation (6).

$$d \left\langle F_n^a, F_n^p, F_m^n \right\rangle = \left| \left(F_n^a - F_n^p\right)^2 - 4tan^2\alpha \left(F_m^n - \frac{1}{2} \left(F_n^a - F_n^p\right)^2 \right)^2 \right|_+ \tag{6}$$

In this context, $|\cdot|_+$ means we take the value as 0 if it is less than 0, also known as the ReLU function or rectified linear activation. The symbol α stands for an angular hyperparameter. It represents the angle formed with the anchor feature F^a, with the positive F^p and negative F^n sample features as its sides. This angle represents the degree of geometric constraint required between the positive and negative samples.

E. INSTANCE LEVEL CLASSIFICATION MODULE

Our ultimate goal is to enable the model to acquire high-quality, distinctive features. So, we need to implement a classification module. This module aims to enhance the feature embedding module's capability to distinguish between categories in various episodic tasks. Like the contrastive embedding module, the classification module starts with a nonlinear transformation network $f_{CLS}(\cdot)$. We aim to protect the general feature expression ability of the upstream feature embedding module from any influence from downstream tasks. In simpler terms, the feature embedding module's ability to express should remain unaffected by the classification and contrastive learning tasks, regardless of changes in episodic tasks. We apply nonlinear projection transformations to the refined support and query set features in Equation (7).

$$\begin{aligned} H^S &= f_{CE}\left(R^s[T]\right), \\ H^q &= f_{CE}\left(R^q[T]\right), \end{aligned} \tag{7}$$

Subsequently, we utilize the commonly used cosine similarity to measure the distance between the features of the support set and query sets' features. This approach effectively mitigates the impact of varying scales. It can also differentiate cases where samples are close regarding Euclidean distance but belong to different categories based on their angular separation. We typically use the cross-entropy loss function in Equation (8) for the classification loss.

$$\mathcal{L}_{CLS} = \frac{1}{NQ} \sum_{i=1}^{NQ} \sum_{n=1}^{N} I\left[y_i = n\right] \cdot \frac{e^{\frac{1}{\sigma} cos\left\langle H_i^q, P_n^s \right\rangle}}{\sum_{l=1}^{N} e^{\frac{1}{\sigma} cos\left\langle H_i^q, P_l^s \right\rangle}} \tag{8}$$

Wherein $I\left[y_i = n\right]$ denotes that for the (i)-th query set sample, if the label is n, then the value is 1, otherwise it is 0; σ represents a scale factor used to control the smoothness after category normalization; and p signifies the class prototype within the support set, which the Equation (9) calculation determines.

$$P_n^s = \frac{1}{K} \sum_{i=1}^{NK} I\left[y_i = n\right] \cdot H_i^s \tag{9}$$

The variable $cos\left\langle H_i^q, P_n^s \right\rangle$ represents the cosine similarity between the i-th query set sample H_i^q and the n-th category prototype P_n^s from the support set in Equation (10).

$$cos\left\langle H_i^q, P_n^s \right\rangle = \frac{H_i^q \cdot P_n^s}{\left\|H_i^q\right\|_2 \cdot \left\|P_n^s\right\|_2} \tag{10}$$

Wherein $\|\cdot\|_2$ denotes the L2 norm, which is the Euclidean norm used to calculate the magnitude of a vector.

F. MODEL REASONING AND TRAINING

We employ an instance-level classification module to categorize samples from the query set directly. We can symbolically represent the inferred category for the i-th query sample in Equation (11).

$$\hat{y} = \underset{y \in [1,N]}{argmaxcos} \left\langle H_i^q, P_y^s \right\rangle \tag{11}$$

During training, we combine contrastive learning loss and classification loss. Then, we take a weighted sum to serve as the global objective function for training our proposed model in Equation (12).

$$l = \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{CLS} \cdot \mathcal{L}_{CLS} + \lambda_\theta \cdot \sum \theta^2 \tag{12}$$

In the formula, λ_{CE}, λ_{CLS}, and λ_θ represent the weight coefficients for the loss components, which are used to control the degree of influence of each type of loss. The final term represents the parameter regularization term.

IV. EXPERIMENTS

In this section, we conducted experiments on three classic and authoritative public remote-sensing image datasets with a small sample size to evaluate the proposed method's superiority. This included comparison experiments with some of the latest advanced techniques, ablation studies, and experiments analyzing sensitivity to hyperparameters. We conducted our experiments using the PyTorch framework on a single NVIDIA GeForce RTX 3090 GPU.

A. EXPERIMENTAL SETUP

1) DATASET INTRODUCTION

We assessed our approach using two prestigious datasets: NWPU [62] and AID [63]. The NWPU dataset is a collection of classified remote sensing scenes gathered and assembled by Northwestern Polytechnical University. The dataset comprises 31,500 color images measuring 256 × 256 × 3 pixels, covering 45 distinct scenes. Each category includes 700 remote-sensing images. Within the NWPU dataset, we divided the 45 categories into 25 for the training set, 10 for the test set, and 10 for the validation set. The AID dataset is a collaborative creation of Huazhong University of Science and Technology and Wuhan University. The collection consists of remote-sensing scene images for classification, comprising 10,000 color images with dimensions of 600 × 600 × 3 pixels, covering 30 scenes. The number of images per category ranges from 220 to 420. We divided the 30 scene categories into 16 for training, 7 for validation, and 7 for testing. The UCM dataset [64] comprises 2,100 images from 21 distinct scenes. Each category includes 100 color images with a resolution of 256 × 256 × 3 pixels. We followed the setup of SCL-MLNet and allocated 10 out of the 21 categories in the UCM dataset for training, 5 for validation, and 6 for testing.

2) EVALUATION PROTOCOLS

To ensure a fair comparison and adhere to the commonly used evaluation framework in few-shot learning, we assess the model's performance regarding the average classification accuracy across different scenarios. Specifically, we evaluate its performance under 5-way 1-shot and 5-way 5-shot settings in Equation (13).

$$A_{CC} = \frac{1}{E}\sum_{t=1}^{E}\frac{N_t}{\varepsilon_t} \tag{13}$$

Herein, E denotes the number of few-shot tasks randomly sampled during the testing phase (which we have uniformly set to 2,000 for our experimental evaluations), N_t represents the count of correctly predicted samples within the t-th few-shot task, and ε_t signifies the total number of samples in the t-th few-shot task.

3) NETWORK ARCHITECTURE

We use the backbone network f_B with a 4-layer fully convolutional neural network for primary feature extraction to ensure a fair comparison. We set the number of channels in each layer to 64, 128, 256, and 512, respectively. Each convolutional block contains three layers: a convolutional layer with a 3 × 3 kernel, a batch normalization layer, and a global average pooling layer. In our cross-attention blocks, the support and query set samples share three identical convolutional layers—Q, K, and V. The number of input and output channels remains consistent across each convolutional layer, with a kernel size of 3 × 3. We construct the instance-level classification layer f_{CLS} with two series-connected, fully connected neural network layers, each followed by a ReLU activation function. Similarly, the comparison module f_{CE} comprises two serially connected, fully connected neural network layers.

4) TRAINING SETUP

First, we sample all images from the three datasets to a resolution of 128 × 128 × 3. From the training set, we randomly sample 200,000 tasks for training. We set the classification loss coefficient λ_{CLS} to 1.0 and the contrastive loss coefficient λ_{CE} to 0.6. We empirically determine the hyper-parameter α in the contrastive loss function to be 45°. We fix the number of cross-attention modules T to 2. We employ the Adam optimizer with an initial learning rate 0.0002 and a weight regularization coefficient of 1e-5. The learning rate decays to 90% of its previous value during the training process every 5,000 tasks.

TABLE 1. Comparative experimental results on NWPU(%) in terms of 5-way 1-shot and 5-way 5-shot scenarios with respect to average class top-1 accuracy computed by Equation (13).

Method	5-way 1-shot	5-way 5-shot
MatchingNet	38.23 ± 0.75	47.40 ± 0.69
ProtoNet	40.35 ± 1.02	69.55 ± 0.55
MAML	47.52 ± 0.70	62.58 ± 0.43
LLSR	51.43	72.90
Meta-LSTM	47.53 ± 0.80	72.36 ± 0.54
Meta-SGD	60.58 ± 0.94	76.04 ± 0.49
RelationNet	61.95 ± 1.73	76.28 ± 1.42
DLA-MatchNet	67.85 ± 0.68	79.97 ± 0.75
RS-MetaNet	46.21 ± 0.58	68.75 ± 0.70
SCL-MLNet	62.21 ± 1.12	80.86 ± 0.76
Ours	64.16 ± 0.62	78.09 ± 0.48

B. COMPARISONS WITH ADVANCED METHODS

To demonstrate the superiority of our proposed method, we start by choosing some of the most classic and advanced few-shot classification approaches recently used in general scenes. These methods include metric-based few-shot learning approaches such as MatchingNet, ProtoNet, and RelationNet. They primarily focus on learning distance metrics between instance-level samples and class prototypes. We also consider optimization-based few-shot methods such as MAML, Meta-LSTM, and Meta-SGD. These methods aim to update the model's gradients for current tasks by using global gradient directions from historical tasks. Moreover, we have included methods tailored explicitly for few-shot classification in remote sensing scene images, such as LLSR, DLA-MatchNet, Sensing-MetaNet, and SCL-MLNet. The methods above utilize 4-layer convolutional blocks to extract shallow features from remote sensing images.

We randomly sample 2,000 tasks from the test set for 5-way 1-shot and 5-way 5-shot scenarios. We calculate the final results by averaging the accuracy of each task via Equation 13 and present the results on three datasets in Tables 1, 2, and 3, respectively. To further demonstrate our proposed model's efficiency, we report a confusion matrix for a randomly sampled 5-way 1-shot 20-query task from the NUPW dataset, as shown in Figure 2.

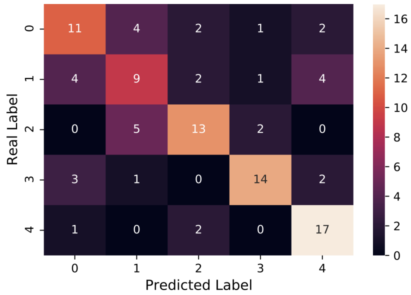


FIGURE 2. Confusion matrix for a randomly sampled 5-way 1-shot 20-query task from the NUPW dataset.

TABLE 2. Comparative experimental results on AID(%) in terms of 5-way 1-shot and 5-way 5-shot scenarios with respect to average class top-1 accuracy computed by Equation (13).

Method	5-way 1-shot	5-way 5-shot
MatchingNet	35.33 ± 0.47	56.41 ± 0.68
ProtoNet	54.79 ± 0.60	69.36 ± 0.10
MAML	47.98 ± 0.72	61.72 ± 0.81
LLSR	-	-
Meta-LSTM	49.79 ± 0.60	67.25 ± 0.89
Meta-SGD	53.14 ± 1.46	66.94 ± 1.20
RelationNet	55.26 ± 1.14	72.81 ± 0.93
DLA-MatchNet	57.21 ± 0.82	73.45 ± 0.61
RS-MetaNet	53.37 ± 0.56	72.59 ± 0.73
SCL-MLNet	59.46 ± 0.96	76.31 ± 0.68
Ours	61.12 ± 0.87	77.20 ± 0.49

TABLE 3. Comparative experimental results on UCM(%) in terms of 5-way 1-shot and 5-way 5-shot scenarios with respect to average class top-1 accuracy computed by Equation (13).

Method	5-way 1-shot	5-way 5-shot
MatchingNet	34.28 ± 0.39	53.64 ± 0.74
ProtoNet	52.42 ± 0.09	67.93 ± 1.01
MAML	49.01 ± 0.58	61.44 ± 0.63
LLSR	39.47	57.40
Meta-LSTM	47.43 ± 1.22	64.72 ± 2.49
Meta-SGD	50.21 ± 2.31	61.35 ± 2.13
RelationNet	48.74 ± 1.33	61.29 ± 0.46
DLA-MatchNet	52.76 ± 0.83	64.57 ± 0.67
RS-MetaNet	52.31 ± 1.12	67.21 ± 0.44
SCL-MLNet	51.37 ± 0.79	68.09 ± 0.92
Ours	55.03 ± 0.84	68.16 ± 0.57

TABLE 4. The Impact of contrastive embedding loss $\mathcal{L}_{CE}$ in terms of 5-way 1-shot and 5-way 5-shot scenarios with respect to average class top-1 accuracy computed by Equation (13).

Dataset	Method	5-way 1-shot	5-way 5-shot
NWPU	w.o. $\mathcal{L}_{CE}$	63.55± 0.87	76.23 ± 0.36
	w. $\mathcal{L}_{CE}$	64.16 ± 0.62	78.09 ± 0.48
AID	w.o. $\mathcal{L}_{CE}$	60.83 ± 0.76	74.96 ± 0.45
	w. $\mathcal{L}_{CE}$	61.12 ± 0.87	77.20 ± 0.49
UCM	w.o. $\mathcal{L}_{CE}$	54.27 ± 0.80	65.31 ± 0.64
	w. $\mathcal{L}_{CE}$	55.03 ± 0.84	68.16 ± 0.57

The comparative experimental results from the three datasets indicate that our proposed method holds a competitive edge. Under the 5-way 1-shot scenario, our method improved 1.6% on the AID dataset and 2.3% on the UCM dataset. However, the advantage of our proposed method is not as pronounced in the 5-way 5-shot scenario. On the NWPU dataset, the DLA-MatchNet method outperforms ours. This improved performance may be due to using five convolutional blocks for feature extraction. In contrast, we incorporated only four convolutional blocks for a fair comparison. Our method demonstrates a distinct advantage when dealing with a smaller number of samples, meaning it can more effectively capture fine-grained features of different categories within remote sensing scenes.

C. ABLATION STUDY AND HYPER-PARAMETER ANALYSIS

1) IMPACT OF CONTRASTIVE EMBEDDING MODULE

To evaluate our contrastive embedding module's effectiveness, we conduct the ablation study for the contrastive loss ($\mathcal{L}_{CE}$) on three datasets. We displayed the results for the 5-way 1-shot and 5-way 5-shot tasks on the three datasets in Table 4. The assessment study results show that our contrastive embedding module has different levels of impact on the three datasets. Specifically, when we integrated the contrastive embedding module in the 5-way 1-shot scenario, it improved performance by about 0.6% on the NWPU dataset, 0.3% on the AID dataset, and 0.7% on the UCM dataset. In the context of the 5-way 5-shot setup, the inclusion of the module resulted in improvements of approximately 1.9%, 2.2%, and 2.9% on NWPU, AID, and UCM, respectively.

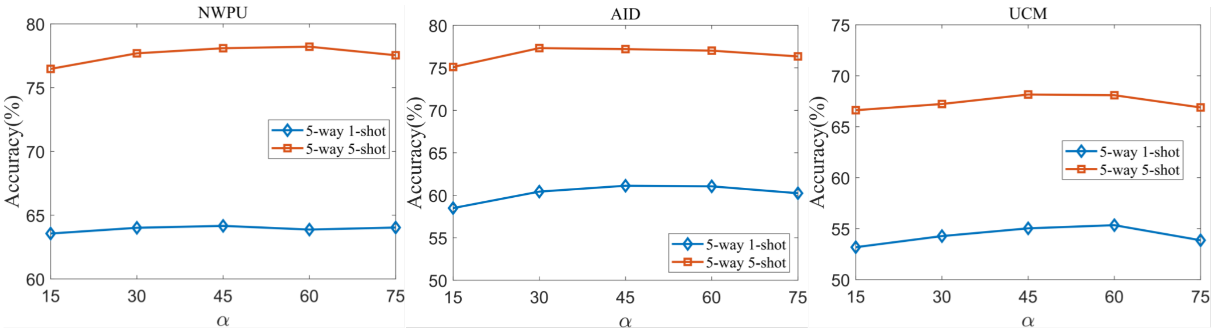
The modest gains in the 5-way 1-shot setting occur because, based on empirical evidence, the contrastive embedding module requires more samples to generate better embeddings. This helps to enhance the distinction between different categories. Our contrastive embedding module, when considered as a whole, is practical and achieves the goals set out at its conception.

2) IMPACT OF THE ANGLE α

The angle α plays a crucial role in determining the embedding orientation and distance of triplet samples within geometric space, and varying this angle directly impacts the embedding outcomes. Therefore, we investigated the effects of different hyper-parameters α on the experimental results. We empirically set the value of α at 15°, 30°, 45°, 60°, and 75°. Then, we ran experiments to analyze the sensitivity of the hyper-parameter α for the 5-way 1-shot and 5-way 5-shot tasks on three datasets in Fig. 3. From the graph, it is observable that the angle α has a significant impact on the experimental outcomes. After thorough consideration, an embedding angle of 45° generally yields the best results across all datasets.

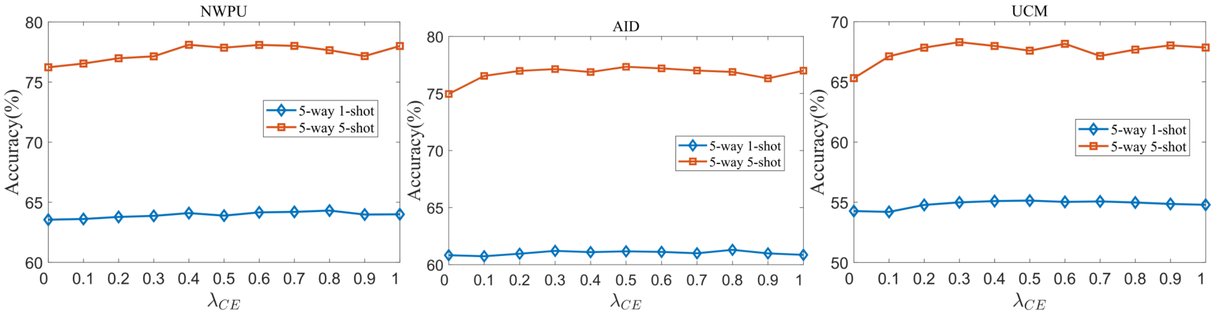
3) IMPACT OF DIFFERENT λ_{CE}

The parameter λ_{CE} determines the dominance of the contrastive embedding loss's influence within the overall loss function. To this end, we examined the impact of different hyperparameters λ_{CE} on experimental outcomes. Empirically, we set λ_{CE} in the range of [0, 1] at intervals of 0.1 and



(a) NWPU (b) AID (c) UCM

FIGURE 3. The impact of the contrastive embedding module.



(a) NWPU (b) AID (c) UCM

FIGURE 4. The impact of different λ_{CE}.

conducted sensitivity analysis experiments for the hyperparameter λ_{CE} in the 5-way 1-shot and 5-way 5-shot settings across three datasets in Fig. 4. The graph reveals that the three datasets are not particularly sensitive to hyperparameter λ_{CE}, with significant effects only when $\lambda_{CE} = 0$, when no contrastive embedding loss is included. In summary, optimal results are generally achieved across all datasets when $\lambda_{CE} = 0.6$.

V. CONCLUSION

In our work, we introduced an innovative model framework to tackle the challenges inherent in remote sensing image recognition. By integrating a cross-attention module with residual connections, we effectively addressed issues associated with feature preservation, redundant information elimination, and the vanishing gradient problem, thereby substantially boosting the capability for feature extraction within limited-sample environments. Employing a contrastive embedding module, enhanced with Angular Loss featuring angle constraints, we intensified the geometric restrictions within the feature space, which elevated the discriminability between different classes.

Furthermore, our model learned features with heightened discriminative power by amalgamating nonlinear projection transformations with instance-level classification based on cosine similarity. We conducted comprehensive evaluations of these three modules across three authoritative optical remote sensing benchmark datasets—NWPU, AID, and UCM—achieving gratifying outcomes. Our model framework is not only model-agnostic but also demonstrates considerable versatility. Moving forward, we plan to delve deeper into our research and extend the application of our findings to additional domains and few-shot learning models, aiming to pioneer new frontiers in the field.

REFERENCES

[1] Y. Fu, Q. Cheng, L. Jing, B. Ye, and H. Fu, "Mineral prospectivity mapping of porphyry copper deposits based on remote sensing imagery and geochemical data in the duolong ore district, tibet," *Remote Sensing*, vol. 15, no. 2, p. 439, 2023.

[2] E. Bedini, "The use of hyperspectral remote sensing for mineral exploration: A review," *Journal of Hyperspectral Remote Sensing*, vol. 7, no. 4, pp. 189–211, 2017.

[3] J. Li, Y. Pei, S. Zhao, R. Xiao, X. Sang, and C. Zhang, "A review of remote sensing for environmental monitoring in china," *Remote Sensing*, vol. 12, no. 7, p. 1130, 2020.

[4] H. Tian, P. Wang, K. Tansey, J. Zhang, S. Zhang, and H. Li, "An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china," *Agricultural and Forest Meteorology*, vol. 310, p. 108629, 2021.

[5] E. S. Zakiev and S. K. Kozhakhmetov, "Prospects for using remote sensing data in the armed forces, other troops and military formations of the republic of kazakhstan," *Vojnotehnicki glasnik/Military Technical Courier*, vol. 69, no. 1, pp. 196–229, 2021.

[6] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7726–7738, 2020.

[7] Q. Ma, X. Zhang, C. Zhang, and H. Zhou, "Hyperspectral image classification based on capsule network," *Chinese Journal of Electronics*, vol. 31, no. 1, pp. 146–154, 2022.

[8] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[9] Y. Liu, H. Zhang, W. Zhang, G. Lu, Q. Tian, and N. Ling, "Few-shot image classification: Current status and research trends," *Electronics*, vol. 11, no. 11, p. 1752, 2022.

[10] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, Lille, 2015.

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

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

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

[14] Y. Zhang, C. Wang, Q. Shi, Y. Feng, and C. Chen, "Adversarial gradient-based meta learning with metric-based test," *Knowledge-Based Systems*, vol. 263, p. 110312, 2023.

[15] Y. Zhang, D. Han, J. Tian, and P. Shi, "Domain adaptation meta-learning network with discard-supplement module for few-shot cross-domain rotating machinery fault diagnosis," *Knowledge-Based Systems*, vol. 268, p. 110484, 2023.

[16] X. Liang, M. Zhang, G. Feng, Y. Xu, D. Zhen, and F. Gu, "A novel deep model with meta-learning for rolling bearing few-shot fault diagnosis," *Journal of Dynamics, Monitoring and Diagnostics*, pp. 1–22, 2023.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[19] L. S.-T. Memory, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 2010.

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

[21] T. Munkhdalai and H. Yu, "Meta networks," pp. 2554–2563, 2017.

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

[23] L. Lu, X. Cui, Z. Tan, and Y. Wu, "Medoptnet: Meta-learning framework for few-shot medical image classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

[24] J. Chen, J. Tang, and W. Li, "Industrial edge intelligence: Federated-meta learning framework for few-shot fault diagnosis," *IEEE Transactions on Network Science and Engineering*, 2023.

[25] A. G. Khoee, A. Javaheri, S. R. Kheradpisheh, and M. Ganjtabesh, "Meta-learning in spiking neural networks with reward-modulated stdp," *arXiv preprint arXiv:2306.04410*, 2023.

[26] F. Zou, S. Sang, M. Jiang, X. Li, and H. Zhang, "Few-shot pump anomaly detection via diff-wrn-based model-agnostic meta-learning strategy," *Structural Health Monitoring*, vol. 22, no. 4, pp. 2674–2687, 2023.

[27] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.

[28] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International conference on learning representations*, 2016.

[29] D. Pal, S. Bose, D. More, A. Jha, B. Banerjee, and Y. Jeppu, "Maml-sr: Self-adaptive super-resolution networks via multi-scale optimized attention-aware meta-learning," *Pattern Recognition Letters*, vol. 173, pp. 101–107, 2023.

[30] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-lingual few-shot hate speech and offensive language detection using meta learning," *IEEE Access*, vol. 10, pp. 14880–14896, 2022.

[31] M. Ren, Z. Wang, and X. Yu, "Personalized federated learning: A clustered distributed co-meta-learning approach," *Information Sciences*, vol. 647, p. 119499, 2023.

[32] X. Guo and R. Zhou, "Data augmentation method for extracting partially occluded roads from high spatial resolution remote sensing images," *IEEE Access*, vol. 11, pp. 79232–79239, 2023.

[33] Y. Yan, Y. Zhang, and N. Su, "A novel data augmentation method for detection of specific aircraft in remote sensing rgb images," *IEEE Access*, vol. 7, pp. 56051–56061, 2019.

[34] A. Forootani, M. Rastegar, and H. Zareipour, "Transfer learning-based framework enhanced by deep generative model for cold-start forecasting of residential ev charging behavior," *IEEE Transactions on Intelligent Vehicles*, pp. 1–9, 2023.

[35] G. M. Dimitri, S. Spasov, A. Duggento, L. Passamonti, P. Lió, and N. Toschi, "Multimodal and multicontrast image fusion via deep generative models," *Information Fusion*, vol. 88, pp. 146–160, 2022.

[36] D. Chen, Y. Chen, Y. Li, F. Mao, Y. He, and H. Xue, "Self-supervised learning for few-shot image classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1745–1749, IEEE, 2021.

[37] X. Li, D. Shi, X. Diao, and H. Xu, "Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[38] T. Gong, X. Zheng, and X. Lu, "Meta self-supervised learning for distribution shifted few-shot scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[39] H. Qi, Z. Jinyuan, H. Dongmei, D. Yanling, X. Huifang, *et al.*, "Raprotonet: A few-shot remote sensing scene classification method based on meta-learning," *Laser & Optoelectronics Progress*, vol. 60, no. 10, p. 1028003–1028003, 2023.

[40] S. Sharma, R. Roscher, M. Riedel, and G. Cavallaro, "Few-shot remote sensing image classification with meta-learning," *Authorea Preprints*, 2023.

[41] X. Ma, W. Wang, W. Li, J. Wang, G. Ren, P. Ren, and B. Liu, "An ultralightweight hybrid cnn based on redundancy removal for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[42] D. Yunya and Z. Qian, "A survey of depth semantic feature extraction of high-resolution remote sensing images based on cnn," *Remote Sensing Technology and Application*, vol. 34, no. 1, pp. 1–11, 2019.

[43] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention. advances in neural information processing systems [c]," in *Proc. of Neural Information Processing Systems (NIPS)*, vol. 2, 2014.

[44] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[45] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[48] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[54] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap

your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[55] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[56] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

[57] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[58] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.

[59] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Dino: Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.

[60] J. Cui, Z. Zhong, Z. Tian, S. Liu, B. Yu, and J. Jia, "Generalized parametric contrastive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[61] J. Yin, H. Wu, and S. Sun, "Effective sample pairs based contrastive learning for clustering," *Information Fusion*, vol. 99, p. 101899, 2023.

[62] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[63] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[64] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010.

YE RONG was born in 1983 and received a master's degree in physics from Hebei Normal University in Shijiazhuang City, Hebei Province, China, in 2017. Since 2020, she has pursued a doctoral degree in optical engineering at Army Engineering University of PLA.

Since 2017, she has been working at Hebei Transportation Vocational and Technical College as a teacher in the electric power supply specialty. During her tenure, she has published more than ten articles, with her research primarily focusing on military optoelectronic system applications, few-shot image processing, and visual image processing, high-speed permanent magnet motors, uncrewed sweeping vehicles.

QING-YI KONG (IEEE) borned in 1983 in Shijiazhuang, Hebei Province, and obtained his bachelor's, master's, and doctoral degrees in electrical engineering from Hebei University of Technology in Tianjin, China, in the years 2006, 2009, and 2014, respectively.

From 2014 to 2017, he was a lecturer and master's supervisor at the School of Electrical Engineering at Hebei University of Science and Technology. Since 2017, he has been a professor at Hebei Transportation Vocational and Technical College. From 2020 to the present, he has held the position of Technical Director at Kingston Technology Co., Ltd. in Hebei, China. He has published more than 30 articles and holds 17 patents. His main research interests include engineering electromagnetic field analysis, high-speed permanent magnet motor design, and the design and application of hydrogen fuel cell air compressors.

Dr. Kong served as an IEEE member of the, a senior member of the Structural Health Monitoring and Early Warning Branch of the China Instrument and Control Society, a senior member of the China Electrotechnical Society, a senior member of the Hebei Electrical Engineering Society, and a peer reviewer for several academic journals.

GUANG-LONG WANG borned in 1964 in Sishui, Shandong, is a professor and doctoral supervisor at Nankai University. He holded a doctorate from the Beijing Institute of Technology, completed postdoctoral research at Tsinghua University, and conducted further postdoctoral work at the University of Cambridge. He is a senior visiting scholar at the University of Science and Technology of China and a senior research fellow at Nanyang Technological University. He teaches a variety of undergraduate and graduate courses.

His long-term research interests include sensors and intelligent instruments, micro-nanotechnology, optoelectronics, and wireless communication. He is a leading academic figure in instrument science, technology, and optical engineering. He has received seven national invention patents, published over 200 scholarly papers, and written two monographs.

He served as a senior member and the Secretary-General of the Structural Health Monitoring and Early Warning Branch of the China Instrument and Control Society. He served as a review committee member for the National Natural Science Foundation of China and a reviewer for multiple academic journals.