Exploring Hard Samples in Multi-View for Few-Shot Remote Sensing Scene Classification

Yuyu Jia, Junyu Gao, *Member, IEEE,* Wei Huang,
Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Few-shot remote sensing scene classification is of high practical value in real situations where data are scarce and annotated costly. The few-shot learner needs to identify new categories with limited examples, and the core issue of this assignment is how to prompt the model to learn transferable knowledge from a large-scale base dataset. Although current approaches based on transfer learning or meta-learning have achieved significant performance on this task, there are still two problems to be addressed: (i) as an essential characteristic of remote sensing images, spatial rotation insensitivity surprisingly remains largely unexplored; (ii) the high distribution uncertainty of hard samples reduces the discriminative power of the model decision boundary. Stimulated by these, we propose a corresponding end-to-end framework termed a Hard Sample Learning (HSL) and Multi-view Integration (MI) Network (HSL-MINet). First, the MI module contains a pretext task introduced to guide the knowledge transfer, and a multiview-attention mechanism used to extract correlational information across different rotation views of images. Second, aiming at increasing the discrimination of the model decision boundary, the HSL module is designed to evaluate and select hard samples via a class-wise adaptive threshold strategy, and then decrease the uncertainty of their feature distributions by a devised triplet loss. Extensive evaluations on NWPU-RESISC45, WHU-RS19, and UCM datasets show that the effectiveness of our HSL-MINet surpasses the former state-of-the-art approaches.**

***Index Terms*—multi-view images, hard samples, meta-learning, and prototypes.**

I. Introduction

THE goal of remote sensing (RS) scene classification is to match descriptive labels to images in line with their visual characteristics. As a significant component within the community of RS image analysis, it has received much interest from many researchers [1], [2], [3]. Recently, approaches thanks to Deep Neural Networks (DNNs) attained remarkable improvements [4], [5] [6], [7], [8], typically driven by massive labeled data. However, it is unrealistic to annotate large amounts of RS images in real applications [9], which becomes a hurdle that plagues DNNs-based methods [10]. Enlightened by the rapid knowledge transfer ability of humans, few-shot learning (FSL) aims to address the challenge of recognizing novel categories with limited annotated examples, which serves as a vital bridge to narrow the gap between artificial intelligence and humans [11].

Generally, the few-shot model learns from a base dataset and leverages its knowledge to discriminate novel categories through a small number of labeled examples (typically 1 or 5) [12], [13], [14], [15]. Different from regular classification tasks, the samples seen in the test phase and training phase are class-disjoint in FSL tasks. Recent reasonable algorithms are proposed to address this challenge that can be roughly separated into two branches. Models in the first branch are pre-trained in the entire base dataset [16], [17], [18], [19], which fall into the fine-tuning paradigm. Another promising branch is in the meta-training schemes, also known as "learning to learn" [14], [20], [21], [22]. Through the execution of multiple episodes sampled from the base dataset, the model can simulate the few-shot scenario during the meta-training phase, thereby enhancing its adaptability to novel tasks. In each episode, the model classifies samples with unknown labels (query samples) based on a few samples with known labels (support samples). With the help of metric learning, many efficient FSL works in the meta-learning scheme designed sophisticated algorithms and achieved encouraging performance. However, they omit an intrinsic characteristic of RS images which can be summarized as **spatial rotation insensitivity** and do not fully mine the potential information between images with different rotation views. Apart from this, the issue of **large intra-class variance and small inter-class discrimination** is seldom alleviated from the perspective of hard samples learning, which is a more targeted solution.

As shown in Fig. 1(a), the **spatial rotation insensitivity** is an obvious dissimilarity between RS images and natural optical images. The former does not contain explicit orientation information, while the latter has obvious up, down, left, and right attitudes. In other words, the category probability distributions of the same RS image with different rotation views should be consistently produced by the FSL classifier. Furthermore, since the contents in different rotation views of each RS image contribute almost equally to the semantic label, the potentially shared information among them is discriminative to provide essential value for matching query samples to category centers.

Affected by background, illumination, distribution discrepancies, *etc*., the scene containing the same semantic contents varies in different RS images [23]. Moreover, the dense distri-

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902 and in part by the National Key Research and Development Program of China under Grant No. 2022ZD01604004.

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

Wei Huang is with the Data Science in Earth Observation, Technical University of Munich, Munich 80333, Germany.

E-mail: jyy2019@mail.nwpu.edu.cn, gjy3035@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com, hw2hwei@gmail.com.

Qi Wang and Junyu Gao are the corresponding authors.

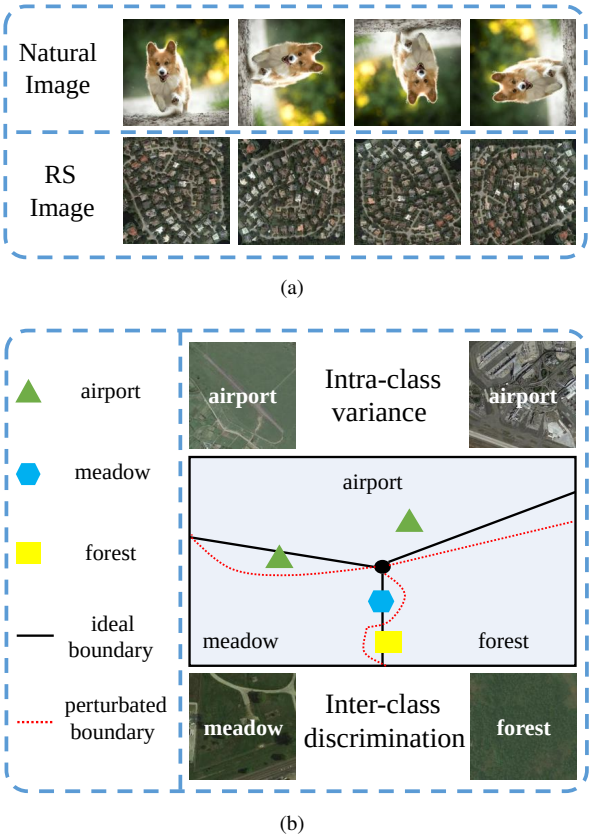


Fig. 1. (a) Spatial rotation insensitivity. (b) Hard samples caused by large intra-class variance and small inter-class discrimination bring perturbations to the model decision boundary.

bution of scenes with different semantics may contain similar objects, leading to a decrease in inter-class discrimination. As we can see in Fig. 1(b), influenced by the common phenomenon of **large intra-class variance and small inter-class discrimination** in RS images, some query samples tend to be in the category ambiguous region in the feature space (*i.e.*, ***hard samples***), which brings perturbation to the model decision boundary and furtherly weakens the generalization of the model. Therefore, to prompt the model to obtain a clearer and more discriminative decision boundary, a targeted learning strategy is needed to calibrate the feature distributions of the selected hard samples.

In response to the two points mentioned above (*i.e.*, spatial rotation insensitivity and hard samples in RS images), the proposed HSL-MINet consists of a Multi-view Integration (MI) module and a Hard Samples Learning (HSL) module to realize few-shot RS scene classification. In the MI module, we jointly optimize two tasks to fully explore the spatial rotation insensitivity of RS images. The first task is not only to distinguish the rotation views of each image [24] but also to keep the distribution of these different views consistent in the semantic space. The second task is to construct the main classifier of the model, in which we devise a multiview-attention mechanism to extract the self-shared information among different views of each image and the cross-correlational attention from each query sample to category centers. In the HSL module, the feature distribution of the poorly classified samples in each category is selected as "anchor distribution" by a class-wise adaptive threshold strategy, and the remaining feature distribution in this category is called "positive distribution". Subsequently, a triplet loss is proposed to lead the "anchor distribution" to attract the "positive distribution" with the same semantic category while pushing away the nearest heterogeneous-semantic "negative distribution". As a result, by reducing the interference of hard samples on the feature distribution, we achieve a model with a clearer and more discriminative decision boundary, which benefits the generalization of the model.

The main contributions of this paper are:

1) Propose a multiview-attention mechanism to extract the self-shared information among different rotation views for each image, and capture the cross-correlational attention between category centers and query features.
2) Present a perspective of hard samples learning to mitigate the issue of large intra-class variance and small inter-class discrimination. Concretely, we propose a class-wise adaptive threshold strategy and a hard-sample triplet loss, which aim to decrease the interference of hard samples on the decision boundary.
3) The proposed HSL-MINet is evaluated on three public benchmarks, where it achieves superior performances compared with the previous leading techniques.

II. Related Works

In this section, relevant research of this paper is briefly reviewed, including remote sensing scene classification, few-shot learning, and few-shot remote sensing scene classification.

A. Remote Sensing Scene Classification

Remote sensing scene classification (RSSC) is a straightforward yet crucial work that involves categorizing images into a label set [3]. Before Deep Neural Networks (DNNs) were widely used, most methods relied on manual feature selection to achieve category prediction [25], [26], [27], [28], [29].

Although these techniques have driven significant advances, their effectiveness strongly relies on manual manipulation and feature filtering. Due to its powerful feature extraction capability, DNNs-based methods have made significant breakthroughs and attracted widespread attention [30]. For instance, Penatti *et al.* [31] first utilized a convolutional neural network to extract features from RS images and achieved promising performance in the RSSC task. A global-local two-branch network was designed in [32] to separately extract the global and local features from the entire image, and hence mitigate the problems caused by the large-scale variation of images. Wang *et al.* [33] embedded the attention mechanism into the recurrent convolutional network and constructed a new end-to-end algorithmic framework for RSSC. Cheng *et al.* [34] devised a novel training framework to train classifiers using samples generated during image reconstruction.

Despite these methods' excellent performance with sufficient source data, data scarcity is a more common scenario in which their performance will drop seriously. The proposed HSL-MINet can significantly alleviate the decline by fully

leveraging the spatial rotation insensitivity and exploring hard samples in a meta-learning manner.

B. Few-Shot Learning

Few-shot learning (FSL) allows for the generalization of knowledge acquired from the base dataset to novel datasets that have dissimilar statistical distributions. Meta-learning-based methods frame FSL in the scheme of episodic training. An episode is constructed by simulating the few-shot scenario of the test phase during the training phase, even with sufficient labeled data [35], [36]. The above FSL studies can be summarized into two groups: metric-based approaches and optimization-based approaches.

Metric-based methods represent images in a low-dimensional space, which allows the FSL task to be seen as a nearest-neighbor retrieval operation with Euclidean [37], cosine [38], [39] earth mover's distance [40], and learnable distance [15]. Vinyals *et al*. [14] aligned the label space of a small labeled support set and an unlabeled query set, building a learning framework that does not require fine-tuning to expeditiously acclimate to new tasks. Snell *et al*. [37] brought in the concept of "prototype", which can more efficiently model the clusterings of samples of the same category in the metric space. In [40], the earth mover's distance is introduced to determine the relevance of two images. Tseng *et al*. [41] addressed the problem of large discrepancies under the feature distribution across domains by simulating various feature distributions in the training stage.

Optimization-based models are optimized to adapt to new tasks under the meta-learning framework. Towards the goal of quick adaptation, Finn *et al*. [42] devised a model dubbed MAML that gets rid of the reliance on gradient steps. Besides, many MAML variants used to be committed to the study of learning a good model initialization. Munkhdalai *et al*. [43] introduced a method using an artificial neural network to acquire the ability for quick adaptation and implemented it in a meta-learning framework. Inspired by the MAML, Park *et al*. [44] proposed meta-curvature (MC), which achieves better generalization to novel tasks by transforming the gradients during the internal optimization process.

Aside from two widespread groups, a few other promising approaches have emerged to tackle the FSL problem, such as graph-based works [45], distribution calibration [46], incremental learning of FSL [47], *etc*. We adopt the metric-based Prototypical Network [37] as a basic classifier in this work due to its simplicity and popularity. Benefiting from our HSL module and the MI module, the classification capability of the model is improved in the few-shot situation.

C. Few-Shot Remote Sensing Scene Classification

The ideas behind few-shot RS scene classification and few-shot natural image classification are similar. Among them, approaches leveraging meta-learning have dominated in recent years. In [48], RS-MetaNet is designed to train the model at the episodic level instead of the sample level, allowing it to acquire a metric space with a robust representation capability. A Siamese prototype network is devised in [49] to enhance the category prototype through self-calibration and inter-calibration. Li *et al*. [23] employed the attention mechanism to investigate the inter-channel and inter-spatial correlations, hence automatically discovering discriminative regions. Gong *et al*. [50] settled the challenge of few-shot RS scene classification from both data augmentation and network architecture perspectives. To mitigate the limitation of a few labeled RS images on classification performance, Zhang *et al*. [51] utilized a graph-matching module to boost the meta-learning model. Liu *et al*. [52] boosted the generalization abilities of models by enforcing models to rank support images according to the support-query similarity. Zeng *et al*. [3] reported a self-attention and mutual-attention to enhance the task-specific discriminative features, hence, improving the classification performance with fewer labeled samples.

Although rotation transformations, intra-class variance, and inter-class discrimination have been referred to by some relevant works [51], [23], [48], the spatial rotation insensitivity and hard samples behind them remain largely untapped. Thus, our work attempts to fill the gap in this research.

III. Methodology

In this part, we will outline the issue being addressed and the basic classifier, followed by an explanation of the overall structure of the proposed approach. Finally, we present the HSL-MINet and its core component.

A. Preliminary

1) Problem definition: In this part, we will give an explicit mathematical definition of the FSL problem. There are such three category sets, which are denoted as $\mathcal{C}_{base}$, $\mathcal{C}_{val}$, and $\mathcal{C}_{novel}$, respectively. Unlike conventional classification tasks, the three category sets are of non-overlapping categories, *i.e*., $\mathcal{C}_{base} \cap \mathcal{C}_{val} \cap \mathcal{C}_{novel} = \emptyset$. There is a dataset that corresponds to each category set, denoted as $\mathcal{D}_{base} = \{(x_i, y_i), y_i \in \mathcal{C}_{base}\}_{i=1}^{N_{base}}$, $\mathcal{D}_{val} = \{(x_i, y_i), y_i \in \mathcal{C}_{val}\}_{i=1}^{N_{val}}$ and $\mathcal{D}_{novel} = \{(x_i, y_i), y_i \in \mathcal{C}_{novel}\}_{i=1}^{N_{novel}}$, respectively, where (x_i, y_i) means the i-th image and its label information.

We solve the FSL task by leveraging an episodic training scheme as in previous meta-learning studies [37], [35]. Each episode $\mathcal{E}$ contains a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N \times K}$ and a query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{N \times M}$, where N represents the number of categories in one episode, K denotes the number of support samples for each category, and M stands for the number of query samples for each category. The above setting is termed as an N-way K-shot task. Similarly, we can organize episodes on $\mathcal{D}_{val}$ and $\mathcal{D}_{test}$ in the validation phase and testing phase. The objective is to promote the model to acquire transferable knowledge through those meta-learning episodes.

2) Basic classifier: The metric-based Prototypical Network (ProtoNet) [37] is employed as the basic classifier for the HSL-MINet. First, all images in each episode are represented to the embedding space through a feature extractor. After that, the support features of the corresponding category are averaged as the prototype of this category, that is,

$$p_c = \frac{1}{K} \sum_{(x_i, y_i) \in \mathcal{S}} f_\theta(x_i) \cdot \mathbb{I}(y_i = c), \tag{1}$$

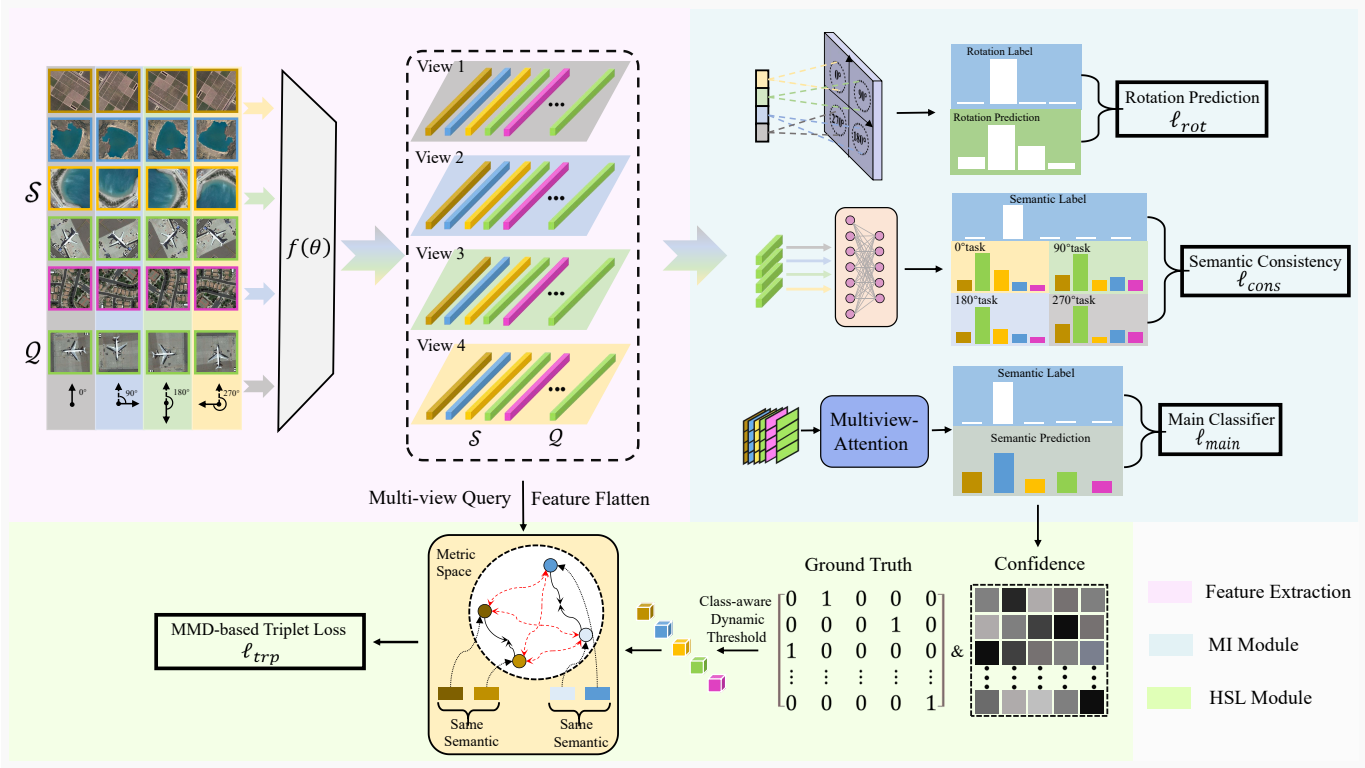


Fig. 2. Illustration of the HSL-MINet for a 5-way 1-shot setting. It includes a feature extractor and two modules. (a) In the Multi-view Integration (MI) module, the rotation prediction loss $\mathcal{L}_{rot}$ and the semantic consistency loss $\mathcal{L}_{cons}$ both fall into a self-supervised manner. Besides, the main classifier loses $\mathcal{L}_{main}$ the model through a multiview-attention mechanism. (b) In the Hard Samples Learning (HSL) module, the hard-sample triplet loss $\mathcal{L}_{trp}$ optimizes the decision boundary of the model by modifying the feature distribution of hard samples.

where $f_\theta(x_i)$ is the feature vector output by the feature extractor of i-th image, $\mathbb{I}$ is the indicator function, and p_c denotes the prototype of the c-th category.

Aiming at classifying a sample $x_i \in \mathcal{Q}$ with an unknown category from the query set, we need to measure the distance between the query feature and the prototypes. Formally, the probability distribution that the query sample x_i is predicted to be category c can be defined as:

$$P_\theta^i(\hat{y}=c|x_i) = \frac{exp\left(-\tau \cdot d\left(f_\theta\left(x_i\right), p_c\right)\right)}{\sum_{\acute{c}}^{N} exp\left(-\tau \cdot d\left(f_\theta\left(x_i\right), p_{\acute{c}}\right)\right)}, \quad (2)$$

where $d(\cdot)$ represents a distance metric (*e.g.*, the Euclidean distance) and τ is a temperature coefficient.

Then, the baseline loss can be defined as:

$$\mathcal{L}_{baseline} = -\frac{1}{|\mathcal{Q}|} \sum_{(x_i, y_i \in \mathcal{Q})}^{|\mathcal{Q}|} log\left(P_\theta^i(\hat{y}=c|x_i)\right), \quad (3)$$

where c denotes the ground truth label of corresponding query sample x_i, and $\hat{y}$ is the class predicted by the model.

B. Method Overview

Starting from the two aspects of RS images (namely, spatial rotation insensitivity and hard samples), HSL-MINet is proposed with ProtoNet as a basic classifier. Fig. 2 depicts the HSL-MINet, which consists of a feature extractor built on ResNet12 [20] and two modules: the MI module and the HSL module. In each episode, we first utilize rotation transformations to generate multiple views for all images. Subsequently, the extracted raw features need to be jointly processed by two tasks in the MI module. To learn more transferable knowledge, we introduce a pretext task, in which the model needs to distinguish the different rotation views of each image and produce consistent semantic probability distributions for these different views. The second task constructs the main classifier, which integrates different rotation views and furtherly captures their self-shared information and cross-correlations through a multiview-attention mechanism. What is more, we focus on hard samples with poor classification performance caused by the large intra-class variance and small inter-class discrimination. These hard samples disrupt the decision boundary of the model and thus drop the generalizability. To this end, the HSL module aims to modify the raw feature distribution of each category by a class-wise adaptive threshold strategy and a hard-sample triplet loss. Of particular importance, this module solely serves the stated objective during the training phase and is not involved in the forward inference. Finally, through the combination of these two modules, HSL-MINet improves the feature representation and generalization ability compared to the basic classifier. Algorithm 1 summarizes the process of the training and inference stages.

C. Multi-view Integration (MI) Module

In the MI module, multiple views of an image are generated with rotation transformations. Formally, a set of rotators are

defined as $\mathcal{G} = \{g_r\}_{r=1}^{R}$, where g_r refers to the process of rotating an image by $(r-1) \times 90$ degrees and R represents the total number of rotators (Our method performs best when R equals 4). Therefore, each image $x_i \in \mathcal{E}$ is extended to R views $X_i = \{x_i^1, \ldots, x_i^R\}$, where $x_i^r = g_r(x_i)$. The MI module aims to explore the spatial rotation insensitivity of RS images by leveraging two tasks, which are called pretext task and multi-view feature integration task. Building upon previous state-of-the-art research in [35], we introduce the "pretext task" to promote the model to acquire transferable knowledge. More importantly, the multi-view feature integration task is an improvement over [35], which is one of the contributions in this paper.

Algorithm 1 Pipeline of the Proposed Method.

Input: Datasets $\mathcal{D}_{base}$, $\mathcal{D}_{val}$, $\mathcal{D}_{novel}$, the rotation operator $\mathcal{G} = \{g_r\}_{r=1}^{R}$, the loss weight hyperparameters β, γ, and λ

Output: The learned ν

▷ Training
Randomly initialize all learnable parameters $\nu = \{\theta, \varphi, \Phi\}$
for Iteration $= 1, \ldots,$ MaxIteration **do**
  Randomly sample episode $\mathcal{E}$ from $\mathcal{D}_{base}$
  Generate the multi-view image set for each image in $\mathcal{E}$ leveraging $\mathcal{G}$
  Compute the loss $\mathcal{L}_{rot}$ for the rotation prediction pretext task with Eq. 4
  Compute the loss $\mathcal{L}_{cons}$ for the consistent pretext task with Eq. 7
  Compute the loss $\mathcal{L}_{main}$ for the main classification with Eq. 11
  Compute the loss $\mathcal{L}_{trp}$ for the hard samples learning with Eq. 15
  $\mathcal{L}_{total} = \mathcal{L}_{main} + \beta\mathcal{L}_{rot} + \gamma\mathcal{L}_{cons} + \lambda\mathcal{L}_{trp}$
  Update ν based on $\nabla_\nu \mathcal{L}_{total}$
end for
Select best ν using $\mathcal{D}_{val}$
Output ν
▷ Inference
Perform classification on the $\mathcal{D}_{novel}$ with Eq. 15

1) Pretext task: The task falls into a self-supervised manner, which is instantiated by employing two loss functions.

First, the model undergoes training to distinguish the multiple views of each image, thus understanding the canonical poses of images. Specifically, given a rotation classifier f_φ, which maps the features to rotation category space. The loss can then be formulated as:

$$\mathcal{L}_{rot} = -\frac{1}{|\mathcal{E}|} \cdot \frac{1}{R} \sum_{i=1}^{|\mathcal{E}|} \sum_{r=1}^{R} log \frac{exp\left(\left[f_\varphi\left(f_\theta\left(x_i^r\right)\right)\right]_r\right)}{\sum_{r'}^{R} exp\left(\left[f_\varphi\left(f_\theta\left(x_i^r\right)\right)\right]_{r'}\right)}, \quad (4)$$

where $x_i^r \in X_i$ denotes the r-th view of the i-th image, $f_\varphi\left(f_\theta\left(x_i^r\right)\right) \in \mathbb{R}^R$ is the vector of predicted rotation score, and $[\cdot]_r$ refers to obtaining the r-th component in a vector.

Second, considering that the semantic information contained in the rotation views of each image is consistent, the corresponding category probability distribution output by the FSL classifier should be consistent. We first need to perform semantic classification operations on episodes under each view. Formally, for an episode $\mathcal{E}^r = \mathcal{S}^r \cup \mathcal{Q}^r$ with r-th rotation view, the probability distribution of the predicted query set $\mathcal{Q}^r$ is defined as $\mathcal{P}^r = \left[p_1^r; \ldots; p_{|\mathcal{Q}|}^r\right] \in \mathbb{R}^{|\mathcal{Q}| \times N}$, where $p_i^r \in \mathbb{R}^N$ is the probability distribution for x_i^r in $\mathcal{Q}^r$. It's c-th element $[p_i^r]_c$ can be similarly computed as (2):

$$[p_i^r]_c = \frac{exp\left(-\tau \cdot d\left(f_\theta\left(x_i^r\right), p_c^r\right)\right)}{\sum_{\acute{c}}^{N} exp\left(-\tau \cdot d\left(f_\theta\left(x_i^r\right), p_{\acute{c}}^r\right)\right)}, \quad (5)$$

where the prototype p_c^r of category c is computed from $\mathcal{S}^r$. Subsequently, the average distribution of probabilities for the query set across all views is:

$$\hat{\mathcal{P}} = \frac{1}{R} \sum_{r=1}^{R} \mathcal{P}^r. \quad (6)$$

Towards the goal of keeping the semantic probability distributions of query sets consistent under different rotation views, we minimize the Kullback–Leibler (KL) divergence of probability distribution between each view and the mean of all views. The consistent loss holds the following form:

$$\mathcal{L}_{cons} = \frac{1}{R} \sum_{r=1}^{R} (\mathcal{D}_{KL}(\hat{\mathcal{P}} || \mathcal{P}^r) + \mathcal{D}_{KL}(\mathcal{P}^r || \hat{\mathcal{P}})). \quad (7)$$

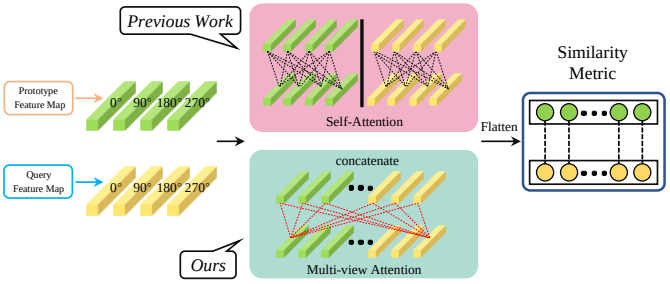


Fig. 3. The comparison between the previous work and ours. Ours can both extract the self-shared information and capture the cross-correlations.

2) Multi-view feature integration task: Since the contents in different rotation views of each image contribute almost equally to the semantic label, the self-shared information among these views plays a crucial role in capturing discriminative features. Furthermore, the classification results are predicted according to the nearest neighbor prototype, necessitating the model's ability to adaptively capture the cross-correlational attention between prototypes and query features. As shown in Fig. 3, different from the feature integration strategy in [35], which only considers the self-shared information in multiple views of each image, we uniform the extraction of self-shared information and the capturing of cross-correlations into a multiview-attention mechanism.

For a certain set of multi-view images $X_i = \{x_i^1, \ldots, x_i^R\}$, after the feature extractor, R feature vectors are outputted with a dimension of d. The multi-view feature map $F_i \in \mathbb{R}^{R \times d}$ is constructed by concatenating these R feature vectors. Similarly, let F_i^S and F_i^Q denote a multi-view support feature map and a multi-view query feature map, respectively.

Specifically for each query sample, an augmented multi-view feature map $\hat{F}_i$ is first obtained by concatenating the query feature map and the N category prototypes along the second dimension, that is,

$$\hat{F}_i = \left\{ [p_c^f; F_i^Q] | c = 1, \ldots, N \right\}, \tag{8}$$

where $\hat{F}_i \in \mathbb{R}^{N \times (2 \times R) \times d}$ and the multi-view prototype feature map $p_c^f = \frac{1}{K} \cdot \sum_{i=1}^{|\mathcal{S}|} F_i^S \cdot \mathbb{I}(y_i = c) \in \mathbb{R}^{R \times d}$ is calculated from the support set in all views.

We then adopt a transformer to integrate self-shared information and capture cross-correlations in the augmented multi-view feature map $\hat{F}_i$. Drawing inspiration from the self-attention mechanism [53], the triplet map $(\hat{F}_i, \hat{F}_i, \hat{F}_i)$ is fed into the feature integration transformer f_Φ as (Query, Key, and Value). With $\hat{F}_i$ representing the augmented multi-view feature map of the i-th sample in $\mathcal{Q}$, the feature integration transformer can be described as:

$$\hat{F}_i^{atten} = atten(\hat{F}_i, \hat{F}_i, \hat{F}_i), \tag{9}$$

where $\hat{F}_i^{atten} \in \mathbb{R}^{N \times (2 \times R) \times d}$ and $atten(\cdot, \cdot, \cdot)$ represents the self-attention operation. Subsequently, we get an integrated prototype feature map $Z_i^P \in \mathbb{R}^{N \times R \times d}$ and an integrated query feature map $Z_i^Q \in \mathbb{R}^{N \times R \times d}$ by splitting $\hat{F}_i^{atten}$ equally along the second dimension.

Afterwards, Z_i^P and Z_i^Q are reshaped to $\mathbb{R}^{N \times L} (L = R \times d)$. The distance set of the query sample can be written as:

$$D_i = \left\{ d(row_j(Z_i^P), row_j(Z_i^Q)) | j = 1, \ldots, N \right\}, \tag{10}$$

where $d(\cdot)$ refers to the Euclidean distance function, $D_i \in \mathbb{R}^N$, and row_j means taking the j-th row in a matrix.

Finally, the distance set is fed into a standard classifier to formulate the main classification loss:

$$\mathcal{L}_{main} = -\frac{1}{|\mathcal{Q}|} \cdot \sum_{i=1}^{|\mathcal{Q}|} log \frac{exp(-\tau \cdot [D_i]_{y_i})}{\sum_c^N exp(-\tau \cdot [D_i]_c)}, \tag{11}$$

where the $[D_i]_c$ means the Euclidean distance between the i-th query sample and the prototype of category c. Note that the calculation of this distance incorporates the learned self-shared information and cross-correlations, which enhances representations with discrimination toward metric space.

D. Hard Samples Learning (HSL) Module

While the MI module promotes the model to learn the transferable emphasis feature extraction ability by exploiting the spatial rotation insensitivity of RS images, it ignores another factor that affects the generalization, that is, the perturbation on the decision boundary of the model brought by hard samples. It is commonly believed that the disturbed and complex decision boundary is detrimental to generalizability (*i.e.*, overly sensitive to intra-class noise). Motivated by this, we improve the intra-class invariance and inter-class discrimination by focusing on the hard samples, and hence make the decision boundary clearer and more discriminative. Firstly, the features of hard samples are filtered out through a class-wise adaptive threshold strategy. Subsequently, on the basis of the distance between feature distributions in metric space, a hard-sample triplet loss is devised to make the feature distributions of the same category more compact and those of different categories more distinct.

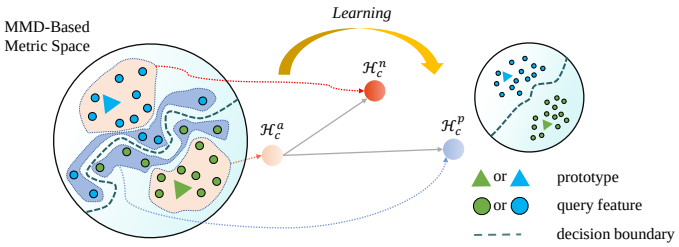


Fig. 4. Visualization of the hard-sample triplet loss $\mathcal{L}_{trp}$ in HSL module. It can be observed that the feature distribution after hard sample learning brings a clearer and more discriminative decision boundary of the model.

1) Class-wise adaptive threshold: The class-wise adaptive threshold is calculated from the prediction confidence scores of the MI module. Since the confidence scores vary by category, it is unfair to the feature distributions of different categories. Additionally, considering the evolution of the model, it is unreasonable to set a fixed threshold. Therefore, the threshold needs to be associated with categories and adaptively adjusted during the training phase.

To ensure that the threshold changes adaptively according to the classification difficulties under different categories, the confidence scores of misclassified samples in each category are first masked to zero, and then the average confidence score of all samples is taken as the class-wise adaptive threshold. Specifically, compared with the one-hot ground truth $\left\{ y_i \in \{0,1\}^N | i = 1, \ldots, |\mathcal{Q}| \right\}$, the correctly-predicted query samples $\bar{\mathcal{Q}} = \left\{ (x_i, \hat{y}_i, s_i) | i = 1, \ldots, |\bar{\mathcal{Q}}| \right\}$ are selected, where $\hat{y}_i$ and s_i separately represent the predicted category and confidence score of the i-th query sample. Then we construct the adaptive threshold $T_c \in \mathbb{R}$ of class c as:

$$T_c = \frac{1}{M} \sum_{i=1}^{|\bar{\mathcal{Q}}|} \mathbb{I}(\hat{y}_i = c) \cdot s_i, \tag{12}$$

where M stands for the number of query samples for each category.

2) Hard-sample triplet loss: According to the class-wise adaptive threshold, the feature distribution in each category with confidence scores above the threshold is referred to as the "anchor distribution" $\{\mathcal{H}_c^a | c = 1, \ldots, N\}$, otherwise we call it "positive distribution" $\{\mathcal{H}_c^p | c = 1, \ldots, N\}$. For evaluating the distance between feature distributions, the Maximum Mean Discrepancy (MMD) [54] is adopted in this paper due to its simplicity and practicality. Denoting $\mathcal{H}_v = \left\{ v^1, v^2, \ldots, v^m \right\}$ and $\mathcal{H}_u = \left\{ u^1, u^2, \ldots, u^n \right\}$ as two groups of query samples in feature space, an estimate of the MMD between $\mathcal{H}_v$ and $\mathcal{H}_u$ is defined as the squared difference between the empirical kernel mean mappings:

$$\mathcal{D}(\mathcal{H}_v, \mathcal{H}_u) = \left\| \frac{1}{m} \sum_{i=1}^{m} \kappa(v^i) - \frac{1}{n} \sum_{i=1}^{n} \kappa(u^i) \right\|^2, \tag{13}$$

where κ refers to the Gaussian kernel, which is widely used in practice.

For a certain category c, we want to ensure that the $\mathcal{H}_c^a$ (*anchor*) is closer to the $\mathcal{H}_c^p$ (*positive*) than it is to the distributions of any other categories $\mathcal{H}_c^n = \left\{ \left\{ \mathcal{H}_{i\neq c}^a, \mathcal{H}_{i\neq c}^p \right\} | i = 1, \ldots, N \right\}$ (*negative*). This is visualized in Fig. 4. Thus we want,

$$\mathcal{D}(\mathcal{H}_c^a, \mathcal{H}_c^p) + \alpha < \mathcal{D}(\mathcal{H}_c^a, \mathcal{H}_c^n), \tag{14}$$

where α adjusts the relative distance between *positive* and *negative* pairs.

In our case, there is only one *anchor* ($\mathcal{H}_c^a$) and one *positive* ($\mathcal{H}_c^p$) in each category. In order to ensure the fast convergence of the model, it is crucial to select an *hard negative* $argmin_{\tilde{\mathcal{H}}_c^n} \mathcal{D}(\mathcal{H}_c^a, \mathcal{H}_c^n)$ that violate the triplet constraint in (14), where $\tilde{\mathcal{H}}_c^n \in \mathcal{H}_c^n$. The hard-sample triplet loss can be defined as:

$$\mathcal{L}_{trp} = \sum_{c=1}^{N} [\mathcal{D}(\mathcal{H}_c^a, \mathcal{H}_c^p) - \mathcal{D}(\mathcal{H}_c^a, \tilde{\mathcal{H}}_c^n) + \alpha]_+ \tag{15}$$

E. Overall Objective

The overall framework of the HSL-MINet contains four losses, which are jointly optimized. For this purpose, a full objective can be interpreted as a weighted summation of the four losses:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \beta \mathcal{L}_{rot} + \gamma \mathcal{L}_{cons} + \lambda \mathcal{L}_{trp}, \tag{16}$$

where β, γ, and λ regulate the weight of respective losses, and their details will be explained later.

IV. EXPERIMENTS

In this part, multiple quantitative and qualitative experiments with two settings (*i.e.*, 5-way 1-shot scenario and 5-way 5-shot scenario) are executed. First, three public benchmarks and implementation details are described. The comparisons to the current leading techniques are then presented. Ablation experiments are also conducted to explore the ingredients that affect the performance of the HSL-MINet. Finally, the feature representations and experimental results are analyzed with visualization.

A. Datasets

The few-shot RS scene classification experiments are conducted on three publicly available datasets: NWPU-RESISC45 [2], WHU-RS19 [55], and UCM [56]. To assure the fairness of the comparative experiments, we follow the common splits and standard evaluation metrics as [3], [23], [50], and so on. The specifics are displayed in Table I.

1) NWPU-RESISC45: The large-scale benchmark is obtained from Google Earth, consisting of 31500 images in 45 distinct scenes, each covering 700 images. The size of each high-resolution RS image is 256×256 pixels, and the resolution of each pixel ranges from 0.2 to 30 meters.

TABLE I
STANDARD SPLITS ON THREE DATASETS

Dataset	Splitting of the dataset			Total categories
	Train	Val	Test	
NWPU-RESISC45	25	10	10	45
WHU-RS19	9	5	5	19
UCM	10	5	6	21

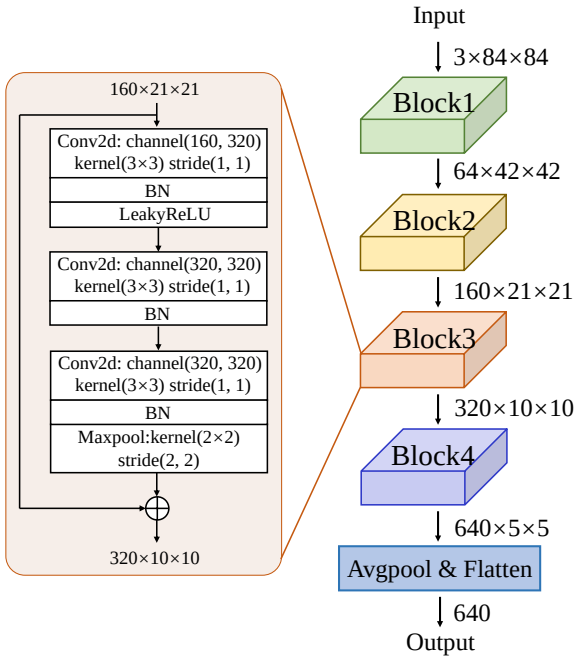


Fig. 5. Schematic diagram of the feature extractor, where the internal skeleton of the third residual block is exhibited as an example.

2) WHU-RS19: The scene classification benchmark focuses on urban areas in China and is released by Wuhan University. 1005 images in total diverge into 19 different scenes unevenly, and each category contains at least 50 pictures of 600×600 pixels in size.

3) UCM: The dataset consists of 21 scene categories, which is published by the UC Merced Computer Vision Laboratory in 2010. These images are sourced from National Map Urban Area Imagery. Each category of scenes includes 100 RS images, each with dimensions of 256 by 256 pixels and a 0.3-meter spatial resolution.

B. Implementation Details

1) Training Setup: The implementation of our model is done in PyTorch [57] and it runs on three NVIDIA GeForce 1080Ti GPUs. For any dataset, the input image is resized to 84×84, and the size of the raw feature vector output is 640. Data augmentations include four angles of rotation (*i.e.*, 0°, 90°, 180°, and 270°). The initial value of the learning rate is set to 3*e*-4, which is scaled by 0.5 every twenty epochs. The SGD optimizer is utilized with a momentum of 0.95 and a weight decay of 1*e*-4. The total number of epochs is fixed at 50 for these three datasets. Following the setting of the majority

TABLE II
5-WAY CLASSIFICATION ACCURACY (%) ON NWPU-RESISC45. THE METHODS WITH "*" ADOPT RESNET12 AS THE BACKBONE, WHICH IS CONSISTENT WITH OURS.

Type	Method	1-shot	5-shot
Optimization-based	MAML	48.40±0.82	62.90±0.69
	Meta-SGD*	60.63±0.90	75.75±0.65
	LLSR	51.43	72.90
Metric-based	MatchingNet*	64.41±0.86	76.33±0.65
	ProtoNet*	65.20±0.84	80.52±0.55
	RelationNet	60.04±0.85	80.39±0.56
	DLA-MatchNet	68.80±0.70	81.63±0.46
	SCL-MLNet	62.21±1.12	80.86±0.76
	SPNet	67.84±0.87	83.94±0.50
	TAE-Net	69.13±0.83	82.37±0.52
	PCFGNet	72.05±0.75	85.07±0.45
	GES-Net	70.83±0.85	82.27±0.55
	IDLN*	75.25±0.75	84.67±0.23
	MKN	65.84±0.89	82.67±0.55
	DANet*	74.30±0.20	87.29±0.11
	SGMNet*	73.01±0.77	84.52±0.50
	TSC*	73.26±0.15	84.62±0.35
	HSL-MINet(ours)	76.71±0.39	88.18±0.23

TABLE III
5-WAY CLASSIFICATION ACCURACY (%) ON WHU-RS19. THE METHODS WITH "*" ADOPT RESNET12 AS THE BACKBONE, WHICH IS CONSISTENT WITH OURS.

Type	Method	1-shot	5-shot
Optimization-based	MAML	49.13±0.65	62.49±0.51
	Meta-SGD	51.54±2.31	61.74±2.02
	LLSR	57.10	70.65
Metric-based	MatchingNet	67.68±0.67	85.01±0.38
	ProtoNet*	76.79±0.47	90.64±0.29
	RelationNet	65.01±0.72	79.75±0.32
	DLA-MatchNet	68.27±1.83	79.89±0.33
	SPNet	81.06±0.60	88.04±0.28
	TAE-Net	73.67±0.74	88.95±0.52
	PCFGNet	72.41±0.91	85.26±0.66
	GES-Net	75.84±0.78	82.37±0.38
	IDLN*	73.89±0.88	83.12±0.56
	DANet*	75.02±0.16	89.21±0.07
	SGMNet*	86.32±0.54	91.02±0.30
	TSC*	70.99±0.74	82.18±0.32
	HSL-MINet(ours)	87.33±0.37	91.22±0.14

of previous studies, there are 15 query samples per category for each episode. The values of temperature coefficient τ and margin α in the proposed hard-sample triplet loss will be analyzed in detail in the following experiments. To balance the four losses, the hyperparameters β, γ, and λ are assigned as 2.5, 1.5, and 1.0, respectively.

2) Evaluation: We organize the evaluation of 1-shot and 5-shot in the 5-way scenario. For the sake of fairness and convenience, we take a random sample of 2000 episodes from the test set and calculate the average accuracy with a 95% confidence interval as the final evaluation result.

3) Network Architecture: The overall framework contains three network structures, including a feature extractor with parameters θ, a rotation classifier with parameters φ, and a feature integration transformer with parameters Φ. For a fair comparison with published results, the proposed HSL-MINet adopts a widely-used feature extractor: ResNet-12 [58], which is not pretrained in advance. As shown in Fig. 5, the feature extractor contains four residual blocks. Every single residual block includes several combinations of Conv, BN, and ReLU layers. The number of the output channel of each block is 64, 160, 320, and 640, respectively. After a max-pooling layer, the feature extractor produces a 640-dimensions feature vector. The rotation classifier is instantiated by a single fully connected layer, which produces four predicted rotation scores (four rotations in this paper). The network structure of the feature integration transformer is encouraged by the standard multi-head attention mechanism.

TABLE IV
5-WAY CLASSIFICATION ACCURACY (%) ON UCM. THE METHODS WITH "*" ADOPT RESNET12 AS THE BACKBONE, WHICH IS CONSISTENT WITH OURS.

Type	Method	1-shot	5-shot
Optimization-based	MAML	48.86±0.74	60.78±0.62
	Meta-SGD	50.52±2.61	60.82±2.00
	LLSR	39.47	57.40
Metric-based	MatchingNet	48.18±0.75	67.39±0.50
	ProtoNet	53.85±0.78	71.23±0.48
	RelationNet	50.07±0.72	65.22±0.52
	DLA-MatchNet	53.76±0.62	63.01±0.51
	SCL-MLNet	51.37±0.79	68.09±0.92
	SPNet	57.64±0.73	73.52±0.51
	TAE-Net	60.21±0.72	77.44±0.51
	DANet*	61.60±0.18	78.62±0.13
	SGMNet*	64.17±0.75	76.63±0.59
	TSC*	55.11±0.68	69.20±0.64
	HSL-MINet(ours)	66.32±0.11	79.83±0.14

C. Comparison to Other State-of-the-Art Works

To investigate the effectiveness of the HSL-MINet, our model is evaluated against the current top-performing methods on the three benchmarks. Comparison approaches can be grouped into two categories: optimization-based and metric-based. The optimization-based methods are represented by MAML [42], Meta-SGD [59], and LLSR [60], while all the others fall under the metric-based category. The experimental results of these algorithms come from the corresponding public papers except for PCFGNet and GES-Net which are borrowed from [50]. To reflect the comparison results more intuitively, the results that rank first and second are displayed in red and green, respectively. In addition, the methods with "*" indicate that their backbones are consistent with ours (*i.e.*, ResNet12), otherwise, the ones without "*" adopt the consistent backbone with their original papers.

1) NWPU-RESISC45: Table II displays that the proposed HSL-MINet attains the highest performance in 1-shot and 5-shot scenarios. Specifically, HSL-MINet surpasses the existing SOTA (IDLN [61]) by about 1.46% in the 1-shot scenario, while having narrower confidence intervals. Compared with DANet [50], HSL-MINet produces 0.89% a greater accuracy in the 5-shot case. Consistent with our work, both IDLN and DANet use ResNet12 as the backbone, which means that the HSL-MINet makes fuller use of the feature extraction capabilities of the network structure.

TABLE V
5-WAY CLASSIFICATION ACCURACY (%) ON CROSS-DOMAIN SETTING ON WHU-RS19 → UCM.

Method	WHU-RS19 → UCM	
	1-shot	5-shot
MatchingNet	59.86±0.70	72.86±0.66
ProtoNet	62.41±0.73	77.81±0.41
RelationNet	60.45±0.76	74.98±0.47
HSL-MINet(ours)	**65.03±0.27**	**81.89±0.14**

TABLE VI
IMPACT OF DIFFERENT MULTIPLE VIEWS' NUMBERS ON NWPU-RESISC45 DATASET.

#	Rotations	NWPU-RESISC45	
		5-way 1-shot	5-way 5-shot
R-1	0°	65.20±0.84	80.52±0.55
R-2	0°, 90°	73.87±0.51	85.62±0.31
R-3	0°, 90°, 180°	76.36±0.44	87.11±0.35
R-4	0°, 90°, 180°, 270°	**76.71±0.39**	**88.18±0.23**
R-6	0°, 60°, 120°, 180°, 240°, 300°	76.12±0.32	87.77±0.28
R-8	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°	75.89±0.41	87.23±0.36

2) WHU-RS19: Table III shows the comparisons of different methods on the WHU-RS19 dataset. In both 5-way 1-shot and 5-way 5-shot settings, the proposed HSL-MINet attains the highest accuracies with an accuracy of 87.33% and 91.22%, respectively. The method that also performs well on this dataset is SGMNet, whose classification accuracy is slightly lower than our method. Both SGMNet and HSL-MINet outperform other methods by a large margin. Furthermore, an interesting observation is that the accuracies of IDLN and DANet have declined sharply from the NWPU-RESISC45 dataset to the smaller WHU-RS19 dataset, while HSL-MINet maintains the best performance on these two datasets. This shows that the proposed HSL-MINet has stronger robustness on different sizes of datasets.

3) UCM: Table IV compares the various techniques on the UCM dataset. From this report, we still find that the proposed HSL-MINet takes a significant lead. Specifically, the HSL-MINet is almost 2.15% and 1.21% more accurate than the previous SOTA (*i.e.*, SGMNet) in the scenario of 1 shot and 5 shots, respectively.

4) Comparative Results for Cross-set FSL: To evaluate our HSL-MINet under the cross-domain setting, we conduct experiments on WHU-RS19 → UCM. Specifically, the original training data of the WHU-RS19 dataset was used as the training set for 9 categories, and the validation and testing sets are the original splits of the UCM dataset. The 5-way 1/5-shot results are shown in Table V. Compared to the other three strong baseline algorithms, our method achieved the best performance while also having smaller performance variances. This indicates that the proposed model achieves effective cross-domain generalization capability and has promising potential for practical applications.

To summarize, the HSL-MINet obtains the best overall performance on the three publicly available benchmarks, which is a big step forward for the few-shot RS scene classification task.

TABLE VII
IMPACT OF DIFFERENT DISTANCE METRICS ON NWPU-RESISC45, WHU-RS19, AND UCM DATASETS.

Dataset	Metric	5-way 1-shot	5-way 5-shot
NWPU-RESISC45	Cosine	74.15±0.47	84.87±0.29
	l_2-Euclidean	74.69±0.33	85.25±0.31
	Euclidean	**76.71±0.39**	**88.18±0.23**
WHU-RS19	Cosine	85.07±0.36	89.63±0.13
	l_2-Euclidean	85.74±0.41	90.12±0.24
	Euclidean	**87.33±0.37**	**91.22±0.14**
UCM	Cosine	64.85±0.12	77.73±0.19
	l_2-Euclidean	65.28±0.13	78.29±0.22
	Euclidean	**66.32±0.11**	**79.83±0.14**

Different from other methods, ours can simultaneously achieve good performance on the three datasets, which validates the excellent generalization ability of our method. Moreover, the proposed HSL-MINet has a faster convergence rate in the training phase (less than 50 epochs), which implies a higher practical value.

D. Ablation Study

This subsection performs ablation studies to examine the impact of each loss, the number of labeled support samples, hyperparameters, the number of multiple views, and different distance metrics.

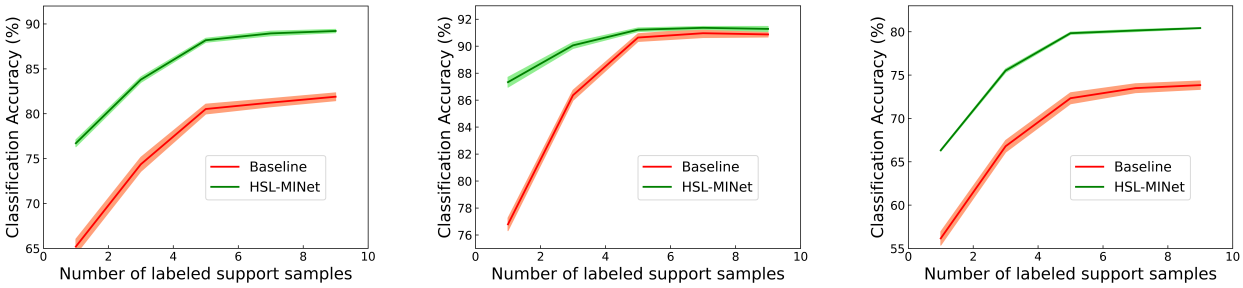
1) Ablation study of multiple views' number: We investigate the effect of multiple views' numbers on performance. In this respect, several extra different combinations of rotations are defined to generate multi-view images. "R-1" denotes a single view, only including 0° rotation, note that in this setting, the proposed model is equivalent to ProtoNet which is introduced as a baseline. "R-2" denotes two views, including 0° and 90° rotations; "R-3" denotes three views, including 0°, 90°, and 180° rotations; "R-4" denotes four views, including 0°, 90°, 180°, and 270° rotations; "R-6" denotes six views, including 0°, 60°, 120°, 180°, 240°, and 300° rotations; "R-8" denotes eight views, including 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315° rotations.

From Table VI, we can make the following analysis: (i) under the single-view setting, the proposed method's self-supervised task and multi-view integration cannot be performed, resulting in a significant drop in model performance. (ii) classification accuracy rises with the number of views when fewer than 4 views are used. We believe that this is because more views bring more self-supervised information to the model. (iii) Conversely, when the number of views exceeds 4, accuracy decreases inversely proportional to the quantity. It is possibly because of the low discrimination between too many views, which brings redundant information to the multi-view feature map. Additionally, when the rotation angle is not a multiple of 90° (*e.g.*, 45° or 60°), the rotated pixel information changes due to various interpolation operations, which may have a negative impact on multi-view integration. Thus, the four rotations setting is employed in this work.

2) Influence of different distance metrics: In this part, we evaluate the performance using Euclidean distance, l_2-

TABLE VIII
ABLATION STUDY OF EACH LOSS ON NWPU-RESISC45, WHU-RS19, AND UCM DATASETS.

Setting	NWPU-RESISC45		WHU-RS19		UCM	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Baseline	65.20±0.84	80.52±0.55	76.79±0.47	90.64±0.29	56.16±0.66	72.33±0.26
$+\mathcal{L}_{rot}$	67.21±0.77	81.96±0.61	77.55±0.37	90.69±0.58	57.33±0.42	74.67±0.31
$+\mathcal{L}_{rot}+\mathcal{L}_{cons}$	70.05±0.72	82.41±0.58	79.94±0.31	90.71±0.43	58.82±0.46	75.11±0.23
$+\mathcal{L}_{rot}+\mathcal{L}_{cons}+\mathcal{L}_{main}$	75.97±0.44	87.25±0.34	87.01±0.42	90.79±0.17	65.86±0.28	78.16±0.36
$+\mathcal{L}_{rot}+\mathcal{L}_{cons}+\mathcal{L}_{main}+\mathcal{L}_{trp}$	**76.71±0.39**	**88.18±0.23**	**87.33±0.37**	**91.22±0.14**	**66.32±0.11**	**79.83±0.14**



(a) NWPU-RESISC45 (b) WHU-RS19 (c) UCM

Fig. 6. Ablation results of varying the number of labeled support samples on the three datasets in a 5-way setting.

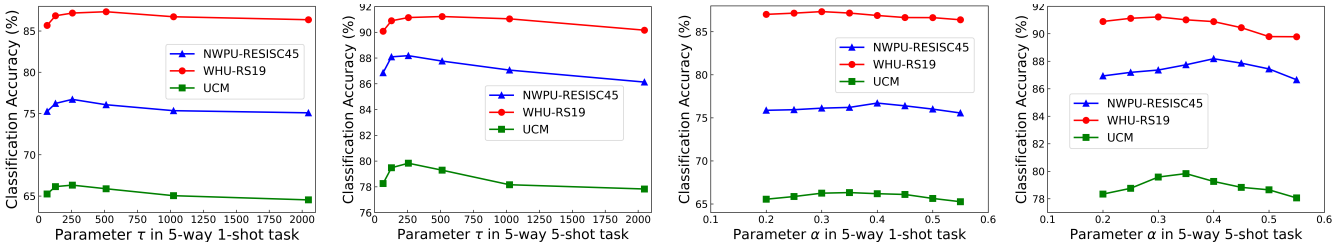


Fig. 7. The impact of hyperparameters in the HSL-MINet on NWPU-RESISC45 dataset.

Euclidean distance, and cosine distance for examining the influence of distance metrics on model classification performance. From Table VII, the accuracy is lower when using the l_2-Euclidean distance and cosine distance compared to the Euclidean distance. Different from Euclidean distance, what l_2-Euclidean distance and cosine distance have in common is that they both use normalization operations, which dilute local features and emphasize the globality of features. Therefore, it is possibly because more attention should be paid to local hard features in learning hard samples.

3) Ablation study of each loss: The full HSL-MINet model is optimized with four losses (see (16)), including the rotation prediction loss $\mathcal{L}_{rot}$, the multi-view consistency loss $\mathcal{L}_{cons}$, the multi-view main classification loss $\mathcal{L}_{main}$, and the hard-sample triplet loss $\mathcal{L}_{trp}$.

In Table VIII, we start with the ProtoNet [37] as a "Baseline"; "+$\mathcal{L}_{rot}$" means that adding $\mathcal{L}_{rot}$ to the "Baseline"; "+$\mathcal{L}_{rot}$+$\mathcal{L}_{cons}$" means that adding $\mathcal{L}_{cons}$ based on the previous step, and we select the classification result of the 0° rotation images (original images); Based on the previous setting, "+$\mathcal{L}_{rot}$+$\mathcal{L}_{cons}$+$\mathcal{L}_{main}$" adds the main classification loss $\mathcal{L}_{main}$; "+$\mathcal{L}_{rot}$+$\mathcal{L}_{cons}$+$\mathcal{L}_{main}$+$\mathcal{L}_{trp}$" means the full model. The performance of the model is seen to be continuously rising as more losses are utilized, indicating that each loss contributes positively to the ultimate performance.

4) Effect of the number of labeled samples on classification performance: With the ProtoNet [37] as a "Baseline", statistical experiments on the three datasets are conducted to analyze the impact of labeled support samples' numbers on accuracy performance. The quantity of support samples (*i.e.*, shot number) for each category is designated as 1, 3, 5, 7, and 9. From Fig. 6, we can get the following analysis results. First, classification accuracies of the baseline algorithm and proposed HSL-MINet consistently increase with the increasing shot number. This rationale is justifiable since prototypes calculated from a larger number of samples tend to be more class-discriminative. Second, for both of the two methods, the performance improvement is more pronounced when the number of shots increases from 1 to 5. This validates the importance of obtaining more discriminative features in situations where there are only a few labeled samples. Finally, the proposed HSL-MINet consistently outperforms the baseline method and has a higher upper bound on performance in few-shot scenarios.

5) Ablation study of hyperparameters: Fig. 7 shows the effects of different temperature coefficients τ and margins α on

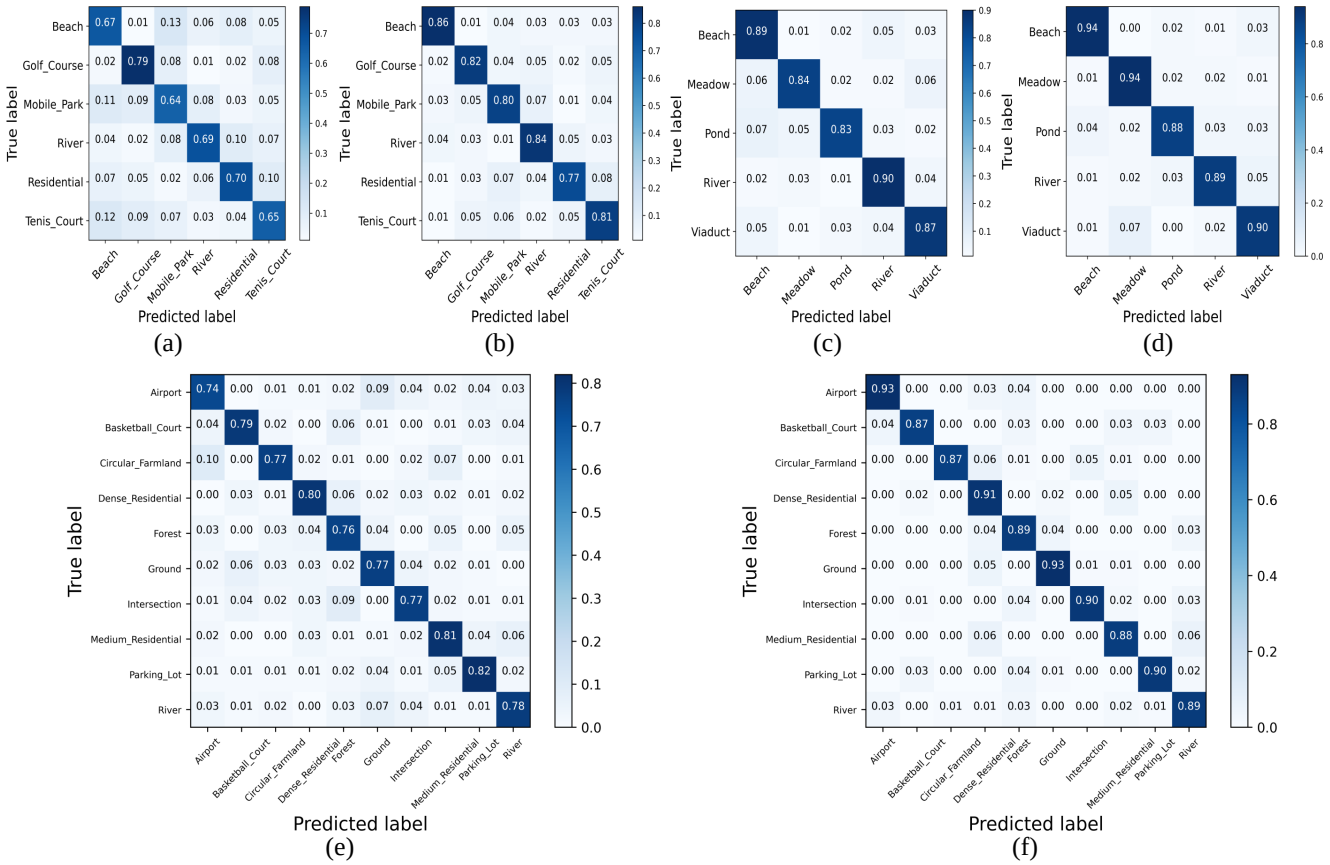


Fig. 8. The visualization of confusion matrices for the three datasets. (a), (c), and (e) present the results on the UCM, WHU-RS19, and NWPU-RESISC45 datasets in the 5-way 1-shot setting, respectively, and (b), (d), and (f) are the results on the UCM, WHU-RS19, and NWPU-RESISC45 datasets in the 5-way 5-shot setting, respectively.

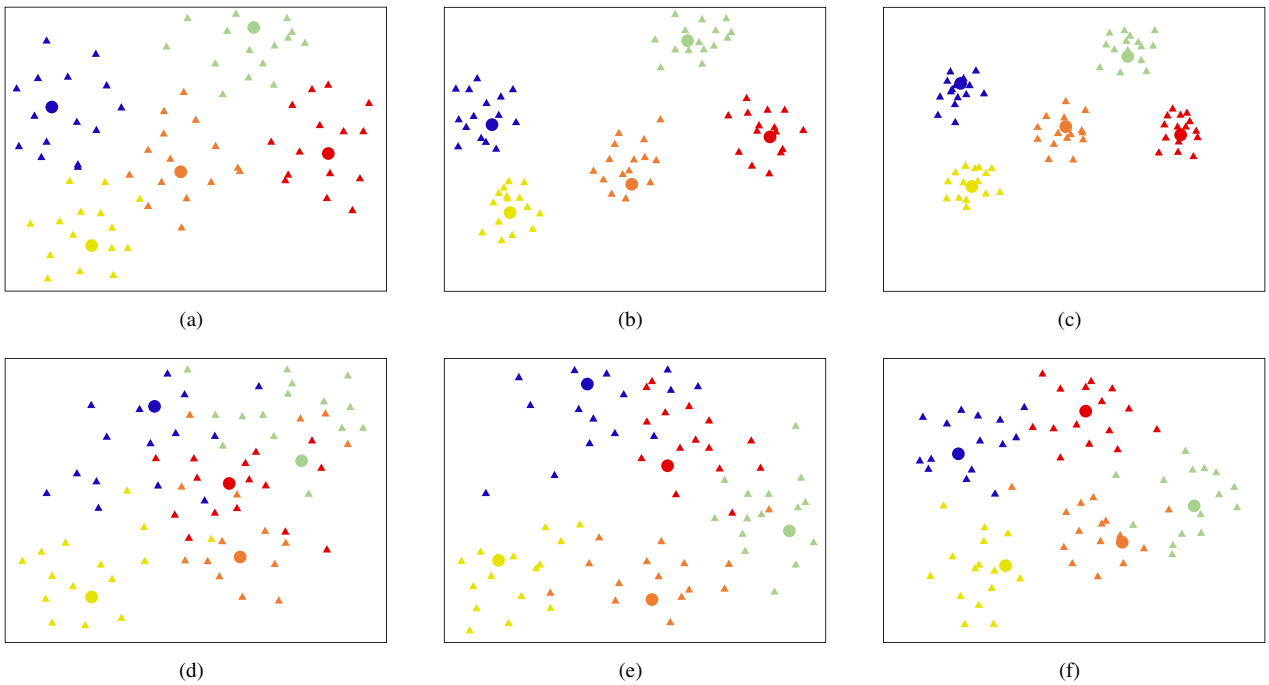


Fig. 9. T-SNE visualization of feature embeddings on NWPU-RESISC45 train (top row) and test (bottom row) dataset. (a) and (d) presents the raw feature distribution output by the backbone, (b) and (e) shows feature distribution applying MI (without HSL module), and (c) and (f) show feature distribution applying the full model (HSL-MINet). Symbols with different colors denote different category features, symbol ● represents features of query samples, and symbol ▲ indicates the prototype feature.

the three datasets. To reduce the difficulty of searching for optimal hyperparameters, when performing ablation experiments on a certain hyperparameter, we fix another hyperparameter to its optimal value. The optimal hyperparameter combinations for different datasets are inconsistent. For NWPU-RESISC45, the optimal combination is $\tau = 256$ and $\alpha = 0.4$. For WHU-RS19, the optimal combination is $\tau = 512$ and $\alpha = 0.3$. For UCM, the optimal combination is $\tau = 256$ and $\alpha = 0.35$.

E. Visualization analysis

1) Visualization of confusion matrix: In Fig. 8, the confusion matrices in distinct scenarios are visualized to intuitively exhibit the model's capability of distinguishing all categories in the test set. The i-th column of j-th row value in each confusion matrix shows the likelihood of a sample belonging to the i-th category being classified as the j-th category. As shown in Fig. 8, the probability of each category being correctly predicted is close to the overall average accuracy under the dataset, which verifies the robustness of the model in different category spaces.

2) Visualization of feature embedding: We construct a series of 5-way 1-shot 15-query experiments on the NWPU-RESISC45 dataset. The feature distributions are visualized by using t-Stochastic Neighbor Embedding (t-SNE) [62]. The first and second rows show the feature distribution of the training and testing data, respectively. The three settings from left to right columns are original distribution (output by the backbone), applying MI (without HSL module), and applying HSL-MINet (our complete model). As shown in Fig. 9, compared with the original distribution, the query feature of the same category is gathered to the corresponding support feature, and those of different categories are far away from each other when applying MI, which demonstrates the RI module has effectively mined class-discriminative features from multi-view images. Beyond this, when the full model (add the HSL module based on the MI module) is applied, the intra-class distribution becomes more compact, while the decision boundary between inter-classes becomes clearer. Furthermore, although the model's performance on unseen (test) data may decrease under the same conditions, our proposed MI and HSL modules effectively form more defined decision boundaries, indicating that the proposed methods have good generalization ability. Such a result verifies that our motivation is reasonable: (i) learning a more discriminative feature representation by exploring the spatial rotation insensitivity of RS images; (ii) optimizing the decision boundary of the model through hard samples learning.

V. Conclusion

In this study, a framework dubbed HSL-MINet is proposed for few-shot RS scene classification. Corresponding to the two aspects (*i.e.*, spatial rotation insensitivity and hard samples) of RS images, the HSL-MINet contains a Multi-view Integration (MI) module and a Hard Samples Learning (HSL) module. In the MI module, the pretext task is able to promote the model to learn transferable knowledge. Benefiting from the proposed multiview-attention mechanism, the multi-view feature integration task is verified to efficiently enhance the discrimination of the prototypes and the query features. In the HSL module, the feature distributions of hard samples are modified through a class-wise adaptive threshold strategy and the proposed hard-sample triplet loss in a more targeted way. Thus, the decision boundary becomes clearer and more discriminative, which is conducive to the generalization of the model. The proposed HSL-MINet is shown to be effective in experiments on three public benchmarks and also provides a reference for the research of few-shot RS images.

References

[1] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[2] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, 2017.

[3] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 143–154, 2022.

[4] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 82–96, 2019.

[5] J. Gao, Q. Wang, and Y. Yuan, "Feature-aware adaptation and structured density alignment for crowd counting in video surveillance," Dec 2019.

[6] J. Gao, T. Han, Y. Yuan, and Q. Wang, "Domain-adaptive crowd counting via high-quality image translation and density reconstruction," *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–13, Nov 2021. [Online]. Available: http://dx.doi.org/10.1109/tnnls.2021.3124272

[7] T. Zhang and X. Huang, "Monitoring of urban impervious surfaces using time series of high-resolution remote sensing images in rapidly urbanized areas: A case study of shenzhen," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2692–2708, 2018.

[8] J. Gao, Y. Yuan, and Q. Wang, "Feature-aware adaptation and density alignment for crowd counting in video surveillance," *IEEE transactions on cybernetics*, vol. 51, no. 10, pp. 4822–4833, 2020.

[9] 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.

[10] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning." *computer vision and pattern recognition*, 2018.

[11] 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*, 2019.

[12] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, and V. Gripon, "Easy: Ensemble augmented-shot y-shaped learning: State-of-the-art few-shot classification with simple ingredients," 2022.

[13] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 836–10 846.

[14] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *neural information processing systems*, 2016.

[15] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," *computer vision and pattern recognition*, 2017.

[16] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," *computer vision and pattern recognition*, 2017.

[17] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," *computer vision and pattern recognition*, 2019.

[18] Y. Chen, X. Wang, Z. Liu, H. Xu, and T. Darrell, "A new meta-baseline for few-shot learning," 2020.

[19] M. Nayeem Rizve, S. Khan, F. Shahbaz Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," *arXiv e-prints*, pp. arXiv–2103, 2021.

[20] B. N. Oreshkin, P. Rodríguez, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *arXiv: Learning*, 2018.

[21] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[22] D. Kang, H. Kwon, J. Min, and M. Cho, "Relational embedding for few-shot classification." *international conference on computer vision*, 2021.

[23] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "Dla-matchnet for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[24] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *international conference on learning representations*, 2018.

[25] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *systems man and cybernetics*, 1973.

[26] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[28] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *european conference on computer vision*, 2010.

[29] J. Chen, C. Wang, and R. Wang, "Using stacked generalization to combine svms in magnitude and shape feature spaces for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, 2009.

[30] B. Niu, Z. Pan, J. Wu, Y. Hu, and B. Lei, "Multi-representation dynamic adaptation network for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[31] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains," *computer vision and pattern recognition*, 2015.

[32] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification." *IEEE Transactions on Neural Networks*, 2020.

[33] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

[34] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[35] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, "Iept: Instance-level and episode-level pretext tasks for few-shot learning," *Learning*, 2021.

[36] H.-J. Ye, L. Ming, D.-C. Zhan, and W.-L. Chao, "Few-shot learning with a strong teacher." *arXiv: Computer Vision and Pattern Recognition*, 2021.

[37] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *neural information processing systems*, 2017.

[38] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *Learning*, 2019.

[39] H. Jung and S.-W. Lee, "Few-shot learning with geometric constraints," *IEEE Transactions on Neural Networks*, 2020.

[40] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," *computer vision and pattern recognition*, 2020.

[41] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," *Learning*, 2020.

[42] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *international conference on machine learning*, 2017.

[43] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler, "Rapid adaptation with conditionally shifted neurons," *arXiv: Learning*, 2017.

[44] E. Park and J. B. Oliva, "Meta-curvature," *neural information processing systems*, 2019.

[45] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," *computer vision and pattern recognition*, 2019.

[46] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," *Learning*, 2021.

[47] S. W. Yoon, J. Seo, D. Kim, and J. Moon, "Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning," *international conference on machine learning*, 2020.

[48] H. Li, C. Zhenqi, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification," *arXiv: Computer Vision and Pattern Recognition*, 2020.

[49] G. Cheng, L. Cai, C. Lang, X. Yao, J. Chen, L. Guo, and J. Han, "Spnet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[50] M. Gong, J. Li, Y. Zhang, Y. Wu, and M. Zhang, "Two-path aggregation attention network with quad-patch data augmentation for few-shot scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[51] B. Zhang, S. Feng, X. Li, Y. Ye, and R. Ye, "Sgmnet: Scene graph matching network for few-shot remote sensing scene classification," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[52] Y. Liu, L. Zhang, Z. Han, and C. Chen, "Integrating knowledge distillation with learning to rank for few-shot scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[53] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and Łukasz Kaiser, "Attention is all you need," 2022.

[54] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *intelligent systems in molecular biology*, 2006.

[55] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International Journal of Remote Sensing*, 2012.

[56] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," *advances in geographic information systems*, 2010.

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv: Computer Vision and Pattern Recognition*, 2015.

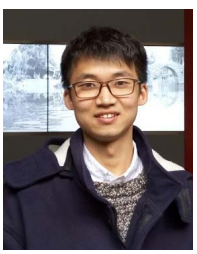
[59] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few shot learning." *arXiv: Learning*, 2017.

[60] M. Zhai, H. Liu, and F. Sun, "Lifelong learning for scene recognition in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2019.

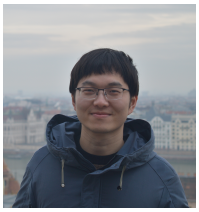
[61] Q. Zeng, J. Geng, W. Jiang, K. Huang, and Z. Wang, "Idln: Iterative distribution learning network for few-shot remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[62] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008.

Yuyu Jia received the B.E. degree and the M.S. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include few-shot learning, deep learning, and remote sensing.



Junyu Gao received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021, respectively. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Wei Huang received the B.E. degree in control theory and engineering and M.E. degree in computer science and technology in the School of Artificial Intelligence, Optics and Electronics (iOPEN) from the Northwestern Polytechnical University, Xi'an, China, in 2018 and 2021, respectively. He is pursuing his Ph.D. degree at Technical University of Munich. His research interests include transfer learning, deep learning, and remote sensing.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.