Multigranularity Decoupling Network With Pseudolabel Selection for Remote Sensing Image Scene Classification

Wang Miao, Jie Geng, *Member, IEEE*, and Wen Jiang

***Abstract*— The existing deep networks have shown excellent performance in remote sensing scene classification (RSSC), which generally requires a large amount of class-balanced training samples. However, deep networks will result in underfitting with imbalanced training samples since they can easily bias toward the majority classes. To address these problems, a multigranularity decoupling network (MGDNet) is proposed for remote sensing image scene classification. To begin with, we design a multigranularity complementary feature representation (MGCFR) method to extract fine-grained features from remote sensing images, which utilizes region-level supervision to guide the attention of the decoupling network. Second, a class-imbalanced pseudolabel selection (CIPS) approach is proposed to evaluate the credibility of unlabeled samples. Finally, the diversity component feature (DCF) loss function is developed to force the local features to be more discriminative. Our model performs satisfactorily on three public datasets: UC Merced (UCM), NWPU-RESISC45, and Aerial Image Dataset (AID). Experimental results show that the proposed model yields superior performance compared with other state-of-the-art methods.**

***Index Terms*— Imbalanced learning (IL), remote sensing image, scene classification, semisupervised learning (SSL).**

I. Introduction

With the remarkable development of remote sensing technology, the application of high-resolution satellite images has expanded to a wide range of fields [1], [2], [3], [4], such as remote sensing scene classification (RSSC), urban planning, environmental monitoring, land cover determination, vegetation mapping etc [5], [6], [7]. High-resolution remote sensing (HRRS) image scene classification has drawn broad attention, which categorizes HRRS images into a set of land use and land cover classes [1].

Recently, a significant number of deep learning approaches have been studied to improve the performance of RSSC [8]. Specifically, Lu et al. [5] presented a feature aggregation convolutional neural network (FACNN) with a progressive aggregation strategy, which aims to learn excellent representation via leveraging semantic label information. Ammour [9] utilized a continual learning model using remote sensing images data regeneration to overcome the shortcoming of catastrophic forgetting. Jun et al. [10] developed a skip-linked covariance network to extract representative features, which embeds differential pooling modules into the deep network. Bai et al. [11] considered the fine-grained and high-order information of the remote dataset simultaneously and applied it to scene classification problems. Xu et al. [12] provided a lie group regional influence network, which takes advantage of lie group feature learning for RSSC. Wang et al. [13] designed a multigranularity canonical appearance pooling (MG-CAP) to extract fine-grained features of remote sensing images automatically. Ma et al. [14] proposed a progressive generative adversarial network, which can generate labeled samples with spatial information. In order to locate the most important area using only image-level labels, a weakly supervised key area detection strategy of structured key area localization (SKAL) is specially designed to connect the global features and local features from the whole image [15].

Current scene classification algorithms primarily focus on learning from a balanced remote dataset with a large number of labeled samples [1], [16], where different classes are evenly distributed. However, as shown in Fig. 1, in real-world applications, remote training samples often exhibit a class-imbalanced distribution, where a small portion of classes have massive samples while the others have only a few samples [1], [17], [18], [19]. With imbalanced training samples, networks trained on a class-imbalanced remote dataset are biased toward the majority classes [20], [21]. To address the problem of class imbalance, there have been some imbalanced learning (IL) methods in the field of computer vision [22], which can effectively alleviate performance degradation due to class imbalance [19]. Zhang et al. [19] defined the imbalance factor as the ratio between the numbers of the most frequent and the least frequent classes, which is capable of reflecting the degree of class imbalance. In addition, there are also some research works that combine the lack of labeling information and class imbalance [23], [24]. However, current methods of RSSC do not take into account the class imbalance.

Meanwhile, it is challenging to obtain sufficient training samples [1], [25], [26], [27], [28], [29]. A deeper neural network may overfit due to insufficient training samples,

Manuscript received 1 July 2022; revised 17 November 2022; accepted 4 February 2023. Date of publication 13 February 2023; date of current version 27 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3900502, in part by the National Natural Science Foundation of China under Grant 62271396, and in part by the Shaanxi Key Research and Development Program under Grant 2023-YBGY-220. *(Corresponding author: Wen Jiang.)*

The authors are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: mw0638@mail.nwpu.edu.cn; gengjie@nwpu.edu.cn; jiangwen@nwpu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3244565

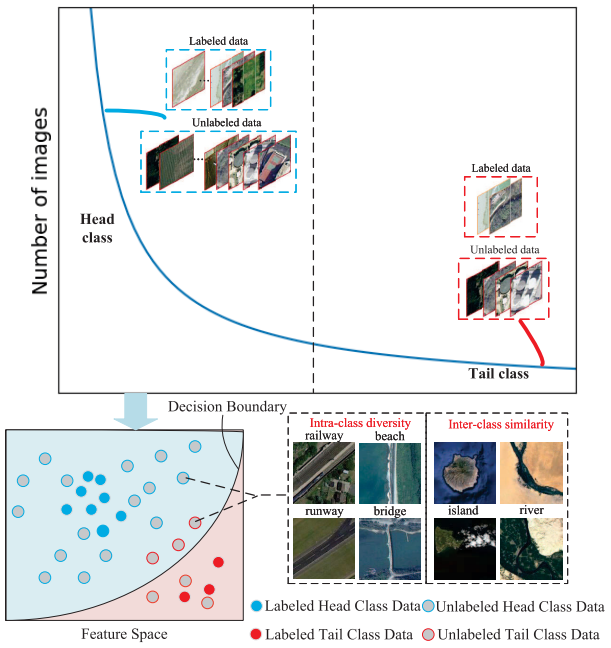


Fig. 1. Real-world remote datasets typically exhibit the phenomenon of class-imbalanced distributions. The extreme long-tailed distribution causes tremendous challenges to the scene classification task.

in which case the trained model performs poorly on test data with a similar distribution. To tackle the above issue of insufficient annotated training samples, semisupervised learning (SSL) methods for scene classification have been studied [1], [25], [30], [31]. Yao et al. [30] employed a self-stepping pseudosamples selection approach with local manifold constraints, which can select easy-training samples for scene classification. Dai et al. [31] introduced an integrated learning strategy to solve the semisupervised scene classification problems. Wei et al. [32] proposed a semisupervised generation framework (SSGF), which integrates deep learning features, self-labeling technology, and a discriminant evaluation approach for scene categorization and dataset updating. It has a strong ability to learn from unlabeled samples and increase classification precision. Miao et al. [25] developed the semisupervised representation consistency Siamese network (SS-RCSN) to reduce average differences of features between the labeled and unlabeled data, which can address the issue of overfitting with limited training samples, and it produces excellent results for scene classification [25].

Although these IL and SSL methods achieve excellent performance, there are setbacks for remote sensing image scene classification. Due to the interclass and intraclass characteristics of remote sensing images, the task is substantially more difficult than general imbalanced and semisupervised scene classification under such circumstances. When the deep model is well-trained with a significant amount of balanced and labeled data, it can learn effective feature representations, resulting in consistent feature expression across samples of the same category. However, there are a large number of imbalanced samples with a lack of annotation utilized for scene classification in the real-world, which severely affects the classification accuracy. When the deep network is insufficiently trained with imbalanced data, the pseudolabels predicted are highly ambiguous due to the poor adaptability of the network. Therefore, the key to improving the performance of the model is obtaining high-confidence pseudolabels with imbalanced unlabeled data.

In this article, we introduce a task of imbalanced SSL (ISSL) for remote sensing image scene classification. Our proposed network aims to improve the classification performance of imbalanced data with less labeled data. Specifically, we propose an end-to-end multigranularity decoupling network (MGDNet) for remote sensing image scene classification. The proposed model mainly consists of three parts: the multigranularity complementary feature representation (MGCFR) method, class-imbalanced pseudolabel selection (CIPS) approach, and diversity component feature (DCF) loss function. The main contributions of our work can be summarized as follows.

1) The MGDNet is proposed for class-imbalanced scene classification, which can address the issue of underfitting with imbalanced training samples. In the proposed MGDNet, we construct a bilateral-branch decoupling network for classifier learning and representation learning, which apply uniform and reversed samplers to each of them separately. The learning focus of the proposed decoupling network is able to shift from representations to classifiers gradually. Based on the bilateral-branch network (BBN), we perform multigranularity feature learning and high-confidence pseudolabel metrics to improve the accuracy of RSSC.
2) To classify and distinguish the subtle differences among categories with similar appearances, the MGCFR method and the DCF loss function are developed to improve the performance of MGDNet. The combination of the MGCFR method and the DCF loss function drives the fine-grained features to be more representative, which can utilize region-level supervision to guide the attention of the decoupling network. The DCF loss function mainly weights the loss corresponding to the samples according to the difficulty of sample discrimination.
3) To enhance learning capacity for imbalanced unlabeled data, the CIPS method is proposed to estimate the confidence of pseudolabel according to the loss distribution component of unlabeled data. We dynamically fit a Gaussian mixture model (GMM) for imbalanced data on its per-sample loss distribution, which can select high confidence pseudolabels set.

The rest of our work is arranged as follows. The related works are briefly introduced in Section II. The proposed MGDNet is detailed in Section III. The comparative results and experimental analyses are presented in Section IV. Conclusions are finally summarized in Section V.

II. Related Work

Numerous SSL and IL studies have been conducted in recent years [33]. In this section, we will briefly introduce the current research work.

A. Imbalanced Learning

IL methods have been extensively studied in many fields over the past decade [22], which can effectively alleviate performance degradation due to class imbalance [19]. The current IL methods are divided into three categories: rebalancing methods, information augmentation methods, and module improvement methods [19], [34].

To begin with, rebalancing methods mostly balance the class distribution in training via changing the sampling probability [35]. One of the commonly used methods in rebalancing methods is oversampling [35]. While oversampling methods can greatly boost the learning capacity for tail class samples, they also come with the potential risk of overfitting [19]. Contrary to oversampling methods, the basic notion of under-sampling [36] is to remove head class samples, which can enable the model to avoid overfitting. However, these approaches are not practicable in some datasets with huge imbalance ratios. Second, information augmentation methods bring additional data structure information into the training, which enables the network to adapt the data distribution in the class-imbalanced classification task [37], [38], [39]. Finally, researchers also explored network module methods in long-tailed learning. For example, Kang et al. [40] proposed a decoupling training schema, which first learns the representations and classifier jointly, then balances the classifier with the resampling method. Zhou et al. [17] also presented a unified BBN based on the decoupling mechanism. Kini et al. [41] formulated the vector-scaling (VS) loss, which incorporates current algorithms and achieves excellent performance on the long-tail dataset.

B. Semisupervised Learning

SSL methods provide a way to learn from unlabeled data, which also have wide applications in many fields [42], [43], [44], [45], [46]. For example, Berthelot et al. [47] developed the MixMatch and ReMixMatch algorithm by introducing mix-up, distribution alignment, and data augmentation. Sohn et al. [48] combined the weak and strong data augmentation techniques to improve the model's learning ability for unlabeled data.

In addition, there have been some studies on ISSL. Specifically, Wei et al. [23] proposed a semisupervised self-training framework, which alleviates class imbalance by selecting pseudolabeled data for the minority classes. Kim et al. [49] designed an iterative algorithm, named distribution aligning refinery of pseudolabel (DARP), to softly refine the pseudolabels predicted from a biased network. Lee et al. [24] presented an auxiliary balanced classifier (ABC) of the representation layer, which can effectively learn unlabeled data in ISSL.

C. Remote Sensing Image Scene Classification

Remote sensing image scene classification is becoming increasingly crucial for interpreting remote sensing images [1]. There has been a lot of remarkable work in some more recent times. For RSSC, Tang et al. [50] suggest a brand-new approach called efficient multiscale transformer and cross-level attention learning (EMTCAL). EMTCAL combines the benefits of CNN and Transformer to completely exploit the information within remote sensing images. In order to extract relevant high-frequency remote sensing image features using a trainable Laplacian operator, Zhang et al. [51] present a Laplacian high-frequency convolutional block (LHCB) based on CNN. Lv et al. [52] propose a spatial-channel feature-preserving Vision Transformer (ViT) model, which considers both the detailed geometric information of the high-spatial-resolution (HSR) imagery and the contribution of the different channels contained in the classification token. To address the issue of few-shot scene classification from both the data and architectural perspectives, Gong et al. [53] introduce a two-path aggregation attention network with patch augmentation, known as the data architecture network (DANet). Xu et al. [54] propose an end-to-end scene classification method by employing the ViT as an outstanding instructor for directing small networks. Although the latest methods have achieved better results in RSSC, there is still no scene classification methods that consider the case of insufficient annotation and class-imbalanced data.

III. Methodology

A. Problem Formulation

The goal of this article is to learn a network with a few class-imbalanced labeled samples. Specifically, assume the remote sensing data as $V = \{v_1, \ldots, v_t, \ldots, v_s\}$, which can be divided into labeled data $X = \{(x_1, p_1), \ldots, (x_i, p_i), \ldots, (x_n, p_n)\}$ and unlabeled data $U = \{u_1, \ldots, u_j, \ldots, u_m\}$, p represent the label of labeled sample and each class have different numbers.

B. Overview

To improve the performance of imbalanced remote sensing image scene classification, we propose the MGDNet. It consists of three parts: the MGCFR method, the CIPS approach, and the DCF loss function, which are shown in Fig. 2.

As shown in Fig. 2, considering intraclass diversity and interclass similarity in remote sensing images, the MGCFR framework is developed, which has the ability to find discriminate local regions that correspond to subtle visual traits. Then, the CIPS approach for unlabeled samples is proposed to metric the confidence of pseudolabels. Finally, the DCF loss function is utilized to focus on the most obvious distinctions between classes.

Concretely, we construct a bilateral-branch decoupling network F for classifier learning and representation learning, which apply uniform and reversed samplers to each of them separately. In this section, the decoupling network divides a deep classification model into two essential parts: 1) the feature extractor and 2) the classifier. As a result, representation learning and classifier learning might be distinguished as different stages of the deep classification network learning process. We develop a two-stage methodology to separately learn deep model representations and classifiers.

The learning focus of the proposed decoupling network is able to shift from representations to classifiers gradually.

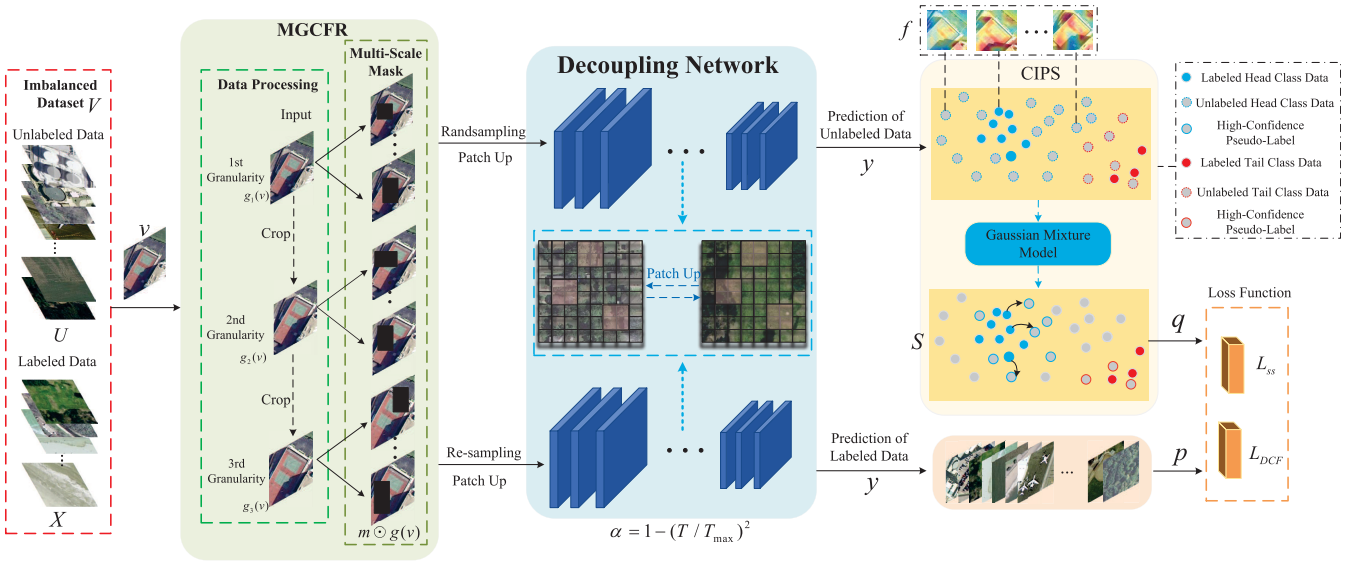


Fig. 2. Overall framework of the proposed MGDNet. We use three different granularities from coarse to fine, which can help improve the performance of scene classification by making it easier to obtain more significant and vivid differences.

The output logits of the BBN F are formulated as

$$z = \alpha W_{F_1}^{\mathrm{T}} v_e + (1-\alpha) W_{F_2}^{\mathrm{T}} v_e \tag{1}$$

$$\alpha = 1 - (T/T_{\max})^2 \tag{2}$$

where z are the output logits, W_F^{T} represents the parameters of the network, the number of total training epochs is described as $T_{\max}$, T is the current epoch, and α will gradually decrease as the training epochs increasing. Based on the BBN, we perform multigranularity feature learning and high-confidence pseudolabel metrics to improve the accuracy of RSSC.

C. Multigranularity Complementary Feature Representation

Many remote sensing images, including class "airport," "playground," "desert," and so on, do not clearly separate foreground from background, so the classification cues for these images usually focus on fine-grained features at the small block level. If the model focuses only on the salient regions, it is likely to ignore the fine-grained information. Thus, we need to jointly consider the salient and relevant regions of each remote sensing image to extract fine-grained information. To classify and distinguish the subtle differences among categories with similar appearances, particularly regarding long-distance links, the classification models should take into account local parts and their contextual linkages.

To address the above issue, we propose the MGCFR method to disperse the model's attention to the local parts of the remote sensing image. An illustration of the representation method is shown in Fig. 2. The core idea of the proposed representation method is to granularly learn multiple fine-grained features in imbalanced samples. For the remote sensing images, we use three different granularities from coarse to fine. In the proposed MGCFR, V stands for the unified representation of the input data, including labeled data X and unlabeled data U. Specifically, as shown in Fig. 2, the labeled input data x are cropped as $g_1(x)$, $g_2(x)$ and $g_3(x)$. The unlabeled data u are cropped as $g_1(u)$, $g_2(u)$ and $g_3(u)$. Meanwhile, we apply masks of diverse scales to images of different granularities. Multiscale masks $m_1()$, $m_2()$, and $m_3()$ are randomly generated, and each image has a different mask, which can erase different local region that covers more subtle interclass differences. Different granularity images of labeled and unlabeled data with different scale masks are input to the MGDNet.

Then, PatchUp is applied both on the different granularity samples with the multiscale masks, which can improve the stability of local feature extraction [55]. In the proposed MGDNet, labeled and unlabeled data are patched up between different classes. Patch up exchanges or interpolates the cut blocks of the middle hidden layer of two samples [55]. By simple linear transformation of the input data, the generalization ability of the model can be increased, and the robustness of the network to obtain fine-grained information can be improved [55], [56]. The operation is divided into two modes: soft operation ψ_{soft} and hard operation ψ_{hard} [55]. The input of class-imbalanced remote sensing data $\psi_{\text{hard}}(g_k(v_i), g_k(v_j))$ is formulated as

$$\psi_{\text{hard}}(g_k(x_i), g_k(x_j)) = \varepsilon \odot m_k \odot g_k(x_i) + (1-\varepsilon) \odot m_k \odot g_k(x_j) \tag{3}$$

$$\psi_{\text{hard}}(g_k(u_i), g_k(u_j)) = \varepsilon \odot m_k \odot g_k(u_i) + (1-\varepsilon) \odot m_k \odot g_k(u_j) \tag{4}$$

$$\psi_{\text{soft}}(g_k(x_i), g_k(x_j)) = \varepsilon \odot m_k \odot g_k(x_i) + \text{Mix}[((1-\varepsilon) \odot m_k \odot g_k(x_i)), ((1-\varepsilon) \odot m_k \odot g_k(x_j))] \tag{5}$$

$$\psi_{\text{soft}}(g_k(u_i), g_k(u_j)) = \varepsilon \odot m_k \odot g_k(u_i) + \text{Mix}[((1-\varepsilon) \odot m_k \odot g_k(u_i)), ((1-\varepsilon) \odot m_k \odot g_k(u_j))] \tag{6}$$

where ε is a binary mask, $\odot$ is the elementwise multiplication operation, $\lambda \tilde{}\text{Beta}(\eta, \eta)$, and η can control the shape of the

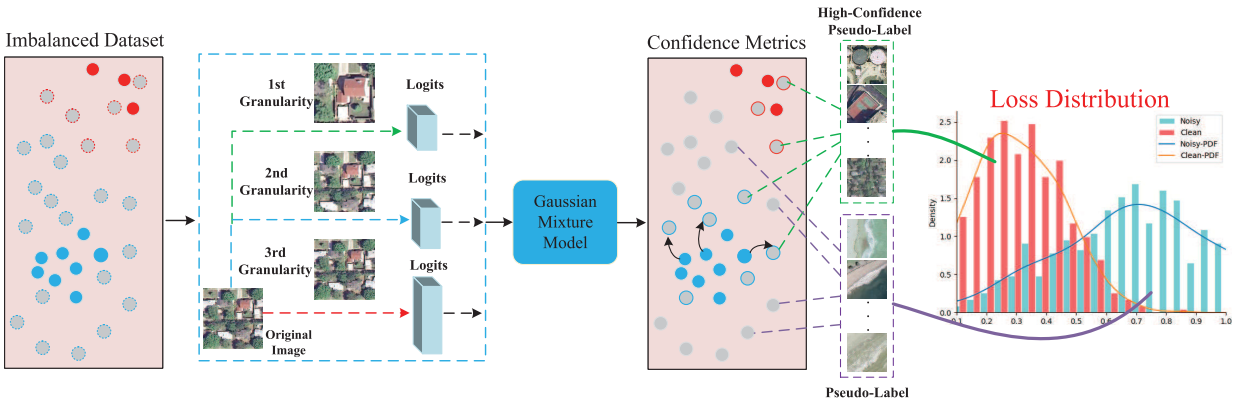


Fig. 3. Processing of selecting with the unlabeled set. The points with different colors represent unlabeled samples with different classes, where the circled points indicate selected samples by CIPS.

Beta distribution, Mix[·] represents the weighted mixture of data [55], [56].

During the training process, the mask ε penalizes the most discriminative part for inducing the model to cover the integral extent of the remote sensing image. The fine-grained feature map f by the proposed BBN is formulated as

$$f_X = F_1(\psi(g_k(x_i), g_k(x_j))) \tag{7}$$

$$f_U = F_2(\psi(g_k(u_i), g_k(u_j))). \tag{8}$$

After extracting features, we can obtain the pseudolabels of remote sensing data through the forward prediction results. High-quality pseudolabels can enable the proposed model to acquire finer-grained information and perform class-IL (more efficiently.

D. Class-Imbalanced Pseudolabel Selection

There are a large number of imbalanced samples with a lack of annotation utilized for scene classification in the real-world, which severely affects the classification accuracy. When the deep model is well trained with a significant amount of labeled data, it can learn effective feature representations, resulting in consistent feature expression across samples of the same category. Nevertheless, when the deep network is insufficiently trained with imbalanced data, the pseudolabels predicted are inaccurate due to the poor adaptability of the network. Therefore, the key to improving the performance of the model is obtaining high-confidence pseudolabels. Our proposed network aims to improve the classification performance of imbalanced data with less labeled data. If the labeled data information is complete, it is easy to study the data manifold in imbalanced data for the deep model.

Deep networks exhibit differing fitting capacities for clean label samples and mistaken label samples, leading to a lower loss for clean label samples [57], [58], [59]. GMM can better distinguish clean and mistaken label samples due to its flexibility in the sharpness of distribution [60], [61], [62]. At the same time, when the network does not fully fit the mistaken label data in the training samples, the loss value of the correct label sample is typically small, while the loss value of the mistaken label sample is generally too large [58].

Inspired by this, we propose a pseudolabel metric method to select high-confidence pseudolabels. The set of loss values for correct label samples can be regarded as a quasi-normal distribution. Similarly, the set of mistaken label sample loss values is also a quasi-normal distribution, and the mean and variance of this normal distribution are unknown, which are connected to the parameters and training of the decoupling network.

As shown in Fig. 3, for class-imbalanced labeled samples X, the label set of forwarding propagation y_i is defined as follows:

$$y_i = e_i^{f_X} \Big/ \sum_{j=1}^{C} e_j^{f_X}. \tag{9}$$

For class-imbalanced unlabeled samples U, the multigranularity pseudolabels set of forwarding propagation q_i is defined as follows:

$$q_i = e_i^{f_U} \Big/ \sum_{j=1}^{C} e_j^{f_U}. \tag{10}$$

We dynamically fit a GMM for imbalanced data on its per-sample loss distribution, which can select high confidence pseudolabels set. The entropy l of the proposed bilateral-branch decoupling networks prediction for the class-imbalanced samples is defined as

$$H = \sum_{c=1}^{C} H(v, (v_i; W)) \tag{11}$$

$$l = \alpha H_{F1} + (1-\alpha) H_{F2} \tag{12}$$

where C is the number of classes, and $H(\cdot)$ is the cross-entropy function. H_{F1} and H_{F2} are the training losses of the proposed BBN.

The GMM is thus expressed as

$$P_m(l) = \sum_{k=1}^{K} w_k \mathrm{GMM}(l|\mu_k, \Sigma_k) \tag{13}$$

where μ is the mean of the Gaussian component, $P_m(l)$ is the prediction probability of GMM. Σ is the variance of the Gaussian component, K is the number of components in the GMM, and w_k is the weight of the Gaussian component.

The sample with high posterior probability is more likely to own the clean pseudolabel and can be selected as high confidence data for expanding the training set. The fitted GMM will metric its confidence according to the loss value, and we will select the pseudolabel by setting a threshold

$$S = \left\{(X^i, p^i)\right\}_{i=1}^{|X|} \cup \left\{(U^j, q^j)\right\}_{j=1}^{|U, P_m(l) > P_{\text{th}}|} \tag{14}$$

where P_{th} is the threshold in the GMM, p represent the label of labeled sample.

E. Diversity Component Feature Loss Function

The key to reducing the intraclass diversity as well as the interclass similarity is finding local regions that correspond to subtle visual traits. To improve the performance of the proposed model for scene classification, our loss function consists of two parts: the DCF loss function and the semisupervised loss function. Here, we propose a DCF loss function to force the decoupling network to learn fine-grained features and find the subtle differences.

In class imbalance learning, the learning difficulty of the samples during training is different. The DCF loss function can deal with the classification of imbalanced samples. It mainly weights the loss corresponding to the samples according to the difficulty of sample discrimination, which means adding a smaller weight to the samples that are easy to discriminate and the samples that are difficult to discriminate. The easy-to-classify head class samples dominate the training process. Therefore, the proposed loss function is reshaped to emphasize the challenging samples while reducing the weight of the easy samples. We propose to add a conditioning factor μ to the proposed loss and define the loss as follows:

$$L_{\text{DCF}} = \mu \cdot \sum_{i}^{S} \sum_{j}^{S} \frac{|f_i \cap f_j|}{|f_i| + |f_j|} \tag{15}$$

$$\mu = (1 - P_{F2} - P_{F1} + P_{F1} P_{F2})^{\gamma} \tag{16}$$

where P_F denotes the predicted probability of the sample, and γ denotes the hyperparameter. $|f_i \cap f_j|$ is the intersection between f_i and f_j, $|f_i|$ and $|f_j|$ represent the number of elements of f_i and f_j. The idea of this part is to make the features more similar, the greater the loss. The DCF Loss function forces all regions of the same class to be representative. Combined with the proposed multigranularity complementary feature extraction framework, the extracted fine-grained features can be more discriminative.

Furthermore, in order to improve the generalization ability of the proposed decoupling network, the cross-entropy loss function is applied to the labeled data and unlabeled data. For the high confidence dataset S, the semisupervised loss function is given as follows:

$$L_{\text{ss}} = \frac{1}{\rho} \sum_{x, p \in S} H(p, P(y/x; W)) + \frac{1}{\tau} \sum_{u, q \in S} H(q, P(y/u; W)) \tag{17}$$

where x and p represent the labeled samples and labels, u and q stand for the unlabeled data and high confidence pseudolabels, respectively. ρ denotes the number of labeled samples, and τ denotes the number of unlabeled samples.

As a result, the overall loss function of our MGDNet can be derived as follows:

$$\text{Loss} = \lambda L_{\text{ss}} + (1 - \lambda) L_{\text{DCF}} \tag{18}$$

where λ is the hyperparameter.

Fig. 4. Some examples from the UCM land use dataset.

IV. Experiments

A. Datasets

In order to evaluate the effectiveness of the proposed MGDNet, three challenging datasets for class-imbalanced semisupervised scene classification are applied in the experiments, including IL, SSL, and ISSL methods. They are UC-Merced (UCM) Land Use Dataset, NWPU RESICS-45 Dataset, and Aerial Image Dataset (AID) [1], which are shown in Figs. 4–6, respectively.

UCM Land Use Dataset was obtained by the U.S. Geological Survey, which contains 21 categories of scene images with the pixel size of 256 × 256.

NWPU-RESISC45 Dataset was released by Northwestern Polytechnical University, which contains 45 categories. The dataset has two notable characteristics: 1) high intraclass diversity and interclass similarity and 2) large scale.

AID Dataset was released by Huazhong University of Science and Technology and Wuhan University in 2017, which contains 30 categories and also had a smaller interclass dissimilarity but higher intraclass variations.

B. Parameter Setup

In the experiments, we employ the long-tailed versions of three public datasets with controllable degrees of data imbalance, and imbalance factors L we use in experiments are 10 and 30. The results for UCM are presented in Table I, and the results for RESISC-45 and AID are presented in Tables II and III. The following is the setup for each dataset.

The 21 classes of UCM are split into seven, seven, and seven classes for the head, medium, and tail classes, respectively. The 45 classes of NWPU-RESICS45 are split into ten, 20, and 15 classes for the head, medium, and tail classes, respectively.

TABLE I
COMPARISON OF ALL METHODS ON THE UCM DATASET

	Method	10% Training Ratios		20% Training Ratios		50% Training Ratios	
		L=10	L=30	L=10	L=30	L=10	L=30
Imbalanced Learning	SMOTEBagging[63]	43.29±1.17	34.58±1.38	48.95±1.56	38.45±1.62	55.66±1.96	45.83±1.77
	BBN [17]	65.87±1.88	62.96±1.25	68.54±0.79	66.12±1.56	73.94±1.39	69.57±1.28
	VS-Loss [41]	66.37±0.59	63.14±1.31	69.32±0.50	68.40±1.29	74.31±0.71	70.48±0.78
Semi-supervised Learning	Mixmatch [33]	84.41±0.96	83.10±1.07	89.68±1.27	84.55±0.28	93.32±0.35	90.21±0.61
	Fixmatch [48]	85.89±1.15	84.27±1.51	90.10±1.33	86.33±1.14	94.88±0.73	91.39±0.73
	SS-RCSN [25]	88.57±1.52	86.71±1.37	92.60±1.16	88.28±0.62	96.91±0.49	94.45±0.22
Imbalanced Semi-supervised Learning	ARCnet[64]	83.41±1.10	81.59±1.05	86.63±1.14	82.54±0.71	91.12±0.82	88.25±0.21
	SKAL [15]	87.81±1.54	85.33±1.26	90.46±0.69	86.33±1.18	94.88±0.36	92.13±0.89
	EMTCAL[50]	86.14±1.29	84.52±0.93	91.42±1.23	87.12±1.05	95.33±0.34	92.42±0.66
	DARP+remixmatch [49]	85.32±0.45	84.19±0.38	90.32±0.47	89.38±0.53	94.59±0.62	91.76±0.43
	DARP+fixmatch [49]	86.65±0.88	84.27±0.39	91.59±1.65	89.07±1.36	95.04±0.76	92.53±0.41
	ABC+remixmatch [24]	87.55±0.69	85.43±0.97	92.36±0.61	89.51±0.55	95.92±0.21	92.28±0.70
	ABC+fixmatch [24]	88.18±0.53	85.80±0.24	93.25±0.58	90.76±0.28	96.65±1.37	93.36±0.58
	Our method	**90.49±0.44**	**88.85±0.68**	**95.74±0.64**	**93.41±0.79**	**97.48±0.32**	**95.89±0.26**

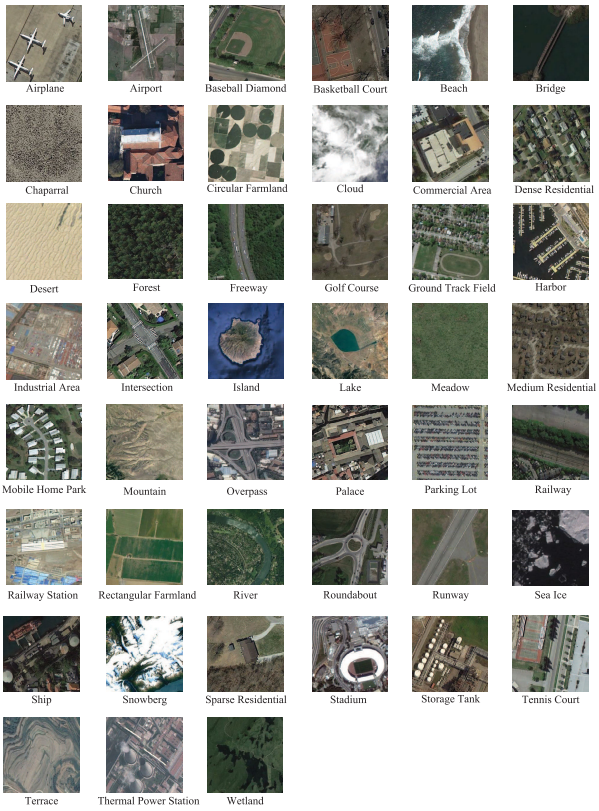


Fig. 5. Some examples from the RESICS-45 dataset.

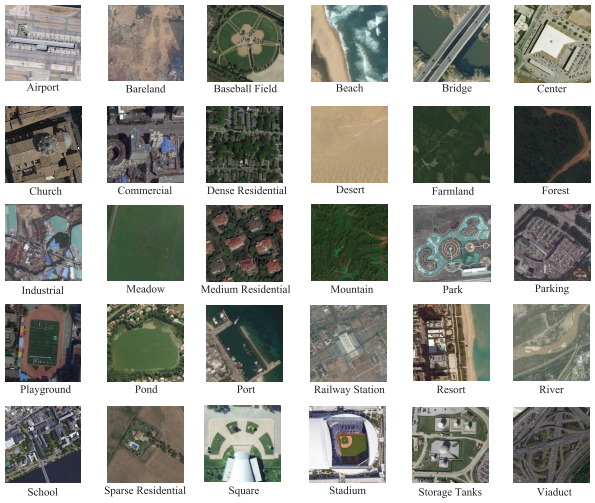


Fig. 6. Some examples from the AID dataset.

The 30 classes of AID are split into ten, ten and ten classes for the head, medium, and tail classes, respectively.

In the proposed model, hyperparameter λ and γ are set to 0.5, the learning rate is set to 0.0005, P_{th} is set to 0.9, and the batch size is set to 32, respectively. The scale of the mask is set from 0.05 to 0.35. All the experiments were conducted on the server equipped with Intel Xeon Silver 4210R, NVIDIA Geforce RTX3080 GPU, and 192-GB DDR4 memory. The Adam optimizer is employed to train the proposed decoupling model. ResNet18 is applied as the classification backbone.

C. Training Setup

In the experiments, the proposed model is compared with the traditional IL methods, the SSL methods, and the ISSL methods, respectively. Each dataset is split into 60% training set, 20% validation set, and 20% test set, respectively. The traditional IL methods are trained with a proportion of 5%–50% of the labeled data. The SSL and ISSL methods are trained with the 10%–50% proportion of labeled data and the remaining unlabeled data.

D. Comparisons on UCM Dataset

Table I shows the accuracy results on the UCM dataset. According to Table I, the proposed MGDNet works effectively

TABLE II
COMPARISON OF ALL METHODS ON THE NWPU RESICS-45 DATASET

	Method	5% Training Ratios		10% Training Ratios		20% Training Ratios	
		L=10	L=30	L=10	L=30	L=10	L=30
Imbalanced Learning	SMOTEBagging[63]	37.51±2.76	33.48±2.59	38.42±1.89	34.46±2.33	40.06±2.41	35.49±1.89
	BBN [17]	62.87±1.79	60.89±1.55	65.04±1.79	63.12±1.32	68.41±1.57	65.38±1.34
	Vector-Scaling Loss [41]	63.55±1.35	61.73±2.02	67.32±1.11	64.05±1.63	69.44±1.58	66.25±2.04
Semi-supervised Learning	Mixmatch [33]	76.41±0.69	72.02±1.34	81.68±1.56	78.99±0.97	84.32±0.81	82.55±0.49
	Fixmatch [48]	78.89±1.06	74.28±0.85	82.10±1.73	80.35±1.27	85.88±0.30	83.42±0.82
	SS-RCSN[25]	80.73±0.75	79.15±0.69	83.30±1.08	81.62±0.85	90.64±0.36	88.04±0.42
Imbalanced Semi-supervised Learning	ARCnet[64]	77.45±1.65	75.72±1.12	78.12±1.32	76.57±1.41	86.62±0.57	81.76±0.54
	SKAL [15]	78.87±1.02	77.23±1.54	82.33±1.42	80.58±1.22	86.41±0.49	83.26±0.25
	EMTCAL[50]	79.14±1.15	78.23±1.38	83.42±1.36	80.55±1.14	89.13±0.34	87.68±0.81
	DARP+remixmatch[49]	78.92±0.83	76.51±0.91	82.32±0.54	81.47±1.31	88.59±0.86	86.22±0.76
	DARP+fixmatch [49]	79.37±0.43	77.26±1.51	82.77±0.73	80.98±0.53	87.92±1.13	85.33±1.37
	ABC+remixmatch[24]	79.80±0.92	78.82±0.72	82.31±1.12	81.65±1.34	88.45±0.28	86.53±0.86
	ABC+fixmatch[24]	80.35±0.46	78.54±0.90	83.54±0.92	81.86±0.39	89.65±0.47	87.38±0.63
	Our method	**81.83±0.85**	**80.46±0.91**	**84.81±0.36**	**82.97±0.81**	**91.41±0.69**	**89.85±0.72**

and achieves the optimal results among all ISSL approaches. Specifically, our model achieves 90.49%, 95.74% and 97.48% accuracy for imbalance ratio = 10% with 10%, 20% and 50% labeled data, respectively. Although ABC-fixmatch and DARP + fixmatch have favorable performance on imbalance ratio = 10, their accuracy drops dramatically as the imbalance ratio rises. In contrast, our proposed decoupling model yields higher accuracies than other ISSL approaches, which improves by 4.66%, 3.42%, 4.58%, and 3.05% compared with DARP + Remixmatch, ABC + Remixmatch, DARP + fixmatch, and ABC + Fixmatch for imbalance ratio = 30% with 10% labeled data, respectively. We argue that other ISSL models are constrained by insufficiently balanced training samples, and our model can properly find more pseudolabel samples from the minority classes.

Compared with other remote sensing methods, our method achieves the best results. For ARCnet, SKAL, and EMTCAL, these methods do not consider the lack of information on unlabeled data and class imbalance, while our method has stronger learning ability for imbalanced unlabeled data. For SS-RCSN, our method can balance the effect of data on the model by decoupling at different granularities.

To compare with the IL methods, EasyEnsemble [65], RUSBoostClassifier [66], SMOTEBagging [63], BBN [17] and VS loss [41] are utilized for class-imbalanced remote sensing image scene classification. According to the comparison results, the performance of the proposed MGDNet performs significantly better than the IL methods, demonstrating the superiority of our proposed method.

E. Comparisons on NWPU RESICS-45 Dataset

Table II shows the accuracy results on the NWPU RESICS-45 dataset. On this dataset, all algorithms have a considerable drop in accuracy. From Table II, we can see that the proposed decoupling network also achieves excellent results for imbalance ratio = 10% and 30% in the test set. This may be because the proposed decoupling classified the tail class data with higher accuracy and produced a significantly more balanced distribution than other ISSL methods. In semisupervised methods, Mixmatch and Fixmatch are semisupervised approaches that highlight certain limitations without DARP and ABC. The reason is that MixMatch, Fixmatch are difficult to learn from biased pseudolabels due to class-imbalanced distribution. Compared with the imbalanced data method, the accuracy also drops significantly, indicating that the imbalanced methods have a weaker learning ability for unlabeled data.

In summary, our proposed MGDNet produces higher accuracies than IL, SSL, and ISSL methods. The experimental results show that the proposed MGDNet model has advantages over current state-of-the-art methods in large-scale datasets.

F. Comparisons on AID Dataset

Table III shows the overall accuracies of AID. All approaches exhibit the same trend on the RESICS-45 dataset. Our proposed decoupling model achieves the best results among all the ISSL comparisons. The specific reasons may be as follows. In the end-to-end ABC module, a significant number of tail data samples were still misclassified as head classes. At the same time, it is difficult for DARP to distinguish between hard and easy samples, and its capacity to learn from unlabeled data has certain limitations. In contrast, our method selects many high-confidence pseudolabels during the training process, which can improve the model's performance for unlabeled data more stably.

Comparisons on three public datasets demonstrate that the proposed ISSL framework can achieve superior performance in scene classification under imbalanced conditions.

G. Ablation Experiment

To analyze the proposed ISSL network, an ablation study was conducted on the UCM dataset with 50% labeled data,

TABLE III
COMPARISON OF ALL METHODS ON THE AID DATASET

	Method	10% Training Ratios		20% Training Ratios		30% Training Ratios	
		L=10	L=30	L=10	L=30	L=10	L=30
Imbalanced Learning	SMOTEBagging [63]	44.65±0.83	35.41±0.75	49.20±0.68	40.39±0.92	57.48±0.85	47.32±1.46
	BBN [17]	63.87±1.53	61.99±2.02	67.71±1.44	64.12±1.61	70.43±1.32	66.24±1.24
	Vector-Scaling Loss [41]	65.32±0.84	63.04±0.69	68.32±0.78	66.11±1.49	71.49±1.28	67.23±0.71
Semi-supervised Learning	Mixmatch [33]	78.41±0.97	75.58±0.82	81.38±0.68	77.68±1.22	85.43±0.92	82.92±0.41
	Fixmatch [48]	80.89±1.21	77.06±1.40	81.25±1.17	79.10±0.73	86.52±0.56	84.65±0.70
	SS-RCSN [25]	82.51±1.13	81.78±0.82	85.45±0.67	83.61±1.04	91.32±0.55	88.41±0.45
Imbalanced Semi-supervised Learning	ARCnet [64]	77.82±1.52	76.55±0.85	78.26±1.31	77.36±1.12	87.64±0.88	85.11±0.91
	SKAL [15]	80.39±0.76	79.24±1.23	83.68±1.53	81.80±0.77	87.31±0.73	85.68±0.87
	EMTCAL [50]	81.24±0.85	77.75±1.51	82.24±1.07	79.62±1.71	88.49±0.97	86.01±0.54
	DARP+remixmatch [49]	81.43±0.51	79.41±1.30	82.32±0.67	81.75±1.39	88.46±0.65	86.72±0.48
	DARP+fixmatch [49]	82.02±0.96	80.19±0.59	82.56±0.80	81.68±1.92	89.55±1.83	87.20±0.76
	ABC+remixmatch [24]	82.18±0.83	81.48±0.93	83.91±0.74	82.38±1.23	90.92±1.08	87.58±1.92
	ABC+fixmatch [24]	83.49±0.57	81.32±1.38	84.42±0.41	83.66±0.48	90.65±1.33	88.79±0.66
	Our method	**84.78±0.32**	**82.41±0.69**	**86.52±0.81**	**85.09±0.61**	**92.41±0.57**	**90.61±0.92**

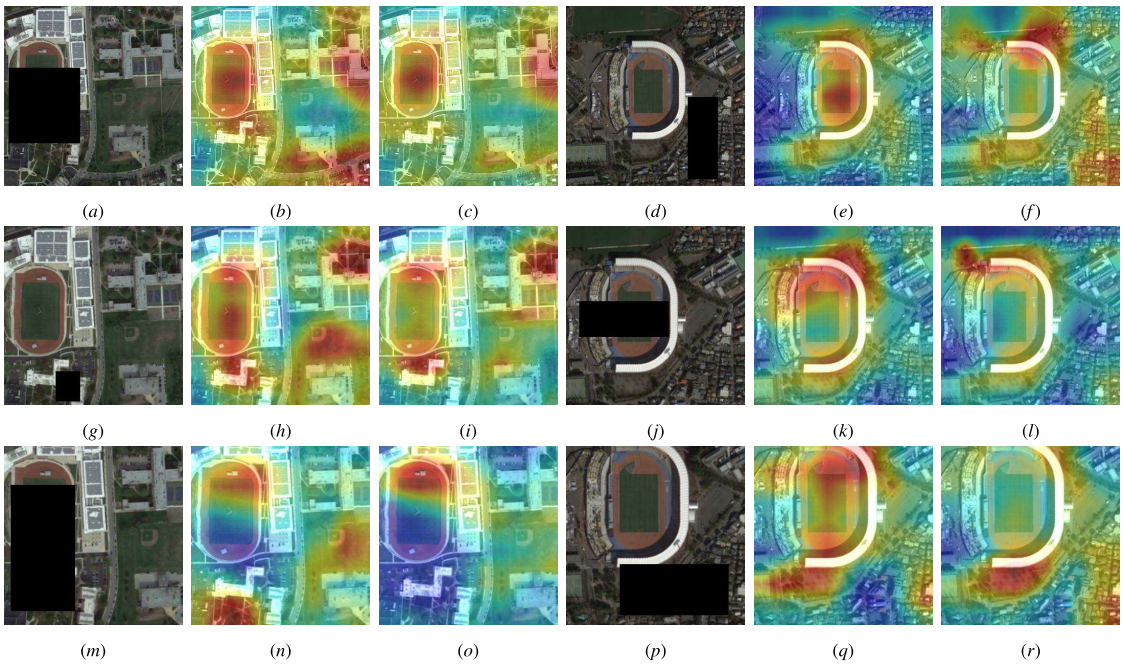
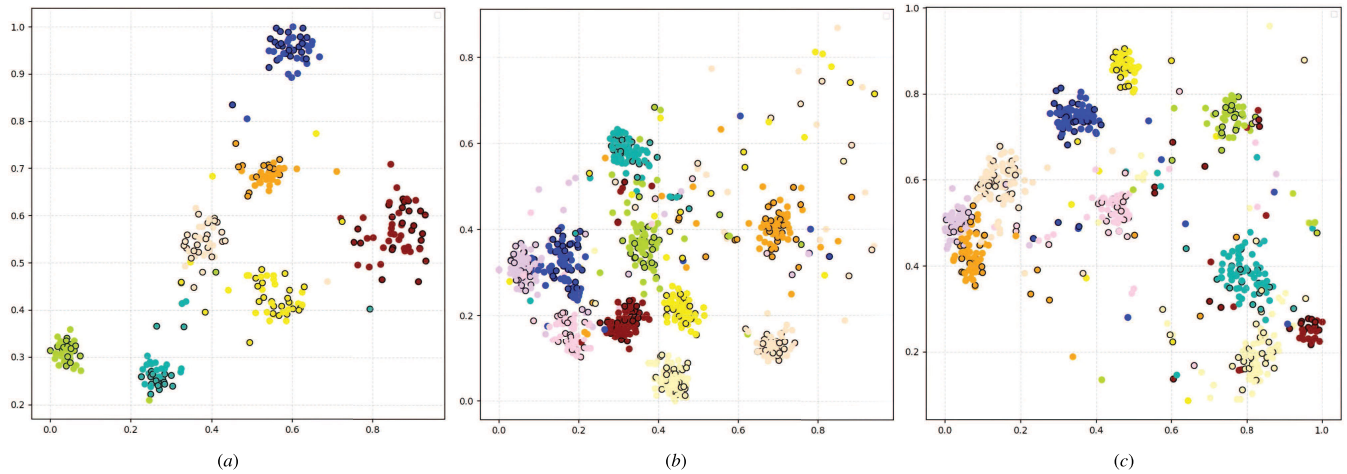


Fig. 7. Example visualization of the minority class "School" and "Playground" on AID. (a) Original image (first granularity). (b) Using the DCF loss function (first granularity). (c) Without the DCF loss function (first granularity). (d) Original image (first granularity). (e) Using the DCF loss function (first granularity). (f) Without the DCF loss function (first granularity). (g) Original image (second granularity). (h) Using the DCF loss function (second granularity). (i) Without the DCF loss function (second granularity). (j) Original image (second granularity). (k) Using the DCF loss function (second granularity). (l) Without the DCF loss function (second granularity). (m) Original image (third granularity). (n) Using the DCF loss function (third granularity). (o) Without the DCF loss function (third granularity). (p) Original image (third granularity). (q) Using the DCF loss function (third granularity). (r) Without the DCF loss function (third granularity).

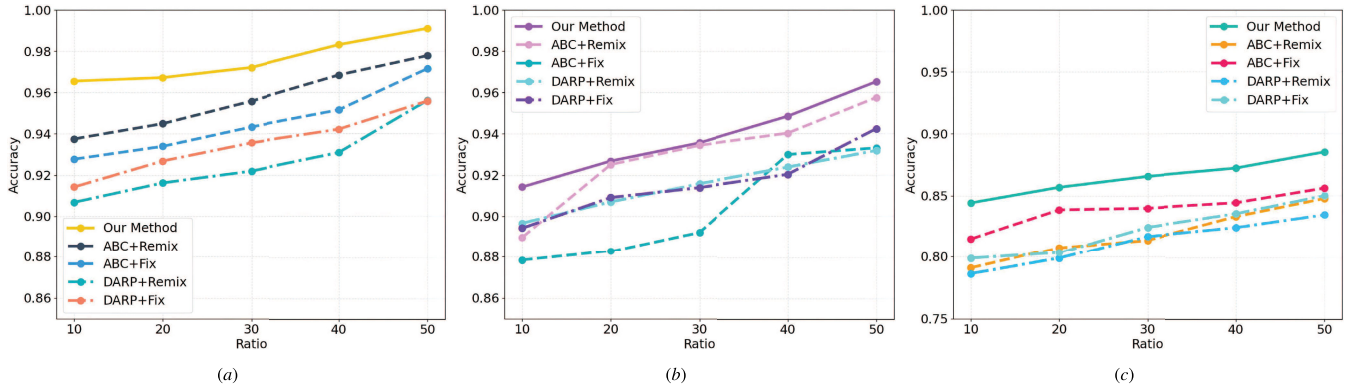
and the imbalance factor L we use in experiments is 30. The experimental results are shown in Table IV, where ✓ means adding the module.

1) *Removing the MGCFR Framework:* In the proposed decoupling module, the MGCFR is introduced to obtain more discriminative fine-grained features. The results are shown in Table IV. For testing the role of the proposed framework in the decoupling network, the MGCFR is removed. It can be seen that the accuracy of scene classification is significantly decreased. The primary reason is that the fine-grained features are lost when the MGCFR is removed. In order to visually illustrate the efficacy of MGCFR, feature maps of different granularity remote images using the MGCFR are shown in Fig. 7. According to Fig. 7, this indicates that the proposed MGCFR indeed forces the decoupling model to find discriminative fine-grained features in the remote sensing image. The model without MGCFR focuses on



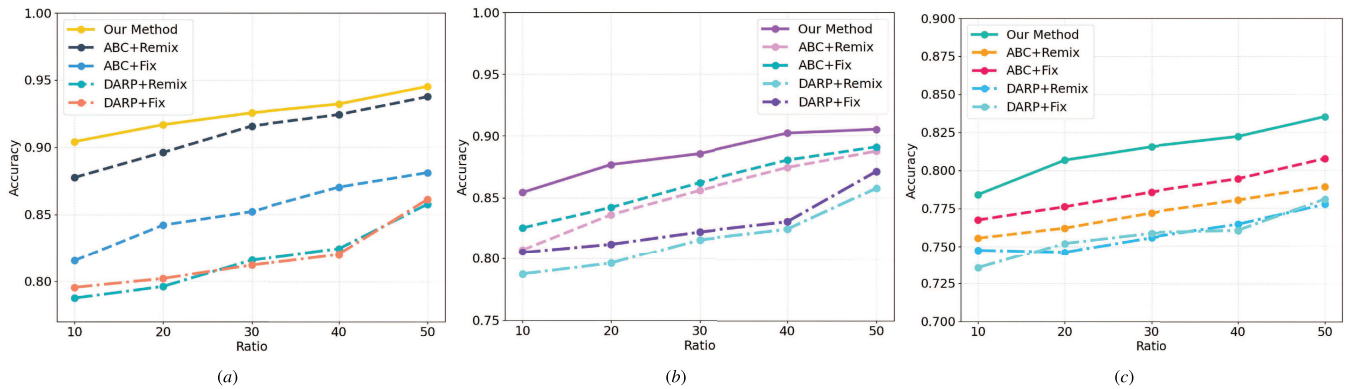
(a) (b) (c)

Fig. 8. T-SNE visualization of the minority class samples selection for the ISSL task. (a) UCM. (b) NWPU-RESICS45. (c) AID.



(a) (b) (c)

Fig. 9. Comparisons on the UCM dataset under different proportions of unlabeled data. (a) Majority classes. (b) Medium classes. (c) Minority classes.



(a) (b) (c)

Fig. 10. Comparisons on the NWPU RESICS-45 dataset under different proportions of unlabeled data. (a) Majority classes. (b) Medium classes. (c) Minority classes.

the most obvious region, and its attention has certain limitations.

2) *Removing the DCF Loss Function:* To illustrate the advantages of the proposed DCF loss function, we removed the proposed loss to implement the visualization for the features of different granularity remote sensing images. Examples of the "stadium" and "school" classes are shown in Fig. 7. According to Fig. 7, the class "stadium" and "school" in Fig. 7 may have less obvious regional differences in the two remote sensing images. The proposed model considers the relevant regions in other parts of each remote sensing image, such as the region next to the salient region in Fig. 7. It can be seen that the use of MGCFR allows the attention of the model to be not only limited to the salient region, while the model using the DCF loss function enables the relevant regional features to be more differentiated. Therefore, the proposed model can learn the potential interactions between different parts of the image with different granularity, and thus the

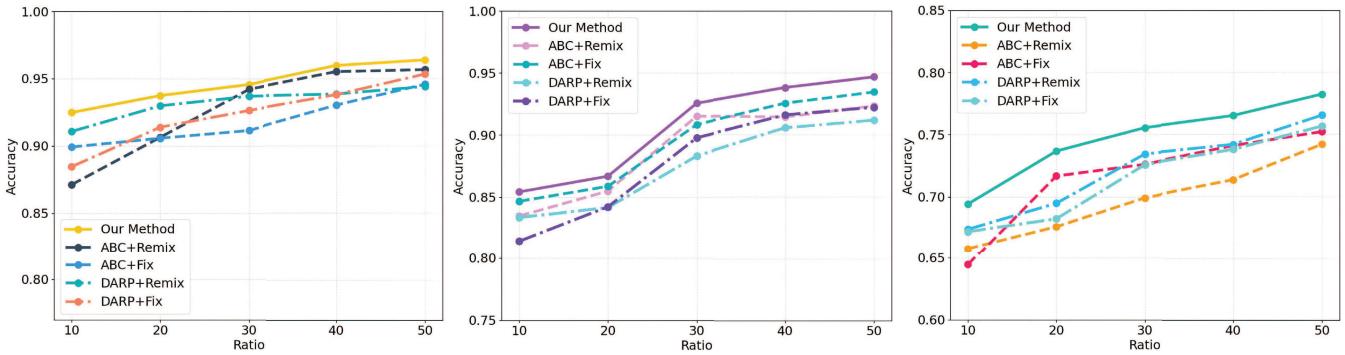


Fig. 11. Comparisons on the AID dataset under different proportions of unlabeled data. (a) Majority classes. (b) Medium classes. (c) Minority classes.

TABLE IV
ABLATION EXPERIMENT RESULTS OF THE PROPOSED MGDNET. ✓ MEANS ADDING THE MODULE

Component	UCM dataset (50% Training Ratios)			
MGCFR	✓	×	✓	✓
CIPS	×	✓	✓	✓
DCF Loss	✓	✓	×	✓
Accuracy(Acc)(%)	**93.58**	**94.23**	**93.96**	**95.89**

TABLE V
CALCULATION TIME ON THE UCM DATASET

Method name	ATT(min)	Test time(ms)
DARP+fixmatch	145.2	37.1
ABC+fixmatch	151.6	38.4
SS-RCSN	112.3	28.2
Our Method	103.7	27.6

interclass and intraclass features of the remote sensing images.

3) *Removing the CIPS Module:* The experimental result without the CIPS component is shown in Table IV. It can be seen that after the removal of the CIPS component, the classification accuracy dropped seriously. In contrast, using the proposed CIPS can effectively improve the performance of class-imbalanced scene classification. This indicates that the proposed method enhances the ability to learn the imbalanced samples with a lack of annotations. Fig. 8 shows the results of embedding high-dimensional features into 2-D space by T-distributed stochastic neighbor embedding (T-SNE). The points with different colors represent unlabeled samples with different classes, while the circled points indicate selected samples by CIPS. It is evident that our CIPS can pick out more credible supplementary samples in the class imbalanced semi-supervised learning (CISSL) training.
4) *Robustness to Varying Label Ratios:* To obtain a complete picture of the performance of the proposed model, we analyze its robustness to varying label ratios in the range of [10, 50] and randomly selected five classes from the majority classes, the middle classes, and the minority classes. The overall results are reported in Figs. 9–11. As presented in Figs. 9–11, the proposed MGDNet exhibited a significant improvement in all the minority classes. This indicates that the proposed decoupling network can mitigate the effects of overfitting by head class samples. Thus, it can achieve robust performance for the tail classes with a small number of labeled samples.
5) *Calculation Time:* The average training time and test time on the UCM dataset are provided in Table V and compared with other approaches to assess the computational effectiveness of the proposed model. In view of the fairness of the comparison, this section compares the SSL method with the ISSL method since the traditional IL method uses only labeled data. As can be observed, the proposed approach required less ATT and processed a test picture in less time than other SSL and ISSL approaches.

V. CONCLUSION

This article proposes the MGDNet for imbalanced remote sensing image scene classification. Experimental results on UCM, RESICS-45, and AID datasets indicate that the proposed ISSL model achieves outstanding performance. It has been proven that the MGCFR method can learn fine-grained features for the ISSL classification task. Furthermore, it is verified that the CIPS method for unlabeled samples can effectively select high-confidence pseudolabel. Finally, the combination of the MGCFR method and the DCF loss function improves the ability to drive the local features to be more discriminative, which significantly contributes to the enhanced feature extraction capability of the decoupling network.

In our future work, we will focus on enhancing MGDNet with information regarding instance-level difficulties to ensure excellent computing performance and improve semisupervised classification accuracy.

REFERENCES

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020.
[2] S.-W. Chen and C.-S. Tao, "PolSAR image classification using polarimetric-feature-driven deep convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 627–631, Apr. 2018.

[3] J. Geng, X. Deng, X. Ma, and W. Jiang, "Transfer learning for SAR image classification via deep joint distribution adaptation networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5377–5392, Aug. 2020.

[4] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526914.

[5] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[6] S.-W. Chen, X.-C. Cui, X.-S. Wang, and S.-P. Xiao, "Speckle-free SAR image ship detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5969–5983, 2021.

[7] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian Heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.

[8] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2020.

[9] N. Ammour, "Continual learning using data regeneration for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[10] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.

[11] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620214.

[12] C. Xu, G. Zhu, and J. Shu, "A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[13] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 5396–5407, 2020.

[14] A. Ma, N. Yu, Z. Zheng, Y. Zhong, and L. Zhang, "A supervised progressive growing generative adversarial network for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618818.

[15] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1414–1428, Apr. 2022.

[16] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.

[17] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9719–9728.

[18] W. Jiang, "A correlation coefficient for belief functions," *Int. J. Approx. Reasoning*, vol. 103, pp. 94–106, Dec. 2018.

[19] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," 2021, *arXiv:2110.04596*.

[20] Z. Deng, H. Liu, Y. Wang, C. Wang, Z. Yu, and X. Sun, "PML: Progressive margin loss for long-tailed age classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10503–10512.

[21] W. Jiang, C. Xie, M. Zhuang, and Y. Tang, "Failure mode and effects analysis based on a novel fuzzy evidential method," *Appl. Soft Comput.*, vol. 57, pp. 672–683, Aug. 2017.

[22] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 27, 2022, doi: 10.1109/TNNLS.2021.3136503.

[23] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10857–10866.

[24] H. Lee, S. Shin, and H. Kim, "ABC: Auxiliary balanced classifier for class-imbalanced semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–13.

[25] W. Miao, J. Geng, and W. Jiang, "Semi-supervised remote-sensing image scene classification using representation consistency Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616614.

[26] W. Jiang, Y. Cao, and X. Deng, "A novel Z-network model based on Bayesian network and Z-number," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 8, pp. 1585–1599, Aug. 2020.

[27] S.-W. Chen, "SAR image speckle filtering with context covariance matrix formulation and similarity test," *IEEE Trans. Image Process.*, vol. 29, pp. 6641–6654, 2020.

[28] 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 Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[29] G. Perantoni and L. Bruzzone, "A novel technique for robust training of deep networks with multisource weak labeled remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[30] X. Yao, L. Yang, G. Cheng, J. Han, and L. Guo, "Scene classification of high resolution remote sensing images via self-paced deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 521–524.

[31] X. Dai, X. Wu, B. Wang, and L. Zhang, "Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 869–873, Jun. 2019.

[32] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 23–43, Nov. 2018.

[33] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5049–5059.

[34] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 162–178.

[35] J. Ren et al., "Balanced meta-softmax for long-tailed visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4175–4186.

[36] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[37] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 694–710.

[38] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "RSG: A simple but effective module for learning imbalanced datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3784–3793.

[39] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6930–6940.

[40] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*.

[41] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis, "Label-imbalanced and group-sensitive classification under overparameterization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[42] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[43] A. Kurakin et al., "RemixMatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–13.

[44] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.

[45] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[46] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (JPSA) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3602–3615, Jul. 2021.

[47] D. Berthelot et al., "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," 2019, *arXiv:1911.09785*.

[48] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.
[49] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14567–14579.
[50] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626915.
[51] W. Zhang, L. Jiao, F. Liu, J. Liu, and Z. Cui, "LHNet: Laplacian convolutional block for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626513.
[52] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.
[53] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4511616.
[54] K. Xu, P. Deng, and H. Huang, "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618715.
[55] M. Faramarzi, M. Amini, A. Badrinaaraayanan, V. Verma, and S. Chandar, "PatchUp: A feature-space block-level regularization technique for convolutional neural networks," 2020, *arXiv:2006.07794*.
[56] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk management," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–13.
[57] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[58] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," 2020, *arXiv:2002.07394*.
[59] K. Huang, J. Geng, W. Jiang, X. Deng, and Z. Xu, "Pseudo-loss confidence metric for semi-supervised few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8671–8680.
[60] M. Lazarou, T. Stathaki, and Y. Avrithis, "Iterative label cleaning for transductive and semi-supervised few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8751–8760.
[61] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 312–321.
[62] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2U-Net: A simple noisy label detection approach for deep neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3326–3334.
[63] F. S. Hanifah, H. Wijayanto, and A. Kurnia, "Smotebagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank X)," *Appl. Math. Sci.*, vol. 9, no. 138, pp. 6857–6865, 2015.
[64] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2018.
[65] T.-Y. Liu, "EasyEnsemble and feature selection for imbalance data sets," in *Proc. Int. Joint Conf. Bioinf., Syst. Biol. Intell. Comput.*, Aug. 2009, pp. 517–520.
[66] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern., A, Syst. Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

Wang Miao is currently pursuing the Ph.D. degree with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.

His research interests include remote-sensing image scene classification and deep learning.

Jie Geng (Member, IEEE) received the B.S. degree in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2013, and the Ph.D. degree from the School of Information and Communication Engineering, Dalian University of Technology, in 2018.

He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests include SAR image processing, deep learning, and few-shot learning.

Wen Jiang received the bachelor's and master's degrees from Information Engineering University, Zhengzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2009.

She is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include information fusion and artificial intelligence with uncertainty.