





Remote Sensing Image Scene Classification by Multiple Granularity Semantic Learning

Weilong Guo , Shengyang Li , Jian Yang, Zhuang Zhou , Yunfei Liu, Junjie Lu, Longxuan Kou, and Manqi Zhao 

Abstract—Remote sensing image scene classification faces challenges, such as the difference in semantic granularity of different scene categories and the imbalance of the number of samples, which cause the wrong features learning for deep convolutional networks (DCNs). This article proposes a multiple granularity semantic learning network (MGSN), including multiple granularity semantic learning (MGSL) and nonuniform sampling augmentation (NUA) modules. Specifically, the MGSL module makes full use of different granularities of semantic information of scenes, guiding the network to learn global and local features simultaneously. And, the relationship between semantic features of different granularity has been explored, based on which the learning of coarse-grained features helps to improve the learning of fine-grained semantic features. It shows that learning fine-grain semantics can inhibit learning coarse-grain semantic features. The NUA module combines sampling and sample augmentation to balance the sample distribution, which can avoid overfitting caused by oversampling. The proposed MGSN achieved state-of-the-art classification accuracy on two large-scale remote sensing image scene classification datasets, Million-AID and NWPU-RESISC45. Under 10% and 20% training samples of the NWPU-RESISC45 dataset, MGSN achieves 91.92% and 94.33% top-1 accuracy, respectively. In experiments conducted on the Million-AID dataset, the proposed MGSN performed best among 18 DCNs. In comparison to the baseline, FixEfficientNet, MGSN improved the accuracy of top-1 and top-5 by 10.63% and 5.47%, respectively, with low complexity costs.

Index Terms—Deep convolutional networks (DCNs), multiple granularity semantic learning (MGSL), remote sensing image scene classification, imbalance of sample number.

I. INTRODUCTION

REMOTE sensing image scene classification [1]–[4] has been widely used in fields [5], such as land surveying,

Manuscript received October 28, 2021; revised January 13, 2022; accepted March 1, 2022. Date of publication March 11, 2022; date of current version April 6, 2022. This work was supported in part by the Space Science and Application of China Manned Space Engineering DataBase, National Basic Science Data Center, under Grant NBSDC-DB-17 and in part by the Director’s Foundation of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, under Grant CSU-JJKT-2020-9. (Corresponding author: Shengyang Li.)

Weilong Guo and Junjie Lu are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: guoweilong19@csu.ac.cn; lujunjie19@csu.ac.cn).

Shengyang Li, Jian Yang, Zhuang Zhou, Yunfei Liu, Longxuan Kou, and Manqi Zhao are with the Technology and Engineering Center for Space Utilization, the Key Laboratory of Space Utilization, Chinese Academy of Sciences and the University of Chinese Academy of Sciences, Beijing 100094, China (e-mail: shyli@csu.ac.cn; yangjian202@mails.ucas.ac.cn; zhouzhuang@csu.ac.cn; liuyunfei@csu.ac.cn; koulongxuan17@mails.ucas.ac.cn; zhao-manqi19@csu.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3158703

nature monitoring, and urban planning [6]. It has made great progress [7], [8] with the development of deep learning [9], [10] and automatic machine learning [11], such as a neural architecture search (NAS) technology [12]. Nevertheless, challenges remain [13], [14], such as differences in semantic granularity and imbalanced samples, across scene categories [15].

Differences of semantic granularity refer to the existence of both fine- and coarse-grained annotations. Fig. 1(a) shows remote sensing image scenes, which are with coarse-grained labels, whose annotations are corresponding to global content information, while Fig. 1(b) shows scenes, which are with fine-grained labels, corresponding to local region content information. The existence of remote sensing scenes labeled with different semantic granularity requires deep convolutional networks (DCNs) to simultaneously learn global and local features. Current image scene classification schemes extract features, and then perform classification. Feature extraction networks, such as AlexNet [16], VGGNet [17], GoogleNet [18], ResNet [19], EfficientNet [20], RegNet [21], and FixEfficientNet [22], turn images into a $C \times H \times W$ feature map, that is, subsampled as a $1 \times C$ feature vector after average or max pooling. A classifier outputs the probability of different scene classes according to the feature vector. It is worth noting that the classifier shares weights for feature vectors from different remote sensing scenes. It is difficult for a classifier to simultaneously learn multiple granularity semantic information and features of different scale regions.

Class sample imbalance is common in remote sensing image scene classification [23], [24]. As shown in Fig. 2, on the Million-AID dataset, some scene classes are with rich samples and some scene classes are with poor samples. There is a great difference in sample numbers across different categories. And, it can easily lead to wrong feature learning for DCNs. Few samples are not enough to support DCNs learning robust features of the corresponding category scene. Sampling methods are usually used to reduce the imbalance in the data or optimization space. Oversampling [25] and undersampling [26] are commonly used to balance category samples in the data space. However, oversampling usually causes overfitting because of repeated samples, and undersampling may miss valuable information. Sampling in the optimization space is used to balance the focus on category scenes. Tan *et al.* [27] believed that the learning of wrong features of categories with poor samples for DCNs is usually due to suppression of categories with rich samples, which get far more negative than positive gradients during training.

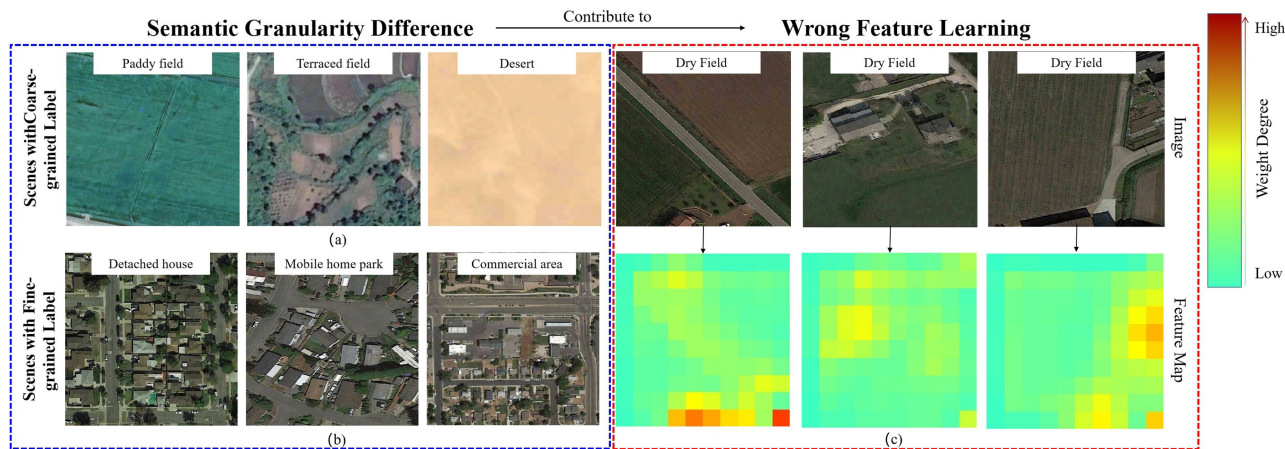


Fig. 1. Semantic granularity difference of different scenes in the Million-AID [13] dataset leads to wrong feature learning of FixEfficientNet. (a) Annotation semantic of scenes with coarse-grained labels corresponds to global region information. (b) Annotation semantic of scenes with fine-grained labels corresponds to local region information. (c) Visualization of feature map shows that FixEfficientNet only learned local feature of “building” and ignored global features, which lead to misclassification.

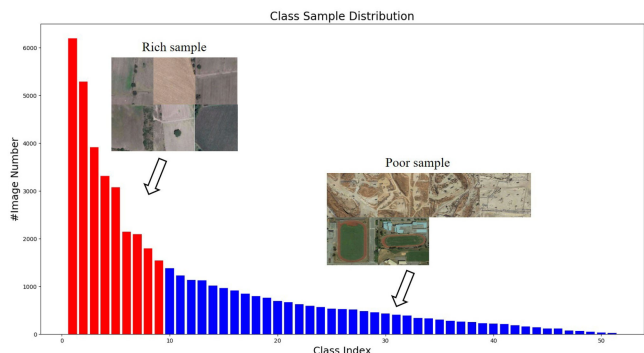


Fig. 2. Class sample imbalance of the Million-AID dataset. Some scene classes are with rich sample (red). And, some scene classes are with poor sample (blue).

To reduce this suppression may help to learn robust features for DCNs.

In this article, we propose a multiple granularity semantic learning network (MGSN), including multiple granularity semantic learning (MGSL) and nonuniform sampling augmentation (NUA) modules. The MGSL module reduces wrong feature learning caused by semantic granularity differences in remote sensing scenes. We regard the classification of remote sensing image scenes with different granularity semantic annotations as a question of multiple semantic granularity feature learning. The combination of semantic granularity features represents different categories of remote sensing scenes. Illustration of generation of multiple granularity semantic annotations is shown in Fig. 3. The node list of each path from the root to leaf in the directed tree corresponds to the multiple annotations of a single remote sensing scene. From root to a leaf node, the annotation semantic granularity is from coarse to fine, and the corresponding semantic region is from global to local. Multiple classifiers are used to predict multiple granularity semantic information in parallel. Different from current methods, the proposed method

learns multiscale region features by setting learning objectives at the semantic level. It learns multiscale region features on a single-scale feature map and does not rely on the heuristic network structure design. The MGSL module has following three advantages.

- 1) It can learn global and local region features on a single-scale feature map at the same time.
- 2) Additional hyperparameters are not needed.
- 3) It does not depend on characteristics of different scenes or statistical information, and it has strong generalization.

The NUA module combines oversampling and sample augmentation, which cooperate. Oversampling makes up for the disadvantage that sample augmentation aggravates class sample imbalance, and sample augmentation makes up for the disadvantage that oversampling easily leads to overfitting.

The main contributions of this article are as follows.

- 1) The proposed MGSL module guides the network to learn multiscale region features on a single-scale feature map. And, we also explored the relation between semantic features of different granularity, founding that coarse-grained feature learning helps to improve the learning of fine-grained semantic features, while fine-grained semantic feature learning can inhibit the learning of coarse-grained semantic features.
- 2) The proposed NUA module reduces class sample imbalance through sample augmentation, which can effectively guide the network to learn robust features of remote sensing scenes with strong generalization.
- 3) MGSN effectively improves DCN performance in the presence of semantic granularity differences and class sample imbalance. It achieved state-of-the-art classification on two large-scale remote sensing image scene classification datasets: 1) Million-AID; and 2) NWPU-RESISC45. In experiments on the Million-AID large-scale remote sensing image scene classification dataset, the proposed method improved top-1 accuracy by 10.63% and top-5 accuracy by 5.47% when compared with the

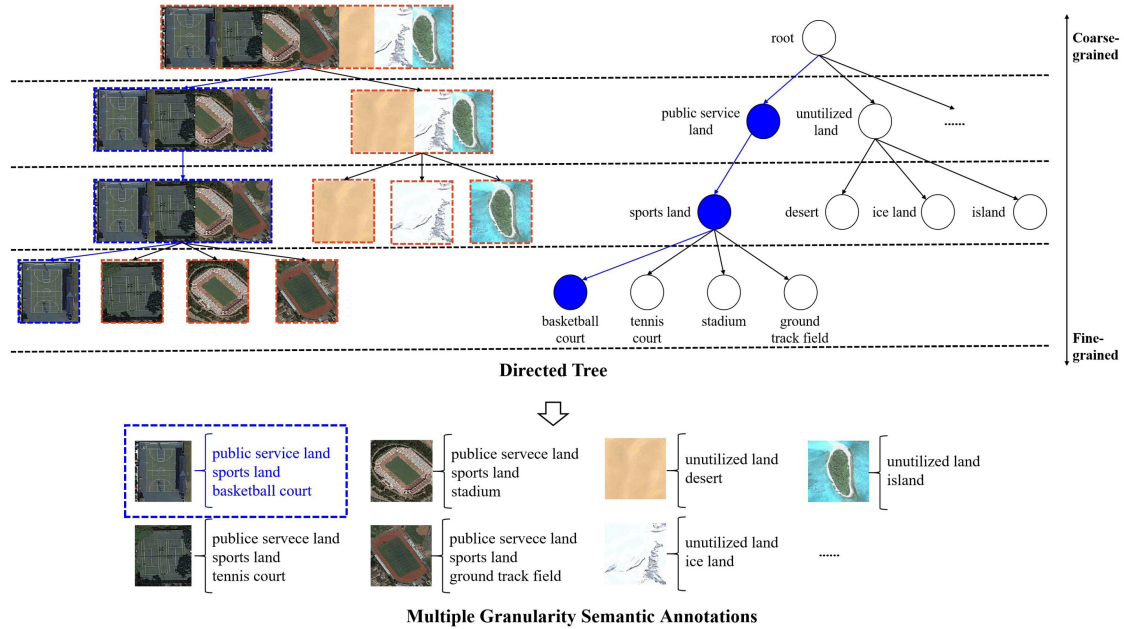


Fig. 3. Generation of multiple granularity semantic annotations of remote sensing scenes in the Million-AID dataset. Nodes of each path from root to leaf in the directed tree correspond to multiple annotations of a single remote sensing scene. From root node to leaf node, the annotation semantic granularity is from coarse to fine. Target prediction classes are in leaf nodes. For example, nodes of the blue path are multiple granularity semantics of the “basketball court” scene. Its multiple granularity semantic annotation is shown in the blue-dotted frame. From coarse- to fine-grained are “public service land,” “sports land,” and “basketball court.” And specifically, “basketball court” is its target prediction class.

baseline FixEfficientNet. Under 10% and 20% training samples of the NWPU-RESISC45 dataset, MGSN achieves 91.92% and 94.33% top-1 accuracy, respectively.

II. RELATED WORKS

There are mainly three types of remote sensing image scene classification methods. One is based on multiscale feature learning, must learn multiple-scale region features and feature aggregation. One is based on discriminative features representation. Another is based on key region feature learning, and usually uses saliency detection and an attention mechanism.

A. Feature Aggregation-Based Methods

Feature aggregation-based methods [30], [31] aggregate multiscale features from different layers of DCNs as the classifier input. This can improve the ability to learn multiscale region features for DCNs, but their aggregation is mainly a heuristic design that involves superparameters, such as the selection and number of layers, merging modes of features, and fusion weights. Yang and Ramanan *et al.* [29] regarded multiscale feature aggregation as the construction of a directed acyclic graph. After average pooling, multiscale features are merged by add. Compared with the use of a single-scale feature, multiscale feature aggregation improves the top-1 accuracy by 4.3%, and it achieves 56.2% top-1 accuracy on the SUN397 dataset. To reduce information redundancy and feature exclusion in feature aggregation, Sun *et al.* [28] proposed a bidirectional gate network for adaptive weighted feature fusion with 95.48% accuracy on the AID [32] remote sensing image scene classification dataset. Multiscale

feature aggregation requires more parameters and more training data, which increases the difficulty of learning for DCNs. In SPP-net [33], spatial pyramid pooling and side supervision strategies are proposed to fuse multiscale features based on pretrained AlexNet. And, it achieved 84.64% classification accuracy under 20% training samples on the NWPU-RESISC45 dataset.

To effectively fuse mid-level and high-level features with the low-level ones, ACR-MLFF [34] proposed a novel multilevel feature fusion network to adaptively reduce channel dimensionality. In [35], a self-attention-based deep feature fusion method is used to aggregate deep layer features and emphasize the weights of the complex objects of remote sensing scenes. It uses a pretrained convolutional neural network to extract the abstract multilayer features, and a nonparametric self-attention layer is proposed for spatial-wise and channel-wise weightings, which enhances the effects of spatial responses of the representative objects. When 10% or 20% samples of the NWPU-RESISC45 dataset are used for training, it achieves 84.38% and 87.86% classification accuracy, respectively. TFADNN [36] uses a two-stream architecture [37] to aggregate deep learning features and general features, named nonlinear encoding bag-of-visual-words.

B. Discriminative Features Representation-Based Methods

Discriminative features representation is important for remote sensing image scene classification. In BoCF [38], a novel feature representation method, named bag of convolutional features, is proposed for scene classification. It uses off-the-shelf convolutional neural networks to generate visual words. And, SCCov [6]

uses covariance pooling to exploit the second-order information contained in multiresolution features, which allows the convolutional neural networks to achieve more representative feature learning. When 20% samples of the NWPU-RESISC45 dataset are used for training, it achieved 92.10% classification accuracy. To address the problems of within-class diversity and between-class similarity, D-CNNs [39] combined deep learning and metric learning together to design a new discriminative objective function that could guide CNNs to learn discriminative features representation. In this mode, images from the same scene are mapped closely to each other, and images of different scene classes are mapped as far as possible. When 20% samples of the NWPU-RESISC45 dataset are used for training, D-CNNs achieved 91.89% classification accuracy. DLA-MatchNet [40] learns discriminative representations and a proper metric for remote sensing scenes in a few-shot manner and it achieves 81.63% classification accuracy under the case of five-way five-shot.

C. Saliency Detection and Attention-Mechanism-Based Methods

Saliency detection and attention-mechanism-based methods distinguish different remote sensing scenes by learning features of key regions. Increasing the weights of some key objects in a scene can reduce misclassification in complex remote sensing environments. Zhang *et al.* [41] proposed a saliency guided information sampling strategy, which removes redundant information and retains saliency region information to improve the performance of DCNs. The proposed method achieved state-of-the-art classification accuracy on the UC Merced dataset [42]. Attention-mechanism-based methods improve classification performance by increasing the focus of DCNs on local key regions. Wang *et al.* [43] proposed an attention recurrent convolutional network to selectively focus on key regions of scenes, extracting only their deep features. When 50% data of UC Merced were used to train the network, it obtained 96.81% classification accuracy. During multilayer and pooling, a large amount of important information is lost, resulting in the insufficient ability of the extracted features to represent objects. To improve the feature extraction and generalization abilities of deep neural networks, EAM [44] proposed an enhanced attention module. It achieved 94.29% accuracy on the NWPU-RESISC45 dataset.

However, these methods ignore the learning of global semantic information in remote sensing scenes. Although focusing on key local regions can improve DCNs' fine-grained classification ability, it leads to misplaced attention on local features of global information-oriented scenes with coarse-grained labels. Scenes in Fig. 1(c) have the coarse-grained label "dry field". From a human perspective, their semantic content is "land" in a global region. However, visualization of feature map FixEfficientNet [22] output shows that the network only learned local semantic information "building" instead of global semantic information "land". Learning global and local semantic information of remote sensing scenes is difficult, but important to improve the robustness of DCNs in complex remote environments.

III. PROPOSED METHOD

Fig. 4 shows the framework of the proposed MGSN, which includes following three parts.

- 1) The NUA module balances samples of different remote sensing scene categories.
- 2) The feature extraction module extracts features from every input remote sensing scene image and outputs its feature map Z^O .
- 3) The MGSL module uses 1×1 convolution (Conv) to split multiple granularity semantic features from Z^O , from which multiple classifiers predict multiple granularity semantic information of corresponding scenes.

Final loss in training consists of multiple granularities semantic losses, making possible multiple granularity semantic and multiple scale region feature learning in backpropagation.

A. NUA Module

The NUA module balances samples of different category scenes, including the sampling and augmentation stages, combining oversampling and augmentation to help DCNs learn robust and distinguishable features of remote sensing scenes. Oversampling can make up for the class sample imbalance caused by sample augmentation, and the latter can make up for overfitting caused by oversampling. This differs from current sampling methods in the calculation of sampling probabilities of different categories.

Sampling probabilities of categories are relative in the NUA module. Fig. 4 shows the calculation pipeline. The maximum number of samples of category k is

$$M_{k-\max} = \max(n_k). \quad (1)$$

And, the sampling probability of the k th category scene is

$$p_k = \begin{cases} \frac{M_{k-\max} - n_k}{n_k}, & \frac{M_{k-\max}}{2} < n_k \leq M_{k-\max} \\ 1, & 1 \leq n_k \leq \frac{M_{k-\max}}{2} \end{cases} \quad (2)$$

where n_k is the number of samples in the k th category.

The fewer the samples of a remote sensing scene category, the higher the sampling probability of each sample in it. When n_k is less than $\frac{M_{k-\max}}{2}$, each sample is selected, and when n_k is equal to $M_{k-\max}$, no sample is selected.

The augmentation increases the number and diversity of category scene samples. The number of augmentations of a sample is calculated as $\lceil \frac{M_{k-\max} - n_k}{n_k} \rceil$. Illustration of different scene images augmentation is shown in Fig. 5, for example, "ground track field" scene image, "oil field" scene image, "apron" scene image, and "quarry" scene image. From left to right, the first column is the original scene image. Columns 2–4 are augmented images. In order of priority, types of augmentation include rotation, center crop, and their combination. The rotation angle of a scene image is an integer multiple of 23. The proportion of center crop order is $\frac{1}{2}$ and $\frac{2}{3}$. After merging augmented and original samples, the numbers of samples tend to be balanced across categories.

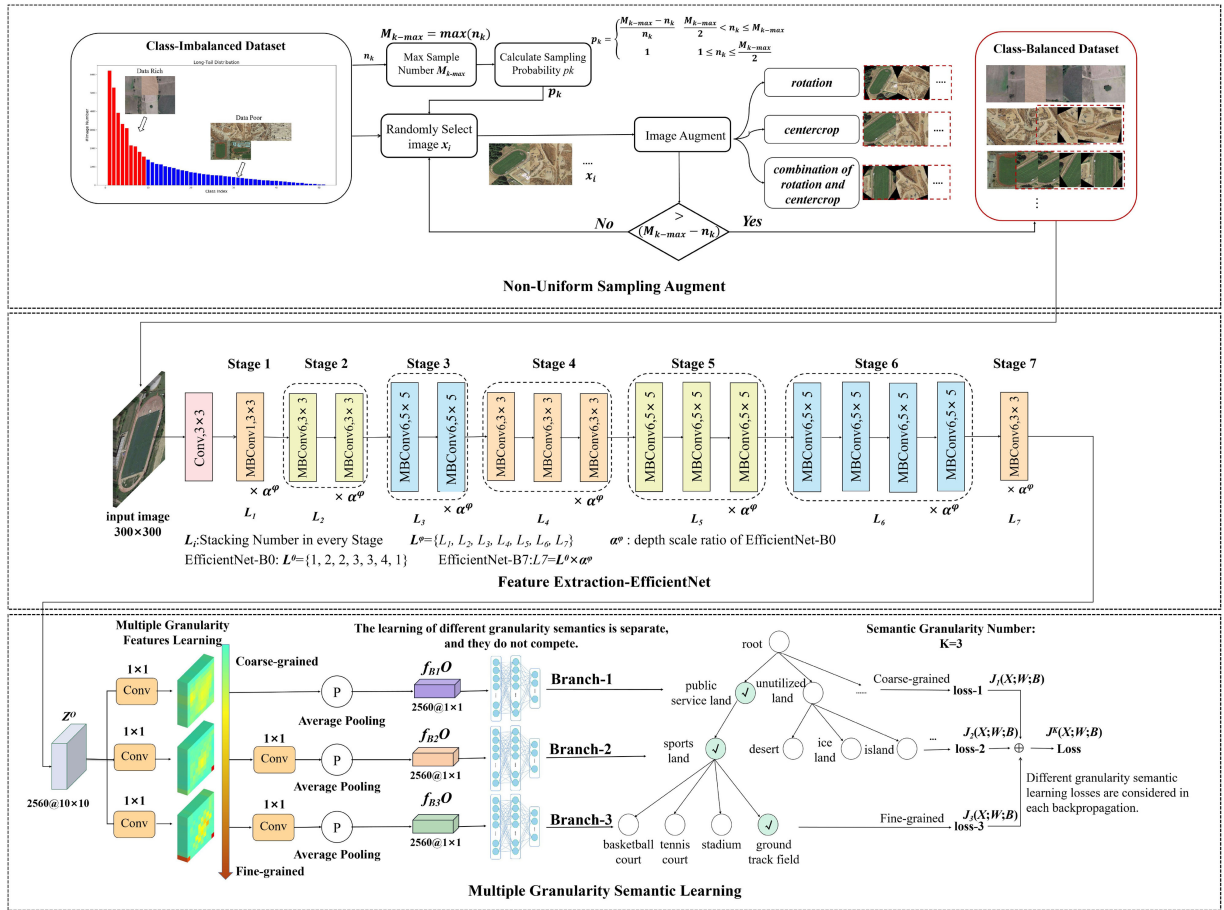


Fig. 4. Framework of the proposed MGSN. (a) Nonuniform Sampling Augment: This module balances samples of different remote sensing scene categories. (b) Feature Extraction-EfficientNet: Features are extracted from every input remote sensing scene image. (c) MGSL: This module helps the network to simultaneously learn global and local semantic features.

B. Feature Extraction

EfficientNet [20] is based on a neural network structure search (NAS), and has excellent feature extraction ability. It is used in many visual tasks, including object detection and tracking. According to image resolution, network depth, and network width, it is divided into EfficientNet-B0–B7. Efficient-B0 is the base, and others can be obtained by scaling it.

We use EfficientNet-B7 as the feature extraction module. Fig. 4 shows its structure, which includes seven stages. An input remote sensing scene image is downsampled by 3×3 Conv to obtain a final feature map Z^O after transformation in each stage. Each stage consists of stacking submodules MBCConv6, $k \times k$, and MBCConv1, $k \times k$, for different times, where k is the size of the Conv kernel. MBCConv6 and MBCConv1 differ according to whether the first 1×1 Conv increases the dimension of the input feature map. The input and output feature map dimensions of the first 1×1 Conv in MBCConv1 are the same, but the dimension of the output feature map of the first 1×1 Conv in MBCConv6 is six times than that of the input feature map.

As shown in Fig. 6, MBCConv6 and MBCConv1 include Conv, depthwise Conv (DWConv) [45], an SE module [46], batch normalization, and the Swish activation function. DWConv includes DWConv in each channel and pointwise Conv at each point on

the feature map. SE is a kind of attention module, consisting of pooling layers, fully-connected layers, Swish activation, and sigmoid and multiply operations.

Let $L = \{L_1, L_2, L_3, L_4, L_5, L_6, L_7\}$ represent the stacking times of MBCConv1 or MBCConv6 in every stage of EfficientNet. In EfficientNet-B0, stacking times of each stage are $L_0 = \{1, 2, 2, 3, 3, 4, 1\}$. The calculation of scaled depth, width, and image resolution from EfficientNet-B0 to EfficientNet-B1–B7 is

$$\begin{cases} \text{depth} : & d = \alpha^\varphi \\ \text{width} : & w = \beta^\varphi \\ \text{resolution} : & \gamma = \gamma^\varphi \\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1 & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \end{cases} \quad (3)$$

where φ is a scale factor. When $\varphi = 7$, the structure of EfficientNet-B7 can be obtained according to EfficientNet-B0, and $\alpha = 1.2$, $\beta = 1.5$, and $\gamma = 1.15$ [20].

C. Multiple Granularity Semantic Learning

In the MGSL module, remote sensing scene image classification is regarded as a multilabel classification task. As shown in Fig. 3, the label class and its father or grandfather class of a single remote sensing scene have multiple annotations corresponding to multiple granularity semantic information from local to global

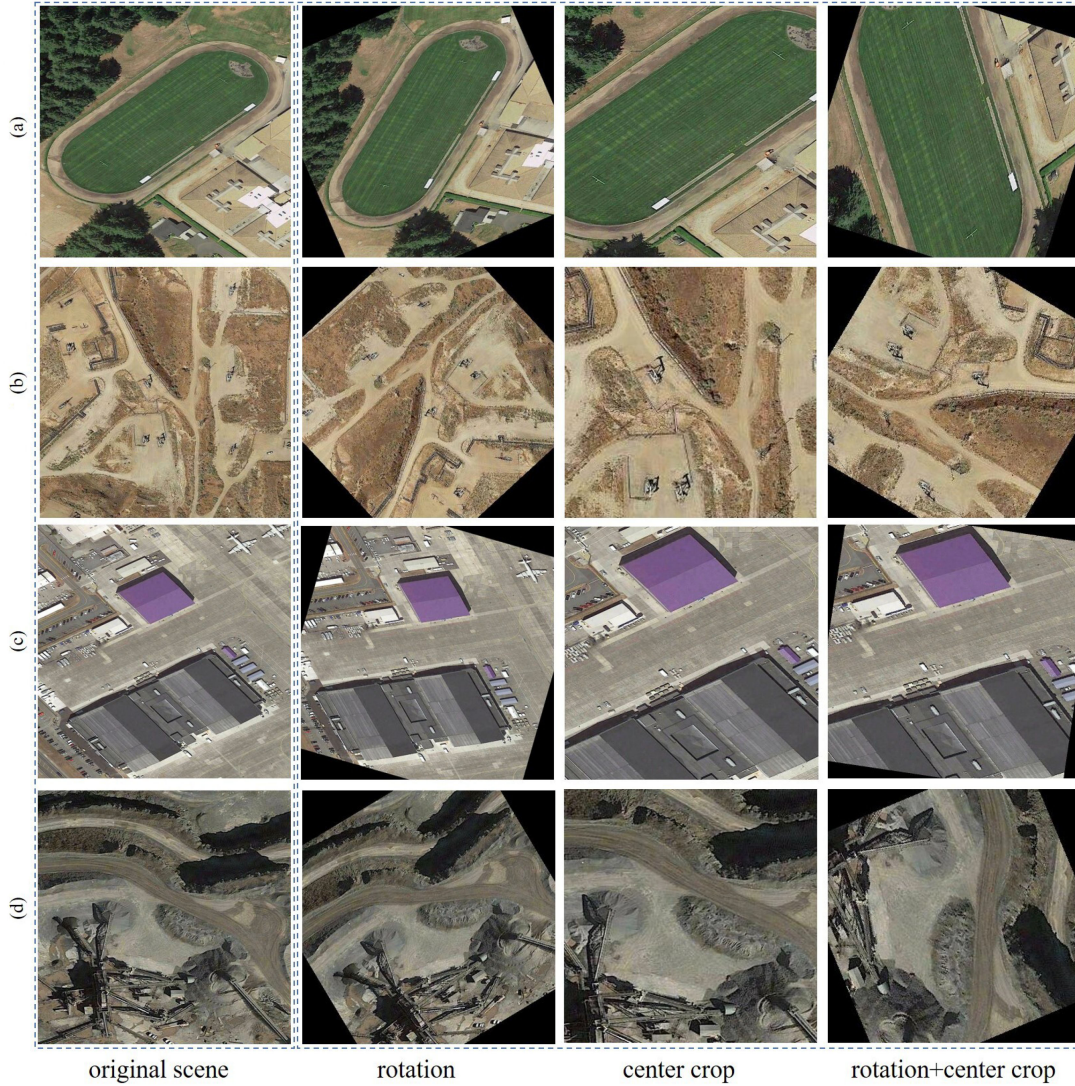


Fig. 5. Illustration of different scene images augmentation. (a) “Ground track field” scene image. (b) “Oil field” scene image. (c) “Apron” scene image. (d) “Quarry” scene image. From left to right, the first column is original scene image. The second–fourth columns are augmented images of rotation, center crop, combination of rotation, and center crop.

regions. The MGSL module includes the generation of multiple granularity annotations, classification model of multiple granularity semantics, and multiple granularity loss function.

- 1) *Generation of Multiple Granularity Annotations*: It is shown as the directed tree in Fig. 3. Let $D = (V, H_d)$ represent the hierarchical relations of label classes and their father and grandfather classes. The target prediction classes of the dataset are in leaf nodes. $V = v_1, v_2, \dots, v_n$ is the node set of the tree, corresponding to classes. $H_d \subseteq V \times V$ is a directed edge set, where $(v_i, v_j) \in H_d$ refers to a hierarchical relation and v_i is the father class of v_j . Let $X = \{x_i | i = 1, 2, \dots, N\}$ be the set of remote sensing scene images. $Y = \{y_i | i = 1, 2, \dots, N\}$ is the corresponding label set of X , where N is the number of samples of all categories. $C = \{c_i | i = 1, 2, \dots, n\}$ is the set of all classes of remote sensing scenes, including target prediction classes in leaf nodes and their father and grandfather

classes, where n is the number of classes. The multiple granularity semantic annotations of one remote sensing scene x_i can be represented as class nodes of one directed path from the root to a leaf node in the directed tree. The map relation is

$$y_i^k = T(x_i; k; C) \quad (4)$$

where $T(\cdot)$ is the map relation, K is the height of the directed tree, and k is the k th level of the tree and the granularity semantic annotation of x_i . The root node is at the zeroth level and is the first node of each directed path. The directed tree transforms the map relation of remote sensing scenes and their labels from one-to-one to one-to-many. Multiple annotations include fine-grained local semantic information and coarse-grained global semantic information of each remote sensing scene image.

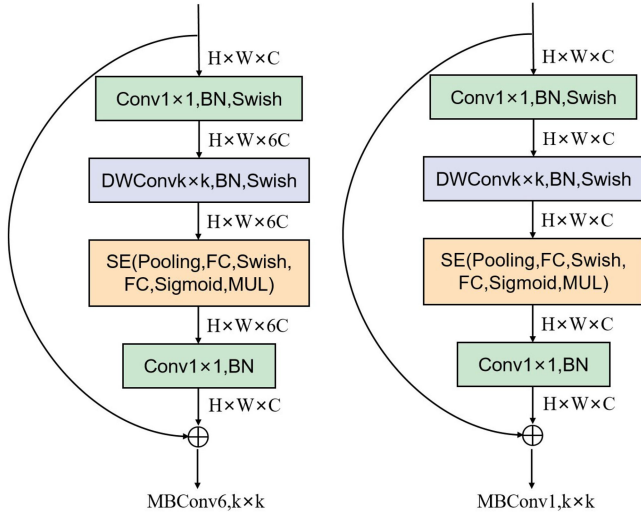


Fig. 6. Submodule structure of EfficientNet, MBCConv6, and MBCConv1, which differ by whether the dimension of the input feature increases after 1×1 Conv.

2) Classification Model of a Multiple Granularities Semantic:

It has multiple branches. As shown in Fig. 4, when the size of an input remote sensing scene image is 300×300 , the dimension of the feature map Z^O and feature extraction module output is 2560. For a directed tree whose height is K , K branches are used to learn K granularity semantic information of remote sensing scenes, each a subclassifier. The k th branch is represented as Branch- k , corresponding to the k th granularity semantic of scenes. We use a K value of 3. The branches are Branch-1, Branch-2, and Branch-3. The corresponding semantic granularity is from coarse to fine. Feature map Z^O is first transformed with 1×1 Conv before it is input to each branch. Because Branch-1 is used to capture the global feature of the scene, the input feature map is directly transformed with average pooling. Branch-2 and Branch-3 capture fine-grained local features. Input features are transformed with 1×1 Conv, followed by average pooling. After these transformations, the different branch output features are $\{f_{B1}^O, f_{B2}^O, f_{B3}^O\}$. After two fully-connected layers and a softmax function, the prediction probabilities of different semantic granularities are obtained. Their combination is the final prediction probability in multidimensional space.

The proposed classification model differs from other remote sensing scene classification methods [47], due to the layering and balance of different granularities of semantic learning. As shown in Fig. 4, the learning of different granularity semantics is separate, and they do not compete. They are in parallel and layering. Different granularity semantic learning losses are considered in each backpropagation, which focuses the network's attention on different granularity semantic and scale regions tend to be balanced.

3) Multiple Granularity Loss Function:

It is an extension of a cross-entropy loss [48], [49] function in high-dimensional space, which helps the network learn multiple granularity semantic information in the training stage. As shown in

(4)–(6), let x_i represent a remote sensing scene image, y_i is the label vector, and p_i is the prediction probability vector. We represent convolutional neural network parameters as W and B . $G(\cdot)$ is the map relation of remote sensing scene x_i and its label y_i . $P(\cdot)$ is the map relation of remote sensing scene x_i and its prediction probability p_i . Most DCN-based remote sensing scene classification methods learn distinguishable features by directly minimizing the gap between $P(\cdot)$ and $G(\cdot)$, $J(X; W; B)$. But these methods can only learn single granularity semantic features, and when the semantic granularities of categories differ, they compete, which leads to unstable and even wrong features learning.

$$y_i = G(x_i; C) \quad (5)$$

$$p_i = P(x_i; W; B) \quad (6)$$

$$J(X; W; B) = \min \left(-\frac{1}{N} \sum_{i=1}^N P(x_i; W; B) \log G(x_i; C) \right). \quad (7)$$

We regard single semantic granularity learning as 1-D optimization. The proposed multiple granularities loss function connects several optimization questions by a certain relationship to construct an optimization model of multiple granularities semantic learning in high-dimensional space. The k th prediction probability of remote sensing scene x_i is

$$p_i^k = P^k(x_i; k; W; B). \quad (8)$$

The k th semantic granularity optimization objective is $J_k(X; W; B)$. To simplify, we use a linear relationship $\sum(\cdot)$ to connect different semantic granularity optimization questions to construct a multiple granularity semantic learning optimization objective. And, the objective is

$$\begin{aligned} J^K(X; W; B) &= \min \left(\sum_{k=1}^K J_k(X; W; B) \right) \\ &= \min \left(\sum_{k=1}^K \left(-\frac{1}{N} \sum_{i=1}^N \right. \right. \\ &\quad \left. \left. (T(x_i; k; C; G) \log P^k(x_i; k; W; B)) \right) \right). \end{aligned} \quad (9)$$

By minimizing the multiple granularities loss function, the proposed method can learn multiple granularity semantic features at the same time, including fine-grained local region features and coarse-grained global region features.

IV. EXPERIMENTS

A. Dataset

We evaluate our approach on two public and popular datasets, which are introduced as follows.

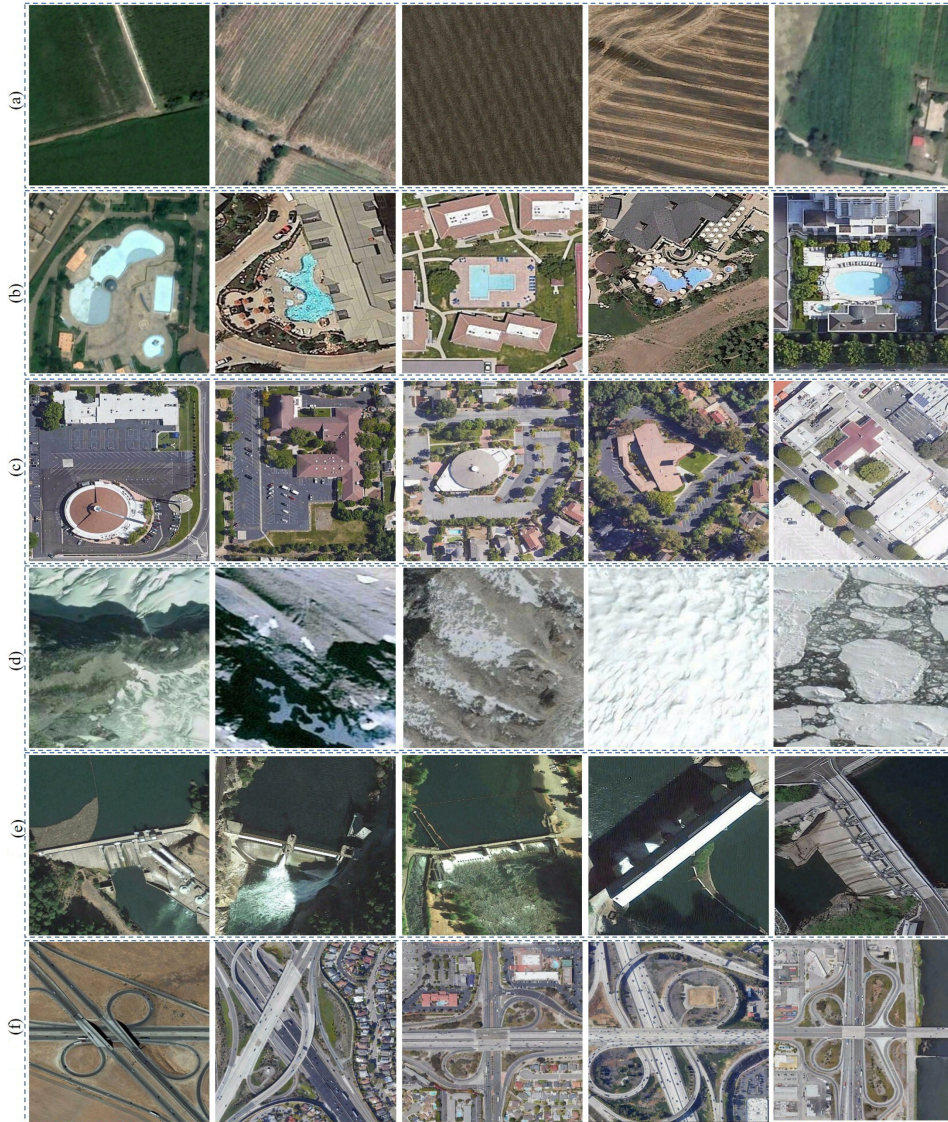


Fig. 7. Some scene image examples of Million-AID. (a) Images of scene “dry field”. (b) Images of scene “swimming pool”. (c) Images of scene “church”. (d) Images of scene “ice land”. (e) Images of scene “dam”. (f) Images of scene “viaduct”.

- 1) *Million-AID* [13]: It is a large-scale remote sensing image scene classification dataset, which includes 51 scene categories. Some scene image examples are shown in Fig. 7, for example, images of scene “dry field,” images of scene “swimming pool,” images of scene “church,” images of scene “ice land,” images of scene “dam,” and images of scene “viaduct”. There were 50000 remote sensing scene images in the training set and 10000 images in the test set. The minimum number of samples of different scene categories was 31 and the maximum was 6197. The minimum size of a scene image was 110×110 and the maximum was 11211×11211 . Some scenes had fine-grained semantic granularity annotation and some were coarse-grained.
- 2) *NWPU-RESISC45*: It is a large-scale remote sensing image scene classification dataset with rich image variations and high interclass similarity. It consists of 31 500 remote sensing images divided into 45 scene classes whose size

is 256×256 . The spatial resolution varies from 30 to 0.2 m per pixel for most scene classes. All images are extracted from Google Earth. Following the setting of previous methods on the NWPU-RESISC45 dataset, we randomly select 10% or 20% samples as the training data and the rest samples are served as the testing data.

B. Experimental Settings

The baseline of the proposed method is FixEfficientNet, whose default train image size is 224×224 and test image size is 256×256 . Its research shows that when the train image size is less than that of the test image size, the performance of a deep model is better.

- 1) *For the Million-AID Dataset*: Since the size of most images is about 300×300 , the train image size in the proposed method is set to 300×300 . Following the scale

TABLE I
EXPERIMENTAL RESULTS OF METHODS ON MILLION-AID. B-1, B-2, AND B-3 ARE DIFFERENT GRANULARITY SEMANTIC LEARNING BRANCHES. THE SEMANTIC GRANULARITY OF BRANCH-1, BRANCH-2, AND BRANCH-3 RANGES FROM COARSE TO FINE. "NUA" DENOTES THE PROPOSED NONUNIFORM SAMPLING AUGMENTATION MODULE

Method	Image Resolution		Modules				Metrics		Params (MB)
	Training size	Test size	B-1	B-2	B-3	NUA	Top-1 (%)	Top-5 (%)	
ResNet50	224 × 224	224 × 224			✓		45.64	62.08	90.07
ResNet101	224 × 224	224 × 224			✓		53.16	68.67	162.52
ResNet152	224 × 224	224 × 224			✓		46.18	63.24	222.20
RegNet-200MF	224 × 224	224 × 224			✓		60.99	88.54	11.14
RegNet-400MF	224 × 224	224 × 224			✓		61.88	89.55	15.37
RegNet-600MF	224 × 224	224 × 224			✓		62.75	89.19	19.51
RegNet-800MF	224 × 224	224 × 224			✓		62.95	89.08	21.11
RegNet-1.6GF	224 × 224	224 × 224			✓		63.30	88.75	42.88
RegNet-3.2GF	224 × 224	224 × 224			✓		63.38	90.48	53.81
EfficientNet-B0	224 × 224	224 × 224			✓		62.07	92.34	15.54
EfficientNet-B1	224 × 224	224 × 224			✓		61.05	92.53	25.10
EfficientNet-B2	224 × 224	224 × 224			✓		61.35	91.52	29.65
EfficientNet-B3	224 × 224	224 × 224			✓		60.68	92.59	41.10
EfficientNet-B4	224 × 224	224 × 224			✓		62.19	93.13	67.29
EfficientNet-B5	224 × 224	224 × 224			✓		61.53	92.57	108.51
EfficientNet-B6	224 × 224	224 × 224			✓		61.81	92.56	155.84
EfficientNet-B7	224 × 224	224 × 224			✓		62.19	92.55	243.83
FixEfficientNet	300 × 300	350 × 350			✓		69.5	92.55	243.83
MGSN (ours)	300 × 300	350 × 350			✓	✓	77.57	97.57	243.83
MGSN (ours)	300 × 300	350 × 350		✓	✓		76.96	97.58	244.03
MGSN (ours)	300 × 300	350 × 350	✓		✓		73.99	97.03	243.98
MGSN (ours)	300 × 300	350 × 350	✓	✓	✓		77.10	97.64	244.18
MGSN (ours)	300 × 300	350 × 350	✓	✓	✓	✓	80.13	98.02	244.18

The top three values of Top-1 and Top-5 accuracy are made in bold.

factor between train and test image size in FixEfficientNet, test image size of the proposed method is set to $300 \times \frac{256}{224} \approx 350$. And, we keep the train and test image size same in baseline and ablation study experiments. For compared methods, we follow their default setting and keep the input image size 224×224 in the training and testing stage.

- 2) *For the NWPU-RESISC45 Dataset:* Since all the image size is 256×256 , following the setting in the baseline method, the train and test image size in the proposed method is set to 224×224 and 256×256 , respectively. And, we compare the proposed method with some state-of-the-art methods published from 2017 to the present.

The initialization parameters of the feature extraction module in the proposed method were from pretrained EfficientNet-B7 on the ImageNet dataset. The height of the directed tree K in the generation of multiple granularity semantic annotations was 3. The proposed method was implemented based on the PyTorch deep learning framework. All model training and testing were on an NVIDIA TITAN X GPU.

C. Evaluation Metrics

The metrics of experiments included top-1 accuracy (overall accuracy) and top-5 accuracy, which are widely used in image classification tasks [10], [12]. Top-1 accuracy was used to evaluate overall performance. In the top-5 accuracy definition, when the top-5 prediction probability includes the target class, the prediction is considered correct. To compare the complexity

of different methods, their parameter size was used as a metric, which was output by the summary function of PyTorch.

D. Comparison With State-of-the-Art Methods

1) *Results on the Million-AID Dataset:* The proposed method was compared with ResNet50, ResNet101, ResNet152, RegNet-200MF, RegNet-400MF, RegNet-600MF, RegNet-800MF, RegNet-1.6GF, RegNet-3.2GF, EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, EfficientNet-B4, EfficientNet-B5, EfficientNet-B6, EfficientNet-B7, and FixEfficientNet, with different structures of ResNet [19], RegNet [21], and EfficientNet [20]. The baseline was FixEfficientNet, a state-of-the-art method, and its feature extraction module was EfficientNet-B7, pretrained on the ImageNet dataset [14].

Experimental results on Million-AID [13] are given in Table I, where B-1, B-2, and B-3 refer to Branch-1, Branch-2, and Branch-3, respectively, which have semantic granularity ranging from coarse to fine. The classifier module for comparison methods is the same as Branch-3. "NUA" indicates the nonuniform sampling augmentation module. Among the methods, our baseline FixEfficientNet achieved the best top-1 accuracy of 69.5%, and its top-5 accuracy was 92.55%. EfficientNet-B4 had the best top-5 accuracy, 93.13%, and its top-1 accuracy was 62.19%. Compared with these methods, our proposed MGSN had 80.13% top-1 accuracy and 98.02% top-5 accuracy, which was best. The gap between top-1 and top-5 accuracy of all methods shows the importance of fine-grained distinguishing

TABLE II
TOP-1 ACCURACY (%) OF STATE-OF-THE-ART METHODS ON THE NWPU-RESISC45 DATASET WITH DIFFERENT TRAINING SAMPLE RATIOS (10% AND 20%).
ALL GRANULARITY SEMANTIC LEARNING BRANCHES ARE USED IN THE PROPOSED METHOD

Method	Year	Publication	Train ratios(%)	
			10%	20%
BoCF [38]	2017	IEEE GRSL	82.65 ± 0.31	84.32 ± 0.17
SPP-net [33]	2017	Remote Sensing	82.13 ± 0.30	84.64 ± 0.23
Fine-tuned AlexNet [50]	2017	Proceedings of the IEEE	81.22 ± 0.19	85.16 ± 0.18
Fine-tuned VGGNet-16 [50]			87.15 ± 0.45	90.36 ± 0.18
Fine-tuned GoogleNet [50]			82.57 ± 0.12	86.02 ± 0.18
MARTAGANs [51]	2017	IEEE GRSL	68.63 ± 0.22	75.03 ± 0.28
MSCP [52]	2018	IEEE TGRS	88.07 ± 0.18	90.81 ± 0.13
D-CNNs [39]	2018	IEEE TGRS	89.22 ± 0.50	91.89 ± 0.22
IORN [53]	2018	IEEE GRSL	87.83 ± 0.16	91.30 ± 0.17
ADSSM [54]	2018	IEEE TGRS	91.69 ± 0.22	94.29 ± 0.14
SF-CNN [55]	2019	IEEE TGRS	89.89 ± 0.16	92.55 ± 0.14
ADFF [56]	2019	Remote Sensing	90.58 ± 0.19	91.91 ± 0.23
CNN-CapsNet [57]	2019	Remote Sensing	89.03 ± 0.21	89.03 ± 0.21
Siamese ResNet50 [58]	2019	IEEE GRSL	—	92.28 ± 3.78
EfficientNet-B0-aux [59]	2019	Remote Sensing	89.96 ± 0.27	92.89 ± 0.16
Attention GANs [60]	2019	IEEE TGRS	72.21 ± 0.21	77.99 ± 0.19
SCCov [6]	2020	IEEE TNNLS	89.30 ± 0.35	92.10 ± 0.25
MF ² Net [37]	2020	IEEE GRSL	85.54 ± 0.36	89.76 ± 0.14
RADC-Net [61]	2020	Neurocomputing	85.72 ± 0.25	87.63 ± 0.28
TFADNN [36]	2020	Information Sciences	87.78 ± 0.11	90.86 ± 0.24
MG-CAP(Sqrt-E) [62]	2020	IEEE TIP	90.83 ± 0.12	92.95 ± 0.11
IB-CNN(M) [63]	2020	IEEE GRSL	90.49 ± 0.17	93.33 ± 0.21
ResNet-50+EAM [44]	2021	IEEE GRSL	90.87 ± 0.15	93.51 ± 0.12
VGG-VD16+SAFF [35]	2021	IEEE GRSL	84.38 ± 0.19	87.86 ± 0.14
VGG-VD16+MICP [64]	2021	Neurocomputing	87.54 ± 0.31	90.49 ± 0.28
AGMFA-Net [1]	2021	Remote Sensing	91.01 ± 0.18	93.70 ± 0.08
Lie Group [65]	2021	Remote Sensing Letters	90.19 ± 0.11	93.21 ± 0.12
MGML-FENet(VGG16) [66]	2021	IEEE TNNLS	90.69 ± 0.14	93.36 ± 0.12
ACR-MLFF [34]	2022	IEEE GRSL	90.01 ± 0.33	92.45 ± 0.20
Fine-tuned ResNet50 [67]	2022	IEEE GRSL	86.40 ± 0.6	90.13 ± 0.4
Fine-tuned ResNet101 [67]			86.47 ± 0.4	89.91 ± 0.6
Fine-tuned Inception V3 [67]			86.90 ± 0.6	90.57 ± 0.5
TResNet-M [67]			88.05 ± 0.8	91.19 ± 0.4
MGSN(ours)	2022	IEEE J-STARS	91.92 ± 0.12	94.33 ± 0.08

ability. Compared with the baseline FixEfficientNet, the proposed MGSN improved top-1 accuracy by 10.63% and top-5 accuracy by 5.47%, showing that MGSN can reduce the semantic granularity difference and class sample imbalance in remote sensing scene image classification.

Table I gives that when remote sensing scenes are complex, structure, such as depth and complexity, of models is not the main and only factor. Comparing ResNet50, ResNet101, and ResNet152, the depth, from 50 to 152 layers, and the parameters, from 90.07 to 222.2 MB, are gradually increasing. Their accuracy increases at first, from 45.64% top-1 accuracy of ResNet50 to 53.16% for ResNet101, and then decreases, from 53.16% top-1 accuracy of ResNet101 to 46.18% for ResNet152. From RegNet-200MF to RegNet-3.2GF, the model complexity increases, as does the classification top-1 accuracy. However, the improvement is small. Among EfficientNet-B0–B7, EfficientNet-B7 achieved the best top-1 accuracy, 62.19%. Compared with EfficientNet-B0, EfficientNet-B7 improved top-1 accuracy by 0.12%, but it is 16-times more complex. These results show that the improvement of careful DCNs structure design is limited in remote sensing scenes with semantic granularity difference and class sample imbalance challenges.

The effectiveness of MGSN shows the superiority of the MGSL and NUA modules in guiding the network to learn robust features with correct semantics, and that an effective learning objective can guide a network to learn the multiscale region features on single-scale feature maps.

2) *Results on the NWPU-RESISC45 Dataset:* The comparison with some state-of-the-art methods published from 2017 to the present on the NWPU-RESISC45 dataset is given in Table II. Under 10% and 20% training samples, the classification accuracy of the proposed method reaches 91.92% and 94.33%, which surpasses all comparison methods. It shows the superiority and stability of the proposed method. ADSSM [54] and MGML-FENet [66] also improve the classification accuracy from the perspective of multiple granularities semantic features learning. ADSSM achieves the second-highest accuracy, 91.69% and 94.29%, when 10% and 20% samples are used to train the model. The difference between the proposed method with ADSSM is that ADSSM regards the features of different levels, such as low-level, mid-level, and high-level features, as multiple granularities semantic information and merges them in feature space. It needs to mix a variety of hand-designed features, such as SIFT, visual dictionary, and so on, and deep

TABLE III
 ABLATION STUDY RESULTS ON THE MILLION-AID DATASET. B-1, B-2, AND B-3 ARE DIFFERENT GRANULARITY SEMANTIC LEARNING BRANCHES, AND THE SEMANTIC GRANULARITY OF BRANCH-1, BRANCH-2, AND BRANCH-3 RANGES FROM COARSE TO FINE. "NUA" DENOTES THE PROPOSED NONUNIFORM SAMPLING AUGMENTATION MODULE.

Model	B-1	B-2	B-3	NUA	Top-1 (%)	Top-5 (%)
EfficientNet-B7			✓		62.19	92.55
FixEfficientNet (baseline)			✓		69.5	92.55
FixEfficientNet (baseline)			✓	✓	77.57	97.57
MGSN (ours)		✓	✓		76.96	97.58
MGSN (ours)	✓		✓		73.99	97.03
MGSN (ours)	✓	✓	✓		77.10	97.64
MGSN (ours)	✓	✓	✓	✓	80.13	98.02

learning features. It is more complicated. MGML-FENet extracts multiple granularities semantic features from different stages of a deep learning model. It achieves 90.69% and 93.36% classification accuracy under 10% and 20% samples training, respectively, which is lower than the proposed method. It shows that our proposed method is better in the learning of multiple granularities semantic features.

E. Ablation Study

Ablation study results on the Million-AID dataset are given in Table III. B-1, B-2, and B-3 refer to Branch-1, Branch-2, and Branch-3, respectively, the semantic granularity from coarse to fine. NUA is the proposed nonuniform sample augmentation module. The feature extraction module of the baseline FixEfficientNet is from EfficientNet-B7.

B-3, B-2+B-3, and B-1+B-2+B-3 refer to models with combinations of different semantic granularity learning branches, and "✓" represents that the corresponding module is used in a model. Compared with a model using only semantic granularity learning branch B-3, the top-1 accuracy of model B-2+B-3 improved from 69.5% to 76.96% and the top-5 accuracy from 93.55% to 97.58%.

When three semantic granularity learning branches were used, i.e., B-1+B-2+B-3, the top-1 accuracy improved from 69.5% to 77.10%, and the top-5 accuracy from 92.55% to 97.64%. The visualization of the confusion matrix is shown in Fig. 8. It is observed that our method performs well in most categories, such as wastewater plant, golf course, stadium, parking lot, pier, and so on. These results show the effectiveness of the proposed MGSL method, and the performance gradually improves with the increase of semantic granularity. Compared with the model using only B-3, the top-1 accuracy of B-2+B-3 improved by 7.46%. Compared with B-2+B-3, the top-1 accuracy of B-1+B-2+B-3 improved by 0.14%. These results show that the relation between the improvement and the semantic granularity is nonlinear. Comparing experimental results of B-2+B-3 versus B-1+B-3 shows that the finer the extra semantic granularity, the more obvious the improvement in classification accuracy.

To verify the effectiveness and generalization of NUA, ablation study experiments on Million-AID were carried out on the baseline FixEfficientNet and the proposed MGSN. The top-1

accuracy of FixEfficientNet with NUA improved from 69.5% to 77.57%, a difference of 8.07%. Compared with MGSN without NUA (B-1+B-2+B-3), the top-1 accuracy of MGSN with NUA (B-1+B-2+B-3+NUA) improved by 3.03%, and the top-5 accuracy by 0.38%. This shows that, based on MGSL, the NUA module can further improve network performance. The NUA module can be directly used in other remote sensing scene image classification methods.

Discussion of the superiority of the NUA module: The goal of the NUA module is to alleviate the imbalance of the sample number of different classes. And, the kernel idea is to unevenly change their sample number and make them balanced. Common strategies that change sample numbers are data augmentation and oversampling. However, data augmentation will enlarge the imbalance. Because it scales the sample of different classes in the same factor. For example, class A is with five samples and class B is with 20 samples. After five different augment transformations, class A is with 25 samples and class B is with 100 samples. The imbalance is larger. And, oversampling usually leads to overfitting since sampled samples are usually repeated. Then, we combined them together, designing the NUA module. There are following two advantages.

- 1) It scales the sample of different classes in a different factor and makes them balanced.
- 2) Different types of augmentation provide additional different samples and it avoids overfitting.

The main innovation of the NUA module is that it alleviates the imbalance challenge in remote sensing image scene classification from the perspective of data augmentation. And, it can boost the classification accuracy and is with good generalization. As given in Table III, the NUA module boosts the top-1 accuracy of baseline by 8.07%. With the NUA module, our MGSN is further improved by 3.03%, from 77.10% to 80.13%. It shows the superiority of the proposed NUA module.

V. DISCUSSION

To fully understand the impact of multiple granularity semantics on the network, we perform an analysis from two aspects on the Million-AID dataset: 1) interaction between different semantic granularity learning; and 2) feature map visualization of models guided by different granularity semantics.

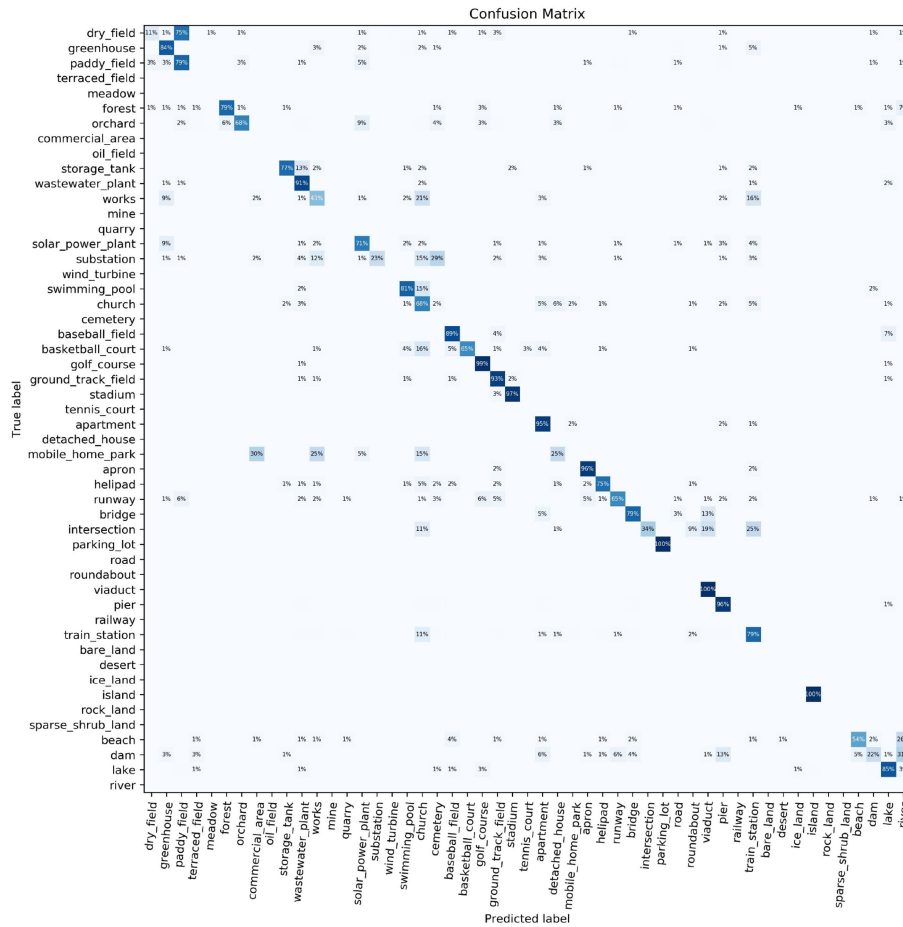


Fig. 8. When three semantic granularity learning branches were used, MGSN with B-1+B-2+B-3, visualization of confusion matrix on the Million-AID dataset.

TABLE IV

EXPERIMENTAL RESULTS OF PROPOSED METHOD GUIDED WITH DIFFERENT SEMANTIC GRANULARITY. THE RESULTS SHOW THE ACCURACY OF EACH SEMANTIC GRANULARITY LEARNING BRANCH INSTEAD OF ONLY THAT OF TARGET CLASS PREDICTION BRANCHES. B-1, B-2, AND B-3 DENOTE BRANCH-1, BRANCH-2, AND BRANCH-3, RESPECTIVELY, WHOSE SEMANTIC GRANULARITY RANGES FROM COARSE TO FINE

Model	Branch	Top-1 (%)	Top-5 (%)
FixEfficientNet (baseline)	B-3	69.5	92.55
MGSN (ours)	B-2	87.68	98.30
	B-3	76.96	97.58
MGSN (ours)	B-1	88.89	99.79
	B-3	73.99	97.03
MGSN (ours)	B-2	87.93	98.73
	B-1	91.09	99.94

and not only target-class prediction accuracy. Comparing the classification accuracy of each semantic granularity when modeled with different combinations of learning branches, their interaction is shown. For FixEfficientNet with B-3, MGSN with B-2+B-3, and MGSN with B-1+B-3, the top-1 accuracy of B-3 was 69.5%, 76.96%, and 73.99%, respectively. This shows that coarse-grained semantic learning, i.e., B-1 or B-2, helps to improve the learning of fine-grained semantic learning B-3. The top-1 accuracy was improved by 4.49% and 7.46%. Hence, the finer the granularity of the extra coarse-grain semantic, the more obvious the improvement in accuracy. Comparing MGSN with B-2 and MGSN with B-2 and B-3, the top-1 accuracy of B-2 decreased by 0.25%. Comparing MGSN with B-1 and MGSN with B-1 and B-3, the top-1 accuracy of B-1 decreased by 2.2%. This shows that fine-grained semantic learning can inhibit the learning of coarse-grained semantic learning.

A. Interaction Between Different Semantic Granularity Learning

Table IV gives experimental results of models guided by different semantic granularity, where B-1, B-2, and B-3 range from coarse- to fine-grain. Different from Tables III and IV gives the classification accuracy of each semantic granularity

B. Feature Map Visualization of the Model Guided by Different Granularity Semantics

To intuitively analyze features of network learning guided by different granularity semantic information, we visualized the output feature maps of their last layers. Figs. 9–12 show scene images of different categories, “wastewater plant,” “golf

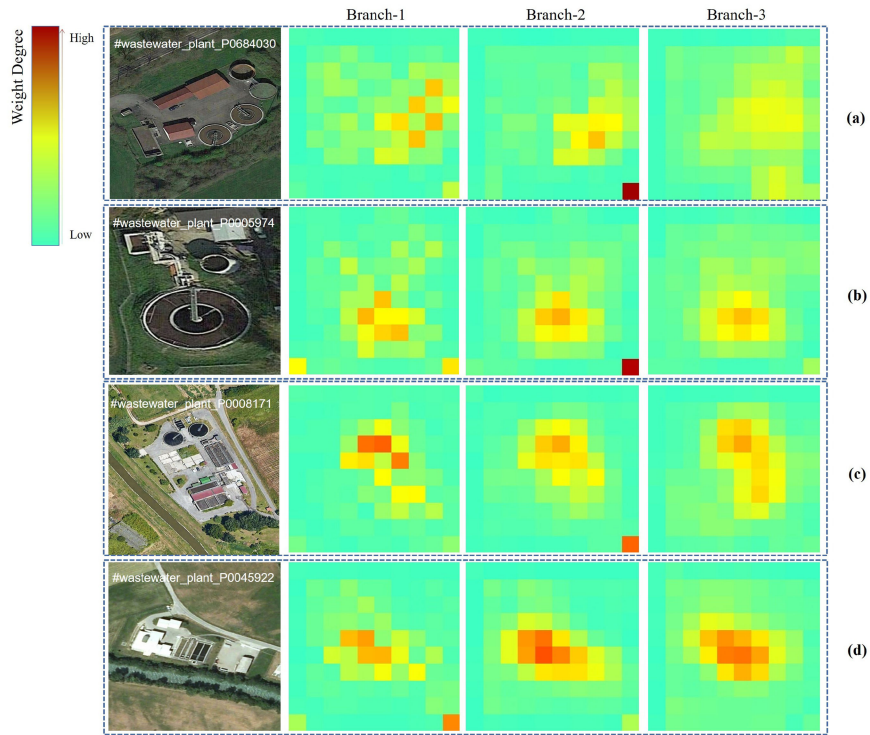


Fig. 9. For scene “wastewater plant” on the Million-AID dataset, visualization of feature map last layer output of models guided by different granularity semantic information. Branch-1, Branch-2, and Branch-3 are different granularity semantic learning branches. (a)–(d) show different remote sensing scene images.

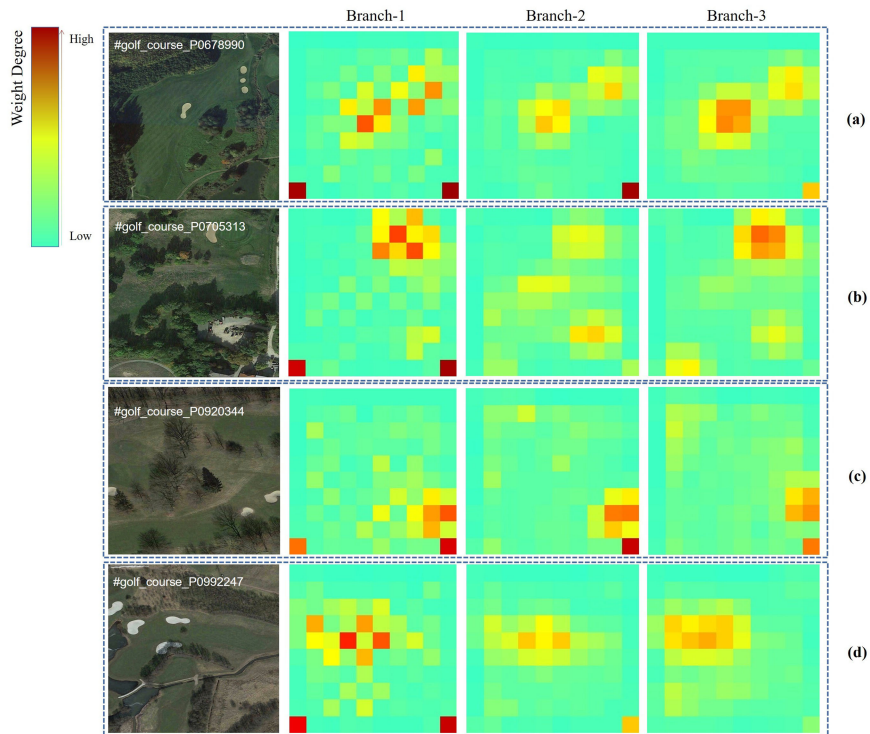


Fig. 10. For scene “golf course” on the Million-AID dataset, visualization of feature map last layer output of models guided by different granularity semantic information. Branch-1, Branch-2, and Branch-3 are different granularity semantic learning branches. (a)–(d) show different remote sensing scene images.

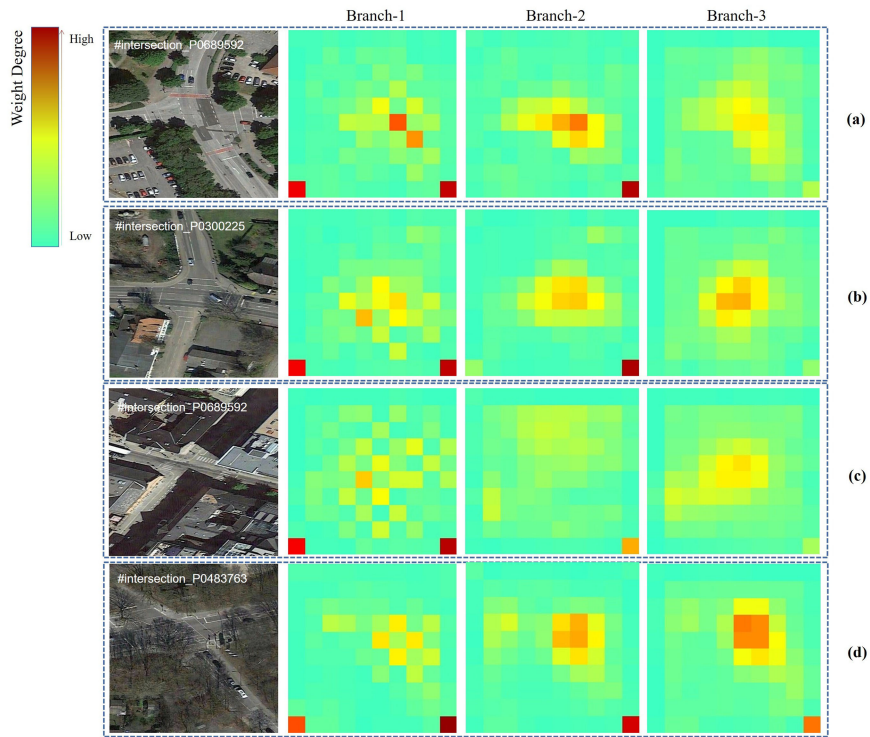


Fig. 11. For scene “intersection” on the Million-AID dataset, visualization of feature map last layer output of models guided by different granularity semantic information. Branch-1, Branch-2, and Branch-3 are different granularity semantic learning branches. (a)–(d) show different remote sensing scene images.

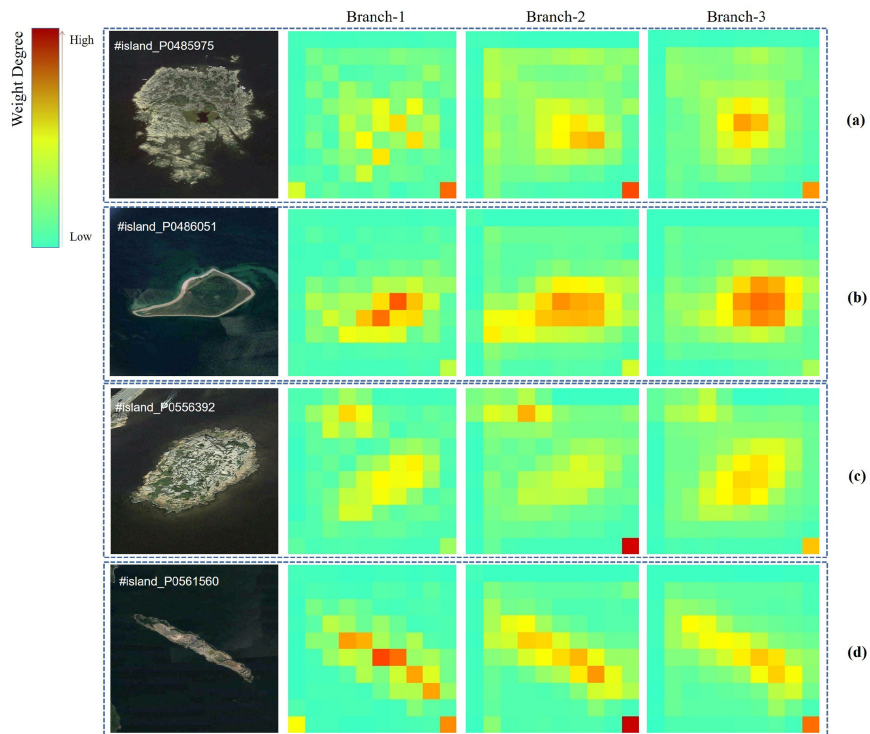


Fig. 12. For scene “island” on the Million-AID dataset, visualization of feature map last layer output of models guided by different granularity semantic information. Branch-1, Branch-2, and Branch-3 are different granularity semantic learning branches. (a)–(d) show different remote sensing scene images.

course,” “intersection,” and “island,” and their output feature map visualization images when using different granularity semantic learning branches, including Branch-1, Branch-2, and Branch-3. In feature map visualization images, a darker color indicates higher attention of the network to objects in the region. The visualization of feature maps shows that coarse-grained semantic learning branches B-1 and B-2 pay more attention to the global content of scenes. The dark region in the feature map visualization image of B-1 is larger than that of B-2. Fine-grained semantic learning branch B-3 focuses on local key region features, and the dark region is concentrated in the local areas. The visualization of feature maps also shows that in the proposed MGSL method, different granularity semantic learning branches learn different scale region features, including global and local region features.

VI. CONCLUSION

We proposed MGSL and NUA to reduce semantic granularity differences and class sample imbalance in remote sensing scene image classification. The MGSL module improves network performance on scenes with semantic granularity differences. We showed that coarse-grain semantic feature learning improves fine-grained semantic feature learning, while fine-grained semantic feature learning can inhibit coarse-grained semantic feature learning. The NUA module combines oversampling and sample augmentation to balance samples with different numbers of categories to avoid overfitting. It is effective at improving the robustness of feature learning and the overall performance of the network. Based on this study, we find that feature fusion may reduce granularity differences in remote sensing scene semantics. Coarse-grained semantics of scenes focus on shallow features, and fine-grained semantics on deep features. Combining them may effectively address the challenge of semantic granularity differences of remote sensing scenes. We plan to carry out related research from this consideration.

ACKNOWLEDGMENT

The authors would like to acknowledge the Key Laboratory of Space Utilization and Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, for offering strong support throughout the experiments.

REFERENCES

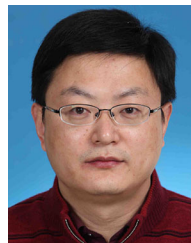
- [1] M. Li, L. Lei, Y. Tang, Y. Sun, and G. Kuang, “An attention-guided multilayer feature aggregation network for remote sensing image scene classification,” *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3113.
- [2] Q. Zeng, J. Geng, K. Huang, W. Jiang, and J. Guo, “Prototype calibration with feature generation for few-shot remote sensing image scene classification,” *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2728.
- [3] Z. Zhou *et al.*, “NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3228–3242, Mar. 2021, doi: [10.1109/JSTARS.2021.3063096](https://doi.org/10.1109/JSTARS.2021.3063096).
- [4] X. Zheng, L. Qi, Y. Ren, and X. Lu, “Fine-grained visual categorization by localizing object parts with single image,” *IEEE Trans. Multimedia*, vol. 23, pp. 1187–1199, Sep. 2021, doi: [10.1109/TMM.2020.2993960](https://doi.org/10.1109/TMM.2020.2993960).
- [5] H. Xie, Y. Chen, and P. Ghamisi, “Remote sensing image scene classification via label augmentation and intra-class constraint,” *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2566.
- [6] 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.
- [7] M. Berman, L. Pishchulin, N. Xu, M. B. Blaschko, and G. Medioni, “AOWS: Adaptive and optimal network width search with latency constraints,” in *Proc. IEEE/CVF Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11214–11223.
- [8] J. Yu *et al.*, “Bignas: Scaling up neural architecture search with big-stage models,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 702–717.
- [9] Z. Lu, K. Deb, and V. N. Boddeti, “MUXConv: Information multiplexing in convolutional neural networks,” in *Proc. IEEE/CF Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12041–12050.
- [10] H. Zhu, Z. An, C. Yang, X. Hu, K. Xu, and Y. Xu, “Efficient search for the number of channels for convolutional neural networks,” in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.
- [11] B. Zhao, H. Xiong, J. Bian, Z. Guo, C.-Z. Xu, and D. Dou, “COMO: Efficient deep neural networks expansion with convolutional MaxOut,” *IEEE Trans. Multimedia*, vol. 23, no. 12, pp. 1722–1730, Jun. 2021.
- [12] T. Elsken, J. Metzger, and F. Hutter, “Neural architecture search: A survey,” *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.
- [13] Y. Long *et al.*, “On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and Million-AID,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, Apr. 2021, doi: [10.1109/JSTARS.2021.3070368](https://doi.org/10.1109/JSTARS.2021.3070368).
- [14] Y. Gu, H. Liu, T. Wang, S. Li, and G. Gao, “Deep feature extraction and motion representation for satellite video scene classification,” *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140307.
- [15] 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, pp. 3735–3756, Jun. 2020, doi: [10.1109/JSTARS.2020.3005403](https://doi.org/10.1109/JSTARS.2020.3005403).
- [16] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” *Adv. Condens. Matter Phys.*, vol. 25, pp. 1097–1105, 2012.
- [17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.
- [18] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [20] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [21] I. Radosavovic, R. Kosaraju, R. Girshick, K. He, and P. Dollár, “Designing network design spaces,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10425–10433.
- [22] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, “Fixing the train-test resolution discrepancy: FixEfficientNet,” 2020, *arXiv:2003.08237*.
- [23] Y. Ren *et al.*, “Full convolutional neural network based on multi-scale feature fusion for the class imbalance remote sensing image classification,” *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3547.
- [24] W. Xia *et al.*, “High-resolution remote sensing imagery classification of imbalanced data using multistage sampling method and deep neural networks,” *Remote Sens.*, vol. 11, no. 21, 2019, Art. no. 2523.
- [25] H. Han, W. Wang, and B. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced datasets learning,” in *Proc. Int. Conf. Intell. Comput.*, 2005, pp. 878–887.
- [26] C. Drummond, “Class imbalance and cost sensitivity: Why undersampling beats oversampling,” in *Proc. ICML-KDD Workshop: Learn. Imbalanced Datasets*, 2003, pp. 1–8.
- [27] J. Tan *et al.*, “Equalization loss for long-tailed object recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11662–11671.
- [28] H. Sun, S. Li, X. Zheng, and X. Lu, “Remote sensing scene classification by gated bidirectional network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
- [29] S. Yang and D. Ramanan, “Multi-scale recognition with DAG-CNNs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1215–1223.
- [30] M. Hayat, S. Khan, M. Bennamoun, and S. An, “A spatial layout and scale invariant feature representation for indoor scene classification,” *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4829–4841, Oct. 2016.
- [31] Y. Kalantidis, C. Mellina, and S. Osindero, “Cross-dimensional weighting for aggregated deep convolutional features,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 685–701.

- [32] G. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
- [33] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 848.
- [34] X. Wang, L. Duan, A. Shi, and H. Zhou, "Multilevel feature fusion networks with adaptive channel dimensionality reduction for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Apr. 2022, doi: [10.1109/LGRS.2021.3070016](https://doi.org/10.1109/LGRS.2021.3070016).
- [35] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2021.
- [36] K. Xu, H. Huang, P. Deng, and G. Shi, "Two-stream feature aggregation deep neural network for scene classification of remote sensing images," *Inf. Sci.*, vol. 539, no. 1, pp. 250–268, 2020.
- [37] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1894–1898, Nov. 2020.
- [38] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing imagescene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
- [39] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [40] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 13, pp. 7844–7853, Sep. 2021.
- [41] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
- [42] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Sigspatial Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
- [43] 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. 2019.
- [44] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1926–1930, Nov. 2021.
- [45] A. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [47] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.
- [48] W. Guo *et al.*, "Spectrum selection and deep feature fusion based hyperspectral image natural scene classification network," in *Proc. Glob. Intell. Ind. Conf.*, 2021, Art. no. 117800A.
- [49] 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.
- [50] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [51] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "Marta GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.
- [52] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.
- [53] J. Wang, W. Liu, L. Ma, H. Chen, and L. Chen, "IORN: An effective remote sensing image scene classification framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1695–1699, Nov. 2018.
- [54] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.
- [55] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.
- [56] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Attention-based deep feature fusion for the scene classification of high-resolution remote sensing images," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 1996.
- [57] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.
- [58] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.
- [59] Y. Bazi, M. Rahhal, M. M. H. Alhichri, and N. Alajlan, "Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification," *Remote Sens.*, vol. 11, no. 18, 2019, Art. no. 2908.
- [60] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.
- [61] Q. Bi, K. Qin, H. Zhang, Z. Li, and K. Xu, "Radc-Net: A residual attention based convolution network for aerial scene classification," *Neurocomputing*, vol. 377, no. 1, pp. 345–359, 2020.
- [62] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 29, no. 2, pp. 5396–5407, Apr. 2020.
- [63] E. Li, A. Samat, P. Du, W. Liu, and J. Hu, "Improved bilinear CNN model for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 3, pp. 1–5, Dec. 2020.
- [64] K. Qi, C. Yang, C. Hu, H. Zhai, Q. Guan, and S. Shen, "A multi-level improved circle pooling for scene classification of high-resolution remote sensing imagery," *Neurocomputing*, vol. 462, no. 1, pp. 506–522, 2021.
- [65] C. Xu, G. Zhu, and J. Shu, "Robust joint representation of intrinsic mean and kernel function of lie group for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 796–800, May 2021.
- [66] Q. Zhao, S. Lyu, Y. Li, Y. Ma, and L. Chen, "MGML: Multigranularity multilevel feature ensemble network for remote sensing scene classification," *IEEE Trans. Neural Nets. Learn. Syst.*, vol. 1, no. 1, pp. 121, Sep. 2021.
- [67] J. M. Haut, A. Alcolea, M. E. Paoletti, J. Plaza, J. Resano, and A. J. Plaza, "GPU-friendly neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Sep. 2022, doi: [10.1109/LGRS.2020.3019378](https://doi.org/10.1109/LGRS.2020.3019378).



Weilong Guo received the B.Eng. degree in software engineering from Jilin University, Jilin, China, in 2018. He is currently working toward the M.Eng. degree in computer applied technology with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

His research interests include intelligent analysis and understanding of image and video.



Shengyang Li received the M.Eng. degree in computer science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2003, and the Ph.D. degree in remote sensing image processing and analysis from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include computer vision, target detection and tracking in video satellite, and remote sensing image analysis and understanding.



Jian Yang received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2020. He is currently working toward the M.S. degree in signal and information processing with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include object tracking for unmanned aerial vehicles and intelligent video analysis.



Junjie Lu received the B.S. degree in information and computing sciences from North China Electric Power University, Beijing, China, in 2019. She is currently working toward the M.Eng. degree in electronics and communication engineering with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

Her research focuses on image and video understanding.



Zhuang Zhou received the B.Eng. degree in electrical engineering and automation from the China University of Mining and Technology, Xuzhou, China, in 2013, and the M.S. degree in cartography and geography information system from Beijing Normal University, Beijing, China, in 2016.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. His research focuses on remote sensing image classification.



Longxuan Kou received the B.S. degree in geographic information science from Beijing Forestry University, Beijing, China, in 2017. She is currently working toward the Ph.D. degree in computer applications with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

Her research interests include satellite video and conventional video analysis, such as object segmentation.



Yunfei Liu received the B.Eng. and M.S. degrees in computer science from Beijing Forestry University, Beijing, China, in 2013 and 2016, respectively.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. His research focuses on remote sensing data analysis.



Manqi Zhao received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include satellite video, unmanned aerial vehicle video, and conventional video analysis, with a focus on object tracking.