applied sciences

MDPI

Article

Semantic Multigranularity Feature Learning for High-Resolution Remote Sensing Image Scene Classification

Xinyi Ma [1], Zhifeng Xiao [2], Hong-sik Yun [1,*] and Seung-Jun Lee [1]

[1] Geo Informatics Lab, School of Civil & Architecture Engineering, Natural Science Campus, Sungkyunkwan University, Suwon 16419, Korea; maxinyi970318@gmail.com (X.M.); issue7942@naver.com (S.-J.L.)
[2] School of Engineering, Penn State Erie, The Behrend College, Erie, PA 16563, USA; zux2@psu.edu
* Correspondence: yoonhs@skku.edu

Abstract: High-resolution remote sensing image scene classification is a challenging visual task due to the large intravariance and small intervariance between the categories. To accurately recognize the scene categories, it is essential to learn discriminative features from both global and local critical regions. Recent efforts focus on how to encourage the network to learn multigranularity features with the destruction of the spatial information on the input image at different scales, which leads to meaningless edges that are harmful to training. In this study, we propose a novel method named Semantic Multigranularity Feature Learning Network (SMGFL-Net) for remote sensing image scene classification. The core idea is to learn both global and multigranularity local features from rearranged intermediate feature maps, thus, eliminating the meaningless edges. These features are then fused for the final prediction. Our proposed framework is compared with a collection of state-of-the-art (SOTA) methods on two fine-grained remote sensing image scene datasets, including the NWPU-RESISC45 and Aerial Image Datasets (AID). We justify several design choices, including the branch granularities, fusion strategies, pooling operations, and necessity of feature map rearrangement through a comparative study. Moreover, the overall performance results show that SMGFL-Net consistently outperforms other peer methods in classification accuracy, and the superiority is more apparent with less training data, demonstrating the efficacy of feature learning of our approach.

Keywords: scene classification; remote sensing; fine-grained; multigranularity

Citation: Ma, X.; Xiao, Z.; Yun, H.-s.; Lee, S.-J. Semantic Multigranularity Feature Learning for High-Resolution Remote Sensing Image Scene Classification. *Appl. Sci.* **2021**, *11*, 9204. https://doi.org/10.3390/app11199204

Academic Editors: Hyung-Sup Jung, Saro Lee, Kyung-Soo Han and No-Wook Park

Received: 16 July 2021
Accepted: 30 September 2021
Published: 3 October 2021

1. Introduction

Remote sensing (RS) refers to the practice of observing, recording, measuring, and deriving information about the Earth's land and water surfaces using images acquired from an overhead perspective [1]. With the development of RS technology in the past decades, tremendous high-resolution RS images are available. Meanwhile, corresponding research efforts towards intelligent understanding, identification, and classification of RS scene content have been actively investigated because the quality of scene interpretation determines the effects of numerous downstream applications, such as urban planning, traffic control, and land resource management. Specifically, the goal of RS image scene classification is to assign a semantic label to an RS image patch. The task is complex and challenging due to interclass similarities and variable shapes of ground objects. For example, Figure 1 shows some image samples in the NWPU-RESISC45 dataset. It is observed that the scene samples of Medium and Dense Residential are similar; further, for the four Palace samples, there is a semantic gap in the color, size, shape, and edge distributions. To better recognize these scene objects, both global and local features are crucial. For the samples in Figure 1, global statistical features help distinguish Medium Residential and Dense Residential, and local features are essential for recognizing the Palace.

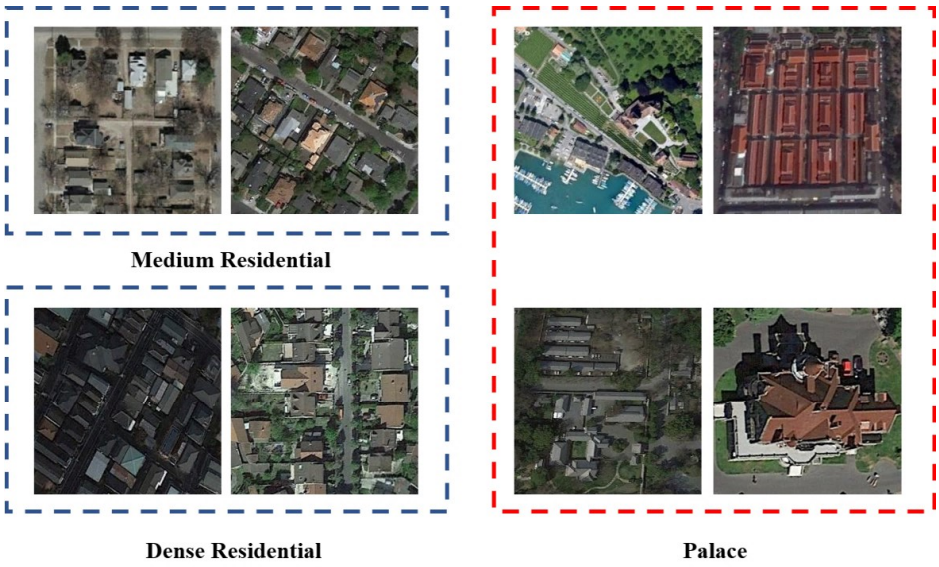

Figure 1. Images from NWPU-RESISC45.

Early representative features, such as Scale-Invariant Feature Transform (SIFT), Gabor filters, and the histogram of oriented gradients (HoG), have been explored for RS image classification. Methods relying on these handcrafted low-level features only perform well on images with uniform spatial arrangements or texture but they are limited to distinguishing RS images with more complex scenes. The rise of deep learning and related hardware advancement has revolutionized every industrial sector. Powered by convolutional neural network (CNN) [2–6], traditional computer vision tasks like image classification and object detection have had tremendous and rapid performance gains in the past ten years. CNNs have demonstrated outstanding capability in the discovery of intricate structures and discriminative information hidden in high-dimensional data, making it suitable for image data. Moreover, CNN models pretrained on large and open-domain datasets, such as ImageNet [7], can be transferred to any domain-specific datasets, only costing additional efforts in fine-tuning. However, the traditional CNN architectures, represented by AlexNet [2], VGG [3], Inception [5,6], and ResNet [4], are still limited in RS image scene classification due to the task-specific challenges mentioned above.

To address these challenges, recent CNN-based RS image scene classification methods mainly focus on how to extract discriminative features. For a fine-grained classification dataset such as NWPU-RESISC45, it would be effective to apply fine-grained visual categorization (FGVC) methods that aim to learn discriminative features either through localization of critical regions [8–12] or via end-to-end feature encoding from the whole input image [13–16]. Besides, there are other methods focusing on image data preprocessing [17,18]. Most of the prior efforts have achieved impressive performance on multiple FGVC datasets, such as CUB-Birds [19], Stanford Dogs, and Stanford Cars [20]. However, as shown in our experiments, these FGVC models do not achieve satisfying results in RS image scene classification. A promising direction to further boost the classification accuracy is multigranularity discriminative feature learning, which aims to discover features at multiple scales along the convolutional network backbone. Prior efforts [21–23] have explored ways to mine multigranularity features that are fused and then fed into the detection head. In [22], a multigranularity progressive training framework is proposed to learn complementary features at different granularities. For each input image, a jigsaw puzzle generator [24] is adopted to partition the image into multiple patches, which are then randomly rearranged to form a reconstructed image. Training with multiple granularities could encourage the network to operate on a patch-level, where patch sizes are specific to a particular granularity. However, the random rearrangement on the patches of the input image could introduce noise, which is harmful to training. As shown in Figure 2,

the random rearrangement operation has destroyed the local construction in the image, creating meaningless edges that prevent the model from learning low-level features.

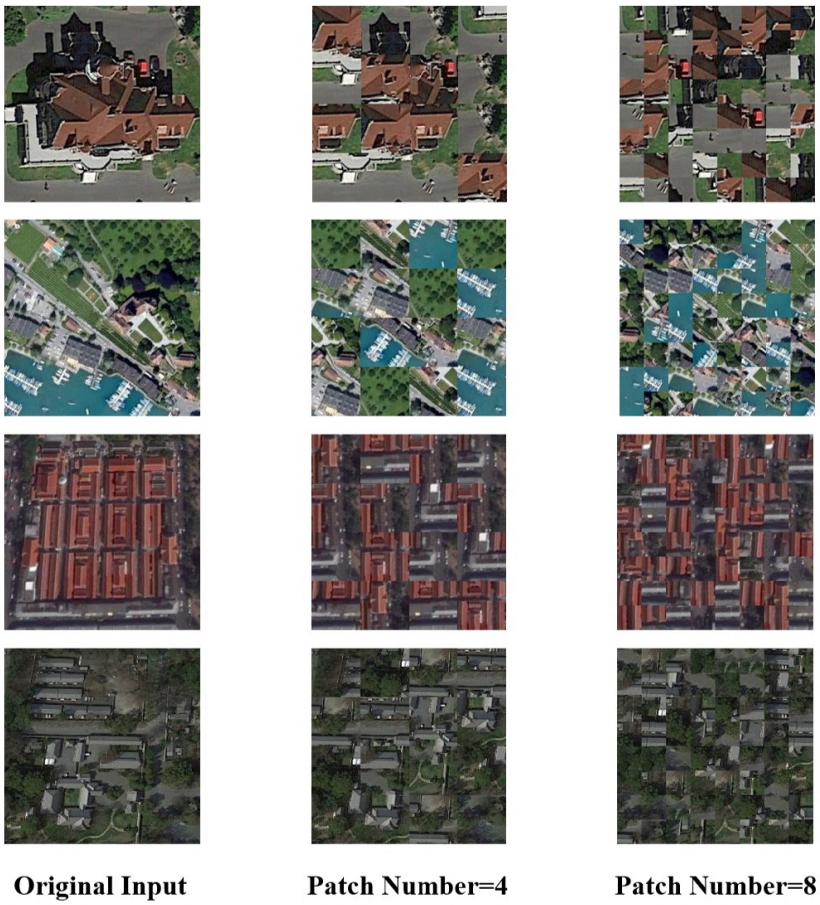

Figure 2. Destruction operation of the spatial layout of input images.

In this paper, we propose a novel model named the Semantic Multigranularity Feature Learning Network (SMGFL-Net) for RS image scene classification. SMGFL-Net is an end-to-end framework that learns multigranularity features and fuses them for final recognition. Instead of rearranging the input image that causes local shape destruction, SMGFL-Net adopts a patch-level rearrangement on the high-level feature maps. We conduct extensive experiments to compare SMGFL-Net with a wide range of peer methods on the NWPU-RESISC45 [25] and on the Aerial Scene Classification (AID) [26] datasets. Results show that SMGFL-Net can effectively learn and leverage global and local features at various granularities, thus, achieving state-of-the-art (SOTA) performance.

2. Related Work

2.1. Fine-Grained Object Recognition

Prior FGVC methods have focused on the learning of discriminative features either from critical regions or the whole image. Yang et al. proposed the Navigator–Teacher–Scrutinizer Network (NTS-Net) [9], a multiagent cooperative learning framework, to identify critical regions. A localization subnetwork is adopted to compute the informativeness of subregions, and activations in the subnetwork correspond to subregions with different sizes and aspect ratios. The informative regions could then be selected through a custom loss function. Sun et al. [27] proposed a one-squeeze multiexcitation module to learn multiple attentive region features, which are then fed into a metric-learning framework with multiattention, multiclass constraints. Discriminative features can also be learned from the whole image via an end-to-end feature encoder. Lin et al. proposed bilinear

feature transformation [28], which allows CNNs to learn fine-grained details over a global image by calculating pairwise interactions between feature channels. Follow-up efforts, including compact bilinear [14] and low-rank bilinear pooling [29], aimed to improve the computational efficiency caused by the exponential growth of feature dimensions appearing in bilinear-CNN. Different from these FGVC methods, the proposed method focuses on learning features at multiple granularities to discover the global and local semantic meaning of the scenes in the RS image.

2.2. Multigranularity Feature Learning

A recently active line of research in FGVC is multigranularity feature learning [21–23]. In [21], a gradually enhanced strategy was introduced to learn multiple granularity-specific experts on limited fine-grained training data. A progressive multigranularity training framework was proposed by [22] to learn complementary features from different granularities, with the central idea of encouraging the network to learn multigranularity features from the whole image as well as the patches obtained and rearranged by a jigsaw puzzle generator. However, the meaningless edges caused by the image patches are harmful to effective feature learning. Thus, this study aims to address this problem by moving the patch generation from the original image to the intermediate feature maps.

2.3. Remote Sensing Image Scene Classification Methods

CNN-based methods have achieved impressive improvement for RS image scene classification. In [30], a feature-selection method was proposed based on the deep belief network (DBN). Due to the effectiveness in feature abstraction, this method works well on a relatively small dataset. [31] proposed an adaptive deep pyramid matching (ADPM) model that took advantage of information from all convolutional layers in the network. ADPM achieved superior performance on the 21-Class Land Use dataset [32] and the 19-Class Satellite Scene dataset [33]. Multiple convolutional layers were also used in [34] to compute a covariance matrix. Each entry of covariance matrix stands for the covariance of two feature maps; this could exploit the complementary information from different convolutional layers. As a complete novel network, the capsule network is applied in RS image scene classification. In [35], the capsule network was added to a CNN without fully connected layer. In [36], a Multigranularity Multilevel Feature Fusion Branch (MGML-FFB) was proposed to extract multigranularity features by a custom module named the channel-separate feature generator (CS-FG). Further, diversified predictions were provided by an ensemble module integrated in the network. Ke et al. conducted a series of studies in RS Image Scene Classification. In [37], a multilayer feature fusion network was developed with a data augmentation approach integrated into training to improve the generalization ability of the model. In [38], the authors proposed a global–local dual-branch structure (GLDBS) that allows a network to explore global and local discriminative features. In addition to CNN, Graph Convolutional Network (GCN)-based methods have also been explored. The paper [39] discussed a deep feature aggregation framework driven by graph convolutional network (DFAGCN) for high-spatial-resolution scene classification. The framework utilizes a pretrained CNN to obtain multilayer features, which are fed into a GCN to obtain patch-to-patch correlations between the feature maps. The output is then passed through a weighted concatenation operation and a linear layer to make a prediction. Our investigation shows that the method involving feature map rearrangement has not been extensively studied.

This method is tested on larger datasets, such as NWPU-RESISC45 [25]. Some of the mentioned efforts are used as baselines of this study.

3. Semantic Multigranularity Feature Learning Network

This section presents the technical details of the proposed SMGFL-Net, an end-to-end deep network for RS image scene classification. Figure 3 shows the neural architecture of SMGFL-Net. At a high level, SMGFL-Net employs a CNN-based backbone that consists

of multiple stages, where feature maps in a stage are of the same size. In the training phase, feature maps at different stages are partitioned into patches and rearranged, and then fed into subsequent convolution and classification blocks. This way, the network is encouraged to learn features at multiple granularities. In the inference phase, however, the intermediate feature maps are not reconstructed since the network has been optimized to learn global and local features during training.

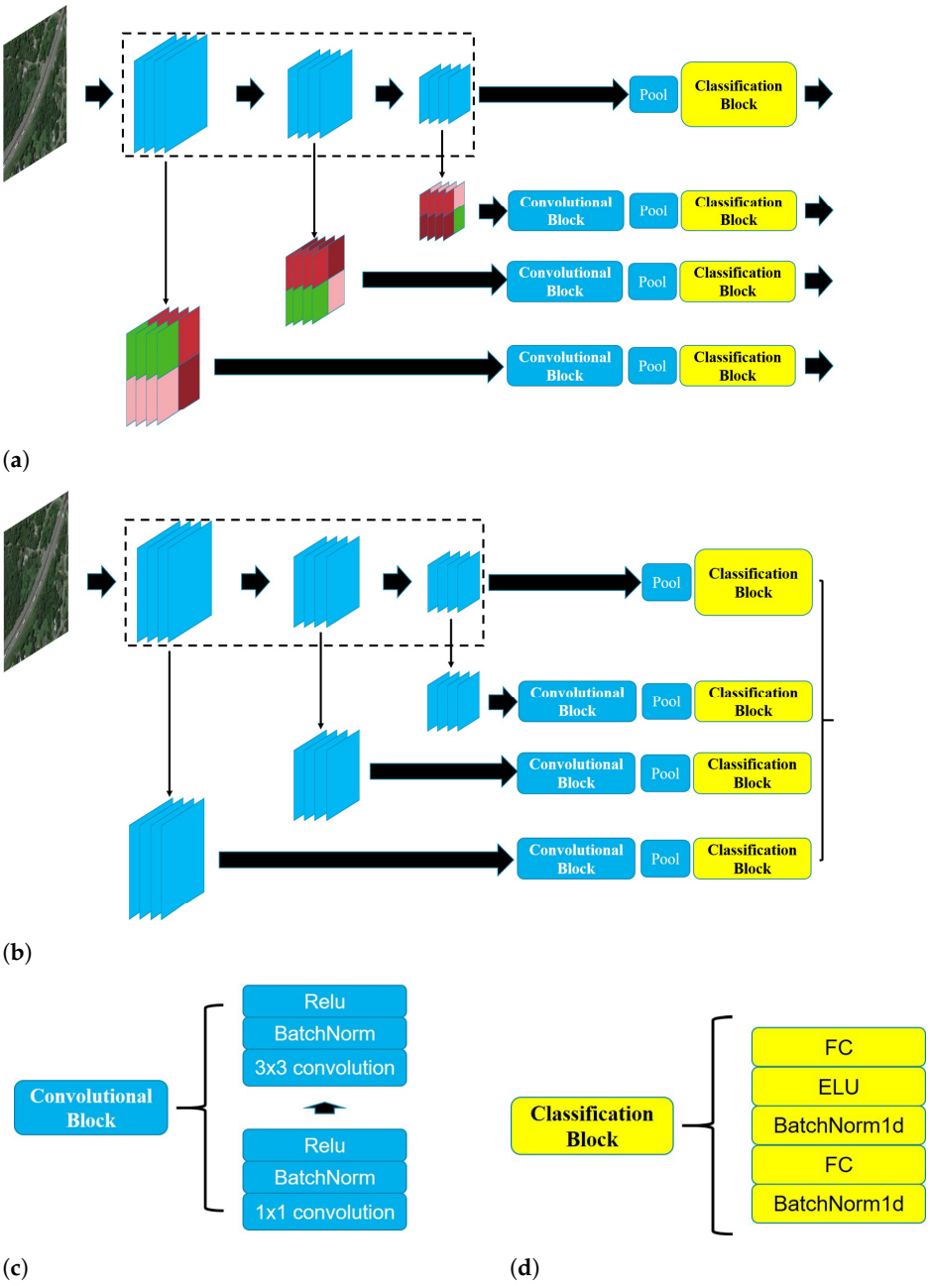

Figure 3. The proposed SMGFL-Net: (**a**) the training phase; (**b**) the inference phase; (**c**) the convolutional block; (**d**) the classification block.

3.1. Network Design

Let F be the backbone feature extractor with L stages, which outputs feature maps of different sizes. Let $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ be the feature map generated at stage S_i, where $i = 1, 2, \ldots, L$, and C_i, H_i, and W_i are the number of feature channels, and height and width of $\mathbf{F}_i$, respectively. In the training phase, an intermediate feature map $\mathbf{F}_i$ is partitioned into multiple patches by a jigsaw puzzle generator and randomly rearranged to

form a reconstructed feature map, denoted by $\mathbf{F}'_i$. In addition to the backbone network, we introduce a convolutional block B_{conv} and a classification block B_{cls} process for each granularity-specific branch. B_{conv} is a two-step operation, as shown in Figure 3c. The first step is a 1×1 convolutional layer followed by a batch normalization operation (BN) and a Rectified Linear Unit (ReLU) activation ($\text{Conv}_{1\times1}$ + BN + ReLU), and the second step is a $\text{Conv}_{3\times3}$ + BN + ReLU layer. The output of a convolutional block is then fed into a pooling block, which could be either max or average pooling. Lastly, the pooling results are sent to a classification block B_{cls}, which consists of two fully connected layers with BN and nonlinear operations, such as Exponential Linear Units (ELU) [40], to predict a probability distribution over the classes. Lastly, to fuse the intermediate feature maps, we define

$$V_{fusion} = \left[B_{conv}(\mathbf{F}'_1); B_{conv}(\mathbf{F}'_2); \ldots; B_{conv}(\mathbf{F}'_L)\right] \tag{1}$$

where $[;]$ refers to the fusion operator. We send V_{fusion} through another classification head to obtain a prediction result $\hat{\mathbf{y}}_{fusion}$. In the experiments, different fusion strategies—including concatenation, summation, and multiplication—are evaluated and compared with each other, and the optimal one is selected.

3.2. Training

In the training phase, a jigsaw puzzle generator is utilized to partition and rearrange the intermediate feature maps. This way, the meaningless edges can be eliminated. When the activation value in the feature map changes, the localization of its receptive field also changes. Further, as the network deepens, the feature map in later layers contains more semantic information and less spatial information.

Given a feature map $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, we equally partition it into $n_i \times n_i$ patches, which are $C_i \times \frac{H_i}{n_i} \times \frac{W_i}{n_i}$. The patches are then randomly shuffled and merged together into a new feature map $\mathbf{F}'_i$. Further, H_i and W_i should be integral multiples of n_i, and the granularity of the convolutional layer at stage S_i is given by $\frac{H_i}{n_i}$. Each $\mathbf{F}'_i$ is subsequently passed through a convolutional block B_{conv}, a pooling block, and a classification block B_{cls}, which outputs a prediction result $\hat{\mathbf{y}}_i$ for the branch belonging to the ith stage. Moreover, we let $\hat{\mathbf{y}}_0$ be the prediction result of the trunk network (i.e., the backbone) and let $\mathbf{y}$ be the ground truth label of an input image; the cost C per image can be defined as the sum of the pairwise cross entropy loss between prediction branch and the ground truth, given in Equation (2).

$$C(\hat{\mathbf{y}}_0, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_L, \hat{\mathbf{y}}_{fusion}, \mathbf{y}) = -\sum_{i=0}^{L}(\mathbf{y} \log \hat{\mathbf{y}}_i) - \mathbf{y} \log \hat{\mathbf{y}}_{fusion} \tag{2}$$

The overall loss function is a summation of the costs across all images in the training set.

3.3. Inference

In the inference phase, an input image is passed through the network in a feed-forward way without the step of feature map rearrangement, because the network has been trained and the parameters have been optimized to discover features at multiple granularities. We consider two types of predictions in this paper. In the first case, only the fused result, defined in Equation (3), is used for the final prediction. Since $\hat{\mathbf{y}}_{fusion}$ is a normalized vector, the element with the highest confidence is selected and returned as the predicted category, which is performed by the arg max operator. This can lead to less computational budget. In the second case, the fused result and results given by each granularity branch are combined together for the final prediction, defined in Equation (4). Due to the complementary nature of different granularity branches, the combined prediction can achieve better performance than the first case. This hypothesis has been proven in the experiments.

$$\hat{y} = \arg\max(\hat{\mathbf{y}}_{fusion}) \tag{3}$$

$$\hat{y} = \arg\max(\sum_{i=0}^{L} \hat{\mathbf{y}}_i + \hat{\mathbf{y}}_{fusion}) \quad (4)$$

4. Experiments

4.1. Dataset

In the experiments, the proposed SMGFL-Net and baselines are evaluated on the NWPU-RESISC45 [25] and AID datasets.

- NWPU-RESISC45 is an open-source RS image scene dataset created by Northwestern Polytechnical University. This dataset contains 31,500 images, covering 45 scene classes. Each category consists of 700 images measuring 256 × 256 pixels. For most scene categories, the spatial resolution varies from 0.2 m to 30 m per pixel. Exceptions are the categories of island, lake, mountain, and snowbank, which are with lower spatial resolutions. NWPU-RESISC45 is divided into 45 scene categories as follows: Airplane, Airport, Baseball diamond, Basketball court, Beach, Bridge, Chaparral, Church, Circular farmland, Cloud, Commercial area, Dense residential, Desert, Forest, Freeway, Golf course, Ground track field, Harbor, Industrial area, Intersection, Island, Lake, Meadow, Medium residential, Mobile home park Mountain, Overpass, Palace, Parking lot, Railway, Railway Station, Rectangular farmland, River, Roundabout, Runway, Sea ice, Ship, Snowbank, Sparse residential, Stadium, Storage tank, Tennis court, Terrace, Thermal power station, and Wetland. In Figure 4, we display two scene samples per category for all categories. In these categories, there are a number of fine-grained scenes with small intervariance and large intravariance. Additionally, both global and local features are necessary for the recognition of certain categories. For example, as shown in Figure 1, images from Palace and Medium residential have similar global shapes, and there is a big difference between images in Palace category. The Medium residential and Density residential samples have local similarities, but differ in global characteristics such as spatial layout.
- As a dataset for aerial scene classification, AID is made up of the following 30 classes: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. The number of sample images varies with different classes—from 220 up to 420. In all, there are 10,000 images within 30 classes in AID. Images in AID are from different remote imaging sensors. Moreover, all the sample images per class are carefully chosen from different regions around the world. AID is a challenging dataset due to the higher intraclass variations and smaller interclass dissimilarity. The pixel resolution in AID changes from about 8 m to about half a meter, and the size of each image in AID is fixed to be 600 × 600.

4.2. Baselines

Baselines in our experiments could be divided into three categories: the methods that extract global features, the methods that extract local features, and the FGVC methods that localize critical regions and fuse global and local features. Some dedicated efforts on RS image scene classification are also taken into account for comparison. These methods are briefly introduced in this subsection.

- **Global methods**. The deep CNNs designed for generic image classification such as VGG [3] and ResNet [4] are used as global methods in our experiments. These networks have achieved SOTA performance on large-scale datasets, such as ImageNet [7].
- **Local methods**. Different from the designed networks for generic object recognition, local CNN-based methods such as bilinear-CNN [16] capture orderless statistical information without spatial information.
- **FGVC methods**. In the experiments, we select several FGVC methods related to the design idea of the proposed SMGFL-Net. In NTS-Net [9], critical regions with different sizes and aspect ratio are automatically selected through the region proposal network.

It could fuse local and global features for recognition. ResNet50 is the backbone network of NTS-Net. DCL [17] encourages the backbone network to extract local features by destroying the spatial distribution of the training images. Another highly relevant study, progressive-MG [22], can learn multigranularity features through jigsaw puzzle generator with different patch size in a progressive way of training.

- **RS image scene classification methods**. The related methods in this category include the following: CNN-CapsNet [35] can achieve better affine transformation invariance; InceptionV3-CapsNet [35] achieves the best performance on NWPU-RESISC45 when the training ratios are 0.1 and 0.2; further, multilayer stacked covariance pooling (MSCP) [34] is applied to the VGG16 backbone and consistently outperforms single-layer models on three challenging datasets. MGML-FENet [36] has achieved SOTA performance on both NWPU-RESISC45 and AID.

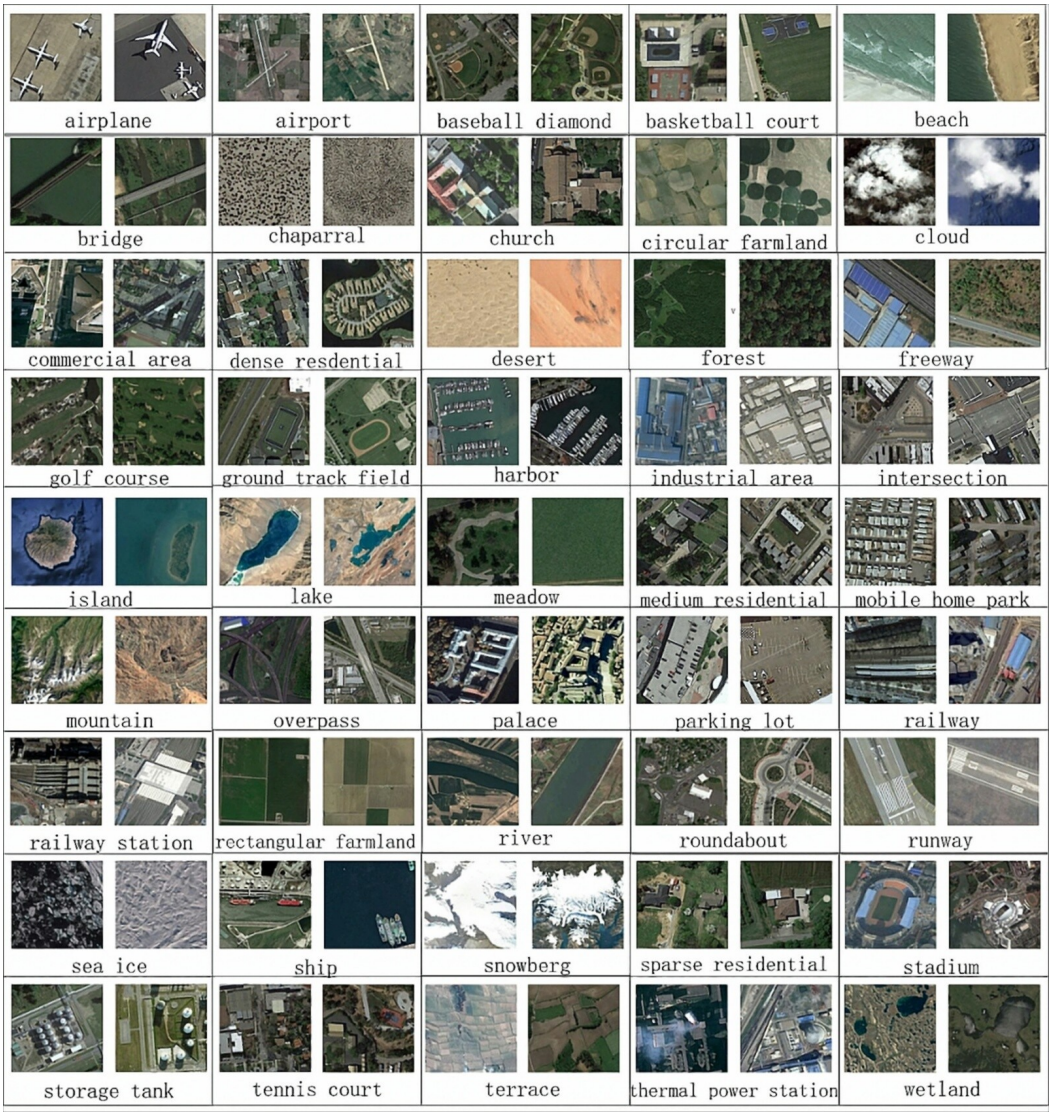

Figure 4. Images samples from NWPU-RESISC45.

4.3. Implementation Details

All the experiments are performed using PyTorch over a cluster of GTX TITAN X GPUs. ResNet50 [4] is used as the backbone network of SMGFL-Net. The input images are resized to 512 × 512, then processed through a series of operations such as flip, random cropping to 448 × 448, and normalization at the training step. In the SMGFL-Net backbone, there are five stages of feature maps with different sizes, and we select stages three to

five, which is a decision based on empirical results. The sizes of feature maps at stages three, four, and five are $512 \times 56 \times 56$, $1024 \times 28 \times 28$, and $2048 \times 14 \times 14$, respectively. Granularity specific to each stage should be integral multiples of height and width of the feature maps. We use stochastic gradient descent (SGD) with momentum 0.9, weight decay 0.0005, and base learning rate 0.0002 for the pretrained parameters and 0.002 for the new parameters of convolutional and classification block. The learning rates of different parameters are reduced by following the cosine annealing schedule during training. The batch size is set to be 32.

4.4. Performance Metric

Since the dataset is well-balanced, we use accuracy (ACC) as the sole performance metric, which is also widely adopted in the literature [17,22,35]. Briefly, ACC measures the percentage of correct predictions on the test set. Formally, $ACC = \frac{TP+FN}{\text{test set size}} \times 100\%$, in which TP and FN stand for the number of true positives and false negatives, which are the two cases of correct predictions.

4.5. Key Design Choices

We evaluate four design choices in the proposed SMGFL-Net, including (1) the pooling operation that connects the convolutional and classification block, (2) the different granularity options, (3) different fusion strategies of granularity-specific branches, (4) with or without the rearrangement of the feature maps. When evaluating each design choice, the accuracies of the two predictions discussed in Section 3.3 are observed on both NEWPU-RESISC45 and AID. The training ratio is set to be 0.5 in NEWPU-RESISC45 and 0.3 in AID. For the pooling operation, we consider average vs. max pooling. For the granularity options, we consider four granularity combinations applied on stages three, four, and five. Let G_{S_i} denote the granularity at stage i in the network. For instance, $G_{S_i} = 8$ means that the patch size in the rearranged feature map at stage i is 8×8 during training. The four granularity options we chose are $\{G_{S_3} = 28, G_{S_4} = 14, G_{S_5} = 7\}$, $\{G_{S_3} = 8, G_{S_4} = 4, G_{S_5} = 4\}$, $\{G_{S_3} = 7, G_{S_4} = 7, G_{S_5} = 7\}$, and $\{G_{S_3} = 2, G_{S_4} = 2, G_{S_5} = 2\}$. For the fusion strategies of granularity-specific branches, three fusion strategies—concatenation, summation, and multiplication—are evaluated. For evaluating the effectiveness of rearranging the feature maps, we keep the feature maps of each granularity-specific branch unchanged to compare with the situation of rearranging feature maps in the training process. We conducted a randomized search on the combinations of the design options and obtained the best combination based on the search result: { max pooling, $\{G_{S_3} = 8, G_{S_4} = 4, G_{S_5} = 2\}$, concatenation, with rearrangement }. To present the effect of a specific design choice, we chose the optimal value for other choices. Results are reported in Tables 1–4. We provide the observations as follows.

- **Pooling strategy**. As shown in Figure 3, all of the global and granularity-specific branches contain pooling operations. The average pooling can process all information in a region and the maximum pooling can obtain the maximum value in a region. Table 1 shows an ACC comparison between average and max pooling on both the NWPU-RESISC45 and AID datasets. It is observed that max pooling consistently outperforms average pooling by 0.2–0.9%. Since average pooling considers all pixels in a region, it tends to smooth out an image and sharp features may not be identified. On the other hand, max pooling works by selecting the brightest pixel in a region, which makes it useful when the image background is dark and the semantic features are in lighter pixels. On both datasets, the majority of scene images are in dark background with light semantic objects that distinguish the categories. This characteristic of the dataset justifies the better effect of max pooling.
- **Branch granularities**. Table 2 reflects the effect of the four granularity options, in which $\{G_{S_3} = 8, G_{S_4} = 4, G_{S_5} = 2\}$ consistently outperform other options on both datasets. From the limited options for the branch granularities that have been tried, we have the impression that a progressively decreasing patch size is favored to accommodate

the feature maps that are also decreasing stage-by-stage. However, to validate this point, more granularity options can be tried, or even a heuristic can be developed to automate the search for an optimal or suboptimal set of granularities. We leave this as a meaningful future study, as described in Section 5.

- **Different fusion strategies**. Table 3 shows the effect of three different fusion strategies—concatenation, summation, and multiplication—in which concatenation consistently outperforms other options in all scenarios. The reason for this result is that the concatenation operation cannot cover the information of the original features and obtain the feature vector with higher dimension. Feature vectors with higher dimension increase the capacity of subsequent classification modules and have stronger feature representation ability.
- **Feature maps rearrangement.** Table 4 reports the effect of feature maps rearrangement. We compare the models with vs. without feature maps rearrangement. A consistent performance gain is observed on both datasets. This is because feature maps rearrangement of different branches can learn subtle fine-grained features of different semantic levels. These features are likely to be complementary to each other to benefit the final scene classification.

Table 1. Effect of average and max pooling. The highest score on each dataset is marked in bold.

	Avg. Pooling	Max Pooling
NEWPU-RESISC45	95.79%	**96.57**%
AID	97.01%	**97.21**%

Table 2. Effect of branch granularities.

	NEWPU-RESISC45	AID
$G_{S_3} = 28, G_{S_4} = 14, G_{S_5} = 7$	96.07%	96.86%
$G_{S_3} = 8, G_{S_4} = 4, G_{S_5} = 2$	**96.57**%	**97.21**%
$G_{S_3} = 7, G_{S_4} = 7, G_{S_5} = 7$	96.34%	96.99%
$G_{S_3} = 2, G_{S_4} = 2, G_{S_5} = 2$	95.97%	96.64%

Table 3. Effect of different fusion strategies.

	Concatenation	Summation	Multiplication
NEWPU-RESISC45	**96.57**%	96.11%	96.03%
AID	**97.21**%	96.99%	96.88%

Table 4. Effect of feature maps rearrangement.

	w/ Rearrangement	w/o Rearrangement
NEWPU-RESISC45	**96.57**%	96.43%
AID	**97.21**%	96.99%

4.6. Results on NWPU-RESISC45

In the experiments, we compare SMGFL-Net with global, local, FGVC, and RS scene classification methods, which are introduced in Section 4.2. The results are shown in Table 5. The first two columns of the table display the evaluated method and its category, and the last three columns show the accuracy of the model that is trained under three training ratio settings—namely, 0.1, 0.2, and 0.5, which specify the training set ratio over the entire dataset. The observations are summarized as follows:

- The proposed SMGFL-Net presents the best overall accuracy compared with all of the benchmarks. Specifically, SMGFL-Net demonstrates the best accuracy of 91.9%

and 96.5%, with training ratios of 0.1 and 0.5, respectively. When the training ratio is 0.2, SMGFL-Net has the second-best accuracy of 93.7%, the same as InceptionV3-CapsNet, and is only 0.4% worse than Progressive-MG. The results demonstrate that SMGFL-Net can achieve superior performance with only 10% of data used for training, outperforming the second best model, ResNet50, by 1%.

- For the global methods, ResNet50 is better than VGG16 under all three training ratio settings, which is reasonable since ResNet50 is featured by shortcut connections, allowing an algorithm to train deep networks; thus, it can be justified that the 50-layer ResNet50 outperforms the 16-layer VGG16. Moreover, our model adopts ResNet50 as a backbone and obtains a gain of 1%, 0.1%, and 0.7% for the three training settings, respectively.
- The only evaluated local method bilinear-CNN performs the worst (91.2%) with a training ratio of 0.5, indicating that the localized features captured by the bilinear-CNN are less effective for this task.
- For the FGVC methods, Progressive-MG stands out, with the best accuracy when the training ratio is 0.2. From a design point of view, Progressive-MG also employs multigranularity feature extraction but with a key design difference from ours, i.e., their jigsaw puzzle generator is applied on the input image, leading to meaningless edges that may destroy certain features. Results show that our method outperforms Progressive-MG by 0.6%, meaning that the patch rearrangement at the feature map level is more effective for the RS scene classification task. On the other hand, with subnetwork specifically designed for localizing critical regions, NTS-Net still fails to achieve satisfying accuracy. When the training ratio reduces to 0.1, NTS-Net shows the worst accuracy of 84.5%; as the training ratio goes up to 0.5, its accuracy is raised to 93.2%, meaning that NTS-Net needs more training data to close the prediction bias.
- Lastly, for the RS methods, VGG16-MSCP does not perform well, with an accuracy less than 90% for training ratios 0.1 and 0.2. InceptionV3-CapsNet, on the other hand, presents the same accuracy as our method with a training ratio of 0.2 but is worse than ours by 2.6% with a training ratio of 0.1. MGML-FENet, as an effective remote sensing image classification method, achieve the best performance when the training ratio is 0.2. However, SMGFL-Net can achieve better performance when the training ratio is 0.1.

Table 5. Comparative results on NWPU-RESISC45.

Method		Ratio = 0.1	Ratio = 0.2	Ratio = 0.5
Global Methods	VGG16 [3]	86.1%	90.0%	93.4%
	Resnet50 [4]	90.9%	93.6%	95.8%
	ResNeXt50 [41]	91.8%	94.0%	96.1%
Local Methods	Bilinear-CNN [16]	-	-	91.2%
FGVC Methods	NTS-Net [9]	84.5%	89.6%	93.2%
	DCL [17]	-	-	94.3%
	Progressive-MG [22]	90.0%	94.1%	95.9%
RS Methods	InceptionV3-CapsNet [34]	89.3%	93.7%	-
	VGG16-MSCP [35]	85.5%	89.1%	-
	MGML-FENet [36]	91.3%	**94.5**%	-
Ours	SMGFL-Net	**91.9**%	93.7%	96.5%

4.7. Results on AID

In the experiments, we compared SMGFL-Net with the same methods in Section 4.6. The results are shown in Table 6. The first two columns of the table display the evaluated method and its category, and the last three columns show the accuracy of the model that is trained under three training ratio settings—namely, 0.1, 0.2 and 0.3, which specify the

training set ratio over the entire dataset. The observations are summarized as follows: The proposed SMGFL-Net shows the best performance compared to all of the benchmarks. Specially, SMGFL-Net achieves the best accuracies of 93.3%, 96.2%, and 97.2% with the training ratios of 0.1, 0.2, and 0.3, respectively. Although MGML-FENet achieves the best accuracy with a training ratio of 0.2 on NWPU-RESISC45, SMGFL-Net outperforms it on AID. For the global methods, ResNeXt50 achieves the best accuracy compared with all of the other global methods in this experiment. It is worth noting that, as the backbone network of SMGFL-Net, ResNet50 achieves 89.8%, 91.3%, and 93.7% with training ratios of 0.1, 0.2, and 0.3. Compared with ResNet50, SMGFL-Net achieves 93.3%, 96.2%, and 97.2% with training ratios of 0.1, 0.2, and 0.3. This result demonstrates the effectiveness of the proposed framework. For the local methods, Bilinear-CNN achieves relatively better accuracy compared with the performance on NWPU-RESISC45. For the FGVC methods, Progressive-MG and MGML-FENet also achieve good performance. This shows that multigranularity information extraction is very important in both remote sensing datasets. On the other hand, NTS-Net achieves better accuracy on AID than on NWPU-RESISC45.

Table 6. Comparative results on AID.

Method		Ratio = 0.1	Ratio = 0.2	Ratio = 0.5
Global Methods	VGG16 [3]	-	86.6%	-
	Resnet50 [4]	89.8%	91.3%	93.7%
	ResNeXt50 [41]	90.1%	92.1%	94.2%
Local Methods	Bilinear-CNN [16]	90.0%	92.6%	93.0%
FGVC Methods	NTS-Net [9]	90.5%	94.0%	94.7%
	DCL [17]	86.3%	91.2%	-
	Progressive-MG [22]	90.0%	94.1%	95.9%
RS Methods	InceptionV3-CapsNet [34]	-	89.0%	-
	VGG16-MSCP [35]	-	91.6%	-
	MGML-FENet [36]	-	95.8%	-
Ours	SMGFL-Net	**93.3**%	**96.2**%	**97.2**%

5. Conclusions

In this paper, we propose a novel deep neural architecture named SMGFL-Net for RS image scene classification. At the training step, the proposed SMGFL-Net learns multigranularity features through a destruction operation on intermediate feature maps by a jigsaw puzzle generator with different sizes, which avoids the meaningless edges appearing in prior studies. The proposed method is validated on the NWPU-RESISC45 and AID datasets. Results show that SMGFL-Net can effectively learn and leverage both global and local features at different granularities, demonstrating the SOTA performance compared with several peer methods from recent literature.

This study has the following limitations, which will be addressed in future work. First, the efficacy of the proposed SMGFL-Net is only validated on two datasets, namely, NWPU-RESISC45 and AID, and its capability on further RS image datasets has not been evaluated. Second, in the experiment of this study, the number of stages with multigranularity branches and the patch sizes are manually determined, which is not cost-effective and provides limited guidance on the choice of these parameters. It is thus essential to develop a heuristic to search for an optimal or suboptimal set of parameters suitable for a problem instance in a systematic way. Third, the challenge of intraclass variance and interclass similarity could be addressed by other CNN design strategies, such as pairwise feature learning and biattention, which can be selectively integrated into SMGFL-Net.

Author Contributions: Conceptualization and methodology, X.M., Z.X., H.-s.Y. and S.-J.L.; software, validation, and original draft preparation, X.M.; review and editing, supervision, funding acquisition, Z.X., H.-s.Y. and S.-J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant (2019-MOIS33-005) of Lower-level and Core Disaster-Safety Technology Development Program funded by the Ministry of Interior and Safety (MOIS, Korea). This research was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C201231911).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets supporting the conclusions of this article are available at https://www.tensorflow.org/datasets/catalog/resisc45 and https://captain-whu.github.io/AID/ (accessed on 2 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Campbell, J.B.; Wynne, R.H. *Introduction to Remote Sensing*; Guilford Press: New York, NY, USA, 2011.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
5. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
6. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
7. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
8. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *arXiv* **2015**, arXiv:1506.02025.
9. Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 420–435.
10. Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; Tian, Q. Picking deep filter responses for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1134–1142.
11. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4438–4446.
12. Zhang, F.; Li, M.; Zhai, G.; Liu, Y. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In Proceedings of the International Conference on Multimedia Modeling, Online, 22–24 June 2021; pp. 136–147.
13. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1449–1457.
14. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact bilinear pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 317–326.
15. Kim, J.H.; On, K.W.; Lim, W.; Kim, J.; Ha, J.W.; Zhang, B.T. Hadamard product for low-rank bilinear pooling. *arXiv* **2016**, arXiv:1610.04325.
16. Liao, Q.; Wang, D.; Holewa, H.; Xu, M. Squeezed bilinear pooling for fine-grained visual categorization. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019.
17. Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and construction learning for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5157–5166.
18. Zhuang, P.; Wang, Y.; Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13130–13137.
19. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
20. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3d object representations for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013; pp. 554–561.
21. Zhang, L.; Huang, S.; Liu, W.; Tao, D. Learning a mixture of granularity-specific experts for fine-grained categorization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8331–8340.

22. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; pp. 153–168.
23. Ni, B.; Paramathayalan, V.R.; Li, T.; Moulin, P. Multiple granularity modeling: A coarse-to-fine framework for fine-grained action analysis. *Int. J. Comput.Vis.* **2016**, *120*, 28–43. [CrossRef]
24. Wei, C.; Xie, L.; Ren, X.; Xia, Y.; Su, C.; Liu, J.; Tian, Q.; Yuille, A.L. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1910–1919.
25. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
26. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
27. Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-attention multi-class constraint for fine-grained image recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 805–821.
28. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1309–1322. [CrossRef] [PubMed]
29. Kong, S.; Fowlkes, C. Low-rank bilinear pooling for fine-grained classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 365–374.
30. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
31. Liu, Q.; Hang, R.; Song, H.; Zhu, F.; Plaza, J.; Plaza, A. Adaptive deep pyramid matching for remote sensing scene classification. *arXiv* **2016**, arXiv:1611.03589.
32. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.
33. Dai, D.; Yang, W. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 173–176. [CrossRef]
34. He, N.; Fang, L.; Li, S.; Plaza, A.; Plaza, J. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6899–6910. [CrossRef]
35. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
36. Zhao, Q.; Lyu, S.; Li, Y.; Ma, Y.; Chen, L. MGML: Multigranularity Multilevel Feature Ensemble Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
37. Xu, K.; Huang, H.; Li, Y.; Shi, G. Multilayer feature fusion network for scene classification in remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1894–1898. [CrossRef]
38. Xu, K.; Huang, H.; Deng, P. Remote Sensing Image Scene Classification Based on Global-Local Dual-Branch Structure Model. *IEEE Geosci. Remote Sens. Lett.* **2021**. [CrossRef]
39. Xu, K.; Huang, H.; Deng, P.; Li, Y. Deep Feature Aggregation Framework Driven by Graph Convolutional Network for Scene Classification in Remote Sensing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
40. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
41. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.