remote sensing

MDPI

Article

SAFFNet: Self-Attention-Based Feature Fusion Network for Remote Sensing Few-Shot Scene Classification

Joseph Kim [1] and Mingmin Chi [1,2,*]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, 2005 SongHu Road, Shanghai 200438, China; Kjoseph18@fudan.edu.cn
[2] Zhongshan Fudan Joint Innovation Center, Zhongshan PoolNet Technology Co. Ltd., 6, Xiangxing Road, Zhongshan 528400, China
* Correspondence: mmchi@fudan.edu.cn

Abstract: In real applications, it is necessary to classify new unseen classes that cannot be acquired in training datasets. To solve this problem, few-shot learning methods are usually adopted to recognize new categories with only a few (out-of-bag) labeled samples together with the known classes available in the (large-scale) training dataset. Unlike common scene classification images obtained by CCD (Charge-Coupled Device) cameras, remote sensing scene classification datasets tend to have plentiful texture features rather than shape features. Therefore, it is important to extract more valuable texture semantic features from a limited number of labeled input images. In this paper, a multi-scale feature fusion network for few-shot remote sensing scene classification is proposed by integrating a novel self-attention feature selection module, denoted as SAFFNet. Unlike a pyramidal feature hierarchy for object detection, the informative representations of the images with different receptive fields are automatically selected and re-weighted for feature fusion after refining network and global pooling operation for a few-shot remote sensing classification task. Here, the feature weighting value can be fine-tuned by the support set in the few-shot learning task. The proposed model is evaluated on three publicly available datasets for few shot remote sensing scene classification. Experimental results demonstrate the effectiveness of the proposed SAFFNet to improve the few-shot classification accuracy significantly compared to other few-shot methods and the typical multi-scale feature fusion network.

Keywords: few-shot learning; scene classification; self-attention; multi-scale features fusion

Citation: Kim, J.; Chi, M. SAFFNet: Self-Attention-Based Feature Fusion Network for Remote Sensing Few-Shot Scene Classification. *Remote Sens.* **2021**, *13*, 2532. https://doi.org/10.3390/rs13132532

Academic Editors: Mohammad Rostami, Senthilnath Jayavelu and Yongshuo Fu

Received: 2 May 2021
Accepted: 23 June 2021
Published: 28 June 2021

1. Introduction

In the remote sensing field, the classification of scene images is a challenging yet important task in real applications, such as natural hazard detection [1–5], geospatial object detection [6–11], geographic image retrieval [12–15], environmental monitoring [16–19] and urban planning [20] from high-spatial resolution (HSR) remote sensing images. Recently [21–23], CNN (Convolutional Neural Network)-based methods have come into the spotlight due to their high-quality feature extraction ability. For CNN-based methods, there is research going on to re-design the whole network [24] to maximize effectiveness rather than apply pre-trained knowledge by various fine-tuning methods [19] for remote sensing scene classification. Regardless of these approaches, CNN-based methods were successfully applied to remote scene classification tasks and achieved outstanding performance [25,26] compared to traditional methods [27,28].

However, in real applications, it is hard yet necessary to classify new scene categories which are not in the class list of a training dataset. To classify unseen classes, traditional methods are to retrain classification models based on newly collected unseen training datasets that contain the existing data and the additional labeled samples of new classes. However, these methods require expensive computing resources and time to retrain the whole model. Moreover, it is time-consuming and expensive to label out-of-bag images for

model retraining. Therefore, this kind of methods is hard to implement in real applications. To solve this problem, few-shot learning is usually utilized to classify unseen classes without retraining whole models [29–31].

There have been two approaches to solve this problem in the recent decade, i.e., zero-shot learning and few-shot learning. For zero-shot learning, usually transfer learning or model-driven methods have been applied. For instance, the zero-shot scene classification method in the remote sensing application attempts to train the model with data from other domains and then semantic information is used to classify new unseen classes without labeled samples [32]. In natural language processing, the Skip-gram model is used to embed each class into semantic vectors [33], where semantic relationships are built between seen and unseen classes by constructing a directed graph over the classes. On the other hand, extracted features by zero-shot learning model are based on a user-designed textual description of the new classes and it could generate biased knowledge to new unseen classes.

in contrast to the zero-shot learning, data-driven methods can be applied in few-shot learning by leveraging very few labeled data instead of the other model-driven methods to recognize unseen classes. Thus, few-shot learning (including one-shot learning) aims to recognize new unseen classes with very few labeled samples together with the known classes available in the large-scale training dataset. A simplified comparison between zero-shot and few-shot learning is demonstrated in Table 1.

Table 1. Comparison of traditional supervised, zero-shot and few-shot learning for classification tasks.

Classification Task	Unseen Classes	Auxiliary Materials
Traditional Supervised Learning	×	None
Zero-Shot learning	✓	Textual description of unseen classes
Few-Shot learning	✓	Few labeled samples of unseen classes

Few-shot learning has been successfully applied to computer vision [29–31,34], natural language processing [35,36], speech recognition [37], gesture recognition [38], medical image classification [39], image translation [40] and drug discovery [41] tasks. Approaches for few-shot learning can be summarized in three ways. The first approach is based on meta learning (i.e., learning to learn) [42–44], where a Model-Agnostic Meta-Learning (MAML) model is proposed to learn good initial parameters of a deep network over multiple tasks [42], and Meta-SGD [44] (Stochastic Gradient Descent) and Meta-LSTM (Long Short Term Memory) [43] try to solve the problem by learning good learning rates and gradient computing functions and it succeeds in fine-tuning the network with only a few unseen samples. The second approach for few-shot learning is based on memory modules [45–48]. This approach uses deep neural networks to extract features for unseen samples and then external memories are updated afterwards for few-shot classification. The last approach is distance metric learning [29–31,49]. CNN networks for the backbone are firstly used as feature extractors, and then the classifiers based on a distance metric to compare between unseen classes are applied, where feature extractors are often pre-trained by a large scale base training dataset.

However, previously researched methods focus more on training strategies rather than an informative feature extraction. Since the major advantage of deep learning lies in the end-to-end feature extraction, it is necessary to develop an optimized feature extractor for remote sensing few-shot scene classification. In particular, remote sensing images are prone to having plentiful texture features rather than shape ones as shown in Figure 1. Accordingly, it is important to extract valuable texture semantic features from a limited number of input images.

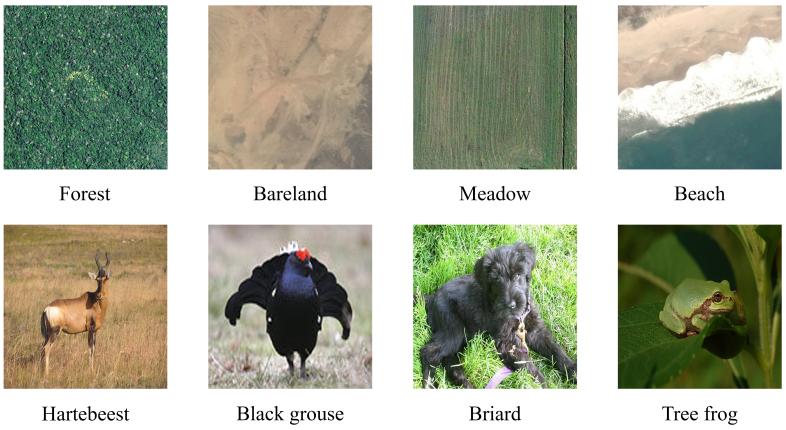
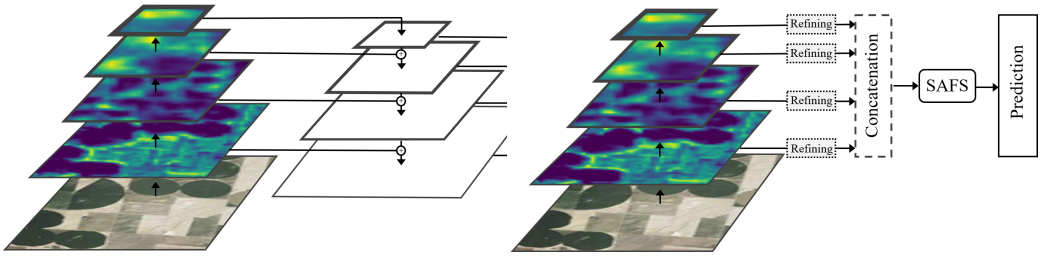


Figure 1. Examples of images in a remote sensing scene classification dataset that has rich texture features. Images on the first row are from a remote sensing scene classification dataset [50] and images at second row are from the Imagenet dataset [51].

Due to a hierarchical feature representation, deep convolutional neural networks have been successfully applied in different applications for the tasks, such as image classification [52], object detection [53] and semantic segmentation [54]. Accordingly, it is better to exploit multi-scale features from a limited number of support sets to obtain a better out-of-bag classification performance. In pyramidal feature models, such as FPN [55] and its extensions [56,57], the combination of multi-scale bottom-up and top-down features is exploited to generate different scales of objects. Here, low-level features contain more accurate locations of small objects and top-down features provide richer semantic information. Top-down features in lower levels are the interpolation of highest-level features through upsampling. If using FPN for classification, features from the highest level are re-emphasized by setting a bigger weighting value. To do so, features for fine-grained scenes are prone to be weakened to lead to poor performance.

To solve the aforementioned problems, a deep feature fusion network for few-shot scene classification is proposed in terms of a multi-scale feature aggregation architecture. Here, multi-scale features with different receptive fields are extracted and kept as stacked feature maps for further feature fusion, rather than FPN in which features are back-propagated from the top layers to the lower layers with lateral connections. Here, multi-scale features with different resolutions in the proposed network are automatically selected for the final decision by a self-attention scheme according to the importance of the features derived from a pyramidal feature hierarchy. Accordingly, the proposed model is called a Self-Attention-based Feature Fusion Network, denoted as SAFFNet. To do so, the support set images can be exploited to "fine-tune" the importance of different-scaled features automatically. The different modules to generating multi-scale feature hierarchy by FPN and the proposed method are shown in Figure 2. Experiments conducted on three benchmark datasets confirm the effectiveness of the proposed deep feature fusion network.



(**a**) FPN (**b**) SAFFNet

Figure 2. The modules used to generate multi-scale feature hierarchies by the feature pyramid network (FPN) shown in (**a**) and by the proposed method SAFFNet shown in (**b**).

2. Related Work

2.1. Few-Shot Scene Classification

Approaches for few-shot scene classification can be summarized into three ways. The first approach is based on a meta learning algorithm (i.e., learning to learn) [42,43]. Since the goal of the meta learning algorithms is to develop a model that can be applied for multiple similar tasks with limited data, there is a huge similarity in training strategy between meta learning and few-shot learning. For detail, a Model-Agnostic Meta-Learning (MAML) model is proposed to learn good initial parameters of a deep network over multiple tasks [42] and Meta-LSTM [43] tried to solve the problem by learning good learning rates and gradient computing functions and it succeeds in fine-tuning the network with only a few unseen samples. The main characteristic of this approach is that proposed methods are focused on the training strategy to solve the problem.

The second approach for few-shot learning is based on memory modules [45,48]. This approach uses deep neural networks to extract features for unseen samples and then external memories are updated afterwards for few-shot classification. Another approach is to apply a graph-based approach. [58] used a Graph Neural Network (GNN) to extract features and added a fully connected layer as the classification head to solve the few-shot problem.

The last approach is distance metric learning [29,30,59]. This approach uses CNN networks for a backbone feature extractor and using a classifier based on a distance metric to compare unseen classes. The feature extractor is pre-trained in the large-scale base training dataset. Since this approach is only revising. the classification head training strategy from meta learning methods also can be utilized. Recently, the distance metric for few-shot classification head has also been substituted for a neural network. RelationNet [31] successfully substituted traditional distance metrics for a neural network and achieved outstanding performance. Other than classification tasks, few-shot learning has been applied to other computer vision tasks successfully [60–62].

2.2. Multi-Scale Feature Fusion

Multi-scale feature fusion is an essential method to improve performance in object detection, instance segmentation and semantic segmentation tasks nowadays. All of the recent models that showed the fine performance are based on this method [53,63]. The most well-known approach is Feature Pyramid Network (FPN)-based methods [55]. FPN [55] tried to fuse the features from various scales and directly predict bounding boxes simultaneously. However, features from the initial stage of the backbone network are not deep enough to extract features from input data. Therefore, top-to-down layers are applied to FPN to solve this problem with upsampling layers. Features from bottom-top stages are fused with upsampled features from top-down layers with corresponding scales by the addition operator. Finally, FPN directly predicts from multiple scales of feature maps.

After FPN, there was a number of studies to develop advanced FPN. PANet [64] improved performance by attaching another bottom-to-top pyramid feature fusion layer. In addition, with significant progress on Neural Architecture Search(NAS) [65], NAS-FPN [66] and BIFPN [67] are introduced into an object detection task. The structures of those feature pyramid networks are automatically searched by reinforcement learning-based algorithms. NAS-based methods show an outstanding improvement in performance, but those methods have challenges, such as heavy storage and large computational complexities. Consequently, it is challenging to implement them in a regular computational environment.

Another approach is to apply multi-scale feature fusion by revising the backbone network itself. For example, HRNet [68] is designed as a feature extraction network by multi-scaling resolutions of input images by up-sampling. On the other hand, Res2Net [69] tries to apply a multi-scale mechanism in a channel-wise manner. In detail, it divides each 3×3 convolutional layer in the ResNet [52] bottleneck block into four scales of the channel.

Through channel multi-scaling, Res2Net improves feature extraction performance with a considerable computational cost.

The last approach is to apply a multi-scale feature fusion mechanism at the detection stage [70]. For example, a multi-scale strategy by dynamic detection heads is adopted in [71]. By applying multiple scale kernel sizes at the detection head with concatenation operation, average precision for object detection tasks can be improved. Other various approaches also show an outstanding performance by multi-scale feature fusion methods regardless of tasks in the computer vision research community [72,73].

2.3. Image Attention

The attention network helps the model to focus more on important features by refining input features. An attention mechanism is firstly adopted at natural language processing tasks. Then, the attention strategy is also successfully applied to image processing and analysis tasks. According to the data properties, there are two popular feature refining strategies, i.e., channel-wise [74] attention and pixel-wise [75].

Since pixels have plentiful spatial information for certain objects, Pixel-wise attention is normally appropriate for shape-biased tasks like object detection, instance segmentation and semantic segmentation [72]. By re-weighting, representative pixels can be extracted and left for final prediction from an input image. The non-local network [75] directly applied the self-attention mechanism from the natural language process research. The deformable convolutional network [76] tried to recalculate the position of every pixel during the convolutional operation rather than using a pre-defined anchored kernel. Empirical attention applies an attention mechanism for each factor in object detection to improve detection results [77].

Whereas channel-wise attention can refine the deep feature of each pixel, the Squeeze and Excitation network (SENet) [78] succeeds in achieving state-of-the-art performance without large computational cost with channel-wise attention. SENet finds that performance gain for image classification can be acquired by applying only channel-wise attention rather than the complicated self-attention network. In addition, SENet can be attached to any convolutional neural network-based feature extractor. The Global Contextual Network (GCNet) [74] revises and simplifies the non-local network [75] in a channel-wise manner based on SENet [78] and proves that channel-wise attention is better from the perspective of the tradeoff between performance gain and computational cost.

Combined attention denotes the attention network that applied both pixel-wise and channel-wise attention. Combined attention can achieve the best result but requires more computational resources. CBAM [79] and Dual Attention [80] successfully combined the two kinds of attention schemes in one model.

Recently, the concept of the attention network has been adopted to select features from an extracted feature map [71] or even utilized attention network itself to recognize images [81], for object detection [82] and conventional remote sensing scene classification [73]. Those methods also achieve comparable performance and computational cost compared to a traditional deep neural network, such as a convolutional neural network and recurrent neural network. Regardless of the task, attention networks come into the spotlight as a major topic in the research community due to their simplicity and effectiveness.

3. Self-Attention-Based Feature Fusion Network (SAFFNet)

In the task of scene classification, different classes in remote sensing images usually contain rich texture semantic features. Accordingly, multi-scale features from different resolutions need to be adopted for further feature or decision fusion. FPN and its extensions are proposed for object detection and then also applied to semantic segmentation. How to integrate the multi-scale features is an important but challenging problem in these tasks. As discussed in Section 2.2, FPN and its extension models, feature or decision fusion, use the features derived from all stages of the models.

The original concept of FPN [55] is to predict bounding boxes of various sizes (small objects and large objects) for the object detection task. Therefore, extracting pixel-wise features (positional-info) is crucial for object detection. FPN [55] tried to apply bottom to top and top-down layers with upsampling layers to reconstruct feature maps from the pooled feature maps. Consequently, it is better to directly regress the bounding box location from reconstructed feature maps (multiple decisions) for the object detection task. However, there is no need to regress bounding boxes in different sizes for remote sensing of few-shot scene classification tasks. In addition, the upsampling layer has a large computational complexity. Therefore, instead of upsampling and predicting directly, our main idea is to let the neural network select and highlight more important features from multiple scales of features to improve few-shot accuracy rather than extracting upsampled pixel-wise feature maps for bounding box prediction.

Our goal is to extract the rich texture semantic features from remote sensing images as shown in the second row of Figure 1. As global features from scene images can better capture contextual semantic information, channel-wise attention for feature selection is designed to "emphasize" the class-specific features in different resolutions derived from few support images for few-shot scene classification. Through cascading the feature selection module, the important features are reactivated and the trivial ones die. To do so, the features for the query image can be better matched with the self-attention fused features contained in the class-specific support images. Accordingly, the proposed model is denoted as SAFFNet (Self-Attention Feature Fusion Network). The overall architecture of SAFFNet is shown in Figure 3.

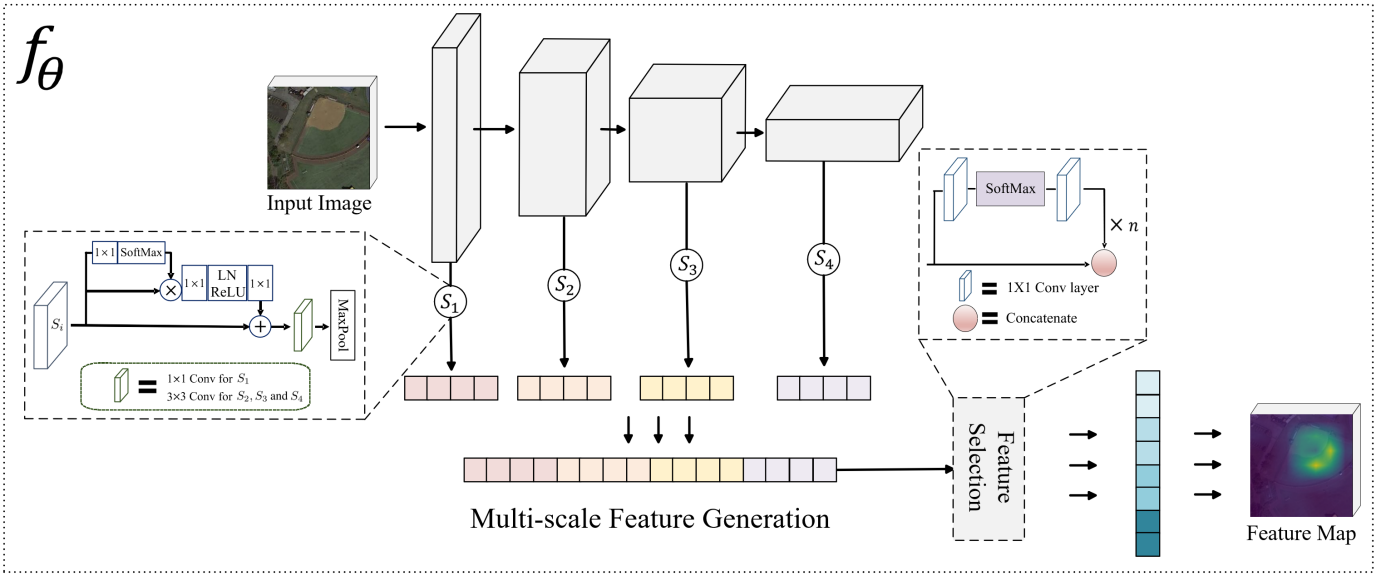


Figure 3. The overall architecture and details for the Self-attention-based Feature Fusion Network (shortened to SAFFNet). f_θ implies the feature extractor. By selecting more representative features (SAFS module) from multiple scales (MFG module), SAFFNet can extract plentiful texture features in remote sensing images. Note that LN denotes Layer Norm.

3.1. Multi-Scale Feature Generation

A multi-scale feature representation contains semantic information in different resolutions of remote sensing images using convolutional neural networks (CNN) for image classification. The corresponding feature maps contain the local salience of scene images and detail contextual information in lower levels of a CNN. However, more extracted features can be generated in higher levels for a pattern discriminant. To better classify remote sensing images, different levels of texture features should be adopted by integrating local semantic-rich and global extracted features in different receptive fields of a CNN.

In each stage, the feature maps are re-weighted by a channel-wise attention based on GCNet [74] in SAFFNet to refining deep features as shown in Figure 3. GCNet caculates

the attention map by a 1×1 convolutional layer and softmax and then transforms features by a 1×1 convolution and layer normalization function before aggregation. After refining, we adopt convolutional and global pooling layers to match the channel and size of feature maps and reduce computational costs by avoiding expensive fully connected (FC) layers in the decision stage of deep networks for few-shot learning. Furthermore, channel-wise features are left by global pooling to focus more on plentiful texture features. Finally, the refined features, i.e., S_1^*, S_2^*, S_3^* and S_4^* from each stage are concatenated as a channel-wise fused feature representation for further feature selection. Note that except for S_1 using 1×1, a 3×3 convolutional filter is adopted for the MFG module according to experimental results shown in Table 2. Note that all of the experimental results with bold text in this paper indicate the best accuracy within the comparison. The combination of $\{1 \times 1, 3 \times 3, 3 \times 3, 3 \times 3\}$ achieved the best accuracy regardless of the dataset. This suggests that emphasizing features with deeper stages (last three stages) by 3×3 kernel size improved performance rather than only using 1×1 or 3×3 kernels.

Table 2. Experimental result on the combination of a 1×1 convolutional layer and a 3×3 convolutional layer at each stage of a Multi-scale Feature Generation module with the AID dataset [50] and NWPU-RESISC45 dataset [20].Number 1 denotes the 1×1 convolutional layer and 3 denotes the 3×3 convolutional layer. For precise measurement, the experiment was conducted without the SAFS module.

Dataset	Combination $S_1S_2S_3S_4$	5-Way Acc (%) 5-Shot	1-Shot
NWPU	1 1 1 1	79.16 ± 0.67	60.88 ± 0.90
	3 3 3 3	76.10 ± 0.71	58.69 ± 0.82
	1 3 3 3	**80.43 ± 0.74**	**63.24 ± 0.87**
	3 3 3 1	75.20 ± 0.74	55.13 ± 0.82
	1 1 3 3	78.36 ± 0.68	58.47 ± 0.89
	3 3 1 1	75.27 ± 0.68	56.84 ± 0.84
AID	1 1 1 1	83.69 ± 0.49	66.61 ± 0.81
	3 3 3 3	83.67 ± 0.52	67.95 ± 0.79
	1 3 3 3	**86.22 ± 0.47**	**70.67 ± 0.75**
	3 3 3 1	85.44 ± 0.50	69.28 ± 0.74
	1 1 3 3	82.72 ± 0.57	65.48 ± 0.80
	3 3 1 1	83.34 ± 0.54	66.15 ± 0.79

3.2. Self-Attention-Based Feature Selection

How to integrate different levels of feature representation is a key concern for a multi-scale feature fusion model. Unlike FPN and its extensions, we adopt the concept of self-attention to select more informative features from the concatenated multi-scale feature after the multi-scale feature generation module as shown in Figure 3.

In detail, a 1×1 convolutional layer followed by the $softMax$ function is obtained to assign feature importance to each channel. Then, an additional 1×1 convolutional layer is utilized to self-emphasize the important features by re-activating the corresponding channels. This module is intended for feature selection by a self-attention strategy. Therefore, the module is denoted as the Self-Attention Feature Selection (SAFS) module.

To further pay more attention to the important features, the SAFS module is cascaded to generate better feature representation by the concatenation operation. Experimental results in Table 3 demonstrate that we can obtain better performance by repeating the SAFS module three times. However, the weights of each convolutional layer in cascaded SAFS modules are shared to give more weight to selected features adaptively. In detail, the convolutional layer of the SAFS module shares the weight with the first convolutional layer in the previous SAFS module except for the first convolutional layer in the first SAFS modules since the input size is different. Furthermore, the performance of is decreased

four times; we supposed that shared weights for each convolutional layer are overfitted four times.

Table 3. Experiments to measure performance according to the number of feature selection layers with the NWPU-RESISC45 dataset [20] since this dataset has the largest volume. n means the number of feature selection blocks.

Dataset	n	5-Way Acc (%)	
		5-Shot	1-Shot
NWPU	1	78.36 ± 0.73	64.00 ± 0.82
	2	80.22 ± 0.65	63.95 ± 0.86
	3	**81.32 ± 0.62**	**67.23 ± 0.85**
	4	79.94 ± 0.74	63.47 ± 0.87
	5	79.40 ± 0.73	62.53 ± 0.92
	6	77.32 ± 0.76	65.11 ± 0.86

In addition, to verify the effectiveness of architecture for the SAFS module, experiments conducted on two datasets are shown in Table 4. The structure of the SAFS module denotes the functions between two 1×1 convolutional layers. Since the $Softmax$ function can be regarded as both the activation and normalization function, other combination activations and normalization functions are experimented with to prove that the improvement of the proposed method is derived from simple normalization or feature selection. Batchnorm [83] (Batch normalization) + ReLU [84] denotes applying the ReLU (Rectified Linear Units) activation function after the batch normalization operation between convolutional layers and Batchnorm+Softmax denotes applying the Softmax function after the batch normalization layer between two convolutional layers. A combination of $Softmax$ with Batchnorm showed the worst few-shot accuracy within our comparison. This clearly denotes that normalization and $Softmax$ itself is not the critical part of improvement. The combination of batch normalization with the ReLU function showed the best result on the AID dataset. However, few-shot classification accuracies were relatively low on the other dataset (NWPU). Single batch normalization improved the performance for 1-shot but showed relatively low few-shot accuracy on the 5-shot task. This denotes that normalization is more critical to a 1-shot (extremely limited sample) task than a 5-shot task. In general, the combination of the $Softmax$ activation function with batch normalization operation showed the best performance within experiments conducted with three datasets. This denotes that the effectiveness of the proposed feature selection method is not only derived from normalization. Therefore, this combination is adopted to the proposed SAFS module.

Other than a combination of functions, an experiment was also conducted to compare the addition concatenation operation and the addition operation at the end of the SAFS module. Experimental results shown in Table 5. Except for the 1-shot task with the AID dataset, all the experiment results for the concatenation operation outperform the addition operation. Features can be automatically selected from concatenated features from the previous layer; however, for addition, features can be just referred to the next layer. Consequently, the basis for adopting a concatenation operation at the end of the SAFS modules was provided by the experiment.

Table 4. Experiment to adopt the best architecture for the SAFS module. Note that all the experiments are based on the SAFFNet50 model.

Dataset	SAFS Structure	5-Way Acc (%)	
		5-Shot	1-Shot
AID	Softmax	**86.91 ± 0.44**	70.80 ± 0.71
	Batchnorm	85.49 ± 0.48	72.35 ± 0.74
	ReLU	86.25 ± 0.47	70.61 ± 0.72
	Batchnorm+ReLU	86.84 ± 0.44	**72.38 ± 0.74**
	Batchnorm+Softmax	84.34 ± 0.47	64.89 ± 0.73
NWPU	Softmax	**81.32 ± 0.62**	**67.23 ± 0.85**
	Batchnorm+ReLU	78.11 ± 0.63	57.89 ± 0.94

Table 5. Experiment to compare the concatenation operation and addition operation at the end of the SAFS module with SAFFNet50.

Dataset	Concatenation	Addition	5-Way Acc (%)	
			5-Shot	1-Shot
NWPU	✓	×	**81.32 ± 0.62**	**67.23 ± 0.85**
	×	✓	79.28 ± 0.68	60.62 ± 0.87
AID	✓	×	**86.91 ± 0.44**	70.80 ± 0.71
	×	✓	84.83 ± 0.46	**71.17 ± 0.71**
UCM	✓	×	**86.79 ± 0.33**	**73.46 ± 0.55**
	×	✓	84.51 ± 0.26	71.17 ± 0.71

3.3. Training and Prediction

The few-shot scene classification problem can be defined as an N-way K-shot problem. N denotes N number of unseen classes, defined as $S = \{s_1, \ldots, s_N\}$, where each class contains K labeled samples, e.g., for the class n with new labeled set. It can be defined as $s_n = \{(x_{n1}, y_{n1}), \ldots, (x_{nK}, y_{nK})\}$. The $Y \in \{1, \ldots, N\}$ is the set of corresponding labels in the support set. Predicting class labels in the list of Y for unlabeled unseen scene samples is the goal of few-shot classification task. Given a query scene image $\hat{x}$ and the label of the query set $\hat{y}$ can be predicted by the support set S as follows

$$\hat{y} = \arg\max_{y_i \in Y} P(y_i|\hat{x}, s_i). \tag{1}$$

For training strategy, a fine-tuning strategy is widely and successfully exploited to refine the feature representation for a small target dataset. Since our model focuses on feature extraction rather than training strategies, we adopt the $Baseline++$ training strategy in [85] due to its simple implementation and effectiveness. In detail, the SAFFNet is fine-tuned by the support set S. At the beginning, the pre-trained model f_θ is obtained by the large volume of in-bag training datasets from scratch. Then, for new classes, the pre-trained backbone model is frozen except for the classifier for a final decision based on the $Softmax$ function. The overall architecture is shown in Figure 4.

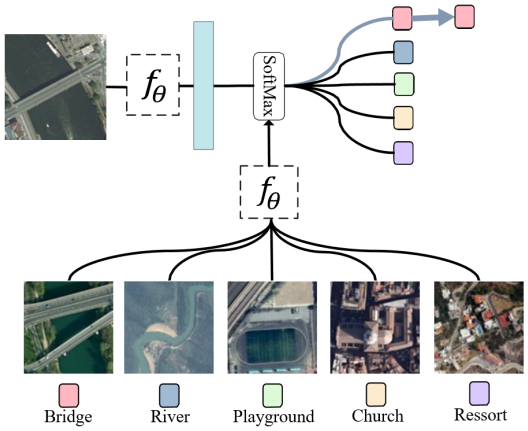


Figure 4. Few-shot learning for SAFFNet. Here, f_θ denotes the feature extractor by SAFFNet shown in Figure 3 and $SoftMax$ denotes a prediction function Equation (4) based on cosine distance Equation (2).

According to [85], the cosine similarity metric with a softmax function is adopted as a classifier for a final decision. Specifically, a cosine function is used to measure the similarity between the feature vector q' of a query image and those from the support images.

In detail, the similarity between the query image qi belonging to the class n and the support image si belonging to the class n can be calculated by

$$d(qi', si_n) = \frac{qi' \cdot si_n}{\sqrt{|\ qi'\ |^2} \times \sqrt{|\ si_n\ |^2}}. \tag{2}$$

To obtain a prediction result, the softmax function is applied to compute the posterior probability as follows.

$$P(y = n|qi', si_n) = \frac{\exp(d(qi', si_n))}{\sum_{j=1}^{N} \exp(d(qi', si_j))}. \tag{3}$$

Based on the similarity between query images and support images, the prediction of the query image can be predicted by maximizing posterior probabilities over C classes.

$$\hat{y} = \arg\max_n P(y = n|qi', si_n). \tag{4}$$

4. Experiments

4.1. Datasets

To verify the effectiveness and performance of the proposed SAFFNet, three scene classification datasets are divided into base training, validation and test sets, respectively. The UC Merced land-use dataset [86] contains 2100 land-use images with 21 classes, and each class has 100 corresponding scene images with the size of 256 × 256 pixels in the RGB color space. This dataset is manually collected from the urban area imageries in various areas around the country acquired by the United States Geological Survey (USGS). For the UC Merced dataset, the number of classes of the training, validation and test set were randomly selected from the 21 scene classes as 11, 5 and 5, respectively.

The AID dataset [50] is a large-scale aerial image database created by collecting sample images from Google Earth Imagery and images are post-processed using RGB rendering from the original optical aerial images. This dataset comprises 10,000 images with 30 scene classes. Those classes are randomly split into 15, 8 and 7. Classes of the test set are beach, commercial, forest, mountain, pond, river and stadium, respectively. Furthermore, all the images of each class in the AID dataset are carefully chosen from different countries

and regions around the world to increase the within-class diversity. Therefore, the AID dataset is a more challenging dataset for the scene classification task compared with the UC Merced dataset.

The NWPU-RESISC45 dataset [20] contains 31,500 remote sensing scene images with 45 scene classes, and each class consists of 700 aerial images as shown in Figure 5. The size of all images is 256 × 256 pixels in the RGB color space. These 45 scene classes are randomly split into 23, 11 and 11, respectively. Classes of the test set are basketball court, church, dense residential, golf course, intersection, medium residential, palace, rectangular farmland, sea ice, stadium and thermal power station. This is the largest publicly available dataset in the remote sensing scene classification task.

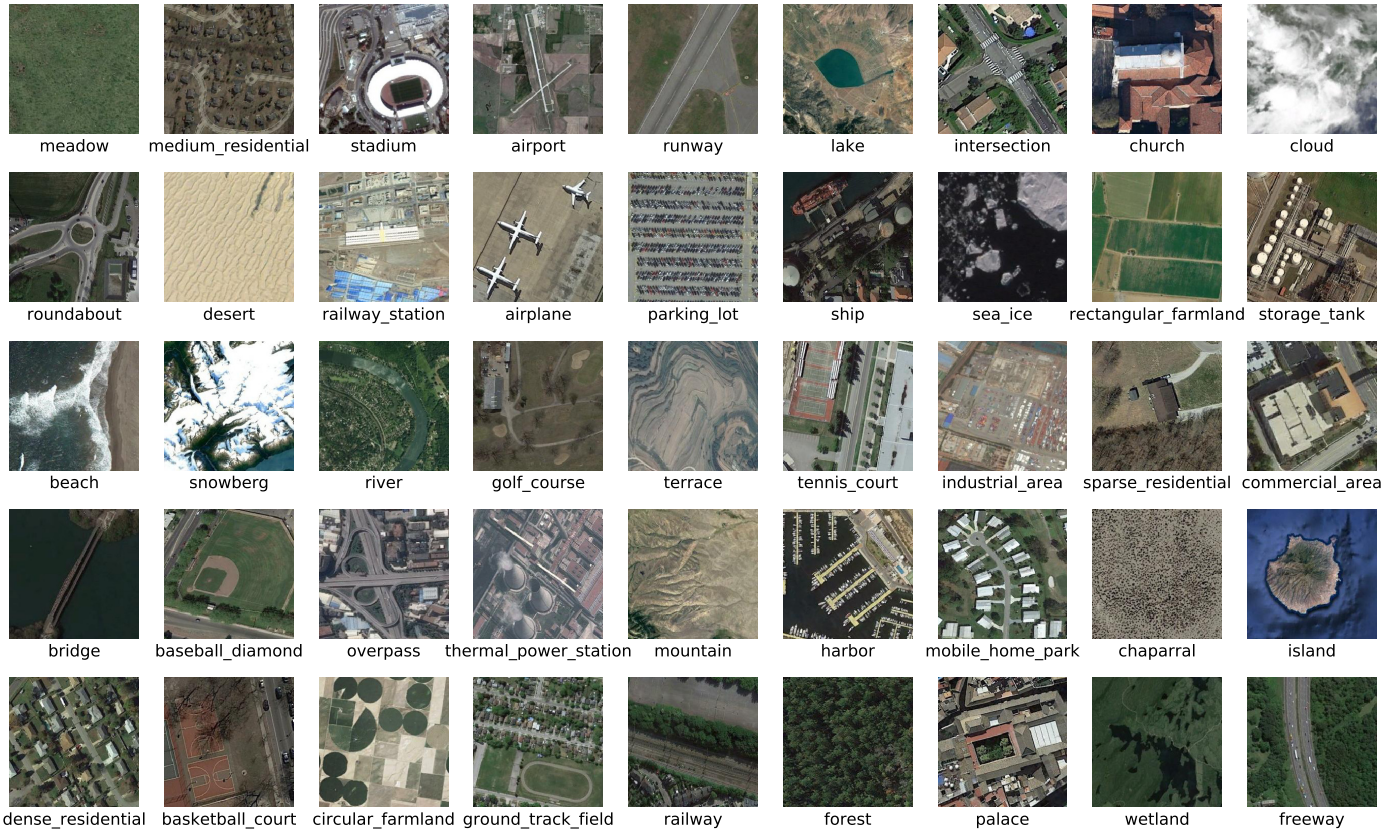


Figure 5. Images (with 256 × 256 pixels) sampled from 45 classes in the NWPU-RESISC45 [20] dataset. We adopted basketball court, church, dense residential, golf course, intersection, medium residential, palace, rectangular farmland, sea ice, stadium and thermal power station as novel classes for few-shot classification.

4.2. Experimental Settings

The SAFFNet is trained based on ResNet 101, 50, 34 and 18 from scratch. The batch size is set to 32 which denotes that each batch contains 32 iterations for training. Standard data augmentation including, random crop, left–right flip and color jitter are applied. Total epochs at the base training level is set to 600. The learning rate is set to 0.001. The models are trained via an Adam Optimizer [87], and $\beta 1$ and $\beta 2$ are set to 0.9 and 0.999, respectively, which control the exponential decay rates. The input size of image is resized to 224 × 224 pixels. Due to the quantity limitation of classes, only 5-way 1-shot and 5-way 5-shot tasks are adopted as a comparison in the following experiments. The loss function for base training and few-shot training is adopted as cross entropy loss for both base training and few-shot training.

4.3. Experimental Results

First of all, the evaluation metric for our experiments as follows. The most similar feature S_n in N class is retrieved based on the similarity function $d(\cdot,\cdot)$ Equation (2) given the embedding feature vectors of the query images $\{q'_i,\ldots,q'_{N_q}\}$. It is assumed that the query samples and their corresponding most similar features belong to the same class. Accordingly, the evaluation metric is defined by

$$Acc = \frac{1}{N_q}\sum_{i=1}^{N_q}[i = \arg\min_j(d(q'_i, S_j))] \tag{5}$$

To validate the effectiveness of the proposed SAFFNet, other few-shot methods are used including prototypical network (denoted as PN) [30], matching net (denoted as MN) [29] and relation net (denoted as RN) [31]. The experiments are carried out on the NWPU, AID and UCM datasets. For a fair comparison, all models are trained from scratch and hyperparameters of the backbones are set to the same as SAFFNet. Predictions results shown in Figure 6. From Table 6, one can see that SAFFNet significantly improves the classification accuracies in both 5-shot and 1-shot learning. This result denotes that feature extraction and selection is a more important factor to improve remote sensing few-shot scene classification rather than training strategy. In addition, we conduct experiments to compare it with other multi-scale deep feature fusion methods. We adopt the Res2Net [69] and FPN (Feature Pyramid Network) [55]. As shown in Table 7, SAFFNet obtains the best results compared with other deep feature fusion methods. Furthermore, the experimental result also proved the effectiveness of the intelligent feature selection strategy by the neural network itself rather than direct feature fusion.

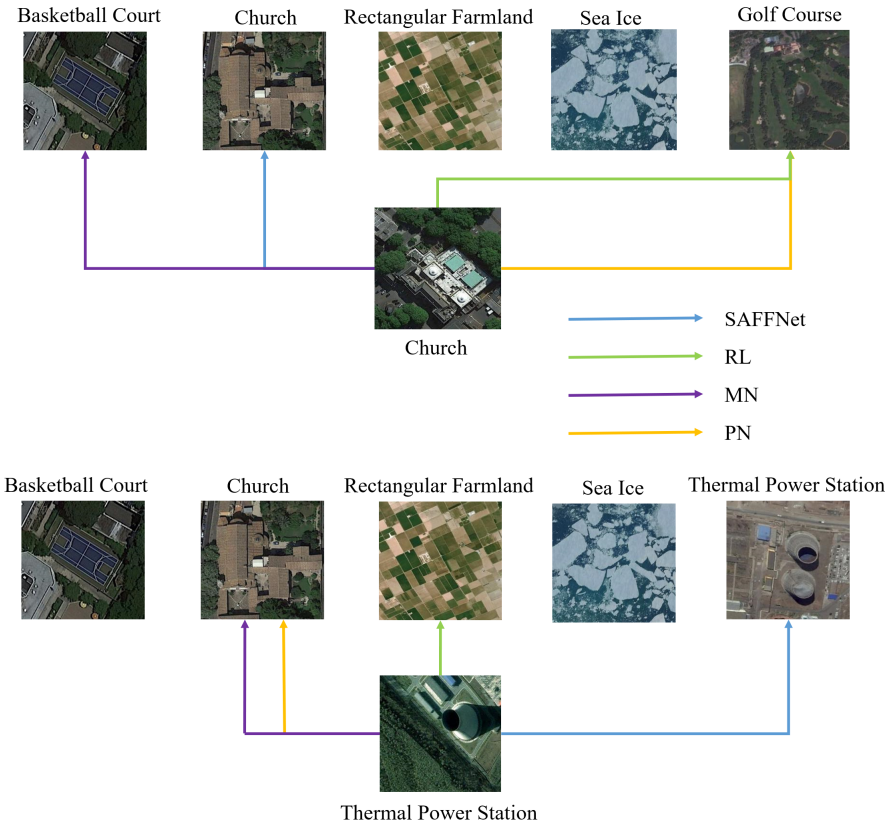


Figure 6. Inference results on the NWPU Dataset.

Table 6. Few-shot scene classification accuracies on three datasets provided by the proposed SAFFNet, ProtoNet [30], MathcingNet [29] and RelationNet [31] models on the ResNet18, denoted as SAFFNet18, PN-ResNet18, MN-ResNet18 and RN-ResNet18, respectively.

Dataset	Methods	5-Way Acc (%)	
		5-Shot	1-Shot
NWPU	PN-ResNet18	76.43 ± 0.74	59.35 ± 0.92
	MN-ResNet18	72.65 ± 0.76	57.40 ± 0.90
	RN-ResNet18	73.97 ± 0.73	56.04 ± 0.91
	SAFFNet18	**79.59 ± 0.66**	**63.83 ± 0.80**
AID	PN-ResNet18	80.09 ± 0.56	55.80 ± 0.83
	MN-ResNet18	70.22 ± 0.69	60.07 ± 0.83
	RN-ResNet18	73.55 ± 0.60	51.78 ± 0.78
	SAFFNet18	**83.77 ± 0.50**	**67.14 ± 0.74**
UCM	PN-ResNet18	77.38 ± 0.31	56.47 ± 0.55
	MN-ResNet18	72.46 ± 0.35	61.51 ± 0.63
	RN-ResNet18	70.02 ± 0.42	57.35 ± 0.60
	SAFFNet18	**81.20 ± 0.57**	**65.02 ± 0.87**

Table 7. Comparison with other multi-scale feature fusion methods, i.e., Res2Net [69] and FPN [55]. Note that we concatenate features instead of direct prediction for FPN to obtain better results for the classification task.

Dataset	Methods	5-way Acc (%)	
		5-Shot	1-Shot
NWPU	ResNet50	77.55 ± 0.69	61.06 ± 0.95
	Res2Net50	81.09 ± 0.66	64.76 ± 0.95
	FPN	81.10 ± 0.64	62.84 ± 0.89
	SAFFNet	**81.32 ± 0.62**	**67.23 ± 0.85**
AID	ResNet50	82.34 ± 0.55	65.50 ± 0.83
	Res2Net50	84.41 ± 0.57	69.02 ± 0.69
	FPN	84.02 ± 0.52	67.71 ± 0.81
	SAFFNet	**86.91 ± 0.44**	**70.08 ± 0.71**
UCM	ResNet50	68.66 ± 0.40	56.45 ± 0.64
	Res2Net50	78.11 ± 0.34	67.06 ± 0.66
	FPN	77.72 ± 0.28	69.57 ± 0.34
	SAFFNet	**86.79 ± 0.33**	**73.46 ± 0.55**

To further validate the effectiveness of the proposed method, four different backbone networks were exploited for further comparison as shown in Table 8. Here, the backbone networks are based on ResNet 101, 50, 34 and 18, respectively. The experimental results show that the proposed SAFFNet can significantly increase both few-shot classification accuracies and robustness in both the 5-way 5-shot task and 5-way 1-shot task in all datasets compared with the backbone models without feature fusion.

Table 8. Comparison results for 5-shot and 1-shot scene classification provided by the proposed SAFFNet and ResNet.

Dataset	Model	5-Way Acc (%)	
		5-Shot	1-Shot
NWPU	ResNet18	77.41 ± 0.72	60.73 ± 0.91
	SAFFNet18	**79.59 ± 0.66**	**63.83 ± 0.80**
	ResNet34	75.25 ± 0.70	58.08 ± 0.92
	SAFFNet34	**80.82 ± 0.54**	**60.24 ± 0.87**
	ResNet50	77.55 ± 0.69	61.06 ± 0.95
	SAFFNet50	**81.32 ± 0.62**	**67.23 ± 0.85**
	ResNet101	78.25 ± 0.68	61.50 ± 0.90
	SAFFNet101	**79.66 ± 0.71**	**63.23 ± 0.86**
AID	ResNet18	80.30 ± 0.63	63.12 ± 0.80
	SAFFNet18	**83.77 ± 0.50**	**67.14 ± 0.74**
	ResNet34	80.32 ± 0.60	64.38 ± 0.78
	SAFFNet34	**84.56 ± 0.49**	**66.96 ± 0.70**
	ResNet50	82.34 ± 0.55	65.50 ± 0.83
	SAFFNet50	**86.91 ± 0.44**	**70.80 ± 0.71**
	ResNet101	80.76 ± 0.56	64.40 ± 0.82
	SAFFNet101	**83.78 ± 0.52**	**67.23 ± 0.77**
UCM	ResNet18	77.26 ± 0.43	61.31 ± 0.61
	SAFFNet18	**84.16 ± 0.37**	**66.57 ± 0.64**
	ResNet34	74.22 ± 0.44	59.85 ± 0.70
	SAFFNet34	**84.41 ± 0.36**	**64.62 ± 0.75**
	ResNet50	68.66 ± 0.40	56.45 ± 0.64
	SAFFNet50	**86.79 ± 0.33**	**73.46 ± 0.55**
	ResNet101	72.04 ± 0.43	59.32 ± 0.61
	SAFFNet101	**83.33 ± 0.37**	**68.66 ± 0.52**

4.4. Ablation Study

For the ablation study, we conducted experiments to validate the influence of MFG and SAFS modules of SAFFNet. Experimental results are shown in Table 9 which prove that two modules improve the classification accuracies for the few-shot scene classification problem. In particular, the improvement by the MFG module is bigger on the 5-shot problem while the SAFS module obtains a larger increase on the 1-shot problem. By aggregating the two modules, SAFFNet achieves the best results on the both 5-shot and 1-shot problems.

Table 9. Ablation study to verify the influence of the MFG and SAFS modules.

Dataset	MFG	SAFS	5-Way Acc (%)	
			5-Shot	1-Shot
NWPU	×	×	77.50 ± 0.69	61.06 ± 0.95
	✓	×	80.43 ± 0.74	63.24 ± 0.87
	✓	✓	**81.32 ± 0.62**	**67.23 ± 0.85**

To measure the influence of fused features at each stage, we conducted an experiment on different combinations of feature fusion and exploited our model to the 5-way 5-shot and 5-way 1-shot tasks by the proposed SAFFNet model based on the ResNet50 network in terms of the NWPU datasets shown in Table 10. Feature vector S_4 denotes the final feature map output without a multi-scale strategy, which is used for the other networks, e.g., the PN, MN and RL models. The outputs from the first stage to last stage, i.e., from the first stage of the base CNN model, are denoted as S_1, S_2, S_3 and S_4 shown in Figure 3. As shown in Table 10, the proposed MFG module using the combination of the features S_1, S_2, S_3 with S_4 showed the best classification accuracy on the NWPU dataset in both 5-way 5-shot and 5-way 1-shot tasks.

This empirically confirms the effectiveness of the proposed refined multi-scale feature fusion network extracting better features from texture biased remote sensing scene images. Moreover, regardless of any combination of multi-scale features fusing from the first stage to the last stage, the classification accuracy can be significantly improved by the proposed MFG and SAFS module in the remote sensing few-shot learning scene classification task. This further proves that the proposed model can achieve better classification accuracy and robustness as well.

Table 10. Accuracies provided by the proposed SAFFNet using the NWPU dataset in terms of different combinations of multi-scale feature fusions.

Dataset	Fused Stages	5-Way Acc (%)	
		5-Shot	1-Shot
NWPU	$\{S_4\}$	76.89 ± 0.70	58.18 ± 0.84
	$\{S_4, S_3\}$	78.05 ± 0.69	59.92 ± 0.87
	$\{S_4, S_3, S_2\}$	78.48 ± 0.73	62.03 ± 0.91
	$\{S_4, S_3, S_2, S_1\}$	**80.43** ± 0.74	**63.24** $\pm$ **0.87**

To further verify the influence of the number of shots on the few-shot learning scene classification results, 5-way K-shot($K = \{1, 2, 3, 4, 5, 6, 7, 8, 9$ and $10\}$) tasks are experimented on the NWPU dataset. To be a fair comparison with the proposed SAFFNet, all experiments were based on the ResNet50 network. The "shot" denotes the number of new training instances of each unseen class. As shown in Figure 7, with the increase in the number of training "shots" K, the few-shot classification accuracies are smoothly improved on the NWPU dataset by the proposed SAFFNet in 5-way K-shot tasks.

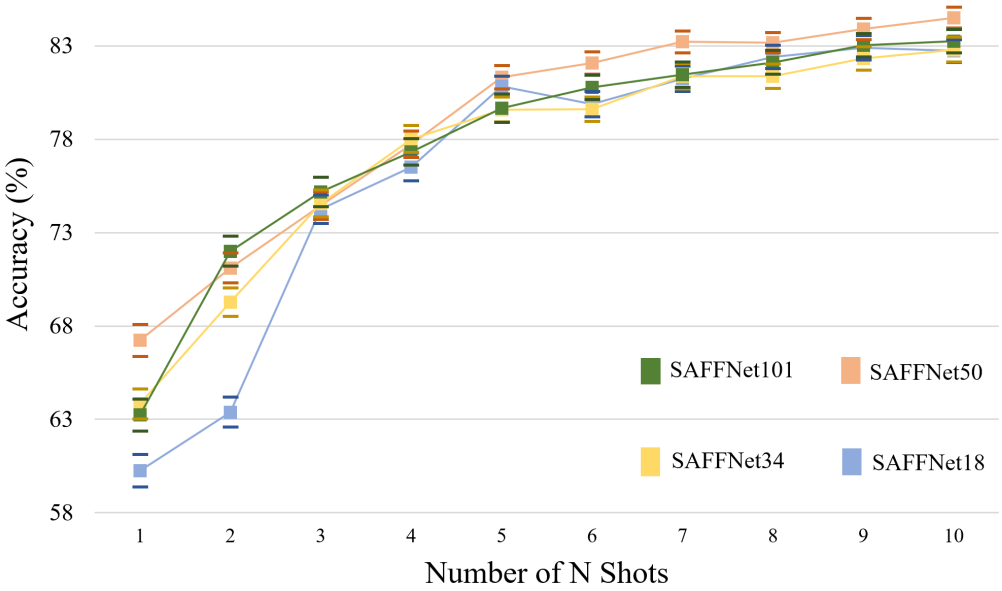


Figure 7. The influence of training "shots" on classification accuracies conducted on the NWPU dataset by the proposed SAFFNet based on each ResNet Network, denoted as SAFFNet101, SAFFNet50, SAFFNet34 and SAFFNet18 in 5-way K-shot tasks when $K = \{1, 2, 3, 4, 5, 6, 7, 8, 9,$ and $10\}$, respectively.

Furthermore, the influence of the number of "ways" on the classification result is also evaluated by N-way 5-shot and 1-shot (where $N = \{3, 4, 5, 6, 7, 8, 9,$ and $10\}$) tasks and the experiments are conducted on the NWPU dataset for fair comparison by the proposed SAFFNet model. The "way" denotes the number of given classes in the test set. As shown in Figure 8, with the increase in the number N, the classification accuracies are gradually decreased on the NWPU dataset.

Regardless of shots and ways, errors for few-shot prediction also gradually decreased according to the increase in shots and ways. This denotes that errors become smaller when there are more data samples for few-shot classification.

Lastly, few-shot classification accuracy according to the depth of backbone network is demonstrated in Figure 9. The experiment confirmed that a deeper backbone network does not always ensure the better few-shot classification accuracy in both 5-shot and 1-shot tasks. Furthermore, the experiment results with the 5-shot task showed a more stable curve according to the depth of the network. This also suggests that additional parameters do not always ensure better few-shot accuracy with an extremely small amount of samples.

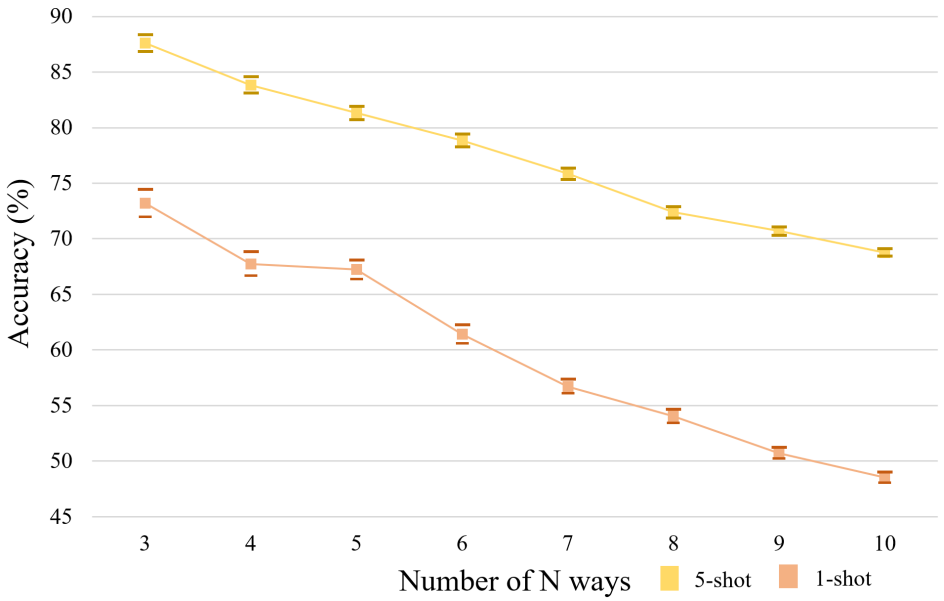


Figure 8. The influence of given "ways" N on classification accuracies by the proposed SAFFNet model based on the ResNet50, in the N-way 1-shot and 5-shot tasks when $N = \{3, 4, 5, 6, 7, 8, 9 \text{ and } 10\}$, respectively.

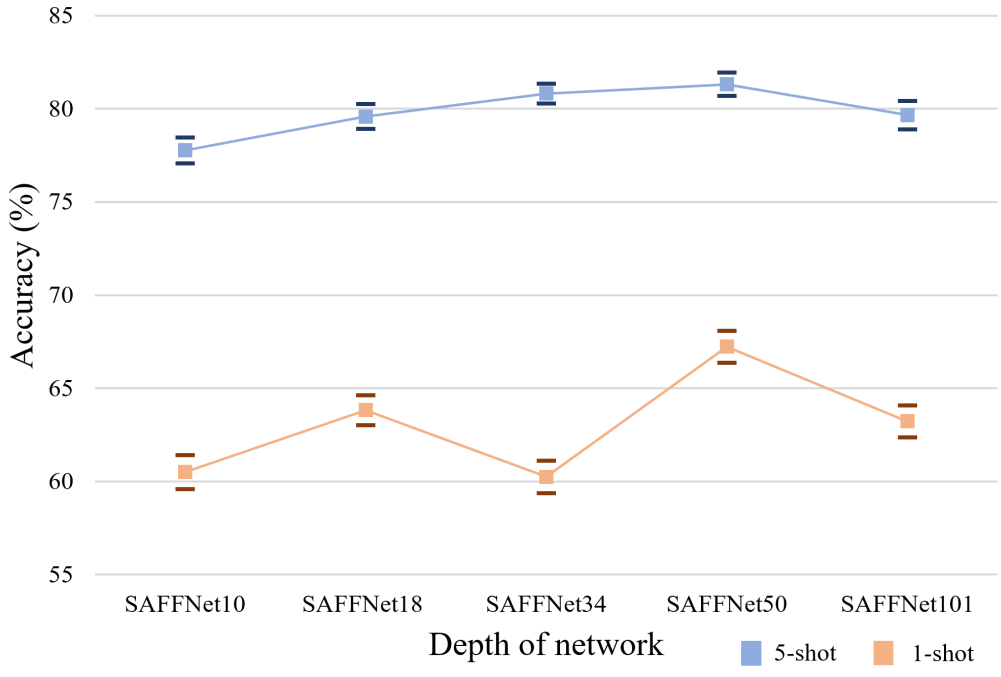


Figure 9. The influence of the depth of the backbone network.

5. Conclusions

In the paper, a self-attention feature selection module was proposed for deep feature fusion in a multi-scale structure for a few-shot remote sensing image classification. The method is denoted as SAFFNet, i.e., a self-attention-based feature fusion network. SAFFNet consists of two modules, i.e., multi-scale feature generation (MFG) and self-attention feature selection modules, respectively. In the MFG module, refining feature maps in different resolutions with richer semantic information are generated by channel-wise attention based on GCNet [74] for a scene image. To do so, the feature importance in different receptive fields can be automatically computed. After that, refined features from different scales are concatenated as input to the SAFS module for further feature fusion. In each SAFS module, the important features with bigger coefficients are highlighted while the trivial ones with smaller parameters decay. Finally, the module is cascaded to concatenate the features generated from the previous module. Experiments conducted on three remote sensing scene image datasets confirm the effectiveness of the proposed SAFFNet by significantly improving scene classification accuracies compared with the existing few-shot scene classification models and multi-scale feature fusion methods as well.

Although the SAFFNet has shown outstanding performances on a few-shot learning scene classification task for remote sensing images, the proposed SAFFNet still requires few unseen training samples to achieve more efficient and meaningful training for fine-tuning of a CNN backbone network. As a future development, a generative adversarial networks-based algorithm [88] or semi-supervised learning approaches [89,90] will be adopted to deal with this problem in a data-driven perspective. Additionally, other efficient CNN architectures will be searched by a neural architecture search [65] strategy for few-shot remote sensing land-cover/use image classification tasks.

Author Contributions: Methodology, software, experiment, and writing—original draft preparation, J.K.; Conceptualization, methodology, writing—review and editing, supervision, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by Science and technology research project of Sinopec under contract (no. pe19003-3) and in part by Zhongshan science and technology development project under contract (no. 2020AG016).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Martha, T.R.; Kerle, N.; van Westen, C.J.; Jetten, V.; Kumar, K.V. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4928–4943. [CrossRef]
2. Cheng, G.; Guo, L.; Zhao, T.; Han, J.; Li, H.; Fang, J. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *Int. J. Remote Sens.* **2013**, *34*, 45–59. [CrossRef]
3. Oštir, K.; Veljanovski, T.; Podobnikar, T.; Stančič, Z. Application of satellite remote sensing in natural hazard management: The Mount Mangart landslide case study. *Int. J. Remote Sens.* **2003**, *24*, 3983–4002. [CrossRef]
4. Gitas, I.; Polychronaki, A.; Katagis, T.; Mallinis, G. Contribution of remote sensing to disaster management activities: A case study of the large fires in the Peloponnese, Greece. *Int. J. Remote Sens.* **2008**, *29*, 1847–1853. [CrossRef]
5. Stumpf, A.; Kerle, N. Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577. [CrossRef]
6. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
7. Cheng, G.; Han, J.; Guo, L.; Qian, X.; Zhou, P.; Yao, X.; Hu, X. Object detection in remote sensing imagery using a discriminatively trained mixture model. *ISPRS J. Photogramm. Remote Sens.* **2013**, *85*, 32–43. [CrossRef]
8. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
9. Yao, X.; Han, J.; Guo, L.; Bu, S.; Liu, Z. A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and CRF. *Neurocomputing* **2015**, *164*, 162–172. [CrossRef]
10. Li, N.; Cao, M.; He, C.; Wu, B.; Jiao, J.; Yang, X. A multi-parametric indicator design for ECT sensor optimization used in oil transmission. *IEEE Sensors J.* **2017**, *17*, 2074–2087. [CrossRef]

11. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, Seoul, Korea, 27–28 October 2019; pp. 8231–8240. [CrossRef]
12. Wang, Y.; Zhang, L.; Tong, X.; Zhang, L.; Zhang, Z.; Liu, H.; Xing, X.; Mathiopoulos, P.T. A three-layered graph-based learning approach for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6020–6034. [CrossRef]
13. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 818–832. [CrossRef]
14. Shyu, C.R.; Klaric, M.; Scott, G.J.; Barb, A.S.; Davis, C.H.; Palaniappan, K. GeoIRIS: Geospatial information retrieval and indexing system—Content mining, semantics modeling, and complex queries. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 839–852. [CrossRef]
15. Li, Y.; Zhang, Y.; Tao, C.; Zhu, H. Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. *Remote Sens.* **2016**, *8*, 709. [CrossRef]
16. Yuan, F.; Sawaya, K.E.; Loeffelholz, B.C.; Bauer, M.E. Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing. *Remote Sens. Environ.* **2005**, *98*, 317–328. [CrossRef]
17. Stefanov, W.L.; Ramsey, M.S.; Christensen, P.R. Monitoring urban land cover change: An expert system approach to land cover classification of semiarid to arid urban centers. *Remote Sens. Environ.* **2001**, *77*, 173–185. [CrossRef]
18. El-Kawy, O.A.; Rød, J.; Ismail, H.; Suliman, A. Land use and land cover change detection in the western Nile delta of Egypt using remote sensing data. *Appl. Geogr.* **2011**, *31*, 483–494. [CrossRef]
19. Lima, R.P.D.; Marfurt, K. Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
20. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
21. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 123–153. [CrossRef]
22. Li, X.; Liu, B.; Zheng, G.; Ren, Y.; Zhang, S.; Liu, Y.; Gao, L.; Liu, Y.; Zhang, B.; Wang, F. Deep-learning-based information mining from ocean remote-sensing imagery. *Natl. Sci. Rev.* **2020**, *7*, 1584–1605. [CrossRef]
23. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
24. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, A. Skip-Connected Covariance Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 1461–1474. [CrossRef]
25. G'omez, P.; Meoni, G. MSMatch: Semi-Supervised Multispectral Scene Classification with Few Labels. *arXiv* **2021**, arXiv:2103.10368.
26. Schmitt, M.; Wu, Y.L. Remote Sensing Image Classification with the SEN12MS Dataset. *arXiv* **2021**, arXiv:2104.00704.
27. Bruzzone, L.; Marconcini, M. Toward the Automatic Updating of Land-Cover Maps by a Domain-Adaptation SVM Classifier and a Circular Validation Strategy. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1108–1122. [CrossRef]
28. Gómez-Chova, L.; Camps-Valls, G.; Muñoz-Marí, J.; Calpe-Maravilla, J. Semisupervised Image Classification With Laplacian Support Vector Machines. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 336–340. [CrossRef]
29. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; pp. 3630–3638.
30. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *arXiv* **2017**, arXiv:1703.05175.
31. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
32. Li, A.; Lu, Z.; Wang, L.; Xiang, T.; Wen, J.R. Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4157–4167. [CrossRef]
33. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems 2013, Stateline, NV, USA, 5–10 December 2013; pp. 3111–3119.
34. Dong, X.; Zheng, L.; Ma, F.; Yang, Y.; Meng, D. Few-shot Object Detection. *arXiv* **2017**, arXiv:1706.08249.
35. Lampinen, A.K.; McClelland, J.L. One-shot and few-shot learning of word embeddings. *arXiv* **2017**, arXiv:1710.10280.
36. Brown, T.B.; Mann, B.P.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
37. Lake, B.; Lee, C.y.; Glass, J.; Tenenbaum, J. One-shot learning of generative speech concepts. In Proceedings of the Annual Meeting of the Cognitive Science Society, Quebec City, QC, Canada, 23–26 July 2014; Volume 36.
38. Wu, D.; Zhu, F.; Shao, L. One shot learning gesture recognition from rgbd images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 7–12.

39. Kim, M.; Zuallaert, J.; De Neve, W. Few-shot Learning Using a Small-Sized Dataset of High-Resolution FUNDUS Images for Glaucoma Diagnosis. In Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, Mountain View, CA, USA, 23 October 2017; ACM: New York, NY, USA, 2017; pp. 89–92.
40. Liu, M.Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-Shot Unsupervised Image-to-Image Translation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
41. Altae-Tran, H.; Ramsundar, B.; Pappu, A.S.; Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **2017**, *3*, 283–293. [CrossRef] [PubMed]
42. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
43. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
44. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *arXiv* **2017**, arXiv:1707.09835.
45. Munkhdalai, T.; Yu, H. Meta networks. In Proceedings of the International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017.
46. Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. One-shot learning with memory-augmented neural networks. *arXiv* **2016**, arXiv:1605.06065.
47. Kaiser, Ł.; Nachum, O.; Roy, A.; Bengio, S. Learning to remember rare events. *arXiv* **2017**, arXiv:1703.03129.
48. Ramalho, T.; Garnelo, M. Adaptive Posterior Learning: Few-shot learning with a surprise-based memory module. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
49. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 2.
50. 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]
51. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–21 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
52. 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.
53. Qiao, S.; Chen, L.C.; Yuille, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *arXiv* **2020**, arXiv:2006.02334.
54. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
55. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 4.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
57. Kim, S.; Kook, H.K.; Sun, J.Y.; Kang, M.C.; Ko, S. Parallel Feature Pyramid Network for Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
58. Satorras, V.G.; Estrach, J.B. Few-Shot Learning with Graph Neural Networks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
59. Ji, Z.; Chai, X.; Yu, Y.; Pang, Y.; Zhang, Z. Improved prototypical networks for few-Shot learning. *Pattern Recognit. Lett.* **2020**, *140*, 81–87. [CrossRef]
60. Wang, K.; Liew, J.; Zou, Y.; Zhou, D.; Feng, J. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 9196–9205.
61. Zheng, Y.D.; Ma, Y.; Liu, R.Z.; Lu, T. A Novel Group-Aware Pruning Method for Few-shot Learning. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–7.
62. Yuan, Z.; Huang, W.; Li, L.; Luo, X. Few-Shot Scene Classification With Multi-Attention Deepemd Network in Remote Sensing. *IEEE Access* **2021**, *9*, 19891–19901. [CrossRef]
63. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. *arXiv* **2020**, arXiv:2012.07177.
64. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
65. Zoph, B.; Le, Q.V. Neural architecture search with reinforcement learning. *arXiv* **2016**, arXiv:1611.01578.
66. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
67. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
68. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.

69. Gao, S.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P.H. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
70. Singh, B.; Najibi, M.; Davis, L.S. Sniper: Efficient multi-scale training. *arXiv* **2018**, arXiv:1805.09300.
71. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11207–11216.
72. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
73. Cao, R.; Fang, L.; Lu, T.; He, N. Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 43–47. [CrossRef]
74. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
75. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
76. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
77. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. *arXiv* **2019**, arXiv:1904.05873.
78. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
79. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018.
80. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
81. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
82. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
83. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
84. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010.
85. Chen, W.Y.; Liu, Y.C.; Kira, Z.; Wang, Y.C.; Huang, J.B. A Closer Look at Few-shot Classification. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
86. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS, San Jose, CA, USA, 3–5 November 2010.
87. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
88. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
89. Li, X.; Sun, Q.; Liu, Y.; Zhou, Q.; Zheng, S.; Chua, T.-S.; Schiele, B. Learning to self-train for semi-supervised few-shot classification. *Adv. Neural Inf. Proc. Syst.* **2019**, *32*, 10276–10286.
90. Ma, T.; Zhang, A. AffinityNet: Semi-supervised few-shot learning for disease type prediction. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 1069–1076.