remote sensing

MDPI

Article

TAE-Net: Task-Adaptive Embedding Network for Few-Shot Remote Sensing Scene Classification

Wendong Huang [1], Zhengwu Yuan [1], Aixia Yang [2,*], Chan Tang [1] and Xiaobo Luo [1]

[1] Chongqing Engineering Research Center for Spatial Big Data Intelligent Technology, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; s190201036@stu.cqupt.edu.cn (W.H.); yuanzw@cqupt.edu.cn (Z.Y.); s200231016@stu.cqupt.edu.cn (C.T.); luoxb@cqupt.edu.cn (X.L.)

[2] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

* Correspondence: yangax@radi.ac.cn; Tel.: +86-10-6480-6256

Abstract: Recently, approaches based on deep learning are quite prevalent in the area of remote sensing scene classification. Though significant success has been achieved, these approaches are still subject to an excess of parameters and extremely dependent on a large quantity of labeled data. In this study, few-shot learning is used for remote sensing scene classification tasks. The goal of few-shot learning is to recognize unseen scene categories given extremely limited labeled samples. For this purpose, a novel task-adaptive embedding network is proposed to facilitate few-shot scene classification of remote sensing images, referred to as TAE-Net. A feature encoder is first trained on the base set to learn embedding features of input images in the pre-training phase. Then in the meta-training phase, a new task-adaptive attention module is designed to yield the task-specific attention, which can adaptively select informative embedding features among the whole task. In the end, in the meta-testing phase, the query image derived from the novel set is predicted by the meta-trained model with limited support images. Extensive experiments are carried out on three public remote sensing scene datasets: UC Merced, WHU-RS19, and NWPU-RESISC45. The experimental results illustrate that our proposed TAE-Net achieves new state-of-the-art performance for few-shot remote sensing scene classification.

Keywords: scene classification; few-shot learning; meta-learning; pre-training; task-adaptive attention

Citation: Huang, W.; Yuan, Z.; Yang, A.; Tang, C.; Luo, X. TAE-Net: Task-Adaptive Embedding Network for Few-Shot Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 111. https://doi.org/10.3390/rs14010111

Academic Editor: Emanuele Frontoni

Received: 2 November 2021
Accepted: 23 December 2021
Published: 28 December 2021

1. Introduction

Remote sensing scene images taken by satellite consist of abundant semantic information of land-cover objects, which have been widely used in various fields, containing urban planning [1], traffic control [2], disaster detection [3], agricultural and environmental modeling [4,5], and other areas [6–9]. For remote sensing images, scene classification aims to divide unseen samples into corresponding scene classes according to the contained semantic information, which has recently attracted growing attention. By contrast, few-shot scene classification of remote sensing images is more inclined to divide remote sensing images into corresponding scene classes with only few labeled samples [10], which has quite vital significance in those areas where only few labeled samples are available. Few-shot scene classification has enormous potential in various fields, containing ecological monitoring, environmental monitoring, road detection, and so on, which can significantly decrease the burden of data collection and manual labeling.

In recent years, deep learning makes great success in traditional remote sensing image recognition [11–15], which typically acquires large amounts of labeled data in the training procedure. To achieve excellent performance, methods based on deep learning tend to contain a great deal of prior knowledge and sophisticated network structure [16,17]. Nevertheless, when training samples are insufficient, the models based on deep learning are susceptible to overfitting and deviation of feature representation [18]. Specifically, in

real remote sensing scenarios, collecting and labeling remote sensing data is pretty time-consuming and laboursome [19]. Moreover, massive computing resources are consumed in the training procedure of the model due to millions of learnable parameters, which further increases the burden on researchers.

Inspired by the way humans classify an object into one of several novel categories, few-shot learning is presented to address a series of few-shot tasks, which aims to make models quickly adapt unseen classes with only limited labeled samples [10]. In general, there are three lines of thinking to resolve the issue of few-shot learning: metric learning [20], meta-learning [21], and transfer learning [22]. In terms of metric learning, the similarity between the query sample and support sample is calculated by a distance function (e.g., Euclidean distance or cosine similarity) and then is compared to judge whether they are from the same class, where the higher the similarity between samples, the greater the probability of deriving from the same class. Recently, various approaches based on metric learning are proposed [23–27], where diverse similarity metrics are employed to address problems of few-shot classification. The goal of meta-learning is to learn how to learn [28], which plays a role in guiding the model to learn how quickly adapt to unseen tasks. For example, Zhai et al. [29] develop a scene classification model on the basis of meta-learning for remote sensing images, called LLSR, which aims to rapidly recognize unseen categories with few labeled samples. Hence, few-shot learning is generally regarded as a special case of meta-learning. As for transfer learning, the line of thinking is to train a model on similar tasks, typically known as pre-training, and then apply the pre-trained model with massive prior knowledge into new tasks.

In view of big scales and unique attributes of remote sensing images, some well-designed neural networks [30] are employed to encode images to acquire deep features with abundant semantic information in earlier work, such as AlexNet, GoogLeNet, VGG16, and so on. Nevertheless, when labeled samples are insufficient, the trained deep neural network typically suffers from overfitting, which is particularly obvious in the few-shot scenarios. Recently, a majority of proposed approaches for few-shot learning have focused more on improving feature representation. In RS-MetaNet, Li et al. [31] propose a training scheme based on meta tasks for few-shot classification of remote sensing images, which can prompt the model to learn how to capture a task-level distribution. In DLA-MatchNet, Li et al. [32] introduce spatial attention and channel attention into the feature extractor to learn robust feature representations, and a learnable matcher is developed to adaptively calculate similarity scores between the samples. In addition, Jiang et al. [33] present a multi-scale metric learning method based on the feature pyramid construction, called MSML, which aims to incorporate multi-scale features to improve feature representation of samples. These methods have been verified to improve the scene classification performance of remote sensing images, but in the case of extremely few available samples for each category, the improvement is not obvious.

The methods mentioned above focus more on employing image-level or class-level features for few-shot image classification, which does not fully excavate discriminative information from the perspective of the entire task. In terms of problems of scene classification, the model may not properly distinguish the correct category of the image when the feature extractor can not capture discriminative features [34]. In particular, in the tasks of few-shot scene classification, there are plenty of similar areas between the images from different categories [35], which further increases the difficulty of few-shot scene classification. Figure 1 presents the scene classification task in the 6-way 1-shot scenario. It is observed from Figure 1 that images in the first row have analogical texture features, and images in the second row have nearly approximate backgrounds, which shows that there are many similar feature areas between various categories. That is to say, the shared semantic features among all categories are not really critical for classifying an unseen sample. Furthermore, a remote sensing image may contain multiple different scene objects, which also has an enormous influence on the scene classification performance of the model.

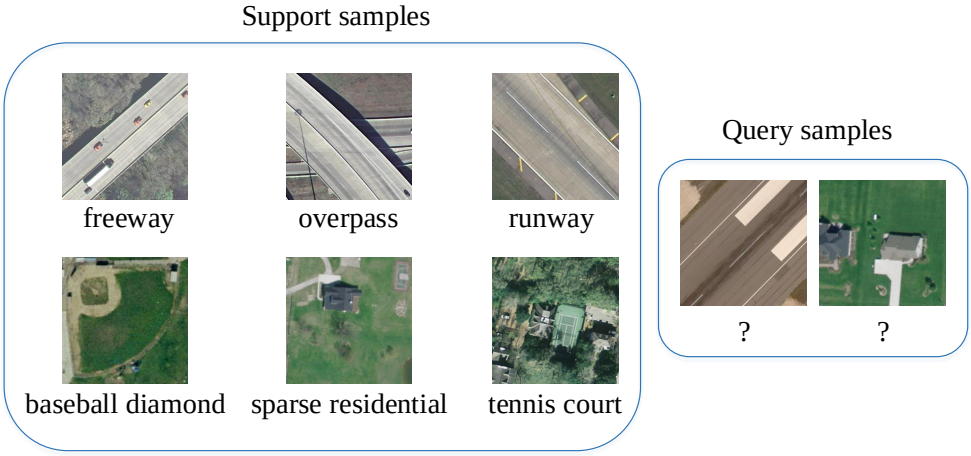


Figure 1. Illustration of few-shot remote sensing scene classification with only one labeled image.

To address the aforementioned issues, a task-adaptive embedding network based on meta-learning is proposed for few-shot remote sensing scene classification, which targets to learn task-relevant semantic features to enhance the discrimination between different scenes in few-shot tasks. In our proposed architecture, a pre-training scheme is employed to learn an effective feature encoder, which aims to make models learn generic feature representations for various scenes. Then, in meta-training stage, a task-adaptive attention module is employed to explore discriminative semantic features from the perspective of the whole task, which enhances weights of salient semantic features and reduces weights of common semantic features shared by diverse scenes. Finally, in the meta-testing stage, new scene categories can be predicted by the trained model. Experimental results on three challenging scene datasets verify state-of-the-art performance of our TAE-Net in few-shot settings.

The primary contributions of this paper are overall summarized as follows:

- A task-adaptive embedding network based on meta-learning, called TAE-Net, is proposed to enhance the generalization performance of the model for unseen remote sensing images in few-shot settings. The proposed TAE-Net can learn generic feature representations by combining pre-training with meta-training, which can allow models to quickly adapt to new categories with extremely few labeled samples.
- A task-adaptive attention module is developed to capture task-specific information, which can remove the effect of task-irrelevant noises. The proposed task-adaptive attention module aims to adaptively and dynamically pick out discriminative semantic features for diverse tasks.
- Comprehensive experiments on three public remote sensing scene classification datasets verify effectiveness of our proposed model, which exceeds existing state-of-the-art few-shot scene classification approaches and acquires new state-of-the-art performance of few-shot scene classification.

2. Related Work

In this section, the work related to our research is reviewed from two perspectives: remote sensing scene classification and few-shot learning.

2.1. Remote Sensing Scene Classification

In the past decades, remote sensing scene classification has been sufficiently researched due to its broad application prospects, which aims to divide images into corresponding scene categories. In recent years, approaches based on deep learning have made considerable progress in the area of remote sensing; particularly, plenty of approaches based on the convolutional neural network (CNN) are proposed [36–38]. To solve the issue of intra-class difference and inter-class similarity in remote sensing images, Cheng et al. [39] design

discriminative CNN (D-CNN). Zhang et al. [40] design a CNN-CapsNet for remote sensing scene classification, which consists of CNN for extracting features and CapsNet for classifying features. Wang et al. [35] design an attention recurrent CNN in an end-to-end manner for remote sensing scene classification, which can capture high-level semantic features and remove irrelevant information. Sun et al. [41] develop a gated bidirectional model based on CNN to implement remote sensing scene classification, which can incorporate multilayer convolutional features and eliminate the distraction information. Pires de Lima et al. [42] study the performance of CNN combined with transfer learning on remote sensing scene classification tasks, which confirms that transfer learning from natural images to remote sensing images is effective. To constrain the distribution of the training data belonging to the same category, Xie et al. [43] design a scene classification network with intra-class constraint; besides that, label augmentation is employed to label each augmented sample with a joint label. Considering the considerable complexity of current CNN, Shi et al. [44] construct a lightweight network that combines the attention mechanism and multi-branch feature fusion strategy for remote sensing scene classification.

2.2. Few-Shot Learning

In recent years, methods based on deep learning have made considerable progress in various areas, especially in fields where a large amount of computing resources and labeled data can be obtained. Nevertheless, methods based on deep learning are susceptible to insufficient data. To resolve the issue, few-shot learning is proposed, which has achieved outstanding success. The goal of few-shot learning is to learn to recognize new categories from only limited labeled data, which is typically regarded as a special meta-learning. Subsequently, some representative few-shot learning methods are introduced, which are divided into two main branches: metric-based approaches and optimization-based approaches.

Metric-based approaches: This line of work targets to solve few-shot image classification problems through learning to compare. The line of thinking aims to learn a feature encoder that transforms input images into embedding representations suitable for comparing; when encoded in the feature space, the similarity between the support sample and the query sample can be calculated explicitly through a common pairwise metric, such as cosine similarity or Euclidean distance. Siamese network [23] is proposed to learn generic embedding features, which is employed to achieve the binary classification task. By an attention mechanism, matching network [24] transforms support samples into embedding space, where the query sample is classified through a nearest-neighbor classifier using cosine similarity. Likewise, prototypical network [25] also learns a metric rule to conduct few-shot classification over embeddings, which calculates Euclidean distance between the category-mean embedding and the query embedding. Instead of common metrics, relation network [26] presents a parameterized relation module as a learnable metric. Inspired by prototypical network, Oreshkin et al. [45] introduce a metric-based framework, called TADAM, that integrates three helpful improvements for few-shot learning, that is, task conditioning, metric scale, and auxiliary task co-training. By adding unlabeled data into each few-shot task, Ren et al. [46] carry out few-shot classification of prototypical network in the semi-supervised setting. Three schemes are tried to improve the prototype representation of each category. MetaOptNet [21] indicates that discriminative linear classifiers, such as support vector machine, may be preferable to nearest neighbor classifiers (e.g., cosine similarity or Euclidean distance) in few-shot settings. Compared with nearest neighbor classifiers, linear classifiers can find more appropriate decision boundaries by negative samples.

Optimization-based approaches: Another family of method is typically known as learning to learn (i.e., meta-learning), which aims to learn generic parameter initialization and then make the model quickly adapt to new tasks by a few gradient steps. Finn et al. [47] propose a model-agnostic architecture based on meta-learning, called MAML, which aims at learning a suitable parameter initialization that can be applied to any neural network. In other words, that network can fast adapt to any unseen few-shot task by only a few

parameter updating steps. In addition, considering the complexity of the algorithm, the authors also introduce a first-order variant of MAML, which can speed up the model optimization by removing second-order derivatives. Likewise, Reptile [48] is also first-order variant of MAML, which employs a Taylor series expansion to perform first-order gradient optimization. A lot of variants [49–52] of MAML are based on an analogous idea that, given appropriate parameter initialization, the model can quickly adapt to the novel task with a few gradient steps. However, these methods suffer from a key challenge that internal optimization contains as massive parameters as external optimization. Additionally, a critical controversy is whether only an initialization condition guarantees quick adaption for various few-shot tasks. Further, whether the upper limit of model performance is affected by this initialization condition.

Recently, several approaches based on pre-training (i.e., transfer learning) have achieved competitive performance [53–55] as methods based on meta-learning make noteworthy progress in few-shot image classification. Additionally, some methods based the graph neural network are also developed for few-shot learning [56,57]. Our study is more inclined to the first kind of work, that is, employing pre-training to extract good embedding features and learning an appropriate similarity metric by encoding task-specific information.

3. Proposed Method

3.1. Overall Architecture

In this work, a task-adaptive embedding network is proposed to tackle scene classification tasks in few-shot settings, which is depicted in Figure 2. The overall architecture is composed of pre-training, meta-training, and meta-testing. A feature encoder is first pre-trained on the base set D_{base} to map inputs to embedding space suitable for comparison. To be specific, a common neural network model is trained on all base classes by minimizing generalization error. To obtain the trained feature encoder, the fully connected (FC) layer is removed. Next, a meta-learning model H is trained across a cluster of episodes in the meta-training phase. Specifically, different from previous works [54,55], f_φ is not frozen to further fine-tune; instead, it is employed to initialize parameters in the meta-learning model and is optimized by minimizing the cross-entropy loss. For each episode, a relation matrix R is calculated to obtain the embedding relation between the support image and the query image. Then, the task-adaptive attention module g_θ yields the task attention matrix A, which can adaptively pick out the discriminative embedding features for a certain query embedding feature within the support set, like visual recognition of humans. It's worth noting that, instead of independent entity, the task attention centers on the relations between the embedding features. Subsequently, the task attention matrix A is incorporated into the relation matrix R by an element-wise multiplication to eradicate noises, such as the relation formed by the similar embedding features within a task, and strengthen the discriminative embedding features. The predicted scores can be then obtained from the incorporated relation matrix using the sum operation. In the meta-testing phase, the meta-learning model H is evaluated across a cluster of episodes, which are sampled at random from the novel set D_{novel}.

3.2. Problem Formulation

In the few-shot settings, remote sensing scene classification can be referred to as a set of N-way K-shot M-query tasks, which indicates that only K labeled samples are used to recognize M unseen samples from N scene classes. A N-way K-shot M-query task $\mathcal{T}$ is also known as an episode. Therein, M samples without labels and K samples with labels are derived from each class, and accordingly, a support set consists of $N \times K$ labeled samples and a query set consists of $N \times M$ samples. To be specific, when M is set to 1, the quantity of all samples in the query set is equivalent to the total quantity of classes. Figure 3 presents the illustration of 5-way 1-shot classification tasks. The dataset for few-shot scene classification is split into three disjoint subsets, that is, base set D_{base}, validation set D_{val},

and novel set D_{novel}. The overall process contains the meta-training phase, meta-validation phase, and meta-testing phase. The specific details are described below.

Different with standard supervised learning, there are extremely few labeled samples available in few-shot learning, which makes it difficult for the model to learn enough prior knowledge. Therefore, intra-class relationship learning and knowledge transfer become quite vital for few-shot scene classification of remote sensing images. In the meta-training phase, the support set $S = (x_i, y_i)(i = 1, 2, 3, \ldots, N \times K)$ and query set $Q = (x_j, y_j)(j = 1, 2, 3, \ldots, N \times M)$ are sampled from D_{base}, where x_i represents the i-th sample and y_i represents its ground truth label. Furthermore, the parameters in the model are iteratively updated through back propagation. After massive iterations, a well-trained feature encoder can be learned.

The validation phase aims to adjust the hyper-parameters in the model. In the meta-validation phase, D_{val} is also divided into the support set S and the query set Q, where the support set is employed to predict labels of query samples. In addition, only forward propagation is performed in the meta-validation phase, which indicates that the parameters in the model are not updated through back propagation.

In the few-shot scene classification, the categories used in the meta-testing phase are new, which indicates that classes in D_{novel} are distinct from those of D_{base} and D_{val}. Query samples in D_{novel} are predicted by a well-trained model given the support set that only contains N classes and K labeled samples per class. The class corresponding to the largest predicted probability is determined as the label of the query sample.

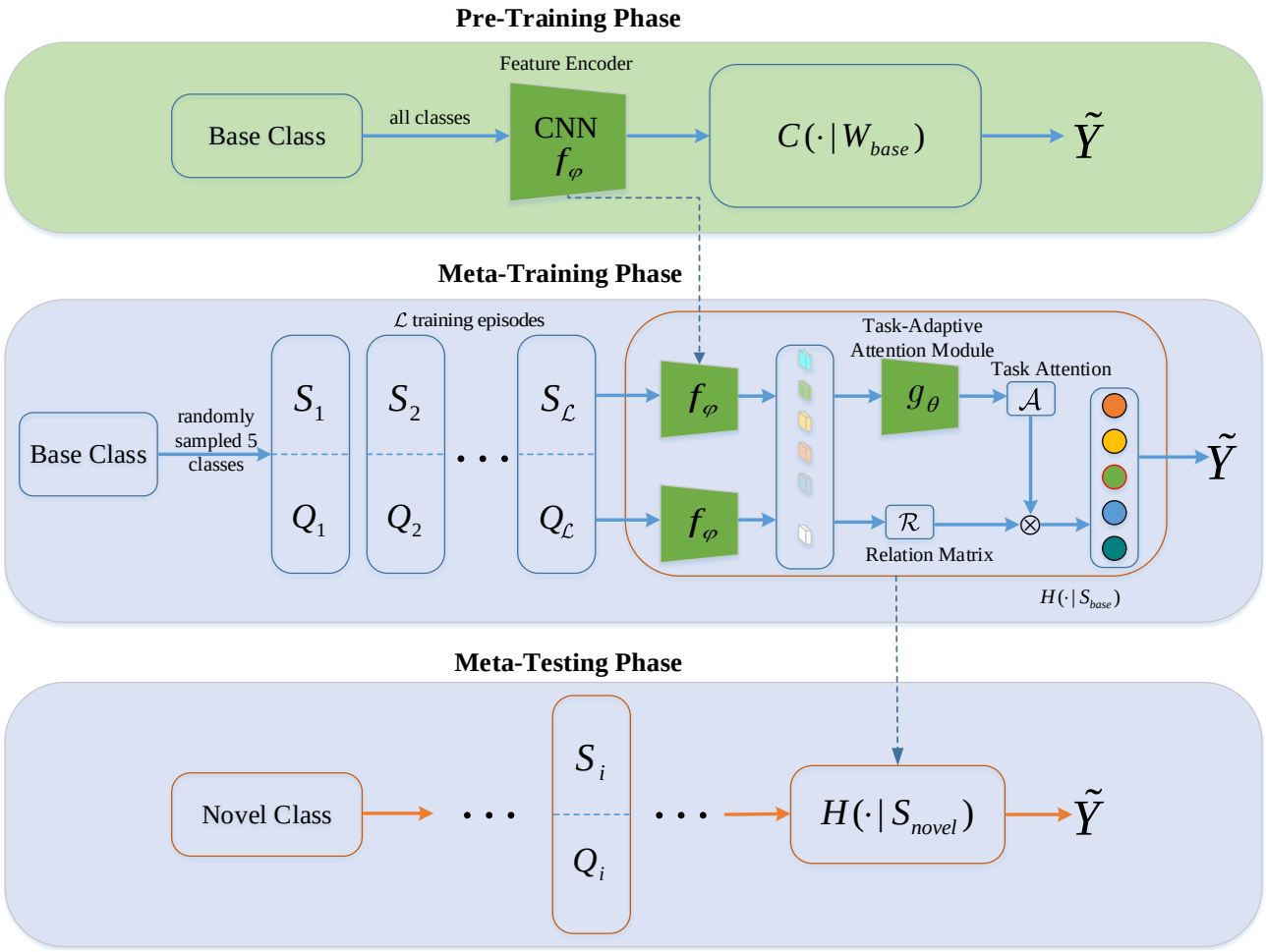


Figure 2. The overall architecture and details of the task-adaptive embedding network (shortened to TAE-Net).

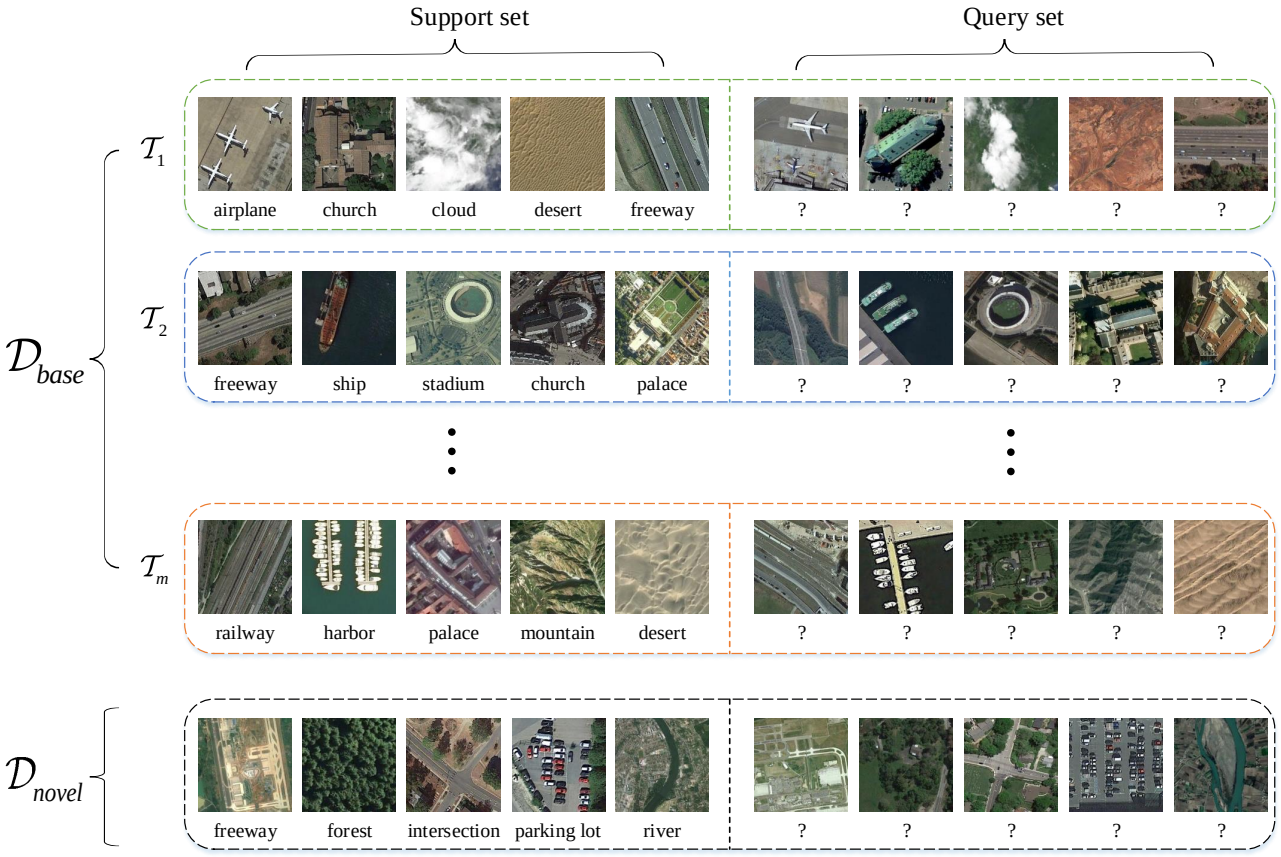


Figure 3. Illustration of data partition of remote sensing scene classification in the 5-way 1-shot setting.

3.3. Pre-Training of Feature Encoder

In the pre-training stage, a feature encoder f_φ with learnable parameters φ is trained on the base set D_{base}, which transforms the input sample into the embedding feature with the shape of $5 \times 5 \times 64$. For a fair comparison, a neural network with 4 convolutional blocks is employed to train a classification model on all base categories. Then, to obtain well-trained f_φ, the fully connected layer and the softmax layer are removed in the neural network. The above process is described in detail as follows. Before starting training, all input samples from D_{base} are adjusted to the size of 84×84. The entire setting of the pre-training model we adopt, presented in Figure 4, is formed from an encoder and a classifier. The encoder contains 4 convolutional blocks (Conv-4), each of which contains 3×3 convolution with 64 kernels, followed by BatchNorm with a momentum value of 0.1, Leaky ReLU activation function, and 2×2 max-pooling. Furthermore, the classifier is made up of a fully connected layer and a softmax layer. To optimize the overall pre-training model, a common cross-entropy loss function is selected as the loss metric in this work. After pre-training, the classifier in the pre-training model is removed to get the encoder for extracting features. Therefore, by inputting a sample into the pre-training model, a embedding feature with the shape of $5 \times 5 \times 64$ can be yielded by the encoder.

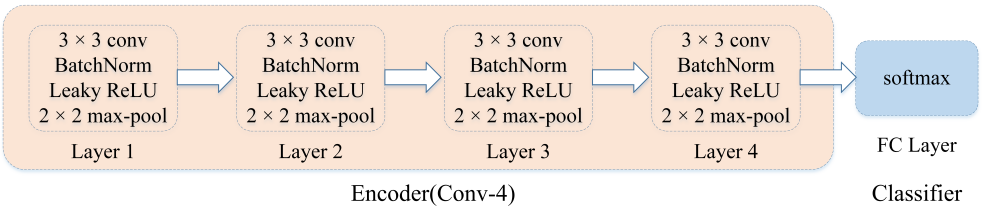


Figure 4. Structure and parameters of the pre-training model.

3.4. Meta-Training with Task-Adaptive Attention Module

The goal of meta-learning is to enhance generalization performance through learning meta-konwledge from multiple tasks, also known as episodes, which is a quite popular method for solving few-shot tasks. In the N-way K-shot setting, a meta-learning model $H(\cdot|S)$ is trained by minimizing the N-way loss. To this end, a set of episodes are sampled at random from the base set. Each episode contains K labeled samples for each class, which means that there are $N \times K$ support samples in total for training and $N \times M$ query samples for testing. Even though each episode only contains a few support samples for training, the model $H(\cdot|S)$ shares the same parameters over multiple episodes. As a result, training $H(\cdot|S)$ across plenty of episodes contributes to decreasing the model demand for data. In the meta-training procedure, the validation set D_{val} is employed to select hyperparameters of $H(\cdot|S)$. Figure 2 presents corresponding meta-training process.

For an episode, a sample x from $S \bigcup Q$ first is inputted into the feature encoder f_φ to attain an embedding representation $f_\varphi(x) \in \mathbb{R}^{W \times H \times C}$. Generally, each input sample is represented as WH C-dimensional embedding features. Thus, the entire support set can be denoted as $NKWH$ C-dimensional support embedding features, that is, $Z^S = f_\varphi(S) \in \mathbb{R}^{NKWH \times C}$ and each query image can be denoted as WH C-dimensional query embedding features, that is, $Z^q = f_\varphi(q) \in \mathbb{R}^{WH \times C}$. Next, the relation matrix R of these embedding features can be formulated as follows:

$$R_{i,j} = s(Z_i^q, Z_j^S) \tag{1}$$

where $i \in \{1, 2, 3, \ldots, WH\}$, $j \in \{1, 2, 3, \ldots, NKWH\}$ and $s(\cdot, \cdot)$ is a similarity function. In this paper, Gaussian similarity function is selected to calculate similarity. Different from prior work that is based on class-level [58] or image-level [26] relation, our proposed approach aims to construct task-level relation while sustaining discriminative relations.

Moreover, a convolution block f_v consisting of two 1×1 convolution layers is applied to the embedding features extracted by the feature encoder. Then another relation matrix $R^{'}$ is calculated by the following operation:

$$R_{i,j}^{'} = s(f_v(Z_i^q), f_v(Z_j^S)) \tag{2}$$

where $i \in \{1, 2, 3, \ldots, WH\}, j \in \{1, 2, 3, \ldots, NKWH\}$. For relation matrix $R^{'}$, each row denotes the adaptive subspace relation of each space position in the query image to all space positions of all support images. Furthermore, to eradicate the noises (e.g., the irrelevant relations), a threshold λ is applied to the relation matrix $R^{'}$, which can generate task attention A by the following operation:

$$A_{i,j} = \frac{M(R_{i,j}^{'})}{\sum_j M(R_{i,j}^{'})} \tag{3}$$

$$M(x) = \begin{cases} 1, & if\ x > \lambda \\ 0, & otherwise. \end{cases} \tag{4}$$

As shown in Equation (3), the universal embedding features tend to exist in multiple categories in the whole task, which will greatly reduce the attention they receive. Hence, their corresponding attention value will also relatively smaller. Additionally, considering that the impact of each irrelevant noise is small, but the total number is quite large, which still has a considerable effect on the distribution of task attention. To this end, Equation (4) is adopted to filter the task attention, which can select informative relations and eliminate negligible relations.

Considering that the threshold λ is fixed in Equation (4), it can not handle different query features flexibly. For this purpose, a task encoder with the adaptive threshold $\lambda^{'}$ is proposed, as shown in Figure 5. Unlike Equation (4), a transformer, consisting of three fully connected layers, is employed to adaptively yield the threshold $\lambda^{'}$ for each C-dimensional

embedding feature of the query image. Meanwhile, considering that Equation (4) is indifferentiable, a variant $M^{'}$ of the sigmoid function is adopted to approximate it:

$$M^{'}(x) = 1/(1 + exp^{\beta(\lambda^{'} - x)}) \tag{5}$$

where $\lambda^{'}$ is the adaptive threshold of x, and x represents one of the elements in A. Conceptually, when β takes a large enough value, $M^{'}$ is equivalent to M.

Subsequently, a weighted relation matrix $A \otimes R$ can be calculated by a Hadamard product between A and R. Then, the predicted score belonging to the c-th category is attained by gathering the weighted relation between the query sample q and the c-th category:

$$Score_c = \frac{1}{WH} \sum_{i=1}^{WH} \sum_{j=r_1^c}^{r_{KWH}^c} (A \otimes R)_{i,j} \tag{6}$$

where r_t^c denotes that the t-th relation of KWH relations belonging to the c-th category in the NKWH relations of all support samples. Hence, the final predicted probability P of the query image q can be formulated as follows:

$$P(y = c|q) = \frac{exp(Score_c)}{\sum_{c^{'}=1}^{N} exp(Score_{c^{'}})} \tag{7}$$

where N represents the total quantity of categories in the N-way K-shot setting.

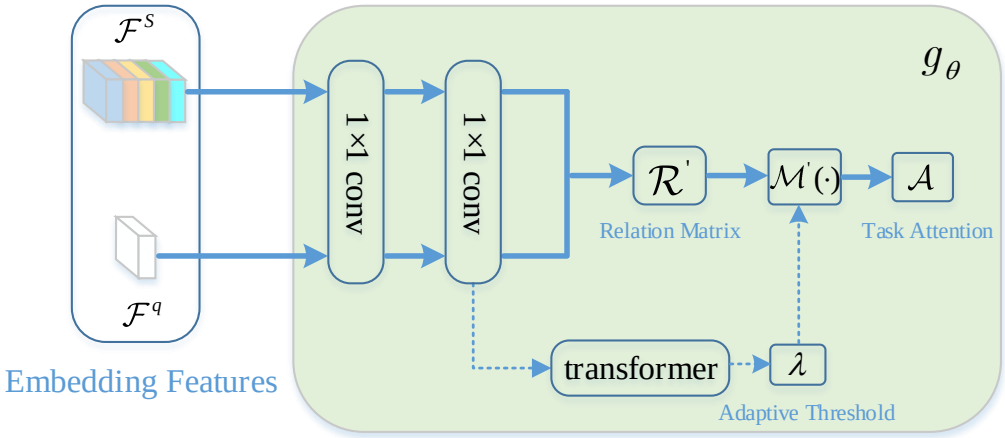


Figure 5. The schematic of task-adaptive attention module.

3.5. Meta-Testing

At the end of meta-training, a novel set D_{novel} is typically utilized to estimate the generalization performance of the trained model $H(\cdot|S_{base})$. It is worth noting that all classes in D_{novel} have never appeared in the meta-training stage. In the meta-testing stage, a set of new episodes are stochastically sampled from D_{novel}, also known as meta-testing set $D_{\mathcal{T}}^{test} = \{\mathcal{T}_i\}_{i=1}^{I}$. Here, $\mathcal{T}$ consists of S_{novel} and Q_{novel}. Then, given the novel support set S_{novel}, categories of samples from Q_{novel} can be predicted by the trained model $H(\cdot|S_{base})$.

4. Results and Discussions

In this section, dataset description and experimental settings are first presented. Then, several state-of-the-art methods for few-shot learning are compared with our proposed approach and the corresponding experimental results are presented below. Moreover, we also conduct a series of ablation experiments to investigate the impact of different components on model performance, containing the pre-training strategy, task-adaptive attention mechanism, and the number of shots.

4.1. Dataset Description

The UC Merced land-use dataset [59] consists of 21 categories with a total of 2100 scene images, each of which corresponds to 100 land-use images containing 256 × 256 pixels. Furthermore, all scene images are of RGB color space. The scene categories contain agricultural, airplane, baseball diamond, beach, buildings, chaparral, and other scenes. This dataset was published by UC Merced Computer Vision Laboratory in 2010, including various urban area imageries from the United States Geological Survey. In this dataset, 10 categories are deemed as the base set, 5 categories are considered as the validation set and the remaining 6 categories are considered as the novel set. In the experiments, the shapes of all images are adjusted to 84 × 84 to fit our proposed feature encoder for feature extraction.

The WHU-RS19 dataset [60] is also composed of remote sensing scenes, which was published by Wuhan University. It consists of 19 scene categories in total, and the quantity of samples for each category is more than or equal to 50 images. A total of 1005 images are contained in this dataset. The entire dataset is divided into three subsets, namely, a base set containing 9 classes, a validation set containing 5 classes, and a novel set containing 5 classes. In order to adapt to our proposed feature encoder, the pixel sizes of all scene images are adjusted to 84 × 84.

The NWPU-RESISC45 is a quite popular scene dataset in the domain of remote sensing, and was published by Cheng et al. [61] in 2017. It contains 45 scene classses and each class has 700 scene images, which consists of a total of 31,500 images, as shown in Figure 6. Each scene image is of 256 × 256 pixels. The scene classes include airplane, beach, circular farmland, dense residential, parking lot, and other scenes. These scene data are gathered by experts from Google Earth, and the spatial resolution of each pixel varies between close to 30 and 0.2. The entire dataset is split into 3 subsets, namely, a base set containing 25 classes, a validation set containing 10 classes, and a novel set containing 10 classes. In the experiments, the pixel sizes of all scene images are adjusted to 84 × 84 , which is to adapt to the designed feature encoder. The segmentation details of the three scene datasets mentioned above are presented in Table 1.

Table 1. The segmentation details of UC Merced dataset, WHU-RS19 dataset, and NWPU-RESISC45 dataset.

Datasets	Base	Validation	Novel
UC Merced	Agricultural; Baseball diamond; Buildings; Parking lot; Harbor; Medium residential; Dense residential; Chaparral; Freeway; Overpass;	Airplane; Forest; Runway; Intersection; Storage tanks;	Beach; River; Golf course; Mobile home park; Sparse residential; Tennis court;
WHU-RS19	Airport; Bridge; Football field; Desert; Mountain; Industrial; Port; Residential; Parking;	Beach; Forest; Farmland; Park; Railway station;	Meadow; Pond; River; Viaduct; Commercial;

Table 1. *Cont.*

Datasets	Base	Validation	Novel
NWPU-RESISC45	Airplane;Church; Baseball diamond; Bridge;Beach; Cloud;Freeway; Desert;Island; Chaparral; Harbor;Lake; Meadow;Mountain; Palace;Ship; Railway;Stadium; Wetland; Golf course; Mobile home park; Sparse residential; Sea ice; Roundabout; Rectangular farmland;	Commercial area; Overpass; Industrial area; Railway station; Snowberg; Runway; Storage tank; Terrace; Thermal power station; Tennis court;	Airport; Dense residential; Basketball court; Circular farmland; Intersection; Forest; Ground track field; Parking lot; Medium residential; River;

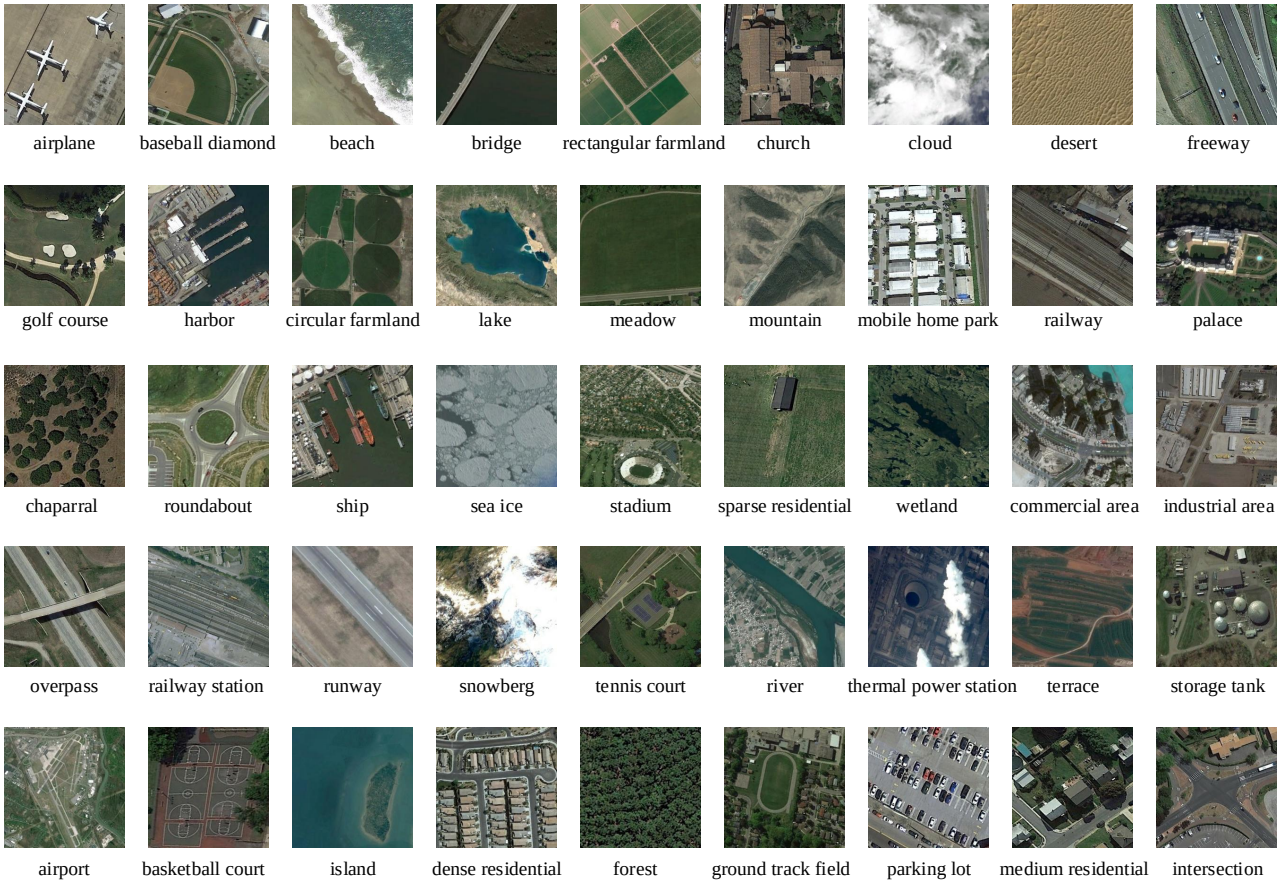


Figure 6. Scene images (with 84 × 84 pixels) derived from 45 categories in the NWPU-RESISC45 dataset [61]. Airport, basketball court, circular farmland, dense residential, forest, ground track field, intersection, medium residential, Parking lot, and river are selected as the novel set for performance evaluation.

4.2. Experimental Settings

In this section, the network architecture and hyperparameter setup are described in detail. To make a fair comparison with other work, the same backbone as other methods [24–26], a 4-layer convolutional network (Conv-4), is employed as our encoder to extract embedding features, which consists of 4 convolution modules, as depicted in Figure 4. For each convolution module, there is a convolutional layer with 64 kernels whose size is 3×3, a batch normalization operation, a Leaky ReLU nonlinearity activation function, and a 2×2 max-pooling layer. The padding and stride sizes are both set to 1 in the convolutional layer while for the 2×2 max-pooling, the padding and stride sizes are separately set to 0 and 2. For each batch normalization operation, the momentum factor is set to 0.1 and the value of epsilon ϵ is set to 0.00001. The learning rate η is initialized to 0.001. In addition, all comparison models adopt the cross-entropy loss to evaluate generalization performance. The models are optimized by Adam Optimizer, where $\beta 1$ and $\beta 2$ are separately set to 0.9 and 0.999 to adjust exponential decay rates. The pixel size of each image is adjusted to 84×84. Classification accuracy is employed to evaluate the performance of all models, which is formalized as:

$$\text{Accuracy} = \frac{1}{\mathcal{L}} \sum_{t=1}^{\mathcal{L}} \frac{u^t}{V} \tag{8}$$

where $\mathcal{L}$ denotes the total quantity of tasks, u^t denotes the quantity of samples that are correctly classified in the t-th task, and V denotes the total quantity of samples in the task.

For traditional deep learning, an epoch indicates that the whole dataset passes the neural network by the forward propagation once. For few-shot learning, each episode randomly sampled from the dataset. Though there are only a few labeled images per episode, when the total quantity of episodes is more enough, all samples have been probably sampled in the whole dataset. In our work, an epoch consists of 5000 episodes.

Pre-training is done with D_{base} containing 10, 9, and 25 categories for UC Merced, WHU-RS19, and NWPU-RESISC45 experiments, separately. For the meta-learning (i.e., meta-training and meta-testing) stage, 5 novel classes are randomly chosen from D_{train} per episode, and 1 or 5 samples (i.e., shots) from 5 randomly chosen novel classes makes up the support set. Following the protocol of few-shot learning [25,62], the query set contains 15 samples per class in each episode. The PyTorch framework is adopted to implement our proposed TAE-Net, which is run with four NVIDIA Tesla V100 GPUs (32G $\times$ 4).

4.3. Experimental Results

4.3.1. Experimental Results on UC Merced Dataset

In few-shot scenarios, the classification results on the UC Merced dataset are presented in Table 2, where classification accuracies are acquired by averaging the results of 600 episodes randomly sampled on the novel set with 95% confidence interval. In the experiments, nine other few-shot scene classification approaches of remote sensing images are employed to make a comparison, in which average accuracies with 95% confidence interval are reported. For Table 2, it is observed that our proposed TAE-Net performs best with accuracies of 60.21% and 77.44% in the scenarios of 5-way 1-shot and 5-way 5-shot , outperforming the accuracies of RS-MetaNet with 2.98% and 1.36% improvements, separately. Compared to DLA-MatchNet, the proposed TAE-Net has 6.45% and 14.43% improvements in the scenarios of 5-way 1-shot and 5-way 5-shot, separately. In addition, our proposed TAE-Net exceeds TPN with 6.85% improvement in the 5-way 1-shot scenario and 9.21% improvement in the 5-way 5-shot scenario, separately. The proposed TAE-Net achieves higher classification accuracies than existing state-of-the-art methods. In consequence, it is illustrated that our proposed TAE-Net can take full advantage of a small amount of information from limited samples, which promotes the performance of scene classification in the cases of limited labeled samples.

Table 2. Classification accuracies and standard deviations (%) of 5-way 1-shot and 5-way 5-shot on the UC Merced dataset. The best results per scenario are marked in bold.

Method	5-Way 1-Shot	5-Way 5-Shot
MatchingNet [24]	46.16 ± 0.71	66.73 ± 0.56
Prototypical Network [25]	52.62 ± 0.70	65.93 ± 0.57
MAML [47]	43.65 ± 0.68	58.43 ± 0.64
Meta-SGD [52]	50.52 ± 2.61	60.82 ± 2.00
Relation Network [26]	48.89 ± 0.73	64.10 ± 0.54
TPN [56]	53.36 ± 0.77	68.23 ± 0.52
LLSR [29]	39.47	57.40
RS-MetaNet [31]	57.23 ± 0.56	76.08 ± 0.28
DLA-MatchNet [32]	53.76 ± 0.60	63.01 ± 0.51
TAE-Net (ours)	**60.21 ± 0.72**	**77.44 ± 0.51**

The proposed TAE-Net is capable of learning the task-specific feature representation by task encoder, which eliminates the deviation of embedding features posed by limited labeled samples. Additionally, pre-training scheme is developed for feature embedding, which can provide the feature encoder of the model with better initialization parameters.

4.3.2. Experimental Results on WHU-RS19 Dataset

Comparative experiments on the WHU-RS19 dataset are performed in few-shot cases, and the corresponding experimental results are presented in Table 3. All experimental results are the average of 600 episodes with 95% confidence interval. It can be seen that our proposed few-shot classification approach achieves the accuracies of 73.67% and 88.95% in the 5-way 1-shot scenario and 5-way 5-shot scenario, separately. From Table 3, it is observed that the proposed TAE-Net achieves superior performance over Prototypical Network, with 2.79% improvement in the 5-way 1-shot scenario and 3.33% improvement in the 5-way 5-shot scenario.

Compared with experimental results in the 5-way 1-shot scenario, the performance of scene classification is better than that in the 5-way 1-shot scenario, which illustrates that prior knowledge is more significant. In our proposed TAE-Net, task-adaptive attention module integrates task-specific information into the relationship matrix, which can seek out more informative prior knowledge from extremely limited labeled samples. Hence, the proposed TAE-Net can significantly improve the classification performance of remote sensing scene images in few-shot cases.

Table 3. Classification accuracies and standard deviations (%) of 5-way 1-shot and 5-way 5-shot on the WHU-RS19 dataset. The best results per scenario are marked in bold.

Method	5-Way 1-Shot	5-Way 5-Shot
MatchingNet [24]	60.60 ± 0.68	82.99 ± 0.40
Prototypical Network [25]	70.88 ± 0.65	85.62 ± 0.33
MAML [47]	46.72 ± 0.55	79.88 ± 0.41
Meta-SGD [52]	51.54 ± 2.31	61.74 ± 2.02
Relation Network [26]	60.54 ± 0.71	76.24 ± 0.34
TPN [56]	59.28 ± 0.72	71.20 ± 0.55
LLSR [29]	57.10	70.65
DLA-MatchNet [32]	68.27 ± 1.83	79.89 ± 0.33
TAE-Net (ours)	**73.67 ± 0.74**	**88.95 ± 0.53**

4.3.3. Experimental Results on NWPU-RESISC45 Dataset

Experimental results on the NWPU-RESISC45 dataset are presented in Table 4, where all evaluation results are the average accuracies on 600 episodes. In Table 4, it is distinctly observed that the proposed TAE-Net obtains best performance with classification accuracies of 69.13% and 82.37% in the 5-way 1-shot scenario and 5-way 5-shot scenario, respectively.

Compared with DLA- MatchNet, our proposed TAE-Net achieves more superior performance, with 0.33% and 0.74% improvements in the scenarios of 5-way 1-shot and 5-way 5-shot. Additionally, the proposed TAE-Net exceeds RS-MetaNet with 16.35% improvement in the 5-way 1-shot scenario and 10.88% improvement in the 5-way 5-shot scenario. Besides, the proposed TAE-Net also obtains greater improvements than other methods. The reason for achieving such significant improvements is that our proposed TAE-Net can reduce the interference between similar embedding features while focusing more on enhancing the discrimination between informative embedding features.

Table 4. Classification accuracies and standard deviations (%) of 5-way 1-shot and 5-way 5-shot on the NWPU-RESISC45 dataset. The best results per scenario are marked in bold.

Method	5-Way 1-Shot	5-Way 5-Shot
MatchingNet [24]	54.46 ± 0.77	67.87 ± 0.59
Prototypical Network [25]	50.82 ± 0.84	74.38 ± 0.59
MAML [47]	37.36 ± 0.69	45.94 ± 0.68
Meta-SGD [52]	60.63 ± 0.90	75.75 ± 0.65
Relation Network [26]	58.61 ± 0.83	78.63 ± 0.52
TPN [56]	66.51 ± 0.87	78.50 ± 0.56
LLSR [29]	51.43	72.90
RS-MetaNet [31]	52.78 ± 0.09	71.49 ± 0.81
DLA-MatchNet [32]	68.80 ± 0.70	81.63 ± 0.46
TAE-Net (ours)	**69.13 ± 0.83**	**82.37 ± 0.52**

4.4. Ablation Study

To further validate the effectiveness of our proposed TAE-Net on few-shot scene classification of remote sensing images, a sequence of ablation experiments are performed to analyze the role of each module in our framework and presented as follows.

4.4.1. Effect of Pre-Training Strategy

Before meta-training, pre-training is introduced to improve the feature representation ability of the model in few-shot settings, which can give the model a good initialization. To illustrate the effect of pre-training, experiments with and without pre-training are conducted, in which several cases containing different shots are designed for comparison. Experimental results on UC Merced dataset are depicted in Figure 7.

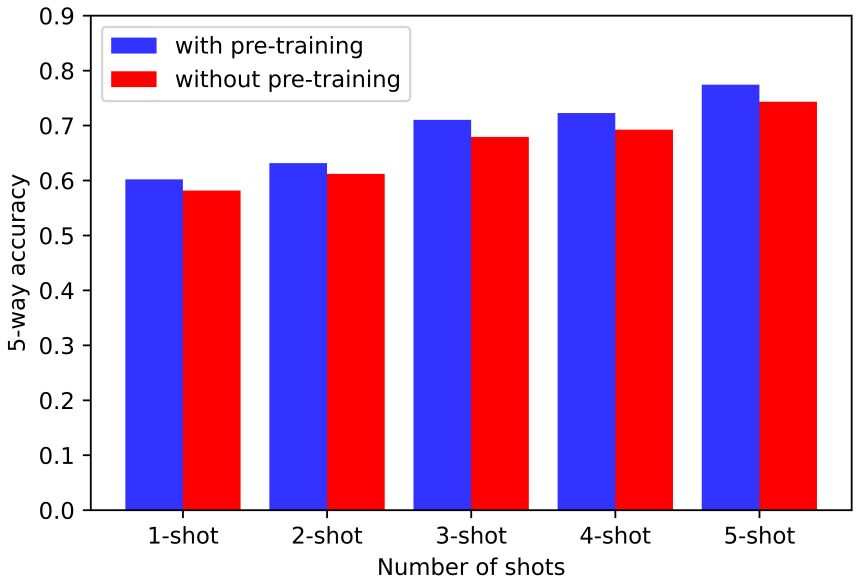


Figure 7. Classification accuracies with different number of shots on UC Merced dataset.

In Figure 7, it can be seen that the classification accuracy without pre-training is only 58.17% in the 5-way 1-shot scenario, which is markedly lower than our proposed TAE-Net by approximately 2%. Likewise, the classification accuracy without pre-training is 74.32% in the 5-way 5-shot scenario, which is obviously lower than our proposed TAE-Net by approximately 3%. Besides, the proposed method with pre-training is superior to that without pre-training in the settings of different shots by a large margin. It can be demonstrated that pre-training before meta-training helps to improve the performance of few-shot scene classification, which can yield a feature encoder with good initialization parameters.

4.4.2. Effect of Task-Adaptive Attention Module

In our proposed approach, a task-adaptive attention module is designed to enhance discriminative embedding features and weaken common embedding features shared by diverse scenes, which can strengthen the discrimination between embedding features from different categories and reduce the distraction of irrelevant information. To verify the effect of the task-adaptive attention module, an ablation experiment is conducted on UC Merced dataset. The results of the ablation experiment for the task-adaptive attention module is depicted in Figure 8. It is observed that when the task-adaptive attention module is ablated, the classification performance of the model notably degenerates in various numbers of shots settings. As the task-adaptive attention module is removed, the model is inclined to treat equally noncritical embedding features and discriminative embedding features, which makes it difficult for the model to pay more attention to informative features. Moreover, it is also noticed that the task-adaptive attention module has a growingly significant contribution to classification performance as the number shots increases. For example, in the case of 5-way 1-shot, the classification accuracy of the model improves by 10% when the task-adaptive attention module is adopted. By contrast, the classification accuracy of the model increases by nearly 13% in the case of 5-way 5-shot when the task-adaptive attention module is adopted. It is verified that our proposed task-adaptive attention module can significantly improve the classification performance of the model in few-shot settings, since it enhances the discriminability between embedding features belonging to diverse categories and reduces the impact of irrelevant information on classification.

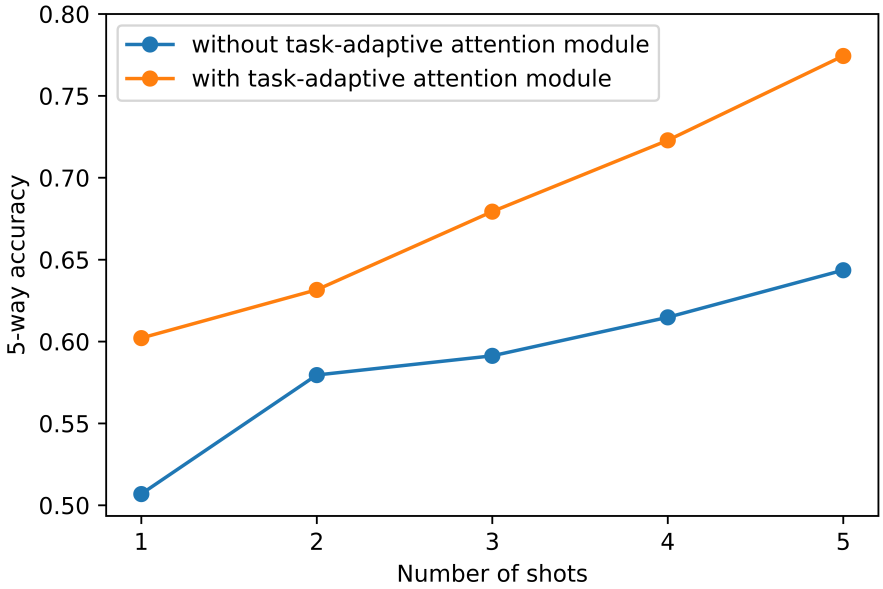
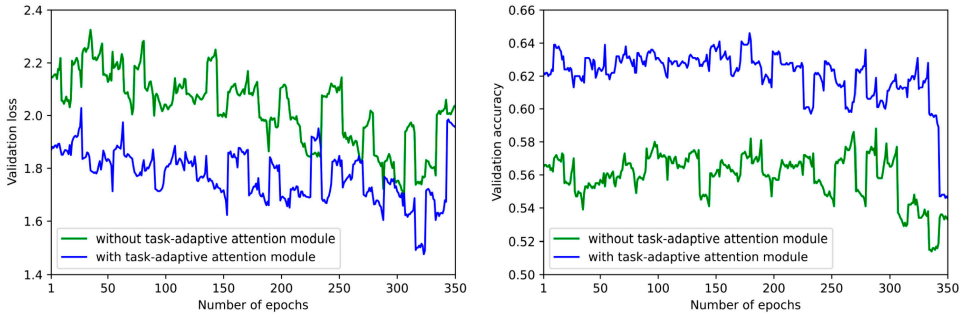


Figure 8. 5-way accuracies of the ablation analysis of task-adaptive attention module. When task-adaptive attention module is removed, the classification performance of the model degenerates dramatically in all scenarios.

To intuitively illustrate the effect of the task-adaptive attention module, 5-way 1-shot classification experiments are carried out on the UC Merced dataset. Figure 9 visualizes

corresponding experimental results, including validation accuracies and losses in the cases of with the task-adaptive attention module and without the task-adaptive attention module. The number of epochs ranges from 1 to 350 with the step size of 1. As can be seen in Figure 9, in the meta-training stage, the validation losses with the task-adaptive attention module are smaller than that without the task-adaptive attention module, and the best validation accuracy obtained by our method is higher than that without the task-adaptive attention module.



(**a**) (**b**)

Figure 9. Validation losses and accuracies on the UC Merced dataset in the scenarios of 5-way 1-shot. (**a**) shows the validation losses, and (**b**) shows the validation accuracies.

4.4.3. Effect of Shots

To further validate the 5-way classification performance in the case of different number of shots, a series of experiments are performed on two public datasets, that is, UC Merced and NWPU-RESISC45. Figure 10 presents the corresponding experimental results, where our model is provided with 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 labeled support images. It is observed from Figure 10 that the classification accuracy significantly improved as the quantity of shots increases from 1 to 5. Nevertheless, the classification performance of the model does not benefit a lot as the number of shots keeps on increasing. It demonstrates that our proposed TAE-Net is well-suited to scene classification tasks in the case of extremely limited labeled samples.

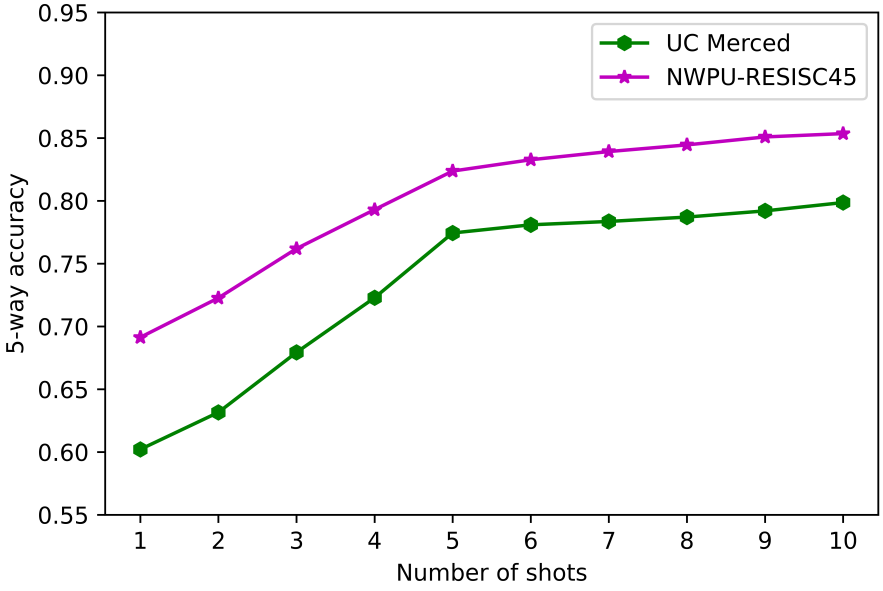


Figure 10. The influence of different shots on 5-way accuracies is presented with 95% confidence interval, in which the experiments are performed on UC Merced and NWPU-RESISC45.

5. Conclusions

In this paper, a task-adaptive embedding network based on meta-learning is proposed for remote sensing scene classification in few-shot settings, which aims to extract more informative embedding features from the perspective of the entire task. Our proposed approach first adopts Conv-4 as an encoder to learn embedding representation from the base set. Next, in the meta-training phase, the whole model is optimized by task-adaptive embeddings. To be specific, a task-adaptive attention mechanism is developed to adaptively filter irrelevant information and retain discriminative information in a certain task, which can avoid distraction from the general features shared by different categories. Experiments on three popular scene datasets illustrate that our proposed approach exceeds existing state-of-the-art few-shot approaches and achieves new state-of-the-art performance. Moreover, a series of ablation experiments are performed to investigate the influence of the pre-training strategy, task-adaptive attention mechanism, and the number of shots.

In the current work, we only use labeled data for few-shot scene classification, while ignoring the impact of unlabeled data on classification performance. Hence, an attractive research focus is the improvement of the performance of few-shot remote sensing scene classification under the semi-supervised setting, which will be the topic of future research.

Author Contributions: Conceptualization, W.H., Z.Y., A.Y. and X.L.; resources, W.H. and Z.Y.; data curation, W.H., Z.Y. and A.Y.; software, W.H. and Z.Y.; investigation, W.H., Z.Y. and A.Y.; methodology, W.H., A.Y. and X.L.; validation, W.H., Z.Y., A.Y. and C.T.; visualization, W.H.; writing—original draft, W.H.; writing—review and editing, W.H. and C.T.; supervision, C.T. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Advance Research Project of Civil Space Technology (grant no. D040402) and the National Natural Science Foundation of China (grant no. 41871226).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author.

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

References

1. Pham, H.M.; Yamaguchi, Y.; Bui, T.Q. A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landsc. Urban Plan.* **2011**, *100*, 223–230. [CrossRef]
2. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
3. 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]
4. Jahromi, M.N.; Jahromi, M.N.; Pourghasemi, H.R.; Zand-Parsa, S.; Jamshidi, S. Accuracy assessment of forest mapping in MODIS land cover dataset using fuzzy set theory. In *Forest Resources Resilience and Conflicts*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 165–183.
5. Li, Y.; Yang, J. Meta-learning baselines and database for few-shot classification in agriculture. *Comput. Electron. Agric.* **2021**, *182*, 106055. [CrossRef]
6. Li, X.; Shao, G. Object-based urban vegetation mapping with high-resolution aerial photography as a single data source. *Int. J. Remote Sens.* **2013**, *34*, 771–789. [CrossRef]
7. Fang, B.; Li, Y.; Zhang, H.; Chan, J.C.W. Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection. *Remote Sens.* **2018**, *10*, 574. [CrossRef]
8. Tai, X.; Li, M.; Xiang, M.; Ren, P. A mutual guide framework for training hyperspectral image classifiers with small data. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1–17. [CrossRef]
9. Denisova, A.Y.; Kavelenova, L.M.; Korchikov, E.S.; Prokhorova, N.V.; Terentyeva, D.A.; Fedoseev, V.A. Tree species classification for clarification of forest inventory data using Sentinel-2 images. In Proceedings of the Seventh International Conference on Remote Sensing and Geoinformation of the Environment, Paphos, Cyprus, 18–21 March 2019; International Society for Optics and Photonics: Bellingham, WA, USA, 2019; Volume 11174, p. 1117408.

10. Alajaji, D.; Alhichri, H.S.; Ammour, N.; Alajlan, N. Few-shot learning for remote sensing scene classification. In Proceedings of the 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 9–11 March 2020; pp. 81–84.
11. Cen, F.; Wang, G. Boosting occluded image classification via subspace decomposition-based estimation of deep features. *IEEE Trans. Cybern.* **2019**, *50*, 3409–3422. [CrossRef]
12. Noothout, J.M.; De Vos, B.D.; Wolterink, J.M.; Postma, E.M.; Smeets, P.A.; Takx, R.A.; Leiner, T.; Viergever, M.A.; Išgum, I. Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE Trans. Med. Imaging* **2020**, *39*, 4011–4022. [CrossRef]
13. Du, L.; Li, L.; Guo, Y.; Wang, Y.; Ren, K.; Chen, J. Two-Stream Deep Fusion Network Based on VAE and CNN for Synthetic Aperture Radar Target Recognition. *Remote Sens.* **2021**, *13*, 4021. [CrossRef]
14. Andriyanov, N.; Dementiev, V.; Gladkikh, A. Analysis of the Pattern Recognition Efficiency on Non-Optical Images. In Proceedings of the 2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Yekaterinburg, Russia, 13–14 May 2021; pp. 319–323.
15. Xu, P.; Li, Q.; Zhang, B.; Wu, F.; Zhao, K.; Du, X.; Yang, C.; Zhong, R. On-Board Real-Time Ship Detection in HISEA-1 SAR Images Based on CFAR and Lightweight Deep Learning. *Remote Sens.* **2021**, *13*, 1995. [CrossRef]
16. Wu, B.; Meng, D.; Zhao, H. Semi-supervised learning for seismic impedance inversion using generative adversarial networks. *Remote Sens.* **2021**, *13*, 909. [CrossRef]
17. Liu, Y.; Zhong, Y.; Fei, F.; Zhang, L. Scene semantic classification based on random-scale stretched convolutional neural network for high-spatial resolution remote sensing imagery. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016; pp. 763–766.
18. Zeng, Q.; Geng, J.; Huang, K.; Jiang, W.; Guo, J. Prototype Calibration with Feature Generation for Few-Shot Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 2728. [CrossRef]
19. Geng, J.; Deng, X.; Ma, X.; Jiang, W. Transfer learning for SAR image classification via deep joint distribution adaptation networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5377–5392. [CrossRef]
20. Chang, H.; Yeung, D.Y. Semisupervised metric learning by kernel matrix adaptation. In Proceedings of the International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 5, pp. 3210–3215.
21. Lee, K.; Maji, S.; Ravichandran, A.; Soatto, S. Meta-learning with differentiable convex optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10657–10665.
22. Shao, L.; Zhu, F.; Li, X. Transfer learning for visual categorization: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *26*, 1019–1034. [CrossRef] [PubMed]
23. 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.
24. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching networks for one shot learning. *Proc. Neural Inf. Process. Syst.* **2016**, *29*, 3630–3638.
25. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical networks for few-shot learning. *Proc. Neural Inf. Process. Syst.* **2017**, *30*, 4077–4087.
26. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; 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.
27. Zhang, C.; Cai, Y.; Lin, G.; Shen, C. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12203–12213.
28. Xu, Z.; Cao, L.; Chen, X. Learning to learn: Hierarchical meta-critic networks. *IEEE Access* **2019**, *7*, 57069–57077. [CrossRef]
29. Zhai, M.; Liu, H.; Sun, F. Lifelong learning for scene recognition in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1472–1476. [CrossRef]
30. Liu, S.; Deng, W. Very deep convolutional neural network based image classification using small training sample size. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 730–734.
31. Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; Tao, C. RS-MetaNet: Deep meta metric learning for few-shot remote sensing scene classification. *arXiv* **2020**, arXiv:2009.13364
32. Li, L.; Han, J.; Yao, X.; Cheng, G.; Guo, L. DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7844–7853. [CrossRef]
33. Jiang, W.; Huang, K.; Geng, J.; Deng, X. Multi-scale metric learning for few-shot learning. *IEEE Trans. Circuits Syst.* **2020**, *31*, 1091–1102. [CrossRef]
34. Ma, H.; Yang, Y. Two specific multiple-level-set models for high-resolution remote-sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 558–561.
35. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
36. Tang, X.; Ma, Q.; Zhang, X.; Liu, F.; Ma, J.; Jiao, L. Attention consistent network for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2030–2045. [CrossRef]

37. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
38. Lu, X.; Gong, T.; Zheng, X. Multisource compensation network for remote sensing cross-domain scene classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2504–2515. [CrossRef]
39. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
40. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
41. Sun, H.; Li, S.; Zheng, X.; Lu, X. Remote sensing scene classification by gated bidirectional network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 82–96. [CrossRef]
42. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
43. Xie, H.; Chen, Y.; Ghamisi, P. Remote Sensing Image Scene Classification via Label Augmentation and Intra-Class Constraint. *Remote Sens.* **2021**, *13*, 2566. [CrossRef]
44. Shi, C.; Zhao, X.; Wang, L. A Multi-Branch Feature Fusion Strategy Based on an Attention Mechanism for Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 1950. [CrossRef]
45. Oreshkin, B.N.; Rodriguez, P.; Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv* **2018**, arXiv:1805.10123.
46. Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J.B.; Larochelle, H.; Zemel, R.S. Meta-learning for semi-supervised few-shot classification. *arXiv* **2018**, arXiv:1803.00676.
47. 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; PMLR: Sydney, Australia, 2017; pp. 1126–1135.
48. Nichol, A.; Achiam, J.; Schulman, J. On first-order meta-learning algorithms. *arXiv* **2018**, arXiv:1803.02999.
49. Sun, Q.; Liu, Y.; Chua, T.S.; Schiele, B. Meta-transfer learning for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 403–412.
50. Jamal, M.A.; Qi, G.J. Task agnostic meta-learning for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 11719–11727.
51. Rusu, A.A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; Hadsell, R. Meta-learning with latent embedding optimization. *arXiv* **2018**, arXiv:1807.05960.
52. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv* **2017**, arXiv:1707.09835.
53. Gupta, A.; Thadani, K.; O'Hare, N. Effective few-shot classification with transfer learning. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 1061–1066.
54. Dhillon, G.S.; Chaudhari, P.; Ravichandran, A.; Soatto, S. A baseline for few-shot image classification. *arXiv* **2019**, arXiv:1909.02729.
55. Chen, W.Y.; Liu, Y.C.; Kira, Z.; Wang, Y.C.F.; 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.
56. Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S.J.; Yang, Y. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv* **2018**, arXiv:1805.10002.
57. Yang, L.; Li, L.; Zhang, Z.; Zhou, X.; Zhou, E.; Liu, Y. Dpgn: Distribution propagation graph network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13390–13399.
58. Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; Luo, J. Revisiting local descriptor based image-to-class measure for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7260–7268.
59. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
60. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [CrossRef]
61. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
62. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.