 doi:10.20944/preprints202010.0033.v1

Article

Few-shot Classification of Aerial Scene Images via Meta-learning

Pei Zhang[1], Ying Li[1,*], Dong Wang[1], and Yunpeng Bai[2]

[1] School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, Northwestern Polytechnical University, Xi'an 710129, China; cszhangpei@mail.nwpu.edu.cn (P.Z.); dongwang@mail.nwpu.edu.cn (D.W.)

[2] School of Computing and Information Systems, The University of Melbourne, Victoria 3010, Australia; yunpengb@student.unimelb.edu.au (Y.B)

* Correspondence: lybyp@nwpu.edu.cn; Tel.: +86-029-8843-1532

Abstract: CNN-based methods have dominated the field of aerial scene classification for the past few years. While achieving remarkable success, CNN-based methods suffer from excessive parameters and notoriously rely on large amounts of training data. In this work, we introduce few-shot learning to the aerial scene classification problem. Few-shot learning aims to learn a model on base-set that can quickly adapt to *unseen* categories in novel-set, using only a few labeled samples. To this end, we proposed a meta-learning method for few-shot classification of aerial scene images. First, we train a feature extractor on all base categories to learn a representation of inputs. Then in the meta-training stage, the classifier is optimized in the metric space by cosine distance with a learnable scale parameter. At last, in the meta-testing stage, the *query* sample in the *unseen* category is predicted by the adapted classifier given a few *support* samples. We conduct extensive experiments on two challenging datasets: NWPU-RESISC45 and RSD46-WHU. The experimental results show that our method outperforms three state-of-the-art few-shot algorithms and one typical CNN-based method, D-CNN. Furthermore, several ablation experiments are conducted to investigate the effects of dataset scale and support shots; the experiment results confirm that our model is specifically effective in few-shot settings.

Keywords: aerial scene classification; remote-sensing image classification; few-shot learning; meta-learning

1. Introduction

Aerial images, taken from the air and space, provide sufficient detail about the earth's surface, such as its landforms, vegetation, landscapes, buildings, and other various resources. Such abundant information is a significant data source for earth observation[1], which opens the door to a broad range of essential applications spanning urban planning[2], land-use and land-cover (LULC) determination[3,4], mapping[5], environmental monitoring[6] and climate modeling. As a fundamental problem in the remote sensing community, aerial scene classification is crucial for these research fields. Xia et al. [7] defined the aerial scene classification as automatically assigning a specific semantic label to each image according to its content.

Over the past few decades, aerial scene classification enjoys much attention from researchers, and many methods have been proposed. The existing approaches to aerial scene classification have mostly fallen into three categories – methods adopting low-level feature descriptors, methods using middle-level visual representations and methods relying on deep learning networks.

Methods adopting low-level feature descriptors. Most early researches[8-10] on aerial image classification fall into this category. These methods use hand-crafted, low-level visual features such as color, spectrum, texture, structure, or their combination to distinguish aerial scene images. Among the hand-crafted features, the most representative feature descriptors include color

 doi:10.20944/preprints202010.0033.v1

histograms[10], texture features[9], and SIFT[8]. While this type of method performs well in certain aerial scenes with uniform structures and spatial arrangements, it has limited performance for aerial images containing complex semantic information.

Methods using middle-level visual representations. In order to overwhelm the insufficiency of low-level methods, many middle-level methods have been explored for aerial scene classification. Such methods mainly aim at combining the local visual attributes extracted by low-level feature methods into high-order statistical patterns to build a holistic scene representation for aerial scenes. Bag of Visual Words (BOVW)[11] and many of its variants have been widely used. Besides the BOVW model, typical middle-level methods include, but not limited to, Spatial Pyramid Matching (SPM)[12], Vector of Locally Aggregated Descriptors (VLAD)[13], Locality-constrained Linear Coding (LLC)[14], Probabilistic Latent Semantic Analysis (pLSA)[15] and Latent Dirichlet Allocation (LDA)[16]. Compared with low-level methods, the scene classification methods using middle-level visual representations have obtained higher accuracy. However, middle-level methods will only go so far; they require hand design features and lack adaptability; their generalization is poor for complex scenes or massive data.

Methods relying on deep learning. Fortunately, with the emergence of deep learning, especially convolutional neural networks[17,18], image classification approaches have seen great success in both accuracy and efficiency, also in remote sensing fields. The methods relying on deep neural networks automatically learn global features from the input data and cast the aerial scene classification task as an end-to-end problem. More recently, while the deep CNNs methods have become the new state-of-the-art solutions[19-22] for the aerial scene classification area, yet, there are clear limitations. Specifically, the most notorious drawback of deep learning methods is that they typically require vast quantities of labeled data and suffer from poor sample efficiency, which excludes many applications where data is intrinsically rare or expensive[23]. In contrast, humans possess a remarkable ability to learn new abstract concepts from only a few examples and quickly generalize to new circumstances. For instance, Marcus, G.F.[24] pointed out that even a 7-month-old baby can learn abstract language-like rules from a handful of unlabeled examples, in just two minutes.

Why do we need few-shot learning? In a world with unlimited data and computational resources, we might hardly need any other technique rather than deep learning. However, we live in a real-world where data are never infinite, especially in the remote sensing community, due to the high cost of collecting. Still, almost all existing aerial scene datasets have several notable limitations.

On the one hand, the classification accuracy is saturated; to be more specific, the state-of-the-art methods can achieve nearly 100% accuracy on the most popular UC Merced dataset[11] and the WHU-RS19[25] dataset. Yet, we argue, such a limited number of categories in the two datasets are critically insufficient for the real world. On the other hand, the scale of the scene categories and the image number per class are limited, and the images lack scene variation and diversity. An intuitive way to tackle this issue is to construct a large-scale dataset for aerial scene classification, and several more challenging datasets, including the AID dataset[7], the PatternNet dataset[26], the NWPU-RESISC45 dataset[19], and the RSD46-WHU dataset[27,28], have been proposed.

Although the aerial scene datasets increase in scale, most of them are still considered small from the perspective of deep learning. For similar situations in the machine learning community, few-shot learning[29] offers an alternative way to address the data-hungry issue from a different standpoint. Instead of expanding the dataset scale, few-shot learning aims to learn a model that can quickly generalize to new tasks from very few labeled examples. Arguably, few-shot learning is a human-like way of learning. It assumes a more realistic situation where not rely on thousands or millions of supervised training data. Namely, few-shot learning can help to relieve the burden of collecting data, especially in some specific domains in which collecting labeled examples is usually time-consuming and laborious, such as aerial scene field or drug discovery. Figure 1 demonstrates a specific 1-shot scenario that it is possible to learn much information about a new category from just one image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

Figure 1. Illustration of using few-shot learning to learn information from just one labeled image.

By seeing the potential that few-shot learning can alleviate the data-gathering effort, improve computing efficiency, and bridge the gap between Artificial Intelligence and human-like learning, we introduce the few-shot paradigm to the aerial scene classification problem. The goal of this work is to classify aerial scene images with only 1 or 5 labeled samples. More specifically, we adopt a meta-learning framework to address this problem. To the best of our knowledge, we are the first to provide a testbed for few-shot classification of aerial scene images. We re-implement three state-of-the-art few-shot learning algorithms, namely Prototypical Networks[29], MAML[30] and Relation Network[31], moreover, a typical CNN-based method D-CNN[21] for comparison.

The main contributions of this article are summarized as follows.

1. This is the first work to provide a testbed for comparison with several different few-shot learning algorithms in the aerial scene field. Our experimental evaluation reveals that it is possible to learn much information for a new category from just a few labeled images, which is a great potential for introducing the few-shot paradigm to the remote sensing community.

2. The proposed method including a feature extraction module and a meta-learning module. First, ResNet-12 is used as a backbone to learn a representation f of input on *base* set. Then, in the meta-training stage, we optimize the classifier by cosine distance with a learnable scale parameter in the feature space.

3. We conduct extensive experiments on two challenging datasets: NWPU-RESISC45 and RSD46-WHU. Besides, we build a mini dataset from the RSD46-WHU to investigate how the scale of the dataset affects the performance. At last, we analyze the performance as a function of the *support* shots. The experimental results demonstrate that our model is specifically effective in few-shot settings.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work on CNN-based methods of aerial scene classification and various state-of-the-art few-shot classification approaches that developed recently. In Section 3, we introduce some preliminary of the few-shot classification as it may be new to some readers. The proposed meta-learning method is described in Section 4. We illustrate the datasets and discuss the experiment results in Section 5. Moreover, finally, Section 6 concludes the paper with a summary and an outlook.

2. Related Work

2.1. CNN-based methods of Aerial Scene Classification

Aerial scene classification has been well studied for the last few decades owing to its broad applications. Since the emergence of the AlexNet[17] in 2012, deep learning-based methods have made an enormous breakthrough, much defeated the traditional methods based on low-level and middle-level methods, and became mainstream in the aerial scene classification task.

One strand study attempted to use a transfer learning method to fine-tune the pre-trained CNNs for aerial image classification. In [32], Yu et al. studied how to transfer the activations of CNNs pre-trained on the ImageNet dataset to high-resolution remote sensing classification. Cheng et al. [19] obtained better performance by using fine-tuned AlexNet[17], VGGNet-16[18], and GoogleNet[33] on the dataset NWPU-RESISC45. Similarly, Nogueira et al. [20] carried out three strategies, namely full training, fine-tuning, and using CNNs as feature extractors, for exploiting six common CNNs in three remote sensing datasets. Their experiment results demonstrate that fine-tuning is generally the best strategy in different situations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

Some further studies utilize the pre-trained CNNs for feature extraction and combine the high-level semantic features with hand-crafted features. Zhao and Du [34] proposed a CNN framework to learn local spatial patterns from multi-scale. Wang et al. [35] presented an encoded mixed-resolution representation framework where multilayer features are extracted from various convolutional layers. The study by Lu et al. [36] introduced an adaptive feature strategy that fuses the deep learning feature and the SIFT feature to overwhelm the scale and rotation variability, which is essential in remote sensing images but cannot be captured by CNN-based methods.

More recent research has begun to concern the problem of within-class diversity and between-class similarity in aerial scene images. For example, to tackle this issue, Cheng et al. [21] trained a discriminative CNN model by optimizing a novel objective function. Beyond a traditional cross-entropy loss, a metric learning regularization term and a weight decay term are added to the proposed objective function. Li et al. [22] constructed a feature fusion network that combining the original feature and attention map feature; besides that, they adopted center loss[37] to improve feature distinguishability.

2.2. Few-shot Classification via Meta-Learning

Deep learning-based approaches have achieved remarkable success in various fields, especially in areas where vast quantities of data can be collected and where substantial computing resources are available. However, deep learning is often suffered from poor sample efficiency. Recently, few-shot learning is proposed to tackle this problem and have been marked by exceptional progress. Few-shot learning aims to learn new concepts from only small amounts of samples and quickly adapt to unforeseen tasks, which can be viewed as a special case of meta-learning. In the following, we introduce some representative few-shot classification literature, gathering into two main streams: optimization-based methods and metric-based methods.

Optimization-based methods. This line of work is most understood as *learning to learn*, which tackles the few-shot classification problem by effectively optimizing model parameters to new tasks. Finn et al. proposed a model-agnostic algorithm named MAML[30], which targets to learn a good initialization of any standard neural network. In such a way, it means to prepare that network for fast adaptation to any novel task through only one or a few gradient steps. The authors also presented a first-order approximation version of MAML by ignoring second-order derivatives to speed-up the network computation. Reptile[38] expands on the results from MAML by performing a Taylor series expansion update and finding a point near all solution manifolds of the training tasks. Many variants[39-41] of MAML follow a similar idea that, with a good initialization, one is just a few gradient steps away from a solution to a new task. These approaches face a critical challenge that the external optimization needs to solve as many parameters as internal optimization. Besides, there is a key debate. That is, whether a single initial condition is sufficient to provide fast adaption for a wide range of potential tasks. And further, whether an initial condition is restricted to relatively narrow distributions.

Metric-based methods. Another family of approach aims to address few-shot classification by *learning to compare*. The key insight of the idea is to learn a feature extractor that mapping raw input into a representation suitable for predicting, such that, when represented in this feature space, the *query* and *support* samples are easy for comparison (e.g., with Euclidean distance or cosine similarity). Matching Networks[42] mapping the support set via an attention mechanism to a function and then classifying the query sample by a weighted nearest-neighbor classifier in an embedding space. Prototypical Networks[29] follows a similar idea that learns a metric-based prediction rule over embeddings. The prototype of each category is represented by the mean embedding of samples, such that the classification can be performed by computing distances to the nearest category mean. Besides a usual embedding module, Relation Network[31] introduces an additional parameterized CNN-based 'relation module' for learnable metric comparison.

While meta-learning approaches have seen great success in few-show classification, some pre-trained methods have recently gained competitive performance[43,44]. Our work is more

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

related to the second line of work by finding a suitable distance metric and taking the pre-trained method's strength by learning good feature embeddings.

3. Preliminary

Before introducing our overall framework in detail, we first look at some preliminary of the few-shot classification as it may be new to some readers.

In standard supervised classification, we are dealing with a dataset $D=\{D_{train}, D_{test}\}$. The training set takes labeled input-output pairs as inputs, denoted as $D_{train}=\{(x_i, y_i)\}_{i=1}^{N}$, $y_i \in \{1,...,C\}$, where N is the number of training samples, C is the number of categories in D_{train}. We are interested in learning a model $\hat{y}=f_\theta(x)$, parameterized by θ on D_{train}, to predict the label $\hat{y} \in \{1,...,C\}$ for an unlabeled sample x_k in the test set $D_{test}=\{(x_k)\}_{k=1}^{K}$.

In few-shot classification, we instead consider learning a model that can generalize effectively to unseen categories in training, given only a few samples, usually 1 or 5, in each new category. Following recent work[29,42], we formalize the few-shot classification paradigm as below. Given a meta-set $\mathcal{D}=\{\mathcal{D}_{base}, \mathcal{D}_{novel}\}$, and $\mathcal{C}_{base} \cap \mathcal{C}_{novel}=\varnothing$, where $\mathcal{C}$ represents the category. The vision is to learn a model $\mathcal{M}$ on $\mathcal{D}_{base}$ that can quickly adapt to unseen categories in $\mathcal{D}_{novel}$ with only a few support samples. To this end, we consider training and evaluating the model on a set of *tasks*, or so-called *episodes*. Here, we treat entire *tasks* as training instances in conventional machine learning. Specifically, we adopt an *N-way K-shot* setting, in which each *episode* has a *support-set* $\mathcal{S}$ and a *query-set* $\mathcal{Q}$. The support-set $\mathcal{S}$ contains N unique categories with K labeled samples in each. The *query-set* $\mathcal{Q}$ holds the same N categories, each with Q unlabeled samples being to classify. The difference between standard supervised classification and few-shot classification is illustrated in Figure 2. More details are described in section 4.3.

Figure 2. (a) Dataset-split for standard supervised classification; (b) Dataset-split for few-shot classification.

4. Proposed Method

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

4.1 Overall Framework

In this work, we propose a meta-learning method for few-shot classification of aerial scene images. The framework consists of a feature extractor, a meta-training stage, and a meta-testing stage. Figure 3 illustrates the overall procedure of our method. First, a feature extractor is trained on base dataset $\mathcal{D}_{base}$ to learn a representation of inputs for further comparison in feature space. To achieve this, we train a typical classifier on all base categories by minimizing a standard cross-entropy loss and removing its last FC Layer to get a 512-dimensional feature representation. Then, we consider training a meta-learning classifier $\mathcal{M}$ over a set of *episodes* in the meta-training stage. Concretely, the objective is to optimize the classifier $\mathcal{M}$ by minimizing the generalization error across *episodes*. For a single *episode*, the query features are compared with the category mean of support features by cosine distance. Finally, in the meat-testing stage, the meta-learning classifier $\mathcal{M}$ is estimated on a set of *episodes* sampled from the novel set $\mathcal{D}_{novel}$, usually referred to as a *meta-test* set.

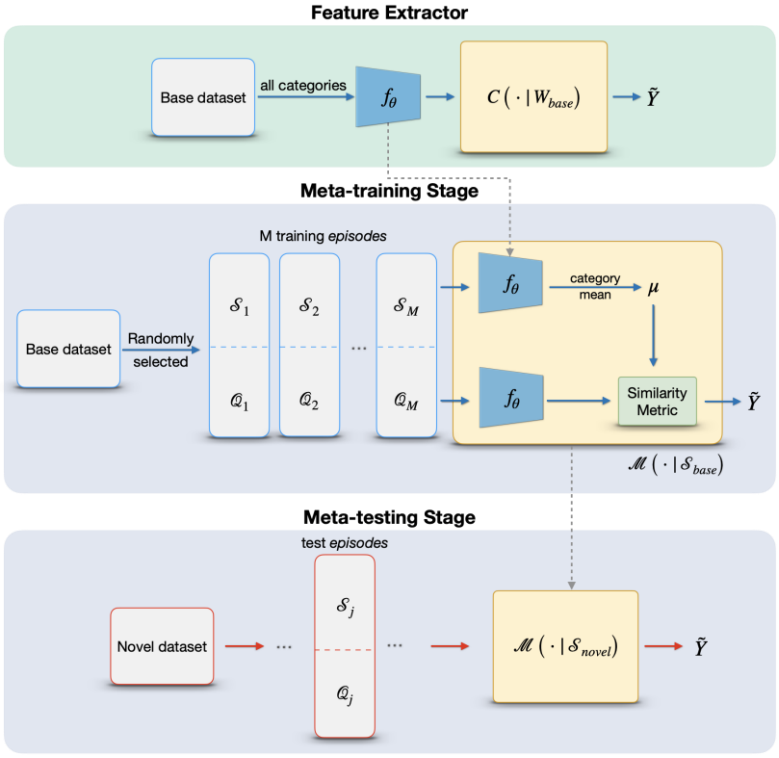

Figure 3. Overall framework of our method. The top represents the feature extractor trained on base dataset.

4.2 Feature extractor

We train a feature extractor f_θ with parameters θ on the base set $\mathcal{D}_{base}$ that encodes the input data to a 512-dimensional feature vector suitable for comparison. Here we employ ResNet-12 to learn a classifier on all base categories and remove the last fully connected layer to get f_θ, which is described below; though, other backbones can also be used. Before feeding to the network, all input images in $\mathcal{D}_{base}$ are resized to 80×80. The architectural setting of ResNet-12 we use, illustrated in Figure 4, consists of four ResNet blocks. Three convolutional layers configure each ResNet block with a 3×3 kernel, followed by BN and Leaky ReLU. As shown in the figure below, $\{C_i\}_{i=1}^{4}$ denotes the channels of convolutional layers in each ResNet block, which is 64, 128, 256, 512, respectively. We then adopt a Leaky ReLU and 2×2 max pooling layer right after each residual block. Lastly, by feeding the $5\times5\times512$ vector generated by the ResNet Block-4 to the 5×5 average pooling layer, we can finally get a 512-dimensional feature representation, as mentioned in the beginning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

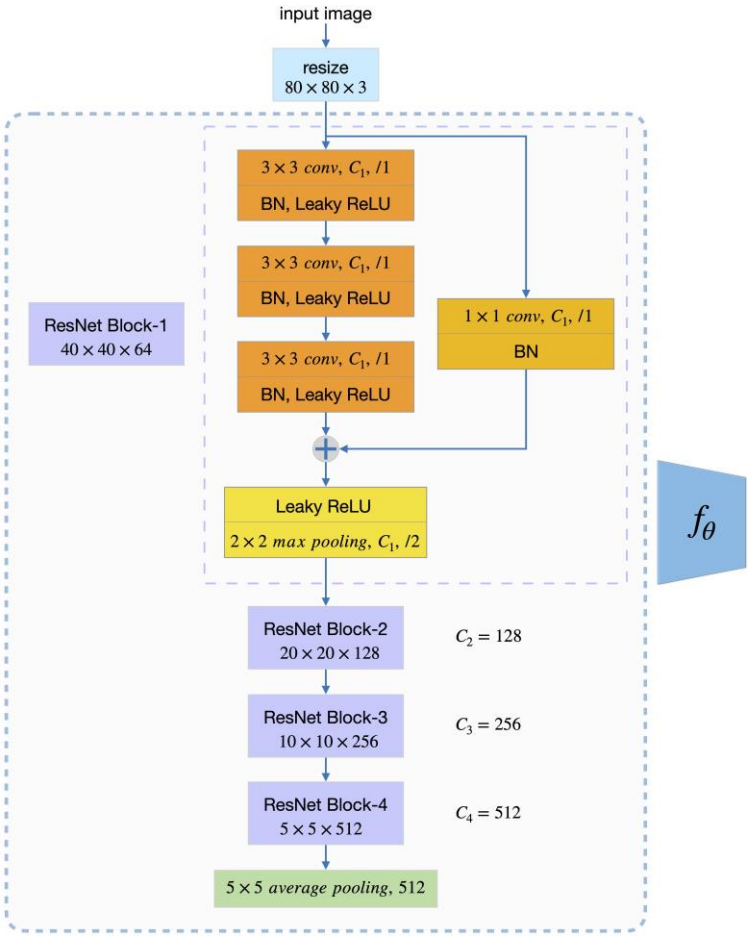

Figure 4. The structure of ResNet-12 with four ResNet blocks.

4.3 Meta-training stage

Section 3 briefly introduces the problem settings in few-shot learning. In this section, we first begin by presenting the problem definition and the key notation more formally in detail, which are useful for understanding the following training procedure.

A dataset we deal with in few-shot learning denotes as meta-set $\mathcal{D}$, which has a split of $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$, meaning base-set and novel-set, respectively. $\mathcal{C}_{total}=\{\mathcal{C}_{base},\mathcal{C}_{novel}\}$, where $\mathcal{C}_{total}$ indicates the total categories in the whole meta-set $\mathcal{D}$, similarly, $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ refer to categories in $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$, respectively, note that $\mathcal{C}_{base}\cap\mathcal{C}_{novel}=\varnothing$. Let $\mathcal{D}_c$ denotes the subset of $\mathcal{D}$ containing all samples (x_i,y_i) belonging to category c, that is $y_i=c$. Few-shot classification aims to train a model $\mathcal{M}(\cdot|\mathcal{S})$ on base-set $\mathcal{D}_{base}$ that learns *meta-knowledge*, which can generalize to unseen categories with only a few *support* samples in novel-set $\mathcal{D}_{novel}$. In this section, we focus on the learning procedure in the meta-training stage that only processes data in the base-set $\mathcal{D}_{base}$.

To train the model in an effective way, we usually assume improving performance by learning from a set of *tasks*, denoted as $\mathcal{D}_T=\{\mathcal{T}_i\}$, also known as *episodes*. In effect, an *episode* $\mathcal{T}_i$ is treated as a data-point in meta-learning. Following the standard few-shot classification paradigm, we often employ the *N-way K-shot* setting to evaluate the model. Thus, we construct an *episode* with *N* randomly selected categories, each with *K support* samples and *Q query* samples. That means each *episode* has both training and test data, denoting as support-set $\mathcal{S}=\{(x_1^s,y_1^s),...,(x_{NK}^s,y_{NK}^s)\}$ and query-set $\mathcal{Q}=\{(x_1^q,y_1^q),...,(x_{NK_Q}^q,y_{NK_Q}^q)\}$, where (x_i,y_i) is an input-output pair, $y_i\in\{1,...,N\}$, and K_Q is the number of *query* images per category. Furthermore, we formulate the set of *M* episodes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

used in the meta-training stage as $\mathcal{D}_{\mathcal{T}}^{train}=\left\{\left(\mathcal{S}_{base},\mathcal{Q}_{base}\right)^{(i)}\right\}_{i=1}^{M}$, we will discuss the *meta-test* set $\mathcal{D}_{\mathcal{T}}^{test}$ in the following Section 4.4.

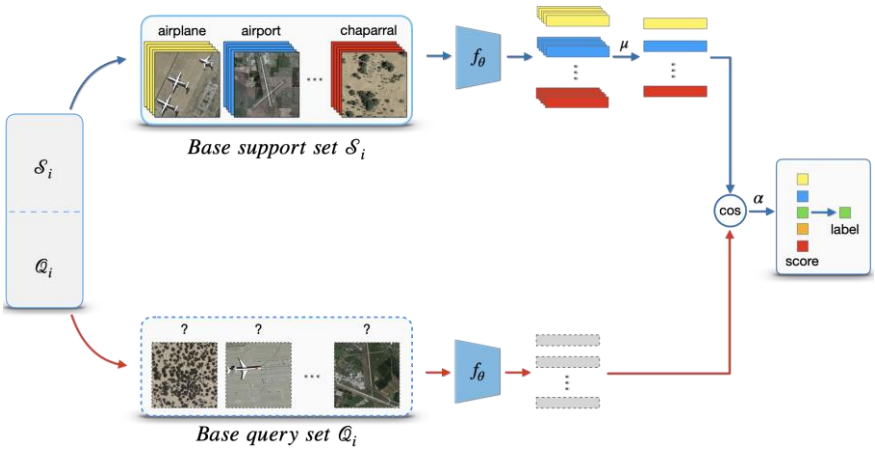

Figure 5. The architecture of meta-training stage for a *N-way K-shot* classification problem.

According to the conventional *N-way K-shot* setting, our goal is to train a meta-learning model $\mathcal{M}(\cdot|\mathcal{S})$ that minimizes the *N-way* prediction loss. To accomplish this, we sample *M episodes* from training data in base categories. An *episode* has *K* input-output pairs randomly selected from each category, namely a total of $N\times K$ samples for *N-way* classification training and $N\times Q$ query samples for test. During meta-training stage, an additional *meta-validation set* is held out to choose the hyper-parameters of the model $\mathcal{M}(\cdot|\mathcal{S})$. Given an episode with the support-set $\mathcal{S}$, we denote $\mathcal{S}_c$ as a subset of $\mathcal{S}$ with all samples in category *c*. Snell, J.[29] defined a *prototype* ω_c as the mean vector over embeddings belonging to $\mathcal{S}_c$ (the centroid of category *c*), an embedding is generate by the pre-trained feature extractor f_θ with learnable parameters θ we described in Section 4.2. We can write down the ω_c as follows:

$$\omega_c=\frac{1}{|\mathcal{S}_c|}\sum_{(x_i)\in\mathcal{S}_c}f_\theta(x_i) \tag{1}$$

One intuitive way to predict the probability that a query sample x belongs to category c is to compare the distance between the feature embedding $f_\theta(x)$ and the centroid ω_c of category c. Two common distance metrics are Euclidean distance and cosine similarity, here we employ the cosine similarity, and thus the prediction can be formalized as follows:

$$p(y=c|x)=\alpha\frac{exp\left(cos\left(f_\theta(x),\omega_c\right)\right)}{\sum_{c'}exp\left(cos\left(f_\theta(x),\omega_{c'}\right)\right)}, \tag{2}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity of two vectors.

Inspired by Gidaris et al.[45], we introduce a learnable scalar parameter α to adjust the original value range $[-1,1]$ of cosine similarity. In our experiments, α is initialized to 10, and we observe that the scaling similarity metric is more appropriate for the following softmax layer. Then, the predictive probability becomes:

$$p(y=c|x)=\frac{exp\left(\alpha\cdot cos\left(f_\theta(x),\omega_c\right)\right)}{\sum_{c'}exp\left(\alpha\cdot cos\left(f_\theta(x),\omega_{c'}\right)\right)}. \tag{3}$$

4.4 Meta-testing stage

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

Once the meta-learning model $\mathcal{M}(\cdot|\mathcal{S}_{base})$ is learned, its generalization is evaluated on a held-out set $\mathcal{D}_{novel}$. Note that all categories in novel-set $\mathcal{D}_{novel}$ are *unseen* in the meta-training stage. The test *episodes* are sampled from $\mathcal{D}_{novel}$, often referred to as a *meta-test* set $\mathcal{D}_{\mathcal{T}}^{test}=\left\{\left(\mathcal{S}_{novel},\mathcal{Q}_{novel}\right)^{(j)}\right\}_{j=1}^{J}$. The learned model is adapted to predict *unseen* categories with the new *support* set $\mathcal{S}_{novel}$.

5. Experiments and Analysis

In this section, we first present some implementation details and dataset description. Then, we compare our method with three state-of-the-art few-shot methods and one typical CNN-based method, D-CNN. In addition, we conduct a new dataset *mini-RSD46-WHU* to investigate how the scale of the dataset impacts the results. At last, we also carry out experiments to evaluate the 5-way accuracy as a function of shots.

5.1. Implementation details

Following the few-shot experimental protocol proposed by Vinyals, O. [42], we carry out the experiments of *N-way* classification with *K* shots, here $N=5$, $K=1$ or 5. In the meta-training procedure, a few-shot training batch is composed of several episodes where an episode is a selection of 5 randomly categories drawn from $\mathcal{D}_{base}$. We set 4 episodes per batch to compute the average loss, namely the batch size is 4. The *support* set in each training episode is expected to match the same number of shots as in the meta-test stage. That is, for example, if we want to perform 5-way 1-shot classification at test-time, then the training episodes could be constituted of $N=5$, $K=1$. Note that each category contains *K query* samples during meta-training stage and 15 query samples during meta-testing.

We employ Resnet-12 as our backbone; by removing the fully connected layer, the network generates a 512-dimensional feature vector for each input image. For this step, we use SGD optimizer with momentum 0.9, the learning rate is initialized to 0.1, and the decay factor is set to 0.1. The feature extractor was trained for 100 epochs with batch size 128 on 2 GPUs, the weight decay for ResNet-12 is 0.0005. For ProtoNet, MAML, and RelationNet, we follow the original literature and adopt a four-layer convolutional backbone (Conv-4). In addition, we re-implement a typical machine learning classification method D-CNN[21], to evaluate its performance in the few-shot scenario. ResNet-12 and the same settings are used in the re-implementation. All our code was implemented in Pytorch and run with *2* NVIDIA *2080ti* GPUs.

5.2. Datasets Description

We evaluate our proposed method on two challenging datasets: NWPU-RESISC45[19] and RSD46-WHU[27,28]. Besides, to answer the question of how the dataset scale impacts the performance, we construct a mini dataset from the RSD46-WHU dataset. The details of the considered datasets are described as follows:

The *NWPU-RESISC45* dataset was proposed by Cheng et al. [19] in 2017 and became a popular benchmark in the RS classification research. It involves 45 categories with 700 remote scene images in each category, each with a size of 256 x 256 pixels. These aerial images are collected by experienced experts from Google Earth; the spatial resolution ranges from approximately 30 to 0.2 m per pixel. According to the split division setting proposed by Ravi et al.[46], we split the 45 categories into 25, 8, 12 for meta-training, meta-validation and meta-testing, respectively. Note that, the validation set was held-out for hyper-parameter selection of the meta-training stage. The set-split for meta-training are the same 25 categories of $\mathcal{D}_{base}$. It is further divided into three sets: *meta_train_support*, *meta_train_val*, *meta_train_query*. The number of images in each category is shown in Table 1.

Table 1. NWPU-RESISC45 Dataset-split.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

Dataset-split		# of categories	Images per category
base	meta_train_support	25	350
	meta_train_val	25	175
	meta_train_query	25	175
val	meta_validation	8	700
novel	meta_test(unseen)	12	700

The *RSD46-WHU*dataset contains 46 categories, each with images ranging from 428 to 3000, for a total of 117,000. Like many other RS datasets, the images are collected by hand from Google Earth and Tianditu, with the ground resolution spanning from 0.5m to 2m. Similar to the NWPU-RESISC45 dataset, that the 46 categories in RSD46-WHU dataset are divided into 26, 8, 12 for meta-training, meta-validation, and meta-testing, respectively. It is relevant to mention that we have dropped about 1200 images in total because some images are not in the size of 256×256 pixels or or contain incorrect content. The details of our modified dataset-split are listed in Table 2.

Table 2. RSD46-WHU Dataset-split.

Dataset-split		meta_train _support	meta_train _val	meta_train _query
base	Airplane	1515	757	757
	Airport	825	413	413
	Artificial dense forest land	1405	702	702
	Artificial sparse forest land	1414	706	707
	Bare land	501	250	250
	Basketball court	1491	745	745
	Blue structured factory building	1536	768	767
	Building	1729	865	864
	Construction site	1639	819	819
	Cross river bridge	1124	562	561
	Crossroads	1024	512	512
	Dense tall building	1534	767	767
	Dock	1574	787	786
	Fish pond	807	403	403
	Footbridge	1312	656	655
	Graff	1505	753	752
	Grassland	1416	708	708
	Low scattered building	1199	600	599
	Lrregular farmland	1568	784	784
	Medium density scattered building	526	263	262
	Medium density structured building	1773	887	886
	Natural dense forest land	1500	750	750
	Natural sparse forest land	1491	746	745
	Oiltank	805	402	402
	Overpass	1252	626	625
	Parking lot	1528	764	764
val	Plasticgreenhouse		1015	
	Playground		1913	
	Railway		3111	
	Red structured factory building		2993	
	Refinery		2657	
	Regular farmland		3209	

Table 2. RSD46-WHU Dataset-split (*cont*).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

	Dataset-split	meta_train _support	meta_train _val	meta_train _query
val	Scattered blue roof factory building		3050	
	Scattered red roof factory building		2936	
novel	Sewage plant-type-one		538	
	Sewage plant-type-two		428	
	Ship		3014	
	Solar power station		3032	
	Sparse residential area		2981	
	Square		3309	
	Steelsmelter		2933	
	Storage land		2114	
	Tennis court		1554	
	Thermal power plant		1263	
	Vegetable plot		2884	
	Water		2713	

We further conduct a new dataset *mini-RSD46-WHU* to investigate how the scale of the dataset impacts the results. The mini-RSD46-WHU dataset is formed from the RSD46-WHU dataset by randomly selecting 500 images in each category. Except for category Sewage plant-type-two only has 428 images, because that is all it holds in the original dataset. We follow the same division setting of the RSD46-WHU dataset; the only change is the number of images in each category. Table 3 shows the details.

Table 3. mini-RSD46-WHU Dataset-split.

Dataset-split		# of categories	Images per category
base	meta_train_support	26	250
	meta_train_val	26	125
	meta_train_query	26	125
val	meta_validation	8	500
novel	meta_test(unseen)	12	500

5.3. Results and Comparisons

Following the most common setting in few-shot classification, namely 5-way 1-shot, and 5-way 5-shot, we conduct experiments to evaluate our method's effectiveness. The proposed method is compared with three state-of-the-art few-shot learning algorithms and one conventional deep learning method. The three few shot methods include the ProtoNet, MAML, and RelationNet. Also, the performance of a conventional classification algorithm D-CNN is analyzed in few-shot classification scenarios.

For 5-way 1-shot experiment, one labeled *support* sample per category is randomly selected as the supervised sample at the test time. Likewise, 5 *support* samples per category are provided for 5-shot setting. Following the setting of [29], 15 *query* images per category are batched in each *episode* for evaluation. We computed the mean classification accuracy of 800 randomly generated episodes from the *novel* (meta-test) set.

On both datasets, the results of average 5-way accuracy (%) with 95% confidence interval of 1-shot and 5-shot are reported in Table 4 and Table 5. As we can see, our method outperforms the other models under both 5-way 1-shot and 5-way 5-shot settings. D-CNN shows inferior performance both in the 1-shot and 5-shot cases, and this result is reasonable due to D-CNN is not designed specifically to few-shot classification. Typical CNNs-based methods most likely lead to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

overfitting when meeting so few supervised samples, whereas meta-based methods have achieved considerable performance.

Table 4. Few-shot classification results on the NWPU-RESISC45 dataset.

Method	Backbone	1-shot	5-shot
ProtoNet[29]	Conv4	51.17% +- 0.79%	74.58% +- 0.56%
MAML[30]	Conv4	53.52% +- 0.83%	71.69% +- 0.63%
RelationNet[31]	Conv4	57.10% +- 0.82%	73.55% +- 0.56%
D-CNN[21]	ResNet12	36.00% +- 6.31%	53.60% +- 5.34%
Ours	ResNet12	**69.46% +- 0.22%**	**84.66% +- 0.12%**

Table 5. Few-shot classification results on the RSD46-WHU.

Method	Backbone	1-shot	5-shot
ProtoNet[29]	Conv4	52.57% +- 0.89%	71.95% +- 0.71%
MAML[30]	Conv4	52.73% +- 0.91%	69.18% +- 0.73%
RelationNet[31]	Conv4	55.18% +- 0.90%	68.86% +- 0.71%
D-CNN[21]	ResNet12	30.93% +- 7.49%	58.93% +- 6.14%
Ours	ResNet12	**69.08% +- 0.25%**	**84.10% +- 0.15%**

A bar chart of few-shot classification results on both datasets is shown in Figure 6. We observe that our method outperforms the other four methods by a significant margin. Similar to our method, ProtoNet and RelationNet are both metric-based methods. ProtoNet uses Euclidean distance while RelationNet compares an *embedding* f_φ and *query* samples using an additional parameterized CNN-based 'relation module.' Our method computes the class centers as the same in ProtoNet. Yet, we employ a cosine distance with a learnable scaling factor for classifying, which contributes a lot to achieve better performance. For MAML, a representative method for model initialization, we adopt a first-order approximation version for the experiments. The original paper of MAML reports that the performance of the first-order approximation is almost identical to the full version. We take the first-order approximation version for its efficiency; the performance of MAML may get narrowly enhance by the full version.

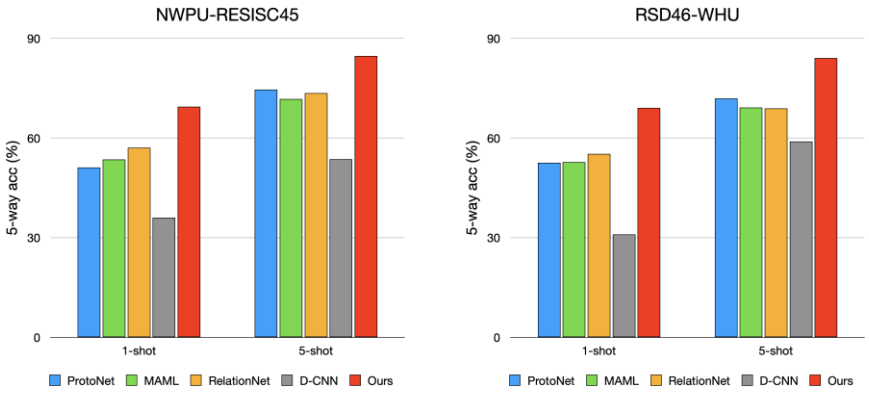

Figure 6. Few-shot classification results on the NWPU-RESISC45 dataset (left) and the RSD46-WHU dataset(right). The performance is reported with 95% confidence intervals.

An interesting phenomenon we observed is shown in Figure 7. We plot the first 90 epochs of the generalization of our model on base and novel categories. Base generalization indicates the training accuracy from unseen data in the base categories, and the novel generalization means test performance from data in novel categories. As shown, while the model achieves better performance on unseen data in the base set, the novel generalization drops instead. Why the test performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

decreases? We suppose lacking supervised data is the reason causing the over-fitting problem, which leads to this phenomenon. This problem will be discussed further in Section 5.4.

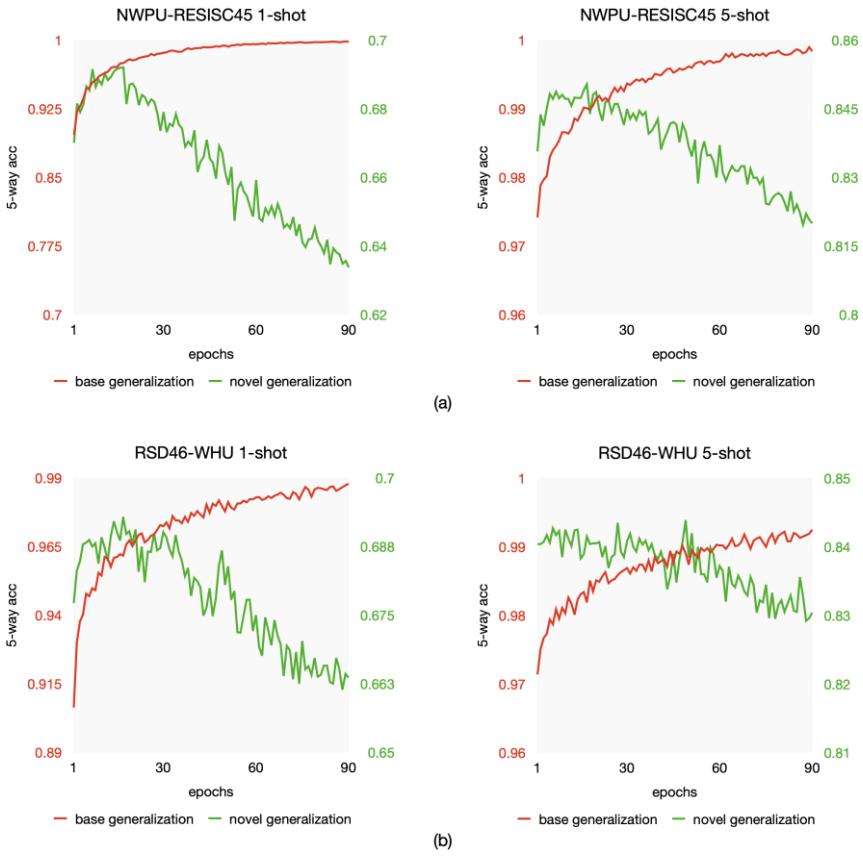

Figure 7. Generalization discrepancy in meta-learning stage.

5.4. Analysis

1. Effect of Dataset scale

To investigate how dataset scale impacts the performance, we conduct a variant of the RSD46-WHU dataset with only 500 images in each category, called mini-RSD46-WHU. The overall accuracies of 5-way 1-shot and 5-shot are reported in Table 6. We adopt the same backbone and training strategy on both datasets. As we can see, apparently, the performance improves when the scale of dataset gets larger. The overall accuracy of 5-way 1-shot and 5-shot on the original dataset increased by 6.86% and 5.78% compared to the mini dataset.

Table 6. Comparison between mini and full RSD46-WHU.

Dataset	1-shot	5-shot
RSD46-WHU	69.08 +- 0.25 (%)	84.10 +- 0.15 (%)
mini-RSD46-WHU	62.22 +- 0.25 (%)	78.32 +- 0.18 (%)

2. Effect of Shots

To further evaluate the 5-way accuracy as a function of shots, we conduct the experiments by providing our model with 1, 5, 10, 15, 20, and 25 labeled *support* samples on both datasets. The results are presented in Figure 8. As we expected, the prediction accuracy is greatly improved when the *shot* is increasing from 1 to 5. However, the performance does not benefit much more when the shot continues to increase. These findings confirm that our model is specifically effective in very-low-shot settings.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

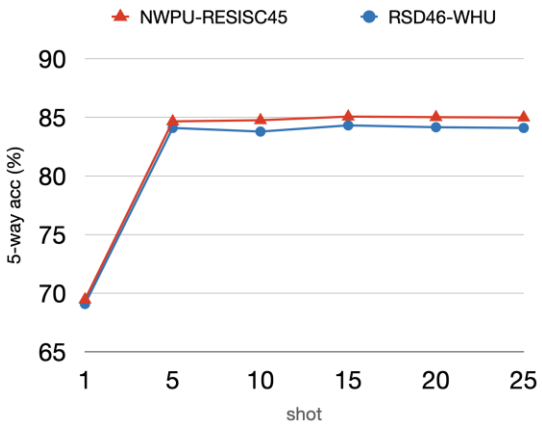

Figure 8. The effect of shots on test performance are reported with 95% confidence intervals when training on NWPU-RESISC45 and RSD46-WHU. All experiments are from 5-way classification with a ResNet-12 backbone.

From the experiments in Section 5.3, we observe from Figure 7 that the model with the best accuracy often appears in the first 40 epochs. For a further analysis of the generalization discrepancy, we plot the generalization curve with different shots on both NWPU-RESISC45 and RSD46-WHU datasets, see Figure 9 and Figure 10. As we can see, the same phenomenon appeared again: when the generalization gets better on the unseen data of base, indicating that the model learns the objective better, whereas the test performance gets worse on the novel task. In other words, this phenomenon still exists when the *support* labeled instances increases; over-fitting may not be the very reason for the test performance drops. This generalization discrepancy may be caused by the objective difference between the novel set and the base set. That is, in the meta-training stage, our model learns too specific on base-set, which has adverse effects on the novel-set. Our investigations suggest that the generalization discrepancy might be a potential challenge in few-shot learning.

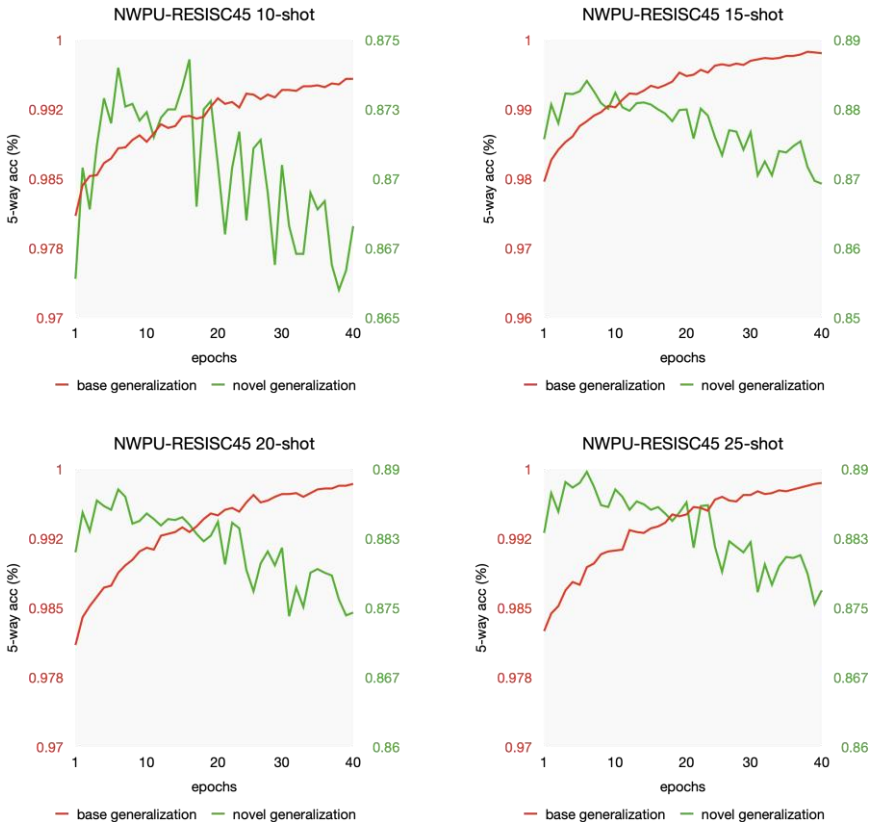

Figure 9. Generalization discrepancy on NWPU-RESISC45 dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2020 doi:10.20944/preprints202010.0033.v1

Figure 10. Generalization discrepancy on RSD46-WHU dataset.

6. Conclusion

The topic of few-shot learning has attracted much attention in recent years. In this paper, we bring few-shot learning to aerial scene classification and demonstrate that useful information may be learned from a few instances. To pursue this idea, we proposed a meta-learning framework which aims to train a model that generalizes well on unseen categories when providing a few samples. The proposed method first employs ResNet-12 to learn a representation on base-set, and then in the meta-training stage, we optimize the classifier by cosine distance with a learnable scale parameter. Our experiments, conducted on two challenging datasets, are encouraging in that our method can achieve a classification performance of around 69% for a new category by just providing one instance, besides approximately 84% for 5 support samples. Furthermore, we have conducted several ablation experiments to investigate the effects of dataset scale and support shots. At last, we observe an interesting phenomenon that there is potentially a generalization discrepancy in meta-learning. We suggest that further research in this phenomenon may be an opportunity to achieve better performance in the future.

Author Contributions: Conceptualization, P.Z., D.W. and Y.L.; Data curation, P.Z., D.W. and Y.B; Investigation, P.Z., D.W. and Y.B; Methodology, P.Z., D.W. and Y.L.; Validation, P.Z., D.W and Y.B; visualization, P.Z.; Writing—original draft, P.Z.; Writing—review & editing, P.Z. and Y.L; supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61871460; in part by the Shaanxi Provincial Key Research and Development Program under Grant 2020KW-003; in part by the Fundamental Research Funds for the Central Universities under Grant 3102019ghxm016.

Conflicts of Interest: The authors declare no competing financial interests. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

References

1. Hu, Q.; Wu, W.; Xia, T.; Yu, Q.; Yang, P.; Li, Z.; Song, Q. Exploring the Use of Google Earth Imagery and Object-Based Methods in Land Use/Cover Mapping. *Remote Sensing* **2013**, *5*, 6026-6042, doi:10.3390/rs5116026.
2. 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. *Landscape and Urban Planning* **2011**, *100*, 223-230.
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. *International Journal of Remote Sensing* **2013**, *34*, 45-59.
4. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.-S.; Zhang, L. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters* **2016**, *13*, 747-751.
5. Li, X.; Shao, G. Object-based urban vegetation mapping with high-resolution aerial photography as a single data source. *International journal of remote sensing* **2013**, *34*, 771-789.
6. Manfreda, S.; McCabe, M.F.; Miller, P.E.; Lucas, R.; Pajuelo Madrigal, V.; Mallinis, G.; Ben Dor, E.; Helman, D.; Estes, L.; Ciraolo, G. On the use of unmanned aerial systems for environmental monitoring. *Remote sensing* **2018**, *10*, 641.
7. 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 Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3965-3981.
8. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **2004**, *60*, 91-110.
9. Manjunath, B.S.; Ma, W.-Y. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence* **1996**, *18*, 837-842.
10. Swain, M.J.; Ballard, D.H. Color indexing. *International journal of computer vision* **1991**, *7*, 11-32.
11. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems; pp. 270-279.
12. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06); pp. 2169-2178.
13. Jegou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Perez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence* **2011**, *34*, 1704-1716.
14. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained linear coding for image classification. In Proceedings of 2010 IEEE computer society conference on computer vision and pattern recognition; pp. 3360-3367.
15. Bosch, A.; Zisserman, A.; Muñoz, X. Scene classification via pLSA. In Proceedings of European conference on computer vision; pp. 517-530.
16. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *Journal of machine Learning research* **2003**, *3*, 993-1022.

17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of Advances in neural information processing systems; pp. 1097-1105.
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
19. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* **2017**, *105*, 1865-1883.
20. Nogueira, K.; Penatti, O.A.; Dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition* **2017**, *61*, 539-556.
21. 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 transactions on geoscience and remote sensing* **2018**, *56*, 2811-2821.
22. Li, J.; Lin, D.; Wang, Y.; Xu, G.; Zhang, Y.; Ding, C.; Zhou, Y. Deep discriminative representation learning with attention map for scene classification. *Remote Sensing* **2020**, *12*, 1366.
23. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *57*, 2290-2304.
24. Marcus, G.F. Rethinking eliminative connectionism. *Cognitive psychology* **1998**, *37*, 243-282.
25. Xia, G.-S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maître, H. Structural high-resolution satellite image indexing.
26. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing* **2018**, *145*, 197-209.
27. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 2486-2498.
28. Xiao, Z.; Long, Y.; Li, D.; Wei, C.; Tang, G.; Liu, J. High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective. *Remote Sensing* **2017**, *9*, 725.
29. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of Advances in neural information processing systems; pp. 4077-4087.
30. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400* **2017**.
31. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp. 1199-1208.
32. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing* **2015**, *7*, 14680-14707.
33. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 1-9.
34. Zhao, W.; Du, S. Scene classification using multi-scale deeply described visual words. *International Journal of Remote Sensing* **2016**, *37*, 4119-4131.
35. Wang, G.; Fan, B.; Xiang, S.; Pan, C. Aggregating rich hierarchical features for scene classification in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2017**, *10*, 4104-4115.
36. Lu, X.; Ji, W.; Li, X.; Zheng, X. Bidirectional adaptive feature fusion for remote sensing scene classification. *Neurocomputing* **2019**, *328*, 135-146.

 doi:10.20944/preprints202010.0033.v1

37. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of European conference on computer vision; pp. 499-515.
38. Nichol, A.; Achiam, J.; Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* **2018**.
39. Rusu, A.A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; Hadsell, R. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960* **2018**.
40. Sun, Q.; Liu, Y.; Chua, T.-S.; Schiele, B. Meta-transfer learning for few-shot learning. In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 403-412.
41. Jamal, M.A.; Qi, G.-J. Task agnostic meta-learning for few-shot learning. In Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp. 11719-11727.
42. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. In Proceedings of Advances in neural information processing systems; pp. 3630-3638.
43. Chen, Y.; Wang, X.; Liu, Z.; Xu, H.; Darrell, T. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390* **2020**.
44. Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C.F.; Huang, J.-B. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232* **2019**.
45. Gidaris, S.; Komodakis, N. Dynamic few-shot visual learning without forgetting. In Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp. 4367-4375.
46. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. **2016**.