Momentum Contrastive Learning for Few-Shot Classification and Segmentation in Remote Sensing

Anonymous submission

Abstract

Classifying and segmenting patterns from a few examples is a key problem in remote sensing and earth observation, as acquiring large amounts of accurately labeled, and ground-truthed data is difficult. Prior works show that meta-learning, based on episodic training on query and support sets, is a promising approach. Yet, direct fine-tuning techniques have drawn scant attention. The objective of this paper is to repurpose contrastive learning as a pretraining method for few-shot learning for both classification and semantic segmentation tasks for remote sensing. We find that fine-tuning of embeddings learned from contrastive methods is crucial to the few-shot learning tasks. With only a few labeled samples, such a simple approach outperforms supervised learning methods. We evaluate our approach on two key remote sensing datasets: Agriculture-Vision and EuroSAT. Combining these contrastive methods with only a few labeled examples, our approach outperforms purely supervised training on the nearly 95,000 images in Agriculture-Vision on both classification and semantic segmentation tasks. Similarly, the proposed few-shot method achieved better results on the land-cover classification task on EuroSAT compared with supervised model training on the fully supervised dataset.

Introduction

Remote sensing (RS) and earth observation (EO) imagery enables the detection and monitoring of critical societal challenges including food security, natural disasters, clean water availability, hunger and poverty, impact of climate change, threats to animal habitats, geopolitical risk, and more. Like other domains, RS-EO has benefited from the significant advances in machine vision systems over the past decade. However, these algorithms' ability to learn without abundant labeled data is far from satisfactory, even when compared to a toddler (Landau, Smith, and Jones 1988; Samuelson and Smith 2005). This fact severely limits the scalability of learned models, with various categories following the long-tail distribution in the real world. The lack of labeled samples is even more severe for RS-EO tasks, as data acquisition often involves concerns around security, ethics, resources, accessibility, and cost (Yang et al. 2022). Furthermore, annotation of remote sensing data often requires high levels of expert knowledge and true ground-truth verification (i.e. physically traveling to locations to confirm predictions). Therefore developing machine-learning applications from RS-EO data to address key societal challenges is often thwarted by the domain's particularly laborious dataset-creation process (Sun et al. 2021; Paliyam et al. 2021).

Inspired by human's highly efficient learning ability, research around learning from unlabeled data, i.e. unsupervised or self-supervised learning, and the ability to generalize from only a few examples, i.e. few-shot learning, have become key areas of interest in the machine learning community (Vinyals et al. 2016; Finn, Abbeel, and Levine 2017; Sung et al. 2018; Gidaris and Komodakis 2018; Sun et al. 2021; Alajaji et al. 2020; Li, Deng, and Fang 2021). Few-shot learning aims to realize the knowledge adaption of embeddings from label-abundant data to label-scarce classes. While the adapted representation aims to discriminate different levels of information (e.g., instance level and semantic level) between classes, the embedding should be invariant to common, irrelevant variations of the image, including different sizes, deformations, and lighting. The question is then: How can we learn a representation invariant to common factors while maintaining differences for diverse classes with limited labels?

As a prevailing and advancing research topic, contrastive learning has demonstrated impressive results on various downstream learning tasks. These methods seek to learn a transferable representation by strongly augmenting large quantities of raw data and pulling views of the same image close together while pushing differing images apart. These methods are often evaluated on the performance of downstream tasks fine-tuned on different fractions of the supervised dataset. However, the focus is rarely on the one-shot or few-shot cases for extremely limited datasets. As raw RS-EO data is highly abundant but ground-truth data is extremely scarce, leveraging constrastive methods for few-shot learning offers a key opportunity in this domain. Additionally, most common contrastive learning and few-shot methods were developed for natural scene imagery; e.g. (He et al. 2020; Chen et al. 2020a; Grill et al. 2020; Zbontar et al. 2021) show that as the statistics of that domain (both source imagery and targets) are extremely different from RS-EO data, there is no guarantee that the same benefit will be observed without adaptation. Therefore, we investigate the improvement in performance of few-shot learning in RS-EO classification and semantic segmentation tasks using

contrastive-learning based pretraining.

Specifically, we focus on pretraining from the Extended Agriculture Vision dataset (AV+) (Wu et al. 2022), which includes high-resolution aerial imagery over agricultural lands in the US Midwest. Obtaining ground-truth annotations for agriculture is particularly challenging due to patterns of interest being small in size, high in number, and often possessing ambiguous boundaries; the ability to identify patterns from only a small number of samples addresses key challenges in precision agriculture and food security.

Inspired by the work of (Gidaris and Komodakis 2018; Qi, Brown, and Lowe 2018; Wang et al. 2020), we adopt a two-stage training scheme: improved momentum contrast, MoCo-V2 (Chen et al. 2020c) pretraining followed by fine-tuning. This pretraining stage should enable the backbones to encode spatially invariant features. Then, we add a final layer, for classification tasks, or a decoder, for semantic segmentation task. During the fine-tuning stage, we fix all backbones parameters and train only on the classification-layer/decoder. Additionally, we apply instance-level feature normalization to the last layer in the classification task (Gidaris and Komodakis 2018; Qi, Brown, and Lowe 2018; Wang et al. 2020) .

We find that the embeddings pretrained from AV+ under this protocol show better adaptability when compared to counterparts pretrained on ImageNet (Deng et al. 2009) and COCO (Lin et al. 2014). Our method outperforms pretrained ImageNet weights by 1 to 6 points on Agriculture-Vision and EuroSAT classification tasks under the same number of supervised training samples. Similarly our embeddings deliver a 5 to 7 point improvement on mIoU compared with embeddings learned from COCO on the Agriculture-Vision semantic segmentation task.

In the meantime, we demonstrate the high learning efficiency of the proposed method for RS-EO imagery. With a few labeled images, we find that the contrastive-learning-based fine-tuning approach shows comparable or even better results under different tasks and datasets when compared with supervised models trained on full labeled data samples; our proposed approach shows matching performance with less than 0.01 percent of labeled data.

In summary, the contributions of this paper can be summarized: (1) We leverage contrastive learning for both classification and segmentation few-shot learning tasks using remote sensing imagery. (2) We demonstrate the successful adaptation of the two-stage contrastive-learning-based few-shot strategy for RS-EO data. (3) We empirically show that instance-level feature normalization benefits classification tasks in RS-EO images. (4) Extensive experiments show that our approach allows us to competitively identify key agricultural and land cover patterns with only a small amount of labeled data. (5) We demonstrate that pretraining on AV+, a high-resolution multi-spectral RS-EO dataset, provides strong benefit to other RS-EO tasks on lower-resolution data such as EuroSAT.

Related Work

Few-Shot Learning

Efficient adaption algorithms have been developed for various few-shot learning tasks such as classification (Fei-Fei, Fergus, and Perona 2006), object detection (Wang et al. 2020), semantic segmentation (Wang et al. 2019), and robot learning (Finn et al. 2017). Generally, previous works can be roughly cast into three categories: metric-based, optimization-based, and hallucination-based.

The key idea of metric-based approaches is to learn good embeddings with appropriate kernels. Previous results from (Koch et al. 2015) propose applying a siamese neural network for few-shot classification. Following that, (Sung et al. 2018) presents a Relation Network by replacing the L1 distance between features with a convolutional neural network (CNN)-based classifier and updating the mean squared error (MSE) with cross-entropy; the triplet loss is utilized to improve the model's performance (Cacheux, Borgne, and Crucianu 2019). (Gidaris and Komodakis 2019) further adds extra self-supervised tasks to enhance generalization capacity.

Optimization-based methods aim to learn through gradient backpropagation. Representative works include MAML (Finn, Abbeel, and Levine 2017), which realizes quick adaption from a good initialization spot, Reptile (Nichol and Schulman 2018), which simplifies the learning process of MAML, and MetaOptNet (Lee et al. 2019), which incorporates the support vector machine (SVM) as a classifier.

Hallucination-based methods seek to learn generators to generate unseen samples. Works from (Wang et al. 2018; Zhang, Zhang, and Koniusz 2019; Li et al. 2020a) show that such a strategy of hallucination improves the test results and enhances the generation of models.

Few-Shot learning for RS-EO has received increased attention in recent years, (Sun et al. 2021). While much of the work is focused on scene classification (Alajaji et al. 2020; Li et al. 2020b; Alajaji and Alhichri 2020; Liu et al. 2018), other recent approaches examine semantic segmentation tasks (Wang et al. 2021; Kemker, Luu, and Kanan 2018; Yao et al. 2021).

Contrastive Learning

Contrastive learning is widely used for pretraining without labeled data and shows superior performance in various self-supervised learning (SSL) tasks (Chen et al. 2020c; Grill et al. 2020; Chen et al. 2020a,b). Specifically, contrastive approaches train architectures by bringing the representation of different views of the same image closer together while spreading representations of views from different images apart. The success of training may rely on large batch sizes (Chen et al. 2020a), a momentum-update memory bank (He et al. 2020), projection heads (Chen et al. 2020c), and/or stop-gradient trick (Chen and He 2021). Barlow Twins (Zbontar et al. 2021) further proposes a new contrastive learning objective without using the trick of stop-gradient, which also brings equivalent results. Among all these methods, MoCo-V2 is one of the most widely used frameworks, given its memory efficiency and promising performance(Chen et al. 2020c). Within the field of RS-EO

specifically, (Manas et al. 2021) takes MoCo-V2 as the basis and uses multiple projection heads to capture desired invariance to seasonality. Therefore, we continue to utilize MoCo-V2 as a pretraining protocol in this work.

Pairing Supervised and Self-Supervised Learning

Supervised and self-supervised learning have been explored jointly in numerous prior works. Some methods utilize the SSL loss as supplemental losses during the supervised training process (Gidaris et al. 2019; Su, Maji, and Hariharan 2020). Often, additional efforts are needed to calibrate (i.e. re-weight) these losses when crossing different domains. More straightforward and effective methods come from supervised fine-tuning (Doersch, Gupta, and Zisserman 2020). While SSL encourages the learning of general-purpose features, the adaption of features on the new task can be realized with only a few labeled samples. Within RS-EO, very recent work has looked to combine SSL and few-shot learning for scene classification (Zeng and Geng 2022) and segmentation (Li et al. 2022).

Method

To start with, we pretrain different backbones with momentum contrastive learning on unlabeled data; to be more specific, there are roughly 1,300,000 images. These images are all randomly cropped from the 3600 raw images from AG+. Additional information for data pre-processing and the dataset are available in the supplementary. Since MoCo-V2 is unsupervised, there is no information on any base classes, unlike the usual settings for few-shot learning. Only k samples are provided for fine-tuning during the evaluation, where k varies from 1-10. The evaluation will include both base and novel classes based on different downstream tasks with the goal to optimize the classification accuracy or mIoU of agricultural patterns on Agriculture-Vision and land covers on EuroSAT.

Pretraining with Momentum Contrast

In this section, we present the pretraining stage of our framework. Concretely, our pretraining is based on MoCo-V2. MoCo-V2 is trained with natural scene images that only contain information about red, green, and blue channels. However, AV+ has extra information in the NIR channel. To fully explore knowledge from the pretraining dataset, we further add one channel to the backbones following the work from (AV+ paper).

In every training epoch of MoCo, a training sample is augmented into two different views named as query x^q and key x^k. These views contain variations introduced from data augmentations, including spatial and color transforms. With an online network and a momentum-updated offline network, the training encourages these two views to be mapped into two similar embedding spaces, i.e., q, k, as a positive pair.

MoCo offers two critical properties for avoiding model collapse (Jing et al. 2021) in contrastive learning. First, a queue data structure is proposed to store a rich set of negative features. This large dictionary enables the training without using large batch sizes. Second, MoCo stabilizes the training stage with a momentum-update strategy instead of using back-propagation. While query features are encoded from the gradient-based backbone, keys are all mapped from the momentum encoder, as shown in Figure 1.

Together, based on positive and negative pairs and a temperature parameter τ for scaling, the training loss function, i.e., InfoNCE (Oord, Li, and Vinyals 2018), is then defined as follows:

$$\mathcal{L} = -\log \frac{\exp(q \cdot k^+/\tau)}{\sum_{k^-} \exp(q \cdot k^-/\tau) + \exp(q \cdot k^+/\tau)} \quad (1)$$

Fine-tuning Approach

Given the strong data augmentations in contrastive learning, the backbone features encoded should be class-agnostic. Ideally, it should adapt to different data classes and types of downstream tasks without much learning effort. Therefore, the critical step of the proposed method is to separate representation learning and downstream task learning into two stages.

With contrastive learning applied in the first stage, we fine-tune models with a few labeled images. We first create a small balanced training set with K images per class, i.e., K shots. These classes can either be seen or be novel classes. In the few-shot classification task, we add one fully connected layer to the backbones without introducing extra nonlinearity. We assign randomly initialized weights to the added classification layer. The intuition to model the classifier in such a naive way is due to the two fundamental properties of features from contrastive learning, alignment, and uniformity(Wang and Isola 2020). Generally, alignment indicates that similar views are close to each other in the embedding space. Uniformity prefers a uniform distribution if features are mapped on a unit hyper-sphere. Therefore, features are linearly separable when classes are well-clustered, as shown in Figure2.

We also consider using cosine similarity between features and weights inspired by work from (Wang et al. 2020; Gidaris and Komodakis 2018). For each class, the per-class weights can be noted as w_t, where t is the index for the classes. Therefore, the weight matrix for all total classes can be formed as [w_1,w_2,...,w_t] and noted as $W \in R^{d \times t}$, where d is the feature dimension. Then, the final similarity score S is computed by the dot product of input feature $F(x)$ and the weight vectors. More specifically, the entries in S are defined as

$$s_{i,j} = \frac{\alpha F(x)_i^{\mathsf{T}} w_j}{\|F(x)_i\| \, \|w_j\|}, \quad (2)$$

which indicates the similarity score between the proposal of the i-th pattern(class) and the weights vector of class j, and α is a hyper-parameter for scaling, which is consistently set to 1 in all experiments. Based on the results in the following sections, we empirically find that the instance-level normalization with such a cosine similarity metric enhances the stability and final performance of the classifier in the classification task on Agriculture-Vision and EuroSAT.

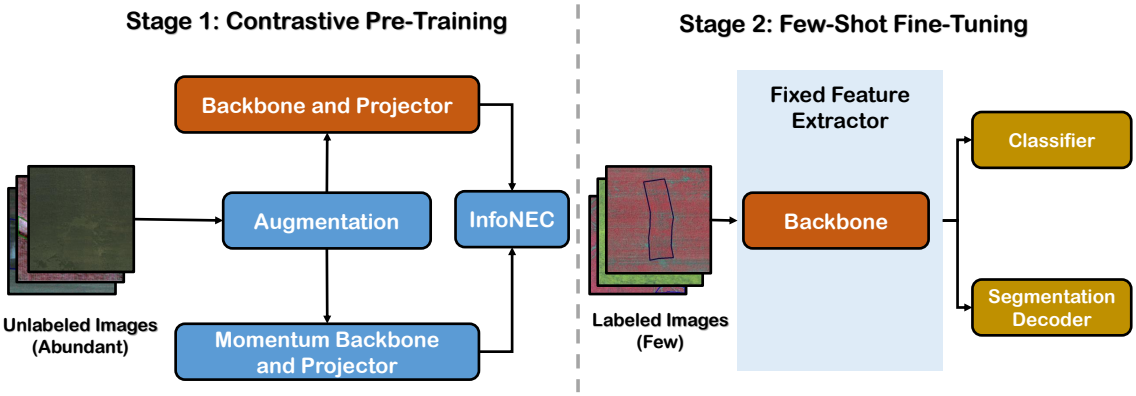

Figure 1: Illustration of the proposed two-stage learning approach. In the pretraining stage, we follow the training strategy of contrastive learning to train the backbones and projectors jointly with abundant unlabeled images. In the fine-tuning stage with limited labeled images, we adapt the pretrained backbone from stage1 for feature extraction. The feature extractors are fixed, and fine-tuning is only involved in the classifier and segmentation decoder.

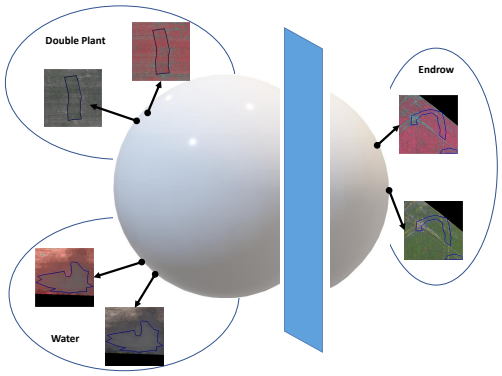

Figure 2: Features learned from momentum contrast learning are uniformly distributed and linearly separable on a unit hyper-sphere.

In the few-shot semantic segmentation task, we choose the lightweight segmentation model U-Net (Ronneberger, Fischer, and Brox 2015) for fine-tuning given limited training samples. Concretely, we add a five-layer decoder based on the encoders (backbones) pretrained from contrastive learning. Empirically, this asymmetric encoder-decoder shows exceptional performance in the segmentation task on Agriculture-Vision.

Experiments and Results

In this study, we conduct various experiments to prove the effectiveness of our proposed methods. We start with the preliminaries of the evaluation method and then give the detailed results in the following sections.

Preliminaries for Evaluation

We evaluate the proposed method from two perspectives, i.e., the quality of embeddings and the required amount of labeled data for model adaptation. First, in the few-shot classification task, we compare the performance of features from the backbone pretrained on ImageNet, and the backbone pretrained on AV+ using MoCo-V2. Similarly, in the few-shot semantic segmentation task, embeddings pretrained on AV+ using MoCo-V2 and embeddings from COCO's weights are compared. Second, to illustrate the learning efficiency of our method, we compare the two models' performances, one of which is fine-tuned on a few samples, and the other is trained in a supervised way with complete labeled data.

Pretraining on Extended Agriculture-Vision

In the first stage, we apply momentum contrast learning to pretrain backbones as shown in the left part of Figure1.

Extended Agriculture-Vision For contrastive pretraining, we use the large-scale remote-sensing dataset, Extend Agriculture-Vision (Wu et al. 2022). While the original Agriculture-Vision dataset contained only 512x512 tiles with semantic segmentation labels for agricultural patterns such as waterways, weeds, nutrient deficiency, etc., AV+ includes several thousand additional raw full-field images (upwards of 10,000 x 10,000 in dimension). Images consist of RGB and Near-infrared (NIR) channels with resolutions as high as 10 cm per pixel. As it also covers data that varies from 2017 to 2020, encoders pretrained on this dataset should capture remote sensing, agriculture, and temporal features. Therefore, the embeddings pretrained on Extend Agriculture-Vision should be adapted well to diverse downstream tasks such as agricultural pattern recognition and land-cover classification. Additional details of this dataset are included in the Supplemental.

Pretraining Implementation Details We adopt MoCo-V2 for pretraining, given its promising performance and memory efficiency. We use different sizes of ResNet as the encoder and two layers of MLP as a projector. Following the original work, each embedding has 128 dimensions, and there are 16,384 negative keys stored in the memory bank. We pretrain MoCo for 200 epochs using an SGD optimizer,

the learning rate of 0.3, and weight decay of 0.0001. The learning rate is adjusted to 0.03 and 0.003 at 140th and 160th epochs accordingly. As this data is hyperspectral, we add one more channel to the encoder during the training for the NIR input, and this extra channel is initialized with the same weights as the red channel.

Few-shot Learning on Agriculture-Vision

In the second stage, we conduct few-shot experiments on classification and semantic segmentation tasks on Agriculture-Vision shown in the right part of Figure 1.

Agriculture-Vision Agriculture-Vision (AV) is a large aerial image database for agricultural pattern analysis. It contains 94,986 high-quality images over 3432 farmlands across the US. Totally, there are nine classes selected in the dataset under the advisement of agronomists, which include double plant, dry-down, endrow, nutrient deficiency, planter skip, storm damage, water, waterway, and weed cluster. With extreme label imbalance across categories, it is a challenge to train well-performance model models for classification and segmentation tasks (Chiu et al. 2020a). The original dataset is designed for semantic segmentation; we also create a "classification" version of the dataset by assigning a positive label if any presence of that class is included in the tile.

Few-shot Classification on Agriculture-Vision The first set of experiments focuses on the classification formulation of the Agriculture-Vision task. We use ResNet-18, ResNet-50 and ResNet-101 as the backbones for fine-tuning. All backbones are fixed, except the last fully connected layer which is learnable. Different from the optimization methods used in (He et al. 2020), we use Adam as an optimizer for all experiments with an initial learning rate set to 0.001. We train the classification models for 100 epochs with a batch size of 64.

Table 1: Comparison of fine-tuning results between weights pretrained on supervised ImageNet and weights from our MoCo on AV+ for the 10-shot classification

Pretrained Weight	Backbone	Accuracy(%)
ImageNet	ResNet-18	55.22
MoCo on AV+	ResNet-18	**65.51**
ImageNet	ResNet-50	54.53
MoCo on AV+	ResNet-50	**64.82**
ImageNet	ResNet-101	54.34
MoCo on AV+	ResNet-101	**64.62**

We first prove the quality and adaptability of pretrained embeddings from the proposed methods. As shown in the Table 1, our pretrained weights show significantly better results than those from ImageNet, with over 10 points improvement on average. These results prove the adaptability of embeddings encoded from our pretrained weights and better generalization capacity in this few-shot classification task for agricultural patterns. The best result is obtained from ResNet-18, instead of the larger ResNet-50 or ResNet-101. With only 10 shots, this observation is due to the last layer attached to ResNet-18 being smaller than the fully connected layers in larger backbones. However, all AV+ pretrained weights enable better performance than any ImageNet weights.

Table 2: Comparison of the classification task between the 10-shot results of the proposed method and end-to-end training using the full Agriculture-Vision on ResNet-18.

Pretrained Weight	Freeze Backbone	Number of Images	Accuracy (%)
Random	False	9,000	57.30
Random	False	94,986	62.31
MoCo on AV+	True	**10**	**65.51**

Table 3: 9-way few-shot classification accuracy on Agriculture-Vision.

Backbone \ Shots	Accuracy 10 shots (%)	Accuracy 5 shots (%)	Accuracy 1 shot (%)
ResNet-18	**65.51**	**61.61**	16.72
ResNet-50	64.82	59.44	**29.64**
ResNet-101	64.62	59.56	28.84

Next, we continue to demonstrate the learning efficiency of our model by comparing it with the model trained with 94,986 labeled images. For models training in an end-to-end manner, there is a noticeable drop once we reduce the number of models for training. However, as shown in the Table 2, the proposed method outperforms model training with numerous images with little computation and much fewer labels for agricultural pattern classification. This observation is important as it illustrates the potential of training diverse deep learning tasks in agriculture and remote sensing with minimum effort but still providing satisfactory results.

Table 3 demonstrates the 9-way few-shot classification results with different sizes of backbones. All results are averaged from 3 trials and use the same training setup for a fair comparison. While ResNet-18 gives the best results when trained with five shots or ten shots, ResNet-50 shows the best results when there is only one labeled sample for each class. The performance of ResNet-50 and ResNet-101 are very similar. Generally, favorable results can be acquired when the number of shots is five or greater.

Few-shot Segmentation on Agriculture-Vision Agriculture-Vision provides dense pixel-level labels for semantic segmentation. Therefore, we also evaluate our proposed method in the semantic segmentation task, which is often given less attention than classification tasks. Following the work from (Chiu et al. 2020b), we ignore storm damage annotations when performing evaluations due to its extreme scarcity. Similar to the fine-tuning strategy

Table 4: Comparison of the segmentation task between the 10-shot results of the proposed method and end-to-end training using the full Agriculture-Vision on ResNet-18 and ResNet-50.

Pretrained Weight	Backbone	Freeze Backbone	Number of Images	mIoU - 8 Classes
Random	ResNet-18	False	9000	19.02
Random	ResNet-18	False	94986	21.37
MoCo on AV+ (ours)	ResNet-18	True	**10**	**23.56**
Random	ResNet-50	False	9000	19.58
Random	ResNet-50	False	94986	21.82
MoCo on AV+ (ours)	ResNet-50	True	**10**	**23.00**

we used in the classification task, we freeze all backbones during training, but with a learnable five-layer decoder. The decoders are randomly initialized and attached to the encoders, forming a lightweight and imbalanced U-Net. We use the AdamW optimizer with the learning rate set to 6e-5 and the one learning rate cycle scheduler proposed by (Smith and Topin 2019). In total, we train the segmentation models for 100 epochs with 300 steps per epoch. For all experiments, we use a batch size of 64 during fine-tuning.

Again, we report the quality and the efficient adaptability of pretrained features from the proposed methods in this segmentation task. Since U-Net's structures contain skip connections from different layers (Ronneberger, Fischer, and Brox 2015), we don't evaluate a single embedding but features from different scales. Concretely, features from our two-stage fine-tuning and features from encoders pretrained on COCO are compared using the mean intersection over union (mIoU) metric. As reported in the Table 5, our proposed method shows around 6-8 points of improvement compared with weights pretrained on COCO. Consistent with the results from the classification task, the best mIoU is reached by ResNet-18 with a smaller decoder attached. The other conclusion we can draw is that the feature distribution pretrained from natural images (COCO) and remote sensing images (AV+) is significantly different. Therefore, we can observe a noticeable improvement in the results pretrained on AV+.

We also study the learning efficiency of the segmentation task in this section. With only ten sampled images per category, we use results of the two-stage fine-tuning technique and the results from models training on the full Agriculture-Vision for comparison. While our approach fixes the backbone, we unfreeze the segmentation model's encoder training on the full dataset. Based on the results of mIoU in the Table4, there is an improvement of 2.19 points and 1.18 points for ResNet-18 and ResNet-50, respectively. While few-shot segmentation still surpasses models trained with a large number of labeled images, we notice that the gain is not salient compared with the improvement in the classification task. This observation is likely because the decoders used for segmentations have more parameters to be tuned than a single-layer classifier. With limited labeled samples, smaller models avoid overfitting, thus showing better results. Therefore, ResNet-18 shows the best results in this few-shot segmentation task.

Table 5: Comparison of fine-tuning results between weights pretrained on COCO and weights from our MoCo on AV+ for the 10-shot semantic segmentation task.

Pretrained Weight	Backbone	mIoU - 8 Classes
COCO	ResNet-18	15.61
MoCo on AV+	ResNet-18	**23.56**
COCO	ResNet-50	15.60
MoCo on AV+	ResNet-50	**23.00**
COCO	ResNet-101	15.19
MoCo on AV+	ResNet-101	**21.04**

Few-shot Learning on EuroSAT

We additionally illustrate that embeddings learned from momentum contrastive learning on AV+ help few-shot learning tasks in the more general remote sensing community. To achieve this, we evaluate our proposed method on the few-shot classification task of EuroSAT (Helber et al. 2019). Following experiments in previous sections, we evaluate the quality and adaptability of pretrained features from the proposed methods in this land-cover classification task.

EuroSAT EuroSAT is a dataset for the classification task of land use and land cover. All the satellite images are collected from Sentinel-2, covering 34 countries. There are 27,000 images in total with ten types of labels corresponding to different land use cases. The class labels are evenly distributed, with each category consisting of 2,000 to 3,000 images. We use the split method for training and evaluation following the work from (Manas et al. 2021; Helber et al. 2019).

Few-shot Classification on EuroSAT Likewise, we add a one-layer network to build the classifier for EuroSAT. We train the model for 100 epochs with an AdamW optimizer and a batch size of 256. The initial learning rate is 0.001.

Our results show that two-stage few-shot learning still leads to better embeddings on this remote sensing dataset. As seen in Table 6, features from MoCo improve one percent of accuracy on average compared to the features trained from ImageNet. Since EuroSAT shares much less similarity with our pretrained dataset, i.e., AV+, the improvement is moderate. However, the gain is still stably earned crossing different sizes of backbones. The most pleasing result is

Table 6: Comparison of fine-tuning results between weights pretrained on supervised ImageNet and weights from our MoCo on EuroSAT for the 10-shot classification

Pretrained Weight	Backbone	Accuracy(%)
ImageNet	ResNet-18	66.90
MoCo on AV+	ResNet-18	**67.92**
ImageNet	ResNet-50	65.01
MoCo on AV+	ResNet-50	**66.11**
ImageNet	ResNet-101	63.34
MoCo on AV+	ResNet-101	**64.79**

achieved by ResNet-18 again.

Table 7: Comparison of the classification task between the 10-shot results of the proposed method and end-to-end training using the full EuroSAT on ResNet-18.

Pretrained Weight	Freeze Backbone	Number of Images	Accuracy (%)
Random	False	2,700	58.81
Random	False	27,000	63.24
MoCo on AV+	True	10	**66.90**

In the experiments on label efficiency, we continue to compare the few-shot classification models with those models randomly initialized and trained on 27,000 labeled images. The proposed method outperforms the end-to-end model by 3.66 points in this classification task with only ten labeled images, as shown in the Table 8. This result is crucial as it proves the effectiveness of our methods across different domains. With remarkably cheap effort for labeling, it re-verifies the vast possibility of deploying our models to various downstream tasks in agriculture and remote sensing.

Table 8: 10-way few-shot classification accuracy on EuroSAT.

Backbone \ Shots	Accuracy 10 shots (%)	Accuracy 5 shots (%)	Accuracy 1 shot (%)
ResNet-18	**67.92**	**63.20**	**11.50**
ResNet-50	65.01	59.40	11.21
ResNet-101	63.70	58.70	11.40

We show the results of the 10-way few-shot classification on EuroSAT in the following Table 8. For a complete and fair comparison, we report the performance of backbones with different sizes and average results over experiments with three random seeds. More specifically, the ResNet-18 shows the most satisfactory performance crossing various backbones and shots. While we can notice a 1 to 4 points drop in accuracy when we increase the size of backbones under 10 or 5 shots settings, one-shot classification shows very similar performance regardless of encoder sizes. This observation is likely because the information is too limited for the adaptation of embeddings.

Ablation Study: Cosine Similarity

We use ResNet-18 as the backbone and conduct experiments of the few-shot classification on Agriculture-Vision for this ablation study. With cosine similarity applied between features and weights, we observe a noticeable improvement in the accuracy. As shown in the Table 9, this observation is especially true under the low-shot settings, i.e., 1 and 5 shots. We also explore the value of the scaling factor α with this hyper-parameter set to 0.1, 1, 5, and 10. All the evaluation settings are the same for fair comparisons.

Table 9: 10-way few-shot classification accuracy on EuroSAT.

Backbone \ Shots	Accuracy 10 shots (%)	Accuracy 5 shots (%)	Accuracy 1 shot (%)
ResNet-18	67.31	62.16	9.43
ResNet-18 (α = 0.1)	67.34	62.41	9.55
ResNet-18 (α = 1)	**67.92**	**63.20**	**11.50**
ResNet-18 (α = 5)	67.55	62.99	11.12
ResNet-18 (α = 10)	67.47	62.63	10.96

Conclusion

In this work, we propose a simple yet effective two-stage fine-tuning approach for few-shot classification and segmentation on remote sensing and earth observation data. The method performs favorably due to the remarkable adaptability of embeddings from pretraining and high learning efficiency. Most importantly, our approach provides an alternative direction to solve the notorious labeling challenge in agriculture and remote sensing domains. With such a few-shot contrastive learning-based approach, we see the possibility of deploying models in the real world with minimal labeled data and training effort.

Similarly, visual search and active learning are closely related to few-shot learning. Because physical ground-truthing is often required, identifying key locations to annotate and subsequently inspect could be enhanced with this approach to minimize the collection burden. Such efforts would be especially important when addressing time-sensitive challenges such as responding to drought, fire, flood, storm damage, illegal deforestation, or pestilence. Therefore, we hope to see further exploration and study of this work.

References

Alajaji, D.; Alhichri, H. S.; Ammour, N.; and Alajlan, N. 2020. Few-shot learning for remote sensing scene classification. In *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, 81–84. IEEE.

Alajaji, D. A.; and Alhichri, H. 2020. Few shot scene classification in remote sensing using meta-agnostic machine. In

2020 6th conference on data science and machine learning applications (CDMA), 77–80. IEEE.

Cacheux, Y. L.; Borgne, H. L.; and Crucianu, M. 2019. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10333–10342.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chiu, M. T.; Xu, X.; Wang, K.; Hobbs, J.; Hovakimyan, N.; Huang, T. S.; and Shi, H. 2020a. The 1st Agriculture-Vision Challenge: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chiu, M. T.; Xu, X.; Wei, Y.; Huang, Z.; Schwing, A. G.; Brunner, R.; Khachatrian, H.; Karapetyan, H.; Dozier, I.; Rose, G.; et al. 2020b. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2828–2838.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33: 21981–21993.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Finn, C.; Yu, T.; Zhang, T.; Abbeel, P.; and Levine, S. 2017. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, 357–368. PMLR.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2019. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8059–8068.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4367–4375.

Gidaris, S.; and Komodakis, N. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21–30.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Jing, L.; Vincent, P.; LeCun, Y.; and Tian, Y. 2021. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Kemker, R.; Luu, R.; and Kanan, C. 2018. Low-shot learning for the semantic segmentation of remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(10): 6214–6223.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 0. Lille.

Landau, B.; Smith, L. B.; and Jones, S. S. 1988. The importance of shape in early lexical learning. *Cognitive development*, 3(3): 299–321.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10657–10665.

Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; and Tao, C. 2022. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Li, K.; Zhang, Y.; Li, K.; and Fu, Y. 2020a. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13470–13479.

Li, L.; Han, J.; Yao, X.; Cheng, G.; and Guo, L. 2020b. DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9): 7844–7853.

Li, X.; Deng, J.; and Fang, Y. 2021. Few-shot object detection on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; and Wang, R. 2018. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4): 2290–2304.

Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; and Rodriguez, P. 2021. Seasonal contrast: Unsupervised pretraining from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9414–9423.

Nichol, A.; and Schulman, J. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3): 4.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paliyam, M.; Nakalembe, C. L.; Liu, K.; Nyiawung, R.; and Kerner, H. R. 2021. Street2Sat: A Machine Learning Pipeline for Generating Ground-truth Geo-referenced Labeled Datasets from Street-Level Images. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Qi, H.; Brown, M.; and Lowe, D. G. 2018. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5822–5830.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Samuelson, L. K.; and Smith, L. B. 2005. They call it like they see it: Spontaneous naming and attention to shape. *Developmental Science*, 8(2): 182–198.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Su, J.-C.; Maji, S.; and Hariharan, B. 2020. When does self-supervision improve few-shot learning? In *European conference on computer vision*, 645–666. Springer.

Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; and Fu, K. 2021. Research progress on few-shot learning for remote sensing image interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 2387–2402.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, B.; Wang, Z.; Sun, X.; Wang, H.; and Fu, K. 2021. DMML-Net: Deep metametric learning for few-shot geographic object segmentation in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9197–9206.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; and Yu, F. 2020. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*.

Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7278–7286.

Wu, J.; Pichler, D.; Marley, D.; Wilson, D.; Hovakimyan, N.; and Hobbs, J. 2022. Under Review: Extended Agriculture-Vision: An Extension of a Large Aerial Image Dataset for Agricultural Pattern Analysis. Forthcoming.

Yang, J.; Guo, X.; Li, Y.; Marinello, F.; Ercisli, S.; and Zhang, Z. 2022. A survey of few-shot learning in smart agriculture: developments, applications, and challenges. *Plant Methods*, 18(1): 1–12.

Yao, X.; Cao, Q.; Feng, X.; Cheng, G.; and Han, J. 2021. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zeng, Q.; and Geng, J. 2022. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191: 143–154.

Zhang, H.; Zhang, J.; and Koniusz, P. 2019. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2770–2779.