

GenCo: An Auxiliary Generator from Contrastive Learning for Enhanced Few-Shot Learning in Remote Sensing

Jing Wu^{a,*}, Naira Hovakimyan^a and Jennifer Hobbs^b

^aUniversity of Illinois Urbana-Champaign

^bIntelinair

ORCID ID: Jing Wu <https://orcid.org/0000-0001-8619-566X>

Abstract. Classifying and segmenting patterns from a limited number of examples is a significant challenge in remote sensing and earth observation due to the difficulty in acquiring accurately labeled data in large quantities. Previous studies have shown that meta-learning, which involves episodic training on query and support sets, is a promising approach. However, there has been little attention paid to direct fine-tuning techniques. This paper repurposes contrastive learning as a pre-training method for few-shot learning for classification and semantic segmentation tasks. Specifically, we introduce a generator-based contrastive learning framework (GenCo) that pre-trains backbones and simultaneously explores variants of feature samples. In fine-tuning, the auxiliary generator can be used to enrich limited labeled data samples in feature space. We demonstrate the effectiveness of our method in improving few-shot learning performance on two key remote sensing datasets: Agriculture-Vision and EuroSAT. Empirically, our approach outperforms purely supervised training on the nearly 95,000 images in Agriculture-Vision for both classification and semantic segmentation tasks. Similarly, the proposed few-shot method achieves better results on the land-cover classification task on EuroSAT compared to the results obtained from fully supervised model training on the dataset.

1 Introduction

Remote sensing (RS) and earth observation (EO) imagery enable the detection and monitoring of critical societal challenges, including food security, natural disasters, clean water availability, hunger and poverty, the impact of climate change, threats to animal habitats, geopolitical risk, and more. Like other domains, RS-EO has benefited from 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. 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 [56]. 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 [42, 37].

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 [46, 16, 43, 19, 42, 1, 30]. 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 contrastive 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. [22, 5, 21, 58] 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 the performance of few-shot learning in RS-EO classification and semantic segmentation tasks using contrastive-learning-based pre-training.

Specifically, we focus on pre-training from the Extended Agriculture Vision dataset (AV+) [54], 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 pat-

* Corresponding Author. Email: jingwu6@illinois.edu.

terns from only a small number of samples addresses key challenges in precision agriculture and food security.

Drawing inspiration from the work of [19, 38, 50], we have adopted a two-stage training approach that involves contrastive-learning-based pre-training followed by fine-tuning. Specifically, we have developed a contrastive learning framework, GenCo, with an auxiliary generator trained. During the pre-training, the generator is tasked with exploring variants of encoded features and formulating additional positive pairs. During the fine-tuning stage, we fixed the parameters of the encoders and trained only the classification layer and decoder for classification and segmentation, respectively. Notably, only a limited number of labeled data samples are provided during the fine-tuning stage. To address this, we introduced the generator from the contrastive learning model to create further samples in feature space and enrich knowledge during the few-shot tasks.

We find that the embeddings pre-trained from AV+ under this protocol show better adaptability when compared to counterparts pre-trained on ImageNet [13] and COCO [31]. Our method outperforms pre-trained 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.

Meanwhile, we demonstrate the high learning efficiency of the proposed method for RS-EO imagery. With a few labeled images, we find that the GenCo shows comparable or even better results under different tasks and datasets when compared with supervised models trained on fully 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 of few-shot learning tasks using remote sensing imagery and demonstrate the successful adaptation of the two-stage contrastive-learning-based few-shot strategy for RS-EO data.

(2) We propose a novel generative-based contrastive learning framework GenCo that benefits both pre-training and the downstream few-shot classification and segmentation. With negligible computation, the designed generator enhances the quality of embeddings and simultaneously provides extra data samples for downstream few-shot tasks.

(3) 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.

(4) We demonstrate that pre-training on AV+, a high-resolution multi-spectral RS-EO dataset, provides strong benefits to other RS-EO tasks on lower-resolution data such as EuroSAT.

2 Related Work

2.1 Few-Shot Learning

Among various methods to speed up training and enhance label efficiency, such as neural network pruning [33, 51], simulation-based training [55, 44], or few-shot learning. Few-shot learning is the most promising one in various application domains like healthcare [7, 4], remote sensing [42], for learning tasks including classification [15, 48] object detection [50], semantic segmentation [49], and robot learning [17]. 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 [25] pro-

pose applying a siamese neural network for few-shot classification. Following that, [43] 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 [3]. [20] further adds extra self-supervised tasks to enhance generalization capacity.

Optimization-based methods aim to learn through gradient back-propagation. Representative works include MAML [16], which realizes quick adaption from good initialization, Reptile [35], which simplifies the learning process of MAML, and MetaOptNet [26], which incorporates the support vector machine (SVM) as a classifier.

Hallucination-based methods seek to learn generators to generate unseen samples. Works from [52, 60, 28] 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, [42]. While much of the work is focused on scene classification [1, 29, 2, 32], other recent approaches examine semantic segmentation tasks [47, 24, 57].

2.2 Contrastive Learning

Contrastive learning is widely used for pre-training without labeled data and shows superior performance in various self-supervised learning (SSL) tasks [8, 21, 5, 6]. 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 [5], a momentum-update memory bank [22], projection heads [8], and/or stop-gradient trick [9]. Barlow Twins [58] 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 [8]. Within the field of RS-EO specifically, [34] takes MoCo-V2 as the basis and uses multiple projection heads to capture the desired invariance to seasonality. Therefore, we continue to utilize MoCo-V2 as a pre-training protocol in this work.

2.3 Data Augmentation

Data augmentation is essential for both contrastive learning and few-shot learning. Contrastive learning seeks to learn invariances in the representations to multiple augmentations so that these representations can be generalized successfully to various downstream tasks; to learn embeddings that are invariant to spatial and lighting transformations, the training of contrastive learning requires strong data augmentations, including color jittering, random flips, grayscale conversion, rotation, cropping, and resizing [8, 22, 21, 5]. Similar data augmentations are used in few-shot learning but with additional techniques like image hallucination [52], manifold mixup [45], and gallery image pool sampling [10]. More recent and related research to our paper comes from [53], which demonstrated effectiveness of feature-level operations in representation learning.

2.4 Pairing Supervised and Self-Supervised Learning

Some methods utilize the SSL loss as supplemental losses during the supervised training process [18, 41]. Often, additional efforts are

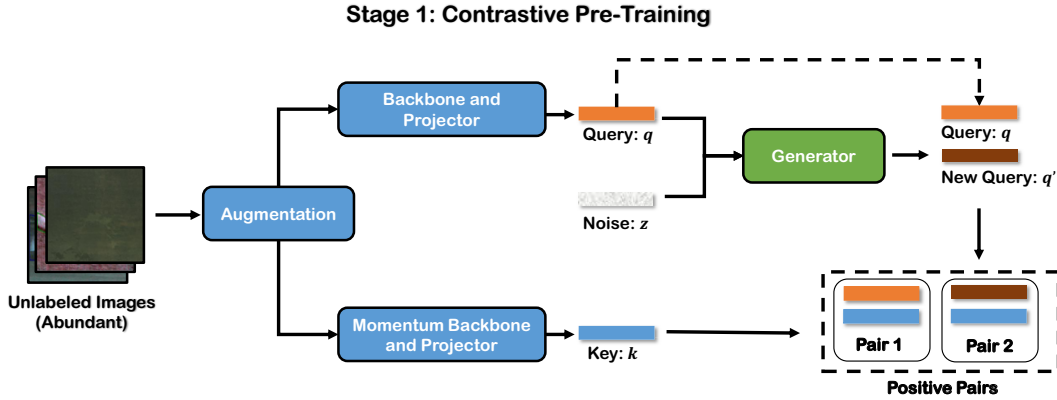


Figure 1. Illustration of the pre-training stage of GenCo. In the pre-training stage, we follow the training strategy of contrastive learning to train the backbones and projectors jointly with abundant unlabeled images. An additional feature vector q' is added by passing the real feature vector q and a noise vector z to the generator. The generator is trained end-to-end along with the contrastive learning framework

needed to calibrate (i.e. re-weight) these losses when crossing different domains. More straightforward and effective methods come from supervised fine-tuning [14]. 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 [59] and segmentation [27].

3 Method

In this section, we first introduce the overall pipeline of pre-training and downstream few-shot learning in Section 3.1. Following this, we describe the proposed GenCo in Section 3.2. Lastly, we demonstrate how the generator from the contrastive learning model helps the downstream few-shot task by generating extra samples in feature space 3.3.

3.1 The Pipeline of Learning

In the first stage, we pre-trained different backbones with contrastive learning models on unlabeled data, as shown in Figure 1. To be more specific, there are roughly 1,300,000 images with shapes 512×512 . These images are all randomly cropped from the raw images from AV+. We use all these unlabeled images as input to pre-train backbones without any supervision. After the pre-training, we move to the second stage, which fine-tunes the pre-trained backbones for the downstream few-shot tasks. Since the proposed contrastive learning 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 to optimize the classification accuracy or mIoU of agricultural patterns on Agriculture-Vision and land covers on EuroSAT.

3.2 Pre-training with Generator-based Contrastive Learning

Basic Framework. The framework of GenCo is shown in Figure 1. Specifically, GenCo can be trained with natural scene images that contain information about red, green, and blue channels similar to previous works[5, 8, 9]. However, AV+ has extra information in

the NIR channel. To fully explore knowledge from the pre-training dataset, we further add one channel to the backbones following the work of from[54].

In every training iteration, a training sample x is augmented into two different views named query x^q and key x^k . These views contain the same semantic meaning but also variations introduced from data augmentations, including spatial and color transforms. With an online network and a momentum-updated offline network proposed in [22], the training encourages these two views to be mapped into two similar embedding spaces, i.e., q, k , as a positive pair. For feature vectors that are not encoded from x , we define them as k^-

Feature Generation. Sufficient positive pairs in feature space help the performance of contrastive learning[5]. However, the large number of positive features relies on large batch sizes that may not always be accessible. To address this, we design a generator G that takes the query feature q and random noise z as input. Entries of z are assumed to follow a normal distribution with mean 0 and variance 0.1. We then generate a new sample q' in the feature space where $q' = G(q, z|\theta)$. Notably, G is a lightweight module with parameters θ . It is instantiated with three linear layers and a ReLU layer between two successive layers.

With this simple design, we eventually obtain an additional positive pair $p(q', k)$. Along with the original positive $p(q, k)$ one, the framework provided enhanced contrast and improve the quality of embeddings during the training with an additional little computation. Together, based on positive and negative pairs and a temperature parameter τ for scaling, the training loss function, i.e., InfoNCE [36], is then defined as follows:

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

3.3 Downstream Few-Shot Learning with Generator

Given the strong data augmentations in contrastive learning, the backbone features encoded can capture the key representation features. It should adapt to different data classes and types of downstream tasks without much learning effort. Therefore, one of the critical steps of the proposed method is to separate representation learning and downstream task learning into two stages.

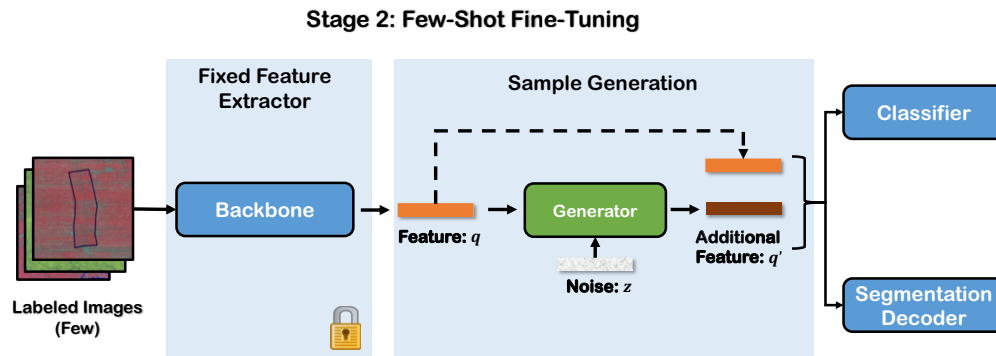


Figure 2. Illustration of the fine-tuning stage. In the fine-tuning stage with limited labeled images, we adapt the pre-trained backbone from stage 1 for feature extraction. The generator is brought from the GenCo to generate additional labeled data. During the tuning, the feature extractors are fixed, and fine-tuning is only involved in the generator, classifier and segmentation decoder.

With contrastive learning applied in the first stage, we freeze the encoder and 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. Then, each image was encoded into feature space, forming a vector q with label y . While the traditional fine-tuning method applied the feature q directly to downstream tasks, we take advantage of the generator G from the contrastive learning framework to generate labeled samples following the equation $q' = G(q, z|\theta)$ as shown in Figure 2. The q' share the same label y with q , forming an extra label data (q', y) . In other words, we eventually obtained $2K$ labeled data points given K images for each class, largely enriching the information for downstream learning tasks. During the fine-tuning stage, the generator G is differentiable and optimized simultaneously with the fine-tuning stage.

In the few-shot classification task, we add one fully connected layer to the backbones without introducing extra non-linearity. We assign randomly initialized weights to the added classification layer. In the few-shot semantic segmentation task, we choose the lightweight segmentation model U-Net [39] for fine-tuning given limited training samples. Concretely, we add a five-layer decoder based on the encoders pre-trained from contrastive learning. In each layer of the decoder, it performs up-sampling by a 2×2 deconvolution layer to recover the original images' sizes. While the size of the feature images increases, the number of channels reduces by half after each up-sampling. Empirically, this asymmetric encoder-decoder shows exceptional performance in the segmentation task on Agriculture-Vision.

4 Experiments and Results

In this study, we conduct various experiments to prove the effectiveness of our proposed methods. In Section 4.1, we introduced the used datasets in this paper. Then, we illustrate the evaluation metrics in Section 4.2. In the following Section 4.3, we demonstrate the necessary details to reproduce the experiments. Lastly, in Section 4.4, we report all the results based on the proposed metrics.

4.1 Datasets

4.1.1 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-performing models for classification and segmentation tasks [11]. 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.

4.1.2 Extended Agriculture-Vision

For contrastive pre-training, we use the large-scale remote-sensing dataset, Extend Agriculture-Vision [54]. While the original Agriculture-Vision dataset contained only 512×512 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 $10,000 \times 10,000$ 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 pre-trained on this dataset should capture remote sensing, agriculture, and temporal features. Therefore, the embeddings pre-trained on Extend Agriculture-Vision should be adapted well to diverse downstream tasks such as agricultural pattern recognition and land-cover classification.

4.1.3 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 [34, 23, 54]. While there is a total of 13 channels in total, in this work, we focus on the RGB channels since it is a more general modality.

4.2 Experiment Setup and Evaluation Metrics

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 pre-trained on ImageNet, and the backbone pre-trained on AV+ using contrastive learning models. Similarly, in the few-shot semantic segmentation task, our embeddings pre-trained on AV+ and embeddings from COCO’s weights are compared[31]. 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.

4.3 Implementations Details

Firstly, we introduce implementation details of the generator-based contrastive learning model to pre-train backbones, as shown in the left part of Figure 1 in Section 4.3.1. Then, we demonstrate details few-shot experiments on classification and semantic segmentation tasks on Agriculture-Vision shown in the right part of Figure 2 in Section 4.3.2. Following this, we list the key parameters for evaluation in the few-shot semantic segmentation task in Section 4.3.3. Lastly, we report the necessary details for EuroSAT in Section .

4.3.1 Pre-training on Extended Agriculture-Vision

GenCo uses different ResNet as its encoder and two layers of MLP as a projector as the basis for our contrastive learning framework. The following generator G is instantiated with three linear layers and a ReLU layer between two successive layers. The output feature vectors, i.e., q , q' , and k have dimensions of 128, and there are 16,384 negative keys stored in the memory bank. We pre-train the generator-based contrastive learning model for 200 epochs using an SGD optimizer, a learning rate of 0.3, and a weight decay of 0.0001. The learning rate is adjusted to 0.03 and 0.003 at the 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.

Besides the proposed GenCo, we also conduct experiments using MoCo-V2[8], SimSiam[9] and SimCLR[5] as alternative methods for pre-training. Results are also reported in the following sections. For MoCo-V2, we train the model with the exact same hyperparameters as the proposed method, given the similarity of the structure. For SimSiam, we train the models for 200 epochs with a batch size of 256. The learning rate = 0.05 with an SGD optimizer. The weight decay is 0.0001 and the SGD momentum is 0.9. For pre-training of SimCLR, we reproduce SimCLR with a smaller batch size of 512 and cosine appealed learning rate of 0.05. We follow the self-supervised learning paradigm in [6] without distillation for a fair comparison of the other two contrastive learning methods. All the experiments are conducted on a server with 8 GPUs.

4.3.2 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 and the generator G , which are learnable. Different from the optimization methods used in [22], 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.

4.3.3 Few-Shot Segmentation on Agriculture-Vision

Following the work from [12], we ignore storm damage annotations when performing evaluations due to their extreme scarcity. Similar to the fine-tuning strategy 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 [40]. 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.

4.3.4 Few-Shot Classification on EuroSAT

We additionally illustrate that embeddings learned from GenCo 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 [23]. Following experiments in previous sections, we evaluate the quality and adaptability of pre-trained features from the proposed methods in this land-cover classification task.

We add one fully connected layer to pre-trained backbones, building 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.

Table 1. Comparison of fine-tuning results between weights from supervised ImageNet and weights from GenCo on AV+ for the 10-shot classification

Pre-trained Weight	Backbone	Accuracy(%)
Sup. on ImageNet	ResNet-18	55.22 ± 0.34
SimSiam on AV+	ResNet-18	63.12 ± 0.71
SimCLR on AV+	ResNet-18	63.59 ± 0.83
MoCoV2 on AV+	ResNet-18	64.17 ± 0.62
GenCo on AV+	ResNet-18	65.51 ± 0.68
ImageNet	ResNet-50	56.53 ± 0.43
SimSiam on AV+	ResNet-50	62.41 ± 0.78
SimCLR on AV+	ResNet-50	63.55 ± 0.77
MoCoV2 on AV+	ResNet-50	63.80 ± 0.72
GenCo on AV+	ResNet-50	64.82 ± 0.70
Sup. on ImageNet	ResNet-101	54.34 ± 0.45
SimSiam on AV+	ResNet-101	62.27 ± 0.89
SimCLR on AV+	ResNet-101	63.50 ± 0.80
MoCoV2 on AV+	ResNet-101	64.12 ± 0.78
GenCo on AV+	ResNet-101	64.62 ± 0.49

4.4 Results of Experiments

4.4.1 Few-shot Learning on Agriculture-Vision

Quality of Pre-Trained Embeddings. We first prove the quality and adaptability of pre-trained embeddings from the proposed methods. As shown in Table 1, our pre-trained 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 pre-trained 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. MoCoV2, SimSiam and SimCLR show sub-optimal results compared with GenCo. However, with our two-stage training strategy, all AV+ pre-trained weights enable better performance than any ImageNet weights.

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

Pre-trained Weight	Freeze Backbone	Number of Images	Accuracy (%)
Random	False	9,000	57.30 ± 0.81
Random	False	94,986	62.31 ± 0.25
GenCo on AV+	True	10	65.51 ± 0.68

Table 3. 9-way few-shot classification accuracy on Agriculture-Vision based on weights pre-trained from GenCo.

Backbone \ Shots	Accuracy 10 shots (%)	Accuracy 5 shots (%)	Accuracy 1 shot (%)
ResNet-18	65.51 ± 0.68	61.61 ± 0.71	16.72 ± 1.32
ResNet-50	64.82 ± 0.70	59.44 ± 0.79	29.64 ± 1.28
ResNet-101	64.62 ± 0.49	59.56 ± 0.66	28.84 ± 1.73

Table 4. Comparison of fine-tuning results between weights pre-trained on COCO and weights from GenCo on AV+ for the 10-shot semantic segmentation task.

Pre-trained Weight	Backbone	mIoU - 8 Classes
COCO	ResNet-18	15.61
GenCo on AV+	ResNet-18	23.56
COCO	ResNet-50	15.60
GenCo on AV+	ResNet-50	23.00
COCO	ResNet-101	15.19
GenCo on AV+	ResNet-101	21.04

Learning Efficiency. Next, we continue to demonstrate the learning efficiency of GenCo 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 Table 2, GenCo 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.

4.4.2 Ablation Study of Few-shot Classification

In this section, we aim to prove the effectiveness of the proposed generator G in the contrastive learning framework. We conduct experiments on GenCo with ResNet-18 as the backbones for pre-training

on AV+. We then report the results of the downstream 10-shot classification tasks on the agriculture vision dataset.

We conducted a comparative analysis of three frameworks: MoCo-V2, MoCo-V2 with a generator, and our proposed GenCo. Our experimental results, as summarized in Table 6, demonstrate that the use of a generator improves the learned embeddings even when introduced solely during pre-training. This improvement can be attributed to the additional variance and contrast introduced by the positive feature vectors. However, the performance gain achieved is relatively marginal. Notably, the largest gain occurs when we adopt the pre-trained generator to few-shot learning, which provides additional labeled data in the feature space. As such, the generator can enhance both pre-training and downstream tasks simultaneously, thereby improving the overall performance of the framework.

4.4.3 Few-shot Segmentation on Agriculture-Vision

Quality of Pre-Trained Embeddings. Since U-Net’s structures contain skip connections from different layers [39], we don’t evaluate a single embedding but features from different scales. Concretely, features from GenCo and features from encoders pre-trained on COCO are compared using the mean intersection over union (mIoU) metric. As reported in Table 4, our proposed method shows around 6-8 points of improvement compared with weights pre-trained 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 pre-trained from natural images (COCO) and remote sensing images (AV+) is significantly different. Therefore, we can observe a noticeable improvement in the results pre-trained on AV+.

Learning Efficiency. We also examine the learning efficiency of the segmentation task. To do this, we use only ten sampled images per category and compare the results of GenCo with those of models trained on the full Agriculture-Vision dataset. While our approach fixes the backbone, we unfreeze the segmentation model’s encoder and the generator training on the full dataset. Based on the results presented in Table 5, we observe an improvement of 2.19 points and 1.18 points for ResNet-18 and ResNet-50, respectively, using the GenCo approach. However, for MoVo-V2, SimSiam, and SimCLR, the results are comparable or only slightly better than those obtained using the end-to-end training method. Importantly, while the GenCo-based few-shot segmentation approach still outperforms models trained with a large number of labeled images, we note that the improvement is not as significant as the improvement achieved in the classification task. This observation is likely because the decoders used for segmentation have more parameters to be tuned than a single-layer classifier. With limited labeled samples, smaller models are better able to avoid overfitting and show better results. Therefore, in this few-shot segmentation task, ResNet-18 performs the most satisfactorily.

4.4.4 Few-shot Classification on EuroSAT

Quality of Pre-Trained Embeddings. Results show that GenCo still leads to better embeddings on this remote sensing dataset. As seen in Table 7, features from GenCo improve one percent of accuracy on average compared to the features trained from ImageNet. Since EuroSAT shares much less similarity with our pre-trained dataset, i.e., AV+, the improvement is moderate. However, the gain is still stably earned, crossing different sizes of backbones. The results from

Table 5. 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.

Pre-trained Weight	Backbone	Freeze Backbone	Number of Images	mIoU - 8 Classes
Random	ResNet-18	False	9000	19.02
Random	ResNet-18	False	94986	21.37
SimSiam on AV+	ResNet-18	True	10	21.30
SimCLR on AV+	ResNet-18	True	10	22.13
MoCo on AV+	ResNet-18	True	10	22.11
GenCo on AV+	ResNet-18	True	10	23.56
Random	ResNet-50	False	9000	19.58
Random	ResNet-50	False	94986	21.82
SimSiam on AV+	ResNet-50	True	10	21.06
SimCLR on AV+	ResNet-50	True	10	21.98
MoCo on AV+	ResNet-50	True	10	21.21
GenCo on AV+	ResNet-50	True	10	23.00

Table 6. Ablation of the proposed generator with different contrastive learning frameworks.

Shots \ Modules	No Generator (MoCo-V2)	Generator with Pre-training	GenCo
10 shot	64.17	64.61	65.51
5 shot	59.32	59.91	61.61
1 shot	15.21	15.23	16.72

MoCo-V2, SimSiam and SimCLR still outperform the results of ImageNet but are sub-optimal compared with results from GenCo, proving the effectiveness of the proposed generator.

Table 7. Comparison of fine-tuning results between weights pre-trained on supervised ImageNet and weights from our GenCo on EuroSAT for the 10-shot classification

Pre-trained Weight	Backbone	Accuracy(%)
ImageNet	ResNet-18	66.90 ± 0.11
SimSiam on AV+	ResNet-18	67.20 ± 0.43
SimCLR on AV+	ResNet-18	67.31 ± 0.51
MoCo on AV+	ResNet-18	67.14 ± 0.32
GenCo on AV+	ResNet-18	67.92 ± 0.31
ImageNet	ResNet-50	65.01 ± 0.14
SimSiam on AV+	ResNet-50	65.61 ± 0.38
SimCLR on AV+	ResNet-50	65.93 ± 0.62
MoCo on AV+	ResNet-50	65.58 ± 0.40
GenCo on AV+	ResNet-50	66.11 ± 0.45
ImageNet	ResNet-101	63.34 ± 0.20
SimSiam on AV+	ResNet-101	63.32 ± 0.35
SimCLR on AV+	ResNet-101	64.17 ± 0.68
MoCo on AV+	ResNet-101	63.79 ± 0.45
GenCo on AV+	ResNet-101	64.79 ± 0.46

Table 8. Comparison of the classification task between the 10-shot results of GenCo and end-to-end training using the full EuroSAT on ResNet-18. *: results referred from [34]

Pre-trained Weight	Freeze Backbone	Number of Images	Accuracy (%)
Random	False	2,700	58.81 ± 0.10
Random	False	27,000	63.21 *
Random	False	27,000	63.34 ± 0.08
GenCo on AV+	True	10	66.90 ± 0.31

Learning Efficiency. 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 Table 8. This result is crucial as it proves the effectiveness of our methods in different domains. With a remarkably cheap effort of labeling, it re-verifies the vast possibility of deploying our models to various downstream tasks in agriculture and remote sensing.

Table 9. 10-way few-shot classification accuracy on EuroSAT based on weights pre-trained from GenCo.

Backbone \ Shots	Accuracy 10 shots (%)	Accuracy 5 shots (%)	Accuracy 1 shot (%)
ResNet-18	67.92 ± 0.31	63.20 ± 0.37	11.50 ± 0.82
ResNet-50	65.01 ± 0.45	59.40 ± 0.51	11.21 ± 1.42
ResNet-101	63.70 ± 0.46	58.70 ± 0.66	11.40 ± 1.71

We show the results of the 10-way few-shot classification on EuroSAT in the following Table 9. 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.

5 Conclusion

In this study, we proposed GenCo, a generator-based two-stage approach for few-shot classification and segmentation on remote sensing and earth observation data. Our method proved to be effective due to the sufficient contrast introduced by the generator during pre-training and the remarkable adaptability of embeddings in downstream few-shot tasks. Furthermore, we demonstrated that the generator could be integrated into the few-shot learning framework to further address the issue of data scarcity and enrich the learning information. Most importantly, our approach provides an alternative solution to the labeling challenges in agriculture and remote sensing domains. With our few-shot contrastive learning-based approach, we believe that it is possible to deploy models in the real world with minimal labeled data and training effort.

References

- [1] Dalal Alajaji, Haikel S Alhichri, Nassim Ammour, and Naif Alajlan, 'Few-shot learning for remote sensing scene classification', in *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pp. 81–84. IEEE, (2020).
- [2] Dalal A Alajaji and Haikel Alhichri, 'Few shot scene classification in remote sensing using meta-agnostic machine', in *2020 6th conference on data science and machine learning applications (CDMA)*, pp. 77–80. IEEE, (2020).
- [3] Yannick Le Cacheux, Herve Le Borgne, and Michel Crucianu, '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*, pp. 10333–10342, (2019).
- [4] Suiyao Chen, Nan Kong, Xuxue Sun, Hongdao Meng, and Mingyang Li, 'Claims data-driven modeling of hospital time-to-readmission risk with latent heterogeneity', *Health care management science*, **22**, 156–179, (2019).
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, 'A simple framework for contrastive learning of visual representations', in *International conference on machine learning*, pp. 1597–1607. PMLR, (2020).
- [6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton, 'Big self-supervised models are strong semi-supervised learners', *Advances in neural information processing systems*, **33**, 22243–22255, (2020).
- [7] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang, 'Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images', *Pattern recognition*, **113**, 107826, (2021).
- [8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, 'Improved baselines with momentum contrastive learning', *arXiv preprint arXiv:2003.04297*, (2020).
- [9] Xinlei Chen and Kaiming He, 'Exploring simple siamese representation learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, (2021).
- [10] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert, 'Image deformation meta-networks for one-shot learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8680–8689, (2019).
- [11] Mang Tik Chiu, Xingqian Xu, Kai Wang, Jennifer Hobbs, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi, 'The 1st agriculture-vision challenge: Methods and results', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, (June 2020).
- [12] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatán Karapetyan, Ivan Dozier, Greg Rose, et al., 'Agriculture-vision: A large aerial image database for agricultural pattern analysis', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2828–2838, (2020).
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).
- [14] Carl Doersch, Ankush Gupta, and Andrew Zisserman, 'Crosstransformers: spatially-aware few-shot transfer', *Advances in Neural Information Processing Systems*, **33**, 21981–21993, (2020).
- [15] Li Fei-Fei, Robert Fergus, and Pietro Perona, 'One-shot learning of object categories', *IEEE transactions on pattern analysis and machine intelligence*, **28**(4), 594–611, (2006).
- [16] Chelsea Finn, Pieter Abbeel, and Sergey Levine, 'Model-agnostic meta-learning for fast adaptation of deep networks', in *International conference on machine learning*, pp. 1126–1135. PMLR, (2017).
- [17] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine, 'One-shot visual imitation learning via meta-learning', in *Conference on robot learning*, pp. 357–368. PMLR, (2017).
- [18] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord, 'Boosting few-shot visual learning with self-supervision', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8059–8068, (2019).
- [19] Spyros Gidaris and Nikos Komodakis, 'Dynamic few-shot visual learning without forgetting', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4367–4375, (2018).
- [20] Spyros Gidaris and Nikos Komodakis, 'Generating classification weights with gnn denoising autoencoders for few-shot learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21–30, (2019).
- [21] Jean-Bastien Grill, Florian Strub, Florent Alché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., 'Bootstrap your own latent—a new approach to self-supervised learning', *Advances in neural information processing systems*, **33**, 21271–21284, (2020).
- [22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, 'Momentum contrast for unsupervised visual representation learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, (2020).
- [23] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, '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, (2019).
- [24] Ronald Kemker, Ryan Luu, and Christopher Kanan, 'Low-shot learning for the semantic segmentation of remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing*, **56**(10), 6214–6223, (2018).
- [25] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al., 'Siamese neural networks for one-shot image recognition', in *ICML deep learning workshop*, volume 2, p. 0. Lille, (2015).
- [26] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, 'Meta-learning with differentiable convex optimization', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, (2019).
- [27] Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, and Chao Tao, '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, (2022).
- [28] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu, 'Adversarial feature hallucination networks for few-shot learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13479, (2020).
- [29] Lingjun Li, Junwei Han, Xiwen Yao, Gong Cheng, and Lei Guo, 'Diamatchnet for few-shot remote sensing image scene classification', *IEEE Transactions on Geoscience and Remote Sensing*, **59**(9), 7844–7853, (2020).
- [30] Xiang Li, Jingyu Deng, and Yi Fang, 'Few-shot object detection on remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–14, (2021).
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *European conference on computer vision*, pp. 740–755. Springer, (2014).
- [32] Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Gang Wan, and Ruihui Wang, 'Deep few-shot learning for hyperspectral image classification', *IEEE Transactions on Geoscience and Remote Sensing*, **57**(4), 2290–2304, (2018).
- [33] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin, 'Thinet: A filter level pruning method for deep neural network compression', in *Proceedings of the IEEE international conference on computer vision*, pp. 5058–5066, (2017).
- [34] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez, 'Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9414–9423, (2021).
- [35] Alex Nichol and John Schulman, 'Reptile: a scalable metalearning algorithm', *arXiv preprint arXiv:1803.02999*, 2(3), 4, (2018).
- [36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, 'Representation learning with contrastive predictive coding', *arXiv preprint arXiv:1807.03748*, (2018).
- [37] Madhava Paliyam, Catherine L Nakalembe, Kevin Liu, Richard Nyirawung, and Hannah R Kerner, '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*, (2021).
- [38] Hang Qi, Matthew Brown, and David G Lowe, 'Low-shot learning with imprinted weights', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5822–5830, (2018).
- [39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, 'U-net: Convolutional networks for biomedical image segmentation', in *International*

- Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, (2015).
- [40] Leslie N Smith and Nicholay Topin, ‘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, pp. 369–386. SPIE, (2019).
- [41] Jong-Chyi Su, Subhansu Maji, and Bharath Hariharan, ‘When does self-supervision improve few-shot learning?’, in *European conference on computer vision*, pp. 645–666. Springer, (2020).
- [42] Xian Sun, Bing Wang, Zhirui Wang, Hao Li, Hengchao Li, and Kun Fu, ‘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, (2021).
- [43] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, ‘Learning to compare: Relation network for few-shot learning’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, (2018).
- [44] Ran Tao, Pan Zhao, Jing Wu, Nicolas F Martin, Matthew T Harrison, Carla Ferreira, Zahra Kalantari, and Naira Hovakimyan, ‘Optimizing crop management with reinforcement learning and imitation learning’, *arXiv preprint arXiv:2209.09991*, (2022).
- [45] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, ‘Manifold mixup: Better representations by interpolating hidden states’, in *International Conference on Machine Learning*, pp. 6438–6447. PMLR, (2019).
- [46] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., ‘Matching networks for one shot learning’, *Advances in neural information processing systems*, **29**, (2016).
- [47] Bing Wang, Zhirui Wang, Xian Sun, Hongqi Wang, and Kun Fu, ‘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, (2021).
- [48] Haoxiang Wang, Yite Wang, Ruoyu Sun, and Bo Li, ‘Global convergence of maml and theory-inspired neural architecture search for few-shot learning’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9797–9808, (2022).
- [49] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng, ‘Panet: Few-shot image semantic segmentation with prototype alignment’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9197–9206, (2019).
- [50] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu, ‘Frustratingly simple few-shot object detection’, *arXiv preprint arXiv:2003.06957*, (2020).
- [51] Yite Wang, Jing Wu, Naira Hovakimyan, and Ruoyu Sun, ‘Double dynamic sparse training for gans’, *arXiv preprint arXiv:2302.14670*, (2023).
- [52] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan, ‘Low-shot learning from imaginary data’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7278–7286, (2018).
- [53] Jing Wu, Jennifer Hobbs, and Naira Hovakimyan, ‘Hallucination improves the performance of unsupervised visual representation learning’, *arXiv preprint arXiv:2307.12168*, (2023).
- [54] Jing Wu, David Pichler, Daniel Marley, David Wilson, Naira Hovakimyan, and Jennifer Hobbs, ‘Extended agriculture-vision: An extension of a large aerial image dataset for agricultural pattern analysis’, *arXiv preprint arXiv:2303.02460*, (2023).
- [55] Jing Wu, Ran Tao, Pan Zhao, Nicolas F Martin, and Naira Hovakimyan, ‘Optimizing nitrogen management with deep reinforcement learning and crop simulations’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1720, (2022).
- [56] Jiachen Yang, Xiaolan Guo, Yang Li, Francesco Marinello, Sezai Er-cisli, and Zhuo Zhang, ‘A survey of few-shot learning in smart agriculture: developments, applications, and challenges’, *Plant Methods*, **18**(1), 1–12, (2022).
- [57] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han, ‘Scale-aware detailed matching for few-shot aerial image semantic segmentation’, *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–11, (2021).
- [58] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, ‘Barlow twins: Self-supervised learning via redundancy reduction’, in *International Conference on Machine Learning*, pp. 12310–12320. PMLR, (2021).
- [59] Qingjie Zeng and Jie Geng, ‘Task-specific contrastive learning for few-shot remote sensing image scene classification’, *ISPRS Journal of Photogrammetry and Remote Sensing*, **191**, 143–154, (2022).
- [60] Hongguang Zhang, Jing Zhang, and Piotr Koniusz, ‘Few-shot learning via saliency-guided hallucination of samples’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2770–2779, (2019).