Extended Agriculture-Vision: An Extension of a Large Aerial Image Dataset for Agricultural Pattern Analysis

Anonymous authors
Paper under double-blind review

Abstract

A key challenge for much of the machine learning work on remote sensing and earth observation data is the difficulty in acquiring large amounts of accurately labeled data. This is particularly true for semantic segmentation tasks, which are much less common in the remote sensing domain because of incredible difficulty in collecting precise, accurate, pixel-level annotations at scale. Recent efforts have addressed these challenges both through the creation of supervised datasets as well as the application of self-supervised methods. We continue these efforts on both fronts. First, we generate and release an improved version of the Agriculture-Vision dataset Chiu et al. (2020b) to include raw, full-field imagery for greater experimental flexibility. Second, we extend this dataset with the release of 3600 large, high-resolution (10cm/pixel), full-field, red-green-blue and near-infrared images for pre-training. Third, we incorporate the Pixel-to-Propagation Module Xie et al. (2021b) originally built on the SimCLR framework into the framework of MoCo-V2 Chen et al. (2020b). Finally, we demonstrate the usefulness of this data by benchmarking different contrastive learning approaches on both downstream classification *and* semantic segmentation tasks. We explore both CNN and Swin Transformer Liu et al. (2021a) architectures within different frameworks based on MoCo-V2. Together, these approaches enable us to better detect key agricultural patterns of interest across a field from aerial imagery so that farmers may be alerted to problematic areas in a timely fashion to inform their management decisions. Furthermore, the release of these datasets will support numerous avenues of research for computer vision in remote sensing for agriculture.

1 Introduction

Massive annotated datasets like ImageNet have fostered the development of powerful and robust deep learning models for natural images Deng et al. (2009); He et al. (2016); Simonyan & Zisserman (2014); Krizhevsky et al. (2012); Russakovsky et al. (2015). However, creating large complex datasets is costly, time-consuming, and may be infeasible in some domains or for certain tasks. Simultaneously, vast amounts of unlabeled data exists in most domains. Contrastive learning has recently emerged as an encouraging candidate for solving the need for large labeled datasets Grill et al. (2020); Caron et al. (2020; 2018); He et al. (2020). Through pre-training, these approaches open up the possibility of using unlabeled images as its own supervision and transferring in-domain images to further downstream tasks Tian et al. (2020); Ayush et al. (2020).

While natural scene imagery largely dominates the research landscape in terms of vision algorithms, datasets and benchmarks, the rapid increase in quantity and quality of remote sensing imagery has led to significant advances in this domain as well Kelcey & Lucieer (2012); Maggiori et al. (2017); Ramanath et al. (2019); Xia et al. (2017). Coupled with deep neural networks, remote sensing has achieved exceptional success in multiple domains such as natural hazards assessment Van Westen (2013), climate tracking Rolnick et al. (2019); Yang et al. (2013), and precision agriculture Mulla (2013); Seelan et al. (2003); Barrientos et al. (2011); Gitelson et al. (2002). However, obtaining large quantities of accurate annotations is especially challenging for remote sensing tasks, particularly for agriculture, as objects of interest tend to be very small, high in number (perhaps thousands per image), possess complex organic boundaries, and may require channels beyond red-green-blue (RGB) to identify.

Many approaches originally developed for natural images work well on remote sensing imagery with only minimal modification, although this is not guaranteed due to the large domain gap. Additionally, they may fail to exploit the unique structure of earth observation data such as geographic consistency or seasonality Mañas et al. (2021). Explicitly benchmarking approaches on domain relevant data is critical. In this work, we focus primarily on the Agriculture-Vision (AV) dataset Chiu et al. (2020b): a large, multi-spectral, high-resolution (10 cm/pixel), labeled remote sensing dataset for semantic segmentation. Unlike low-resolution public satellite data, this imagery enables within-field identification of key agronomic patterns such as weeds and nutrient deficiency. While this dataset is noted for its size, most aerial agriculture dataset are quite small. Therefore we leverage the large amounts of *un-annotated data* which is readily available in this domain, benchmark several self-supervised approaches whose inductive bias reflect the structure of this data, and evaluate the impact of these approaches in more data-limited settings.

Together, our contributions are as follows:

- We release a full-field version of the Agriculture-Vision dataset to further encourage broad agricultural research in pattern analysis and semantic segmentation.
- We release over 3 terabytes of unlabeled, full-field images from more than 3600 full-field images to enable unsupervised pre-training.
- We benchmark self-supervised pre-training methods based on momentum contrastive learning and evaluate their performance on downstream classification *and* semantic segmentation tasks with variable amounts of annotated data.
- We perform benchmarks using both CNN and Swin Transformer backbones.
- We incorporate the Pixel-to-Propagation Module Xie et al. (2021b) (PPM), originally built on SimCLR Chen et al. (2020a), into the MoCo-V2 Chen et al. (2020b) framework and evaluate its performance.
- We adapt the approach of Seasonal Contrast (SeCo) Mañas et al. (2021) for this dataset, which contains imagery only during the growing season, and fuse this approach with PPM to specifically address the spatiotemporal nature of the raw data and the desire to perform downstream segmentation tasks.

2 Related Work

2.1 Constrastive Learning

Unsupervised and self-supervised learning (SSL) methods have proven to be very successful for pre-training deep neural networks Erhan et al. (2010); Bengio (2012); Mikolov et al. (2013); Devlin et al. (2018). Recently, methods like MoCo He et al. (2020); Chen et al. (2020b), SimCLR Chen et al. (2020a), BYOL Grill et al. (2020) and others Bachman et al. (2019); Henaff (2020); Li et al. (2020) based on contrastive learning methods have achieved state-of-the-art performance. These approaches seek to learn by encouraging the attraction of different views of the same image ("positive pairs") as distinguished from "negative pairs" from different images Hadsell et al. (2006). Several approaches have sought to build on these base frameworks by making modifications that better incorporate the invariant properties and structure of the input data or task output. Specifically pertinent to the current work, Xie et al. (2021b) extended the SimCLR framework through the incorporation of pixel-to-propagation module and additional pixel-level losses to improve performance on downstream tasks requiring dense pixel predictions. Mañas et al. (2021) Mañas et al. (2021) combined multiple encoders to capture the time and position invariance in downstream remote sensing tasks.

2.2 Remote Sensing Datasets

Aerial images have been widely explored over the past few decades Cordts et al. (2016); Everingham et al. (2010); Gupta et al. (2019); Lin et al. (2014); Zhou et al. (2017), but the datasets for image segmentation typically focus on routine, ordinary objects or street scenes Deng et al. (2009). Many prominent datasets including Inria Aerial Image Maggiori et al. (2017), EuroSAT Helber et al. (2019), and DeepGlobe Building Demir et al. (2018) are built on low-resolution satellite (e.g. Sentinel-1, Sentinel-2, MODIS, Landsat) and only have limited resolutions that vary from 800 cm/pixel to 30 cm/pixel and can scale up to 5000×5000

Figure 1: Left: Full-field imagery (RGB-only) constructed from the AV dataset. A field of this size is approximately 15,000×15,000 pixels which can yield many smaller tiles. Right: Sample imagery and labels for the fine-grained segmentation task.

pixels. Those datasets featuring segmentation tend to explore land-cover classification or change detection Daudt et al. (2018); Sumbul et al. (2019).

Pertaining to aerial agricultural imagery, datasets tend to be either low-resolution (>10 m/pixel) satellite Tseng et al. (2021); Feng & Bai (2019) or very high-resolution (<1 cm/pixel) imagery taken from UAV or on-board farming equipment Haug & Ostermann (2014); Olsen et al. (2019). The Agriculture-Vision dataset Chiu et al. (2020b;a) introduced a large, high-resolution (10 cm/pixel) dataset for segmentation, bridging these two alternate paradigms.

3 Datasets

3.1 Review and Reprocessing of Agriculture-Vision Dataset

The original AV dataset Chiu et al. (2020b) consists of 94,986 high-resolution (10-20 cm/pixel) RGB and near-infrared (NIR) aerial imagery of farmland. Special cameras were mounted to fixed-wing aircraft and flown over the Midwestern United States during the 2017-2019 growing seasons, capturing predominantly corn and soybean fields. Each field was annotated for nine patterns described in the supplemental material. After annotation, 512×512 *tiles* were extracted from the full-field images and *then* pre-processed and scaled. While this pre-processing produces a uniformly curated dataset, it naturally discards important information about the original data.

To overcome this limitation, we obtained the original raw, full-field imagery. We are releasing this raw data as full-field images without any tiling, as it has been demonstrated to be beneficial to model performance Chiu et al. (2020a). A sample image is shown in Figure 1 (left). The original dataset can be recreated from this new dataset by extracting the tiles at the appropriate pixel coordinates provided in the data manifest.

3.2 Raw Data for Pre-training

We identified 1200 fields from the 2019-2020 growing seasons collected in the same manner as in Section 3.1. For each field we selected three images, referred to as *flights*, taken at different times in the growing season, resulting in 3600 raw images available for pre-training. We elect to include data from 2020 even though it is not a part of the original supervised dataset because it is of high quality, similar in distribution to 2019, and we wish to encourage exploration around incorporating different source domains into modeling approaches as this is a very central problem to remote sensing data. We denote this raw imagery plus the original supervised dataset (in full-field format) as the "Extended Agriculture-Vision Dataset" (AV+); it will be made publicly available.

One characteristic of remote sensing is data revisiting: capturing images from the same locations over multiple times. Through data revisiting, the temporal information can serve as an additional dimension of variation beyond the spatial information alone. In AG+, a typical revisit time ranges from seven days to six months, capturing a field at different points during the pre-planting, growing, and harvest seasons. We provide an example of a revisit in Figure 3.

3.3 Fine-Grained Segmentation Dataset

Fine-grained segmentation tasks for high-resolution remote sensing data, particularly for agriculture, are often overlooked because of the difficulty in collecting sufficient amounts of annotated dataMonteiro & von Wangenheim (2019); Haug & Ostermann (2015). To explore the transferability of the AG+ dataset and SSL methods to a very challenging, data-limited, in-domain (i.e. same sensor and geography) task, we construct a densely annotated dataset. We collected 68 flights from the 2020 growing season that were not included in AV+ for this task. From these flights, 184 tiles with shape 1500×1500 were selected and densely annotated with four classes: soil, weeds, crops, and un-managed area (e.g. roads, trees, waterways, buildings); an "ignore" label was used to exclude pixels which may unidentifiable due to image collection issues, shadows, or clouds. The annotations in this dataset are much more fine-grained than those in the AV+ dataset. For example, whereas the AV+ dataset identifies regions of high weed density as a "weed cluster", this dataset identifies each weed individually at the pixel level and also labels any crop or soil in those regions by their appropriate class. A sample image and annotation are shown in Figure 1 (right).

The fine-grained nature and small dataset size make this a very challenging segmentation task: very young crop often looks like weeds, weeds growing among mature crop are only detectable through an interruption in the larger structure of the crop row, unmanaged areas often contain grasses and other biomass which closely resemble weeds but is not of concern to the grower, and classes are highly imbalanced.

4 Methodology for Benchmarks

In this section, we present multiple methods for pre-training a transferable representation on the AV+ dataset. These methods include MoCo-V2 Chen et al. (2020b), MoCo-V2 with a Pixel-to-Propagation Module (PPM) Xie et al. (2021b), the multi-head Temporal Contrast based on SeCo Mañas et al. (2021), and a combined Temporal Contrast model with PPM. We also explore different backbones based on ResNet He et al. (2016) and the Swin Transformer architecture Liu et al. (2021a).

4.1 Momentum Contrast

MoCo-V2 is employed as the baseline module for the pre-training task. Unlike previous work focusing only on RGB channels He et al. (2020); Mañas et al. (2021); Chen et al. (2020b), we include the information and learn representations from RGB and NIR channels. In each training step of MoCo, a given training example x is augmented into two separate views, query x^q and key x^k. An online network and a momentum-updated offline network, map these two views into close embedding spaces $q = f_q(x^q)$ and $k^+ = f_k(x^k)$ accordingly; the query q should be far from the negative keys k^- coming from a random subset of data samples different from x. Therefore, MoCo can be formulated as a form of dictionary lookup in which q and k are the positive and negative keys. We define the instance-level loss $\mathcal{L}_{inst}$ with temperature parameter τ for scaling Wu et al. (2018) and optimize the dictionary lookup with InfoNCE Oord et al. (2018):

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

4.2 Momentum Contrast with Pixel-to-Propagation Module

Compared with classical datasets such as ImageNet Deng et al. (2009), COCO Lin et al. (2014), and LVIS Gupta et al. (2019) in the machine learning community, low-level semantic information from AV+ is more abundant, with regions of interest corresponding more closely to "patterns" (i.e. areas of weed clusters, nutrient deficiency, storm damage) and less to individual instances. Therefore, pre-training MoCo-V2 beyond image-level contrast should be beneficial to downstream pattern analysis tasks.

Xie et al. (2021) Xie et al. (2021b) added a Pixel-Propagation-Module (PPM) to the SimCLR framework and achieved outstanding results on dense downstream tasks. In PPM, the feature of a pixel x_i is smoothed

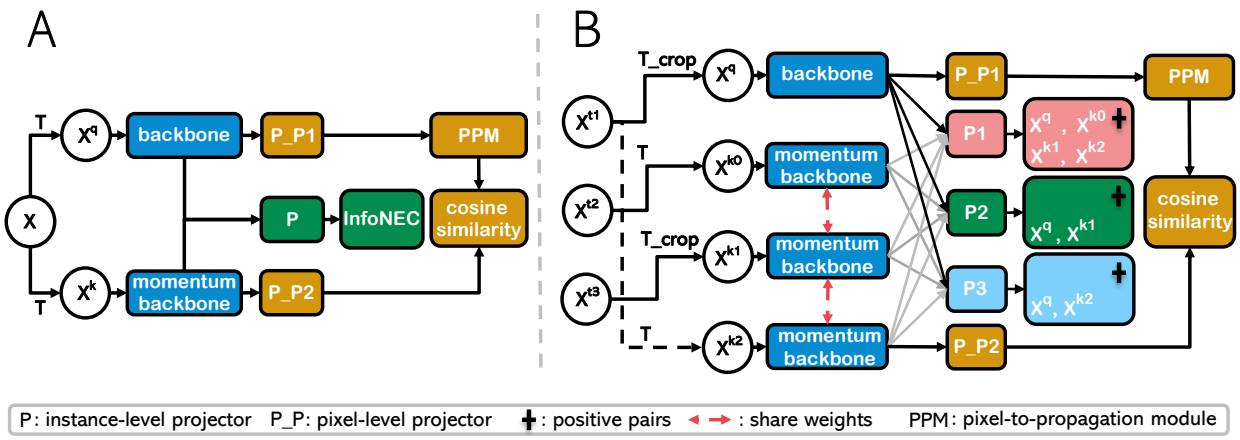

Figure 2: **A**. Diagram of MoCo-V2 with Pixel-to-Propagation Module (MoCo-PixPro). P includes a normally updated projector and a momentum updated projector. For pixel-level pre-task, P_P1 is updated by gradient descent and P_P2 is momentum projector. **B**. Diagram of Temporal Contrast with Pixel-to-Propagation Module (TemCo-PixPro). Query view x^q and key view x^{k_0} contain both artificial and temporal variance. Query view x^q and key view x^{k_1} contain only temporal variance. Query view x^q and key view x^{k_2} only contain artificial variance. Identical cropping T_{crop} is applied to x^{t_1} and x^{t_3}. Pixel-level contrast in only computed on x^q and x^{k_2}.

to y_i by feature propagation with all pixels x_j within the same image I following the equation:

$$y_i = \sum_{x_j \in I} S(x_i, x_j) \cdot G(x_j), \tag{2}$$

where $G(\cdot)$ is a transformation function instantiated by linear and ReLU layers. $S(\cdot, \cdot)$ is a similarity function defined as

$$S(x_i, x_j) = (max(cos(x_i, x_j), 0))^{\gamma} \tag{3}$$

with a hyper-parameters γ to control the sharpness of the function. However, SimCLR requires a large batch size, which is not always achievable. To generalize the PPM and make the overall pre-training model efficient, we incorporate the pixel-level pretext tasks into basic MoCo-V2 models to learn dense feature representations. As demonstrated in Figure 2A, we add two extra projectors for pixel-level pretask compared with MoCo-V2. The features from the backbones are kept as features maps instead of vectors to ensure pixel-level contrast. With those two projectors, we can compute the similarity between two pixel-level feature vectors, i.e., smoothed q_i^s from PPM and k_j for each positive pair of pixels i and j. Since two augmentation views both pass the two encoders, we use a loss in symmetric form following Xie et al. (2021b):

$$\mathcal{L}_{PixPro} = -cos(q_i^s, k_j) - cos(q_j^s, k_i) \tag{4}$$

During the training, the loss $\mathcal{L}_{PixPro}$ from the PPM is integrated with the instance-level loss as show in the equation 5. These two complementary losses are balanced by a factor α, set to 0.4 in all the experiments experiments (see Supplemental: Additional Results- Balance Factor).

$$\mathcal{L} = \alpha \mathcal{L}_{inst} + \mathcal{L}_{PixPro} \tag{5}$$

Figure 3: Visualization of temporal contrast in AG+.

4.3 Temporal Contrast

While a pixel-level pretext task learns representations useful for spatial inference, we would further like to learn a representation that takes advantage of the temporal information structure of AV+. Following the work of SeCo Mañas et al. (2021), additional embedding sub-spaces that are invariant to time are created. Since the backbones learn temporal-aware features through extra sub-spaces, it offers a more precise and general pattern analysis in downstream tasks. More specifically, we define a positive temporal pair by obtaining one pair of images from the same area (i.e. of the same field) but at different times as shown in Figure 3. We explore whether the structure provided by the temporal alignment of positive temporal pairs provides more semantically significant content than naive artificial transformations (i.e. flipping, shifting) for pre-training.

Unlike in SeCo where images were separated with a constant time (3 months), the time difference between images from our data varies from 1 week to 5 months. We adapt SeCo as follows. First, we randomly select three tiles with 512×512 from the same field at identical locations but different times, which will be defined as x^{t_1}, x^{t_2} and x^{t_3}. Only random cropping T_{crop} is applied to the query image to generate the query view, i.e., $x^q = T_{crop}(x^{t_1})$. The first key view that contains both temporal and artificial variance is defined as $x^{k_0} = T(x^{t_1})$, where the T is the typical data augmentation pipeline used in MoCo. The second key contains only temporal augmentation compared with the query view. Therefore, we apply the exact same cropping window applied to the query image, $x^{k_1} = T_{crop}(x^{t_2})$. The third key contains only artificial augmentations, $x^{k_2}=T(x^{t_0})$. Following the MoCo and SeCo learning strategy He et al. (2020); Mañas et al. (2021), these views can be mapped into three sub-spaces that are invariant to temporal augmentation, artificial augmentation and both variance. In this way, we fully explore the multi-time scale information in AV+ to improve the temporal sensitivity of encoders further. Since the temporal contrast does not necessarily cross seasons or enforce alignment of seasonality within a sub-space, we denote our approach as Temporal Contrast (TemCo).

4.4 Temporal Contrast with Pixel-to-Propagation Module

We create an integrated model (TemCo-PixPro) to capture the dense, spatiotemporal structure of AV+. Concretely, we merge PPM and TemCo into a single model to increase the encoders' spatial and temporal sensitivity.

To ensure efficient computation, we do not compute pixel-wise contrastive updates in each temporal sub-space. Instead, we assign two extra projectors for pixel-level contrastive learning. We include the PPM after the online backbone and one of the pixel-level projectors to smooth learned features. Then, we calculate the

similarity of the smooth feature vectors and the momentum encoder features through a dot product. We illustrate the overall architecture of this model in Figure 2B.

4.5 Swin Transformer-Based Momentum Contrast

While the Swin Transformer achieves superior performance on various computer vision tasks Liu et al. (2021a;b), only very recent work has focused on self-supervised training for vision transformers (ViT)Xie et al. (2021a); Li et al. (2021). To the best of our knowledge, no study has investigated Swin Transformer's performance on remote sensing datasets using self-supervised methods. Therefore, we explore a Swin Transformer-based MoCo for pre-training of AV+. Specifically, we adopt the tiny version of the Swin Transformer (Swin-T) as the default backbone.

Following most transformer-based learning tasks, we adopt AdamW Kingma & Ba (2014) for training. Additionally, we incorporate the multiple-head projectors from TemCo and PPM to capture temporal knowledge and pixel-level pretext tasks.

5 Experiments and Results

We benchmark the performance and transferability of the learned representations on four downstream tasks: agricultural pattern classifications, semantic segmentation on AV+, fine-grained semantic segmentation and land-cover classification.

We report the implementation details and results from four models covering two kinds of backbones: basic Momentum Contrast model (MoCo-V2), basic Momentum Contrast with Pixel-to-Propagation Module (MoCo-PixPro), Temporal Contrast (TemCo), Temporal Contrast with Pixel-to-Propagation Module (TemCo-PixPro), Swin Transformer-Based MoCo-V2, Swin Transformer-Based TemCo, and Swin Transformer-Based TemCo-PixPro.

5.1 Pre-training Settings

All the artificial data augmentations used in this paper are follow MoCo-V2, including random color jitter, gray-scale transform, Gaussian blur, horizontal flipping, resizing, and cropping. We train each model for 200 epochs with batch size 512. For ResNet-based models we use SGD as the optimizer with a weight decay of 0.0001, and momentum of 0.9. The learning rate is set to 0.03 initially and is divided by 10 at epochs 120 and 160. Swin-T models use the AdamW optimizer, following previous work Xie et al. (2021a); Liu et al. (2021a). The initial learning rate is 0.001, and the weight decay is 0.05.

We use all four channels, RGB and NIR, to fully extract the features contained in the dataset. When testing ImageNet-initialized backbones for comparison, we copy the weights corresponding to the Red channel of the pre-trained weights from ImageNet to the NIR channel for all the downstream tasks following the method of Chiu et al. (2020b).

5.2 Downstream Classifications Performance

The downstream classification task considers nine patterns: Nutrient Deficiency, Storm Damage, Drydown, Endrow, Double plant and Weeds. More details descriptions of these patterns can be found in the supplementary. We benchmark and verify the performance of our basic model (MoCo-V2) and its variants on classification task of the labeled portion of AV+, following three protocols: (i) linear probing He et al. (2020); Chen et al. (2020b;a); Tian et al. (2020), (ii) non-linear probing Han et al. (2020), and (iii) fine-tuning the entire network for the downstream task.

5.2.1 Linear Probing

Following standard protocol, we freeze the pre-trained backbone network and train only a linear head for the downstream task. We train the models for 50 epochs using Adam optimizer with an initial learning rate of 0.0001 and report the top-1 classification validation set.

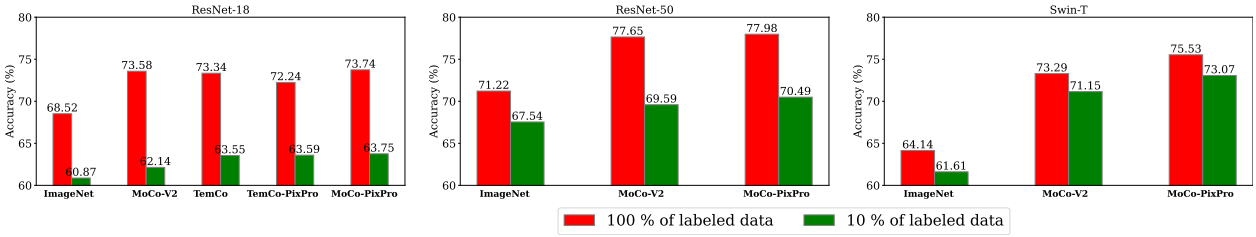

Figure 4: Accuracy under the linear probing protocol on AV+ classification. Results are shown from different pre-training approaches with different backbones (Left: ResNet-18, Middle: ResNet-50, Right: Swin-T) under different fractions of data for the downstream task (red: 10%, blue: 100%).

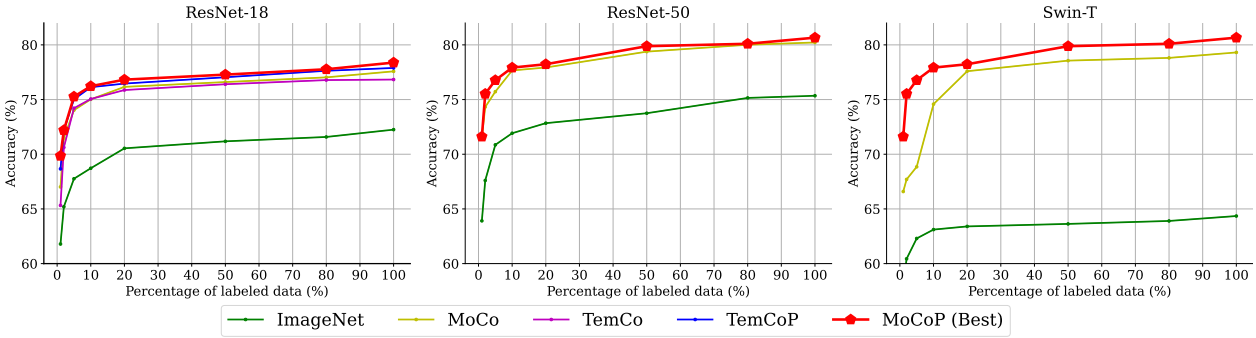

Figure 5: Accuracy under non-linear probing protocol on AV+ classification. Results are shown from different pre-training approaches with different backbones, ResNet-18 (left), ResNet-50 (middle), and Swin-T (right), under different percentages of labeled data for the downstream task.

Figure 4 shows the impact of different weight initialization and percentages of labeled data in the downstream task. Consistent with previous research Mañas et al. (2021), there is a gap between remote sensing and natural image domains: ImageNet weights are not always an optimal choice in this domain. MoCo-PixPro obtain the highest accuracy for the ResNet-18 backbone. As we compare the results of ResNet50 and Swin-T with full labeled data, all Swin-T models underperformed their CNN counterparts.

5.2.2 Non-Linear Probing

We evaluate the frozen representations with non-linear probing: a multi-layer perceptron (MLP) head is trained as the classifier for 100 epochs with Adam optimization.

Classification results on AV+ classification under non-linear probing are shown in Figure 5. Consistent with results in the natural image domain Han et al. (2020), non-linear probing results surpass linear probing. Our SSL weights exceed ImageNet's weights by over 5% regardless of the amount of downstream data or backbone type. From the results of ResNet-18, the optimal accuracy between different pre-training strategies comes from either MoCo-PixPro or TemCo-PixPro, different from linear probing. Overall, MoCo-PixPro performs better than the basic MoCo model across different backbones.

5.2.3 Fine-Tuning

Finally, we examine end-to-end fine-tuning with different percentages of labeled AV+ data for classification. We use the same architecture, learning schedule and optimizer as non-linear probing.

Our SSL weights show outstanding results in the low-data regions (<10% of data). For ResNet-18, MoCo-PixPro is better than the other models in all cases, whereas other SSL models demonstrate similar performance to ImageNet when labeled data is abundant. As we increase the backbone size to ResNet-50, our MoCo and MoCo-PixPro stably outperform ImageNet's model across all amounts of data, suggesting a greater capacity to learn domain-relevant features.

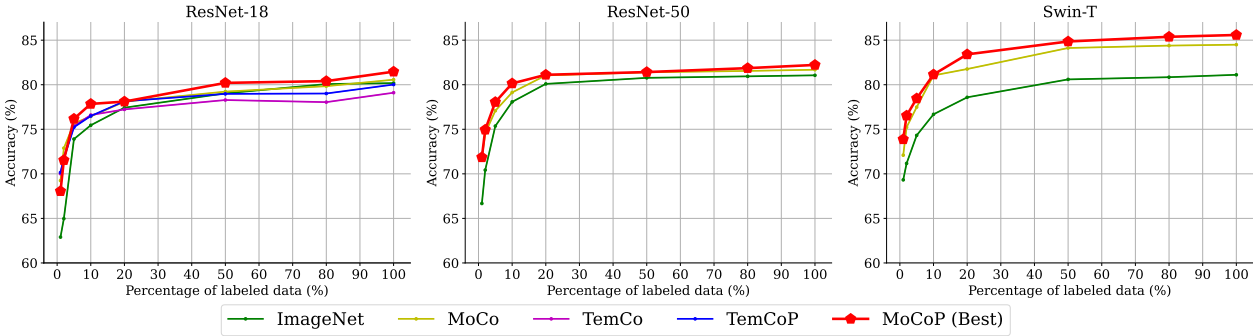

Figure 6: The accuracy under the end-to-end classification protocol on AV+. Results cover different pre-training approaches and backbones, varying from ResNet-18 (left), ResNet-50 (middle), and Swin-T (right). We also report the model's performance tuned with different percentages of the fully labeled dataset, ranging from one percent to a hundred percent.

In Figure 6 (right), all models perform agreeably well in the Swin-T framework compared with weights from ImageNet. While fine-tuning was performed in the same manner as the ResNet models for fair comparisons, Swin-T shows the most promising performance in this end-to-end setting.

5.3 Semantic Segmentation on Extended Agriculture-Vision

We continue our benchmarking study by examining their impact on the semantic segmentation approach on AV+ as originally formulate in Chiu et al. (2020b). Again we apply two protocols for evaluating the learned representations: maintaining a fixed encoder, or fine-tuning the entire network.

To naively assess the pre-trained representations, we adopt the simple yet effective U-Net Ronneberger et al. (2015) framework. Unlike previous work on AV Chiu et al. (2020b), we report results over all the patterns in AV+, including storm damage, to ensure an integrated and comprehensive analysis.

First, we evaluate the representation by holding the pre-trained encoder fixed and fine-tuning only the decoder during the supervised learning phase. Similarly, we evaluate pre-training impact on segmentation tasks allowing for fine-tuning of both the encoder and decoder during the supervised learning phase. We train the models using Adam optimization with an initial learning rate of 0.01. The one-cycle policy Smith (2017) is used to update the learning rate as in Chiu et al. (2020b). ResNet-18 models are trained for 30,000 steps while the larger ResNet-50 and Swin-T models are trained for 120,000 steps to allow for sufficient training.

5.3.1 Segmentation Results

Results are shown in Table 1. MoCo-PixPro performs best for the ResNet-18 backbone when encoder remains fixed during supervised training; this result is similar to that seen for classification. This result supports our hypothesis that AV+ has abundant low-level semantic information and including pixel-level pre-task is critical for downstream learning tasks. When the encoder is unfrozen during supervised training, the basic MoCo-V2 shows the best results, but not significantly better than TemCo or TemCo-PixPro. By scaling from ResNet-18 to ResNet-50, MoCo-PixPro outperforms ImageNet, especially when the encoder remains fixed. Importantly, unlike the ResNet-based models, the Swin Transformer-based MoCo-PixPro shows the best results across all variations in the setting.

5.3.2 Comparison with Agriculture-Vision Results

The AV dataset was benchmarked on a downstream segmentation task with architectures based on the DeepLabV3 Chen et al. (2018) framework. Because the previous results report mean Intersecion-over-Union (mIoU) for 8 agricultural patterns, we re-trained our models using a simple U-Net architecture Ronneberger et al. (2015) and without considering the excluded pattern storm damage to appropriately compare the results. With a lightweight U-Net, smaller backbone, and much less training, our SwinT-based model

Table 1: Results of Downstream Segmentation Task on AV+ using mean-IoU metric

Pretrained Weights	Backbone	mIoU (%) Fixed 1%	mIoU (%) Fixed 100%	mIoU (%) Fine-Tuned 1%	mIoU (%) Fine-Tuned 100%
Random	ResNet-18	18.89	21.37	19.02	26.94
ImageNet	ResNet-18	19.02	23.39	19.73	29.23
MoCo-V2	ResNet-18	22.36	27.83	**22.53**	**31.80**
MoCo-PixPro	ResNet-18	**23.71**	**30.60**	20.04	30.56
TemCo	ResNet-18	23.71	26.85	21.09	31.76
TemCo-PixPro	ResNet-18	22.97	28.60	21.32	31.66
Random	ResNet-50	19.42	21.82	18.71	26.37
ImageNet	ResNet-50	21.21	25.94	20.31	30.52
MoCo-V2	ResNet-50	24.25	31.03	**21.47**	**31.87**
MoCo-PixPro	ResNet-50	**25.76**	**32.35**	21.36	31.58
Random	Swin-T	15.89	20.10	22.68	37.14
ImageNet	Swin-T	20.00	22.40	30.96	43.01
MoCo-V2	Swin-T	25.51	30.60	28.12	41.02
MoCo-PixPro	Swin-T	**27.61**	**32.96**	**32.06**	**43.33**

outperforms the best results from Chiu et al. (2020b) in the Table 2, demonstrating the effectiveness of our approach.

More specifically, we train the nine-class models using an Adam optimization with an initial learning rate of 0.01 and the one-cycle policy Smith (2017) for the learning rate adjustment. For fair comparisons, we use a batch size of 40 and 25,000 training steps, consistent with the supervised learning settings in Chiu et al. (2020b).

Table 2: Comparision of mIoUs between the Agriculture-Vision model and our propsed U-Net-based model on Agriculture-Vision validation set.

Methods	Pre-trained Weights	Backbone	mIOU(%)	Number of Parameters
FPN-basedChiu et al. (2020b)	ImageNet	ResNet-101	43.40	45.11M
U-Net	MoCo-V2	Swin-T	46.15	32.40M
U-Net	MoCo-PixPro	Swin-T	**48.75**	32.40M

5.4 Fine-Grained Semantic Segmentation

Unlike AV+, this dataset is severely limited by the availability of fine-grained segmentation labels. There are 184 tiles, from 68 flights, in this dataset that are split into training (70%), validation (15%), and test (15%). Again, we use a U-Net architecture with ResNet-18 encoder. For training, we use a multi-class focal loss Lin et al. (2017) to account for the strong class imbalance.

Results are shown in Table 3 and a sample output is shown in Figure 7. Results improve across board when both encoder and decoder are fine-tuned. Although less dramatic than the results seen on the AV+ classification and segmentation tasks, some improvement over ImageNet weights is seen using the MoCo-v2 framework with ResNet-18 backbone for fixed-weights. As seen on the other tasks, when the entire network undergoes fine-tuning, the ImageNet and SSL weights, specifically MoCo-PixPro, produce roughly the same performance on the downstream task. Additional per-class analysis is provided in the Supplemental. The ResNet-50 and Swin-T models performed relatively worse compared to the ResNet-18 models which is unsurprising given the extremely small size of this dataset.

Table 3: IoU for each model in the fine-grained semantic segmentation task considering different encoder weight initializations, architectures, and weight fixing schemes.

Weights	Architecture	IoU (Fixed-Weights)	IoU (Fine-Tuned)
Random	ResNet-18	39.05	42.19
ImageNet	ResNet-18	40.81	**45.47**
MoCo-v2	ResNet-18	**44.05**	43.97
MoCo-PixPro	ResNet-18	42.03	44.62
TemCo	ResNet-18	42.30	44.48
TemCo-PixPro	ResNet-18	43.45	43.91
MoCo-v2	ResNet-50	40.03	40.54
MoCo-v2	Swin-T	40.25	40.00
MoCo-PixPro	Swin-T	37.56	40.67
TemCo.	Swin-T	39.52	40.26

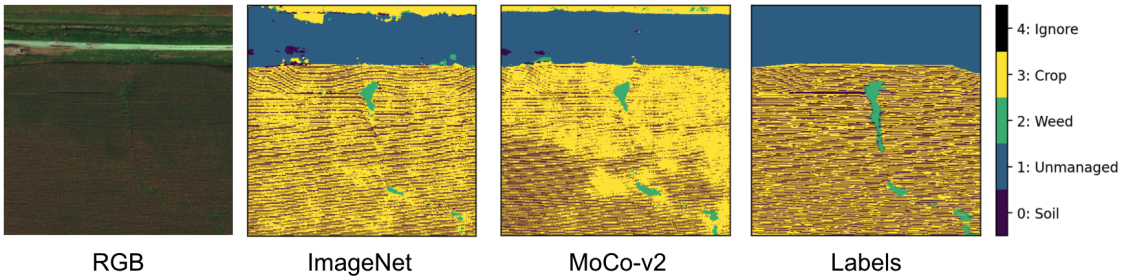

Figure 7: A sample output on the fine-grained segmentation task using fixed-encoder weights from ImageNet and MoCo-V2. The segmentation outputs are compared with both the original RGB image and the segmentation labels.

5.5 Land-Cover Classification on EuroSAT

We further prove pretraining on the AV+ dataset benefits the downstream task in the broader remote sensing community. We conduct downstream classification experiments on EuroSATHelber et al. (2019). EuroSAT addresses the classification challenge of land use and land cover with images from Sentinel-2. It consists of 27,000 labeled images and 10 classes over 34 European countries. We use the splits protocol of train/val following the work of Neumann et al. (2019); Mañas et al. (2021).

We freeze the pre-trained backbones and add a linear layer to evaluate the learned representation in this classification task. Totally, the linear layer is tunned with 100 epochs using the Adam optimizer. The initial learning rate is set to 0.001 and is divided by 10 at the 60th and 80th epochs.

The results shown in the Table 4 compare weights pre-trained from AG+ against other baselines. We notice that MoCo-V2 and our proposed MoCo-PixPro achieve 1.21% and 6.25% higher accuracy compared with ImageNet's weights accordingly. These results confirm not only the effectiveness of pre-training on AG+ but also AG+'s significant potential to generalize to the broader remote sensing field.

Table 4: Accuracy of the EuroSAT land-cover classification task using ResNet-18

Weights	Random	ImageNet	MoCo-V2	MoCo-PixPro
Accuracy (%)	63.21	86.32	87.53	**89.97**

5.6 Ablation Study: Number of Flights

We use a ResNet-18 backbone and basic MoCo-V2 for experiments. When the number of flights used for SSL is increased from 300 to 3600, we observe stable improvement in the downstream classification task under

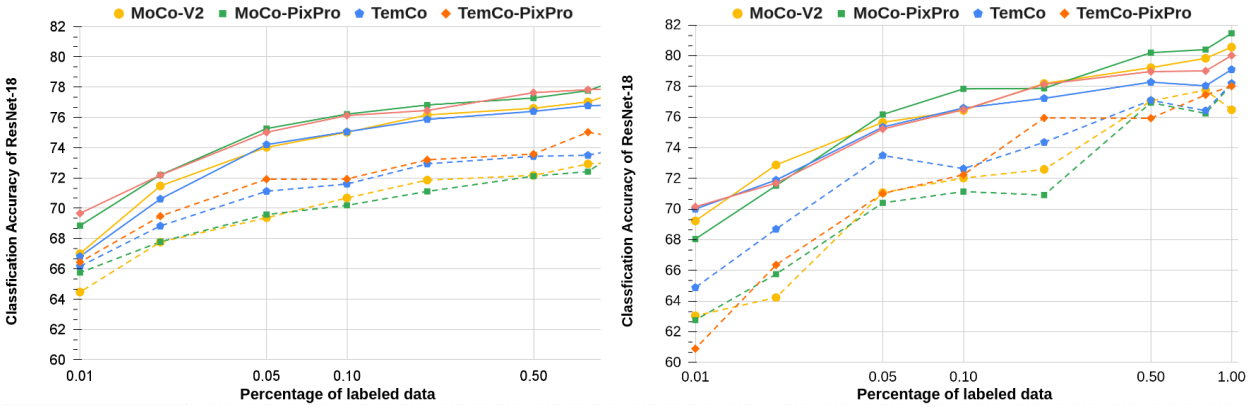

Figure 8: Ablation study on the pre-training size of data on different pre-training methods on two downstream tasks. Solid lines represent accuracy from 3600 flights while dashed lines represent accuracy from 1200 flights. Left: results from non-linear probing on downstream classification. Right: fine-tuning results on entire networks for downstream AV+ segmentation.

the non-linear probing setting; this gain is confirmed regardless of the fraction of labeled dataset for tuning. See Supplemental: Additional Results for more detailed results.

This improvement is seen for all examined SSL methods Figure 8 when the raw dataset is increased from 1200 to 3600 flights and evaluated under non-linear probing for classification and full-network fine-tuning for AV+ segmentation. Our SSL models' performance steadily grows as raw data size increases, suggesting that even more data may lead to even greater performance.

6 Conclusion

Large, high-quality datasets are opening tremendous new opportunities for computational agriculture, but they are extremely difficult to obtain. As in other domains, remote sensing and earth observation data is marked by huge amounts of unlabeled data and relatively few annotations; leveraging the information in this unlabeled data therefore becomes a critical task. In this work we contribute to the advancement of these efforts by releasing the AV+ dataset which contains annotated full-field imagery based on the original AV dataset Chiu et al. (2020b), supplemented by more than 3TB of raw full-field images taken at different times in the season. The improved supervised component of the AV dataset will allow for greater flexibility in training and augmentation protocols and enable additional possible lines of study around long-range context and large-scale imagery. The raw unlabeled data will enable continued exploration in self, semi, and weakly supervised methods which we have begun to benchmark here. This extension of an already important dataset in the computational agriculture will open up many lines of research and investigation which benefit both the agriculture and computer vision communities.

Next, we conduct a thorough benchmark study on self-supervised pre-training methods based on contrastive-learning which capture the fine-grained, spatiotemporal nature of this data. We analyze a classification formulation of the AV+ dataset under linear probing, non-linear probing, and fine-tuning. We also examined segmentation tasks, which are often overlooked in remote sensing approaches, based on the original segmentation formulation of AV+ with a frozen and unfrozen encoder and an extremely small fine-grained segmentation task under the same formulations. Our benchmark study explores both traditional CNN architectures (ResNet-18 and ResNet-50) as well as the more recent Swin Transformer, which offers unique potentials for computer vision, but requires huge amounts of data to train.

Importantly, we incorporate the Pixel-to-Propagation Module, originally built in the SimCLR framework, into the MoCo-V2 framework which allows for training on larger batch sizes. Our results show that this module is key for downstream segmentation and *classification* tasks, even though it was designed primarily for dense detection and segmentation tasks. As our dataset contains richer low-level, high-frequency, fine-grained

features than traditional natural imagery like COCO or ImageNet, this suggests that PPM is beneficial for learning dense, fine-grained *features* in addition to dense label structure.

We further combine this module with a TemCo, a modification of SeCo, into a rich framework which captures the dense, spatiotemporal structure of our data. While this combined framework was not the highest-performing on the various task, it again may have been at a disadvantage since it is a larger model and the number of steps was fixed for fair comparison. Additionally, extending how *positive* samples are generated could prove beneficial. These improvements are the focus of future analysis.

Self-supervised methods will be crucial for unlocking opportunities in remote sensing, particularly for agriculture, and this dataset release and benchmark study offers a significant step in that direction.

References

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *arXiv preprint arXiv:2011.09980*, 2020.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Antonio Barrientos, Julian Colorado, Jaime del Cerro, Alexander Martinez, Claudio Rossi, David Sanz, and Joao Valente. Aerial remote sensing in agriculture: A practical approach to area coverage and path planning for fleets of mini aerial robots. *Journal of Field Robotics*, 28(5):667–689, 2011.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36. JMLR Workshop and Conference Proceedings, 2012.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

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, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

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 Workshops*, pp. 48–49, 2020a.

Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan 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, 2020b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118, 2018.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2018.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 201–208. JMLR Workshop and Conference Proceedings, 2010.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Min Feng and Yan Bai. A global land cover map produced through integrating multi-source datasets. *Big Earth Data*, 3(3):191–219, 2019.

Anatoly A Gitelson, Yoram J Kaufman, Robert Stark, and Don Rundquist. Novel algorithms for remote estimation of vegetation fraction. *Remote sensing of Environment*, 80(1):76–87, 2002.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pp. 312–329. Springer, 2020.

Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. 2014.

Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *European conference on computer vision*, pp. 105–116. Springer, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

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.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Joshua Kelcey and Arko Lucieer. Sensor correction of a 6-band multispectral imaging sensor for uav remote sensing. *Remote sensing*, 4(5):1462–1493, 2012.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. URL `http://arxiv.org/abs/1708.02002`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021a.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021b.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3226–3229. IEEE, 2017.

Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *arXiv preprint arXiv:2103.16607*, 2021.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

A.A.O. Monteiro and A. von Wangenheim. Orthomosaic dataset of rgb aerial images for weed mapping. `http://www.lapix.ufsc.br/weed-mapping-sugar-cane`, 2019.

David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013.

Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019.

Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):1–12, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Anushree Ramanath, Saipreethi Muthusrinivasan, Yiqun Xie, Shashi Shekhar, and Bharathkumar Ramachandra. Ndvi versus cnn features in deep learning for land cover clasification of aerial images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6483–6486. IEEE, 2019.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Santhosh K Seelan, Soizik Laguette, Grant M Casady, and George A Seielstad. Remote sensing applications for precision agriculture: A learning community approach. *Remote sensing of environment*, 88(1-2):157–169, 2003.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017. doi: 10.1109/WACV.2017.58.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904, 2019. doi: 10.1109/IGARSS.2019.8900532.

Shu Tian, Lihong Kang, Xiangwei Xing, Zhou Li, Liang Zhao, Chunzhuo Fan, and Ye Zhang. Siamese graph embedding network for object detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 18(4):602–606, 2020.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Cees J Van Westen. Remote sensing and gis for natural hazards assessment and disaster risk management. *Treatise on geomorphology*, 3:259–298, 2013.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021a.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021b.

Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jingming Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10):875–883, 2013.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

A Appendix

You may include other additional sections here.