Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data

Oscar Mañas[1,2] Alexandre Lacoste[1] Xavier Giró-i-Nieto[2] David Vazquez[1] Pau Rodriguez[1]
[1]Element AI [2]Universitat Politècnica de Catalunya
oscmansan@gmail.com, pau.rodriguez@servicenow.com

Abstract

Remote sensing and automatic earth monitoring are key to solve global-scale challenges such as disaster prevention, land use monitoring, or tackling climate change. Although there exist vast amounts of remote sensing data, most of it remains unlabeled and thus inaccessible for supervised learning algorithms. Transfer learning approaches can reduce the data requirements of deep learning algorithms. However, most of these methods are pre-trained on ImageNet and their generalization to remote sensing imagery is not guaranteed due to the domain gap. In this work, we propose Seasonal Contrast (SeCo), an effective pipeline to leverage unlabeled data for in-domain pre-training of remote sensing representations. The SeCo pipeline is composed of two parts. First, a principled procedure to gather large-scale, unlabeled and uncurated remote sensing datasets containing images from multiple Earth locations at different timestamps. Second, a self-supervised algorithm that takes advantage of time and position invariance to learn transferable representations for remote sensing applications. We empirically show that models trained with SeCo achieve better performance than their ImageNet pre-trained counterparts and state-of-the-art self-supervised learning methods on multiple downstream tasks. The datasets and models in SeCo will be made public to facilitate transfer learning and enable rapid progress in remote sensing applications.[1]

1. Introduction

Remote sensing is becoming increasingly important to many applications, including land use monitoring [12], precision agriculture [29], disaster prevention [37], wildfire detection [11], vector-borne disease surveillance [20], and tackling climate change [33]. Combined with recent advances in deep learning and computer vision, there is enormous potential for monitoring global issues through the automated analysis of remote sensing and other geospatial data streams.

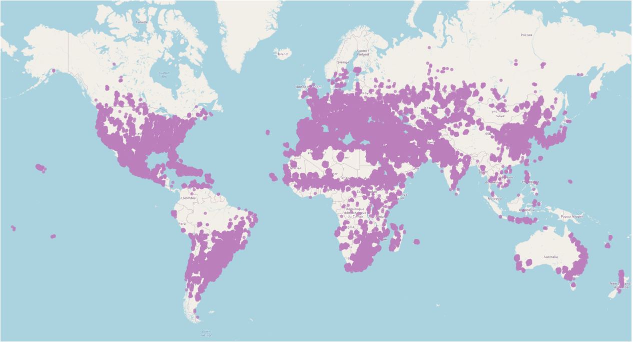

Figure 1. **Distribution of the Seasonal Contrast (SeCo) dataset.** Each point represents a sampled location. Images are collected around human settlements to avoid monotonous areas such as oceans and deserts.

Remote sensing provides a vast supply of data. The number of Earth-observing satellites is continuously growing, with over 700 satellites currently in orbit generating terabytes of imagery data every day [30]. However, many downstream tasks of interest are constrained by a lack of annotations, which are particularly costly to obtain since they often require expert knowledge, or expensive ground sensors. In recent years, a number of techniques have been developed to mitigate the need for labeled data [24, 26, 25], but their application to remote sensing images is largely underexplored.

Furthermore, existing remote sensing datasets [38, 19, 42] are highly curated to form well-balanced and diversified classes. Simply discarding the labels does not undo this careful selection of examples, which also requires considerable human effort. Our goal is to exploit the massive amount of publicly available remote sensing data for learning good visual representations in a truly unsupervised way. To enable this, we construct a remote sensing dataset from Sentinel-2 [10] tiles without any human supervision, neither for curating nor annotating the data.

[1] Code, datasets and pre-trained models are available at https://github.com/ElementAI/seasonal-contrast

Another characteristic unique to remote sensing data is satellite revisit, which describes the ability of the system to make repeated image captures of the same point of the Earth's surface over time. For publicly funded satellite constellations such as Sentinel [10] or Landsat [35], the revisit time is of the order of days. This temporal dimension provides an additional source of natural variation which complements the artificial augmentation of images. For instance, no amount of artificial augmentation can show how a snowy mountain summit looks like when the snow melts down, or how the different stages of a crop change through the seasons.

Self-supervised learning methods have recently emerged as an effective methodology to learn from vast amounts of unlabeled data. Contrastive methods push together representations of images that are semantically similar (i.e. positive pairs). Since no labels are available, traditional contrastive learning methods that work with natural images use different artificial augmentations of the same image (views) as positive pairs. In the case of remote sensing images, we propose to leverage the temporal information to obtain pairs of images from the same location at different points in time, which we call *seasonal positive pairs*. We argue that seasonal changes provide more semantically meaningful content than artificial transformations, and remote sensing images provide this natural augmentation for free.

We propose Seasonal Contrast (SeCo), a novel methodology for pre-training rich, transferable representations for remote sensing applications. SeCo consists of two parts, an unsupervised data acquisition procedure and a self-supervised model to learn from the acquired data. The self-supervised learning model is designed based on the observation that encouraging the representation to be invariant to seasonal changes is a strong inductive bias. This property can be beneficial for certain downstream tasks where the prediction will not change with seasonal variations (*e.g.*, land-cover classification, agricultural pattern segmentation, building detection), but harmful for downstream tasks where seasonal variations are important (*e.g.*, deforestation tracking, change detection). We want to learn good representations of remote sensing images that are agnostic to the downstream tasks where they could be applied.

To leverage temporal information without limiting the visual representations to be always invariant to time, we use the idea of multiple embedding sub-spaces [47]. Instead of mapping an image to a single embedding space which is invariant to all augmentations, we construct separate embedding sub-spaces and optimize them to be variant or invariant to seasonal changes. We use a multi-head architecture with a shared backbone which produces a common representation that encodes the different variances and invariances. Once the model is trained, this representation can be applied to a wide range of remote sensing downstream tasks, where the model can selectively utilize the different factors of variation captured in the representation.

We evaluate SeCo on several remote sensing datasets and tasks. Our experiments on land-cover classification with BigEarthNet [38] and EuroSAT [19], and change detection with OSCD [8] demonstrate that SeCo pre-training is more effective for remote sensing tasks than the common ImageNet [36] and MoCo [18] pre-training.

In summary, our contributions are:

- We describe a general method for collecting uncurated and unlabeled datasets of remote sensing images. We use this method to construct a remote sensing dataset from Sentinel-2 tiles without any human supervision.
- We combine recent contrastive self-supervised learning methods with the temporal information provided by satellites to learn good visual representations which are simultaneously variant and invariant to seasonal changes.
- We obtain state-of-the-art results on BigEarthNet and EuroSAT land-cover classification, and on OSCD change detection.

2. Background

Self-supervised learning is the branch of unsupervised learning where the data itself provides the supervision. The main idea is to occlude or perturb part of the data and task the network with predicting it from the visible data. This defines a pretext task (or proxy loss) and the network is forced to learn what we care about in the data (*e.g.*, a semantic representation) in order to solve it. A variety of pretext tasks have been proposed for images, such as predicting the relative position of patches [9], solving jigsaw puzzles [31], predicting rotations [13] or colorization [48].

More recently, contrastive pretext tasks [46, 32, 41, 18, 28, 5, 16, 4] have dominated the subfield of self-supervised learning, demonstrating superior performance in various downstream tasks. Intuitively, contrastive learning methods pull together the representations of similar examples while pushing apart the representations of dissimilar examples. Since the examples are not labeled, these methods make the assumption that each example defines and belongs to its own class. Hence, positive pairs are generated by applying random augmentations to the same example, while negative pairs come from other instances in the dataset.

Formally, this task can be formulated as a dictionary look-up problem, where a given example x is augmented into two views, query x^q and key x^k, an encoder network f maps the examples into an embedding space, and the representation of the query $q = f(x^q)$ should be closer to the representation of its designated key $k^+ = f(x^k)$ than to the representation of any negative key k^- coming from a set of randomly sampled instances different from x. To this end, a contrastive objective is optimized over a batch of positive/negative pairs. A common choice is the InfoNCE loss [32]:

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

where τ is a temperature hyper-parameter scaling the distribution of distances.

3. Method

We propose a methodology for pre-training rich, transferable representations for remote sensing imagery, consisting of a general procedure for collecting an unsupervised pre-training dataset (Section 3.1) and a self-supervised learning method (Section 3.2) for leveraging this data.

3.1. Unsupervised Dataset Collection

Remote sensing provides a vast amount of imagery data, but annotations are usually scarce, and domain expertise or ground sensors are often required [21]. In order to train on a large amount of satellite images, we collect a new dataset of Sentinel-2 [10] patches without any human supervision.

The Sentinel-2 imagery consists of 12 spectral bands (including RGB and NIR) at 10 m, 20 m and 60 m resolution, with a revisit time of around 5 days. We use Google Earth Engine [15] to process and download image patches from about 200K locations around the world, where each patch covers a region of roughly 2.65×2.65 km. At each location, we download 5 images from different dates separated by approximately 3 months, which capture the seasonal changes that occurred in the region over a year. To avoid getting images from the same periods of the year, at each location we jitter the dates for up to a year. We also filter out Sentinel-2 tiles with a cloud percentage higher than 10%. In total, we obtain about 1 million multi-spectral image patches, which amount to a total of over 387 billion pixels.

Sampling Strategy Our objective is to learn an encoder that can be used on a wide variety of downstream tasks. To this end, we need to sample from a wide variety of regions on the Earth. Uniform sampling would lead to a large amount of redundancy in the types of images. For example, oceans cover 71% of the planet, forests cover 31% of land, and deserts cover 33% of land. To work around this, we make the assumption that most of the variability can be observed in the greater areas around cities. The cities themselves contain a wide range of constructions, a few kilometers away from cities we often observe a variety of crops and industrial facilities. Finally, in the range of 50 km-100 km away from cities, we usually observe natural environments. Hence, we sample around cities following this heuristic (see results in Figure 1):

1. Sample uniformly among the 10k most populated cities, and then sample a set of coordinates from the Gaussian distribution spanning a standard deviation of 50 km around the center of the city.
2. Randomly select a reference date over the past year. Add periodic increments of 3 months to obtain the sampling dates.
3. For a 15-day range around each date, check if there exists a Sentinel-2 tile with less than 10% of cloud coverage that intersects with the coordinates.
4. If there exists a valid Sentinel-2 tile for this location on all dates, process and download all the image patches. Otherwise, go to step 1.

We do not perform any additional data cleaning to ensure that the obtained images are diverse, informative and free of clouds. Because our dataset is constructed automatically, we can easily gather more data (more locations, more dates per location). In this work, however, we limit the scale to a total of 1M images to make it more comparable to ImageNet [36].

3.2. Seasonal Contrast

Given an unsupervised dataset of remote sensing images with temporal information, we learn a representation that takes advantage of the structure of the data. We get inspiration from [47] to develop a multi-augmentation contrastive learning method. This approach can selectively prevent information loss incurred by artificial augmentations, and extend it with natural augmentations provided by the seasonal changes on remote sensing images. Instead of projecting every view to a common embedding space which is invariant to all augmentations, a common representation is projected into several embedding sub-spaces which are variant or invariant to time (see Figure 2). Hence, the shared representation will contain both time-varying and invariant features, which will transfer efficiently to remote sensing downstream tasks regardless of whether they involve temporal variation.

3.2.1 Views Generation

Given a reference image (query), we produce multiple positive pairs (keys) with seasonal and artificial augmentations. Let $\mathcal{T}$ be a set of commonly used artificial augmentations [18], such as random cropping, color jittering, and random flipping. We first obtain 3 images from the same location at different times, x^{t_0}, x^{t_1} and x^{t_2}, which are randomly selected among all the available ones for that location. No additional transformations are applied to the query image, i.e. $x^q = x^{t_0}$. Hence, x^{t_1} and x^{t_2} can be considered seasonal augmentations (or temporal views) of x^q. The first key view contains both seasonal and artificial transformations, $x^{k_0} = \mathcal{T}(x^{t_1})$, the second key view contains only

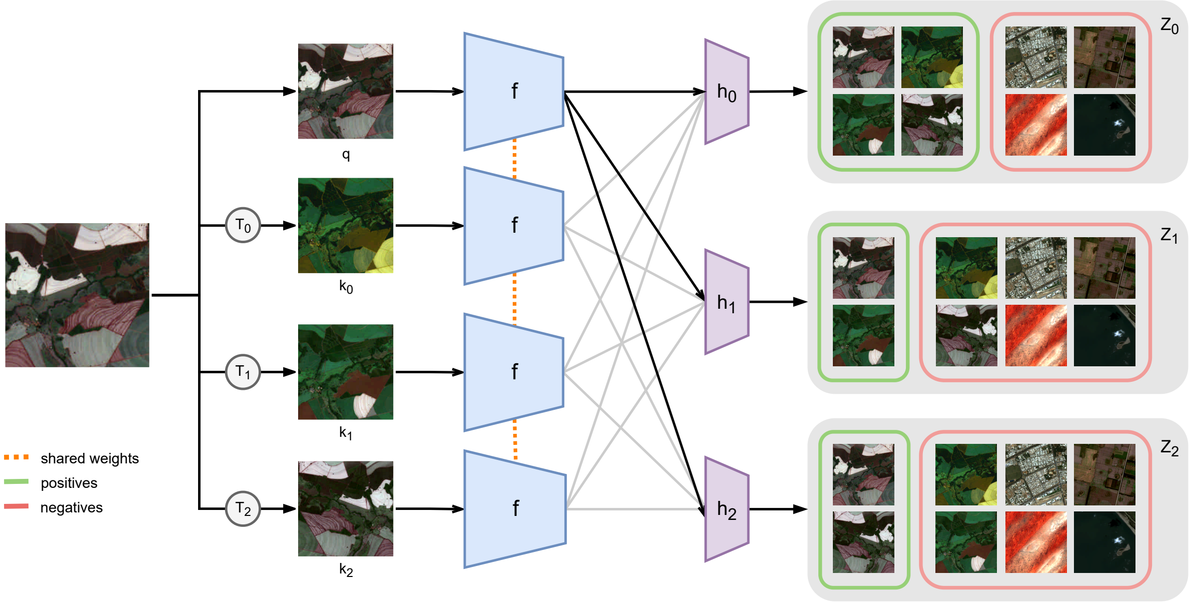

Figure 2. **Diagram of the Seasonal Contrast method.** A query image (q) is augmented with temporal (k_0, k_1) and synthetic (k_0, k_2) transformations $\mathcal{T}$. Image embeddings produced by the encoder f are projected into three different sub-spaces by heads h_0, h_1, h_2. Green boxes represent positive pairs while red boxes represent negative pairs (i.e. including images from other locations). Sub-space $\mathcal{Z}_0$ is invariant to all transformations, thus all keys belong to the same class as the query. $\mathcal{Z}_1$ is invariant to seasonal augmentations, while $\mathcal{Z}_2$ is invariant to synthetic augmentations.

seasonal augmentations, $x^{k_1} = x^{t_2}$, and the third view contains only artificial augmentations, $x^{k_2} = \mathcal{T}(x^{t_0})$.

3.2.2 Multiple Embedding Sub-spaces

The query and key views are encoded by a neural network f into representations v^q, v^{k_0}, v^{k_1}, v^{k_2} in a common embedding space $\mathcal{V} \in \mathbb{R}^d$. Next, each intermediate representation is projected into 3 different sub-spaces $\mathcal{Z}_0, \mathcal{Z}_1, \mathcal{Z}_2 \in \mathbb{R}^{d'}$ by non-linear projection heads h_0, h_1, h_2, where $h_i : \mathcal{V} \mapsto \mathcal{Z}_i$. Following recent literature [44], the embedding sub-spaces are l_2-normalized, effectively restricting them to the unit hypersphere.

The embedding sub-space $\mathcal{Z}_0$ is invariant to all augmentations, $\mathcal{Z}_1$ is invariant to seasonal augmentations but variant to artificial augmentations, and $\mathcal{Z}_2$ is invariant to artificial augmentations but variant to seasonal augmentations. Namely, in $\mathcal{Z}_0$ all embeddings z_0^i should be pulled together, in $\mathcal{Z}_1$ only z_1^q and $z_1^{k_1}$ should be pulled together and pushed apart from $z_1^{k_0}$ and $z_1^{k_2}$, and in $\mathcal{Z}_2$ only z_2^q and $z_2^{k_2}$ should be pulled together and pushed apart from $z_2^{k_0}$ and $z_2^{k_1}$. This is represented visually in Figure 2.

A contrastive learning objective is optimized on each embedding sub-space based on Equation 1, where the definition of positive (and negative) pairs depends on the invariances (and variances) that are encoded. In $\mathcal{Z}_0$, the positive pair for the query z_0^q is $z_0^{k_0}$, and the negative pairs are embeddings of other instances in this embedding sub-space. For embedding sub-space $\mathcal{Z}_1$, the positive pair for the query z_1^q is $z_1^{k_1}$, while the negative pairs are embeddings of other instances in this embedding sub-space, plus $z_1^{k_0}$ and $z_1^{k_2}$. Note that $z_1^{k_0}$ and $z_1^{k_2}$ are harder negative pairs for z_1^q as they come from the *same* instance but have a different artificial augmentation. Positive and negative pairs in embedding space $\mathcal{Z}_2$ are analogous to $\mathcal{Z}_1$.

The final learning objective is the sum of all the embedding sub-space losses. The encoder f must preserve time-varying and invariant information in the general embedding space $\mathcal{V}$ in order to optimize the combined contrastive learning objectives of all normalized embedding sub-spaces $\mathcal{Z}_i$. Note that the original contrastive learning objective [32] is a particular case of multi-augmentation contrastive learning when only the embedding sub-space $\mathcal{Z}_0$ is used.

The representation for transfer learning is taken from the general embedding space $\mathcal{V}$, since we do not assume any *a priori* knowledge about the downstream tasks. In case the right invariances for downstream tasks were known, the representation could be extracted from a particular embedding sub-space $\mathcal{Z}_i$ (see the supplementary material).

4. Experiments

In this study, we evaluate the learned representations on three downstream tasks: two land-cover classification tasks,

		100k images				1M images			
Pre-training	Backbone	Linear probing		Fine-tuning		Linear probing		Fine-tuning	
		10%	100%	10%	100%	10%	100%	10%	100%
Random init.	ResNet-18	43.05	45.95	68.11	79.80	43.05	45.95	68.11	79.80
ImageNet (sup.)		65.69	66.40	78.76	85.90	65.69	66.40	78.76	85.90
MoCo-v2	ResNet-18	69.70	70.90	78.76	85.17	69.28	70.79	78.33	85.23
MoCo-v2+TP		70.20	71.08	79.80	85.71	72.58	73.60	80.68	86.59
SeCo (ours)		**74.67**	**75.52**	**81.49**	**87.04**	**76.05**	**77.00**	**81.86**	**87.27**
Random init.	ResNet-50	43.95	46.92	69.49	78.98	43.95	46.92	69.49	78.98
ImageNet (sup.)		70.46	71.82	80.04	86.74	70.46	71.82	80.04	86.74
MoCo-v2	ResNet-50	71.85	73.27	79.23	85.79	73.71	75.65	80.08	86.05
MoCo-v2+TP		72.61	73.91	79.04	85.35	74.50	76.32	80.20	86.11
SeCo (ours)		**77.49**	**79.13**	**81.72**	**87.12**	**78.56**	**80.35**	**82.62**	**87.81**

Table 1. Mean average precision on the BigEarthNet land-cover classification task. Results cover different pre-training approaches and different ResNet backbones. We also explore the effect of the unlabeled pre-training set size between 100k and 1M images, and the size of the BigEarthNet training set between 10% and 100%.

where the representation should be invariant to seasonal changes, and a change detection task, where the representation should be variant to seasonal changes.

Pre-training Implementation Details We adopt Momentum Contrast (MoCo-v2) [6] as the backbone for our method due to its combination of state-of-the-art performance and memory efficiency. We apply the same artificial augmentations as MoCo-v2, i.e. color jittering, random grayscale, Gaussian blur, horizontal flipping, and random-resized cropping. We use a ResNet [17] architecture as the feature extractor, and a 2-layer MLP head with a ReLU activation and 128-dimensional output for each embedding sub-space. We also use separate queues [18] for each embedding sub-space, containing 16,384 negative embeddings at a time. We pre-train the network for 200 epochs with a batch size of 256. We use an SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. We set an initial learning rate of 0.03 and divide it by 10 at 60% and 80% of the epochs. A temperature scaling τ of 0.07 is used in the contrastive loss. Although the collected dataset contains up to 12 spectral bands, in this work we focus on the RGB channels since it is a more general modality.

Methods We compare our unsupervised learning approach against several baselines, including random initialization, ImageNet supervised pre-training, and self-supervised pre-training. For the latter, we provide results for MoCo-v2 pre-training on our unsupervised dataset without exploiting the temporal information. In this case, the length of the dataset depends on the total number of images and not the number of geographical locations, so we divide the number of pre-training epochs by the number of images per location. We also provide results for MoCo-v2 pre-training on our dataset leveraging the temporal information for generating positive pairs (MoCo-v2+TP), i.e. positive image pairs come from the same location at different times, and MoCo-v2 artificial augmentations are then applied to the spatially aligned image pairs (similar to Ayush et al. [1]). Note that SeCo sees more sample views than the baselines since it uses more embedding sub-spaces. We evaluate all methods with linear probing (freezing the encoder and training only the classifier) and fine-tuning (updating the parameters of both the encoder and the classifier).

4.1. Land-Cover Classification on BigEarthNet

BigEarthNet [38] is a challenging large-scale multi-spectral dataset of Sentinel-2 [10] images, captured with similar sensors to the ones in our unsupervised dataset, i.e. 12 frequency channels (including RGB) are provided. It consists of 125 Sentinel-2 tiles acquired between June 2017 and May 2018 over 10 European countries, which are divided into 590,326 non-overlapping image patches, each covering an area of 1.2 × 1.2 km with resolutions of 10 m, 20 m, and 60 m per pixel. We discard about 12% of the patches which are fully covered by seasonal snow, clouds or cloud shadows. This is a multi-label dataset where each image is annotated by multiple land-cover classes, so we measure the downstream performance in terms of mean average precision (mAP). We adopt the new class nomenclature introduced in [39], and we use the same train/val splits proposed in [30].

Implementation Details We evaluate the learned representations by training a linear classification layer with supervised learning. We initialize the ResNet backbone with a pre-trained representation and add a single fully-connected layer which maps from the intermediate representation to class logits. We fine-tune the network for 100 epochs with a batch size of 1024, and report the best validation results for each run. We use an Adam optimizer with default hyperparameters. We set the initial learning rate to 1e-3 and 1e-5 for linear probing and full fine-tuning, respectively. During

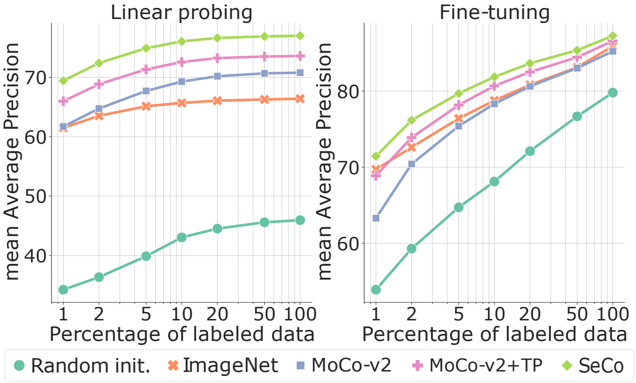

Figure 3. Label-efficient land-cover classification on BigEarthNet. We use a ResNet-18 backbone pre-trained on 1M images.

training, the learning rate is divided by 10 at 60% and 80% of the epochs.

Quantitative Results Table 1 compares the accuracy of SeCo pre-training on BigEarthNet with other pre-training methods. The comparison is done by linear probing or fine-tuning with different backbones, number of pre-training images, and percentage of BigEarthNet labeled data available. For linear probing, we observe that SeCo consistently outperforms MoCo-v2+TP. We also observe that temporal positives (TP) improve the performance of MoCo-v2 by a narrow margin. We find that SeCo features significantly improve over ImageNet pre-trained features, which confirms our hypothesis that there is a gap between remote sensing and natural image domains. We also find that this gap decreases when fine-tuning an ImageNet pre-trained feature extractor on the whole BigEarthNet training set. Nonetheless, with 1M images and a ResNet-50 backbone, SeCo features achieve 1.1% higher accuracy than ImageNet features. To the best of our knowledge, this is the first time an unsupervised method obtains higher accuracy than ImageNet pre-training on BigEarthNet with 100% of the labels. Regarding the backbone size, we observe a wider performance gap between ResNet-18 and ResNet-50 when linear probing than when fine-tuning the whole network. We also find that pre-training with 1M images yields better performance regardless of the backbone used. In all cases, we find that SeCo is more efficient than the baselines when only using 10% of BigEarthNet's labeled data; we provide more details in the next section. In the supplementary material, we include more results comparing to similar published works.

Study on Label-Efficient Transfer Learning Figure 3 shows the linear probing and fine-tuning performance of SeCo and the different baselines for different percentages of labeled data on BigEarthNet. For linear probing, we observe that, with only 1% of the BigEarthNet labels, SeCo outperforms ImageNet pre-training with 100% of the labels and matches MoCo-v2 with 20% of the labels. We also observe that the gap between ImageNet pre-training and self-supervised learning increases with the amount of labeled data, while the gap between self-supervised methods does not change significantly. For all percentages of labeled data, SeCo achieves a constant ~4% improvement gap with respect to MoCo-v2+TP. When fine-tuning, the performance gap between self-supervised methods and ImageNet narrows down when increasing the percentage of labeled data. Nevertheless, SeCo is more label-efficient than all the baselines, matching the performance of ImageNet pre-training using all the available labels with only 50% of the labels.

Sampling	Linear probing		Fine-tuning	
	10%	100%	10%	100%
Gaussian	**74.67**	**75.52**	**81.49**	**87.04**
Uniform	71.63	72.59	79.65	85.75

Table 2. Comparison of SeCo dataset sampling strategies. We use a ResNet-18 backbone pre-trained on 100k images.

Ablation on the Locations Sampling Strategy In order to evaluate the effectiveness of our sampling strategy to collect uncurated images for pre-training remote sensing representations, we download an alternative version of the SeCo dataset where Earth locations are sampled uniformly from within the continents. We download 100k images following this approach and pre-train a ResNet-18 with SeCo. Table 2 compares transfer learning performance on the BigEarthNet downstream task when using each sampling scheme. We observe that a SeCo representation pre-trained on images sampled from a mixture of Gaussians centered around human settlements (see Section 3.1) provides better downstream performance than sampling the images uniformly. We argue this is because populated regions tend to be more diverse due to human activity and thus collected images contain more information for learning good representations. In the supplementary material, we provide more results when restricting the geographical area of sampling.

4.2. Land-Cover Classification on EuroSAT

EuroSAT [19] also addresses the challenge of land use and land cover classification using Sentinel-2 satellite images. The images correspond to 34 European countries, and they consist of 10 classes corresponding to different land uses. Each of the classes is composed of 2,000 to 3,000 images, making a total of 27,000 labeled images. The images size is 64 × 64 pixels, covering an area of 640 × 640 m. All 13 Sentinel-2 spectral bands are included. We adopt the same train/val splits proposed in [30].

Implementation Details On this task, we also evaluate the learned representations by learning a linear classifier with supervised learning. We initialize a ResNet-18 backbone with a pre-trained representation and add a single fully-connected layer on top. In this case, we initialize

Pre-training	Accuracy
Random init.	63.21
Imagenet (sup.)	86.44
MoCo-v2	83.72
MoCo-v2+TP	89.51
SeCo (ours)	**93.14**

Table 3. Fine-tuning accuracy on the EuroSAT land-cover classification task. We use a ResNet-18 backbone pre-trained on 1M images.

Pre-training	Precision	Recall	F1
Random init.	**70.53**	19.17	29.44
Imagenet (sup.)	**70.42**	25.12	36.20
MoCo-v2	64.49	30.94	40.71
MoCo-v2+TP	69.14	29.66	40.12
SeCo (ours)	65.47	**38.06**	**46.94**

Table 4. Fine-tuning results on the Onera Satellite change detection task. We use a ResNet-18 pre-trained on 1M images.

the backbone with representations pre-trained on 1M satellite images (except when using random weights or loading an ImageNet pre-trained model). We freeze the backbone weights and train the classifier for 100 epochs with a batch size of 32, reporting the best validation accuracy for each run. We use an Adam optimizer with default hyperparameters, setting the initial learning rate to 1e-3 and dividing it by 10 at 60% and 80% of the epochs.

Quantitative Results Table 3 compares the linear probing accuracy of SeCo representations against the different baselines. We see that SeCo achieves 6.7% higher accuracy than ImageNet pre-training and 3.6% higher accuracy than MoCo-v2+TP. These results confirm that the learned representation is not only effective on BigEarthNet, but also generalizes to other remote sensing datasets such as EuroSAT.

4.3. Change Detection on Onera Satellite

The Onera Satellite Change Detection (OSCD) dataset [8] is composed of 24 pairs of multispectral images from Sentinel-2. The images were recorded between 2015 and 2018 from locations all over the world with various levels of urbanization, where urban changes were visible. Each location contains aligned pairs covering all 13 Sentinel-2 spectral bands. Images vary in spatial resolution between 10 m, 20 m and 60 m, with approximately 600 × 600 pixels at 10 m resolution. The goal is to detect changes between satellite images from different dates. Pixel-level change ground truth is provided for all training and validation image pairs. We use the same train/val splits proposed by Daudt et al. [8]: 14 images for training and 10 images for validation. We measure the the downstream performance in terms of F1 score, as it is common in the image segmentation literature.

Implementation Details For every pair of images from a given location at two different timestamps, we produce segmentation masks by following a procedure similar to Daudt et al. [7]. First, a ResNet-18 backbone extracts features from each image. We keep the features after each downsampling operation in the backbone network. Then, we compute the absolute value of the difference between the two sets of features in each pair, and use the feature differences as input to a U-Net [34] in order to generate binary segmentation masks. The backbone network is initialized with representations pre-trained on 1M satellite images. To avoid overfitting, we freeze the backbone and only train the weights of the U-Net, add a 0.3 dropout rate after each upsampling layer in the U-Net, and augment the training images with random horizontal flips and 90° rotations. In addition, since the images in the OSCD dataset have variable size, we split them into non-overlapping patches of 96 × 96 pixels. We train the decoder for 100 epochs with a batch size of 32, and report results on the validation set from the point of view of the ”change” class. We use an Adam optimizer with a weight decay of 1e-4. We set the initial learning rate to 1e-3 and decrease it exponentially with a multiplicative factor of 0.95 at each epoch.

Quantitative Results Table 4 compares SeCo with random initialization, ImageNet pre-training, MoCO-v2, and MoCo-v2+TP. We observe that SeCo initialization achieves higher recall and F1 score than all the baselines. In particular, SeCo outperforms MoCo-v2+TP by 6.8% F1 score. This might be due to MoCo-v2+TP representations being invariant to temporal variations, which is not a desirable property in a change detection task. Interestingly, although both SeCo and MoCo-v2 consider image patches from the same location at different timestamps as negative pairs (i.e. their learned representations are variant to time), SeCo attains a 6.2% higher F1 score. This indicates that the multiple embedding sub-spaces make SeCo more effective at detecting temporal changes by disentangling image augmentations from temporal variations.

Qualitative Results Figure 4 compares the change detection masks produced by our method and all the baselines on two samples from the OSCD validation set. We observe that SeCo pre-training produces higher quality masks which cover more of the changed pixels without excessive false negatives. We also notice some discrepancies in the performance of MoCo-v2 with and without leveraging temporal information (TP). We hypothesize these might be due to the different treatment of temporal invariance by each approach, and the image differences resembling more artificial augmentations or temporal changes. SeCo overcomes this problem by learning a representation that preserves time-varying and invariant factors.

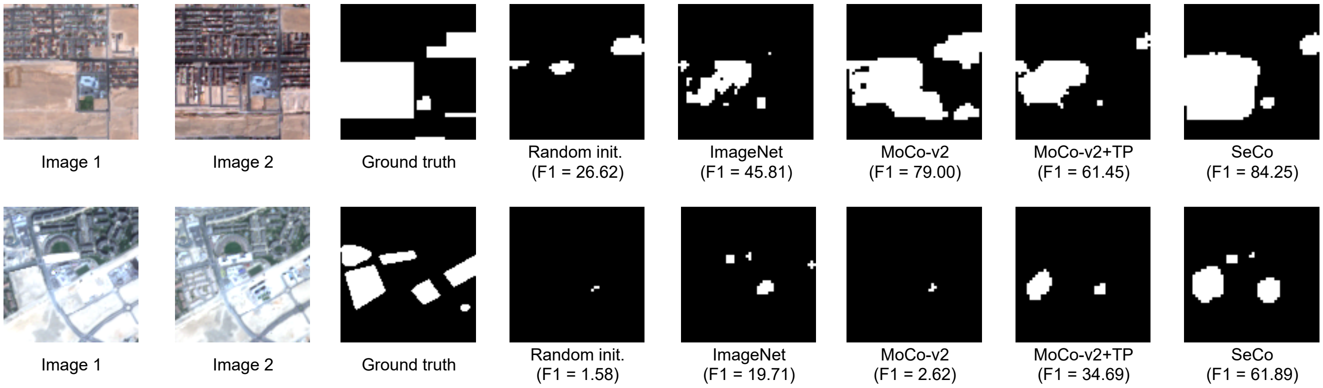

Figure 4. Comparison of qualitative results on the Onera Satellite change detection task. Each row contains the input images, the ground truth mask, and the generated change detection masks for a validation sample.

5. Related Work

Learning from Uncurated Data Recent efforts in unsupervised feature learning have focused on either small or highly curated datasets like ImageNet, whereas using uncurated raw datasets was found to decrease the feature quality when evaluated on a transfer task [9, 2]. Caron et al. [3] propose a self-supervised approach which leverages clustering to improve the performance of unsupervised methods trained on uncurated data. Other methods use metadata such as hashtags [23, 40, 27], geolocation [45] or the video structure [14] as a source of noisy supervision. In our work, we leverage the geographical and temporal information of remote sensing data to learn unsupervised representations from uncurated datasets.

Multi-augmentation Contrastive Learning Recent self-supervised contrastive learning methods have been able to produce impressive transferable visual representations by learning to be invariant to different image augmentations. However, these methods implicitly assume a particular set of representational invariances, and can perform poorly when a downstream task violates this assumption. Xiao et al. [47] propose Leave-one-out Contrastive Learning (LooC), a multi-augmentation contrastive learning framework that produces visual representations able to capture varying and invariant factors by constructing separate embedding spaces, each of which is invariant to all but one augmentation. In our work, we use a similar approach to learn representations that are variant and invariant to the seasonal changes present in remote sensing images.

Unsupervised Learning in Remote Sensing While unsupervised learning has been extensively studied on natural image datasets (*e.g.*, ImageNet), this subfield remains underexplored on the remote sensing domain. This is quite surprising given the importance of remote sensing for Earth observation, the vast amount of readily available data, and the many opportunities for self-supervision from the unique characteristics of satellite images. For instance, Jean et al. [22] use the geographical information of images to sample positive and negative pairs and build a pretext task based on the triplet loss. Uzkent et al. [42] pair georeferenced Wikipedia articles with satellite images of the corresponding locations, and learn representations by predicting properties of the articles from the images. Vincenzi et al. [43] leverage the multi-spectrality of remote sensing images to build a colorization task, where they reconstruct the visible colors from the other spectral bands. More similar to our work, Ayush et al. [1] also propose to exploit the temporal information in satellite imagery to generate positive pairs and train a contrastive objective. However, their representations are always invariant to temporal changes, which might be detrimental for downstream tasks involving temporal variation. We overcome this problem by using multi-augmentation contrastive learning, where the representations preserve time-varying and invariant information.

6. Conclusions

We presented Seasonal Contrast (SeCo), a new transfer learning pipeline for remote sensing imagery consisting of (1) a data collection strategy and (2) a self-supervised learning algorithm that leverages this data. First, we sample locations around populated regions over multiple timestamps, which provides a diverse set of satellite images. Then, we extend multi-augmentation contrastive learning methods to take into account the seasonal changes and learn rich and transferable remote sensing representations. In our experiments, we found that SeCo consistently outperforms the considered pre-training baselines on several remote sensing downstream tasks.

It is worth noting that we strive for impact rather than groundbreaking novelty by building on top of existing ideas. Thus, we hope to improve performance on Earth monitoring algorithms and help address global challenges.

References

[1] K. Ayush, B. Uzkent, C. Meng, M. Burke, D. Lobell, and S. Ermon. Geography-aware self-supervised learning. *arXiv preprint arXiv:2011.09980*, 2020. 5, 8

[2] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 8

[3] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019. 8

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[6] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5

[7] R. C. Daudt, B. Le Saux, and A. Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. 7

[8] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE, 2018. 2, 7

[9] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2, 8

[10] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 1, 2, 3, 5

[11] F. Filipponi. Exploitation of sentinel-2 time series to map burned areas at the national level: A case study on the 2017 italy wildfires. *Remote Sensing*, 11(6):622, 2019. 1

[12] G. M. Foody. Remote sensing of tropical forest environments: towards the monitoring of environmental resources for sustainable development. *International journal of remote sensing*, 24(20):4035–4046, 2003. 1

[13] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[14] D. Gordon, K. Ehsani, D. Fox, and A. Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 8

[15] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017. 3

[16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 5

[19] P. Helber, B. Bischke, A. Dengel, and D. 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*, 2019. 1, 2, 6

[20] C. Ippoliti, L. Candeloro, M. Gilbert, M. Goffredo, G. Mancini, G. Curci, S. Falasca, S. Tora, A. Di Lorenzo, M. Quaglia, et al. Defining ecological regions in italy based on a multivariate clustering approach: A first step towards a targeted vector borne disease surveillance. *PloS one*, 14(7): e0219072, 2019. 1

[21] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 3

[22] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019. 8

[23] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 8

[24] I. Laradji, P. Rodriguez, O. Manas, K. Lensink, M. Law, L. Kurzman, W. Parker, D. Vazquez, and

D. Nowrouzezahrai. A weakly supervised consistency-based learning method for covid-19 segmentation in ct images. 2021. 1

[25] I. H. Laradji, D. Vazquez, and M. Schmidt. Where are the masks: Instance segmentation with image-level supervision. 2019. 1

[26] I. H. Laradji, R. Pardinas, P. Rodriguez, and D. Vazquez. Looc: Localize overlapping objects with count supervision. 2020. 1

[27] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 8

[28] I. Misra and L. v. d. Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[29] D. J. Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013. 1

[30] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 1, 5, 6

[31] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[32] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3, 4

[33] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019. 1

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7

[35] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145:154–172, 2014. 2

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge (2014). *arXiv preprint arXiv:1409.0575*, 2014. 2, 3

[37] G. J. Schumann, G. R. Brakenridge, A. J. Kettner, R. Kashif, and E. Niebuhr. Assisting flood disaster response with earth observation data and products: A critical assessment. *Remote Sensing*, 10(8):1230, 2018. 1

[38] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 1, 2, 5

[39] G. Sumbul, J. Kang, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, and B. Demir. Bigearthnet dataset with a new class-nomenclature for remote sensing image understanding. *arXiv preprint arXiv:2001.06372*, 2020. 5

[40] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 8

[41] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2

[42] B. Uzkent, E. Sheehan, C. Meng, Z. Tang, M. Burke, D. Lobell, and S. Ermon. Learning to interpret satellite images in global scale using wikipedia. *arXiv preprint arXiv:1905.02506*, 2019. 1, 8

[43] S. Vincenzi, A. Porrello, P. Buzzega, M. Cipriano, P. Fronte, R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara. The color out of space: learning self-supervised representations for earth observation imagery. *arXiv preprint arXiv:2006.12119*, 2020. 8

[44] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 4

[45] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016. 8

[46] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[47] T. Xiao, X. Wang, A. A. Efros, and T. Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 2, 3, 8

[48] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2