Change-Aware Sampling and Contrastive Learning for Satellite Images

Utkarsh Mall Bharath Hariharan Kavita Bala
Cornell University
{utkarshm, bharathh, kb}@cs.cornell.edu

Abstract

Automatic remote sensing tools can help inform many large-scale challenges such as disaster management, climate change, etc. While a vast amount of spatio-temporal satellite image data is readily available, most of it remains unlabelled. Without labels, this data is not very useful for supervised learning algorithms. Self-supervised learning instead provides a way to learn effective representations for various downstream tasks without labels. In this work, we leverage characteristics unique to satellite images to learn better self-supervised features. Specifically, we use the temporal signal to contrast images with long-term and short-term differences, and we leverage the fact that satellite images do not change frequently. Using these characteristics, we formulate a new loss contrastive loss called Change-Aware Contrastive (CACo) Loss. Further, we also present a novel method of sampling different geographical regions. We show that leveraging these properties leads to better performance on diverse downstream tasks. For example, we see a 6.5% relative improvement for semantic segmentation and an 8.5% relative improvement for change detection over the best-performing baseline with our method.

1. Introduction

Our planet is surrounded by a large number of satellites constantly collecting images of the world. This massive trove of visual information can help monitor phenomena at the world-scale, and inform solutions to global problems such as climate change or loss of biodiversity. Automatic vision tools can help by, for example, monitoring land-use change over time [36] or the evolution of urban areas [6].

However, training all these models requires *labeled data*. Unfortunately, labeling the massive trove of satellite images is expensive, more so than internet images because of the expertise necessary. This issue is exacerbated by the different label requirements of many monitoring applications.

One way to alleviate this problem of limited labeled data is to use self-supervised learning techniques to learn a good feature representation from unlabeled satellite imagery. This representation can then be further finetuned with much fewer labels for specific applications. Modern self-supervised learning approaches are based on contrastive learning. These techniques train a feature space so that each image in the dataset is embedded close to augmented versions of itself (e.g., with jittered colors) but far from other images of the dataset. A possible approach is to directly apply these techniques on satellite image datasets. However, the spatio-temporal structure of satellite imagery is much richer than the unstructured collections of internet images typically used in standard self-supervised learning. In this work, we ask, *how can we best leverage the structure of satellite images for better self-supervised learning?*

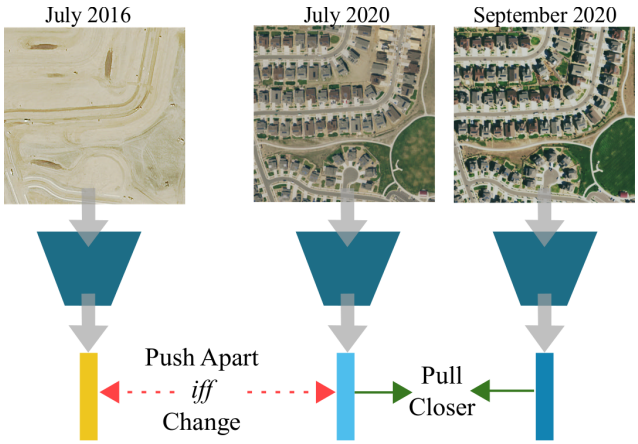

Figure 1. Images of the **same location** at three different times. Changes from 2016 to 2020 are due to major urban development, while those from July to September are seasonal variations. Our approach, CACo, learns features that are sensitive to the former but invariant to the latter.

The first important aspect of the spatio-temporal structure of satellite images is the availability of multiple temporally-spaced images for the same location. Past work has used this structure to sample images spread over a few weeks or months from each location to encourage invariance to seasonal variations [25]. However, we can access images not just over a few months, but over *many years*. Over such long time spans, we often see significant, per-

manent change, such as the construction of houses, the drying of lakes, or the logging of forests (Fig. 1). While we want feature representations to be invariant to temporary, seasonal change, we want the representation to be *sensitive* to permanent, long-term change. To capture this intuition, we sample multiple images from the same location covering both short and long time spans and encourage invariance to the former but *sensitivity* to the latter.

A second and more important aspect of satellite images is that permanent change is spatially rare. Change is concentrated near urban areas, but many parts of the planet see little change. When there is no change, we want our feature representation to be the same even over long time spans. We capture this insight with a novel strategy to robustly estimate whether or not there is a long-term change, even in the middle of training, by comparing long-term feature differences to short-term variations. We then design a loss function that is *conditional* on this change estimate: it encourages invariances to long-term differences depending on whether a change occurs or not. We call this novel loss function *Change-aware Contrastive Loss (or CaCo)*.

The above loss function uses the temporal structure of satellite images. We can also utilize geographical structure by carefully sampling the most informative locations on the planet. We provide an improvement over a previously proposed geographical sampling [25]. We show that sampling closer to cities, and ignoring samples completely in the ocean can result in a dataset much more useful for learning a general representation for various downstream tasks. We evaluate our new representation on a diverse set of downstream tasks such as landcover classification, semantic segmentation, and change detection. Our method achieves significant relative improvements (ranging from 6.5% to 8.5%) over the state-of-the-art for a variety tasks such as segmentation and change detection.

To summarize, we make the following contributions:

- We propose a new self-supervised learning loss that uses long-term temporal information in satellite imagery to encourage invariance to seasonal variations but sensitivity to permanent, long-term changes.
- We introduce a novel approach to robustly estimate whether a location has undergone significant change by comparing long-term changes to seasonal variations. Our new change-aware loss function (CACo) uses this to decide when to encourage invariance.
- We use an improved geographical sampling that provides more diverse data for representation learning.

2. Related Work

Self-supervised Learning. Self-supervised learning methods can be used to learn a good general representation from a dataset without any labels. These methods obtain supervisory signals or inductive bias from the unlabelled data itself. Earlier self-supervised works used "pretext tasks" such as rotation prediction [15], solving jigsaw puzzles [28], colorization [44], or missing data completion [17]. More recently contrastive learning methods have been shown to learn a better representation than pretext tasks. Methods for contrastive learning such as NPID [43], PIRL [26], MoCo [8, 18], SimCLR [7], and BYOL [16], use instance discrimination to learn a feature representation. While these methods can be applied even to satellite images, in this work we leverage the structure of satellite images that can be used for self-supervised learning.

Self-supervised Learning in Remote Sensing. Increasingly, self-supervised learning is being applied to the area of remote sensing. Prior works have used signals from unlabeled data such as location [2, 20], seasonal variations [25], or texture [1], as a signal for self-supervision. Recent methods use transformers [12] and diffusion models [4] for self-supervised representation [10, 32].

Remote Sensing Applications and Data Several remote sensing applications require vision tools for automatic recognition at scale. Applications such as landcover classification [19, 34] or segmentation [36, 40] benefit from the advances in vision algorithms for scene classification and semantic segmentation. Several object detection applications have also been developed using object detection algorithms such as for detecting constructions [9], buildings [38], trees [41], or floating objects [14]. Segmentation methods also help in extracting structural information like road networks [3, 22, 39], crop types [37], or clouds [27] from remote sensing images. Many applications such as change detection [11] require temporal information for regions as well. Temporal information and changes are also used to find long-term semantic events [24]. In this work, we aim to learn a representation without supervision, that leads to better performance on all these downstream tasks.

The image data for these algorithms and datasets are obtained from a few different sources. Sentinel-2 [13] provides multispectral images (13 bands) at 10m resolution with a temporal revisit of 5 days. Several other satellites provide higher resolution information such as WorldView-3 and PlanetScope. We use Sentinel-2 satellite imagery as it provides frequent temporal information that we use in our self-supervised formulation. Furthermore, many of the downstream tasks [11, 19, 24, 36] use Sentinel-2 imagery.

3. Method

3.1. Overview

We propose a self-supervised method that can leverage properties that are unique to satellite images. One such

property is the availability of long-term temporal information for any location on earth. For example, two images of the same location captured months apart will likely be the same except for seasonal variations; variations our features should be invariant to. In contrast, two images captured decades apart might show big, structural changes such as new buildings or sparser forests. Of course, this may not be true for all locations. Indeed, such long-term change is rare, and is concentrated in regions of human activity. But these long-term changes are important for many downstream applications such as modeling urban growth, or monitoring climate change. Our key insight is therefore to find these rare but meaningful long-term changes, and ensure that they are reflected as changes in the feature space as well.

Past work on using temporal data for representation learning on satellite imagery only uses short-term seasonal variations but not long-term change [25]. We review this work, which is called Seasonal Contrast (or SeCo) [25], in Sec. 3.2. We then explain how we can use long-term temporal information in Sec. 3.3, and address the challenge of spatial rarity in Sec. 3.4. Finally, we look at how we can efficiently sample more informative locations when collecting unlabelled training data in Sec. 3.5.

3.2. Background: Seasonal Contrast (SeCo)

Contrastive learning techniques train feature representations from unlabeled images by discriminating between individual instances. In particular, they pull augmentations of the same image closer together in feature space and push apart representations of two different images. SeCo is a contrastive learning-based representation learning framework for remote sensing images. SeCo uses MoCo v2 [8], to perform contrastive learning. In MoCo v2, in each iteration, one generates two views for an image using random augmentations. One is the "query" image I_q and the other is the "positive key" I_{k^+} that is pulled closer to the query in the representation space. It also has a set of other images from the dataset, $\mathcal{I}_{k^-}$, that are treated as negative keys and are pushed apart from query image I_q in the latent space. The precise loss function in MoCo v2 called InfoNCE [29], for a representation function f, can be written as:

$$\mathcal{L} = -\log \frac{exp(f(I_q) \cdot f(I_{k^+})/\tau)}{\sum\limits_{I_k \in \mathcal{I}_{k^-} \cup \{I_{k^+}\}} \exp(f(I_q) \cdot f(I_k)/\tau)} \quad (1)$$

The key idea in SeCo is to also use short-term temporal differences (over a few months) at a location as augmentations. Since these images are a few months apart, these images represent different seasons (hence the name seasonal contrast). The negative keys are then images from other locations. More specifically SeCo trains a representation with 3 subspaces, i) a subspace invariant to both season and artificial augmentations, ii) a subspace invariant to season only, and iii) a subspace invariant to augmentation only. The additional invariance to seasonal differences results in a better representation than simply using MoCo v2 on this data.

3.3. Using Long-term Temporal Information

SeCo introduces invariance to seasonal variations. However, invariance is not enough. We also want our feature representation to be informative. For this, simply pushing apart images from different locations (which already look quite different) may not be enough.

As discussed above, we argue that to produce a richer feature representation, we should use information from longer temporal spans. Over the span of years, we might observe big, structural changes such as new construction or urban growth. We want our feature representation to be *sensitive* to these major changes. Thus, while short-term temporal differences are typically seasonal and should be treated as augmentations as in SeCo, long-term temporal differences might correspond to permanent change, and should be reflected as a significant change in the underlying feature representation. This suggests that for a query image from a particular location, an image captured at the same location several years later should be treated as a *negative* in the contrastive learning objective. We formalize this below.

Data: Let $\{l_1, l_2 \cdots l_n\}$ be n different locations in our dataset. For each location, we have 2 sets of images captured around time points t_1 and t_2 that have a gap of 4 years between them. We represent these sets by $\mathcal{I}_{l_i}^{t_1}$ and $\mathcal{I}_{l_i}^{t_2}$. Each set $\mathcal{I}_{l_i}^{t_j}$, contains images sampled over a few months around t_j, and thus these images have only short-term or seasonal differences between them. See Fig. 2 (a) for a schematic of this data sampling strategy.

Loss: Let $I_{l_i}^{t_j+\Delta_1}, I_{l_i}^{t_j+\Delta_2} \sim \mathcal{I}_{l_i}^{t_j}$ be two images sampled randomly from $\mathcal{I}_{l_i}^{t_j}$ (j is 1 or 2). $I_{l_i}^{t_j+\Delta_1}$ and $I_{l_i}^{t_j+\Delta_2}$ are a few months apart. So like SeCo, we still want their representations to be *pulled closer*. However, images with larger time differences, for example, $I_{l_i}^{t_1+\Delta_1}$ and $I_{l_i}^{t_2+\Delta_1}$ should potentially be *pushed apart*. Finally, images from two different locations such as $I_{l_i}^{t_j+\Delta_1}$ and $I_{l_j}^{t_j+\Delta_1}$, should always be pushed apart irrespective of time. This intuition is reflected in Fig. 2 (c).

Following SeCo, we implement this loss function using MoCo v2 by adapting the positive key image I_{k^+} and negative key set $\mathcal{I}_{k^-}$ in Eq. (1). As discussed above, SeCo has 3 subspaces. We add our modifications to all 3 subspaces. However, for ease of explanation, we only describe the change to the subspace that is invariant to seasons. For this subspace, for a query image, $I_q = I_{l_i}^{t_1+\Delta_1}$, the positive key is the same as in SeCo: the seasonal pair ($I_{k^+} = I_{l_i}^{t_1+\Delta_2}$). But we modify the negative key set to not only include images from other locations but also images from the same

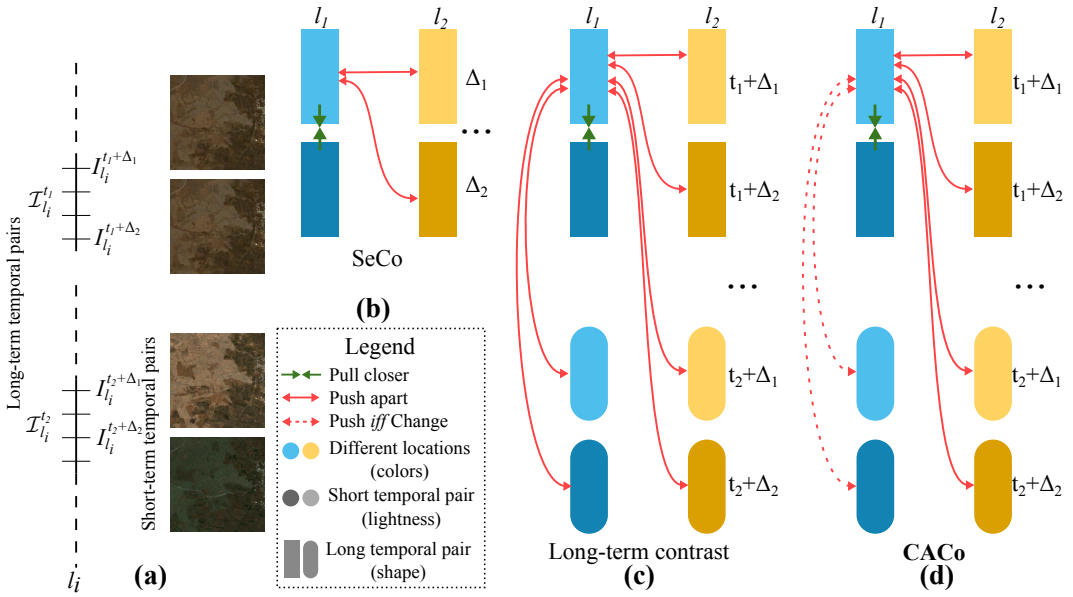

Figure 2. Overview of our self-supervised framework. (a) Shows the notation for images and the temporal sampling strategy we use to create our dataset for training. (b) Shows the contrastive learning performed by SeCo in the temporally invariant subspace. The red arrows show examples that need to be pushed apart from the top left example. The green arrows indicate examples to be pulled closer. Different lightness (value) represents different short-term temporal pairs and different colors (hue) indicate different locations (see Sec. 3.2). (c) Shows our method when using long-term temporal pairs. Different shapes indicate a difference of 4 years between examples (see Sec. 3.3). (d) Shows the contrastive learning performed by our CACo loss term. The dashed red arrows indicate examples, that are pushed apart iff our method estimates a change has occurred between the long-term pairs (see Sec. 3.4).

location with a large time difference as follows:

$$\mathcal{I}_{k^-} = \{I_{l_i}^{t_2+\Delta_k} : k \in \{1,2\}\} \cup \{I_{l_h}^{t_j+\Delta_k} : h \neq i, j \in \{1,2\}, k \in \{1,2\}\} \quad (2)$$

In each training iteration, we sample 4 images (long and short-time pairs) for each location instead of 2 like SeCo, so each batch contains half as many locations.

3.4. Change-aware Contrastive Learning

The training method in the previous section assumes that regions *always* change after a long time (a few years). However, this assumption is not always true: in remote areas away from human activity, there may be no change even over several years. For such locations, pushing apart features of temporally distant image pairs may be counterproductive and destroy needed invariances.

If we know whether a location has changed or not, we can make our loss function conditional on this information. If it has changed between $I_{l_i}^{t_1+\Delta_1}$ and $I_{l_i}^{t_2+\Delta_1}$, we push them apart as described in Sec. 3.3. If not, we refrain from adding the temporally distant image to the negative set. This conditional loss function is illustrated in Fig. 2 (d).

However, the challenge is knowing which locations have changed. Change detection can be performed by looking at feature differences between temporally separated image pairs [21, 23, 35, 42], but this requires a feature representation we do not have. We solve this chicken-and-egg problem by bootstrapping our feature representation. We start from a randomly initialized model, using which we extract features for different locations and use feature differences as an estimate for change. Using these change values, we use our conditional loss to train the model. We then keep alternating between training the model using changes and estimating the changes using the partially trained model. In the initial stages, the change estimates could be very poor as the model has random weights. However, as training progresses we expect better features, and, in turn, better estimates of changes. One might worry about degenerate solutions, but the other contrastive loss terms based on location (that do not depend on this inferred change) ensure that the feature representation avoids any pathological cases.

Obtaining changes using features. The absolute distance between features scales differently for different types of locations, and does not directly reflect the magnitude of actual change. For example, urban locations with less vegetation look the same over the seasons, but temperate forests will change a lot with the seasons. So instead of using the absolute distance between two long-term images, we instead use the ratio of distances between long-term images and short-term images (images with a seasonal difference).

This ratio normalizes the scaling differences between images from different locations. If f_k is the partially trained feature extractor at epoch k, we use the following ratio as an indicator of change for location l_i

$$R^k_{l_i} = \frac{||f_k(I^{t_1+\Delta_1}_{l_i}) - f_k(I^{t_2+\Delta_1}_{l_i})||^2}{||f_k(I^{t_1+\Delta_1}_{l_i}) - f_k(I^{t_1+\Delta_2}_{l_i})||^2} \quad (3)$$

Since, at each epoch, the seasonal images are selected randomly for each location, we use an exponential moving average over ratios from previous epochs to obtain a more stable estimate for change as follows

$$\mathbf{r}^k_{l_i} = (1-\beta)R^k_{l_i} + \beta \mathbf{r}^{k-1}_{l_i}$$

We use a Gaussian Mixture Model (GMM) on $\mathbf{r}^k_{l_i}$ values to find our two clusters for change and no change. Locations with smaller $\mathbf{r}^k_{l_i}$ are treated as having no long-term change, and those with larger $\mathbf{r}^k_{l_i}$ values are deemed to have undergone significant change (see Fig. 3).

For MoCo v2, we use features from the momentum encoder for images to calculate the ratio estimate. To maintain training speed, we compute and store the ratios during the training iterations themselves, even as the feature extractor is being updated. In preliminary experiments, we found that the exponential moving average was sufficient to regularize against the noise induced by the changing feature extractor.

Our change-aware formulation uses change information in long-term pairs. On the other hand, there could be short-term pairs with sudden changes such as due to disasters like landslides or earthquakes. However, we notice that in practice such changes are extremely rare. Out of the top 200 examples with the highest short-term change score only 3 pairs showed a sudden change. Because short-term changes are so rare, it is not useful to consider them in our pipeline.

3.5. Improving Geographical Sampling

Another property specific to satellite imagery is that we can control the geographic distribution of sampled images. Leveraging this freedom, SeCo collects data from a 100 km radius around the 10k most populated cities of the world. However, we argue that 100 kms is too large a radius.

Sampling far from cities results in uninformative images (such as images entirely over oceans, (see supplementary Fig. 3)). In fact, 22% of the dataset has such uninformative images and this hurts the quality of the representation.

The median area of the 10k cities sampled is approximately 24 km^2, which represents a radius of $\sim$ 5 km. Based on this observation, we use 2d-gaussian sampling with $\sigma = 5$ km. This means that 95% of the data comes from a radius of $2\sigma = 10$ kms around the city center. Additionally, if the sampled location is completely in the ocean, we reject that sample and sample again (We use a landmass map of the earth to check this). Our improved sampling methods result in a dataset with a more diverse set of locations that are more informative. As we show in experiments, this additionally improves the quality of the representation.

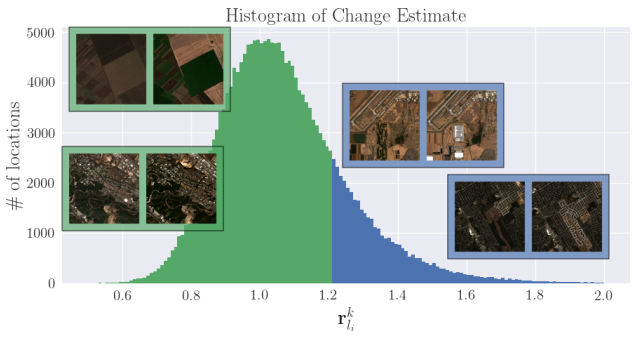

Figure 3. The histogram of change estimates separated into two groups (blue and green) by the GMM. Pairs of examples with a higher estimate can be seen to have major construction and land-cover changes (2 pairs on the right). Whereas pairs with a lower estimate only have seasonal changes.

4. Results

4.1. Implementation details and Baselines

We evaluate our model by measuring its representation ability on various downstream tasks. We evaluate the following baselines for comparison.

Random init.: is randomly initialized and is either fine-tuned or linear probed for downstream tasks.

ImageNet: is pre-trained on supervised ImageNet [31].

MoCo v2: uses self-supervised learning without the temporal or change information.

SeCo: is Seasonal Contrast [25] and uses only short-term changes as augmentations.

GSSL: uses temporal and geographical information for contrastive learning [2]. Since the data this model was trained on is of different resolution retrain it on our data.

SatMAE [10]: is a Masked autoencoder [17] that uses ViT-Large as backbone. The comparison is not apples-to-apples as the model has larger capacity. The model is trained on 713k images therefore it is comparable to our 1m images trained model.

DDPM-CD [4]: uses a denoising diffusion model as backbone. It is also trained on a million scale dataset.

We show results with both a ResNet-18 and a ResNet-50 backbone. Similar to SeCo our models are trained with a batch size of 256 and with 16,384 negative embeddings. We also use the same optimizer, learning rate, schedule, and temperature scaling as in the SeCo implementation.

We experiment with two datasets for self-supervised pre-training: one with 100k images and the other with 1 million

Data	Pre-training	ResNet-18	ResNet-50
-	Random init.	64.21	55.32
	ImageNet.	86.16	89.08
100k	MoCo v2	87.22	89.75
	GSSL	87.74	90.19
	SeCo	90.05	93.12
	CACo (ours)	**93.08**	**94.48**
1m	SeCo	93.99	95.63
	SatMAE (ViT-L)	-	93.03
	DDPM-CD	-	87.67
	CACo (ours)	**94.72**	**95.90**

Table 1. Performance of our representation on the EuroSat classification task with linear probing, in top-1 Accuracy. Our method provides a more accurate classification, with different backbones.

Data	Pre-training	ResNet-18		ResNet-50	
		10%	100%	10%	100%
-	Random init.	42.87	45.95	44.76	45.22
	ImageNet.	65.43	66.40	70.36	71.37
100k	MoCo v2	65.43	67.20	70.38	72.88
	GSSL	65.78	67.36	70.65	72.86
	SeCo	65.80	67.43	70.69	73.42
	CACo (ours)	**67.89**	**69.43**	**71.55**	**73.63**
1m	SeCo	67.68	69.95	72.89	74.82
	SatMAE	-	-	67.93	69.45
	DDPM-CD	-	-	65.47	67.31
	CACo (ours)	**68.64**	**70.41**	**73.40**	**74.98**

Table 2. Performance of our method on BigEarthNet land-cover classification, in mean Average Precision (mAP). The two columns use different percentages of data (10% and 100%) of BigEarthNet data for training the linear layer.

images. These datasets contain RGB images from Sentinel-2 [13]. The long-time difference between the two sets $\mathcal{I}_{l_i}^{t_1}$ and $\mathcal{I}_{l_i}^{t_2}$ is 4 years. The maximum time difference within a set $\mathcal{I}_{l_i}^{t_j}$ is 1 year. We train our model and baselines for 1000 and 200 epochs on the 100k and 1m dataset respectively.

4.2. Landcover classification

Nine (out of seventeen) of the UN's Sustainable Development Goals require global monitoring of land cover [5]. Thus, improving the performance of models for landcover classification would result in better monitoring of these goals. We evaluate our method on two datasets for landcover classification: EuroSat [19] and BigEarthNet [34].

EuroSat has 10-classes with a total of 27k images 64×64 images from Sentinel-2. We use the same train/val split proposed by the dataset. We add a linear layer to the frozen pre-trained backbone, to perform the linear evaluation. More details about the training are in the supplementary.

Tab. 1 shows the top-1 classification accuracy of various pre-trained backbones on the EuroSat dataset. We first note that all satellite image-specific methods outperform generic pretraining on ImageNet, suggesting the importance of training a representation specifically for satellite imagery. They also outperform the MoCo v2 baseline trained on satellite images, indicating the importance of well-designed loss functions that use the structure of satellite imagery.

Compared to SeCo, our approach (CACo) results in 3 points of improvement with the linear classifier (with ResNet-18 and 100k data). This validates our insight that invariance to seasonal changes is not enough; the features must also be trained to be sensitive to long-term change. In fact, CACo trained on 100k images produces a representation competitive with SeCo trained on 1m images. Thus our method can save an order of magnitude of images and training time over the SeCo baseline while producing representations of similar quality.

We also evaluate our method on the BigEarthNet dataset, which is a multi-label classification dataset with 17 classes. The dataset is significantly larger than EuroSat, containing 590k sentinel-2 image patches. We use the RGB bands to perform classification on this dataset. Since this is a multi-label classification problem, we use mean Average Precision (mAP) to evaluate the performance.

Tab. 2 compares the performance of our model to the baselines on BigEarthNet[1]. Even on BigEarthNet, our method performs better than the baselines.

4.3. Change detection

We also evaluate our model on the OSCD change detection dataset [6]. It contains 24 pairs of images from the Sentinel-2 satellites between 2015 and 2018. The dataset is split into 14 training images and 10 testing images.

Our model architecture follows past work [11, 25]. The input to the model is an image pair, each of which is fed into the frozen, pre-trained feature extractor (we use the ResNet-18 models). We extract feature maps after each downsampling layer for each image. We then take the absolute difference between the corresponding feature maps from the two images. These differences are passed as input to a U-Net Decoder [30]. The U-Net decoder is then trained with change supervision. More details about the training and architecture can be found in the supplementary.

Results are shown in Tab. 3. We report the F1-score obtained by using a threshold of 0.5. Our method results in better features for change detection resulting in about 4

[1]We obtained the SeCo number by training a model using the authors' publicly released code and dataset, but this does not match the published number [25]. Repeated attempts to contact the authors were unanswered.

Backbone	Pre-training	Prec.	Rec.	F1-score
ResNet-18	Random init.	69.73	18.24	28.91
	ImageNet.	70.22	23.58	35.30
	MoCo v2	62.21	27.57	38.21
	GSSL	62.29	34.08	44.06
	SeCo	64.15	36.89	46.84
	CACo (ours)	60.68	42.94	**50.29**
ResNet-50	Random init.	65.46	22.43	33.41
	ImageNet.	70.05	30.96	42.94
	GSSL	62.16	36.83	46.26
	SeCo	63.21	38.26	47.67
	CACo (ours)	62.87	44.49	**52.11**

Table 3. Performance of our method on the Change Detection Task. We report the precision, recall and F1-score at a threshold of 0.5 for various ResNet backbones, trained on the 100k dataset.

Pre-training	linear	finetuning
Random init.	41.53	38.62
ImageNet.	43.75	43.78
MoCo v2	47.97	47.15
GSSL	46.77	48.10
SeCo	46.83	48.18
CACo (ours)	**50.20**	**51.29**

Table 4. Performance of our method on the DynamicEarthNet segmentation task. We report the mIoU score for the ResNet-18 backbone, trained on the 100k dataset.

points of improvement in the F1 scores for both ResNet-18 and ResNet-50. This large improvement on change detection is likely because we explicitly train our representation to be sensitive to long-term change.

4.4. Semantic segmentation

We also evaluate our method on a satellite images segmentation task. DynamicEarthNet [36] is a dataset used to evaluate landcover segmentation over 7 classes. The dataset contains images from Planet satellites for 65 locations. Each location has 24 1024×1024 images with labels. The locations are split into 55 for training and 10 for testing.

Similar to change detection model, we use a U-net with ResNet-18 as the encoder. But unlike them, we input a single image and use feature maps instead of feature differences in the decoder. Due to the dataset imbalance, we use Dice Loss [33] for training (see supplementary for details).

Tab. 4 shows the mIoU (intersection over union averaged over classes) for our method on DynamicEarthNet validation images. Our method is better than SeCo by more than 3 points, and ImageNet pre-training by more than 6 points.

Pre-training	CaiRoad		CalFire	
	AP@50	AP@400	AP@10	AP@40
ImageNet	41.79	32.26	52.61	44.99
MoCo v2	41.78	33.02	54.83	48.20
SeCo	39.63	34.72	62.87	51.87
CACo (ours)	**44.38**	**35.99**	**65.71**	**53.44**

Table 5. Performance of our method on change event retrieval for the CaiRoad and CalFire benchmarks. We show average precision@K for ResNet-18 backbones trained on the 100k dataset.

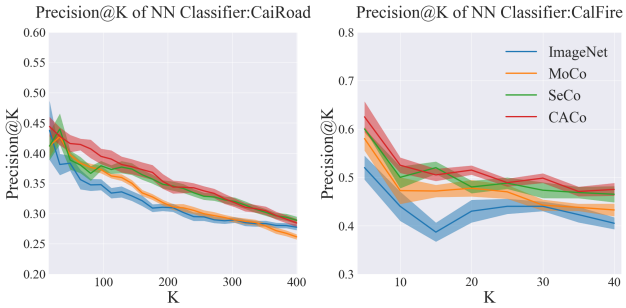

Figure 4. Precision@K of our backbone on CaiRoad (Left) and CalFire (Right) retrieval benchmark. Our method has better Precision@K than baselines for most K values on both benchmarks.

4.5. Evaluation on Change Events

We evaluate our method on a recently released *change events* benchmark [24]. Change events are spatio-temporal semantic structures that are caused by a real-world event. The goal is to classify these change events. The benchmark contains two datasets: one from Cairo where we find road construction events (called CaiRoad), and the other from California where we find forest fire events (CalFire).

We use the model proposed by Mall et al. [24], taking in a sequence of images and binary change masks for consecutive pairs. The model encodes image pairs using ResNet-18, average pooling the feature map with the downscaled change mask as weights. The final feature vector is obtained by temporal averaging. The representation is learned using SimCLR on change events (see supplementary).

We follow the same procedure, but the ResNet-18 backbone is initialized with the trained weights, and then *fine-tuned* per [24]. We evaluate by measuring average precision@K (AP@K) for different backbone architectures.

Tab. 5 shows the AP@K for different backbones, on the CaiRoad and CalFire benchmarks. Our method results in better retrieval for both types of events. Fig. 4 shows the precision@K for various methods on these benchmarks.

4.6. Ablation

We evaluate the design choices made by our method in Sec. 3. We look at improvements made by different com-

Pre-training	EuroSat (Acc.)	DynamicEarthNet (mIoU)
SeCo	90.05	46.83
+ Improved Sampling	91.42	47.24
+ Long-term temporal Contrast	92.56	49.68
+ Change-awareness (Ours)	**93.08**	**50.20**

Table 6. Performance with different sub-components of CACo on EuroSat classification, with ResNet-18 pre-trained on the 100k dataset. All three of our novel contributions lead to an improvement in performance on downstream tasks.

Pre-training	EuroSat (Acc.)	OSCD (F1)
SeCo	87.97	44.05
CACo	**90.83**	**47.50**

Table 7. Accuracy on EuroSat when using a ResNet-18 backbone pre-trained using SimCLR instead of MoCo v2. Our insights generalize to other self-supervised frameworks like SimCLR as well.

ponents such as better data, long-term temporal information, and change estimates. We also show that our method generalizes to other self-supervised frameworks.

Are the new components of our method essential? In Tab. 6 we evaluate the impact of each component of our approach on the final performance as measured on EuroSat and DynamicEarthNet. We find that each component adds a significant improvement. It is telling that incorporating long-term temporal contrast has a particularly large impact on DynamicEarthNet, indicating that long-term contrast is especially needed for pixel-level localization tasks.

Can our method work with other self-supervised frameworks? We replace the MoCo v2 framework with SimCLR [7]. SimCLR requires a large batch size to find more hard negative examples; we use a batch size of 512 instead of 256 for all the methods. Tab. 7 shows the performance of SeCo and CACo using SimCLR. Our model performs better than SeCo by 3 points when using SimCLR as well. This indicates that our insights are general and can be applied in the future to new self-supervision frameworks.

How useful is the ratio estimate? In Sec. 3.4 we proposed to estimate using $\mathbf{r}_{l_i}^k$ to estimate changes. We now look at other estimates for change in Tab. 8. Directly using the feature distance between the long-term pair, instead of the ratio leads to an almost 2-point drop in performance. Removing the exponential moving average (i.e., using $R_{l_i}^k$ instead of $\mathbf{r}_{l_i}^k$; see Eq. (3)) also reduces performance by about 1 point. Another alternative is to measure the ratio

Change Estimate	EuroSat (Acc.)
long-term distance	91.17
$R_{l_i}^k$	91.98
Align	92.92
$\mathbf{r}_{l_i}^k$ (ours)	**93.08**

Table 8. Performance of different change estimates with ResNet-18 on 100k dataset. $R_{l_i}^k$ (Eq. (3)) is the estimate without the moving average. Long-term distance is the numerator of the $R_{l_i}^k$. *Align* uses seasonally aligned long-term pairs without smoothing.

with long-term pairs that are seasonally aligned. *Align* gives similar results to $\mathbf{r}_{l_i}^k$, but requires a more sophisticated sampling. In sum, these results suggest that how one identifies changed regions can have a big impact on performance.

5. Discussion and Conclusions

Potential Negative Societal Impact and Limitations. As in all visual recognition, there is the possibility of negative impact through violations of privacy. To mitigate this concern we intentionally use low-resolution satellite images (1 pixel $\sim$ 10m). The use of our proposed techniques for surveillance should be appropriately regulated. Our work currently is limited in its ability to generalize to multiple resolutions. Though object detection tasks use higher-resolution images for better object delineation, our representation trained on medium-resolution images cannot directly generalize to them; although we can retrain our method on high-resolution images for these tasks.

Conclusions. We present a novel self-supervised framework to learn representations well suited for remote sensing applications. We introduce a new loss that robustly leverages the long-term temporal information readily available for satellite images. We also propose a better location sampling method to provide more informative data. Our evaluation on diverse downstream tasks shows that our approach (CACo) is very versatile and leads to better downstream accuracy than prior art (with relative improvements of 8.5% on change detection and 6.5% on semantic segmentation).

Acknowledgements. This research is based upon work supported in part by the ODNI (IARPA) via 2021-20111000006, NSF 1900783, NSF 2144117, and NSF 2212084. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *CVPR*, 2022. 2

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *CVPR*, 2021. 2, 5

[3] Gaetan Bahl, Mehdi Bahri, and Florent Lafarge. Single-shot end-to-end road graph extraction. In *CVPR EarthVision*, 2022. 2

[4] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *CoRR*, 2022. 2, 5

[5] Sarah Carter and Martin Herold. Specifications of land cover datasets for SDG indicator monitoring. https://ggim.un.org/ Accessed: 11-8-2022. 6

[6] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS*, 2018. 1, 6

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 8

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, 2020. 2, 3

[9] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 2

[10] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022. 2, 5

[11] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *ICIP*, 2018. 2, 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, 2020. 2

[13] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 2012. 2, 6

[14] Jan Gąsienica-Józkowy, Mateusz Knapik, and Bogusław Cyganek. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering*, 2021. 2

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 2

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 5

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTAEORS*, 2019. 2, 6

[20] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *AAAI*, 2019. 2

[21] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. Self-supervised pre-training enhances change detection in sentinel-2 imagery. In *ICPR*, 2021. 4

[22] Yahui Liu, Jian Yao, Xiaohu Lu, Menghan Xia, Xingbo Wang, and Yuan Liu. RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images. *TGRS*, 2018. 2

[23] William A Malila. Change vector analysis: an approach for detecting forest changes with landsat. In *LARS symposia*, 1980. 4

[24] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change event dataset for discovery from spatio-temporal remote sensing imagery. In *NeurIPS*, 2022. 2, 7

[25] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, 2021. 1, 2, 3, 5, 6

[26] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2

[27] Sorour Mohajerani and Parvaneh Saeedi. Cloud and cloud shadow segmentation for remote sensing imagery via filtered jaccard loss function and parametric augmentation. *J-STARS*, 2021. 2

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, 2018. 3

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *ICMICCAI*, pages 234–241. Springer, 2015. 6

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[32] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-

cover segmentation and classification. In *CVPR EarthVision*, 2022. 2

[33] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA*, 2017. 7

[34] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS*, 2019. 2, 6

[35] Xu Tang, Huayu Zhang, Lichao Mou, Fang Liu, Xiangrong Zhang, Xiao Xiang Zhu, and Licheng Jiao. An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning. *TGRSS*, 2021. 4

[36] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *CVPR*, 2022. 1, 2, 7

[37] Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *NeurIPS*, 2021. 2

[38] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *CoRR*, 2018. 2

[39] Adam Van Etten, Jacob Shermeyer, Daniel Hogan, Nicholas Weir, and Ryan Lewis. Road network and travel time extraction from multiple look angles with spacenet data. In *IGARSS*, 2020. 2

[40] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2

[41] Ben G Weinstein, Sergio Marconi, Stephanie A Bohlman, Alina Zare, Aditya Singh, Sarah J Graves, and Ethan P White. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *Elife*, 2021. 2

[42] Chen Wu, Hongruixuan Chen, Bo Do, and Liangpei Zhang. Unsupervised change detection in multi-temporal vhr images based on deep kernel pca convolutional mapping network. *IEEE Transactions on Cybernetics*, 2019. 4

[43] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[44] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2