Supplementary Material for "Change-Aware Sampling and Contrastive Learning for Satellite Images"

1. Overview

In this supplementary material, we present more details and extensive results that we could not fit in the main paper. In Sec. 2 we provide the additional implementation details for all the models that we use. We compare our geographical sampling with SeCo in Sec. 3. In Sec. 4 and Sec. 5 we present more qualitative and quantitative results respectively.

2. Implementation Details

In this section we look at the model architecture and training details of the models we use for linear probing and finetuning with our representations.

2.1. Landcover Classification

As stated in the main paper for landcover classification, we add a linear layer to the pre-trained backbone. The backbone is either frozen for linear evaluation or is finetuned. For EuroSat we perform the optimization by minimizing the cross-entropy loss over 10 classes. Since BigEarthNet is a multi-label classification problem we use multi-label soft margin loss. We use an Adam optimizer with default hyperparameters for both linear evaluation and finetuning. For linear evaluation, we use a learning rate of 10^{-3}, whereas we use a smaller learning rate of 10^{-5} for finetuning. We train the classifier for 100 epochs and reduce the learning rate by a factor of 10 at epochs 60 and 80. Following SeCo, we use a batch size of 32 for EuroSat and a batch size of 1024 for BigEarthNet.

2.2. Change Detection

Our model architecture follows past works [2, 4], using a U-Net [5] architecture. The pre-trained ResNet backbones are used as the encoder for the U-Net. The U-Net decoder uses the absolute feature differences at different resolutions. The decoder follows the architecture of the U-Net decoder [5]. The upsampling layer of the U-net uses the encoded feature maps at multiple resolutions to obtain a feature map at the highest resolution. This feature map is then passed through a 1×1 convolution layer with 1 output channel to obtain the logit map for changes.

We use a binary cross-entropy loss, to learn change vs no change per-pixel. Similar to SeCo, we train the decoder for 100 epochs and report results on the validation set. Since the images in the OSCD dataset have variable sizes, we also split them into non-overlapping patches of 96×96 pixels. We use a batch size of 32 with an Adam optimizer with a weight decay of 1e-4. The initial learning rate is set to 10^{-3} and is decreased exponentially with a multiplicative factor of 0.95 at each epoch. All these settings are unchanged from SeCo for fairness.

2.3. Semantic Segmentation

The architecture for semantic segmentation is similar to change detection. We use a U-Net with our pre-trained ResNet as the encoder. The decoder takes the feature maps as input instead of absolute values of feature differences. This final feature map is also passed through a 1×1 convolution layer with 7 output channels to obtain the logit map for the 7 semantic classes of Dynamic EarhNet.

Since the dataset is highly imbalanced we use Dice Loss [6] for training instead of cross-entropy. Our model is trained for 20 epochs. For both finetuning and frozen backbone, the learning rate is set to 10^{-3}, and reduced by a factor of 10 at epochs 12 and 16. We use 1024×1024 images in their original resolution with a batch size of 12. For training, we use random rotations, flipping, and brightness as augmentations. For fairness, all baselines are trained with the same training settings.

2.4. Change Event Retrieval

The change event retrieval model is trained on temporal slices of temporal change events. The training procedure follows [3]. $\langle V_{1\cdots l}, C_{1\cdots l-1} \rangle$ is a change event, where $V_{1\cdots l}$ are a sequence of l images and $C_{1\cdots l-1}$ are a sequence of $l-1$ change masks between successive pairs. The change event representation learning method learns by contrasting individual slices, i.e. $\langle V_t, V_{t+1}, C_t \rangle$. More specifically we use SimCLR on instances of $\langle V_t, V_{t+1}, C_t \rangle$. The features of a temporal slice are obtained by passing the pair of images V_t, V_{t+1} through a siamese ResNet backbone. The feature maps are concatenated in channel dimensions and averaged across the spatial dimensions using the downscaled change

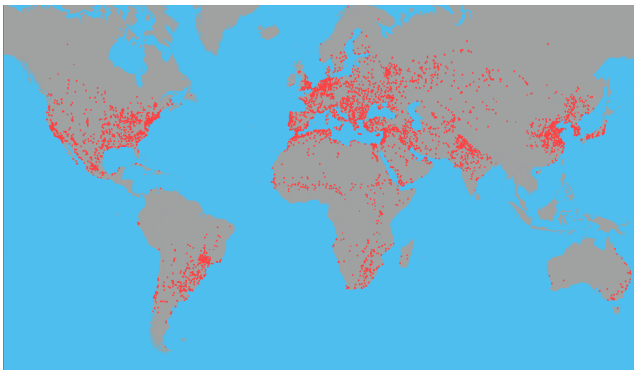

Figure 1. Distribution of locations sampled by our geographical sampling. We sample from a diverse set of regions near urban areas around the world.

mask C_t. During inference we need a feature representation for whole events ($\langle V_{1\dots l}, C_{1\dots l-1} \rangle$) and not just temporal slices. This feature is obtained by averaging features of temporal slices weighted with the fraction of change at each time $\sum C_t$.

We use a temperature of $\tau = 0.07$ and a stochastic gradient descent optimizer with a learning rate of 10^{-3}, for training the network, with a batch size of 256.

3. Geographical Sampling

Fig. 1 shows the locations sampled across the world. The samples are diverse across all 6 continents. Some populous regions such as regions in Southeast Asia cannot be sampled due to high cloud density in these regions.

Fig. 2 compares locations sampled in a coastal region (United Arab Emirates) by our method and SeCo. For coastal cities, SeCo samples many locations in the ocean. Other samples are very far from urban areas. Our method prevents sampling from oceans and focuses in the vicinity of urban areas.

Fig. 3 shows examples of satellite images from oceans when SeCo samples far away from cities. These images are not very informative.

4. Qualitative Results

During the training process, CACo automatically estimates locations where there is a significant long-term change (Sec. 3.4 in the main paper). Fig. 4 shows long-term image pairs with a high value of estimated change. These pairs show major real-world changes. For example, for pairs in the first, fourth, and fifth rows we see new constructions. In the second row, we can see changes in the water level. In the third row, we can see changes in land use patterns. These pairs are used as negative examples during contrastive learning of our features, and they act as hard negative examples.

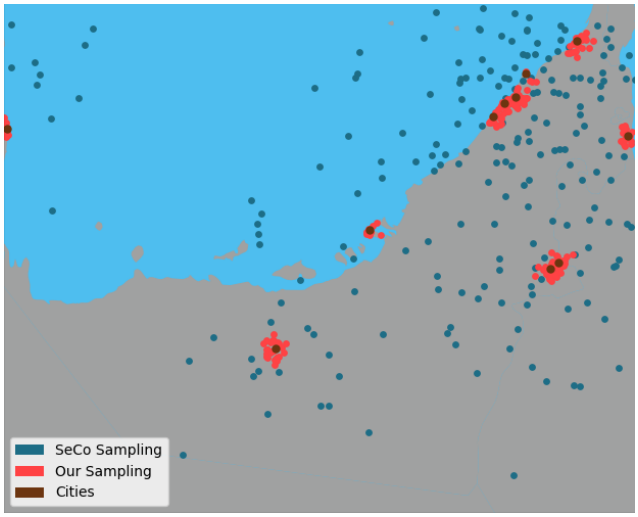

Figure 2. Comparison of sampling strategies of SeCo vs Ours near a coastal region. Many SeCo samples fall in the ocean and many other samples are very far from urban areas. Our method prevents both these sampling issues.

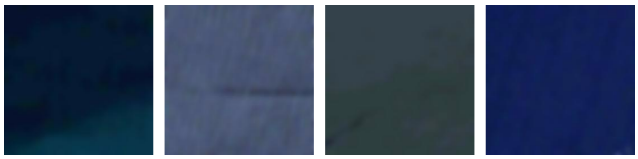

Figure 3. Less informative samples from SeCo geographical sampling. By investigating a small random subset of the SeCo dataset, we estimate that 22% of the sampled locations are uninformative.

Fig. 5 shows image pairs with a low value of change estimate. No major difference can be seen between the pairs other than slight season changes. For example, regions look more or less greener after the interval. Our contrastive learning framework learns to be invariant to such changes.

5. Quantitative Results

Functional Map of the World. We present a few more results that we could not include in the main paper. We evaluate our method on the scene recognition task on Functional Map of the World (FMoW) [1]. The dataset contains satellite views of images at multiple times for 62 different scenes such as parks, airports, shipyards, etc. Since the dataset, is at a different resolution we downscale images to match with the resolution of Sentinel-2 imagery. Additionally, since the dataset is originally designed for a different task of object detection using multitemporal images, we only use a single image for each scene.

We add a linear layer to the pre-trained backbone. Similar to EuroSat Evaluation we perform the optimization by minimizing the cross-entropy loss over 62 classes. We use

Data	Pre-training	ResNet-18		ResNet-50	
		top-1	top-5	top-1	top-5
-	Random init.	16.04	36.36	13.39	31.85
	ImageNet.	32.41	61.22	37.31	65.03
100k	MoCo v2	34.33	63.17	38.27	67.25
	SeCo	34.57	63.12	38.32	66.68
	CACo (ours)	**36.00**	**64.72**	**39.90**	**68.59**
1m	SeCo	38.84	67.35	43.64	71.89
	CACo (ours)	**39.13**	**68.06**	**44.12**	**72.52**

Table 1. Performance of our representation on the Functional Map of the World (FMoW) scene recognition task with linear probing, in top-1 and top-5 Accuracy. Our method provides a more accurate classification, with different backbones.

Backbone	Data	Pre-training	Fine-tuning
ResNet-18	-	Random init.	80.08
		ImageNet.	92.08
	100k	MoCo v2	94.94
		SeCo	96.71
		CACo (ours)	**97.02**
	1m	SeCo	97.25
		CACo (ours)	**97.47**
ResNet-50	-	Random init.	79.20
		ImageNet	93.41
	100k	SeCo	96.56
		CACo (ours)	**97.17**
	1m	SeCo	97.34
		CACo (ours)	**97.77**

Table 2. Performance of our method on the EuroSat landcover classification task when we finetune the whole network.

an Adam optimizer with default hyperparameters with a learning rate of 10^{-3}. We train the classifier for 100 epochs and reduce the learning rate by a factor of 10 at epochs 60 and 80. We use a batch size of 32 for training.

Tab. 1 shows the top-1 and top-5 classification accuracy of various pre-trained backbones on the FMoW dataset. Our method results in better features that are better suited for scene recognition.

EuroSat finetuning. Tab. 2 shows the performance of our method on EuroSat classification when we perform fine-tuning instead of linear classification. Even in the case of fine-tuning we see gains, albeit small. This shows that not only does our method learn a good representation, but it can also be used as a better initialization for transfer learning.

References

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

[2] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *ICIP*, 2018. 1

[3] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change event dataset for discovery from spatio-temporal remote sensing imagery. In *NeurIPS*, 2022. 1

[4] 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

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *ICMICCAI*, pages 234–241. Springer, 2015. 1

[6] 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. 1

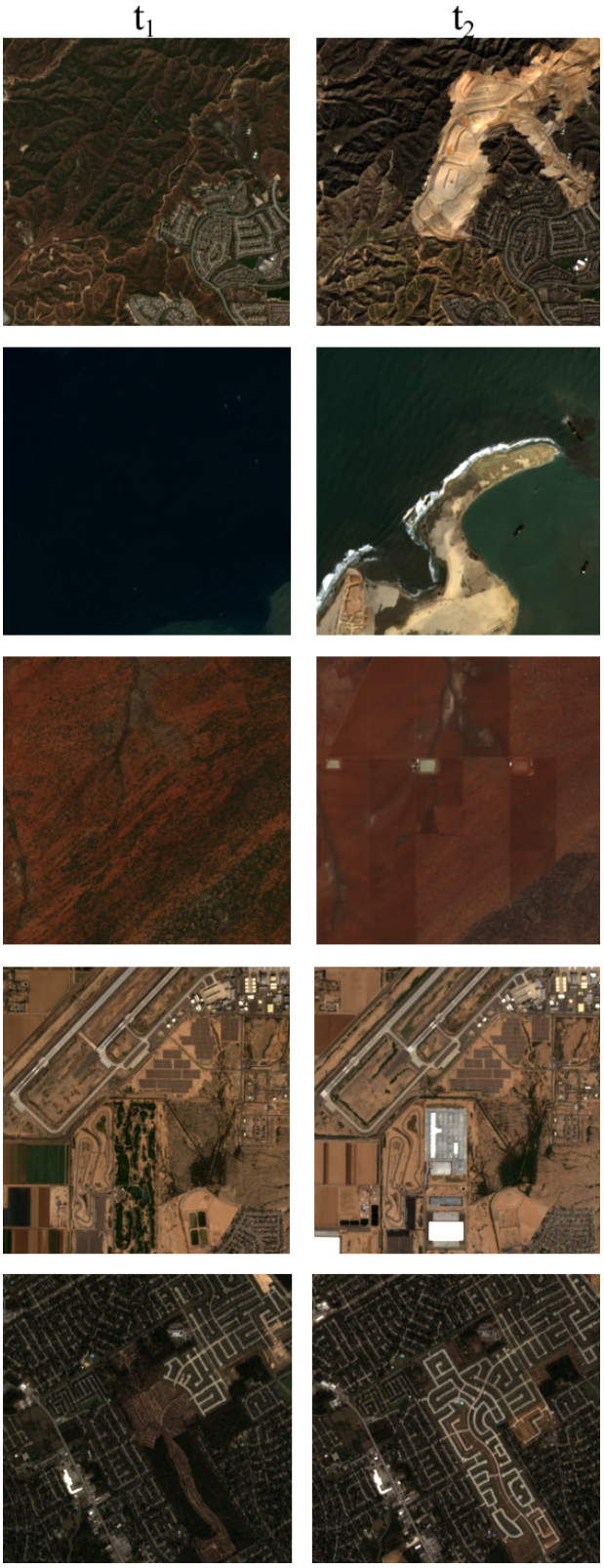

Figure 4. Examples of locations with a very high value of change estimate (> 2). These pairs show images at time t_1 and t_2 that are at least 4 years apart. Big changes can be seen between the pairs due to a change in land use or new constructions.

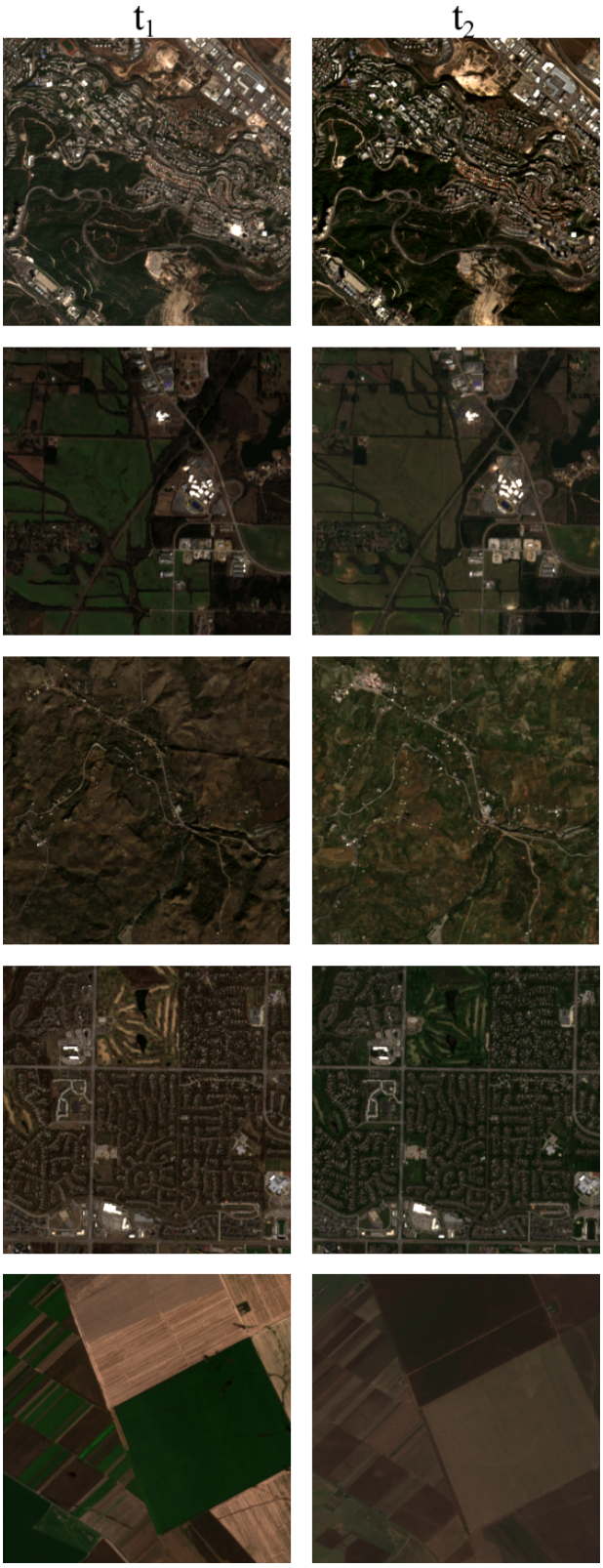

Figure 5. Examples of locations with a very low value of change estimate (close to 1). These pairs show images at time t_1 and t_2 that are at least 4 years apart. Only seasonal change can be observed between each pair.