# Evaluating the Label Efficiency of Contrastive Self-Supervised Learning for Multi-Resolution Satellite Imagery

Jules Bourcier, Gohar Dashyan, Jocelyn Chanussot, Karteek Alahari

# Evaluating the label efficiency of contrastive self-supervised learning for multi-resolution satellite imagery

Jules Bourcier[a,b], Gohar Dashyan[b], Jocelyn Chanussot[a], and Karteek Alahari[a]

[a]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[b]Preligens (ex-Earthcube), 75000 Paris, France

## ABSTRACT

The application of deep neural networks to remote sensing imagery is often constrained by the lack of ground-truth annotations. Adressing this issue requires models that generalize efficiently from limited amounts of labeled data, allowing us to tackle a wider range of Earth observation tasks. Another challenge in this domain is developing algorithms that operate at variable spatial resolutions, e.g., for the problem of classifying land use at different scales. Recently, self-supervised learning has been applied in the remote sensing domain to exploit readily-available unlabeled data, and was shown to reduce or even close the gap with supervised learning. In this paper, we study self-supervised visual representation learning through the lens of label efficiency, for the task of land use classification on multi-resolution/multi-scale satellite images. We benchmark two contrastive self-supervised methods adapted from Momentum Contrast (MoCo) and provide evidence that these methods can be perform effectively given little downstream supervision, where randomly initialized networks fail to generalize. Moreover, they outperform out-of-domain pretraining alternatives. We use the large-scale fMoW dataset to pretrain and evaluate the networks, and validate our observations with transfer to the RESISC45 dataset.

**Keywords:** deep learning, computer vision, remote sensing, self-supervised learning, land use classification, label-efficient learning, optical imagery

## 1. INTRODUCTION

The application of deep learning techniques to remote sensing imagery presents many challenges, one of the most important being the scarcity of annotations. Although large amounts of satellite imagery are readily available, curating and annotating them for a specific Earth observation tasks is usually very expensive, time-consuming and requires fine domain expertise. This implies that it is impractical in many real-world contexts to acquire the data needed to effectively leverage classical supervised learning methods. From this perspective, it is necessary to develop label-efficient approaches, i.e., models that are able to learn with few annotated samples. Self-supervised learning (SSL) is a promising approach for this purpose, as it pretrains representations without requiring human labeling. Inspired by the success of recent methods on natural images benchmarks,[1–5] SSL has been applied in the remote sensing domain to exploit the plentiful unlabeled data, and was shown to reduce or even close the gap with supervised learning and transfer from ImageNet[6,7] (IN).

Another common problem in remote sensing is to process images covering various spatial scales, e.g., for the task of classifying land use, where categories can range from individual storage tanks to full harbors. The capacity of SSL methods to generalize from few labels on this important problem was not explored by previous works, to the best of our knowledge.

To address this, in this paper we study self-supervised visual representation learning through the lens of label efficiency, for the classification of land use at different spatial resolutions. We benchmark two contrastive self-supervised methods adapted from Momentum Contrast (MoCo)[1] to assess their capacity to learn generic, multi-resolution features, which do not need many labeled examples for downstream image classification. We use the large-scale and diverse fMoW dataset[8] to pretrain and evaluate the networks, and validate our observations with transfer to the RESISC45 dataset,[9] using diverse evaluation methods and amounts of labels. We provide

Further author information: (Send correspondence to J. Bourcier)
J. Bourcier: E-mail: jules.bourcier@preligens.com

evidence that these methods can be trained effectively in few-label settings that are insufficient for randomly initialized networks to generalize. Thanks to MoCo with temporal positives,,[6] when finetuning the pretrained models to RESISC45 with only about 4 examples per class, we reach 5× the accuracy of a classifier trained from scratch. Moreover, with simple linear probing on frozen representations, we surpass from-scratch networks in every label setting on fMoW and RESISC45. Additionally, the MoCo variants applied on fMoW images outperform out-of-domain pretraining on IN by significant margins, despite being pretrained on 3× less data. We also reveal that a basic $k$-nearest neighbors ($k$-NN) classifier on the learned representations provides out-of-the-box efficient generalization, and competes or outperforms finetuning with other methods when only few labels are available.

Our main contributions are as follows:

- We experiment with two contrastive SSL methods, MoCo[1] and MoCoTP,[6] on multi-resolution land use classification on the fMoW dataset, and observe their efficiency in terms of annotations required, on the common evaluation settings for representation learning with $k$-NN, linear, and finetuned classifiers.
- To demonstrate the transferability of the pretrained representations for land use prediction at different spatial resolutions, we study how the models pretrained on fMoW generalize to the smaller RESISC45 dataset, including extremely few images per class.

## 2. RELATED WORK

### 2.1 Self-supervised learning and contrastive learning

SSL methods learn representations of data without relying on manual annotations. It consists in *pretraining* a neural network to solve a *pretext task* on unlabeled data, for the purpose of extracting semantic representations that allow for effective *transfer* to downstream predictive tasks such as classification, segmentation or object detection. In computer vision, the popularity of SSL is due to recent methods that have shown to perform comparably well or even better than their supervised counterparts on natural image benchmarks.[1–5]

*Contrastive learning* has established itself as the staple framework for SSL of visual representations, with approaches such as MoCo,[1] SimCLR,[2] and SwAV.[3] These methods work by attracting embeddings of pairs of sample images known to be semantically similar (*positive pairs*) while simultaneously repelling pairs of dissimilar samples (*negative pairs*). The most common way to define similarity is with the *instance discrimination* pretext task,[10] in which positives are generated as random data augmentations on the same image, and negatives are simply generated from different images. Thanks to this objective, the encoder learns to encode close representations for different views of the same object instance in an image and distant representations for other instances. In this work, we use the strong contrastive method of MoCo[1] as well as an extension proposed in Ref. 6 that adapts the learning objective to the spatio-temporal structure of satellite imagery.

### 2.2 SSL in remote sensing

Following their success in computer vision, several works have applied SSL methods to remotely sensed imagery for Earth observation tasks. Ref. 11 was one of the first to use of contrastive learning for remote sensing representation learning. Ref. 12 applies a spatial augmentation criteria on top of MoCo.[1] These works exploit a relevant assumption about the remote sensing domain: images that are geographically close should be semantically more similar than distant images. Another way of making the learning procedure *geography-aware* is to exploit the spatio-temporal nature of satellite imagery. Ref. 6 uses spatially-aligned images over time to construct *temporal positive pairs* with MoCo. The resulting temporally-aligned features were shown to improve generalization for classification, segmentation and object detection downstream tasks. In this work, we adopt this model with notable improvements, to study how such temporal invariance can benefit learning from few labels. In the same vein, Ref. 7 proposes a method that learns representations that are simultaneously variant and invariant to temporal changes. One can also exploit the multi-spectral and multi-sensor nature of remote sensing. Ref. 13 applies CMC[14] on multi-spectral images, using different subsets of channels as augmented (positive) views. Ref. 15 extends this to co-located images from multiple sensors, combining different sensor channels to construct positive pairs.

### 2.3 Label efficiency of remote sensing representations

Acquiring large amounts of labeled remote sensing data is often hard and prohibitively expensive. Surprisingly, few works have specifically studied how models are able to learn to solve Earth observation tasks with few annotated examples. The SSL method of Ref. 7 was shown to be useful for *label-efficient* transfer learning on multi-label land cover classification on the BigEarthNet dataset.[16] It was benchmarked on Sentinel-2 image chips from a fixed resolution of 10m and a rather different semantic domain (mainly land cover in natural environments) from the functional land use classification task of fMoW[8] studied in the present work. Ref. 17 shows that in-domain supervised pretraining can improve the performance in low-labeled settings on downstream classification datasets. However, it shows that transfer performance is very dependent on the labeling and data curation quality in the upstream dataset and leaves unresolved the problem of obtaining generic representations with less dependence on labels. Ref. 18 proposes a weakly-supervised multi-modal pretraining method using paired satellite images and geo-located Wikipedia articles, and outperforms other pretraining strategies on fMoW classification when finetuned on small amounts of labeled data. However this is a rather different direction than SSL for learning representations and requires each satellite image to be paired to its corresponding article, which severely limits the amount of images that can be used. In this work, we investigate contrastive SSL for label-efficient learning on multi-resolution imagery, which, to the best of our knowledge, has not been previously done. The three-faceted evaluation we perform with linear, finetuning and $k$-NN classification, is also more complete that what previous works have employed in their protocols.

## 3. METHOD

### 3.1 Datasets

We study the task of land use classification from satellite images, on the fMoW[8] and RESISC45[9] datasets.

fMoW[8] is a large-scale dataset containing 363,571 RGB training images, across 62 fine-grained and diverse categories of functional land use. It provides several images from same locations over time, over 83,412 locations across 207 countries for training. Notably, the objects represented cover a varying range of ground resolution, from 0.5m for small structures (e.g., wind turbine) to 35m for large facilities (e.g., airport). As fMoW is a big, diverse, and multi-resolution dataset, we use it for self-supervised pretraining with the hope to learn rich semantic representations for remote sensing. We also use it for evaluation of the pretrained networks on the land use classification task with the included labels.

For complementarity, we selected RESISC45 for evaluation by transfer from fMoW pretraining to a separate downstream dataset. RESISC45 is of much smaller scale, with 18,900 RGB training images divided into 45 land use classes. The images are extracted from Google Earth from over 100 countries. The categories are similar or overlap with the ones in fMoW, and the data is characterized by a multi-resolution distribution close to fMoW, ranging from 0.2 to 30m. Despite this, there is a degree of domain gap between fMoW and RESISC45, at least through some unrelated categories, which is interesting for the evaluation of the transferability of features pretrained on fMoW.

Therefore, these two datasets are well suited for studying label-efficient representation learning on the task of multi-resolution and multi-scale land use classification.

### 3.2 Models

We employ the MoCo[1] framework for contrastive SSL, specifically two different variants of the proposed MoCo-V2[19] model. They differ in the selection of the positive views: we take (i) the usual MoCo setup of selecting two artificially augmented samples of the same image as positive views, and (ii) the MoCo with Temporal Positives (MoCoTP) setup,[6] which leverages the temporal resolution of geospatial data for generating positive pairs, i.e., positive views are sampled in temporal sequences of spatially aligned images, and the same augmentations of the usual MoCo setup are then applied to these temporal views. Compared to the MoCoTP framework from Ref. 6, we further add two improvements: (i) additional augmentations for rotational invariance; (ii) a fixed loss function that removes the false temporal negatives in the learning process. These improvements are detailed in Appendix B.

| Method | F1-score | | Accuracy | |
|---|---|---|---|---|
| | Linear († / ∗) | Finetune († / ∗) | Linear († / ∗) | Finetune († / ∗) |
| Random init | - | 64.71 / 65.39 | - | 69.05 / 69.33 |
| IN-sup | - / 50.25 | 64.72 / 66.01 | - / 54.56 | 69.07 / 70.80 |
| IN-MoCo | 31.55 / 53.47 | 57.36 / 65.37 | 37.05 / 57.28 | 62.90 / 70.17 |
| MoCo | 55.47 / 65.55 | 60.61 / 67.23 | 60.69 / 69.62 | 64.34 / 71.82 |
| MoCoTP | **64.53** / **68.89** | **67.34** / **68.96** | **68.32** / **72.56** | **71.55** / **73.01** |

Table 1. Baseline improvements on fMoW classification. We report top-1 F1-score and accuracy (in %) on the fMoW validation set, on linear probing and finetuning. Methods are different pretrainings: "IN-" means using ImageNet transfer, "sup" means supervised pretraining, and "Random init" represents no pretraining. "†" denote results from Ref. 6. "∗" denote our improved reproductions.

We compare the two self-supervised methods against the following baselines: supervised pretraining on IN, pretraining with MoCo on IN, and random initialization (i.e., no pretraining). For RESISC45, we also add the baseline of supervised pretraining on fMoW (with the available land use labels), in addition to our SSL models on fMoW. We use a ResNet-50[20] encoder in all experiments.

The methods are evaluated for classification with three common evaluation procedures: (i) *linear probing*, that trains a logistic regression on top of frozen features from the pretrained encoder that maps to the output logits for each class of the target task; (ii) *finetuning*, which trains a logistic regression and also updates all the parameters of the network end-to-end with the target labels; (iii) *k-NN*, that matches the frozen feature of an image to the $k$ nearest stored training features that votes for the label.

The SSL models are trained on all the images of the fMoW dataset (without labels). We then evaluate their performance against supervised baselines on the target classification task, by training a classifier on three subsets containing 1, 10, and 100% of the labeled training data. When sampling a subset of the full training set (1% and 10% labels), we conduct 3 Monte Carlo experiments, so that the results account for variance induced by data selection. Sampling is stratified, i.e., we ensure that the overall distribution of classes is maintained in every subset.

The implementation details for datasets and trainings are provided in Appendix C.

# 4. RESULTS

## 4.1 fMoW classification

### 4.1.1 Baseline improvements

We first show the performance enhancements brought by our modifications to the SSL framework of Ref. 6. Table 1 shows the results of linear probing and finetuning on the 62-class land use classification task of fMoW of the methods with the proposed improvements, compared to the results from Ref. 6. Models are evaluated with single-image top-1 F1-score averaged over classes as well as accuracy on the validation set of fMoW.

With 100% labels, we see that our improved reproduction of MoCoTP increases performance by 4.36% F1 compared to Ref. 6 in linear probing and 1.62% in finetuning. This shows that the use of the rotation augmentations and the correction of false negatives in the loss function are helpful, especially for linear probing, which closes the gap with finetuning completely. As a note, the additional augmentations also improve all other baselines, e.g. it improves random initialization by 1.29% F1, showing the general relevance of rotational invariance for satellite imagery.

### 4.1.2 Label efficiency

We now study label-efficient classification on fMoW by evaluating the models with $k$-NN, linear, and finetuning classifiers, for fractions of 1, 10, and 100% labeled training data. Table 2 shows the results. Figure 1 in the appendix provides a graphical view of the same results.

| Method | 1% labels | | | 10% labels | | | 100% labels | | |
|---|---|---|---|---|---|---|---|---|---|
| | $k$-NN | Linear | Finetune | $k$-NN | Linear | Finetune | $k$-NN | Linear | Finetune |
| Random init | - | - | 19.29 (1.65) | - | - | 51.87 (0.50) | - | - | 65.39 |
| IN-sup | 21.04 (0.48) | 32.41 (0.17) | 39.43 (1.53) | 31.92 (0.25) | 43.86 (0.07) | 57.32 (0.07) | 39.65 | 50.25 | 66.01 |
| IN-MoCo | 22.59 (0.19) | 35.10 (0.26) | 37.93 (0.75) | 32.71 (0.40) | 46.85 (0.47) | 56.73 (0.52) | 40.72 | 53.47 | 65.37 |
| MoCo | 47.34 (0.22) | 53.61 (0.30) | 52.38 (0.45) | 52.89 (0.15) | 61.13 (0.18) | 60.80 (0.31) | 57.57 | 65.55 | 67.23 |
| MoCoTP | **56.75 (0.41)** | **60.05 (0.11)** | **60.00 (0.43)** | **61.58 (0.09)** | **66.15 (0.11)** | **66.35 (0.75)** | **64.86** | **68.89** | **68.96** |

Table 2. Label-efficient land use classification on fMoW. We report top-1 F1-score (in %) on the fMoW validation set, for 1, 10 and 100% labeled training data, with $k$-NN, linear probing and finetuning classifiers. Methods are different pretrainings: "IN-" means using ImageNet transfer, "sup" means supervised pretraining, and "Random init" represents no pretraining. For 1% and 10% labels, the values are 'mean (sd)' of 3 runs with random sampling of a subset of the full training set.

On linear probing, we observe that, while the supervised baseline shows poor generalization from few labels, the MoCo variants retain a much higher performance in comparison as the fraction of labels is reduced. On the fewest number of labels, MoCo and MoCoTP respectively give +30 and +35% accuracy against random initialization. In the semi-supervised settings of 1% and 10% labels, we see that MoCoTP respectively give 96% and 87% of the performance reached by 100% labels, and furthermore outperforms the IN-pretrained methods by large margins. These results indicate that in-domain contrastive SSL with the use of temporal priors is very efficient at learning semantic features from the upstream dataset.

When finetuning, interestingly, we see that MoCo and MoCoTP do not significantly improve performance against linear probing on the 1% and 10% settings. This means that they have learned to the best of their capacity to represent the classes linearly, which implies high-level semantic representations. On the contrary, IN-pretrained models are significantly improved by finetuning due to the domain gap between IN and fMoW; even so, they show inferior label efficiency that MoCo and MoCoTP, e.g., IN-supervised weights retains only 60% of its maximum performance on 1% of labels vs. 87% for MoCoTP. Perhaps surprisingly, we also remark that both MoCo and MoCoTP improve over the all supervised counterparts in the many-label regime of 100%, e.g., by +1.84% and +3.57% F1 respectively vs. random initialization, showing than SSL can also be useful if we have large quantities of annotations for a task.

Finally, under the $k$-NN evaluation, we see that MoCoTP outperforms all other methods, and that for 1% and 10% labels, it provides even better classifier than other pretrainings with finetuned networks – e.g., on 1% fMoW, MoCoTP with $k$-NN is better than finetuned IN-supervised weights by 18.82% F1. This further confirms that MoCoTP produces high quality and label-efficient features.

### 4.2 Transfer to RESISC45 classification

A main goal of representation learning is to improve generalization on new tasks and datasets through feature re-use. We then evaluate label-efficient transfer learning on the RESISC45 dataset, with the same three classifiers and fraction of labeled samples as in Sec. 4.1.2. Table 3 shows the results on the 45-class land use classification task of RESISC45. Models are evaluated with top-1 accuracy on the testing set. Figure 2 in the appendix provides a graphical view of the same results.

In the case of linear probing, we see that MoCo and MoCoTP pretrained on fMoW are label efficient with features adapted for linear classification on this different downstream dataset. While the supervised end-to-end baseline shows a large gap of performance of 74.83% accuracy between training of 100 and 1% of labels, due to the relatively small size of the dataset, the accuracy of MoCoTP for 1% of labels is only 23.64% accuracy lower than for 100% of labels, achieving 68.93%.

When finetuning, self-supervised features do not provide a significant advantage against IN pretraining when increasing the percentage of labeled data: we see that all the pretrained networks saturate on the full dataset at around 95% accuracy. Nevertheless, fMoW-MoCoTP pretraining is more label-efficient than other methods in the very-low label regime of 1%, outperforming IN-supervised pretraining by 5.28% accuracy. This shows the capability of specialized pretraining with MoCoTP to improve generalization from very few labels, as in this setting there is only 175 examples, about 4 examples per class.

| Method | 1% labels | | | 10% labels | | | 100% labels | | |
|---|---|---|---|---|---|---|---|---|---|
| | $k$-NN | Linear | Finetune | $k$-NN | Linear | Finetune | $k$-NN | Linear | Finetune |
| Random init | - | - | 13.76 (3.83) | - | - | 53.07 (1.04) | - | - | 88.59 |
| IN-sup | 49.45 (1.20) | 53.91 (0.20) | 64.87 (1.27) | 70.51 (0.27) | 77.01 (0.34) | 87.44 (0.33) | 79.32 | 86.48 | 95.38 |
| IN-MoCo | 43.67 (1.28) | 49.79 (0.50) | 59.45 (1.04) | 65.94 (0.32) | 77.21 (0.53) | 84.98 (0.84) | 75.40 | 87.32 | **95.41** |
| fMoW-sup | 62.12 (2.11) | 65.86 (1.70) | 67.66 (2.17) | 77.13 (0.37) | 81.42 (0.34) | 88.14 (0.63) | 83.65 | 87.94 | 94.14 |
| fMoW-MoCo | 60.23 (0.60) | 64.22 (1.20) | 65.94 (1.00) | 79.61 (0.46) | 84.33 (0.32) | 87.69 (0.33) | 85.68 | 91.40 | 94.84 |
| fMoW-MoCoTP | **65.47 (1.10)** | **68.93 (0.86)** | **71.15 (0.33)** | **83.16 (0.17)** | **87.18 (0.28)** | **89.16 (0.15)** | **88.37** | **92.57** | 95.30 |

Table 3. Label-efficient land use classification on RESISC45. We report top-1 accuracy (in %) on the RESISC45 testing set, for 1, 10 and 100% labeled training data, with $k$-NN, linear probing and finetuning classifiers. Methods are different pretrainings: "IN-" means transfer from ImageNet, "fMoW-" means transfer from fMoW, "sup" means supervised pretraining, and "Random init" represents no pretraining. For 1% and 10% labels, the values are 'mean (sd)' of 3 runs with random sampling of a subset of the full training set.

With $k$-NN classifiers, fMoW-MoCoTP features provide the best performance over all annotations regimes. When given 10% of labels, it is on par with fMoW-supervised pretraining on 100% labels, with 83.16 vs. 83.65% accuracy. Furthermore, on 100% of labels, it gives similar performance to from-scratch training (88.37 vs. 88.59% accuracy). Note that $k$-NN is the simplest and weakest classifier one can build on top of pretrained networks. These results indicate that in-domain SSL with temporal priors is effective and label-efficient for out-of-the-box transfer to a new downstream dataset.

## 5. CONCLUSION

We present a study of the annotation-efficiency of self-supervised contrastive learning for remote sensing, on the specific problem of multi-resolution and multi-scale land use classification. The results demonstrate the potential of methods based on MoCo, which provides a generalization capability from few labels that is not achievable with classical supervised predictors. We show that the use of MoCo with temporal positives further improves label-efficient learning. Our observations indicate that SSL is a promising direction for solving Earth observation tasks that have previously been inaccessible due to the scarcity of annotations. We can suppose that improvements to label efficiency could otherwise be promoted by scaling up the amount of pretraining data, performing more pretraining epochs, or using higher-capacity network architectures. Further work could include additional studies such as extending the sampling of sizes for the downstream datasets, investigating the multi-scale aspects of features, expanding to object detection tasks, and experimenting with different SSL methods, such as SwAV[3] or BYOL.[4]

## APPENDIX A. GRAPHICAL VIEW OF LABEL-EFFICIENT CLASSIFICATION

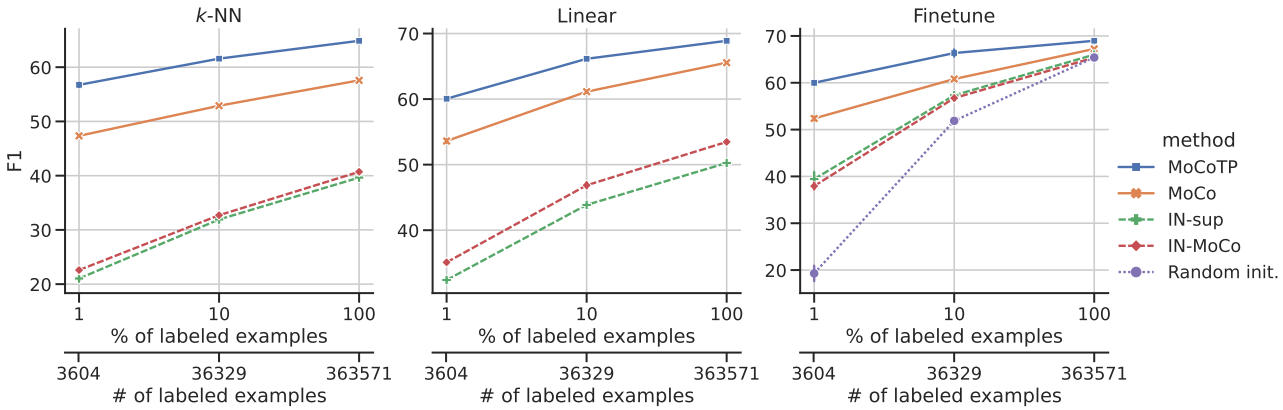


Figure 1. Label-efficient land use classification on fMoW. These values correspond to Tab. 2 in the main text, see the table caption for description.

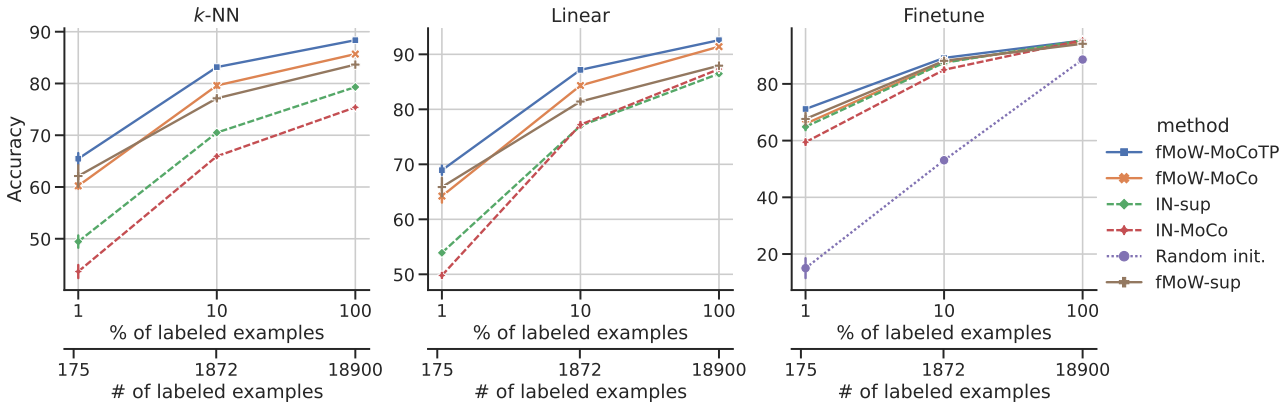


Figure 2. Label-efficient land use classification on RESISC45. These values correspond to Tab. 3 in the main text, see the table caption for description.

# APPENDIX B. BASELINE IMPROVEMENT DETAILS

Compared to the framework of Ref. 6, we add two modifications to improve the models in our experiments.

## B.1 Additional augmentations

In addition to the geometric and color perturbations of MoCo-V2, we apply random horizontal flips and rotations by multiples of 90°. Since the data augmentation scheme plays a leading role in contrastive learning,[21] we aim to learn representations more suited to overhead images thanks to rotational invariance. These additional augmentations are also used for supervised finetuning of all MoCo models and baselines.

## B.2 False temporal negatives removal in MoCoTP

MoCo learns to match an input *query* $q$ to a *key* $k^+$ (representing the encoded views of the same sample) among a set of negative keys $k^-$, using the instance discrimination pretext task[10]–we refer readers to Ref. 1 for details on MoCo. MoCo uses the popular choice of InfoNCE[22] for the contrastive loss:

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

where $\tau$ is a temperature scaling parameter.

MoCoTP extends the instance discrimination pretext task to use spatially aligned images from different times as positives. This is implemented as a drop-in replacement for $q$ and $k^+$ in Eq. (1). However, this can introduce *false negatives*. Indeed, at each iteration of training, it may happen that the set of negatives $k^-$ contains temporal views for samples of the current mini-batch of queries. Such false negatives will cause an incorrect repulsion between the embeddings of similar samples. To what extent this is detrimental to the learned representations depends on the probability of sampling temporal pairs in the training set, as well as on the size of the queue. To avoid the false negatives to interfere with the learning objective, we simply mask out the logits $q \cdot k^-$ in the InfoNCE loss in (1) for every $k^-$ that happens to be a temporal view of $q$.

# APPENDIX C. IMPLEMENTATION DETAILS

Here we describe the datasets and training settings used for experiments.

## C.1 Datasets details

### C.1.1 fMoW

Following Refs. 8 and 6, we use the fMoW-RGB products in our experiments, which provides 3-bands imagery at 0.5m ground resolution. We use the official train and validation splits, composed of 363,571 images and 50,041 images respectively. Preprocessing is applied identically to Ref. 8 and input images are resized to 224×224 pixels.

### C.1.2 RESISC45

We use the 224×224 images of RESISC45 without specific preprocessing. We use the train/validation/test splits defined in Ref. 17, composed of 18,900, 6,300 and 6,300 images respectively.

## C.2 Training details

In all our experiments, we use a ResNet-50 architecture for the encoder. For linear, finetuning and $k$-NN evaluation, all our hyperparameters apply to both fMoW and RESISC45 experiments, and are the same regardless of the training subset size. We use the `PyTorch` framework in our code which is based on the official implementation of Ref. 6*. All trainings are performed on compute nodes with 4 Tesla V100 GPUs.

### C.2.1 MoCo pretraining

Pretraining with the two MoCo variants is performed with the following hyperparameters: learning rate of 3e-2 with a cosine schedule, batch size of 256, dictionary queue size of 65536, temperature scaling of 0.2, SGD optimizer with a momentum of 0.9, weight decay of 1e-4. We use MoCo-V2 as the contrastive framework.[19] The data augmentation scheme is the random composition of resized cropping, horizontal flipping, color jittering, Gaussian blur, horizontal flipping, and 90° rotations. Pretraining is performed for 200 epochs.

### C.2.2 Linear probing

For linear evaluations, we train a supervised linear classifier on top of frozen features from the output of the ResNet global average pooling layer. For MoCo methods, we use a learning rate of 1 reduced by a multiplicative factor of 0.5 when the validation loss plateaus for 5 epochs, a batch size of 256, no weight decay, and only random resized cropping for the augmentations. For supervised methods, hyperparameters are identical except we use a learning rate of 1e-3. All models are trained with cross-entropy loss until convergence of the validation loss, and evaluated on epoch with the best top-1 accuracy on the validation set.

### C.2.3 Finetuning

For finetuning evaluations, we initialize networks with the pretrained weights and adapt them during training. For MoCo methods, we use a learning rate of 3e-4 for ResNet weights and 1 for the final classification layer, reduced reduced by a multiplicative factor of 0.5 when the validation loss plateaus for 2 epochs; a batch size of 256, no weight decay, and the same augmentations used for MoCo pretraining. For supervised methods, hyperparameters are identical except we use a single learning rate of 1e-3 and weight decay of 1e-4. All models are trained with cross-entropy loss until convergence of the validation loss, and evaluation on epoch with the best top-1 accuracy on the validation set.

### C.2.4 $k$-NN

We adopt the weighted nearest neighbors classifier of Ref. 10, following common practice. We freeze the pretrain model to compute features at the output of the ResNet global average pooling layers. We do not apply any data augmentation. The classifier has only one hyperparameter: the number of nearest neighbors $k$. We tune $k$ and find that a value of 200 consistently works best for our MoCo methods; we then use this value across all experiments.

### C.2.5 ImageNet-pretrained weights

We use available pretrained networks on IN-1K[23] for baselines. For the IN-supervised method, we use the standard weights available in the `torchvision`† library. For the IN-MoCo method, we take the official weights of MoCo-V2 pretrained for 200 epochs‡.

*https://github.com/sustainlab-group/geography-aware-ssl

†https://pytorch.org/vision

‡available at https://github.com/facebookresearch/moco

## ACKNOWLEDGMENTS

We thank Thomas Floquet, Tugdual Ceiller, and other colleagues at Preligens for fruitful discussions and for providing important feedback. Karteek Alahari was supported in part by the ANR grant AVENUE (ANR-18-CE23-0011). This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011013097 made by GENCI.

## REFERENCES

[1] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R., "Momentum contrast for unsupervised visual representation learning," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 9729–9738 (2020).

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

[3] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A., "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems* **33**, 9912–9924 (2020).

[4] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., "Bootstrap your own latent - a new approach to self-supervised learning," *Advances in Neural Information Processing Systems* **33**, 21271–21284 (2020).

[5] Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S., "Barlow twins: Self-supervised learning via redundancy reduction," in [*International Conference on Machine Learning*], 12310–12320, PMLR (2021).

[6] Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., and Ermon, S., "Geography-aware self-supervised learning," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 10181–10190 (2021).

[7] Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., and Rodriguez, P., "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 9414–9423 (2021).

[8] Christie, G., Fendley, N., Wilson, J., and Mukherjee, R., "Functional map of the world," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 6172–6180 (2018).

[9] Cheng, G., Han, J., and Lu, X., "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE* **105**(10), 1865–1883 (2017).

[10] Wu, Z., Xiong, Y., Yu, S. X., and Lin, D., "Unsupervised feature learning via non-parametric instance discrimination," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 3733–3742 (2018).

[11] Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., and Ermon, S., "Tile2vec: Unsupervised representation learning for spatially distributed data," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **33**(01), 3967–3974 (2019).

[12] Kang, J., Fernandez-Beltran, R., Duan, P., Liu, S., and Plaza, A. J., "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing* **59**(3), 2598–2610 (2021).

[13] Stojnic, V. and Risojevic, V., "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 1182–1191 (2021).

[14] Tian, Y., Krishnan, D., and Isola, P., "Contrastive multiview coding," in [*European conference on computer vision*], 776–794, Springer (2020).

[15] Swope, A. M., Rudelis, X. H., and Story, K. T., "Representation learning for remote sensing: An unsupervised sensor fusion approach," *arXiv preprint arXiv:2108.05094* (2021).

[16] Sumbul, G., Charfuelan, M., Demir, B., and Markl, V., "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in [*IGARSS-IEEE International Geoscience and Remote Sensing Symposium*], 5901–5904, IEEE (2019).

[17] Neumann, M., Pinto, A. S., Zhai, X., and Houlsby, N., "In-domain representation learning for remote sensing," in [*AI for Earth Sciences Workshop at International Conference on Learning Representations (ICLR)*], 1–20 (Apr. 2020).

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

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

[20] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[21] Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P., "What makes for good views for contrastive learning?," *Advances in Neural Information Processing Systems* **33**, 6827–6839 (2020).

[22] Oord, A. v. d., Li, Y., and Vinyals, O., "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748* (2018).

[23] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, Ieee (2009).