This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Geography-Aware Self-Supervised Learning

Kumar Ayush*[1], Burak Uzkent*[1], Chenlin Meng[1], Kumar Tanmay[2], Marshall Burke[1], David Lobell[1], and Stefano Ermon[1]

[1]Stanford University [2]IIT Kharagpur

Abstract

Contrastive learning methods have significantly narrowed the gap between supervised and unsupervised learning on computer vision tasks. In this paper, we explore their application to geo-located datasets, e.g. remote sensing, where unlabeled data is often abundant but labeled data is scarce. We first show that due to their different characteristics, a non-trivial gap persists between contrastive and supervised learning on standard benchmarks. To close the gap, we propose novel training methods that exploit the spatio-temporal structure of remote sensing data. We leverage spatially aligned images over time to construct temporal positive pairs in contrastive learning and geo-location to design pre-text tasks. Our experiments show that our proposed method closes the gap between contrastive and supervised learning on image classification, object detection and semantic segmentation for remote sensing. Moreover, we demonstrate that the proposed method can also be applied to geo-tagged ImageNet images, improving downstream performance on various tasks.

1. Introduction

Inspired by the success of self-supervised learning methods [2, 12], we explore their application to large-scale remote sensing datasets (satellite images) and geo-tagged natural image datasets. It has been recently shown that self-supervised learning methods perform comparably well or even better than their supervised learning counterpart on image classification, object detection, and semantic segmentation on traditional computer vision datasets [20, 9, 12, 2, 1]. However, their application to remote sensing images is largely unexplored, despite the fact that collecting and labeling remote sensing images is particularly costly as annotations often require domain expertise [33, 15, 4].

In this direction, we first experimentally evaluate the performance of an existing self-supervised contrastive learning method, MoCo-v2 [12], on remote sensing datasets, finding a performance gap with supervised learning using labels. For instance, on the Functional Map of the World (fMoW) image classification benchmark [4], we observe an 8% gap in top 1 accuracy between supervised and self-supervised methods.

To bridge this gap, we propose geography-aware contrastive learning to leverage the spatio-temporal structure of remote sensing data. In contrast to typical computer vision images, remote sensing data are often geo-located and might provide multiple images of the same location over time. Contrastive methods encourage closeness of representations of images that are likely to be semantically similar (positive pairs). Unlike contrastive learning for traditional computer vision images where different views (augmentations) of the same image serve as a positive pair, we propose to use *temporal positive pairs* from spatially aligned images over time. Utilizing such information allows the representations to be invariant to subtle variations over time (e.g., due to seasonality). This can in turn result in more discriminative features for tasks focusing on spatial variation, such as object detection or semantic segmentation (but not necessarily for tasks involving temporal variation such as change detection). In addition, we design a novel unsupervised learning method that leverages geo-location information, i.e., knowledge about where the images were taken. More specifically, we consider the pretext task of predicting where in the world an image comes from, similar to [10, 11]. This can complement the information-theoretic objectives typically used by self-supervised learning methods by encouraging representations that reflect geographical information, which is often useful in remote sensing tasks. Finally, we integrate the two proposed methods into a single geography-aware contrastive learning objective.

Our experiments on the functional Map of the World [4] dataset consisting of high spatial resolution satellite images show that we improve MoCo-v2 baseline significantly. In particular, we improve it by $\sim 8\%$ classification accuracy

*Equal Contribution. Contact: {kayush, buzkent}@cs.stanford.edu

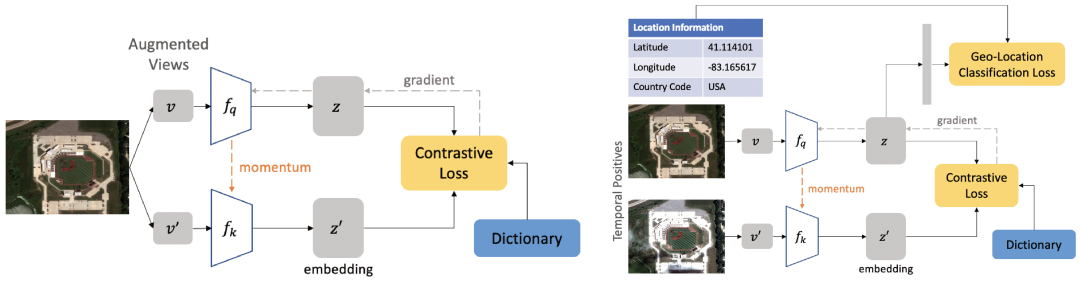

Figure 1: **Left** shows the original MoCo-v2 [2] framework. **Right** shows the schematic overview of our approach.

when testing the learned representations on image classification, $\sim$ 2% AP on object detection, $\sim$ 1% mIoU on semantic segmentation, and $\sim$ 3% top-1 accuracy on land cover classification. Interestingly, our geography-aware learning can even outperform the supervised learning counterpart on temporal data classification by $\sim$ 2%. To further demonstrate the effectiveness of our geography-aware learning approach, we extract the geo-location information of ImageNet images using FLICKR API similar to [6], which provides us with a subset of 543,435 geo-tagged ImageNet images. We extend the proposed approaches to geo-located ImageNet, and show that geography-aware learning can improve the performance of MoCo-v2 by $\sim$ 2% on image classification, showing the potential application of our approach to any geo-tagged dataset. Figure 1 shows our contributions in detail.

2. Related Work

Self-supervised methods use unlabeled data to learn representations that are transferable to downstream tasks (*e.g.* image classification, object detection, semantic segmentation). Two commonly seen self-supervised methods are *pre-text task* and *contrastive learning*.

Pre-text tasks Pre-text task based learning [21, 35, 28, 42, 37, 27] can be used to learn feature representations when data labels are not available. [8] rotates an image and then trains a model to predict the rotation angle. [41] trains a network to perform colorization of a grayscale image. [25] represents an image as a grid, permuting the grid and then predicting the permutation index. In this study, we use *geo-location classification* as a pre-text task, in which a deep network is trained to predict a coarse geo-location of where in the world the image might come from.

Contrastive Learning Recent self-supervised contrastive learning approaches such as MoCo [12], MoCo-v2 [2], SimCLR [1], PIRL [21], and FixMatch [30] have demonstrated superior performance and have emerged as the fore-runner on various downstream tasks. The intuition behind these methods are to learn representations by pulling positive image pairs from the same instance closer in latent space while pushing negative pairs from difference instances further away. These methods, on the other hand, differ in the type of contrastive loss, generation of positive and negative pairs, and sampling method.

Although growing rapidly in self-supervised learning area, contrastive learning methods have not been explored on large-scale remote sensing dataset. In this work, we provide a principled and effective approach for improving representation learning using MoCo-v2 [12] for remote sensing data as well geo-located conventional datasets.

Unsupervised Learning in Remote Sensing Images Unlike in traditional computer vision areas, unsupervised learning on remote sensing domain has not been studied comprehensively. Most of the studies utilize small-scale datasets specific to a small geographical region [3, 16, 29, 14, 18], a few classes [24] or a highly-specific modality, i.e. hyperspectral images [22, 40]. Most of these studies focus on the UCM-21 dataset [39] consisting of less than 1,000 images from 21 classes. A more recent study [33] proposes large-scale weakly supervised learning using a multi-modal dataset consisting of satellite images and paired geo-located wikipedia articles. While being effective, this method requires each satellite image to be paired to its corresponding article, limiting the number of images that can be used.

Geography-aware Computer Vision Geo-location data has been studied extensively in prior works. Most of these studies utilizes geo-location of an image as a prior to improve image recognition accuracy [31, 13, 23, 19, 5]. Other studies [38, 10, 11, 36] use geo-tagged training datasets to learn how to predict the geo-location of previously unseen images at test time. In our study, we leverage geo-tag information to improve unsupervised and self-supervised learning methods.

3. Problem Definition

We consider a geo-tagged visual dataset $\{((x_i^1, \cdots, x_i^{T_i}), \text{lat}_i, \text{lon}_i)\}_{i=1}^N$, where the ith datapoint consists of a sequence of images $\mathcal{X}_i = (x_i^1, \cdots, x_i^{T_i})$

"gsd":	2.10264849663	2.06074237823	1.9968634	2.2158575	1.24525177479	1.4581833	1.2518295
"img_width":	2421	2410	2498	2253	4016	3400	4003
"img_height":	2165	2156	2235	2015	3592	3041	3581
"country_code":	IND	IND	IND	IND	IND	IND	IND
"cloud_cover":	6	0	1	0	0	2	0
"timestamp":	2015-11-02 T05:44:14Z	2016-03-09 T05:25:30Z	2017-02-02 T05:47:02Z	2017-02-27 T05:24:30Z	2015-04-09 T05:36:04Z	2016-12-28 T05:57:06Z	2017-04-12 T05:51:49Z

Figure 2: Images over time concept in the fMoW dataset. The metadata associated with each image is shown underneath. We can see changes in contrast, brightness, cloud cover etc. in the images. These changes render spatially aligned images over time useful for constructing additional positives.

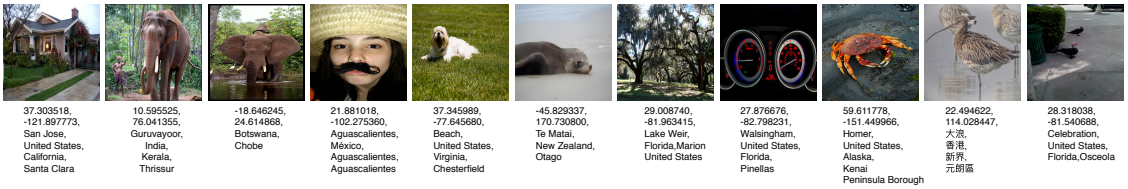

Figure 3: Some examples from GeoImageNet dataset. Below each image, we list their latitudes, longitudes, city, country name. In our study, we use the latitude and longitude information for unsupervised learning. We recommend readers to zoom-in to visualize the details of the pictures.

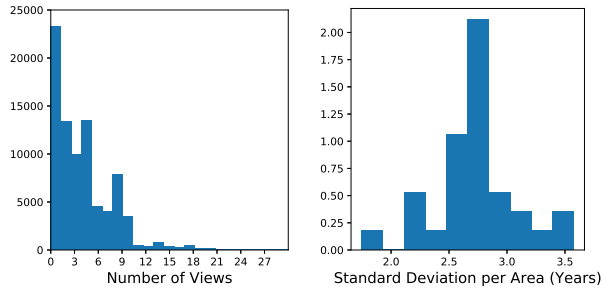

Figure 4: **Left** The histogram of number of views. **Right** the histogram of standard deviation in years per area in fMoW.

at a shared location, with latitude and longitude equal to $\text{lat}_i, \text{lon}_i$ respectively, over time $t_i = 1, ..., T_i$. When $T_i > 1$, we refer to the dataset to have temporal information or structure. Although temporal information is often not available in natural image datasets (*e.g.* ImageNet), it is common in remote sensing. While the temporal structure is similar to that of conventional videos, there are some key differences that we exploit in this work. First, we consider relatively short temporal sequences, where the time difference between two consecutive "frames" could range from months to years. Additionally unlike conventional videos we consider datasets where there is no viewpoint change across the image sequence.

Given our setup, we want to obtain visual representations $z_i^{t_i}$ of images $x_i^{t_i}$ such that the learned representation can be transferred to various downstream tasks. We do not assume access to any labels or human supervision beyond the $\text{lat}_i, \text{lon}_i$ geo-tags. The quality of the representations is measured by their performance on various downstream tasks. Our primary goal is to improve the performance of self-supervised learning by utilizing the geo-coordinates of geo-tagged datasets with remote sensing and traditional computer vision images.

3.1. Functional Map of the World

Functional Map of the World (fMoW) is a large-scale publicly available remote sensing dataset [4] consisting of approximately 363,571 training images and 53,041 test images across 62 highly granular class categories. It provides images (temporal views) from the same location over time $(x_i^1, \cdots, x_i^{T_i})$ as well as geo-location metadata $(\text{lat}_i, \text{lon}_i)$ for each image. Fig. 4 shows the histogram of the number of temporal views in fMoW dataset. We can see that most of the areas have multiple temporal views where T_i can range from 1 to 21, and on average there is about 2.5-3 years of difference between the images from an area. Also, we show examples of spatially aligned images in Fig. 2. As seen in Fig. 5, fMoW is a global dataset consisting of images from seven continents which can be ideal for learning global remote sensing representations. Such representations can be used for transfer learning on different remote sensing tasks for different regions.

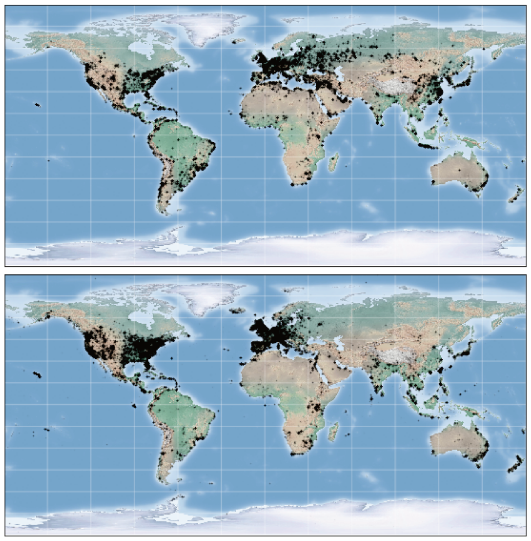

Figure 5: **Top** shows the distribution of the fMoW and **Bottom** shows the distribution of GeoImageNet.

3.2. GeoImageNet

Following [6], we extract geo-coordinates for a subset of images in ImageNet [7] using the FLICKR API. More specifically, we searched for geo-tagged images in ImageNet using the FLICKR API, and were able to find 543,435 images with their associated coordinates $(\text{lat}_i, \text{lon}_i)$ across 5150 class categories. This dataset is more challenging than ImageNet-1k as it is highly imbalanced and contains about 5× more classes. In the rest of the paper, we refer to this geo-tagged subset of ImageNet as *GeoImageNet*. Upon publication, we will release the GeoImageNet dataset publicly for the research community.

We show some examples from GeoImageNet in Fig. 3. As shown in the figure, for some images we have geo-coordinates that can be predicted from visual cues. For example, we see that a picture of a person with a Sombrero hat was captured in Mexico. Similarly, an Indian Elephant picture was captured in India, where there is a large population of Indian Elephants. Next to it, we show the picture of an African Elephant (which is larger in size). If a model is trained to predict where in the world the image was taken, it should be able to identify visual cues that are transferable to other tasks (e.g., visual cues to differentiate Indian Elephants from the African counterparts). Figure 5 shows the distribution of images in the GeoImageNet dataset.

4. Method

In this section, we briefly review contrastive loss functions for unsupervised learning and detail our proposed approach to improve Moco-v2 [2], a recent contrastive learning framework, on geo-located data.

4.1. Contrastive Learning Framework

Contrastive [12, 1, 2, 32, 26] methods attempt to learn a mapping $f_q : x_i^t \mapsto z_i^t \in \mathbb{R}^d$ from raw pixels x_i^t to semantically meaningful representations z_i^t in an unsupervised way. The training objective encourages representations corresponding to pairs of images that are known a priori to be semantically similar (positive pairs) to be closer to each other than typical unrelated pairs (negative pairs). With similarity measured by dot product, recent approaches in contrastive learning differ in the type of contrastive loss and generation of positive and negative pairs. In this work, we focus on the state-of-the-art contrastive learning framework MoCo-v2 [2], an improved version of MoCo [12], and study improved methods for the construction of positive and negative pairs tailored to remote sensing applications.

The contrastive loss function used in the MoCo-v2 framework is InfoNCE [26], which is defined as follows for a given data sample:

$$L_z = -\log \frac{\exp(z \cdot \hat{z}/\lambda)}{\exp(z \cdot \hat{z}/\lambda) + \sum_{j=1}^{N} \exp(z \cdot k_j/\lambda)}, \quad (1)$$

where z and $\hat{z}$ are query and key representations obtained by passing the two augmented views of x_i^t (denoted v and v' in Fig. 1) through query and key encoders, f_q and f_k parameterized by θ_q and θ_k respectively. Here z and $\hat{z}$ form a positive pair. The N negative samples, $\{k_j\}_{j=1}^N$, come from a dictionary of representations built as a queue. We refer readers to [12] for details on this. $\lambda \in \mathbb{R}^+$ is the temperature hyperparameter.

The key idea here is to encourage representations of positive (semantically similar) pairs to be closer, and negative (semantically unrelated) pairs to be far apart as measured by dot product. The construction of positive and negative pairs plays a crucial role in this contrastive learning framework. MoCo and MoCo-v2 both use perturbations (also called "data augmentation") from the same image to create a positive example and perturbations from different images to create a negative example. Commonly used perturbations include random color jittering, random horizontal flip, and random grayscale conversion.

Temporal Positive Pairs Different from many commonly seen natural image datasets, remote sensing datasets often have extra temporal information, meaning that for a given location $(\text{lat}_i, \text{lon}_i)$, there exists a sequence of spatially aligned images $\mathcal{X}_i = (x_i^1, \cdots, x_i^{T_i})$ over time. Unlike in traditional videos where nearby frames could experience large changes in content (*e.g.* from a cat to a tree), in remote sensing the content is often more stable across time due to the fixed viewpoint. For instance, a place on ocean is likely to remain as ocean for months or years, in which case satellite images taken across time at the same location should share high semantic similarities. Even for locations

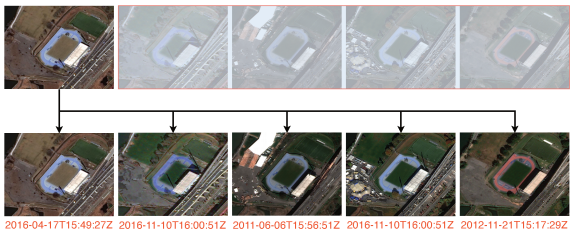

Figure 6: Demonstration of temporal positives in eq. 2. An image from an area is paired to the other images including itself from the same area captured at different time. We show the time stamps for each image underneath the images. We can see the color changes in the stadium seatings and surrounding areas.

where non-trivial changes do occur over time, certain semantic similarities could still remain. For instance, key features of a construction site are likely to remain the same even as the appearance changes due to seasonality.

Given these observations, it is natural to leverage temporal information for remote sensing while constructing positive or negative pairs since it can provide us with extra semantically meaningful information of a place over time. More specifically, given an image $x_i^{t_1}$ collected at time t_1, we can randomly select another image $x_i^{t_2}$ that is spatially aligned with $x_i^{t_1}$ (*i.e.* $x_i^{t_2} \in \mathcal{X}_i$). We then apply perturbations (*e.g.* random color jittering) as used in MoCo-v2 to the spatially aligned image pair $x_i^{t_1}$ and $x_i^{t_2}$, providing us with a *temporal positive pair* (denoted v and v' in Figure 1) that can be used for training the contrastive learning framework by passing them through query and key encoders, f_q and f_k respectively (see Fig. 1). Note that when $t_1 = t_2$, the *temporal positive pair* is the same as the positive pair used in MoCo-v2.

Given a data sample $x_i^{t_1}$, our TemporalInfoNCE objective function can be formulated as follows:

$$L_{z_i^{t_1}} = -\log \frac{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda)}{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda) + \sum_{j=1}^{N} \exp(z_i^{t_1} \cdot k_j/\lambda)}, \quad (2)$$

where $z_i^{t_1}$ and $z_i^{t_2}$ are the encoded representations of the randomly perturbed temporal positive pair $x_i^{t_1}, x_i^{t_2}$. Similar as equation 1, N is number of negative samples, $\{k_j\}_{j=1}^{N}$ are the encoded negative pairs and $\lambda \in \mathbb{R}^+$ is the temperature hyperparameter. Again, we refer readers to [12] for details on construction of these negative pairs.

Note that compared to equation 1, we use two *real* images from the same area over time to create positive pairs. We believe that relying on real images for positive pairs encourages the network to learn better representations for real data than the one relying on synthetic images. On the other hand, our objective in equation 2 enforces the representations to be invariant to changes over time. Depending on the target task, such inductive bias can be desirable or undesirable. For example, for a change detection task, learning representations that are highly sensitive to temporal changes can be more preferable. However, for image classification or object detection task, learning temporally invariant features should not degrade the downstream performance.

4.2. Geo-location Classification as a Pre-text Task

In this section, we explore another aspect of remote sensing images, *geolocation metadata*, to further improve the quality of the learned representations. In this direction, we design a pre-text task for unsupervised learning. In our pre-text task, we cluster the images in the dataset using their coordinates $(\text{lat}_i, \text{lon}_i)$. We use a clustering method to construct K clusters and assign an area with coordinates $(\text{lat}_i, \text{lon}_i)$ a categorical geo-label $c_i \in \mathcal{C} = \{1, \cdots, K\}$. Using the cross entropy loss function, we then optimize a geo-location predictor network f_c as

$$L_g = \sum_{k=1}^{K} -p(c_i = k) \log(\hat{p}(c_i = k | f_c(x_i^t)), \quad (3)$$

where p represent the probability of the cluster k representing the true cluster and $\hat{p}$ represents the predicted probabilities for K clusters. In our experiments, we represent f_c with a CNN parameterized by θ_c. With this objective, our goal is to learn location-aware representations that can potentially transfer well to different downstream tasks.

4.3. Combining Geo-location and Contrastive Learning Losses

In the previous section, we designed a pre-text task leveraging the geo-location meta-data of the images to learn location-aware representations in a standalone setting. In this section, we combine geo-location prediction and contrastive learning tasks in a single objective to improve the contrastive learning-only and geo-location learning-only tasks. In this direction, we first integrate the geo-location learning task into the contrastive learning framework using the cross-entropy loss function where the geo-location prediction network uses features z_i^t from the query encoder as

$$L_g = -\sum_{i=1}^{K} p(c_i = k) \log(\hat{p}(c_i = k | f_c(z_i^t)). \quad (4)$$

In the combined framework (see Fig. 1), f_c is represented by a linear layer parameterized by θ_c. Finally, we propose the objective for joint learning as the linear combination of TemporalInfoNCE and geo-classification loss with coefficients α and β representing the importance of contrastive learning and geo-location learning losses as

$$\arg\min_{\theta_q, \theta_k, \theta_c} L_f = \alpha L_{z^{t_1}} + \beta L_g. \quad (5)$$

By combining two tasks, we learn representations to jointly maximize agreement between spatio-temporal positive pairs, minimize agreement between negative pairs and predict the geo-label of the images from the positive pairs.

5. Experiments

In this study, we perform unsupervised representation learning on fMoW and GeoImageNet datasets. We then evaluate the learned representations/pre-trained models on a variety of downstream tasks including image recognition, object detection and semantic segmentation benchmarks on remote sensing and conventional images.

Implementation Details for Unsupervised Learning *For contrastive learning*, similar to MoCo-v2 [2], we use ResNet-50 to paramaterize the query and key encoders, f_q and f_k, in all experiments. We use following hyper-parameters in the MoCo-v2 pre-training step: learning rate of 1e-3, batch size of 256, dictionary queue of size 65536, temperature scaling of 0.2 and SGD optimizer. We use similar setups for both fMoW and GeoImageNet datasets. Finally, for each downstream experiment, we report results for the representations learned after 200 epochs.

For geo-location classification task, we run k-Means clustering algorithm to cluster fMoW and GeoImageNet into $K = 100$ geo-clusters given their latitude and longitude pairs. We show the clusters in Fig. 7. As seen in the figure, while both datasets have similar clusters there are some differences, particularly in North America and Europe. In Fig. 8 we analyze the clusters in GeoImageNet and fMoW. The figure shows that the number of clusters per class on GeoImageNet tend to be skewed towards smaller numbers than fMoW whereas the number of unique classes per cluster on GeoImageNet has more of a uniform distribution. For fMoW, we can conclude that each cluster contain samples from most of the classes. Finally, when adding the geo-location classification task into the contrastive learning we tune α and β to be 1.

Methods We compare our unsupervised learning approach to **supervised learning** for image recognition task. For object detection, and semantic segmentation we compare them to pre-trained weights obtained using (a) **supervised learning**, and (b) **random initilization** while fine-tuning on the target task dataset. Finally, for ablation analysis we provide results using different combinations of our methods. When appending only geo-location classification task or temporal positives into **MoCo-v2** we use **MoCo-v2+Geo** and **MoCo-v2+TP**. When adding both of our approaches into **MoCo-v2** we use **MoCo-v2+Geo+TP**.

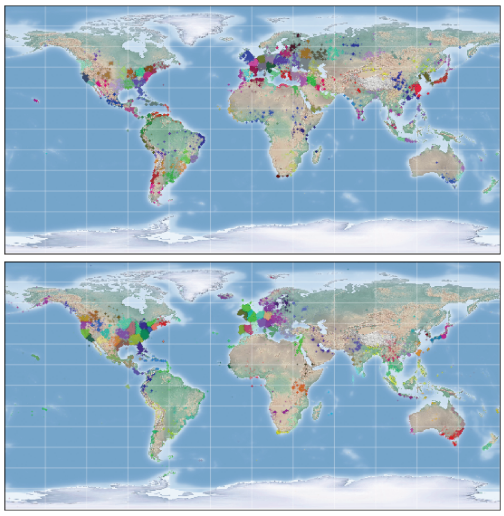

Figure 7: **Top** and **Bottom** show the distributions of the fMoW and GeoImageNet clusters.

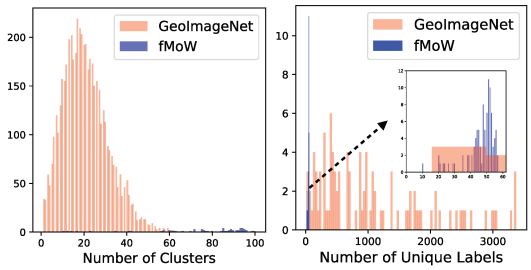

Figure 8: **Left** shows the number of clusters per label and **Right** shows the number of unique labels per cluster in fMoW and GeoImageNet. Labels represent the original classes in fMoW and GeoImageNet.

5.1. Experiments on fMoW

We first perform experiments on fMoW image recognition task. Similar to the common protocol of evaluating unsupervised pre-training methods [2, 12], we freeze the features and train a supervised linear classifier. However, in practice, it is more common to finetune the features end-to-end in a downstream task. For completeness and a better comparison, we report end-to-end finetuning results for the 62-class fMoW classification as well. We report both top-1 accuracy and F1-scores for this task.

Classifying Single Images In Table 1, we report the results on single image classification on fMoW. We would like to highlight that in this case we classify each image individually. In other words, we do not use the prior information that multiple images over the same area $(x_i^1, x_i^2, \ldots, x_i^{T_i})$ have the same labels $(y_i, y_i, \ldots, y_i)$. For evaluation, we use 53,041 images. Our results on this task (linear classification on frozen features) show that MoCo-v2 performs reasonably well on a large-scale dataset with 60.69% accuracy, 8% less than the supervised learning methods. *Sup. Learning (IN wts. init.)* and *Sup. Learning (Scratch)* correspond to supervised learning method starting from ima-

	Backbone	F1-Score ↑ (Frozen/Finetune)	Accuracy ↑ (Frozen/Finetune)
Sup. Learning (IN wts. init.)*	ResNet50	-/64.72	-/69.07
Sup. Learning (Scratch)*	ResNet50	-/64.71	-/69.05
Geoloc. Learning*	ResNet50	48.96/52.23	52.40/56.59
MoCo-V2 (pre. on IN)	ResNet50	31.55/57.36	37.05/62.90
MoCo-V2	ResNet50	55.47/60.61	60.69/64.34
MoCo-V2+Geo	ResNet50	61.60/66.60	64.07/69.04
MoCo-V2+TP	ResNet50	**64.53/67.34**	**68.32/71.55**
MoCo-V2+Geo+TP	ResNet50	63.13/66.56	66.33/70.60

Table 1: Experiments on fMoW on classifying single images. * indicates a model trained up to epoch with the highest accuracy on the validation set. We use the same set up for Sup. Learning and Geoloc. Learning in the remaining experiments. **Frozen** corresponds to linear classification on frozen features. **Finetune** corresponds to end-to-end finetuning results for the fmow classification.

genet pre-trained weights and random weights respectively. This result aligns with MoCo-v2's performance on the ImageNet dataset [2]. Next, by incorporating geo-location classification task into MoCo-v2, we improve by 3.38% in top-1 classification accuracy. We further improve the results to 68.32% using temporal positives, bridging the gap between the MoCo-v2 baseline and supervised learning to less than 1%. However, when we perform end-to-end finetuning for the classification task, we observe that our method surpasses the supervised learning methods by more than 2%. For completeness, we also include results for MoCo-v2 pre-trained on Imagenet dataset (4th row in Table 1) and find that the distribution shift between Imagenet and downstream dataset leads to suboptimal performance.

Classifying Temporal Data In the next step, we change how we perform testing across multiple images over an area at different times. In this case, we predict labels from images over an area i.e. make a prediction for each $t \in \{1, \ldots, T_i\}$, and average the predictions from that area. We then use the most confident class prediction to get area-specific class predictions. In this case, we evaluate the performance on 11,231 unique areas that are represented by multiple images at different times. Our results in Table 2 show that doing area-specific inference improves the classification accuracies by 4-8% over image-specific inference. Even incorporating temporal positives, we can improve the accuracy by 6.1% by switching from image classification to temporal data classification. Overall, our methods outperform the baseline Moco-v2 by 4-6% and supervised learning by 1-2%. Here we only report temporal classification on top of frozen features.

5.2. Transfer Learning Experiments

Previously, we performed pre-training experiments on fMoW dataset and quantified the quality of the representations by supervised training a linear layer for image recognition on fMoW. In this section, we perform transfer learning experiments on different low level tasks.

	Backbone	F1-Score ↑	Accuracy ↑
Sup. Learning (IN wts. init.)*	ResNet50	68.72 (+4.01)	73.22 (+4.15)
Sup. Learning (Scratch)*	ResNet50	68.73 (+4.02)	73.24 (+4.19)
Geoloc. Learning*	ResNet50	52.01 (+3.05)	56.12 (+3.72)
MoCo-V2 (pre. on IN)	ResNet50	35.93 (+4.38)	42.56 (+5.51)
MoCo-V2	ResNet50	63.96 (+8.49)	68.64 (+7.95)
MoCo-V2+Geo	ResNet50	66.93 (+5.33)	70.48 (+6.41)
MoCo-V2+TP	ResNet50	**70.11 (+5.58)**	**74.42 (+6.10)**
MoCo-V2+Geo+TP	ResNet50	69.56 (+6.43)	72.76 (+6.43)

Table 2: Experiments on fMoW on classifying temporal data. In the table, we compare the results to the ones on single image classification. Here we present results corresponding to linear classification on frozen features only.

pre-train	AP_{50} ↑
Random Init.	10.75
Sup. Learning (IN wts. init.)	14.44
Sup. Learning (Scratch)	14.42
MoCo-V2	15.45 (+4.70)
MoCo-V2-Geo	15.63 (+4.88)
MoCo-V2-TP	17.65 (+6.90)
MoCo-V2-Geo+TP	**17.74 (+6.99)**

Table 3: Object detection results on the xView dataset.

5.2.1 Object Detection

For object detection, we use the xView dataset [15] consisting of high resolution satellite images captured with similar sensors to the ones in the fMoW dataset. The xView dataset consists of 846 very large (∼2000×2000 pixels) satellite images with bounding box annotations for 60 different class categories including airplane, passenger vehicle, maritime vessel, helicopter etc.

Implementation Details We first divide the set of large images into 700 training and 146 test images. Then, we process the large images to create 416×416 pixels images by randomly sampling the bounding box coordinates of the small image and we repeat this process 100 times for each large image. In this process, we ensure that there is less than 25% overlap between any two bounding boxes from the same image. We then use RetinaNet [17] with pre-trained ResNet-50 backbone and fine-tune the full network on the xView training set. To train RetinaNet, we use learning rate of 1e-5 and a batch size of 4 and Adam optimizer.

Qualitative Analyis Table 3 shows the object detection results on the xView. We achieve the best results with the addition of temporal positive pair, and geo-location classification pre-text task into MoCo-v2. With our final model, we can outperform the randomly initialized weights by 7% AP and the supervised learning on the fMoW by 3.3% AP.

5.2.2 Image Segmentation

In this section, we perform downstream experiments on the task of Semantic Segmentation on SpaceNet dataset [34]. The SpaceNet datasets consists of 5000 high resolution satellite images with segmentation masks for buildings.

Implementation Details We divide our SpaceNet dataset into training and test sets of 4000 and 1000 images respectively. We use PSAnet [43] network with ResNet-50 backbone to perform semantic segmentation. We train PSAnet network with a batch size of 16 and a learning rate of 0.01 for 100 epochs and use SGD optimizer.

Qualitative Analysis Table 4 shows the segmentation performance of differently initialized backbone weights on the SpaceNet test set. Similar to object detection, we achieve the best IoU scores with the addition of temporal positives and geo-location classification task. Our final model outperforms the randomly initialized weights and supervised learning by 3.58% and 2.94% IoU scores. We observe that the gap between the best and worst models shrinks going from the image recognition to object detection, and semantic segmentation task. This aligns with the performance of the MoCo-v2 pre-trained on ImageNet and fine-tuned on the Pascal-VOC object detection and semantic segmentation experiments [12, 2].

pre-train	mIOU ↑
Random Init.	74.93
Imagenet Init.	75.23
Sup. Learning (IN wts. init.)	75.61
Sup. Learning (Scratch)	75.57
MoCo-V2	78.05 (+3.12)
MoCo-V2-Geo	78.42 (+3.49)
MoCo-V2-TP	78.48 (+3.55)
MoCo-V2-Geo+TP	**78.51 (+3.58)**

Table 4: Semantic segmentation results on Space-Net.

pre-train	Top-1 Accuracy ↑
Random Init.	51.89
Imagenet Init.	53.46
Sup. Learning (IN wts. init.)	54.67
Sup. Learning (Scratch)	54.46
MoCo-V2	55.18 (+3.29)
MoCo-V2-Geo	**58.23 (+6.34)**
MoCo-V2-TP	57.10 (+5.21)
MoCo-V2-Geo+TP	57.63 (+5.74)

Table 5: Land Cover Classification on NAIP dataset.

5.2.3 Land Cover Classification

Finally, we perform transfer learning experiments on land cover classification across 66 land cover classes using high resolution remote sensing images obtained by the USDA's National Agricultural Imagery Program (NAIP). We use the images from the California's Central Valley for the year of 2016. Our final dataset consists of 100,000 training and 50,000 test images. Table 5 shows that our method outperforms the randomly initialized weights by 6.34% and supervised learning by 3.77%.

5.3. Experiments on GeoImageNet

After fMoW, we adopt our methods for unsupervised learning on fMoW for improving representation learning on the GeoImageNet. Unfortunately, since ImageNet does not contain images from the same area over time we are not able to integrate the temporal positive pairs into the MoCo-v2 objective. However, in our GeoImageNet experiments we show that we can improve MoCo-v2 by introducing geo-location classification pre-text task.

Qualitative Analysis Table 6 shows the top-1 and top-5 classification accuracy scores on the test set of GeoImageNet. Surprisingly, with only geo-location classification task we can achieve 22.26% top-1 accuracy. With MoCo-v2 baseline, we get 38.51 accuracy, about 3.47% more than supervised learning method. With the addition of geo-location classification, we can further improve the top-1 accuracy by 1.45%. These results are interesting in a way that MoCo-v2 (200 epochs) on ImageNet-1k performs 8% worse than supervised learning whereas it outperforms supervised learning on our *highly imbalanced* GeoImageNet with 5150 class categories which is about $5\times$ more than ImageNet-1k.

	Backbone	**Top-1 (Accuracy)** ↑	**Top-5 (Accuracy)** ↑
Sup. Learning (Scratch)	ResNet50	35.04	54.11
Geoloc. Learning	ResNet50	22.26	39.33
MoCo-V2	ResNet50	38.51	57.67
MoCo-V2+Geo	ResNet50	**39.96**	**58.71**

Table 6: Experiments on GeoImageNet. We divide the dataset into 443,435 training and 100,000 test images across 5150 classes. We train MoCo-V2 and MoCo-V2+Geo for 200 epochs whereas **Sup. and Geoloc. Learning are trained until they converge**.

6. Conclusion

In this work, we provide a self-supervised learning framework for remote sensing data, where unlabeled data is often plentiful but labeled data is scarce. By leveraging spatially aligned images over time to construct temporal positive pairs in contrastive learning and geo-location in the design of pre-text tasks, we are able to close the gap between self-supervised and supervised learning on image classification, object detection and semantic segmentation on remote sensing and other geo-tagged image datasets.

Acknowledgement

This research was supported by Stanford Data for Development Initiative, HAI, IARPA SMART, ONR (N00014-19-1-2145), and NSF grants #1651565 and #1733686.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 4

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 4, 6, 7, 8

[3] Anil M Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451, 2013. 2

[4] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 1, 3

[5] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[6] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 2, 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[9] 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. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[10] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1, 2

[11] James Hays and Alexei A Efros. Large-scale image geolocalization. In *Multimodal location estimation of videos and images*, pages 41–62. Springer, 2015. 1, 2

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 4, 5, 6, 8

[13] Hayate Iso, Shoko Wakamiya, and Eiji Aramaki. Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750*, 2017. 2

[14] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019. 2

[15] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018. 1, 7

[16] Yansheng Li, Chao Tao, Yihua Tan, Ke Shang, and Jinwen Tian. Unsupervised multilayer feature learning for satellite image scene classification. *IEEE Geoscience and Remote Sensing Letters*, 13(2):157–161, 2016. 2

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[18] Xiaoqiang Lu, Xiangtao Zheng, and Yuan Yuan. Remote sensing scene classification by unsupervised representation learning. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5148–5157, 2017. 2

[19] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9596–9606, 2019. 2

[20] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1

[21] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[22] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):391–406, 2017. 2

[23] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 2

[24] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016. 2

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[27] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 2701–2710, 2017. 2

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[29] Adriana Romero, Carlo Gatta, and Gustau Camps-Valls. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2015. 2

[30] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2

[31] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pages 1008–1016, 2015. 2

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 4

[33] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David B Lobell, and Stefano Ermon. Learning to interpret satellite images using wikipedia. In *IJCAI*, pages 3620–3626, 2019. 1, 2

[34] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 8

[35] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[36] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2621–2630, 2017. 2

[37] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 2

[38] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016. 2

[39] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 2

[40] Lefei Zhang, Liangpei Zhang, Bo Du, Jane You, and Dacheng Tao. Hyperspectral image unsupervised classification by robust manifold matrix factorization. *Information Sciences*, 485:154–169, 2019. 2

[41] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 2

[43] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 8