Extending global-local view alignment for self-supervised learning with remote sensing imagery

Xinye Wanyan
x.wanyan@unimelb.edu.au

Sachith Seneviratne
sachith.seneviratne@unimelb.edu.au

Shuchang Shen
chuchangs@student.unimelb.edu.au

Michael Kirley
mkirley@unimelb.edu.au

Abstract

Since large number of high-quality remote sensing images are readily accessible, exploiting the corpus of images with less manual annotation draws increasing attention. Self-supervised models acquire general feature representations by formulating a pretext task that generates pseudo-labels for massive unlabeled data to provide supervision for training. While prior studies have explored multiple self-supervised learning techniques in remote sensing domain, pretext tasks based on local-global view alignment remain underexplored, despite achieving state-of-the-art results on natural imagery. Inspired by DINO [6], which employs an effective representation learning structure with knowledge distillation based on global-local view alignment, we formulate two pretext tasks for self-supervised learning on remote sensing imagery (SSLRS). Using these tasks, we explore the effectiveness of positive temporal contrast as well as multi-sized views on SSLRS. We extend DINO and propose DINO-MC which uses local views of various sized crops instead of a single fixed size in order to alleviate the limited variation in object size observed in remote sensing imagery. Our experiments demonstrate that even when pre-trained on only 10% of the dataset, DINO-MC performs on par or better than existing state-of-the-art SSLRS methods on multiple remote sensing tasks, while using less computational resources. All codes, models, and results are released at https://github.com/WennyXY/DINO-MC.

1. Introduction

Computer vision models provide promising solutions for remote sensing image analysis to be applied in numerous real-world tasks, including disaster prevention [46], forestry [25], agriculture [29], land surface change [18], biodiversity [45]. However, training large-scale deep learning models in a traditional supervised manner requires extensive labeled datasets. Labeling large datasets is costly and error-prone, particularly within the remote sensing domain, which requires expert knowledge [38]. Therefore, reducing the reliance of the model on labeled images is crucial for resolving specific downstream tasks. Self-supervised learning (SSL) paradigm is a common solution to learn useful features from copious data without labels, which comprises two distinct phases: the initial pretext task pre-training phase followed by the subsequent downstream task fine-tuning phase. Since different pretext tasks prompt the model to acquire diverse feature aspects, devising an appropriate task for which labels can be automatically derived from the data is essential for fostering generalized representations.

Based on the pretext tasks, SSL models can be categorized into three distinct categories. Generative tasks allow the model learn feature representations by reconstructing or generating original images. For example, the image inpainting [33] uses the original data as labels to train the model to recover several masked parts of the original image. The discriminative tasks [7, 14] train the network to distinguish certain characteristics or properties. For example, [13] trains the model to predict the relative position of a patch to its neighbors, and each image patch is defined to have up to eight adjacent patches. However, the feature representations learned by generative and discriminative methods highly rely on the designed pretext tasks, and an ineffective pretext task might reduce the transfer-ability of the pre-trained model [49]. In stead of solving a straightforward and specific pretext task, contrastive SSL approaches extract features by maximising the similarity between the latent representations of two positive instances (e.g., two augmented views of the same image) and the distance between two negative instances. However, simply following this approach will easily lead to an identity map for each pair of positive samples, i.e., model collapse [49].

DINO [6] utilizes knowledge distillation and centering and sharpening of the teacher network [23] to handle this issue. DINO exhibits impressive performance in numerous

computer vision tasks, including image retrieval, copy detection, and video instance segmentation. Although some previous work has applied DINO to remote sensing domain [36, 48, 50], a comprehensive evaluation and extension of the self-supervised objective for remote sensing imagery deserves further exploration. In particular, the size of instances observed during pre-training on traditional natural scenes shows wider variation than seen in remote sensing imagery. Thus, the alignment of the latent representation of the multi-sized local augmented crops against the global augmented views is of interest in remote sensing imagery as it leads to a more challenging pretext task. Remote sensing data holds spatial-temporal heterogeneity offering rich time series information due to the repetitive capture of imagery by satellites over the same geographic area. Some studies [3, 27, 51] have explored employing the temporal information for SSL and their results indicate promising potential. [3] integrated temporal information of video sequences into instance discrimination based contrastive SSL. [27, 51] constructed large-scale SSLRS datasets involving a wide spatio-temporal range of remote sensing imagery.

Inspired by the inherent characteristics of the size variation of semantic content observed between traditional natural scenes and remote sensing imagery, we propose DINO-MC which uses size variation in local crops to drive better representation learning of the semantic content in remote sensing imagery. Additionally, we explore applying temporal views as positive instances into contrastive SSL method. We evaluate our model with different backbones including two transformer-based models and two convnets. Although the majority of current SSL methods for remote sensing imagery rely on ResNet and Vision Transformers (ViTs) as backbone architectures, there is growing interest in exploring the self-supervised feature extraction capabilities of Wide ResNets (WRN) [53] and Swin Transformers [26] in the remote sensing domain. We incorporate these two architectures into our analysis. In the linear probing evaluation, the findings demonstrate that DINO-MC has great transfer-ability which achieves 2.56% higher accuracy with a smaller pre-trained dataset than SeCo. When fine-tuned on two remote sensing classification tasks and change detection task, DINO-MC outperforms DINO and SeCo.

Our main contributions are summarized as follows:

- We apply temporal views as positive instances to recent contrastive self-supervised models (DINO-TP). We analyze different backbone networks to explore their effectiveness on different remote sensing tasks when pre-trained under this setting.
- We combine a new multi-sized local cropping strategy with DINO and propose DINO-MC. We pre-train DINO-MC with different backbones on a satellite imagery dataset SeCo-100K to learn general representations.
- DINO-MC outperforms SeCo with only 10% pre-training dataset, and achieves state-of-the-art results on BigEarthNet multi-label and EuroSAT multi-class land use classification, as well as Onera Satellite Change Detection (OSCD) task.

2. Related Work

2.1. Contrastive Self-supervised Learning.

SSL methods create task-agnostic signals to instruct the model in learning versatile features which can effectively generalise across various downstream tasks [42]. In generative and discriminative SSL, the loss function is computed based on the disparity between the ground truth and the model's prediction. Conversely, in contrastive models, the loss is determined by the dissimilarities among feature representations within the latent space [49]. Contrastive SSL exhibits promise in transferability and gains increasing attention. Furthermore, the statistics of recent studies shows that nearly half of the remote sensing self-supervised studies applied contrastive learning models [49].

Instance discrimination [14] proves to be a highly effective pretext task commonly used in contrastive learning. In this task, augmented views originating from the same input are labeled as positive, whereas distinct instances are designated as negative. The model is then trained to maintain proximity between positive pairs and create separation between negative pairs in the representation space [36, 44, 52]. Several studies have explored positive/negative samples generating strategies. [31] regards the diagonally connected patches as the positive instances otherwise negative, and [43] incorporates same spot views from different sensory as the positive instances. [21] proposes MoCo-V1 with a momentum encoder to effectively maintain multiple negative samples for representation learning. SimCLR [7] employs very large batch sizes instead of momentum encoders, and provides experimental results on 10 forms of data augmentation. Inspired by SimCLR, MoCo-V1 is extended to MoCo-V2 [9], and MoCo-V3 [10] integrates ViT as the backbone architecture.

Clustering method is another line of contrastive learning, which demands no precise indication from the inputs [4]. [5] introduces a contrasting cluster assignment that learns features by Swapping Assignments between multiple Views of the same image (SwAV). SwAV can be readily applied to different sizes of datasets due to its utilization of clustering rather than pairwise comparisons. [5] proposes a data augmentation named multi-crop to increase the instances without drastically additional memory and computation. Specifically, when doing multi-crop, images are cropped to a collection of views with lower resolutions instead of the full-resolution views. [1] maps the feature representations of two randomly augmented views (with one view having randomly masked patches) to a shared cluster

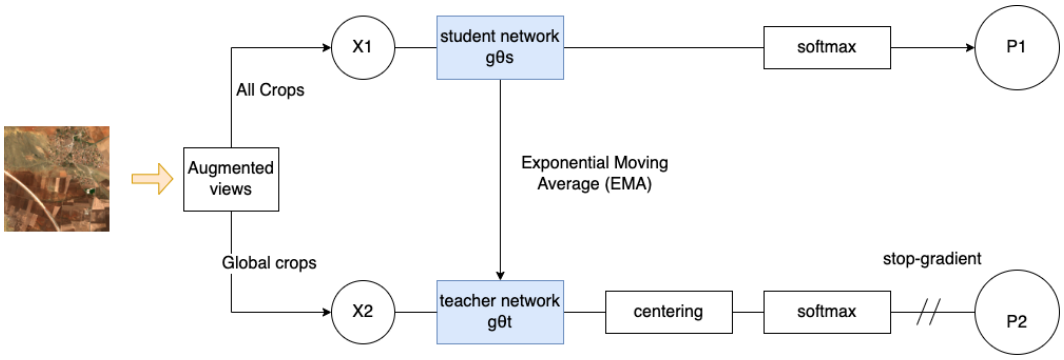

Figure 1. *DINO*: the self-supervised contrastive algorithm with knowledge distillation. It is the basic structure of both DINO-TP and DINO-MC. For DINO-TP, we use three temporal views to generate global crops and multi-sizes local crops as the input to do positive contrastive representation learning. For DINO-MC, we generate global and local crops from one imagery, then apply two different augmentations to global views and multi-sizes local views, respectively, to get the input of teacher and student network.

space to maximize their similarity.

Knowledge distillation in contrastive learning can be employed for feature extraction without the need to differentiate between images [6]. [19] proposes Bootstrap Your Own Latent (BYOL) based on the online and target networks whose weights are updated with each other. Inspired by BYOL, [6] explores the further synergy between SSL and different backbones, especially ViTs, and proposes a simple form of self-distillation with no labels (DINO). DINO uses same networks architecture as BYOL but with different loss function and backbones. When pre-trained on ImageNet [35], DINO achieves better linear and KNN probing evaluation results than other self-supervised methods with fewer computation resources [5, 6, 19]. [37] suggests incorporating two distillation modes within SSL: self-distillation and cross-distillation, to improve the semantic feature alignment across two networks. [24] enables the intermediate layers to learn from the final layer by employing contrastive loss through, i.e., self-distillation.

2.2. Self-supervised Learning in Remote Sensing.

[41] experiments with SSL models on different pretext tasks, including image inpainting [33], context prediction [13], and instance discrimination [52] on remote sensing imagery tasks. [47] proposes to learn practical representations from satellite imagery by reconstructing the visible colors (RGB) from its high-dimensionality spectral bands (Spectral). [2] explores the application of contrastive SSL to large-scale satellite image datasets. [27] leverages seasonal information of remote sensing imagery and demonstrates that the performance of MoCo-V2 is enhanced with the incorporation of temporal positives (TP). TP regards the temporal views as the positive instances and trains models match their representations allowing the model to learn essential features that do not change over time. Besides, they introduce a negative temporal contrastive model named SeCo designed to capture variations or discrepancies resulting from temporal changes. [48] uses Swin Transformer as the backbone of DINO and applies it to remote sensing imagery tasks. DINO-MM [50] extends DINO by combining synthetic-aperture radar (SAR) and multispectral (optical) images and is applied to BigEarthNet land use classification task.

Current studies have demonstrated the utility and viability of SSL models for practical applications in tasks involving remote sensing images. But the full potential of SSL in remote sensing warrant further investigation. Our work targets to bridge this gap and extend existing SSL model by generating more effective contrastive instances.

3. Methodology

Our work is mainly based on a contrastive SSL model named DINO, which has been applied into both natural and remote sensing imagery tasks [6, 50, 51]. We aim to learn useful, transferable features for remote sensing imagery by exploring positive temporal contrastive SSL model DINO-TP (Sec. 3.1) and contrastive SSL model with multi-sized crops DINO-MC (Sec. 3.2). In addition, we experiment with the feature extraction ability of different backbones in SSL.

3.1. DINO-TP

Architecture Fig. 1 shows the model structure. The student and teacher networks in DINO have same architecture g but different weights θ_t and θ_s. Two sets of augmented views x_1 and x_2 are generated from the same image. The global view covers the majority of the initial image and the local view only covers a small part. The student network receives both global and local crops as inputs, whereas the teacher network only receives global crops. The model is trained to match the individual local views to global views in the feature space.

Rather than a pre-trained and frozen teacher network [8, 17], DINO dynamically updates the teacher weights by using EMA on student weights, i.e.,

$$\theta_t \leftarrow \theta_t \lambda + (1 - \theta_s) \tag{1}$$

The λ values adhere to a cosine schedule between 0.996 and 1, indicating that the teacher network is less dependent on the present student and more dependent on the integration of the student network in each round; hence, the weights are updated slowly.

Centering and sharpening are used to prevent model collapse. Centering is adding a bias term c to the features of the output of the teacher model, i.e.,

$$g_t(x) = g_t(x) + c \tag{2}$$

The bias term c is dynamically updated by the EMA, where $m > 0$ denotes the update rate and B is the batch size.

$$c \leftarrow mc + (1 - m)\frac{1}{B}\sum_{i=1}^{B} g\theta_t(x_i) \tag{3}$$

Sharpening is achieved by softmax normalization using low temperature in the teacher network to avoid consistent distribution. The teacher network consistently outperforms the student network during pre-training, which is used as the feature extractor in downstream tasks [6].

Temporal Contrastive This work performs self-supervised pre-training on SeCo-100K [27] to integrate temporal information. SeCo-100K is an unlabeled remote sensing imagery dataset collected from Sentinel-2 [16] with 100K instances, and each instance is composed of five different temporal views of the same location. DINO-TP regards the temporal views of the same location as the positive instances, i.e., the teacher and student networks are trained to match temporal views in the feature space. In the pre-training phase, we randomly select three temporal views and name them as t_0, t_1, t_2. As shown in Fig. 2, t_1, t_2 can be regarded as the temporal augmentation views of t_0. The temporal image t_0 is directly used as q to generate 6 local crops of various sizes, including $18^2, 16^2, 14^2, 12^2, 10^2$ and 8^2. After applying color jittering, grayscale, and Gaussian blur to each of the three temporal perspectives, we acquire k_0, k_1, k_2 views. k_0, k_1, k_2 are scaled and resized into 22^2 serving as three global crops utilized as the inputs. In prior work, different techniques are paired with different crop scaling ranges. DINO-TP not only learns the relationship between the whole and the pieces of the image, but also discovers the interrelation among various temporal perspectives.

3.2. DINO-MC

The structure of DINO-MC is nearly identical to DINO-TP. This work changes the cropping strategy for local crops in DINO to create a more challenging pretext task. We use multi-sized crops instead of the fixed-size local crops used in DINO. Keeping the number of global and local crops same as DINO, DINO-MC outperforms DINO on both linear and KNN probing as well as end-to-end evaluation on multiple downstream tasks. Color-related pretext tasks have been proven to be effective in multiple image tasks in the field of remote sensing [1, 47, 54]. Due to the strong connection between color and semantics, we set two strategies of color transformations for global and local crops respectively. The global views are augmented by random color jittering and GaussianBlur. The local views are augmented by random color jittering, random grayscale shifting, and random Gaussian blur all together. Random color jittering modifies the brightness, contrast, saturation and hue of the image with a specified probability within a certain range. This process aims to replicate the effects of capturing images in diverse lighting conditions and real-world shooting scenarios. Random grayscale converts an image into a grayscale image randomly, aiming to diminish the influence of color while capturing additional image characteristics beyond color properties. We use different settings for cropping and color transformation. The experiments verify that our strategy is effective and DINO-MC outperforms DINO on both linear probing and end-to-end fine-tuning on three downstream tasks (classification and change detection).

4. Experiments

In this study, the features learned from the self-supervised pre-training are evaluated on three downstream tasks: EuroSAT [22] and BigEarthNet [39] land use classification tasks, and OSCD change detection task [12]. Due to its flexibility and computational efficiency, we utilize DINO as the foundation for our experiments and further extend its capabilities. SeCo-100K [27] serves as the dataset for self-supervised representation learning in this experiment, containing 100K unlabeled remote sensing images. During the pre-training phase, the student and teacher networks receive two separate sets of temporal views depicting the same geographical region. The SSL model is trained to aligns these views within the feature space to generate representations that exhibit temporal consistency. In DINO-MC and DINO-TP, the scaling range of multi-crop is $(0.05, 0.32)$ for local crops and $(0.32, 1)$ for global crops.

4.1. DINO with Different Backbones

When employed as the backbone of DINO, ViT demonstrates superiority over ResNet [6]. In this work, there are four different networks applied as the backbone of DINO, DINO-TP, and DINO-MC, including ViT, Swin Transformer, ResNet, and WRN.

ViT preserves the Transformer structure used in NLP as much as possible and exhibits promising performance in

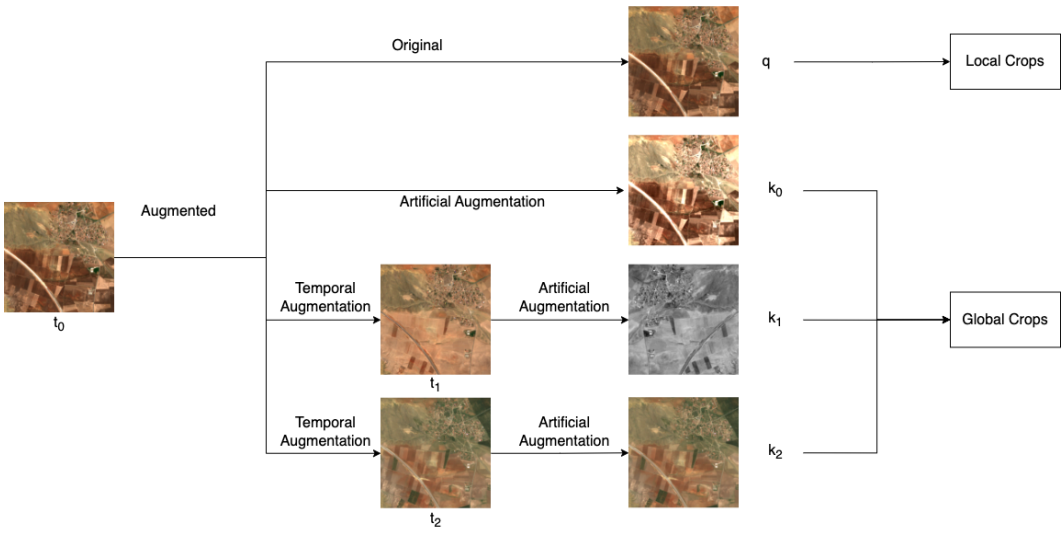

Figure 2. The process of handling temporal views of contrastive learning in DINO-TP. We randomly select three temporal views of the same location and augment them to obtain global and local crops, which is the input of the teacher and student network. Different temporal views of the same location in DINO-TP are considered as positive examples, and we train the model to match their representations in the feature space.

computer vision tasks [15]. The input of transformer encoder is the embedding formed by adding the patch embedding with the position encoding. The transformer encoder of ViT consists mainly of layer normalization (LN) which is applied before each block, multi-head attention, and multilayer perceptron block (MLP). Inspired by ViT, Swin Transformer is proposed with a hierarchical transformer, which computes representation with both regular and shifted windows to capture features from different levels and resolutions [26].

A deep residual learning framework (ResNet) is proposed to overcome the degradation issue of the deep neural networks [20]. They add shortcut connections to transmit the information of shallow layers directly to the deeper layers of the neural network and require no additional computation. With this mechanism, the residual net is deepened to 152 layers and achieves state-of-the-art results in multiple computer vision tasks. Furthermore, with the emerging and development of self-supervised representation learning, ResNet also serves as an effective backbone widely used in multiple SSL architectures [7, 19, 28, 31, 43]. WRN [53] is proposed to improve the performance of ResNet by increasing the width of the residual block, i.e., widening the convolutional layers. When employing an equal number of parameters, wide residual blocks yield superior results with reduced training time compared to the original residual blocks. Despite the quadratic increase in parameter numbers and computational complexity associated with the width factor, it proves to be a more efficient approach compared to expanding the number of layers in ResNet. This is because larger tensors can better leverage the parallel computing capabilities of GPUs [53]. WRN has been proved to achieve advanced results in a supervised manner in terms of classification F1-Score and training efficiency [32]. This project intends to explore the feature extraction capability of WRN in SSL framework.

Implementation Details ViT-small is used as the backbone and the implementation follows DINO. In the experiments, we load different backbones as the teacher and student network without pre-trained weights. WRN-50-2 has similar structure to ResNet, with the exception of the bottleneck number of channels, which is twice as large in each block. We follow the recommendation from DINO to use AdamW optimizer. Cross-entropy is employed as the loss function to measure the disparity between feature representations generated by two networks. We evaluate the learned representations by applying KNN and linear probing on EuroSAT land use classification task.

Quantitative Results A common method for assessing feature representations is training a classifier on top of the frozen feature extraction model for a supervised learning assignment. In order to evaluate the representations independently, we freeze the pre-trained backbone model and only train the linear and KNN classifiers on EuroSAT land use classification task. The results are shown in Tab. 1. From the table, DINO-MC performs better than both DINO and DINO-TP when using ViT-samll, WRN-50-2, and ResNet-50 backbones. DINO-MC with WRN-50-2 pre-trained on 100K data is even 2.56% higher than the linear probing accuracy of SeCo pre-trained on 1 million data. Additionally,

Model	Arch	#images	KNN	Linear
MoCo-V2	ResNet-50	1M	-	83.72
SeCo-1M	ResNet-50	1M	-	93.14
DINO	ResNet-50	100K	90.09	89.65
DINO-TP	ResNet-50	100K	79.05	86.70
DINO-MC	ResNet-50	100K	93.94	95.59
DINO	WRN-50-2	100K	92.74	91.65
DINO-TP	WRN-50-2	100K	86.37	88.15
DINO-MC	WRN-50-2	100K	**94.65**	**95.70**
DINO	ViT-small	100K	93.35	91.50
DINO-TP	ViT-small	100K	93.15	93.89
DINO-MC	ViT-small	100K	93.41	94.09
DINO	Swin-tiny	100K	92.15	86.87
DINO-TP	Swin-tiny	100K	92.83	91.94
DINO-MC	Swin-tiny	100K	93.22	90.54

Table 1. Linear and KNN probing classification on EuroSAT. We evaluate our models with different backbones on EuroSAT and record the top-1 accuracy of KNN and linear probing on the validation set. MoCo-V2 [9] and SeCo-1M [27] are pre-trained on SeCo-1M dataset, and their linear probing results on EuroSAT are from SeCo [27]. Other listed models are pre-trained on SeCo-100K dataset with only 10% images of SeCo-1M.

DINO-MC with ViT-samll has 2.59% higher accuracy than DINO with ViT-samll which indicates the effectiveness of our strategy. In comparison to the other three backbones, Swin-tiny performs less well. One possible reason could be that the model size of Swin-tiny is much smaller than the other three models. Another interesting observation is that the DINO-TP performs worse with two convnets than the DINO, but better with two transformer models. Among the different self-supervised models, ViT-small and Swin-tiny performed more consistently than ResNet-50 and WRN-50-2. Overall, DINO-MC is particularly effective.

4.2. Land Use Classification on EuroSAT

EuroSAT is a widely used benchmark dataset for remote sensing land use classification tasks. It is used for representation evaluation in this study, allowing the results of models based on it to be easily compared to those of other models. The dataset, which collects 27,000 remote sensing images from the Sentinel-2 satellite, is divided into 21,600 and 5,400 images for training and evaluation respectively in our experiments.

Implementation Details The pre-trained backbones of the SSL models are evaluated as feature extractors in this land use classification downstream task. Based on the pre-trained backbone models, we incorporate a fully-connected layer as the classifier, leveraging the extracted features to predict the classification results. Both the feature extractor and the classifier are fine-tuned on this supervised classification task. We train the models for around 200 epochs with a batch size of 32. The learning rate is $1e-3$ or $3e-3$ for different models. We use the *CosineAnnealingLR* scheduler to update the learning rate. Same as SeCo [27], we use SGD optimizer without weight decay to update weights.

Model	Backbone	Accuracy
Supervised	WRN-50-2	98.72
Supervised	ResNet-50	98.78
Supervised	ViT-base	98.83
DINO	ViT-small	97.98
DINO-MC	ViT-small	98.15
DINO-MC	Swin-tiny	98.43
DINO-MC	ResNet-50	98.69
DINO-MC	WRN-50-2	**98.78**

Table 2. End-to-end accuracy results on EuroSAT land use classification task. The three supervised models listed are pre-trained on ImageNet-1K [35], and we load and fine-tune them on EuroSAT. While DINO and DINO-MC are only pre-trained on SeCo-100K dataset, which has only 10% the number of images of ImageNet-1K.

Quantitative Results Tab. 2 compares our pre-trained models against three supervised baseline models on EuroSAT land use classification task. DINO-MC with WRN-50-2 achieves similar results to the three supervised models, in particular, it is pre-trained on only 100K images, while the three supervised models are pre-trained on 1M images. This further confirms the effectiveness of the representations learned by DINO-MC.

4.3. Land Use Classification on BigEarthNet

BigEarthNet [39] is a widely used benchmark dataset for land use classification task, containing a total of 590,326 images. This work uses BigEarthNet-S2 [39], which specifically gathers remote sensing images from Sentinel-2 only, and each image is annotated by multiple land use categories. The dataset provides 12 spectral bands for each image and a JSON file with its multi-labels and metadata information. Following SeCo [27], we employ a new nomenclature of 19 classes introduced in [40], and around 12% of the patches that are totally masked by seasonal snow, clouds, or cloud shadows are eliminated in this experiment. We used the training/validation splitting strategy suggested by [30] with 311,667 instances for training and 103,944 images for validation.

Implementation Details We add a linear classification layer on top of the backbone model as the output layer and then fine-tune it on 10% and 100% BigEarthNet, respectively, to evaluate the features learned by the SSL models. In this experiment, the Adam and AdamW optimizer with

Model	Backbone	Param.	10%	100%
SeCo-100K	ResNet-50	23M	81.72	87.12
SeCo-1M	ResNet-50	23M	82.62	87.81
DINO	ResNet-50	23M	79.67	85.38
DINO-TP	ResNet-50	23M	80.10	85.20
DINO-MC	ResNet-50	23M	82.55	86.86
DINO-MC	WRN-50-2	69M	82.67	87.22
DINO-MC	Swin-tiny	28M	83.84	**88.75**
DINO-MC	ViT-small	21M	**84.20**	88.69

Table 3. Mean average precision (MAP) results on BigEarthNet-S2 land use classification. We use the same train/validation splits as SeCo [27]. We fine-tune the pre-trained SSL models with different backbones on 10% and 100% training dataset, and evaluate them on the identical validation dataset. SeCo-100K and SeCo-1M represent SeCo pre-trained on SeCo-100K and SeCo-1M, respectively, and their results listed are from [27]. DINO, DINO-MC, and DINO-TP are pre-trained on SeCo-100K only.

default hyper-parameters are used to update the weights of the models. Identical to SeCo, we set the learning rate to $1e-5$ and scale it down by ten in epochs of 60% and 80%, respectively. We use the MultiLabelSoftMarginLoss as the loss function, which allows assigning a different number of target classes to each sample.

Quantitative Results Tab. 3 provides the results of pre-trained models in the end-to-end BigEarthNet classification task. We assess the performance of each model by mean average precision (MAP). When fine-tuned on the 10% BigEarthNet, DINO-MC outperforms SeCo-100K with the same backbone. Additionally, DINO-MC with ViT-small achieves 1.65% higher MAP than with ResNet-50, 2.48% higher MAP than SeCo-100K, and even 1.58% higher MAP than SeCo-1M. The performance of DINO-MC with each of the four backbone models exceeds that of SeCo-100K, and even outperforms SeCo-1M except for ResNet-50.

When fine-tuned on the whole BigEarthNet, DINO-MC achieves comparable result to SeCo-100K with the same backbone. Interestingly, DINO-MC with Swin-tiny achieves 1.89% higher MAP than with ResNet-50, 1.63% higher MAP than SeCo-100K, and 0.94% higher than SeCo-1M. The results demonstrate that the representations learned by DINO-MC generalize well in the multi-label classification task on BigEarthNet.

4.4. Change Detection on OSCD

OSCD is a benchmark dataset of which the images are collected from Sentinel-2, which mainly focuses on urban growth and disregards natural changes [12]. Change detection is a fundamental problem in the field of earth observation image analysis. The input is a pair of images captured at the same location, while the label is a mask map highlighting the change parts. It is a binary classification task in which labels are assigned to each pixel based on a series of images sampled at different times: change (positive) or no change (negative). The results are measured by F1 Score, precision, and recall in this experiments.

Model	Backbone	Pre.	Rec.	F1
Supervised	ResNet-50	56.49	43.63	48.61
Supervised	WRN-50-2	53.76	47.11	49.72
PatchSSL (21M param.)		40.44	69.10	51.00
MoCo-V2	ResNet-50	64.49	30.94	40.71
SeCo-1M	ResNet-50	65.47	38.06	46.94
DINO	ResNet-50	57.37	44.21	49.53
DINO-MC	ResNet-50	51.94	54.04	52.46
DINO-TP	ResNet-50	51.10	49.03	49.74
DINO	WRN-50-2	53.58	52.28	52.41
DINO-MC	WRN-50-2	49.99	56.81	**52.70**
DINO-TP	WRN-50-2	55.77	47.30	50.61

Table 4. Fine-tuning accuracy results on OSCD change detection task. We adopt the same train/validation splits as SeCo [27]. Following SeCo, we freeze the pre-trained backbone and only update the weights of U-net [34]. Our WRN-50-2 and ResNet-50 models are pre-trained on 100K satellite images. The result of PatchSSL is from [11]. The listed MoCo-V2 [9] and SeCo-1M are pre-trained on 1M satellite images and are results provided by [27].

Implementation Details The OSCD dataset is divided into 14 and 10 pairs for training and validation separately, as recommended by prior research [12, 27]. In addition, the categories of change and unchanged in this dataset are unbalanced due to the property of the task. We use the same U-net architecture as SeCo for the change detection task, which employ the pre-trained backbone models to extract features as the encoder in the U-net. We apply the pre-trained WRN-50-2 and ResNet-50 backbone models as the encoder and select the first convolution layer, and Layer 1 to 4 as the copy and cropping layers. As for the decoder module, it is constructed to rebuild an image with the same width and height as the input. And the input and output sizes of its layers are set according to the input and output of the specific encoder layers, in order to concatenate them in the direction of the channel. Upsampling here is achieved by interpolate function, which can be understood simply as a technology to increase the resolution of the output. The last layer of the U-net is a 1×1 convolutional layer, which is used to map the channel of the feature vector to the number of output classes. During fine-tuning, in order to avoid overfitting, only the U-net weights are updated. In this task, the pre-trained WRN are fine-tuned with a batch size of 32 and a learning rate of 0.0006. The loss function used for this task is $BCEWithLogitsLoss + SoftDiceLoss$.

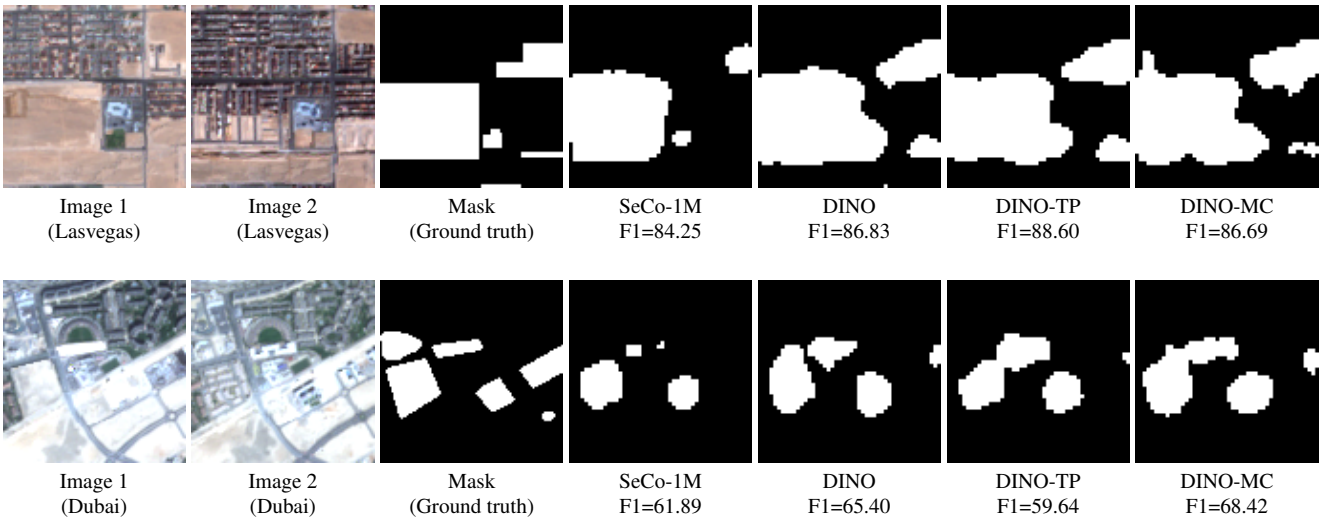

Figure 3. Same as SeCo [27], our visualization is based on two instances of 'Losvegas' and 'Dubai' from the OSCD change detection dataset. SeCo-1M is SeCo model pre-trained on SeCo-1M dataset, while DINO, DINO-TP, and DINO-MC are pre-trained on only SeCo-100K. We visualize the outputs of DINO-TP and DINO-MC for comparison with DINO and SeCo. We present results on the same images as SeCo for comparison and the two outputs of SeCo is from [27]. We also provide F1 score for each output on the bottom of output image.

Quantitative Results Tab. 4 provides the results of DINO, DINO-MC, and DINO-TC with ResNet-50 and WRN-50-2 respectively, compared against some supervised and self-supervised baselines on OSCD dataset. The F1 score of DINO-MC with ResNet-50 is 5.52% higher than that of SeCo and almost 3% higher than that of DINO. When using WRN-50-2 as the backbone, the F1 score of DINO-MC is 5.76% higher than that of SeCo and similar with DINO. It is no surprise that DINO-TP does not perform as well as DINO and DINO-MC, because DINO-TP receives images of the same location taken at different times as positive instances, so it aims to learn features that do not change over time, which is not very suitable for change detection task [27]. In the results of OSCD task, the backbones pre-trained in DINO-MC can capture the subtle differences of image changes over time after a simple fine-tuning on the change detection dataset, which indicates that DINO-MC is able to learn general and effective features using only very simple data augmentation methods.

Qualitative Results Fig. 3 gives the visualization masks of DINO-MC on OSCD task. To do comparison to SeCo, we select two identical examples from the OSCD validation set and present their results. The masks generated by DINO-MC outperform SeCo-1M since they cover more changed pixels without excessive false predictions. We observe that the performance of DINO-TP on OSCD task is unstable since it achieves particularly high F1 score on the first instance, which is 4.35 higher than SeCo-1M, but much lower in the second one, which is 2.25 lower than SeCo-1M. Although positive temporal contrast used in DINO-TP has been proved to be undesirable for change detection task [27], the performance of DINO-TP is comparable to SeCo-1M when evaluated on the whole validation set.

5. Conclusions

This study introduces DINO-TP and DINO-MC, which expand upon the contrastive SSL framework in two ways: (1) by adding a positive temporal contrast strategy, and (2) by introducing a novel cropping and color transformation strategy for local views. The results of KNN and linear probing evaluation on EuroSAT demonstrate the effectiveness of our cropping strategy, as well as the unstable performance of the positive temporal contrast. We evaluate and compare our models with established supervised and self-supervised models on two land use classification tasks and one change detection task in remote sensing domain, revealing the superior performance and effectiveness of DINO-MC.

This study only focus on utilizing temporal views as the positive instances, which does not demonstrate its superiority in our experimental results. In future work, we will further investigate the performance of applying temporal views as negative instances in the contrastive SSL methods.

Acknowledgement This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese

networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022. 2, 4

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 3

[3] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 2

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2, 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 3, 4

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 5

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 4

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

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2

[11] Yuxing Chen and Lorenzo Bruzzone. Self-supervised remote sensing images change detection at pixel-level. *arXiv preprint arXiv:2105.08501*, 2021. 7

[12] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018. 4, 7

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 1, 3

[14] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 1, 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[16] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 4

[17] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 4

[18] M. K. Firozjaei, A. Sedighi, H. K. Firozjaei, M. Kiavarz, and Ska Panah. A historical and future impact assessment of mining activities on surface biophysical characteristics change: A remote sensing-based approach. *Ecological Indicators*, 122(2021), 2021. 1

[19] 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*, 33:21271–21284, 2020. 3, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] 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. 2

[22] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[24] Jiho Jang, Seonhoon Kim, Kiyoon Yoo, Chaerin Kong, Jangho Kim, and Nojun Kwak. Self-distilled self-supervised representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2829–2839, 2023. 3

[25] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:24–49, 2021. 1

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 5

[27] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2, 3, 4, 6, 7, 8

[28] 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. 5

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

[30] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 6

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[32] Ioannis Papoutsis, Nikolaos-Ioannis Bountos, Angelos Zavras, Dimitrios Michail, and Christos Tryfonopoulos. Efficient deep learning models for land cover image classification. *arXiv preprint arXiv:2111.09451*, 2021. 5

[33] 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. 1, 3

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

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3, 6

[36] Sachith Seneviratne, Kerry A Nice, Jasper S Wijnands, Mark Stevenson, and Jason Thompson. Self-supervision. remote sensing and abstraction: Representation learning across 3 million locations. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 01–08. IEEE, 2021. 2

[37] Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multi-mode online knowledge distillation for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11848–11857, 2023. 3

[38] Vladan Stojnic and Vladimir Risojevic. 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*, pages 1182–1191, 2021. 1

[39] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 4, 6

[40] Gencer Sumbul, Jian Kang, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, and Begüm Demir. Bigearthnet dataset with a new class-nomenclature for remote sensing image understanding. *arXiv preprint arXiv:2001.06372*, 2020. 6

[41] Chao Tao, Ji Qi, Weipeng Lu, Hao Wang, and Haifeng Li. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geoscience and Remote Sensing Letters*, 2020. 3

[42] Chao Tao, Ji Qi, Mingning Guo, Qing Zhu, and Haifeng Li. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2, 5

[44] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020. 2

[45] Woody Turner, Sacha Spector, Ned Gardiner, Matthew Fladeland, Eleanor Sterling, and Marc Steininger. Remote sensing for biodiversity science and conservation. *Trends in ecology & evolution*, 18(6):306–314, 2003. 1

[46] CJ Van Westen. Remote sensing for natural disaster management. *International archives of photogrammetry and remote sensing*, 33(B7/4; PART 7):1609–1617, 2000. 1

[47] Stefano Vincenzi, Angelo Porrello, Pietro Buzzega, Marco Cipriano, Pietro Fronte, Roberto Cuccu, Carla Ippoliti, Annamaria Conte, and Simone Calderara. The color out of space: learning self-supervised representations for earth observation imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3034–3041. IEEE, 2021. 3, 4

[48] Xuying Wang, Jiawei Zhu, Zhengliang Yan, Zhaoyang Zhang, Yunsheng Zhang, Yansheng Chen, and Haifeng Li. Last: Label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2, 3

[49] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022. 1, 2

[50] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142. IEEE, 2022. 2, 3

[51] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 2, 3

[52] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2, 3

[53] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2, 5

[54] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 4