Semester Project on Contrastive Representation Learning For Remote Sensing Datasets and Downstream Applications

Ata Bartu Soyuer
ETH Zurich, D-ITET
bsoyuer@student.ethz.ch

Sotiris Anagnostidis
ETH Zurich, D-INFK, DA Lab

Supervised By:

Prof. Dr. Luc Van Gool
ETH Zurich, D-ITET, CV Lab

Prof. Dr. Thomas Hofmann
ETH Zurich, D-INFK, DA Lab

Abstract

In this project, we focus on an unsupervised learning framework, namely, contrastive learning, to obtain representations of remote sensing datasets and use the learnt embedding space for various downstream tasks on satellite imagery. Recently, the commercialization of large scale remote sensing datasets has enabled the training of more complex models, yet the nature, scarcity and quality of their labels introduces more challenges. Thus, the use of unsupervised learning methods to perform accurate inference on such datasets becomes an appealing alternative. Hence, we propose to use a state-of-the-art contrastive learning method to learn better generalizing representations which in turn are expected to yield better performances when used for transfer learning. We present results from various downstream tasks on remote sensing data to argue how contrastive learning methods can in fact provide transferability to other tasks on par with their supervised counterparts.

1 Introduction

In this section, we will be presenting the problem we have at hand, which is to investigate the evaluation and transfer learning performance of contrastive learning methods in comparison to classical supervised learning methods. Moreover, we will discuss some preliminary work related to contrastive learning.

1.1 Problem Definition

Remote sensing data has become increasingly abundant in the fields of Computer Vision and Deep Learning with advancements in aerial imagery. Moreover, many interesting tasks have emerged related to processing remote sensing data in the fields of health, socio-economic studies, environmental/geographic studies, infrastructure, development and many more. Despite this spike of attraction in remote sensing data in the community, this type of data comes with its challenges. Most notably, remote sensing data often has completely or partially missing labels. This may result from either challenges in annotating typically very high resolution aerial images with fine details or necessity of domain knowledge for niche tasks such as the ones listed above. [1]. Therefore, in this project, we experimented with contrastive learning as an unsupervised learning framework in order to be able to learn representations that generalize over such a wide range of classification, detection and segmentation tasks without requiring complete labels.

In the field of remote sensing, many novel and interesting downstream tasks have started to emerge[28]. As more open source satellite data becomes publicly available, governments or third-party organizations start funding new challenges or competitions for various learning tasks related to this wide range of data. While this attracts many researchers in the community, these tasks also have significant impact on the environment and society. Numerous humanitarian work related to environmental planning, infrastructure, security or even population health rely on processing a wide stream of remote sensing data. Moreover, environmental studies often seek state-of-the-art solutions to make inference on this vast source of data in order to offer more viable living conditions for the whole society. Bearing in mind the large-scale impact of remote sensing data in tackling real world problems of utmost importance, we hope to offer a viable and general framework to improve our inference-making capabilities in various related downstream tasks.

1.2 Related Work and Background

In the work of [29; 32], it is argued that classical supervised learning methods may occasionally fail to capture well generalizing features on the dataset they are trained with, despite their reasonable testing performances. Therefore, in order to reliably transfer models to downstream tasks, it is vital to perform the pre-training procedure with a model capable of obtaining well-generalizing representations. Contrastive learning models, are suitable choices for such purposes as they've been reported in numerous work [1; 2; 3; 4] to achieve downstream performance superior than their supervised counterparts. Moreover, these papers further present near-supervised learning accuracies when used in linear evaluation protocols, which is a well-motivated indicator that they indeed output useful representations.

Contrastive learning methods rely on the idea of learning useful representations by training to output representations in some latent space such that image pairs of the same instance (often referred to as positive pairs) are pulled closer whereas pairs belonging to different instances (often referred to as negative pairs) are pushed further apart according to some contrastive loss function operating on that latent space[1]. This way, in the learnt embedding space, the semantically similar positive pairs are 'closer' by the virtue of some metric or similarity kernel used in that space. While this is the core principle behind contrastive methods, each model differs in the exact used contrastive loss, the mining procedure for the negative and positive pairs or the model architectures, which we will address shortly.

In most cases, contrastive learning models are comprised of either a single or Siamese Twin networks[13; 4; 3] each of which consists of an encoder followed by a projection head. Therefore, the encoder which is fed batches of augmented image pairs, learns a mapping to the representation space and the projection head learns a mapping from the representation space to the embedding space where the contrastive loss is applied. Then, the learnt mapping to the representation space is used for downstream tasks. In one of the earliest works of contrastive learning, Chen et al.'s SimCLR [2], it is hypothesized that the decoupling of latent and representation spaces is crucial as the nonlinear projection head is trained to maximize agreement with a contrastive loss under numerous augmentations. Hence, it may strip the representations of information content that may be useful for downstream tasks and thus, the use of a projection head forces the model to accumulate more information in the preceding encoder layer representations.

It is also shown in the work of Chen et al.[2] and Dosovitskiy et al.[9] that augmentation of the image pairs is a crucial component of contrastive models. Otherwise, it is argued that the task of differentiating between positive and negative pairs becomes trivial and the model fails to learn representations generalizing over basic stochastic augmentations or views of the same instances. It is generally an art to fine-tune these augmentations and they can be somewhat dependent on the dataset the model is trained on and we will come back to this issue in the following sections.

Under such considerations, many contrastive models are known to perform reliably well on various common computer vision benchmarks [5; 6; 7]. However, a key issue to consider which stems from the Siamese architecture of these models or the contrastive nature of the task is the *representation collapse*[31] problem. Representation collapse refers to the phenomenon where the two branches of the model trivially satisfy contrastive agreement of positive pairs by outputting constant or non-informative embeddings regardless of the input. Hence, the model trivially learns to maximize agreement between views of each instance without actually discovering a useful embedding. In order to circumvent this, different contrastive models proposed different training or architectural 'tricks' so

the model avoids these non-informative solutions. We will address these proposed solutions as we go through some of the main works on contrastive learning below.

SimCLR.[2] Chen et al., build upon the pioneering work of Hadsell et al.[8], Dosovitskiy et al.[9], Oord et al.[10], Bachmann et al.(2019)[11] on discriminative learning by contrastive methods, to propose one of the first general and widely applicable contrastive learning frameworks to outperform its predecessors significantly, known as SimCLR. In their work, they define the similarity kernel on the representation space via cosine similarity as $sim(u, v) = u^T v/\|u\|\|v\|$ where u and v are generic representation vectors. Then, for all positive pairs and their corresponding representations (z_i, z_j) in a drawn batch of size 2N, the loss *NT-Xent* is defined as:

$$\ell_{i,j} = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \tag{1}$$

It is immediately clear that this loss draws inspiration from the INFONCE Loss[10; 12] and the Cross-Entropy Loss. In this sense, SimCLR sets a contrastive loss which explicitly pushes negative pairs further apart to avoid collapse. The paper further motivates this loss by arguing that its gradient sends a more useful supervision signal by automatically weighting negative samples by their 'hardness' and thus does not require additional mining for hard negative pairs. However, apart from these heuristics, the choice for this loss can still be ambiguous and is there to ensure the model avoids collapse. Moreover, it has been empirically found in [2] that this loss requires a large number of negatives in a batch to actually work which in turn requires very large batch sizes to mine the negatives.

MoCo.[12] In this work, He et al. propose a 'twin' architecture where they use a regular encoder backbone as in SimCLR and a second branch with a special 'momentum encoder'. Moreover, all input samples to both branches are stored as queries of key-value pairs. The basic principle behind this is that during training, positive pairs are mined from the queries of keys of current batch, while negative pairs are mined from the queries of the current batch and the keys from previous batches. This procedure allows having a large number of negatives in each training step and decouples the number of negatives and batch size while doing so owing to the look-up dictionary. The encoder is updated via gradient backpropagation of the contrastive loss as before whereas the momentum encoder is updated via the linear interpolation below:

$$\theta_{k+1} = m\theta_k + (1 - m)\theta_q \tag{2}$$

where θ_k and θ_q are the regular and momentum encoders respectively. The paper argues that this moving average has an exponential dampening effect[3] such that the momentum branch is a low pass filtered version of the other and can produce consistent keys from one batch to the other. This is found to be an empirically required trick since otherwise inconsistent pair of keys again result in collapse.

BYOL.[4] Like MoCo, BYOL also utilizes a pair of twin networks, namely an online and target network. The paper presents that the representations of the target network are bootstrapped such that the online network can learn to predict these bootstrapped representations whereas the target network is updated via a running average as in MoCo. The paper argues that this allows the online network to learn 'enhanced' representations which have stronger robustness against the choice of augmentations by not relying on comparing against negative examples during training. This architectural design is also motivated by distillation methods [16] where the teacher, by providing soft labels, 'distills' its knowledge into the student. While these claims are backed up by empirical findings, the model also requires many tricks such as a predictor head on top of the online network or a stop gradient layer in the target network to prevent collapsed solutions and it is not clear how distillation methods are effective in avoiding collapse.

Barlow Twins.[13] Barlow Twins adopts a single encoder and projector network like SimCLR to which the two batches of augmented instances are fed. However, in order to prevent dimensional collapse, Barlow Twins enforce the following loss function:

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{i \neq j} \mathcal{C}_{ij}^2 \tag{3}$$

$\mathcal{C}$ is the covariance matrix of the output embeddings over a batch with entry $\mathcal{C}_{ij} \in [-1, 1]$ defined as:

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}} \tag{4}$$

where $z^B_{b,j}$ is the j^{th} coordinate of the b^{th} output representation vector in batch B. The paper then builds the intuition that the first term, namely, an 'invariance' term, aims to learn representations that generalize over and are invariant towards different sets of augmentations and the latter term, namely, the 'redundancy reduction term' performs 'soft-whitening' on the covariance matrix. While this loss is also well-motivated by Information Bottleneck Principle[15] in the sense that it filters out the redundant information making the representations as generic as possible while keeping essential information in, it also draws well-founded heuristics from neuroscience[14]. Thus, the loss effectively tries to correlate the representation of opposite views from the two batches maximally while it decorrelates off-diagonal variables of the representation vector and in doing so, filters out redundant or repetitive information in the representations. Hence, it rules out non-informative or constant valued collapsed solutions without having to require additional tricks as the ones listed for other models.

The idea behind this loss is crucial, as it acts as a regularized against collapsed solutions for the overall contrastive loss we will be adopting in our method, which is to be explained in the next section.

Geography Aware Self Supervised Learning.[1] While the previously noted work on contrastive learning report experiments on popular Deep Learning benchmarks such as Imagenet, MS_COCO or PASCAL_VOC [5; 6; 7], this paper is among the pioneering work in terms of analysing contrastive methods in the scope of remote sensing datasets. In this spirit, the paper focuses on MoCo_v2 (a variation of MoCo) as the base contrastive framework and experiments with various large scale geo-tagged datasets such as Functional Map of The Word(fMoW) or GeoImagenet. The paper further discovers a striking performance gap between supervised and contrastive learning schemes in such remote sensing datasets despite how much this gap is closed in popular benchmarks with the recent advancements in contrastive learning. Then, the paper proposes two novel additions to the MoCo_v2 contrastive framework which leverages basic properties of many remote sensing data and in this spirit presents one of the first widely applicable and motivated contrastive learning frameworks in the field of remote sensing.

The first proposal is to make use of the temporally captured views of the same image as positive pairs in the contrastive loss. It is motivated that this allows for much more natural and realistic augmentations, allowing the model to learn more descriptive representations. The second proposal is to introduce a pretext task which is to predict various metadata in the geo-tag of the images, in this case, the latitude and longitude coordinates. This generates additional supervision signal for the network encoder. We however, refrained from using this supervision in our work. The main motivation behind this choice was that such location tags may not be available for all remote sensing data in general and the fact that this method heads in the supervised training regime (almost identical to training a supervised model under labels of locations) whereas we aim to discover representations in a strictly unsupervised manner. In this report, we will also give appropriate reference to this paper as these ideas are explored and become detrimental to our work.

2 Method

In this section we will be introducing the contrastive learning method of our choice, namely, Variance-Invariance-Covariance Regularization (VICReg) [3] and provide a detailed description of its working principles. We will argue that unlike its predecessor contrastive models, it provides a much more intuitive and effective way of avoiding collapse by enforcing an innovative and self-regularizing loss without requiring additional and cumbersome tricks. We note that this is highly desirable as we do not wish any additional technicalities while working with remote sensing datasets which can already be tricky to work with as mentioned. We will finally elaborate on the temporal positive pairs idea from [1], as we shall combine this method with the default VICReg model in our later experiments.

2.1 The VICReg Model

The VICReg model consists of a Siamese network where the two architectures are completely symmetric and share weights.[1] Following the previous work, each branch is comprised of an encoder-expander pair, f_θ and h_ϕ respectively. As in SimCLR[2], the expander strips the representations of the positive pairs of the augmented information that distinguishes them so that contrastive loss indeed

[1] Although, this is not a design constraint for the model and is only set this way for convenience.

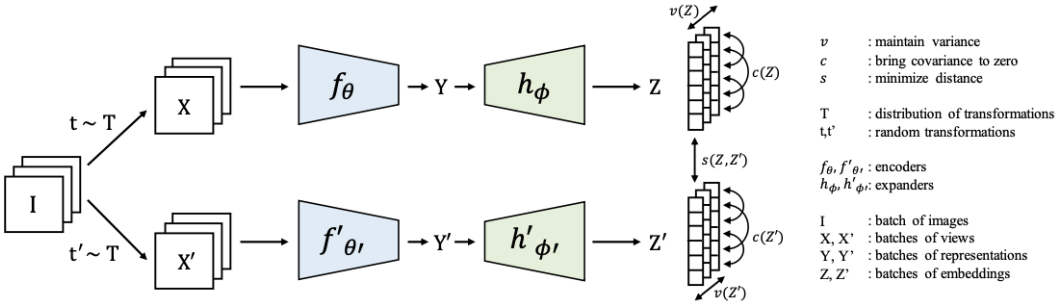


Figure 1: The VICReg Model. Figure retrieved from [3]. A batch of images *I* are augmented into two batches of different views of the same images (by possibly two different set of stochastic augmentations) and the resulting batches *X,X'* are mapped to their corresponding representations *Y,Y'* and embeddings *Z,Z'* respectively. The contrastive loss acting on *Z,Z'*: 1) minimizes the distance between positive pairs of embeddings from the two batches, 2) maintains the variance of each entry of the representation vector in a batch above some lower threshold and 3) shrinks the covariance between different set of entries of the representation vectors. As illustrated here, the two branches need not share weights although they will be so in our experiments for convenience. The encoder backbones will be chosen as an appropriate ResNet model and the expander will be comprised of 3 fully connected layers each of which has a size of 8192.

identifies them as positive pairs in the representation space. Additionally, in VICReg, it projects the representations non-linearly into a higher dimensional space such that the contrastive loss which we will introduce shortly can break the dependencies between entries of the embedding vectors by in fact, decorrelating them similarly to the mechanism in Barlow Twins[13].

The network is fed a batch of images which are then augmented into two separate batches of views via some set of augmentations τ.Then, the two views, namely, x and x' are mapped consecutively into the representation spaces via the encoder as: $y = f_\theta(x)$ and $y' = f_\theta(x')$ respectively. Then, the expander maps these representations to their corresponding embeddings as $z = h_\phi(y)$ and $z' = h_\phi(y')$ where the contrastive loss is applied. As always, only the encoder backbone is used for downstream tasks, utilizing the learnt representations. Technical details regarding the implementation will be detailed in the Experiments section of the report. The overall model is illustrated below in Figure 1.

One of the most striking novelties introduced by the VICReg model, is its unique loss function which can be regarded as a contrastive loss with regularization against collapsed solutions. The overall loss has three additive terms: The **variance**, **covariance** and the **invariance** terms.

Let $Z = [z_1, ..., z_n]$ and $Z' = [z'_1, ..., z'_n]$ be representations of the two augmented views of the same batch and $z_i, z'_i \in \mathbb{R}^d$. Then, the **variance regularization** term is defined via the hinge loss as:

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} max(0, \gamma - S(z^j, \epsilon)) \tag{5}$$

where $S(x, \epsilon) = \sqrt{Var(x) + \epsilon}$. Here, γ is the threshold value below which the variance term inflicts a penalty and ϵ is a small number to prevent vanishing gradients in the variance term during training. $Var(x)$ computes the variance of the batch of embeddings over the x^{th} dimension of the embeddings and the variance loss term encourages it to be above γ which is usually fine-tuned. Hence, the mapping of input instances into constant or very similar representations is explicitly prevented. The **covariance regularization** term is defined as:

$$c(Z) = \frac{1}{d}\sum_{i \neq j} [C(Z)]_{ij}^2 \tag{6}$$

where $C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$ with $\bar{z}$ being the empirical mean of representations over a batch. We note that this is the same term as the off-diagonal regularization term in Barlow Twins[13] except now, the covariance matrix is computed within the output of each Siamese and not cross-correlating them. Likewise, it explicitly decorrelates the dimensions of the embedding vector over a batch as to prevent informational collapse where each dimension would encode similar and thus redundant information. It is also shown in the paper empirically that, this also results in some decorrelation effect at the representation level as desired which also seems to be positively correlated with the performance of the model. Thirdly, the **invariance** term is given by:

$$s(Z, Z') = \frac{1}{n}\sum_{i}(\|z_i - z_i'\|)^2 \tag{7}$$

As in classical contrastive methods, this term purely acts to bring the positive pairs closer in the embedding space. Note that the model doesn't explicitly compare against negative examples in the loss [2] and it is argued that the role of negative pairs is played by the variance regularization term which ensures a certain amount of variance within a batch of negative samples such that they cant collapse into a single vector. Finally, the overall loss over one augmented pair of batches is:

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] \tag{8}$$

where λ, μ, ν are hyperparameters. In overall, owing to this cleverly designed loss, VICReg doesn't require any additional tricks such as large batch sizes or momentum encoders to mine negative samples, no feature-wise or batch-wise normalizations for the embeddings or architectural tricks such as stop gradients or predictor heads in order to avoid collapsed solutions. Instead, it naturally and reliably avoids collapse by the notions of variance and covariance regularization introduced in the loss and ensures useful representations are learnt via the MSE term. These advantages over alternative contrastive methods make VICReg a simple yet generally applicable method for contrastive learning. This ease of interpretability and reliability in VICReg is key to our applications in this report as remote sensing data already consists of challenges such as very high resolution images, heavy class imbalances or semantic intra-class similarities in numerous downstream tasks. Thus, a model naturally capable of producing uncollapsed and descriptive representations for any input source without having to incorporate additional cumbersome tricks and technicalities, such as VICReg, is highly suitable for our experiments on unsupervised representation learning in remote sensing datasets. Moreover, the original paper [3] reports linear evaluation and downstream results which surpass its predecessor contrastive methods and we are hoping to utilize this potential to its full extent in our remote sensing applications as well.

2.2 Temporally Positive Pairs

As introduced in [1], using temporally aligned images as positive pairs is a very natural, effective yet inexpensive method. Suppose τ is the set of hand-crafted augmentations used in our model and x^{t_1}, x^{t_2} be temporally captured views of the same instance in the remote sensing dataset. Here, τ encodes a 'mild' set of augmentations as the choice of temporal images as positive pairs already encodes some natural notion of augmentation. In practice, this method can even work with no additional augmentations (i.e, τ is the identity transform) and we have investigated with this option in our experiments. Then, using these pairs and cascading them with augmentations τ, we get that $Z = [f_\theta(h_\phi(\tau(x_1^{t_1}))), ..., f_\theta(h_\phi(\tau(x_n^{t_1})))]$ and $Z' = [f_\theta(h_\phi(\tau(x_1^{t_2}))), ..., f_\theta(h_\phi(\tau(x_n^{t_2})))]$ in the above loss terms. Hence, we note that, VICReg is perfectly capable of incorporating these temporally positive pairs and in fact, we hope that the use of these temporal augmentations along with the strengths of the VICReg model will allow us to get results on par with [1] in the remote sensing data that we shall use. Some illustrative pairs of temporal positives taken from the FmoW Dataset[20] are depicted in Figure 2 in Appendix A.1.

2.3 Used Datasets

In this section, we will elaborate shortly on the two main datasets we have used to train the VICReg model. The default dataset we had initially used in our experiments is the SpaceNet-7 Dataset[19].

[2] Much like BYOL

The training data for this dataset contains a set of 24 temporal images taken in monthly intervals for 60 different and developing locations worldwide from a total of 31 countries. Each image is a 4-band 1024 x 1024, medium resolution (4 meters per pixel) satellite image of the associated location. Although the dataset originally belongs to an urban development detection challenge, for VICReg training purposes we ignored these masks and used the country labels extracted manually from the GeoJSON files for the supervised training experiments. Moreover, since the original SpaceNet dataset had perhaps too few images for our experiments, we have divided each images to 16 patches of size 256x256 and used them as separate samples in our training. Secondly, we have used the Functional Map of The World (fMoW)[20] dataset to train our backbones. This is the same dataset used in the inspirational work of [1]. The dataset contains numerous 3-band, very-high-resolution satellite imagery of varying sizes and of $\sim$0.5 meters per pixel ground truth distance resolution. Each image belongs to widely selected set of locations under 63 categories of land-covers and infrastructure. Each location consists of a varying number of temporally aligned images as well. At a total of 360K samples, this dataset allowed us to replicate our experiments at a large scale.

2.4 Implementation Details

Now we will be addressing the technical details related to the training setup of our model in our experiments. The setting described here will be the default training setup with the SpaceNet-7 Dataset[19] used in our experiments unless stated otherwise. Firstly, the hyperparameter coefficients λ, μ, ν are set to 25, 25 and 1 respectively. This choice stems from the experiments carried out in the original VICReg paper[3] as this setting was reported to be most optimal in the sense of linear performance. We also briefly experimented with different value combinations for these coefficients, however, we didn't see a noticeable difference in the overall performance of the model. The encoder network is chosen as an appropriate ResNet backbone which in our experiments will be either ResNet-18 or ResNet-50. The decoder is a three layer fully connected network of sizes 8192 which we have borrowed directly from the original paper. We also observed that decreasing the layer sizes worsened the overall linear performance of the model whereas increasing it had negligible effect. For the optimization, we used a LARS Optimizer[17] with 10^{-6} weight decay for 100 epochs under cosine delay scheduling[18] where we used a linear warmup for the learning rate for the first 10 epochs after which the learning rate is gradually decreased until a final value of 0.002. We have also used a learning rate given as $learning\_rate = base\_lr * batch\_size/128$ where *batch_size* is set to 256 by default. Finally, for training on the SpaceNet dataset, we use the exact set of augmentations in [3] which is a stochastic cascade of crop, color distortion, Gaussian blurring and normalization transforms. We note that this is a well founded training regimen used in notable prior work[3; 4; 13]. For training on the fMoW dataset, we use the set of augmentations used in [1] which also were also reported in this work to yield satisfactory results on remote sensing datasets. Although these set of augmentations, the color distortions in particular, may occasionally generate temporal pairs which are counter-intuitive or unnatural seeming for aerial imagery, we mostly stick to these augmentations due to arguments pointing in favor of such distortions in previous work[2].

3 Experiments and Results

In this section, we will detail the experiments we conducted with the trained VICReg encoder backbone. Any changes in the default implementation or the used datasets will be addressed in the experiment details. We shall report results from various downstream evaluations including linear and semi-supervised evaluation protocols, KNN and K-Means evaluation protocols and transfer learning results on classification, detection and segmentation tasks on various remote sensing data. Moreover, an analysis and discussion of the results will be presented in the later section of the report.

3.1 Linear and Semi-Supervised Evaluations

In this section we use our trained VICReg encoder in a linear evaluation protocol[21] on the validation split of the Spacenet-7 Dataset. There are 2 procedures that we followed here. We either froze the encoder weights and trained the linear classifier layer on top of the whole training dataset or we finetuned both the encoder and the linear layer on a partition[3] of the training dataset in a semi-supervised manner. The training sessions for the linear and semi-supervised evaluations last for 40

[3]We have used 1% or 10% partitions

and 20 epochs respectively. We finally evaluate the Top-1 and Top-5 accuracies on the validation split of the dataset. For the finetuning setup, we exactly followed the linear and SSL evaluation setup detailed in [3]. The results are depicted in Table 1. We finally note that for the SpaceNet-7 dataset, due to the number of samples being limited, the results obtained on Semi-Supervised Evaluation using 1% of the labels were inconclusive as the fine-tuned model would have a lot of variance. Thus, for convenience, we simply chose to exclude the 1% evaluation results for the cases where SpaceNet-7 dataset was used to train the model.

Table 1: Linear and Semi-Supervised Evaluation Results obtained on trained VICReg encoder. The used backbone architecture for the encoder and the dataset it is trained which is specified in parentheses. The use of Temporal Positives(if any) is specified by the '+TP' sign. The results include: 1) Linear classification on top of frozen representation learnt on the associated dataset. 2) Classification Results after training a linear layer and fine-tuning the learnt representations using 1% or 10% of the training samples of the associated dataset. Top-1 and Top-5 accuracies are obtained from the validation split of the dataset used in the associated experiment. We perform an evaluation on the validation set after each epoch and report the best performance attained by the model throughout the training.

Method	Linear		Semi-Supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised(ResNet18+SN7)	0.7674	0.9774	-	0.4007	-	0.8050
Supervised(ResNet50+SN7)	0.8862	0.9927	-	0.2524	-	0.6463
VICReg(Resnet18+SN7)	0.8184	0.9873	-	0.7667	-	0.9633
VICReg(Resnet50+SN7)	0.9116	0.9977	-	0.7942	-	0.9693
VICReg(Resnet50+TP+SN7)	0.9461	0.9983	-	0.8441	-	0.9767
VICReg(Resnet50+TP+fMoW)	0.5670	0.8278	0.4983	0.5818	0.7638	0.8310

3.2 KNN and K-Means Evaluations

In this section, using heuristics inspired by [22], we try to evaluate the 'compactness' of the learnt representation space via running KNN and K-Means algorithms on the learnt representations. For the related experiments, we have two different protocols. First protocol is that we first obtain the representations of all the samples in the training and validation splits of the dataset by a single forward pass. Then, we fit the KNN classifier on the representations belonging to the training set and evaluate the performance on the representations of the validation set. The second protocol we apply specifically for the SN7 dataset is that we split the test set into 'seen' and 'unseen' locations and apply the first protocol where 'seen' and 'unseen' test sets replace the training and validation sets respectively. The idea is that while Protocol-1 evaluates the performance of representations for unseen samples, Protocol-2 evaluates the performance under unseen classes. For the K-Means experiments, we simply treat the training and validation datasets or the 'seen' and 'unseen' test sets as a whole and fit the K-Means model on these overall datasets. The accuracies of the cluster indices are reported after matching them with the well known Hungarian Algorithm. The KNN evaluation results are reported for $K = 20$ and $K = 200$ neighbours. For SpaceNet-7 dataset related K-Means experiments, we set number of clusters to 31 in Protocol-1 since all images used come from 31 countries in total and in Protocol-2 we set number of clusters to 20 as test images used here do not have geo-tags so we simply treat each set of temporal images as a separate ground truth cluster. Moreover, we carry out KNN protocol-2 for fMoW and Imagenet datasets as well, where we fit the KNN classifier in the representations of the training set and evaluate on those of the validation set. We also note that, for the case of Imagenet and fMoW trained models, we cannot perform Protocol-1 since in the case of these datasets, the trained model already observes samples from all classes during training unlike the case of SpaceNet-7 dataset where the test set purely contains images from new locations. For the K-Means based evaluations, we set number of clusters as 1000 and 63 in the ImageNet and fMoW datasets respectively as per the number of ground truth classes. Also, for computational limitations and time constraints, we perform the K-Means evaluations on the validation set alone for these two large-scale datasets. Since the original VICReg paper doesn't include KNN based evaluations of the model, the results obtained from Imagenet trained models serve as a nice sanity check. The results

are documented on Table 2. For completeness, we have also visualized a portion of the obtained representations of each model on 2 dimensions using t-SNE[30]. The plots are given in Figure 3 of Appendix A.2.

Table 2: KNN and K-Means based evaluation results obtained on trained VICReg encoder. The used backbone architecture for the encoder and the dataset it is trained with are specified in parentheses. The use of Temporal Positives(if any) is specified by the '+TP' sign. Protocol-1 refers to evaluation of KNN and K-Means models under unseen samples(using training and validation sets) and Protocol-2 refers to evaluation under unseen classes (using the test set).

Method	KNN				K-Means	
	Protocol-1		Protocol-2		Protocol-1	Protocol-2
	K=20	K=200	K=20	K=200		
Supervised(ResNet18+SN7)	0.9074	0.8413	0.8026	0.7028	0.5744	0.4455
Supervised(ResNet50+SN7)	0.8720	0.8007	0.6978	0.5691	0.4687	0.3907
Supervised(ResNet50+ImageNet)	-	-	0.7326	0.7158	-	0.4836
VICReg(Resnet18+SN7)	0.9702	0.5153	0.9091	0.5568	0.1611	0.2388
VICReg(Resnet50+SN7)	0.9652	0.5611	0.9253	0.6510	0.1782	0.3071
VICReg(Resnet50+TP+SN7)	0.9950	0.5840	0.9464	0.6577	0.1632	0.2895
VICReg(Resnet50+TP+fMoW)	-	-	0.5058	0.5083	-	0.2261
VICReg(Resnet50+ImageNet)	-	-	0.5920	0.5697	-	0.3199

3.3 Downstream Transfer Learning Tasks

In this section, we will present the results obtained from downstream classification, detection and segmentation tasks on different remote sensing data performed by using the pretrained encoders from either the VICReg of supervised models.

3.3.1 Land-Cover Classification Task

In this task we have used the EuroSat[23] dataset which is a novel geo-tagged dataset using a subset of Sentinel-2 imagery. The dataset consists of $\sim$27K 13-band imagery of 10 meters per pixel resolution under 10 categories of land-covers. We note that only the RGB channels will be relevant for our applications. For this task, we have simply retrieved our trained ResNet-50 encoders, cascaded the encoder with a fully connected linear layer and trained the overall network with Cross Entropy Loss. During training, we have applied common ResNet training heuristics such as gradient clipping and weight decay and trained for 10 epochs with *Adam* optimizer. The best accuracies reported on the validation split of the dataset can be seen in Table 3. We have also plotted a sample batch of images from this dataset for completeness and the associated plots can be found in Figure 4 of Appendix A.3.

Table 3: Linear Evaluation Results of trained encoder backbones on the downstream task of land-cover classification on the EuroSat Dataset. The used backbone architecture for the encoder and the dataset it is trained with are specified in parentheses. The use of Temporal Positives(if any) is specified by the '+TP' sign. Accuracies are obtained on the validation split of the dataset.

Method	Accuracy (Top-1)
Supervised(ResNet50+SN7)	0.9447
Supervised(ResNet50+ImageNet)	0.9832
VICReg(Resnet50+SN7)	0.9651
VICReg(Resnet50+TP+SN7)	0.9731
VICReg(Resnet50+TP+fMoW)	0.9720

3.3.2 Building Detection Task

In this task, we have conducted large-scale experiments with the CrowdAI Mapping Challenge Dataset[24]. The data consists of 280K training and 60K validation images each of which are

300x300 RGB images. The task in this dataset is to detect and mask building in a given input image. We also note that the images in the dataset are of considerably lower ground truth distance per pixel. The annotations are provided in MS-COCO[7] format. Our choice of model was a MASK-RCNN with FPN backbone[25] where the backbone uses the pre-trained weights of our encoder. For the training setup, we have used an *SGD* optimizer for 90K iterations of batches. We have used a *weight_decay* of 10^{-6} and a base *learning_rate* of 0.01 along with some gradient clipping. The learning rate scheduling was as follows: We have used a linear warmup for 1.5K iterations and used *MultiStepScheduler* which multiplies the learning rate by a factor of $gamma = 0.1$ at iterations 60K and 80K similar to the detection task training setup in[3]. We report the detection and segmentation results in Table 4.

Table 4: Transfer Learning results of trained encoders on the downstream task of building detection on CrowdAI Mapping Challenge Dataset. Mask-RCNN model with FPN backbone and pre-trained encoder weights are utilized. The used backbone architecture for the encoder and the dataset it is trained with are specified in parentheses. The use of Temporal Positives(if any) is specified by the '+TP' sign. Average Precision metrics for various detection thresholds are reported on the validation split of the dataset.

Method	Bounding Box			Segmentation		
	APm	AP@50	AP@75	APm	AP@50	AP@75
Supervised(ResNet50+SN7)	0.6248	0.9011	0.7173	0.5758	0.8844	0.6722
Supervised(ResNet50+ImageNet)	0.6586	0.9206	0.7572	0.6009	0.9012	0.7067
VICReg(Resnet50+SN7)	0.5967	0.8898	0.6895	0.5580	0.8751	0.6483
VICReg(Resnet50+TP+SN7)	0.6089	0.8967	0.7035	0.5659	0.8801	0.6603
VICReg(Resnet50+TP+fMoW)	0.6231	0.9015	0.7185	0.5760	0.8848	0.6748
VICReg(Resnet50+ImageNet)	0.5984	0.8898	0.6944	0.5617	0.8746	0.6554

3.3.3 Flood Damage Segmentation Task

In this final downstream task, we perform semantic segmentation of flooded buildings on a Sentinel dataset captured from Hurricane Harvey[27]. The dataset consists of 1973x2263 VHR images which we have resized and divided up to patches of size 512x512 for our training purposes. Each image has ground truth distance resolution of 0.5 meters per pixel. The dataset has provided 2 pixel-wise masks of all the buildings in the image and a separate mask of all the flooded/damaged buildings in the image. The training protocol we followed was to use a simple U-Net[26] model whose backbone was our pre-trained encoder and we have first used the building masks to fine-tune the model and then trained on the actual masks of flooded buildings. As training objective, we have used the DICE coefficient. In Table 5, we report the DICE coefficient and IoU results on the validation split of the dataset, both for the building masks after fine-tuning and the flooded building masks after the overall training. Moreover, visual segmentation results for the overall building masks after the fine-tuning stage and the segmentation results for the flooded building masks after the actual training stage are provided in Figures 5 and 6 of Appendix A.4.

4 Discussion

In this section, we will analyze the results given in the Experiments section of the report and try to justify the obtained results for each of our evaluation protocols and downstream tasks. To begin with, as we glance at the results of Table 1 for linear and semi-supervised evaluation protocols of trained ResNet backbones, we immediately notice that the VICReg backbones perform much better in both evaluations compared to their supervised counterparts. In particular, the original paper[3] reports similar results on the ImageNet dataset with the exception of VICReg model being slightly ($\sim$3%) worse in linear accuracy compared to the supervised model. In our case, VICReg backbones surpass their supervised counterparts significantly as well. While this result is a welcome one, we also suspect that this additional performance of the VICReg backbone over the supervised one may be an artifact of the dataset. To elaborate, unlike ImageNet, the SpaceNet-7 dataset consists of images which are temporally aligned, resulting in more intra-class similarities between images. Therefore, we believe that the generalized features learnt by the VICReg model in fact can boost the linear

Table 5: Transfer Learning results of trained encoders on the downstream task of building and flooded building segmentation on Hurricane Harvey Dataset. U-Net architecture with pre-trained encoders as backbone are used. The used backbone architecture for the encoder and the dataset it is trained with are specified in parentheses. The use of Temporal Positives(if any) is specified by the '+TP' sign. DICE Coefficient and IoU metrics at a detection threshold of 0.5 are reported on the validation split of the dataset.

Method	Building		Flooded Building	
	DICE Coefficient	IoU@50	DICE Coefficient	IoU@50
Supervised(ResNet50+SN7)	0.6674	0.5158	0.6901	0.626
Supervised(ResNet50+ImageNet)	0.6983	0.551	0.6897	0.6797
VICReg(Resnet50+SN7)	0.6505	0.4961	0.7156	0.6558
VICReg(Resnet50+TP+SN7)	0.6525	0.5002	0.6835	0.6375
VICReg(Resnet50+TP+fMoW)	0.6565	0.5075	0.7034	0.6472

evaluation performance of the model above the supervised model as well. Unsurprisingly, we also notice that ResNet50 backbones have better overall performances compared to ResNet18 backbones as reported in [3]. Finally, we observe that the use of temporal positives in VICReg training increases the accuracies for both evaluations. As linear evaluation on top of frozen representations as in our case is a well motivated metric for assessing the learnt embeddings by a model, we concur that the claims in [1] on temporal positives are partly justified since the model indeed performs better, necessarily on linear evaluation. Moreover, the overwhelming performance of the VICReg model on semi-supervised evaluations emphasize the advantage of the contrastive learning framework over supervised learning in the missing/incomplete annotations scenario as motivated in the Introduction. Lastly, we have repeated the same evaluations for the fMoW trained VICReg model as well for completeness. The results were comparable to those reported in [3] for the ImageNet dataset. As both datasets are quite rich and large-scale datasets, we were satisfied with these results which served as a nice sanity check. Due to lack of time, we were unable to replicate these experiments for a supervised model trained on fMoW, however, we expect similar results to hold.

Secondly, we will argue the results on K-Means and KNN based evaluations we performed on our learnt representations. In the contrastive learning literature, there is only few notable work (such as [22]) that carry out these evaluations on the trained model. So, we aimed to demystify the renowned performance of contrastive models by exploring the learnt representation space of the models. The paper [22] experimentally presents that compared to state-of-the art contrastive models, supervised models surprisingly learn representations that are more compact in the sense of KNN evaluations in the representation space. In the KNN-based results presented in Table 2, we notice that the VICReg model in fact performs better on the K=20 neighbours case as opposed to the findings of [22] ImageNet. Once again, we strongly believe that this is a phenomenon due to the existence of temporal images in the dataset. Unlike ImageNet, for each patch of image we use for this evaluation there exists 23 other temporally captured images of the same patch. We hypothesize that the model, more naturally and easily learns to bring the representations of these temporal pairs of the same class closer together as per the contrastive loss. Thus, for a smaller number of neighbour used, the VICReg model actually seems to perform better on KNN evaluations. However, for K=200 neighbours, our results indeed agree with [22] and we notice how the performance of the supervised models are clearly better. This leads us to hypothesize that: 1) The representation space learnt by supervised models is ultimately much better grouped into representations of the same classes as we take in to account all the samples and 2) The representations learnt by the VICReg model have their temporal positives as some of their closest neighbours but ultimately, the grouping structure of the representation space is poorer. Our hypothesis is further supported in the K-Means based evaluations where the supervised models significantly surpass VICReg models in the cluster assignment accuracies. Moreover, as the most noticeable conclusion from Figure 3 where we visualize a partition of samples and their representations, the supervised models yield representations with much better grouping structure overall. This finding also supports our hypothesis on how the learnt representation spaces may be structured. Additionally, we observe in Table 2 that the gap between the supervised and the VICreg model seems to be less for protocol-2 where we obtain and evaluate on the representations of the test set which come from unseen classes (i.e, unseen locations) during training time. Based on this result, we infer that VICReg model in fact produces better grouped representations for unseen classes. We

believe that this may translate into better transferability of the model to unseen/other tasks which is our ultimate expectation. Finally, for completeness, we repeat the evaluations for fMoW trained VICReg model. As we did for linear evaluation case, we take the ImageNet evaluations as a point of reference and notice that once more the results are quite comparable between the two. We explain the gap by the fact that the ImageNet pretrained VICReg model which we use in our experiments here is originally trained for 1000 epochs in [3] which is 10 fold what we train our models for. This would naturally aid the model to learn more compact representations on the dataset it is trained on (owing to the VICReg Loss described on Equation 8) and hence possibly yield better KNN and K-Means evaluations. Nevertheless, we notice that fMoW trained VICReg model doesn't suffer from a drastic drop in KNN performance with increasing number of neighbours as with other VICReg models trained on SpaceNet-7 dataset. Hence, we have reason to believe that the fMoW trained model learns richer representations owing to the large-scale fMoW dataset and thus, more generalized features along the way. We shall analyze briefly how this belief translates into the performance of the models in the downstream tasks.

Now, we shall elaborate on the result obtained on the three downstream tasks we pursued in this work. We begin our discussion with the land-cover classification task. In this task, the results as presented in Table 3 turned out to be almost as anticipated. We see from the results that all variations of the VICReg models surpassed their supervised counterpart in terms of the Top-1 Accuracy quite significantly. We also observe that the additional utilization of temporal positives boosts the performance by nearly 1%. We had further anticipated that the fMoW trained VICReg model would actually perform better than the other variations of the VICReg model. However, it seems to be very slightly behind the SpaceNet-7 trained model with temporal positives although still well above the regular VICReg model. We suspect that this gap which is uncalled for could be associated with the ground-truth distance-per-pixel resolutions of the datasets. As mentioned before, the fMoW, SpaceNet-7 datasets used for training and the EuroSat dataset used in the downstream task have different distance-per-pixel resolutions. Moreover, as can be seen from Figures 2 and 4, the EuroSat and SpaceNet-7 datasets are closest in terms of the distance-per-pixel resolutions whereas fMoW dataset has considerably higher distance-per-pixel resolution. We believe that this might have allowed the model with temporal positives trained on the SpaceNet-7 dataset to transfer its learnt features more easily and naturally in to the task without requiring any re-scaling and hence, resulting in the small gap. Finally, as opposed to the results reported for the downstream tasks in [3], we notice that the VICReg models fall behind the ImageNet pre-trained Supervised model on this task. We suspect that this might have to do with the differences in the training schemes of both models. More specifically, the ImagNet pre-trained model is conventionally trained for a very large number of epochs, in a meticulous manner and using a dataset on much larger scale compared to SpaceNet-7. Thus, it is not far-fetched to expect that the ImageNet pre-trained supervised model could actually learn richer set of features, allowing it to out-perform other models on this task. Lack of time has prevented us from comparing this model against its contrastive counterpart to test this hypothesis, however, based on the results of [1], we believe that with the utilization of temporal positives, this gap could indeed be closed given that we train on a comparable number of epochs to that of the ImageNet pre-trained model.

We will now proceed to analyze the results presented for the building detection task whose results are presented in Table 4. To begin with, we see that the VICReg models are out-performed by their supervised counterparts with the ImageNet pre-trained supervised model once again taking the lead both in the bounding box regression and mask segmentation metrics. On the bright side, we notice that the utilization of temporal positives or the fMoW dataset for the training of the model indeed boosts the model performance quite significantly which is in line with our claims. In fact, the VICReg model trained on fMoW with temporal positives seems to be almost on par with the supervised network, although below our expectations. At first, we had conjectured that this dramatic performance drop of the VICReg models could have been an artifact of the large learning rate used by the optimizers utilized during the training of VICReg models. We believed that the pre-trained VICReg models would have larger weights as they were trained on *LARS* optimizers rather than *SGD* optimizers which we used conventionally for training the supervised models. Hence, we tried running the downstream tasks with different learning rates on the VICReg models, however, the results haven't improved significantly. Hence, we ruled out this hypothesis. Moreover, when we used the ImageNet pre-trained VICReg model [4] on the downstream task as a sanity check, we saw that the

[4] provided on the Github repository of [3]

results were also below its supervised counterpart. This points to the performance gap of contrastive models against their supervised counterparts for remote sensing datasets as claimed in [1]. Thus, we believe that these performance gaps could be closed with the proper usage of temporal positives as demonstrated in [1]. In fact, in this particular detection task, we have both experimented with using and discarding augmentations along with the VICReg model trained on SpaceNet-7 with temporal positives. Thus, specifically for this task, we observed that the use of temporal positives without any artificial augmentations seemed to provide better end results which we chose to provide in Table 4. We note that the use of temporal positives in the absence of additional augmentations didn't yield results significantly different than the case utilizing hand-crafted augmentations for other downstream tasks. Nevertheless, we believe that the leap in the performance for this specific task highlights the significance of relying on natural and intuitive augmentations on remote sensing datasets as provided by temporal pairs rather than choosing arbitrary, hand-crafted augmentations.

Finally, we will discuss the results presented in Table 5 for our third and final downstream task. We first go through the results after the first fine-tuning stage, i.e, the results obtained for segmenting all the buildings on the images after the U-Net model with the pre-trained backbone is fine-tuned using the masks of all the buildings in the images. We notice that the results are quite inconclusive for this fine-tuning stage. Most notably, we see that the ImageNet pre-trained supervised model once again demonstrates the best performance both in terms of the DICE Coefficient and IoU scores. We again suspect this is due to the ImageNet dataset and the training scheme of this model as explained in the previous downstream tasks. The VICReg models variations also seem to fall behind their supervised counterpart in both metrics. As in the previous task, we once again notice that both the use of temporal positives and the fMoW dataset during training boost the performance of the model as expected. We however, would like to focus on the results regarding the actual training stage, i.e, the segmentation results of only the flooded buildings after the fine-tuned U-Net model is trained on the masks of flooded buildings. Here, as observed numerous times before, the ImageNet pre-trained supervised model shows the best performance in IoU scores although its DICE score is not the best among other candidates. Initially, we believed this might indicate that the other models are moderately better on most examples although significantly worse on few samples. However, as we went through the visualization of segmentations for each model as exemplified in Figure 6, we did not come across any such significantly poorer examples either. On the other hand, from these results we observe a very clear and irrefutable conclusion which is that the VICReg models indeed provided better end-results in terms of both metrics against their supervised counterpart. We believe that this might be an indication of how better the generalized features learnt by the VICReg models are transferred into the final training stage, thus resulting in better overall performance for our task. We hypothesize that although the VICReg models are not particularly promising after the fine-tuning stage, the fine-tuning of the U-Net models using VICReg backbones in fact prepared the models better for the subsequent training stage and along with the more enhanced features they have acquired, better translated this into out ultimate task of detecting flooded buildings.

5 Conclusion

In this report, we have presented a simple yet widely applicable unsupervised learning framework intended to tackle common issues in processing remote sensing data and yield state-of-the-art transfer learning performance for aerial imagery. Inspired by the recent line of work, we have combined the promising contrastive learning framework, namely VICReg[3], and the idea of utilizing temporally positive pairs[1] for the choice of augmentations in the contrastive framework to learn rich and well-generalizing representation spaces for remote sensing data. Our method yielded a mix of satisfactory and inconclusive results on our experiments with the chosen downstream tasks. Although our method may require additional empirical evidence, we believe that our work fully motivated and experimentally demonstrated the absolute necessity of temporal positives, meticulous choice of hand-crafted augmentations and the promise of contrastive learning frameworks for downstream applications concerning remote sensing data.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell & Stefano Ermon. Geography-Aware Self-Supervised Learning. *arXiv preprint arXiv:2011.09980, 2020*.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, & Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709, 2020*.

[3] Adrian Bardes, Jean Ponce, & Yann LeCun. VICREG: Variance-Invariance-Covariance Regularization For Self Supervised Learning *arXiv preprint arXiv:2105.04906, 2022*.

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec& Pierre H. Richemond. Bootstrap Your Own Latent A New Approach to Self-Supervised Learning. *arXiv preprint arXiv:2006.07733, 2020*.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li& 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

[6] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn & Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. In 2009, *International Journal of Computer Vision*, pages 303–338.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick & Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv preprint arXiv:1405.0312, 2014*.

[8] Hadsell, R., Chopra, S.& LeCun, Y. Dimensionality reduction by learning an invariant mapping. In 2006, *IEEE Computer So- ciety Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

[9] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M.& Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *In Advances in neural information processing systems*, pp. 766–774, 2014.

[10] Oord, A. v. d., Li, Y.,& Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748, 2018.*

[11] Bachman, P., Hjelm, R. D.,& Buchwalter, W. Learning rep- resentations by maximizing mutual information across views. *In Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick& Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297, 2020*.

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun& Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv preprint arXiv:2103.03230, 2021*.

[14] H.B. Barlow. Possible Principles Underlying the Transformations of Sensory Messages. *in Sensory Communication*. The MIT Press, 1961.

[15] Naftali Tishby, Fernando C. Pereira& William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057, 2000*.

[16] Geoffrey Hinton, Oriol Vinyals& Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[17] Yang You, Igor Gitman& Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888, 2017*.

[18] Ilya Loshchilov& Frank Hutter. Sgdr: stochastic gradient descent with warm restarts. *In ICLR, 2017*.

[19] Adam Van Etten, Daniel Hogan, Jesus Martinez-Manso Jacob Shermeyer, Nicholas Weir& Ryan Lewis. The Multi-Temporal Urban Development SpaceNet Dataset. *arXiv preprint arXiv:2102.04420, 2021*.

[20] Gordon Christie, Neil Fendley, James Wilson& Ryan Mukherjee. Functional Map of the World. *arXiv preprint arXiv:1711.07846, 2018*.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li& Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *In CVPR, 2009.*

[22] Mathilde Caron, Hugo Touvron Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski& Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.14294, 2021*.

[23] Patrick Helber, Benjamin Bischke, Andreas Dengel& Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *arXiv preprint arXiv:1709.00029, 2017*.

[24] Sharada Prasanna Mohanty, Jakub Czakon, Kamil A Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad& Sascha Fleer. Deep Learning for Understanding Satellite Imagery: An Experimental Survey. *Frontiers in Artificial Intelligence*, vol. 3, 2020.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár,& Ross Girshick. Mask r-cnn. *In ICCV, 2017.*

[26] Olaf Ronneberger, Philipp Fischer& Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv preprint arXiv:1505.04597, 2015*.

[27] Hurricane Harvey Flood Dataset. *Available at: https://s3.eu-central-1.amazonaws.com/corupublic/AAAI_harvey_data/harvey.zip*.

[28] Robin Cole. satellite-image-deep-learning. *Available at: https://github.com/robmarkcole/satellite-image-deep-learning.*

[29] Le Lei Andrew Patterson& Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in neural information processing systems*, vol. 31, 2018

[30] Laurens van der Maaten& Geoffrey Hinton. Visualizing Data using t-SNE. *In Journal of Machine Learning Research, vol. 9, 2008.*

[31] Li Jing, Pascal Vincent, Yann LeCun& Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. *In International Conference on Learning Representations, 2022*

[32] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari& Diane Larlus. Improving the Generalization of Supervised Models. *arXiv preprint arXiv:2206.15369, 2022.*

A Appendix

A.1 Temporal Positive Pairs

In this section of the Appendix, we will visualize the temporal positives extracted from the two datasets we have used in our experiments to train the VICReg model, namely, the fMoW and SpaceNet-7 datasets. The results are illustrated in Figure 2.

A.2 K-Means and KNN Evaluations

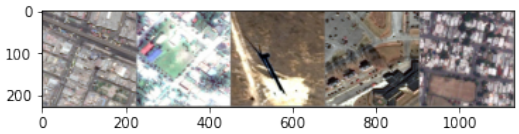
In this section, we will depict the plots of the obtained representations of various contrastive and supervised models trained on SpaceNet-7, fMoW or ImageNet datasets. The results are illustrated in Figure 3. In agreement with the quantitative results of Table 2, these qualitative results also suggest that the supervised models interestingly yield more compact representations. This however, is no surprise as previous work such as [22] also explored this phenomenon and found the supervised models to yield more compact representations in comparison to their unsupervised counterparts.

A.3 EuroSat LandCover Classification Dataset Sample Batch

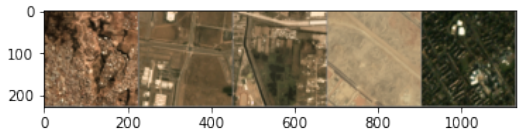
In this section of the Appendix, we present a sample batch of the EuroSat dataset used for our first downstream task along with their labels for the convenience of the reader. The labelled images are depicted in Figure 4.

A.4 Hurricane Harvey Flood Detection Segmentation Results

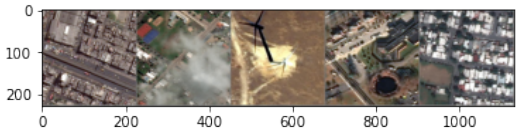
In this section of the Appendix, we visualize the segmentation results after both the fine-tuning and training stages of the Hurricane Harvey Flood Detection Task. The results are presented in Figures 5 and Figure 6 respectively.



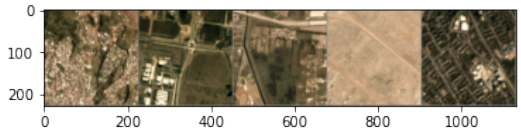
(a) First set of fMoW images of the temporal pairs at time t_1



(b) First set of SN7 images of the temporal pairs at time t_1

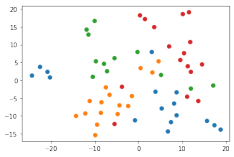


(c) Second set of fMoW images of the temporal pairs at time t_2

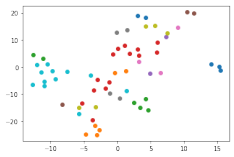


(d) Second set of SN7 images of the temporal pairs at time t_2

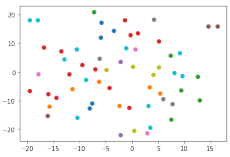
Figure 2: Set of 5 temporal pairs taken at different time instances from both fMoW and SpaceNet-7 datasets. Each pair is aligned column-wise in top and bottom rows. We note that for visual interpretability, most of the augmentations were removed during this visualization except for the random cropping and resizing transformations which explains why each temporal pair denotes a different patch of the original image. We notice the effects of weather, seasonality, development over time or even the perspectives at which the images are captured generate a natural sense of augmentation for our positive pairs.



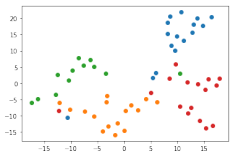
(a) Visualized Representations of VICReg model with Resnet50 under SpaceNet-7 dataset and protocol-1



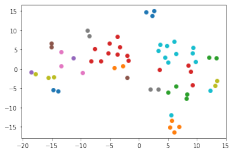
(b) Visualized Representations of supervised model with Resnet50 under SpaceNet-7 dataset and protocol-1



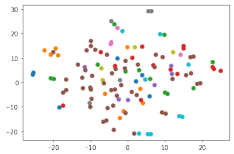
(c) Visualized Representations of VICReg model with Resnet50 under SpaceNet-7 dataset and protocol-2



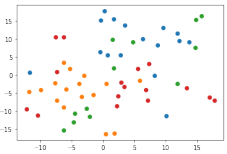
(d) Visualized Representations of supervised model with Resnet50 under SpaceNet-7 dataset and protocol-2



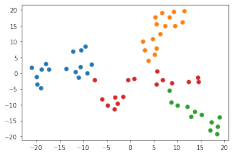
(e) Visualized Representations of VICReg model with Resnet18 under SpaceNet-7 dataset and protocol-1



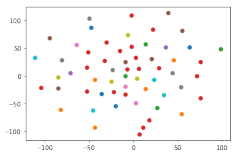
(f) Visualized Representations of supervised model with Resnet18 under SpaceNet-7 dataset and protocol-1



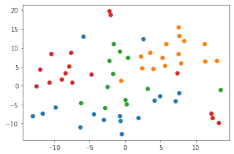
(g) Visualized Representations of VICReg model with Resnet18 under SpaceNet-7 dataset and protocol-2



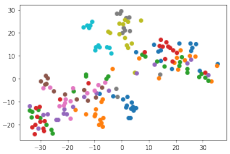
(h) Visualized Representations of supervised model with Resnet18 under SpaceNet-7 dataset and protocol-2



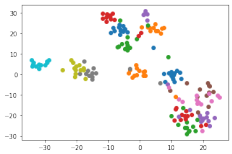
(i) Visualized Representations of VICReg model with Resnet50+TP under SpaceNet-7 dataset and protocol-1



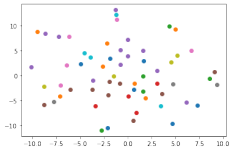
(j) Visualized Representations of VICReg model with Resnet50+TP under SpaceNet-7 dataset and protocol-2



(k) Visualized Representations of VICReg model with Resnet50 under ImageNet dataset and protocol-2



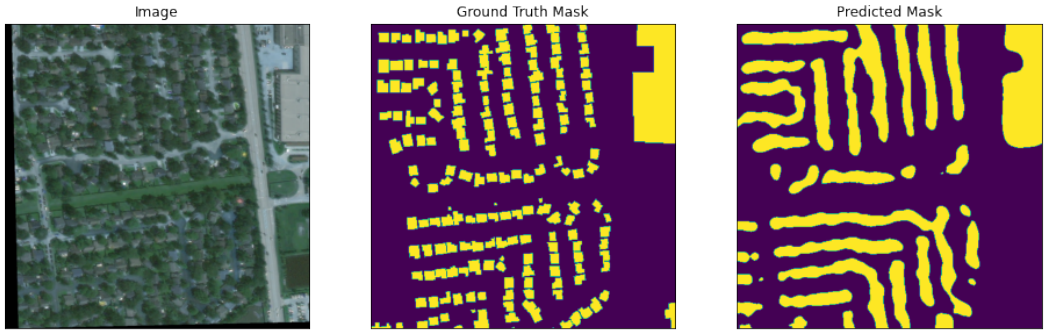
(l) Visualized Representations of supervised model with Resnet50 under ImageNet dataset andpProtocol-2



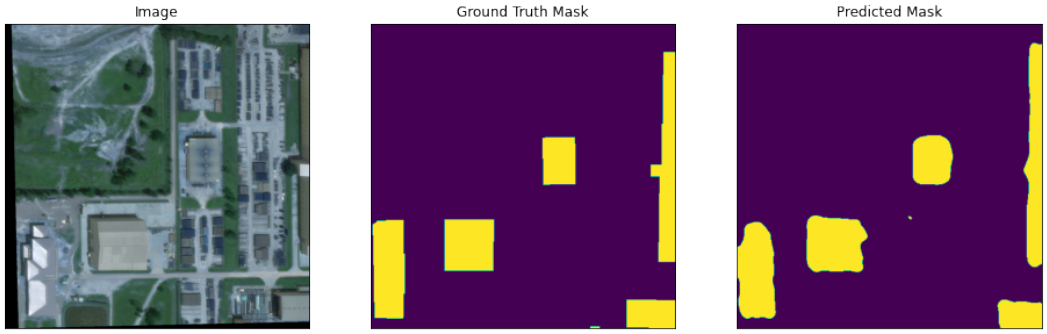
(m) Visualized Representations of VICReg model with Resnet50+TP under fMoW dataset and protocol-2

Figure 3: Visualization of the representations obtained via forward pass of different models under different datasets used. Results are provided both for protocol-1 and protocol-2 where applicable. In line with previous work[22], we qualitatively show that Supervised models tend to yield representations that are much more compact or clustered within classes.

Figure 4: A sample batch of images obtained from the EuroSat landCover classification dataset. The associated labels belong to any one of the following classes:'annualCrop', 'forest', 'herbaceousvegetation', 'Highway', 'Industrial', 'Pasture', 'permanentcrop', 'Residential', 'River', 'sealake'. We also visually notice that the images of the dataset have a ground-truth distance resolution per pixel comparable to that of the SpaceNet7 dataset.



(a)



(b)

(c)

Figure 5: Segmentation results of all of the buildings after the fine-tuning stage. As depicted in the plots, we first use the model at hand with the chosen pre-trained ResNet backbone in order to fine-tune the model first using the training masks consisting of all the buildings in the image. The aim of this stage is to boost the final performance of the model at detecting flooded buildings. Then, the predictions performed by the fine-tuned model to segment all the buildings in the chosen validation images are as given in the plots above.

(a)

(b)

(c)

Figure 6: Segmentation results of the flooded buildings alone after the actual training stage. We train the fine-tuned model on the masks depicting flooded buildings which is our ultimate task. Then, the predictions performed by the fully trained model to segment only the flooded buildings in the chosen validation images are as given in the plots above.