Self-Supervised Learning for Invariant Representations From Multi-Spectral and SAR Images

Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross

Abstract—Self-supervised learning (SSL) has become the new state of the art in several domain classification and segmentation tasks. One popular category of SSL are distillation networks, such as Bootstrap Your Own Latent (BYOL). This work proposes RS-BYOL, which builds on BYOL in the remote sensing (RS) domain where data are nontrivially different from natural RGB images. Since multispectral (MS) and synthetic aperture radar (SAR) sensors provide varied spectral and spatial resolution information, we utilize them as an implicit augmentation to learn invariant feature embeddings. In order to learn RS-based invariant features with SSL, we trained RS-BYOL in two ways, i.e., single channel feature learning and three channel feature learning. This work explores the usefulness of single channel feature learning from random 10 MS bands of 10–20 m resolution and VV-VH of SAR bands compared to the common notion of using three or more bands. In our linear probing evaluation, these single channel features reached a 0.92 F1 score on the EuroSAT classification task and 59.6 mIoU on the IEEE Data Fusion Contest segmentation task for certain single bands. We also compare our results with ImageNet weights and show that the RS-based SSL model outperforms the supervised ImageNet-based model. We further explore the usefulness of multimodal data compared to single modality data, and it is shown that utilizing MS and SAR data allows better invariant representations to be learnt than utilizing only MS data.

Index Terms—Optical-synthetic aperture radar (SAR) fusion, satellite images, self-supervised learning (SSL), unsupervised learning.

I. INTRODUCTION

Computer vision applications to satellite imagery have progressed tremendously over recent years with the availability of multimodal imagery. Nevertheless, the remote sensing (RS) community faces two major problems: 1) labeled data scarcity and 2) dealing with the multisensor and multimodal nature of data obtained from different satellites.

Primarily, RS are broadly divided into optical and synthetic aperture radar (SAR) sensors. On the one hand, optical sensors provide easy-to-read information and vary in spectral and spatial resolution levels, but are negatively affected by weather conditions, such as clouds. On the other hand, SAR sensors provide high resolution and weather independent data but do not provide readily interpretable images. Over the past decade, computer vision in the RS domain has extended its work from supervised learning to unsupervised learning methods for several tasks in MS and SAR data, such as crop monitoring and assessments [1], [2], building mapping [3], [4], [5], scene classification [6], [7], sea-land segmentation [8], change detection [9], [10], and many more. These works used various deep learning networks, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), attention networks, autoencoders, and GANs to solve RS tasks based on spectral, spatial, and temporal information [9], [11], [12], [13], [14], [15], [16], [17], [18]. Some applications use only MS or RGB data, and some use SAR data, such as flood detection. However, since both have their benefits, several research efforts have investigated the possibility of combining data from these two sensors to obtain better quality information [19], [20], [21], [22]. Typically, there are three types of optical-SAR fusion, which are pixel-level, feature-level, and decision-level [19], [23], [24]. However, simply learning the joint features of both modalities has also shown significant advantages in previous works [25], [26].

Although a considerable amount of work has been performed in the RS domain, and raw satellite data are also freely available, the use of a large amount of labeled data remains a challenge due to the expense of human annotation and domain expertise [27]. To address this problem, transfer learning is one of the solutions that we can leverage from the machine learning community. This, in fact, has shown excellent generalization capability in several domains [13], [28], [29]. Unfortunately, however, most ML pretrained models are based on RGB and non-RS images, which are different in spectral and spatial aspects from satellite images. This problem has motivated many researchers in the RS domain to obtain a large amount of labeled land cover data to pretrain a supervised model [30], or to apply unsupervised or self-supervised learning to the RS data. With the challenges associated with labeled data, self-supervised learning has become quite popular in the RS community [25], [26], [31], [32], [33].

In recent years, self-supervised learning (SSL) has shown competitive results compared to supervised learning [34], [35]. Self-supervised learning utilizes the unlabeled data to learn invariant features within the data, and the learnt features are then
leveraged for use on downstream tasks with labeled data. In most common SSL architectures, as in Fig. 1, encoders are trained along with the projection layer, and after training, encoder weights are used with a task-specific classifier or decoder.

SSL can be broadly divided into three categories: 1) pretext task learning, 2) contrastive learning, and 3) distillation networks. Pretext task (PTT) learning is based on proxy task prediction, such as rotation prediction, solving jigsaw puzzles, colorization, etc. [36], [37], [38]. Although PTT learning has shown great advantages in learning different invariant properties, such as rough boundaries and semantic information about the images, it does not generalize well in the latter layers of the network [39]. Meanwhile, contrastive learning involves the use of positive and negative pairs of images, such as in SimCLR [35] or MoCo [40]. This usually leads to an increase in the need for memory requirements and an increase in the computational cost since they depend on a high batch size for the contrastive approach to be most effective. Recently, distillation networks, such as BYOL [34] and DINO [41] have shown the potential of using only positive pairs to learn better representations, ultimately reducing the need for a large batch size compared to other contrastive learning approaches. Although these methods have claimed to add several factors to models that should prevent the common collapse problem in Siamese networks, this is still an active area of research.

In SSL type learning, apart from network architectures, the design of data augmentation, such as random crops, random rotation, color jitters, random greyscale, etc., is a crucial part of the learning process. These augmentations preserve the semantic aspect of the data and provide better invariant representations [42]. The goal of distillation networks, such as BYOL is to learn invariant feature embedding by bringing these augmented views closer. Recent work in the RS domain with SSL is being used for classification, segmentation, or change detection tasks, yet there are very few works which utilize the multimodality of the remote sensing domain [43], [44]. We argue that having the use of multimodal satellite data can provide implicit augmentation to the model when we use MS and SAR data as augmented pairs of images. Motivated by this, we trained a remote sensing-based SSL distillation network (RS-BYOL) against the MS and SAR data to learn invariant spectral and spatial features. Although both data sources vary in terms of resolution and modality, our hypothesis is that feature-level information should remain the same and provide valid inputs to the distillation network process.

In summary, the main contributions of this work are then as follows.

1) We investigate the importance of having RS-based pretrained models compared to traditional RGB or non-RS-based pretrained models. For this, we compared our BYOL-based RS-BYOLs weights with popular ImageNet and MS COCO pretrained weights. We also compare our results with another RS-based contrastive learning model, SeCo [45]. This further investigates the suitability of self-supervised distillation networks for RS-based learning.

2) We explore the usefulness of multimodality for feature learning rather than learning by a single modality. For that, we train RS-BYOL with MS-SAR data and, also, with only MS data, which further confirms the effectiveness of learning invariant features from MS-SAR data.

3) Finally, we also validate single-band feature learning efficiency against three-band feature learning in satellite imaging. In general, each MS band has its own reflectance property to differentiate between geographical features, such as water, land cover, and more. This motivates us to explore single-band invariant feature learning, which we believe can learn powerful features through a model combining three or more bands.

II. RELATED WORK

This section discusses previous relevant work on self-supervised learning and its application in the remote sensing domain.

A. Optical-SAR Fusion and Computer Vision

Computer vision in the remote sensing domain has progressed greatly in both classification, object detection, and segmentation tasks. With that, the number of applications in remote sensing has also increased and is not only limited to scene classification, but also extends to crop monitoring [46], [47], flood detection [48], or building detection [23], [49] and many more. Considering these applications, many supervised and unsupervised learning methods have been proposed over the years, and extensive use of CNNs, attention networks, autoencoders, and GANs has been seen [9], [11], [12], [13], [14], [15], [16], [17], [18].

Each band in MS has its own characteristics, e.g., red-edge bands that are good at detecting vegetation and soil [50], [51], SWIR and NIR bands that are good at detecting water and parsing thin clouds [52], [53], [54], while SAR provides images regardless of weather conditions, which helps in the study of terrain or disaster related tasks [55], [56]. These properties of MS and SAR bands are utilized in many applications, such as crop or weed detection, flood detection, or oceanography, which requires more than just an RGB image. Although SAR has its benefits and provides additional information missed in MS, it...
is highly susceptible to speckle noise, which encouraged researchers to use them jointly. The optical-SAR fusion technique is divided into three levels: these are the 1) pixel level, 2) feature level, and 3) decision level. Where pixel-level fusion is computationally costly and requires a strict registration technique, feature-level fusion uses feature extraction, where feature maps of optical and SAR data are combined, decision-level fusion is less sensitive to registration and combines individually classified images [21], [57]. Many works showcased the different efficiencies of these fusion techniques, such as Rajah et al. who showed a feature-level fusion framework for the detection and mapping of plant species, while Zhang et al. [58] used a framework for classification of land cover in mountainous areas. Multi3Net by Rudner et al. [23] demonstrated pixel-level and decision-level fusion for flooded building mapping. All these works showed that fusing optical and SAR images is advantageous in comparison to using a single modality. However, we believe that looking at optical and SAR images as augmented counterparts and learning joint representations is potentially an excellent source for feature learning.

B. Contrastive Learning and Distillation Networks

Self-supervised learning (SSL) performance has closed the gap with supervised learning [34], [41], which has helped solve the big problem of labeled data requirements for deep learning. Many contrastive and distillation networks have been proposed with increasing interest and performance enhancements. These networks are based on the general class of Siamese networks, which learn feature embedding based on the similarity between two images. Unfortunately, the early Siamese networks often suffered from model collapse to trivial solutions [34], [59]. However, with careful control over the input data, model architecture, and loss function, contrastive, and distillation network approaches have been able to avoid this collapse.

Contrastive learning, such as SimCLR [35], deals with positive and negative pairs of images, with the intention of maximizing the agreement in projections between similar images while minimizing for dissimilar images. Similarly, MoCo [40] leverages the concept of a momentum encoder and a memory bank for negative feature vectors. Although these learning methods show good performance, they are computationally costly due to the use of negative pairs that require large batch sizes and memory requirements.

Meanwhile, distillation networks, such as BYOL [34] and DINO [41] use only positive pairs and have become quite popular in terms of performance and memory efficiency. BYOL and DINO are similarity-based networks whose architecture is inspired by concepts from MoCo [40] and follow a student–teacher approach. Here, teacher encoders do not have error-signal back-propagation; instead, they utilize slow moving averages from the student network weights. This makes them asymmetric and prevents the embeddings from collapsing.

In addition to contrastive and distillation networks, Barlow Twin (BT) [60] has been proposed, which employs a redundancy reduction approach. This model architecture is similar to SimCLR except that it utilizes only positive pairs and the loss is calculated based on a cross-correlation matrix between the projections. BT shows competitive results with BYOL on certain aspects, such as batch size and asymmetry requirements, but BT networks are not robust toward augmentation and require higher projection output. Although this network trains well with a smaller batch size (1024) compared to BYOL, this still requires us to train both sides of the network. This ultimately makes it more computationally costly, where hardware availability is still a constraint.

Although continuous improvement has been seen in the architectures of SSL networks, this area is still being explored thoroughly, with efforts focusing on examining each network layer, the usefulness of stop gradient in collapsing, and investigation of appropriate loss functions. As all network performance fluctuates across different datasets and with different parameters and network settings, it is not yet possible for us to determine which of these models ultimately performs best.

C. SSL and Remote Sensing

Due to the lack of available labeled data in the remote sensing (RS) domain, and the gap which exists with non-RS RGB images in terms of spectral and spatial context, work in the RS domain is now leveraging self-supervised learning to pretrain RS-based models for downstream tasks.

Recent literature has shown the popularity of applying both pretext tasks (PTT) and contrastive learning methods. The work by Vincenzi et al. [32] utilized colorization as a PTT to learn spectral information using multispectral images. Meanwhile, the work by Tao et al. applies instance discrimination-based PTT to RGB remote sensing data, which performs well in downstream tasks with small datasets [61].

Some works use contrastive learning with the application of MoCo-based models to temporal [62] and spatially augmented RGB remote sensing images [33]. SeCo [45] is one such approach, which utilizes the MoCo architecture with the multi-augmentation contrastive learning method taking into account the seasonal contrast in the RGB remote sensing data. Also, quite a few works have attempted to fully exploit both sensor modalities, i.e., multispectral and SAR images, for representation learning with contrastive or distillation networks. One such work uses contrastive multiview coding (CMC) by leveraging multiple high- and low-resolution remote sensing datasets [26]. Chen and Bruzzone [43] showed the use of both modalities using contrastive loss for pixel-level learning. Another work, by Chen and Bruzzone [10], concatenates MS and SAR temporal images to obtain pixel-level discrimination embedding with the distillation network for change detection. Meanwhile, work by Jain et al. demonstrated that single channel features can also provide significant advantages in learning invariant representations for satellite data [25]. Similar results are shown by Montanaro et al. [63], which uses a combination of contrastive learning and the text task with single-channel feature learning. Though it has been seen that similarity-based networks, such as SimSiam do not perform well in the case of remote sensing [25], [63], we
believe that distillation networks provide additional advantages for the learning of invariant spatial and spectral representations.

In addition, all these works highlighted the benefit of learning RS-based representations for many downstream tasks rather than utilizing RGB natural image-based pretrained models. Moreover, those works have shown that SSL has been established as providing a solution to the problem of large labeled data without compromising performance compared to supervised learning. This motivates us to provide better experimental analysis of self-supervised distillation networks for RS representation learning by utilizing both MS and SAR data.

III. METHODOLOGY

This section presents our framework for multimodal satellite image representation learning based on distillation networks in self-supervised learning. This framework builds on the concept of distillation networks and, in particular, the BYOL architecture, but adopts a particular strategy to the application and training process for multimodal data as shown in Fig. 2.

Self-supervised learning is based on the weak smoothness assumption, which states that distortion or the addition of perturbation to an image does not change its labels [64]. This work revolves around the same concept with multiview representation similarity. To put this in another way, since MS and SAR images are two views of the same location, our modeling leverages the implication that having different sensor images does not change the images’ geographical properties.

Let us assume that the MS and SAR images are in practice two views $V_1$ and $V_2$, and that $f(V)$ is a representation of a view. The goal of this work is then to increase the similarity between the representations of the two views, say

$$ f(V_1) \Leftrightarrow f(V_2). $$

As indicated, the most appropriate network for such learning is arguably the siamese networks due to the duality of the data. However, the trivial solution to siamese networks is prone to collapse to constant representations, which led to the rise of contrastive and distillation networks.

Considering that our goal is to learn invariant representations of MS ($V_1$) and SAR ($V_2$), we trained a remote sensing network (RS-BYOL) based on the BYOL architecture. The architecture follows the student–teacher type asymmetry, where students and teachers have identical encoders ($f(T(V))$). Furthermore, the weights are updated only for the student side of the network, whereas the weights of the teacher network are modified based on the moving average of the student weights. In order to add more asymmetry to the network, an extra predictor, i.e., a multilayer perceptron (MLP) head, has been added to the student side of the network. The network compares the L2 normalized projection from the student and teacher sides of the network, which is an output of a predictor head ($q(\theta z)$) and a projection head ($z$), respectively. The projection and prediction heads are two-layer multilayer perceptrons (MLP) of hidden projection size 256 with batch normalization and a nonlinear (ReLU) layer.

The loss is calculated between the L2 normalized prediction head output $q(\theta z)$ from the student side and the projection head output $z$ from the teacher side. Here $\theta$ is the weight of the student network, $\xi$ is the weight of the teacher network, which is an exponential moving average of $\theta$. As shown in (1), $\xi$ is updated by $\theta$ and the decay rate ($\tau$), which in this network is 0.9

$$ \xi \leftarrow \tau \xi + (1 - \tau) \theta. \quad (1) $$

This implicitly provides asymmetry and prevents the model from learning the same latent representations, which in turn prevents collapse.

The loss function for the network is represented as in (2), where $\| \cdot \|_2$ is the normalization of $L_2$, and $t$ and $t'$ are the augmentations applied to the views $V_1$ and $V_2$

$$ L(t(V_1), t'(V_2)) = 2 - 2 \cdot \frac{q(z_{\theta,1}) \cdot z_{\xi,2}}{\|q(z_{\theta,1})\|_2 \cdot \|z_{\xi,2}\|_2}. \quad (2) $$

Fig. 2. (a) MS bands and SAR polarization are two views $V_1$ and $V_2$ respectively, where single channel model utilizes three bands stacked randomly. We utilized 10 MS bands of 10–20-m resolution and discarded 60 m atmospheric bands, for SAR we utilized both polarization, where $i \neq j \neq k$ and represents random three channel whereas $j, m$ represents two polarization in random order with third as repetitive of either $j$ or $m$. (b) BYOL architecture with RS data, which consist of an Encoder ($f$), an MLP projection head ($z$), and an MLP prediction head ($q(\theta z)$). $sg$ represents stop gradient, $\theta$, and $\xi$ are different weights, where $\xi$ is updated based on $\theta$ and $\tau$. Input images do not represent actual band here, but used only for multiple band representation.
Our loss function is the mean square error (mse) between the two projections for the augmented views of the MS and SAR images. The total loss is calculated by interchanging views between the student and teacher network, which results in the total mse as

$$\text{Loss}_{\text{total}} = L(t(V_1), t'(V_2)) + L(t'(V_2), t(V_1)).$$ (3)

This architecture not only prevented the collapsing issue but also provided better performance than some of the contrastive learning approaches and helped reduce the batch size requirement for training. Although this architecture performs well, many research works have shown that different analyses, such as the use of the batch normalization layer, do not help to avoid collapse [65], but rather that the stop gradient is the key to avoiding collapse [59]. In other words, these works show that there is still a need for more analysis in terms of the reasons for the outstanding performance of BYOL-like distillation networks and further need to explore them in other domains, such as remote sensing.

In order to better understand the representation learnt from the RS data, we pretrained our models into two variants: 1) a single-channel model, and 2) a three-channel model. For the single channel model, two views were randomly selected from the MS and SAR bands, that is, one MS band ($V_i$) out of ten bands, i.e., $i = \{1, 2, \ldots, 10\}$, and one SAR polarization ($V_j$), i.e., $j = \{1, 2\}$. The three-channel model utilizes the same approach, except that it draws upon a random three bands from the MS data and generates three-channel SAR images by randomly copying a third channel from the two polarizations. These two views are then randomly augmented with the transformations $t$ and $t'$ before being sent to the encoder.

After pretraining of the model, MLP heads are removed, and we use only student encoder weights for the downstream task, which in our case we evaluate on different remote sensing datasets.

This work demonstrates the usefulness of RS-BYOL, which is trained to learn single and three channel RS-based invariant features. We compare our work with models trained from scratch, random initialization, ImageNet and MS COCO pretrained ResNet50, BYOL-ImageNet [66], and SeCo [45] weights to explore the usefulness of RS-BYOL models against models trained with non-RS data and RS-based contrastive models. For BYOL-ImageNet pretrained weights, we utilized the weights published by Yao Xin [66], the author trained the BYOL architecture with a ResNet50 encoder, with batch size of 128 and 300 epochs.

### IV. Experiment

In this section, we discuss the implementation details of the RS-BYOL models and give details on the datasets and data augmentation used.

#### A. Datasets

We pretrained our models with the Sen12MS dataset [67], which consists of 180 662 triplets of Sentinel-1, Sentinel-2 image patches, and MODIS land cover map images of size 256 × 256 pixels. Whereas patches from Sentinel-1 are a dual-polarization synthetic aperture radar (SAR), and have VV and VH polarization, Sentinel-2 has 13 multispectral (MS) bands. Among the 13 MS bands, in this work, we used 10 bands of 10–20-m resolution and discarded three atmospheric bands of 60-m resolution, and for VV and VH, we used both polarizations. We utilized 50% of the data for our work, that is, 90 K pairs of images. The reason for using 50% data is due to hardware limitation, as a larger dataset adds computational cost.

For evaluation of the representations learnt by the models, we utilized two benchmark datasets for a land cover classification task, i.e., EuroSAT [68] and RESISC45 [11]. EuroSAT consists of 64 × 64 size images and is similar to Sen12MS in terms of sensor and spectral information. It consists of ten land cover classes with 27 000 images. Meanwhile, the RESISC45 dataset consists of high-resolution RGB images of 256 × 256 pixels in size and contains 31 500 images, covering 45 scene classes. We also evaluated our models on 5000 images from Sen12MS.

Finally, we use the 2020 IEEE Data Fusion Contest (DFC) land cover mapping dataset [69] for segmentation evaluation, which is a subset of Sen12MS data and contains 12 228 pairs of labeled Sentinel-1/2 data. Of these, we used 900 images for the linear probing evaluation. The dataset consists of ten land cover classes based on the IGBP classification scheme. Among the ten classes, we excluded “Savanna” and “Snow/Ice,” as these classes are rare in this subset of Sen12MS, and utilized eight land cover classes, that is, forest, scrubland, grassland, wetlands, croplands, urban / built-up, woodland, and water.

#### B. Data Augmentation

We used all the augmentation techniques applied in BYOL [34], except color jitter, as this work implicitly applies color jitters to images through the use of random MS and SAR bands. We also utilized random pixel erasing to provide a stronger augmentation to the images. Thus, our augmentations include random flips, rotation, Gaussian blur, random pixel erasing, and resized crops with bicubic interpolation. We also applied random greyscale to three-band images. An example of a batch of randomly generated augmented images is shown in Fig. 3.

#### C. Experimental Settings and Parameters

For pretraining the model, we used ResNet50 as the backbone encoder and the SGD optimizer with a cosine warm restart scheduler. The warm restart of 10 epochs was given with a multiplier of 2, and the learning rate ranges between 0.002 and 0.2. We utilized a batch size of 32 and trained all our models for 400 epochs with weight decay of 1e-5.

The following subsections explain individual settings and variations for single channel and three channel model variants.
For the three-channel remote sensing-based feature learning model, we utilized two different input techniques. The first technique, RS-BYOL-3, is similar to our single-channel model (RS-BYOL-1), which takes three random MS bands as one view \( V_1 \) and SAR’s VV & VH bands as another view \( V_2 \); we randomly copied the third channel for the third SAR view. Another input technique which we applied is RGB-SJS2, which keeps one view \( V_1 \) as red, green, and blue bands (RGB), and another view \( V_2 \) with one SAR and two random non-RGB MS bands, such as red edge, narrow infrared, or shortwave infrared bands.

We utilized a single Nvidia RTX 2080Ti GPU for pretraining and evaluation of the models.

V. EVALUATION

For the evaluation of RS-BYOL’s learnt representations, we performed a linear probing. That is, we froze the encoder weights and added a simple classifier head. The classifier head was an MLP with a hidden dimension of 128, followed by batch normalization layers, ReLU, and a softmax layer. We trained all the model variants for 100 epochs and utilized cross-entropy loss as our loss function. For downstream classification tasks we compared our RS-BYOL weights with models trained from scratch, random initialization, and ImageNet pretrained ResNet50 and BYOL weights for both single channel and three channel models. We also compared three channel RS-BYOL weights with SeCo weights, which are trained on 1 million RGB remote sensing images with a contrastive learning architecture.

Additionally, in order to evaluate how well representations learnt pixel level information, we also provided an evaluation on land cover mapping with the DFC 2020 Sen12MS dataset. Both single and three channel RS-BYOL utilized the DeepLabV3 model architecture with a ResNet50 encoder. As was the case with linear probing, we froze the weights of the encoder and trained only the decoder. We also compared three channel RS-BYOL weights with an MS COCO based model trained on DeepLabV3 and SeCo weights to analyse the benefit of using multimodal MS and SAR data learning with an SSL distillation network.

A. Single Channel Models

Fig. 4 shows the F1 score for the bandwise evaluation on the EuroSAT Full, EuroSAT 7 K, and Sen21MS 5 K datasets. To evaluate the ImageNet trained ResNet50 and BYOL model, we copied single band information into three channels, which is the standard solution available for greyscale or single band data. It is evident that RS-BYOL-1 and RS-BYOL-1 w/150 K perform almost similarly, but RS-BYOL-1 w/150 K outperformed all other versions in the EuroSAT full and Sen12MS 5 K datasets. Especially in the SAR bands, the gap between RS-BYOL-1 w/150 K and other versions is quite significant.

With these results, it can be seen that models pretrained in a supervised fashion with a large amount of data, such as ImageNet, show average performance. At the same time, a slight increase in remote sensing data leads to a competitive result. This opens the possibility of better representation learning in the remote sensing domain with larger datasets in a self-supervised fashion. However, the performance for a subset of EuroSAT, that is, EuroSAT 7 K, shows different results, as RS-BYOL-1

**1) Single Channel Models:** For single channel feature learning, RS-BYOL-1 was trained with random bands from MS as view \( V_1 \) and a random selection of VV and VH from SAR as view \( V_2 \).

We trained RS-BYOL-1 in multiple variations to look at the impact of each RS-BYOL’s representation learning. These variations were: 1) increasing batch size, 2) increasing dataset size, and 3) utilizing only the multispectral dataset.

Most of the best performing SSL networks, such as BYOL, DINO, and MoCo utilize much larger batch sizes than 32 to attain the maximum performance; thus, we also trained RS-BYOL-1 with 64 batch size. This variation was made to analyze the impact of batch size on feature learning. As our dataset is not as large as the datasets on which typically large networks are trained on, making use of a large batch size would likely not be a benefit for RS-BYOL-1 in practice. This upper batch size of 64 was influenced by our hardware limitations.

Furthermore, we analyze the impact of training with different dataset sizes on RS-BYOL-1. For this training, we increased the size of the dataset by 30% using 150 K pairs of Sen12MS.

Finally, to better understand the usefulness of SAR data in learning invariant representation, we also trained our model with a single modality, that is, MS data only. This study examines the benefit of using multisensor data for remote sensing representation learning.

Each of these variations was trained with the same architectural and parameter settings, that is, encoder, optimizer, scheduler, epochs, and batch size, except for the case of the 64 batch size variation.

**2) Three Channel Models:** For the three-channel remote sensing-based feature learning model, we utilized two different input techniques. The first technique, RS-BYOL-3, is similar to our single-channel model (RS-BYOL-1), which takes three random MS bands as one view \( V_1 \) and SAR’s VV & VH bands as another view \( V_2 \); we randomly copied the third channel for the third SAR view. Another input technique which we applied is RGB-SJS2, which keeps one view \( V_1 \) as red, green, and blue bands (RGB), and another view \( V_2 \) with one SAR and two random non-RGB MS bands, such as red edge, narrow infrared, or shortwave infrared bands.

We utilized a single Nvidia RTX 2080Ti GPU for pretraining and evaluation of the models.

**Fig. 3. Example: A batch of randomly augmented three channel MS and SAR images. (a) MS. (b) SAR.**
Fig. 4. Bandwise linear probing (F1 Score) for single channel models on EuroSAT and Sen12MS Datasets. (a) EuroSAT Full. (b) EuroSAT 7K. (c) Sen12MS.

outperforms RS-BYOL-1 w/150 K with a performance gap of 1–3% for most bands. To better understand if these results are significant, we also performed a t-test for the EuroSAT results, but found no significant differences between RS-BYOL results, whereas bands with the lower resolution of 20 m or SAR images show a more significant gap in performance with RS-BYOL-1. This may also be due to the small data size of EuroSAT 7K and Sen12MS, which is not surprising as training from scratch is strongly impacted by data size. This also confirms the usefulness of RS pre-trained models, as many applications, such as crop monitoring or oceanoigraphy require more spectral information in the RS domain than is available in RGB images.

Table I illustrates that, for pixel-level information, increased batch size does aid performance. This can be seen in the mean intersection over union (mIoU) and actual accuracy (AA) values of each multispectral band. However, the overall difference in results in comparison to other model versions is not very large. These results also reveal that the model performs well on both SAR and MS data, as the SAR results are also competitive with MS data. It also needs to be taken into account that these results are highly skewed due to water class detection. The accuracy of the water pixels is around 90% and above for all bands in the DFC 2020 dataset. This can be seen in Fig. 6, since water classes are well predicted by the three variations of RS-BYOL-1 for the NNIR and SWIR bands. This can be assumed to be due to the bands’ reflection property, since NNIR and SWIR are both able to detect water pixels more profoundly. Meanwhile, the classwise performance analysis on each band in Fig. 7 shows that almost all bands show good results for water pixels compared to other classes. In addition, notably, almost all bands show similar performances for each class, including SAR bands.

B. Three Channel Models

Table II presents the F1 score of our three-channel model evaluation on the EuroSAT Full, EuroSAT7K, Sen12MS, and RESISC45 datasets. We compared the RS-BYOL-3 weights with the ResNet50 model trained from scratch, the ImageNet trained ResNet50, and the BYOL and SeCo weights. For all
Fig. 5. ROC-AUC curve for Non-RGB bands RE5, RE5-6-7, SAR (VV), and SAR (VV-VH-VV). The comparison among models for a single channel is analogous to the results in Fig. 4. For three channel model, as the performance gap in Table II for RE5-6-7 is not too large, the AUC for RS-BYOL-3 is also almost equivalent to SeCo and ImageNet trained models. However, as for SAR (VV-VH-VV), the gap is quite significant, and this can be seen in the ROC curves as well. (a) EuroSAT 7 K on RE5 band. (b) EuroSAT Full on RE5 band. (c) Sen12MS on RE5 band. (d) Sen12MS on SAR (VV) band. (e) EuroSAT 7 K on RE5-6-7 band. (f) EuroSAT Full on RE5-6-7 band. (g) Sen12MS on RE5-6-7 band. (h) Sen12MS on SAR (VV-VH-VV) band.

Fig. 6. DFC Segmentation linear probing prediction for single channel model with NNIR and SWIR11 band. Results are skewed due to higher water pixel in the data.

three datasets, we evaluated RS-BYOL-3’s on red, green, and blue (RGB) bands, and for non-RGB bands, we utilized Red Edge bands, i.e., RE5-RE6-RE7 and SAR (VV-VH-VV) bands in the Sen12MS dataset.

The results in Table II show that ResNet50 trained from scratch performs better than pretrained models when the size of the data set is large enough. This is not a surprise, as the size of the dataset is the limiting factor in training from scratch, leading to the rise of the concept of transfer learning. Although it can be seen that for RGB images, SeCo shows competitive results with RS-BYOLs, and also outperforms ImageNet trained ResNet50 and BYOL, except in the case of the Sen12MS datasets, where RS-BYOL-3 performed better. The good performance of ImageNet and SeCo on RGB images is not particularly surprising, as they are predominantly trained on large RGB image datasets. However, for non-RGB-based images, such as RE5/6/7 and...
and SAR, RS-BYOL-3 performs better than ImageNet and SeCo trained models. This emphasizes the need for RS-specific models, where the data are typically not RGB images, but MS or SAR images. In fact, it is notable that while ImageNet-based models outperformed the RS-based model for RGB images, the difference in results was not statistically significant and could potentially be reduced with further research on self-supervised learning.

In the case of RESISC45, the performance gap between ImageNet-supervised and RS-BYOL-3 weights is due to very high-resolution RGB data, while our RS-BYOLs are trained with low-resolution data. This particularly applies to SeCo, as their trained weights are also based on low-resolution Sentinel-2 data. However, SeCo does perform better than RS-BYOL. This evaluation suggests that SSL type learning requires more research to generalize the models across the RS domain. One solution to such a problem is to utilize both low- and very-high-resolution satellite images. However, this needs to be further evaluated on datasets other than the Sen12MS subset, as RS-BYOLs are pretrained on the Sen12MS dataset, which might make these results skewed.

The classification and segmentation results for the single-channel models in Fig. 4 and Table I report competitive performance compared to any three channel feature learning weights in Tables II and III, including models pretrained on ImageNet and MS COCO and SeCo weights. On the basis of these results, we argue that single bands can also provide good invariant representations in remote sensing, compared to the popular notion of having a model with three or more bands.

![Image](43x592 to 284x726)

Fig. 7. Classwise, performance boxplot on each band across single-channel RS-BYOLs.

### VI. Discussion

This work proposes RS-BYOL which utilizes the SSL distillation architecture (BYOL) in the RS domain with the use of multimodality, that is, MS and SAR data. We found that multimodality enhances the performance of the SSL-based learning due to the stronger variance in data, which makes representations more invariant. The results in Fig. 5 also show that when random MS and SAR bands are utilized for single channel feature learning, it can provide competitive results compared to the common notion of three channel feature learning.

The efficiency of RS-BYOLs are further validated by the t-SNE plots provided in Fig. 9 for ten classes of the EuroSAT 7K dataset. We compared the feature embedding from ImageNet trained ResNet50 and BYOL, SeCo, RS-BYOL-3, and RS-BYOL-3 RGB-S1S2 on RE5-6-7 band images. For single channel RS-BYOLs feature embedding, we utilized RE5 band images. It can be seen that RS-BYOLs trained with the multimodal approach form tight clusters in the embedding space compared to ImageNet and SeCo variants. However, SeCo shows better clustering than ImageNet, which further shows the benefit of using RS based data with SSL types of learning. In addition to this, it can be seen that RS-BYOLs trained with MS dominant data did not form clean clusters on embedding. This underperformance suggests that in order to learn invariant representation in low-resolution data, it is advantageous to leverage multiple modalities.

We also looked at the full gradient class activation maps (CAM) to understand where RS-BYOLs, SeCo, and ImageNet trained ResNet50 and BYOL weights focus for class prediction in EuroSAT RGB and Blue band images. In Fig. 8, it can be seen that the single channel feature embedding strongly focuses on class specific regions along with the semantic separation.
between two classes, such as river and land area. However, CAM for single-modality RS-BYOL-1 MS only shows out of region focus, which aligns with under performance for classification and segmentation tasks. Similar analysis can be seen in three channel MS dominant models, i.e., RS-BYOL-3 RGB-S1S2, which shows that while making decision, its region of focus is spread across the image.

The CAM for ImageNet, SeCo, and RS-BYOL-3 weights on river class are distributed across the image and also generally fail to find the stronger region of focus except for BYOL-ImageNet. For industrial classes, all three weights have stronger focus on a smaller region of the class, despite the fact that 90% of the image covers the class. However, in contrast, it should be noted that the same model is very precise for the highway class for ImageNet and RS-BYOL-3. While these are illustrative examples, we believe that such analysis can provide insight into the learning process.

Both single channel and three channel RS-BYOLs performed well in downstream classification tasks, but need to improve at segmentation tasks, at least in the case of three channel RS-BYOLs. However, it should be noted that we did not evaluate these with a fine-tuning process, which might provide performance enhancement. This limitation can be further explored in the future to provide generalized network weights.

Finally, we note that RS-based SeCo showed competitive results with RS-BYOLs, which shows the potential of SSL approaches in the RS domain to overcome the scarcity of labeled data. However, we argue that, in terms of computational efficiency, the BYOL-based RS-BYOLs are more efficient than contrastive learning based. This is due to the fact that training with BYOL style networks require only positive pairs, and use of stop gradient, which in return requires less batch size and memory requirements without compromising the performance.

VII. CONCLUSION

In this work, we applied the distillation network concept to build and analyze single channel and three channel features learning for MS and SAR data. MS and SAR data provide...
implicit augmentation with spectral and structural-level differences. This work showed that using MS and SAR as different views for network training provides better representation learning than utilizing MS data alone.

We also showed that RS domain-based learning provides better results than non-RS-based learning by comparing our results with results based on natural RGB datasets, such as ImageNet and MS COCO pretrained models. Even though we trained our RS-BYOL with only 90 K instances, in comparison, for example, with ImageNet (14 million), MS COCO (328 K) dataset or SeCo (1 million), we achieved competitive results even with single channel feature learning alone. Another noticeable aspect of RS-BYOLS is that we trained the network with only 32 batch size, unlike other networks trained on 256 or more. This may open the path for further research work in satellite data, aiming at achieving better generalization across the domain.

In continuation of our work, we will further explore other computationally efficient SSL networks and other RS-based pretrained networks. Additionally, we would like to evaluate our models on further RS-based datasets in order to better understand the learnt representations.

REFERENCES

[1] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, “Machine learning applications for precision agriculture: A comprehensive review,” IEEE Access, vol. 9, pp. 4843–4873, 2021.
[2] M. Ouhami, A. Haifane, Y. Es-Saady, M. El Hajji, and R. Canals, “Computer vision. IoT and data fusion for crop disease detection using machine learning: A survey and ongoing research,” Remote Sens., vol. 13, no. 13, 2021, Art. no. 2486.
[3] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, “MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 6169–6181, Jul. 2021.
[4] C. Robinson et al., “Fast building segmentation from satellite imagery and few local labels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2022, pp. 1463–1471.
[5] J. Kang et al., “DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4706315.
[6] S. Mei, K. Yan, M. Ma, X. Chen, S. Zhang, and Q. Du, “Remote sensing scene classification using sparse representation-based framework with deep feature fusion,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 5867–5878, 2021.
[7] P. Deng, K. Xu, and H. Huang, “When CNNs meet vision transformer: A joint framework for remote sensing scene classification,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 802035.
[8] R. Li et al., “DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 11, pp. 3954–3962, Nov. 2018.
[9] L. Khelifi and M. Mignotte, “Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis,” IEEE Trans. Geosci. Remote Sens., vol. 8, pp. 126385–126400, 2020.
[10] Y. Chen and L. Bruzzone, “Self-supervised change detection by fusing SAR and optical multi-temporal images,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2021, pp. 3101–3104.
[11] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and State of the Art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[12] M. Gong, H. Yang, and P. Zhang, “Feature learning and change feature classification based on deep learning for ternary change detection in SAR images,” ISPRS J. Photogramm. Remote Sens., vol. 129, pp. 212–225, 2017.
[13] R. P. de Lima and K. Marfurt, “Convolutional neural network for remote-sensing scene classification: Transfer learning analysis,” Remote Sens., vol. 12, no. 1, 2019, Art. no. 86.
