Self-supervised learning for joint SAR and multispectral land cover classification

Antonio Montanaro, Diego Valsesia, Member, IEEE, Giulia Fracastoro, Member, IEEE, and Enrico Magli, Fellow, IEEE

Abstract—Self-supervised learning techniques are gaining popularity due to their capability of building models that are effective, even when scarce amounts of labeled data are available. In this paper, we present a framework and specific tasks for self-supervised training of multichannel models, such as the fusion of multispectral and synthetic aperture radar images. We show that the proposed self-supervised approach is highly effective at learning features that correlate with the labels for land cover classification. This is enabled by an explicit design of pretraining tasks which promotes bridging the gaps between sensing modalities and exploiting the spectral characteristics of the input. When limited labels are available, using the proposed self-supervised pretraining and supervised finetuning for land cover classification with SAR and multispectral data outperforms conventional approaches such as purely supervised learning, initialization from training on Imagenet and recent self-supervised approaches for computer vision tasks.

Index Terms—Self-supervised learning, synthetic aperture radar, multispectral images, land cover classification

I. INTRODUCTION

Deep learning is nowadays an established way of designing powerful models that are able to effectively solve problems in a wide variety of fields, from natural language processing [1], to computer vision [2] and remote sensing [3]. The most striking successes, such as surpassing human performance on image classification, are due to supervised learning, where huge annotated datasets are used to learn end-to-end models addressing a specific problem. However, supervised learning has been increasingly under scrutiny due to such data requirements. Unfortunately, not all the domains can rely on huge datasets like ImageNet. This is the case of remote sensing imagery, where carefully annotating satellite images requires domain experts, and doing so for large amounts of data can be expensive and error-prone.

The emerging field of self-supervised learning (SSL) addresses this data bottleneck, studying techniques that can be used to train deep models to extract features that are relevant to the problem of interest, without requiring labeled data [4]. Even more interestingly, such models can then be finetuned in a supervised fashion with a small amount of labeled data to reach performance close to or better than that of purely supervised models trained on huge amounts of labels.

This paper addresses the problem of developing SSL techniques that are effective for the land cover classification problem in remote sensing. This is not a trivial objective since there are several challenges that are unique to this problem and find no correspondence in the computer vision field. In particular, in Earth observation several imaging modalities (e.g., optical and radar) can be used to acquire a scene of interest, and it is not obvious how to train a model that is capable of exploiting both; in this paper we address the problem of using multiple imaging modalities, namely multispectral and synthetic aperture radar (SAR) images, to infer the land cover classes.

Recent works in the context of the 2020 IEEE GRSS Data Fusion Contest [5] have shown difficulties in building end-to-end models based on deep learning for land cover classification with both SAR and multispectral data. In particular, the results of the contest show a clear prevalence of traditional machine learning methods like random forests, as well as issues in achieving better performance when adding SAR to the optical images. This is a symptom of deep models being unable to extract high-quality features due to a variety of reasons such as difficulties in integrating two widely different imaging modalities, lack of large labeled datasets, pretraining techniques (e.g., classification on Imagenet) suffering from large domain gaps with respect to remote sensing data, and more.

In this paper we propose a method, named Spatial-Spectral Context Learning (SSCL), which is composed of a generic modular architecture for neural networks and two self-supervised pretraining approaches, that allow to effectively train models for multichannel data having an arbitrary number of channels representing imaging modalities (multispectral bands, SAR polarizations, etc.). SSCL is a universal framework that can be used whenever the available input data have many channels and it is more effective than transferring models from computer vision datasets due to the large existing domain gaps. For example, image classification on ImageNet deals with RGB instead of multichannel images, its classes are mostly object-centric and require reasoning about spatial geometry rather than spectral characteristics of materials. Instead, the self-supervised tasks in SSCL are explicitly designed to account for the existence of multiple channels with possibly very different representations, and promote learning a model of the correlations across channels, as the spectral properties of materials can be jointly inferred from the visible and infrared spectral bands in multispectral images, and from the microwave wavelengths captured by SAR. Since the classes of interest in problems such as land cover classification involve discriminating materials, this multichannel approach is more effective at extracting features for remote sensing problems than the more geometric one promoted in computer vision.

In summary, our novel contributions include:
• SSCL: a framework to design neural networks, comprising a high-level architecture design and self-supervised training tasks, for the land cover classification task with multiple imaging modalities; specifically SAR and multispectral are analyzed in this paper;
• UniFeat (Uniform Features): a self-supervised pretraining method for the first portion of the SSCL architecture; it employs contrastive learning to promote similarity between the feature vectors of two patches representing the same area from different input channels, thus bridging the gap between imaging modalities;
• CoRe (Context Reconstruction): a self-supervised pretraining method for the whole SSCL architecture; it degrades the input images according to specific processes designed so that learning to restore the image from such degradations promotes features that capture the spectral information and thus correlate with land cover classes, while also promoting high spatial resolution;

Extensive experiments show how the proposed approach is more effective than approaches that are commonly used such as purely supervised learning, pretraining from ImageNet and a recent self-supervised approach that is state-of-the-art for computer vision tasks (SimSiam [6]), when labels are scarce. We also show how effective SAR-multispectral fusion is achieved, whereby adding SAR to multispectral images actually improves the overall performance.

II. RELATED WORK

Recently, many researchers have started investigating SSL approaches that do not require labelled data [4]. In SSL, the training task is created out of the unlabelled input data, leveraging their underlying properties. This allows neural networks to learn useful data representations that can be then exploited in any downstream task, while eliminating or reducing the need for annotated data. Combining a pretraining phase with SSL and supervised finetuning has been shown to be significantly more effective than pure supervised learning [7], when the number of annotated samples is small and the models can exploit huge unlabeled datasets.

Several different SSL techniques have been presented over the last years. The most popular approach consists in learning to capture relevant image features by solving a pretext task. A wide variety of pretext tasks have been proposed [8]. Some of them involve geometric transformations such as guessing the rotational angle of an image [9] or solving jigsaw puzzles [10], others consider generation-based tasks such as image inpainting [11] and colorization [12]. As discussed in [13] and [14], choosing the best pretext task is challenging and can depend on the target downstream task, since some transformations are suitable for some tasks but perform poorly for others. In Sec. III-A we also emphasize that, while some pretext tasks based on geometric transformations can be effective for object recognition in photographs, they are not suitable for land cover classification tasks in remote sensing because they are not focused on capturing the spectral features that discriminate material properties.

More recently, contrastive learning is emerging as a new appealing paradigm for SSL [15]. This approach aims at embedding augmented views of the same input close to each other, while trying to separate embeddings from different inputs. Augmenting the same input in two different ways can be seen as a way to generate multiple data that are known to belong to the same class (albeit which class is unknown) and promoting similarity among them tends to cluster their representations in the feature space. All the methods following this approach employ a siamese network [16] and a contrastive loss [15], but they differ from each other mainly in the way they collect negative samples. In SimCLR [17], the authors perform training with large batch size in order to have a contrastive loss more representative of the image domain. Instead, MoCo [18] considers contrastive learning as dictionary look-up, where the keys are the embeddings of the input samples, and proposes to employ a dynamic dictionary with a queue and a moving-averaged encoder, decoupling the dictionary size from the batch size. A different approach is proposed by SimSiam [19], which reduces the batch size by not collecting any negative samples while still reaching comparable or better results than previous techniques. Other similar approaches can be found in [20], [21].

Remote sensing is strongly affected by limited data availability, where publicly available datasets are several but sparsely annotated. Therefore, supervised deep learning approaches are severely hampered in this context. For this reason, many deep learning methods for remote sensing employ models pretrained on large-scale computer vision datasets [22], [23]. Even though this can reduce the need for large annotated datasets, domain gap issues may arise since pretrained models are tuned on RGB photographs from general purpose datasets used in computer vision. Instead, remote sensing applications usually consider multispectral or hyperspectral images, since the spectral information is important in many inference problems. A model pretrained on RGB images can thus limit the full exploitation of the spectral information. In addition, these domains also differ with respect to the image content: images from many well-known computer vision datasets (e.g., ImageNet) are usually object-centric, while in remote sensing imagery there may be various complex objects in a single scene.

In order to overcome these issues, a limited number of works have started to explore using SSL approaches in remote sensing applications, in particular for land scene classification as downstream task. In [24], the authors propose to use colorization as pretext task for remote sensing imagery, leveraging the spectral bands to recover the visible colors. Instead, in [25] the authors compare three different SSL techniques, namely image inpainting, relative position prediction and instance discrimination, showing that the latter one provides better performance for scene classification as downstream task. Another work [26] extends the constrastive approach proposed by MoCo to remote sensing imagery, defining the augmented views as randomly shifted patches of the same image.

However, little attention has been paid to develop self-supervised deep learning models that can effectively combine information from different spectral channels or sensing modalities, such as multispectral and SAR. To the best of the authors’ knowledge, Contrastive Multiview Coding (CMC) [27] is the
In this section we present the proposed approach to land cover mapping from joint SAR and multispectral imagery, which we call Spatial-Spectral Context Learning (SSCL). At a high-level, SSCL processes an input stack of images composed of $C$ channels, representing multispectral bands and multiple SAR polarizations.

The main novelty of the proposed method lies in the development of self-supervised pretraining strategies, that are able to train feature extractors for the land cover classification task without the need of ground truth labels. If labeled data are available, further supervised finetuning can be performed, and, thanks to effect of pretraining, superior performance can be achieved with respect to not using the self-supervised step.

The proposed self-supervised approach comprises two stages of pretraining, accomplishing different goals. First, it promotes feature robustness with respect to variations in the imaging modality (i.e., SAR or optical). Second, it creates a latent space where material features can be easily separable so that dissimilarities in this space correlate with different land cover labels, while not having access to the true labels themselves. These two stages are trained sequentially, so that the latter benefits from the increased robustness learned by the former. Sec. III-A describes the self-supervised tasks in detail.

An additional important concept that we introduce regards the overall neural network architecture, which is illustrated in Fig. 1 State-of-the-art semantic segmentation models are often developed for single-band or RGB images. It is important to carefully adapt them to the scenario where multiple channels, possibly from multiple imaging modalities, are available. For this reason, we also present a high-level design of a neural network architecture. In this design, a preprocessing stage, composed by a few convolutional layers, acts on the individual channels, sharing its model weights across them. The goal is to slowly extract features from the single channels themselves, before merging them. Slow, single-channel feature extraction, compared to immediate fusion, allows to build a richer feature space and ties into the working of the first stage of self-supervised pretraining, which promotes a convergence of the statistics of the various channels to reduce their domain gap.

### A. Self-supervised pretraining tasks

The objective of self-supervised pretraining is to learn a model that is capable of extracting features which cluster in their space according to the true land cover labels. However, this must be done without the knowledge of such labels, and, as such, it must exploit some internal information of the data distribution. If this objective is accomplished, further supervised training of the model via the labels allows to achieve increased efficiency with respect to a method that does not employ the pretraining stage, i.e., provide higher accuracy for an equal amount of labeled data, or reduce the data requirements for a given target performance. In the context of land cover mapping from joint SAR and optical images, it is imperative to address two issues that are not typically found in classic SSL computer vision problems [33]. For this reason, our SSCL method comprises two novel self-supervised pretraining steps, which we call Unifeat and CoRe and that are described in the following.
1) UniFeat: contrastive uniforming of sensing modalities: The first issue is the multi-channel nature of the input and the domain gap that exists between the channels, and, particularly, between SAR and optical images due to coherent and incoherent imaging modalities. A smaller gap also exists among the spectral bands of a multispectral image. Since the same scene is being imaged across the modalities, it is desirable for the features that are derived to be robust to low-level variations which do not carry discriminative information to infer the class label. Examples of such low-level nuisances can be the different noise characteristics of each channel, the local patch statistics, and so on. Promoting similarity of low-level features across the input channels can help bridge the domain gaps, and avoid large distances between points in the feature space representing the same class. This is the goal of the first self-supervised task we propose, namely UniFeat, depicted in Fig. 2a This task addresses the pretraining of the single-channel feature extractor depicted in Fig 1. We consider the features extracted by the single-channel encoder, consisting in one vector with $F$ features for each spatial location $(i, j)$ and each channel $c$. We use a contrastive learning approach where we promote similarity between the feature vectors of two patches representing the same area from different input channels. Conversely, dissimilarity is promoted if the patches do not represent the same geographical area. Several contrastive losses have been studied for this kind of tasks in computer vision problems [15]. We choose to follow the SimCLR approach [17], where we consider the single-channel feature extractor as the base encoder $f(\cdot)$ and we introduce an additional projection head $g(\cdot)$ that maps the output features of the single-channel encoder to the space where the discriminative loss is applied. Notice that, contrary to the base encoder adopted in SimCLR which targets whole-image classification, the proposed single-channel encoder does not pool all the feature vectors of the patch into a single representation to be further projected, but rather produces a pixel-wise mapping of the input. This promotes features with higher spatial resolution, as shown in Sec. IV-B which is particularly useful for semantic segmentation tasks. Also notice that the projection head depicted in Fig. 2a is removed after pretraining.

Given a minibatch of $N$ image patches, we define two correlated views $x_{c_1}^i$ and $x_{c_2}^i$ of the same input patch $x_i$ in the minibatch by randomly selecting two channels $c_1$ and $c_2$. We then promote similarity between their feature representations by minimizing the Normalized-Temperature Cross-Entropy (NT-Xent) loss [34], defined as:

$$
\ell(c_1, c_2) = \frac{1}{N} \sum_{i,j,k} \log \frac{\exp(sim(z_{(i,j),k}^{c_1}, z_{(i,j),k}^{c_2})/\tau)}{\sum_{l \neq k} \exp(sim(z_{(i,j),l}^{c_1}, z_{(i,j),l}^{c_2})/\tau)}
$$

where $z_{(i,j),k}^{c_1}$ is the value of $z_k^{c_1}$ at spatial location $(i,j)$, $z_{(i,j),l}^{c_2}$ is the value of $z_l^{c_2}$ at $(i,j)$, $z_l^{c_2}$ is a view of the input image $x_i$ (i.e., $z_l^{c_2}$ corresponds to $x_{c_1}^i$ or $x_{c_2}^i$), $\text{sim}(u, v) = \frac{u^T v}{\|u\|\|v\|}$ is the cosine similarity between the feature vectors $u$ and $v$, and $\tau$ is a temperature hyperparameter which controls the rate of convergence. Notice that this task is applied not only to promote similarity between SAR and optical but also between different optical bands.

Since this pretraining task is applied to the outputs of the single-channel encoder, a relatively shallow preprocessor, the feature space is still mostly affected by low-level image characteristics, as desired. However, we remark that UniFeat can also learn how to cluster the feature space according to materials, albeit weakly, in addition to promote uniformity across sensing modalities. In fact, if a single material were present in all the pixels of each patch, UniFeat would promote similarity between multiple realizations of the same class and dissimilarity between different classes, just like methods for image classification do when contrasting different augmentations [15]. However, we argue that this learning is only weak in our scenario since it is highly likely that patches contain more than one material, thus introducing noise in the form of false positives.

2) CoRe: context reconstruction to promote material features: The second issue we address is also something that is specific to the remote sensing problem we are tackling. The well-studied image classification task in the computer vision literature is mostly concerned with classes that represent objects. As such, properties like shape and geometric appearance are extremely important for class discrimination. This is in contrast with the land cover mapping problem we are addressing, where the class label is mostly related...
to the spectral properties of the scene, and only weakly to its geometric appearance. This suggests that the recent self-supervised approaches used in computer vision with contrast views obtained via geometric augmentations (e.g., rotations, distortions, etc.) can only have limited success in extracting features representing material useful for land cover mapping.

For this reason, we propose CoRe (Context Reconstruction), depicted in Fig. 2b, a pretext task that can be solved in a self-supervised manner and whose solution promotes features that capture material properties and thus cluster according to land cover labels. In this pretext task, the input image is first corrupted using a given degradation process, then the network learns to reconstruct the clean image by minimizing the $\ell_2$ or $\ell_1$ distance between the output of the network and the original image. In contrast to UniFeat, which only pretrained the early layers of the network, this task pretrains the entire architecture of Fig. 1. Notice that a projection head with $C$ output channels is used during pretraining and then discarded, to be replaced with the actual head estimating the class probabilities. The input degradation process, also depicted in Fig. 3, consists in:

- **Channel dropout**: randomly dropping a number of input channels (multispectral bands and/or SAR polarizations);
- **Cutout**: delete small patches at randomly placed locations;
- **Gaussian blur**.

By solving this reconstruction task, the network is able to learn features that accurately represent the spectrum, which is highly informative for material discrimination, in particular thanks to the random band dropout. The additional cutout and blurring also add robustness, improving resilience to noise, avoiding convergence to trivial solutions and forcing the network to reason across spatial neighborhoods due to the missing regions. We remark that it might happen that different channels have different spatial resolutions (e.g., in a Sentinel 1-2 fusion problem, the multispectral bands can have resolutions of 10m, 20m or 60m, and 10m or more for SAR). In the case where all the channels at higher resolutions are dropped, the pretraining task becomes an inter-band super-resolution problem, which further promotes the emergence of features with high spatial resolution. Additionally, in a SAR-optical fusion setting, the task also requires to predict one modality from one other, possibly without having any input examples of the target modality and this further enhances the creation of a shared feature space across modalities.

### IV. Experimental Results and Discussions

This section presents experimental results showing the effectiveness of the proposed SSCL approach on a recent dataset for land cover classification jointly exploiting SAR and multispectral images. We first present the main results in terms of segmentation accuracy, both without using the label information (only self-supervised training) and fitting a linear classifier to the representations produced by the models, and after supervised fine-tuning. After that, we examine the various components of the proposed method to understand their relative contribution to the final performance. We also report results for two different state-of-the-art segmentation networks to show that the proposed approach is mostly agnostic to the specific neural network architecture, once combined with the single-channel feature extractor.

#### A. Dataset Description and Experimental Setting

We test the proposed SSCL method on the dataset used for Track 2 of DFC2020 challenge [5] organised by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society, which is a subset of the SEN12MS dataset [35]. The DFC2020 dataset for Track 1 has labels with low-resolution maps generated by the MODIS instrument, while Track 2 additionally discloses 986 samples with semi-manually annotated high-resolution maps.

The input images are acquired by 2 sensors: Sentinel 1 (S1) SAR [36] with 2 channels corresponding to the VV and VH polarizations and Sentinel 2 (S2) multispectral [37] with 13 channels. All data are provided at a ground sampling distance equal to 10m and a fixed image size of $256 \times 256$ pixels. The high-resolution semantic maps have a resolution of 10m and follow a simplified version of IGBP classification scheme, which originally comprised 17 classes, but it was aggregated to 10 less fine-grained classes. For the DFC2020 dataset, the Savanna and Snow/Ice classes are not accounted for since they do not appear in the images.

For our experiments, we only use the 986 “validation” scenes with high-resolution labels for training purposes and test on the 5128 reserved for testing. For self-supervised pretraining we also use the unlabeled test set. We choose not to exploit the labels in Track 1, nor the larger SEN12MS dataset to fit the condition of having a limited amount of available data. This is the typical setting where SSL is expected to improve performance. We use the widely adopted overall
TABLE I: Test AA for the linear protocol of DeepLab (DL) and Dranet (Dra) at different initializations.

|        | Random init | ImageNet  | SimSiam | SSCL  |
|--------|-------------|-----------|---------|-------|
| DL AA  | 28.0 ± 1.2  | 20.0 ± 1.2| 31.6 ± 0.1| 30.8 ± 0.9|
| OA     | 41.8 ± 4.7  | 90.2 ± 1.5| 43.0 ± 0.2| 46.4 ± 1.0|
| Dra AA | 21.4 ± 2.1  | 19.3 ± 1.2| 28.9 ± 0.2| 34.4 ± 0.7|
| OA     | 30.8 ± 2.6  | 28.0 ± 1.2| 32.7 ± 0.3| 50.7 ± 1.5|

accuracy (OA) and the average accuracy (AA) as evaluation metrics.

The single-channel feature extractor is based on ResNet-18 with $F = 64$ features.

For the Unifeat pretraining stage, we randomly crop patches of size $64 \times 64$ and use a batch size of 128. The learning rate is set to $10^{-3}$ and the temperature hyperparameter to $\tau = 0.1$. We train the single-channel FE using SGD optimizer [38] (weight decay equal to $5 \times 10^{-3}$ and momentum equal to 0.9) for one epoch. We decided to train only for one epoch to avoid collapsing solutions and to preserve spectral information belonging to each band.

Finally, for the CoRe pretraining stage, the network is trained for 200 epochs, Adam optimizer [39], learning rate of $10^{-3}$ and batch size equal to 40. The loss is the $\ell_1$ distance between the reconstructed image and the true uncorrupted one. At this stage, the input images are randomly cropped to $224 \times 224$. The implementation details of the input degradation process are the following:

- **Channel dropout**: a channel is removed with probability $p = 0.5$;
- **Cutout**: 10 masks of random size (ranging from 10 to 30 pixels) are randomly positioned within the images;
- **Gaussian blur**: a Gaussian kernel is applied with a standard deviation chosen uniformly between 0.1 and 2.0.

Finally, for the SimSiam baseline used in the comparisons, we use a batch size of 100 with patches of size $100 \times 100$. The learning rate is $10^{-3}$ with the SGD optimizer. All the experiments have been conducted using two Nvidia Titan-RTX GPUs. Code is available at [https://github.com/diegovalsesia/sscl](https://github.com/diegovalsesia/sscl).

B. Main results

We first assess the effectiveness of the self-supervised learning stages. This can be measured by how well the feature vectors associated to the image pixels are separated when the pixels belong to different classes. The established method to evaluate this is the linear protocol, which consists in fitting a linear classifier to the network output, i.e. the weights of the neural network are frozen and only a final linear layer is trained to perform classification (a $1 \times 1$ convolution implementing a linear projection for each pixel is used for this segmentation problem). We compare the proposed method against a randomly initialized network with the same architecture and with respect to using a pretraining on Imagenet or a self-supervised pretraining method which is state-of-the-art on computer vision tasks, namely SimSiam [19]. Table I reports the results for two different architectures used for the segmentation part (DeepLabv3 [40] and Dranet [41]). We can observe that the pretraining on Imagenet is not effective. This is due to the fact that there is a significant domain gap between images used in standard computer vision problems and remote sensing imagery. Instead, both the SSL methods are better than the random initialization, confirming that they learn useful features for the land cover classification task. However, the proposed method shows higher accuracy with respect to SimSiam, confirming our conjecture that the proposed self-supervised tasks are able to better capture the information related to material properties, which is discriminative for the land cover labels, with respect to the, mostly geometric, transformations adopted by SimSiam. We also notice that the result is stable across the two different architectures.

We then focus our attention on evaluating the finetuning performance, i.e., when the entire pretrained model is optimized using the available labels. We compare against different initialization schemes: random initialization, which covers the case of training a model in a fully-supervised manner; initialization from a supervised pretraining of the ResNet-18 backbone on ImageNet, a popular choice despite the domain gaps with remote sensing data; initialization from SSL via SimSiam; initialization with the proposed SSL pretraining. Tables I and III report the results for the DeepLab and Dranet segmentation architectures, respectively. It can be noticed that the proposed approach is the only one that is able to significantly improve over random initialization. In general, this can be explained by the fact that there exists a semantic gap between the pretraining process and the labels of the final task. The finetuning process has thus to use the precious labels to partially undo what has been learned during pretraining, losing efficiency. In particular, pretraining on ImageNet does not allow to learn feature extractors that are aware of the spectral information, or that can effectively process SAR data, being their statistics so different from optical data. On the other hand, SimSiam is also not very effective because the mostly geometric augmentations used by the SSL do not lend themselves to learning features that correlate with materials and thus do not generalize well beyond the limited improvements seen in the linear protocol. Finally, notice, once again, how the observed gains are agnostic to the specific segmentation architecture. These results suggest that the proposed method is highly effective to improve the performance of end-to-end deep learning models for land cover classification when SAR and multispectral data are jointly used.

A qualitative comparison is shown in Fig. 4 which shows some examples of predicted maps obtained using the different methods considered in the evaluation. We can observe that the proposed SSCL is able to segment finer details than existing methods. Also notice that, according to visual inspection, in some cases, SSCL seems to be even more accurate than the ground truth due to mislabeling issues in the dataset, especially for similar classes such as Shrubland, Grassland and Forest.

The improvement of the proposed method is confirmed by evaluating its performance when considering a different number of training labels, as shown in Fig. 5. The difference with respect to the baseline in terms of AA ranges from 2 to 4 percentage points for all the number of training samples considered in the evaluation. It is interesting to notice
TABLE II: Class-wise average and overall accuracies achieved on the DFC2020 test set (Track 2) for a single-channel FE
DeepLab with different initializations.

| Class          | Random init. | ImageNet | SimSiam | SSCL     |
|----------------|--------------|----------|---------|----------|
| Forest         | 74.6 ± 5.7   | 77.7 ± 4.7| 75.5 ± 1.3| 83.8 ± 1.7 |
| Shrubland      | 56.3 ± 1.9   | 50.7 ± 7.0| 46.4 ± 7.1| 60.1 ± 4.5 |
| Grassland      | 33.3 ± 5.9   | 35.4 ± 8.1| 37.4 ± 4.2| 39.5 ± 8.0 |
| Wetlands       | 11.0 ± 2.1   | 9.3 ± 3.7 | 13.4 ± 3.6| 12.1 ± 2.4 |
| Croplands      | 32.1 ± 5.3   | **36.8 ± 5.4**| 28.1 ± 4.1| 29.4 ± 1.1 |
| Urban          | 79.6 ± 3.9   | 77.1 ± 3.5| **79.8 ± 6.9**| 78.9 ± 4.1 |
| Barren         | 40.9 ± 5.2   | 42.2 ± 6.7| 43.4 ± 10.3| **44.8 ± 5.3** |
| Water          | 99.2 ± 0.2   | **99.8 ± 0.1**| **99.3 ± 0.1**| **99.3 ± 0.1** |
| AA             | 53.4 ± 1.3   | 53.6 ± 0.9| 52.9 ± 1.1| **56.0 ± 1.1** |
| OA             | 65.1 ± 1.9   | 66.4 ± 0.8| 64.6 ± 1.3| **67.8 ± 0.9** |

TABLE III: Class-wise average and overall accuracies achieved on the DFC2020 test set (Track 2) for a single-channel FE
Dranet with different initializations.

| Class          | Random init. | ImageNet | SimSiam | SSCL     |
|----------------|--------------|----------|---------|----------|
| Forest         | 71.2 ± 12.5  | 75.2 ± 3.9| 68.5 ± 5.4| **83.5 ± 2.4** |
| Shrubland      | **52.4 ± 5.7**| 47.7 ± 1.2| 49.6 ± 8 | 49.6 ± 9.5 |
| Grassland      | 29.9 ± 6.8   | **38.1 ± 2.5**| 37.7 ± 3.4| 35.9 ± 1.9 |
| Wetlands       | 7.3 ± 2.3    | 13.2 ± 7.1| 10.2 ± 4 | **14.3 ± 5.0** |
| Croplands      | 29.9 ± 4.6   | 32.2 ± 3.4| 30.8 ± 5.1| **32.7 ± 2.5** |
| Urban          | 75.4 ± 13.4  | 67.3 ± 16.8| **78.6 ± 13**| 75.3 ± 6.1 |
| Barren         | 38.7 ± 6.1   | 38.8 ± 2.0| 32.7 ± 5.7| **50.6 ± 7.5** |
| Water          | **99.4 ± 0.1**| 99.3 ± 0.2| 99.2 ± 0.2| **99.2 ± 0.1** |
| AA             | 50.5 ± 1.9   | 51.5 ± 2.4| 50.9 ± 1.9| **55.2 ± 1.6** |
| OA             | 62.7 ± 3.0   | 63.9 ± 2.2| 63.0 ± 2.4| **67.2 ± 0.8** |

that even with very few samples (48 in the minimum case) the pretrained model can provide better performance than a randomly initialized network.

In conclusion, we must remark that the previously presented results are not directly comparable to the results in Track 2 of the DFC2020 challenge, where AAs up to 61% are reached, for a number of reasons. In particular, the dataset suffers from mislabeling and many works presented in the challenge manually fixed part of the labels. Additionally, we do not use the low-resolution labels from Track 1 to avoid confounding factors related to weak, imprecise supervision. Finally, most of the final results of the challenge are based on ensembles of methods and post-processing techniques which can be applied to any method.

C. Results Analysis

In this section, we analyze how our design choices lead to performance observed in Sec. [IV-B].

First of all, we introduced a contrastive learning approach in UniFeat with the goal of bridging the gap between the local features of multispectral and SAR images, and, to a lesser extent, between bands. What we expect is that, by means of contrastive learning, two patches representing the same ground texture lie closer to each other in the feature space rather than the raw pixel values. At the same time, patches extracted from different locations, either from the same channel or from different channels, should be well separated in the feature space. UniFeat pretraining might reach degenerate solutions if trained for too long, i.e., it will try to make all patches be the same point in the feature feature. To avoid this we pretrained for just one epoch, thus preserving part of the unique information brought by each channel. Fig. [6] shows the effect of contrastive learning by showing the cosine similarity between feature vectors of matching patches (same location, different channel). The histograms are then calculated for a sample of 200 images, and Bands 13 and 14 refer to SAR VH and VV polarization while all the others refer to multispectral Sentinel 2. Notice how the average similarity increased (or equivalently distance decreased) between all SAR-optical pairs. Similar results (not shown in figure) also hold for SAR-SAR, optical-optical pairs, thus reaching the desired objective.

Moreover, we want to analyze the characteristics of the feature maps induced by pretraining with CoRe as well. Fig. [7] shows some representative feature maps from the last network layer and compares them between the proposed method and SimSiam SSL. We can immediately notice that the spatial resolution of the feature maps obtained with CoRe is much higher than SimSiam, and finer details are preserved. This could be explained by the fact that the reconstruction task promotes high-resolution solutions since it has to solve problems that amount to super-resolution/deblurring (e.g., when the highest-resolution channels are dropped, in addition to the input blur) or inpainting, thus heavily relying on fine spatial clues. Instead, SimSiam is designed for a more global reasoning from the traditional whole-image classification perspective, and even if we contrast features in a pixel-by-pixel fashion, there are no
(a) Input  (b) Baseline  (c) Image-Net  (d) SimSiam  (e) SSCL  (f) Ground Truth

Fig. 4: Land cover maps generated by different methods. We can see that the proposed method is able to segment finer details than existing methods. Also notice that, according to visual inspection, it sometimes is even more accurate than the ground truth due to mislabeling issues.

Fig. 5: Test average accuracy over different number of training samples for DeepLab pretrained using our SSCL and using random initialization.

TABLE IV: Test average and overall accuracy with or without UniFeat for DeepLab.

|          | CoRe UniFeat + CoRe (SSCL) |
|----------|-------------------------------|
| AA       | 54.7 ± 0.3                   |
| OA       | 66.4 ± 1.7                   |

|          | 56.0 ± 1.1                   |
| OA       | 67.8 ± 0.9                   |

significant improvement in the case when UniFeat is applied. This confirms that uniforming the features from different sensing modalities is beneficial for the downstream task.

Then, the importance of the single-channel feature extractor is assessed in Table V. The results show that Deeplab with single-channel feature extractor outperforms a standard Deeplab with the first layer merging all the input channels and comparable number of parameters. Note that, for a fair comparison, we show the models without any pretraining in the first two columns and the model with SSCL pretraining in the last column. In addition, the same table shows the performance difference when those models do or do not process SAR images, in order to evaluate how well they are able to exploit this information. We can observe that all the models have a drop in performance when the SAR images are not considered. This means that this information contributes significantly to the characterization of materials in satellite imagery.

D. Ablation Study

In this section, we explore how some design choices can affect the performance of SSCL.

We first investigate the impact of UniFeat. Table IV reports the results of SSCL with and without UniFeat, showing a guarantees to preserve high spatial resolution in the resulting embeddings.

Next, we evaluate the impact of band dropout in CoRe. To this aim, in Table VI we test SSCL with and without band dropout. The results clearly show that band dropout
TABLE V: Test average accuracy for DeepLab with or without single-channel FE and with or without SAR images. Std Deeplab and Single-ch. FE do not have any pretraining.

|               | Std Deeplab | Single-ch. FE | SSCL       |
|---------------|-------------|---------------|------------|
| with SAR      | 52.7 ± 1.5  | 53.4 ± 1.3    | 56.0 ± 1.1 |
| w/o SAR       | 46.0 ± 2.7  | 47.1 ± 1.1    | 51.4 ± 0.8 |

Fig. 6: Examples of histograms over the cosine similarity of feature maps for different channel pairs and for the proposed method without and with UniFeat: (a) channels 13-6, weighted median 0.43 and 0.51; (b) channels 13-11, weighted median 0.35 and 0.40; (c) channels 14-7, weighted median 0.53 and 0.63; (d) channels 14-10, weighted median 0.38 and 0.59. Bands 13 and 14 refer to SAR VH and VV polarizations.

significantly improves the performance of SSCL. We repeat the same test considering a state-of-the-art self-supervised contrastive pretraining method, namely SimSiam, but in this case the results do not show any relevant improvement when we employ band dropout. This highlights the importance of the image reconstruction task over the simple contrastive procedure. We then test different dropout probabilities. Fig. 8 shows that a dropout probability of 50% (i.e., dropping 7 bands out of 15) reaches the highest average accuracy, while performance degrades at high dropout probabilities, which result in a less effective pretraining task due to excessive deterioration of the spectrum.

Finally, we investigate the generalization ability of SSCL by considering test samples never seen during pretraining. Indeed, one could argue that the higher test accuracy obtained using SSL methods could be correlated to the fact that the test samples have already been processed during the self-supervised pretraining, even though without using their labels. For this reason, we evaluate the performance of SSCL using in pretraining only 4096 samples out of the total 5128 samples of the test set and testing the finetuned model only on the remaining 1036 samples. Table VII shows the results compared to the standard model studied so far. We can observe that the performance are equal or within one standard deviation, thus confirming the generalization capability of SSCL to samples never seen before.

V. CONCLUSIONS

In this paper, we proposed a framework for self-supervised pretraining of deep neural networks for the task of land

TABLE VI: Test average accuracy for the SimSiam and SSCL with and without channel dropout.

|               | SimSiam | SimSiam w/o ch.drop. | SSCL       | SSCL w/o ch.drop. |
|---------------|---------|-----------------------|------------|------------------|
|               | AA      | OA                    | AA         | OA               |
| SimSiam       | 52.9 ± 1.1 | 64.6 ± 1.3   | 56.0 ± 1.1 | 67.8 ± 0.9       |
| SSCL w/o ch.drop. | 43.4 ± 1.1 | 46.2 ± 2.0       | 54.3 ± 0.7 | 67.0 ± 0.4       |

TABLE VII: Generalization performance of SSCL over 1036 test samples (included / not included in self-supervised pretraining).

|               | SSCL with 5128 samples | SSCL with 4096 samples |
|---------------|------------------------|------------------------|
| Linear Protocol | 31.6 ± 1.1              | 46.2 ± 2.0              |
| AA            | 28.9 ± 0.5              | 55.2 ± 1.3              |
| OA            | 43.4 ± 1.1              | 55.0 ± 0.8              |
| Finetune      | 46.2 ± 2.0              | 66.1 ± 0.7              |
| AA            | 66.8 ± 1.1              | 66.1 ± 0.7              |
| OA            | 66.8 ± 1.1              | 66.1 ± 0.7              |
cover classification. We showed how the proposed method is effective at jointly processing images from multiple sensing modalities, such as SAR and multispectral. The proposed self-supervised training strategies are able to learn feature extractors that compute features that correlate with the land cover classes without any labeled data. Moreover, the performance of the pretrained feature extractors can be improved if labels are available, resulting in superior performance with respect to purely supervised training.

REFERENCES

[1] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, “Natural language processing advancements by deep learning: A survey,” arXiv preprint arXiv:2003.01200, 2020.

[2] A. Rasouli, “Deep learning for vision-based prediction: A survey,” arXiv preprint arXiv:2007.00095, 2020.

[3] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” ISPRS journal of photogrammetry and remote sensing, vol. 152, pp. 166–177, 2019.

[4] A. Kolesnikov, X. Zhai, and L. Beyer, “Revisiting self-supervised visual representation learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1920–1929.

[5] C. Robinson, K. Malkin, N. Jojic, H. Chen, R. Qin, C. Xiao, M. Schmitt, P. Ghamisi, R. Hänsch, and N. Yokoya, “Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3185–3199, 2021.

[6] X. Chen and K. He, “Exploring simple siamese representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 750–15 758.

[7] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, “Scaling and benchmarking self-supervised visual representation learning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6391–6400.

[8] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” IEEE transactions on pattern analysis and machine intelligence, 2020.

[9] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015.

[10] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in European conference on computer vision. Springer, 2016, pp. 69–84.

[11] D. Pathak, P. Krizhevsky, I. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[12] G. Larsson, M. Maire, and G. Shakhnarovich, “Colorization as a proxy task for visual understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6874–6883.

[13] T. Xiao, X. Wang, A. A. Efros, and T. Darrell, “What should not be contrastive in contrastive learning,” arXiv preprint arXiv:2008.05659, 2020.

[14] S. Yamaguchi, S. Kanai, T. Shioda, and S. Takeda, “Multiple pretext-task for self-supervised learning via mixing multiple image transformations,” arXiv preprint arXiv:1912.10803, 2019.

[15] A. Jaisswal, A. R. Baha, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” Technologies, vol. 9, no. 1, p. 2, 2021.

[16] G. Koch, R. Zemel, R. Salakhutdinov et al., “Siamese neural networks for one-shot image recognition,” in ICML deep learning workshop, vol. 2. Lille, 2015.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[19] X. Chen and K. He, “Exploring simple siamese representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 15 750–15 758.

[20] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” arXiv preprint arXiv:2006.09882, 2020.

[21] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, 2020, pp. 776–794.

[22] D. Marmanis, M. Datcu, T. Esch, and U. Still, “Deep learning earth observation classification using imagenet pretrained networks,” IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 1, pp. 105–109, 2015.

[23] M. Mahdianpari, B. Salehi, M. Rezaee, F. MohammadiMarinat, and Y. Zhang, “Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery,” Remote Sensing, vol. 10, no. 7, p. 1119, 2018.

[24] S. Vincenzi, A. Porrello, P. Buzzega, M. Cipriano, P. Fronte, R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara, “The color out of space: learning self-supervised representations for earth observation imagery,” in 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 3034–3041.

[25] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, “Remote sensing image scene classification with self-supervised paradigm under limited labeled samples,” IEEE Geoscience and Remote Sensing Letters, 2020.

[26] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, “Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast,” IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 2598–2610, 2021.

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

[28] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, “ Dense contrastive learning for self-supervised visual pre-training,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3024–3033.

[29] P. Satija and V. B. Deepa, “Analysis of supervised image classification method for satellite images,” International Journal of Computer Science Research (IJCSR), vol. 5, no. 2, pp. 16–19, 2017.

[30] C. Lo and J. Choi, “A hybrid approach to urban land use/cover mapping using landsat 7 enhanced thematic mapper plus (etm+) images,” International Journal of Remote Sensing, vol. 25, no. 14, pp. 2687–2700, 2004.

[31] A. Mellor, A. Haywood, C. Stone, and S. Jones, “The performance of random forests in an operational setting for large area sclerophyll forest classification,” Remote Sensing, vol. 5, no. 6, pp. 2838–2856, 2013.

[32] A. W. Abbas, N. Minalll, N. Ahmad, S. A. R. Abid, and M. A. A. Khan, “K-means and isodata clustering algorithms for landcover classi-
[33] J. E. Ball, D. T. Anderson, and C. S. Chan Sr, “Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community,” Journal of Applied Remote Sensing, vol. 11, no. 4, p. 042609, 2017.

[34] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in Advances in neural information processing systems, 2016, pp. 1857–1865.

[35] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, “Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion,” arXiv preprint arXiv:1906.07789, 2019.

[36] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown et al., “Gmes sentinel-1 mission,” Remote Sensing of Environment, vol. 120, pp. 9–24, 2012.

[37] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort et al., “Sentinel-2: Esa’s optical high-resolution mission for gmes operational services,” Remote sensing of Environment, vol. 120, pp. 25–36, 2012.

[38] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in Proceedings of COMPSTAT’2010. Springer, 2010, pp. 177–186.

[39] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in ECCV, 2018.

[41] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, “Scene segmentation with dual relation-aware attention network,” IEEE Transactions on Neural Networks and Learning Systems, 2020.