Continual Barlow Twins: Continual Self-Supervised Learning for Remote Sensing Semantic Segmentation

Valerio Marsocci and Simone Scardapane, *Member, IEEE*

***Abstract*—In the field of earth observation (EO), continual learning (CL) algorithms have been proposed to deal with large datasets by decomposing them into several subsets and processing them incrementally. The majority of these algorithms assume that data are, first, coming from a single source, and second, fully labeled. Real-world EO datasets are instead characterized by a large heterogeneity (e.g., coming from aerial, satellite, or drone scenarios), and for the most part they are unlabeled, meaning they can be fully exploited only through the emerging self-supervised learning (SSL) paradigm. For these reasons, in this article, we present a new algorithm for merging SSL and CL for remote sensing applications that we call continual Barlow twins. It combines the advantages of one of the simplest self-supervision techniques, i.e., Barlow twins, with the elastic weight consolidation method to avoid catastrophic forgetting. In addition, we evaluate the proposed continual SSL approach on a highly heterogeneous EO dataset, showing the effectiveness of this strategy on a novel combination of three almost non-overlapping domains datasets (airborne Potsdam, satellite US3D, and drone unmanned aerial vehicle semantic segmentation dataset), on a crucial downstream task in EO, i.e., semantic segmentation. Encouraging results show the superiority of SSL in this setting, and the effectiveness of creating an incremental effective pretrained feature extractor, based on ResNet50, without the need of relying on the complete availability of all the data, with a valuable saving of time and resources.**

***Index Terms*—Continual learning (CL), remote sensing, self-supervised learning (SSL), semantic segmentation.**

I. Introduction

IN RECENT years, improvements in speed and acquisition technologies have drastically increased the amount of available earth observation (EO) images [1]. These improvements bring challenging issues to the widespread use of remote sensing (RS) classification [2] and semantic segmentation techniques, due to 1) the continuous arrival of new data, generally belonging to partially overlapping domains and 2) the increasing quantity of images, which has not been labeled by a domain expert. Most of the time power consumption issues are raised, too. The main aim of this article is to propose a model capable to deal with these characteristics simultaneously that we call continual Barlow twins (CBT). This algorithm combines the strengths of two separate lines of research: Continual learning (CL) for processing a large heterogeneous dataset in an incremental way, tackling the first highlighted problem, and self-supervised learning (SSL) to deal with the lack of labeling information, solving the issues raised by the second problem. In the following, we describe briefly the two issues separately, before introducing our proposed solution, shaped to be easy and intuitive, being a first step toward what we call continual self-supervised learning (CSSL) in EO.

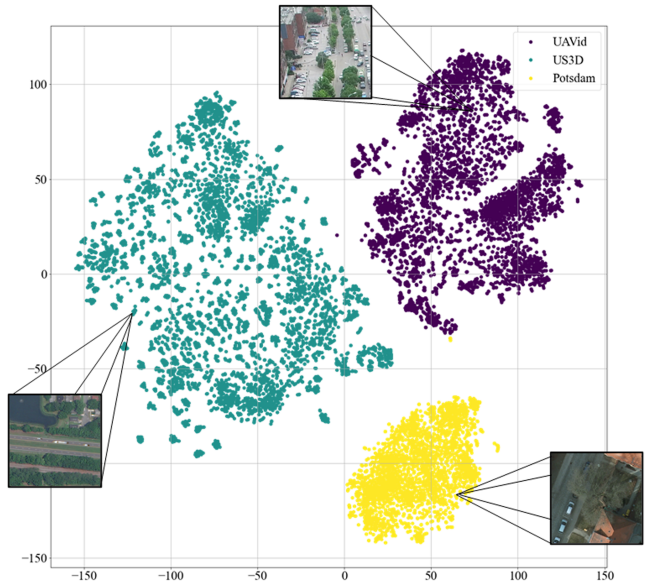

Fig. 1. t-Stochastic Neighbor Embedding (t-SNE) visualization results of the features of three selected RS datasets (see Section IV for a description of the datasets). It can easily be seen that the images from the different settings can be considered independent but non-identically distributed.

A. Problem #1: EO datasets are heterogeneous

Consider the situation where we trained a semantic segmentation network on a dataset of Italian satellite images. If we receive a new labeled dataset of similar images from a different nation, ideally, we would like our network to be able to segment equally well images coming from the two countries. However, most of the algorithms developed for classification or segmentation in EO suffer from catastrophic forgetting problems in this context, requiring to discard the acquired knowledge to retrain the model from scratch on the combination of the two datasets [3]. In a wide range of EO applications, the strategy of retraining the whole

Manuscript received 9 January 2023; revised 13 March 2023 and 24 April 2023; accepted 23 May 2023. Date of publication 25 May 2023; date of current version 12 June 2023. *(Corresponding author: Valerio Marsocci.)*

Valerio Marsocci is with the Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, 00185 Rome, Italy (e-mail: valerio.marsocci@uniroma1.it).

Simone Scardapane is with the Department of Information Engineering, Electronics and Telecommunication, Sapienza University of Rome, 00184 Rome, Italy (e-mail: simone.scardapane@uniroma1.it).

Digital Object Identifier 10.1109/JSTARS.2023.3280029

model is computationally expensive and costly [4]. Therefore, there is a need to ensure that the newly-developed models have the ability to learn new tasks while retaining satisfying performance on previous ones. The cause of the catastrophic forgetting problem is that different datasets are independent but not identically distributed in the feature domain, as it is known that the distribution of RS data varies greatly [5], due to different resolutions, acquisitions, textures, and captured scenes. This is even more evident in urban scenes and in datasets made of images acquired from different types of sensors, e.g., drone, airborne, and satellite (see Fig. 1 for a visualization of this phenomenon).

In this article, we leverage a CL algorithm [3] to mitigate the catastrophic forgetting problem and allow our algorithm to generalize to different feature distributions without the requirement of accessing already seen data.

B. Problem #2: EO datasets are largely unlabeled

Most methods, especially in EO applications, are framed as supervised systems, relying on annotated data. More than in other fields, for drone, aerial, and satellite images, it is difficult to rely on a labeled dataset, in light of the high cost and the amount of effort and time that are required, along with domain expertise [6]. In computer vision (CV), SSL has been proposed to handle this problem, reducing the amount of annotated data needed [7], [8]. The goal of SSL is to learn an effective visual representation of the input using a massive quantity of data provided without any label [9]. We can see the task as the need to build a well-structured and relevant set of features, able to represent an image, which is useful for several downstream tasks. There is a growing research line demonstrating how SSL techniques increase performance in EO applications [10], [11], [12], although most of the evaluations have been limited to a single dataset or domain. More recent works employed different datasets for their SSL solutions [13], [14], [15], but focused their effort toward multimodal models, based on existing CV approaches [16].

C. Contributions of This Article

While CL and SSL have been explored in isolation in RS (as described in Section II), we propose to investigate a novel CSSL scenario wherein multiple heterogeneous datasets arrive *continuously* and a *single* backbone network must be updated for the downstream tasks. This poses new and interesting challenges, both in terms of data heterogeneity (as we build each task from a completely different data source) and effectiveness of existing CL solutions in the SSL task. To this end, we design and experimentally evaluate a novel strategy, built on two popular algorithms that is able to train a deep network for RS by exploiting vast amounts of unlabeled, continually arriving data.

Our experimental results show that it is possible to exploit the potential of SSL *incrementally*, to obtain an efficient and effective pretrained model trained in several successive steps, without the need to retrain it from scratch every time new data is added [17]. The proposed CBT algorithm trains a feature extractor (ResNet50) with Barlow twins (BT) [7], whose loss is integrated with a regularization term targeted for CL, borrowed from elastic weight consolidation (EWC) [18], to avoid catastrophic forgetting. With the obtained feature extractor, we train a UNet++ [19] to perform semantic segmentation. We underline that while our method is built on the combination of two easy and widespread methods from the CV literature, the proposed CSSL scenario has not, to the best of authors' knowledge, been investigated in RS. The selection of EWC and BT was based on the current RS literature on SSL [11], [13], [20], [21], [22], and CV more in general [7], [8], [16], [23]. We strongly believe that EWC and BT are an effective combination, due to their properties. The former proposes an easy weight regularization term, based on the importance of the weights in solving the tasks. The latter, avoiding typical issues of contrastive frameworks [24], aims to reduce the redundancy of the embeddings. Thus, when acquiring new RS data, CBT can be trained quickly, as it will be necessary to update it on the new data only, discarding all old data. Our method also provides computational efficiency, potentially allowing small realities to train a large model on huge amounts of data that would be unfeasible otherwise [15].

Since the generalization capabilities and benefits of SSL on RS data from non-overlapping domains (as shown in Fig. 1) are still unexplored, we also propose a new benchmark by combining three datasets with images captured with different sensors (drone, airborne, and satellite data), with different resolutions and acquired under different conditions, representing different objects and scenes. In fact, in RS, there are several CL benchmarks focused on incremental annotation of land cover classes [5], [25], [26], but the case where images with different characteristics arrive in a continual scenario is underexplored. On the other hand, datasets based on the bundle of already existing datasets, are gaining more and more interest in SSL for RS [13], [14], [27], [28]. Thus, we show that SSL targeted to RS images can outperform standard pretrained strategies (e.g., ImageNet), and we expect this to become a useful benchmark scenario for further research in SSL and CL in RS. We also show that the proposed CBT algorithm offers significantly more versatility and less computing time compared to a standard approach.

II. Related Works

In this section, we briefly review the relevant literature on SSL (see Section II-A) and CL (see Section II-B) in EO. In CV, the combination of these two technologies is starting to be explored, and some works have already started to demonstrate how SSL methods are feasible to learn incrementally [17], [29], [30]. In EO, this combination has not yet been explored, to the best of authors' knowledge, apart from a few contributions combining weakly-supervision [31] and contrastive losses [32], [33] with CL. On the other hand, we observe an increasing interest in foundation models [15], [34], based also on new extensive datasets [14], [27], [28].

A. Self-Supervised Learning in EO

In [11] and [12], we can find two reviews of SSL in RS. SatMAE [13] trains a masked autoencoder [16], properly modified for RS purposes. In [21], the authors train CMC [8] on three large datasets both with RGB and multispectral bands, evaluating the effectiveness of the learned features on several classification downstream tasks. The same authors, in [10], apply a split-brain autoencoder on aerial images. In [35], a global style and local matching contrastive learning network is proposed. FALSE [36] sets an effective strategy for negative sampling. Also in [37], a contrastive method for EO semantic segmentation is proposed. GeoKR [38] uses metadata for an efficient pretraining strategy on a wide dataset. Marsocci et al.[22] performed a semantic segmentation downstream task on the Vaihingen dataset [39] to learn the features of the encoder of the network that solves the segmentation. Ayush et al. [40] introduced a loss term based on the geolocation of the tiles. Similarly, SeCo [41] is based on seasonal difference among same views. Other contrastive strategies are proposed by [42] and [43]. Vincenzi et al.[44] learned visual representations inferring information on the visible spectrum from the other bands on BigEarthNet [45]. Dong et al. [46] proposed a GAN discriminator, which has to identify patches taken from two temporal images. A similar approach, with multiview images, is proposed in [47]. Yuan and Lin [48] showed the effectiveness of an SSL pretraining for time-series classification. In [49], SSL strategy for transfer learning super-resolution purposes is used. Finally, several semisupervised approaches have been proposed [50], [51], [52], often along with new extensive datasets for effective pretraining [27], [28]. Recently, referring to CV in general, several innovations have been proposed [53], [54], reviewed in many surveys [24], [55], [56], [57], [58].

B. Continual Learning in EO

Ammour [25] proposed a two-block network for RS land cover classification tasks, where one module minimizes the error among classes, while another one learns how to effectively distinguish among tasks, based on a linear memory. In [59], continual prototype calibration is proposed for few-shot classification CL. The authors in [32] and [33] make use of contrastive learning to learn effective representations that can reduce catastrophic forgetting. A fine-grained CL algorithm for SAR incremental target recognition is presented in [60]. In [61], a CL network for pansharpening is proposed for the first time. The authors in [62] and [63] proposed adapting and remembering strategies based on a memory that holds the previous-task net. Continual learning benchmark for remote sensing [64] is a large-scale remote sensing image scene classification database based on three CL scenarios. CILEA-Net [65] proposes a CL strategy, based on the incremental learning of new classes ordered according to the similarity with the old ones. In [66], an incremental learning with open-set recognition framework and a new loss are proposed for RS image scene classification. Alqahtani and Ammour [67] trained two subnetworks for continually learning classes from RS generated images. Lightweight incremental approach proposes a small feature transfer module, to align representations continually. Focused on small object segmentation, class incremental learning proposes a diversity distillation loss. Shaped for semantic segmentation, Feng et al. [5] proposed two regularization components: representation consistency structure loss and pixel affinity structure loss. The first retains the information in the isolated pixels. The second saves the high-frequency information throughout the tasks. Recently, a geospatial foundation model [15] has been proposed, to obtain an effective pretrained model, similarly to [13], but in a continual fashion. In the wider deep learning landscape, several surveys reviewed the best CL approaches and nets [3], [68], [69].

III. Methodology

A. Overview of the Components

We consider an RS scenario where:

- data are coming incrementally from multiple domains (e.g., drone, airborne, and satellite images);
- we cannot retrain from scratch the model when new data are received;
- the majority of the data is unlabeled.

We refer to each domain (or subset of the dataset) as a *task*, in accordance with the CL literature. Differently from several CL for RS benchmarks [5], [25], [26], we are not interested in the continual annotation of new classes on images arriving from the same source (i.e., sensor), but on the incremental acquisition of images from different sensors (i.e., domains). Thus, when dealing with multiple datasets, most of the classes can change and we want our method to naturally adapt to this shift by considering separate heads in a task incremental scenario, common in CL [68], [69]. To achieve our compound objective, the intuition is to embed a CL strategy in a self-supervised framework, by combining two algorithms that are considered state-of-the-art in their respective fields: (i) BT [7], which trains a network based on measuring the cross-correlation matrix between the outputs of two identical networks fed with augmented versions of a sample, and making it as close to the identity matrix as possible; (ii) EWC [18] consisting of constraining important weights of the network to stay close to the value obtained in previous tasks. We decided to put together two methods that would simultaneously work well in terms of the effective regularization of the internal embeddings of the network, and in reducing their redundancy across tasks. Our intuition is confirmed by the experiments (see Section VI). In the following section, we highlight how CBT works. A schematic overview of the method is provided in Fig. 2.

B. Continual Barlow Twins

Consider for now a single task, and denote by X a batch of unlabeled images. Our main training step, taken from BT, produces two disturbed views of X, Y_A, and Y_B, based on a set of data augmentations strategies S (e.g., random rotations and scalings). In this article, we consider standard sets of data augmentations (see Section V), although augmentations specific to RS could also be considered. The two views are fed to a convolutional neural network with weights θ that produces, respectively, two

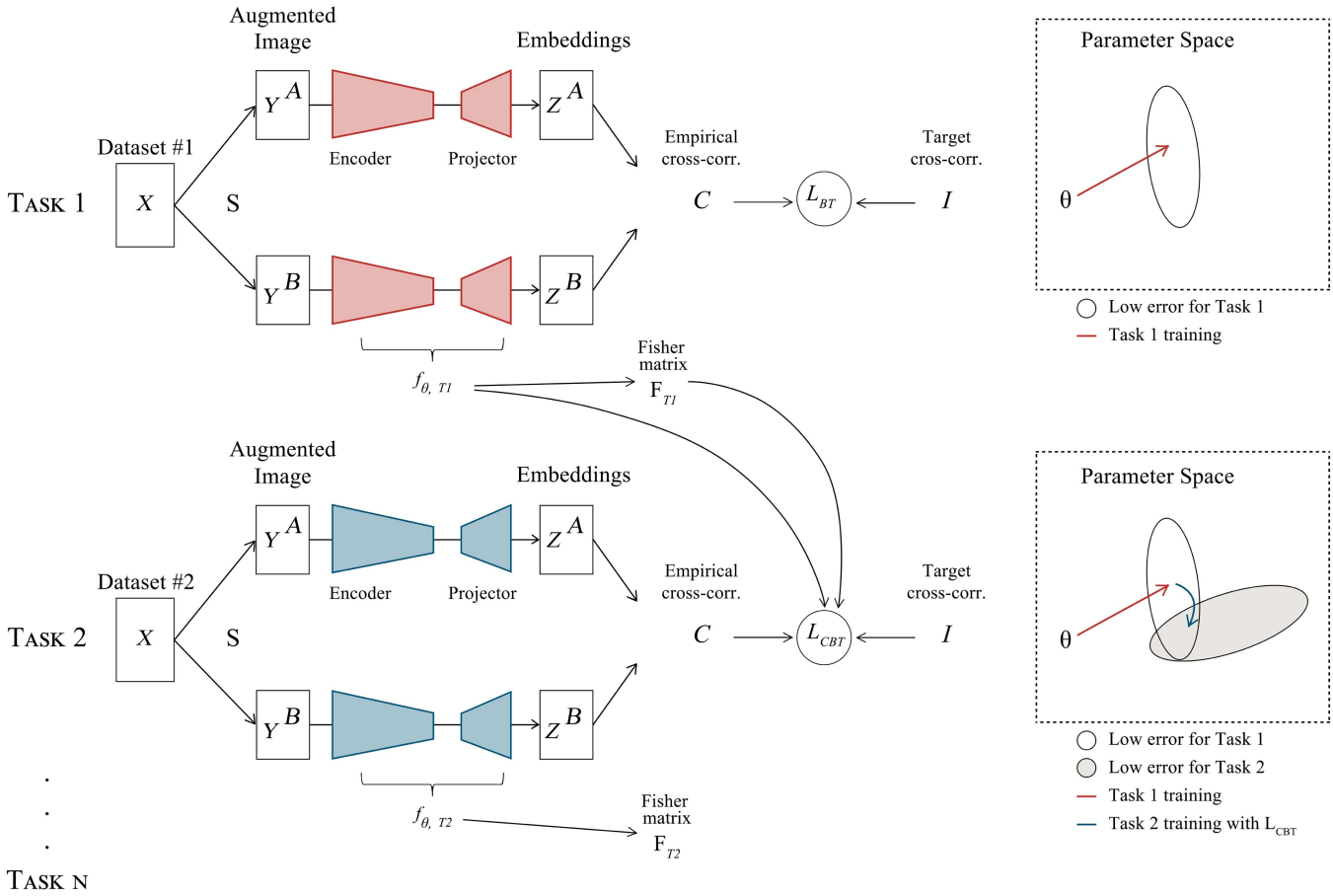

Fig. 2. Schematic representation of the CBT algorithm. When computing $\mathcal{L}_{\mathcal{CBT}}$, C, and I contribute to the BT loss term, $f_{\theta,T1}$, F, and $f_{\theta,T2}$ to the EWC regularization term.

embeddings Z_A and Z_B (assumed to be mean-centered along the batch dimension). To learn effective representations of the input images in a self-supervised fashion we leverage the BT loss, which is composed of two terms called *invariance* and *redundancy reduction* terms

$$\mathcal{L}_{\mathcal{BT}}(X) = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \mu \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}} \tag{1}$$

where μ is a positive constant balancing the invariant and the redundancy reduction terms of the loss, and $\mathcal{C}$ is the cross-correlation matrix, with values comprised between -1 (i.e., total anticorrelation) and 1 (i.e., total correlation), computed between the outputs of the two networks along the batch dimension (i.e., for each mini-batch, the vector Z_b^A compute the cross-correlation against Z_b^B, where $b = 1, \ldots, N$ is the index of the mini-batch and N is the mini-batch size). Practically, the first term of the loss has the goal to make the diagonal elements of $\mathcal{C}$ equal to 1. In this way, the embeddings become invariant to the applied augmentations. On the other hand, the second term of the loss has the aim to bring to 0 the off-diagonal elements of $\mathcal{C}$. This ensures that the various components of the embeddings are decorrelated with each other, making the information non-redundant, enhancing the representations of the images.

Suppose now that the network has been trained on images coming from a task T_1 using the loss (1) (e.g., drone images), and we receive a new dataset of images coming from a second task T_2 (e.g., satellite images). We denote the weights obtained at the end of the first training as θ^{T_1}, and the data of the two tasks, respectively, as D_{T_1} and D_{T_2}. To retain old knowledge from T_1 and avoid catastrophic forgetting, we complement the BT loss (1) with an EWC regularization term [18], which forces the weights to stay close to θ^{T_1} depending on their importance, given by the diagonal of the Fisher information matrix F, which is a positive semidefinite matrix corresponding to the second derivative of the loss near the minimum. In our scenario, the loss cannot be decomposed for each individual data point, as it depends on the cross-correlations between data in a mini-batch and its corresponding augmentations. To this end, denoting by B_{T_1} the number of mini-batches X that can be extracted from D_{T_1}, we approximate the ith element of the diagonal Fisher information matrix as

$$F_i = \frac{1}{B_{T_1}} \sum_{X \in D_{T_1}} \left[\frac{\partial \mathcal{L}_{\mathcal{BT}}(X)}{\partial \theta_i^{T_1}} \right]^2 \tag{2}$$

where $\mathcal{L}_{\mathcal{BT}}(X)$ denotes the BT loss computed on mini-batch X, as in (1). Intuitively, each weight of the network is given an importance that depends on the square of the corresponding loss

TABLE I
SUMMARY OF THE DATASETS USED FOR THE EXPERIMENTS

Dataset	Type of images	Number of images
Potsdam [39]	Aerial	∼5000
UAVid [70]	UAV	∼7500
US3D [71]	Satellite	∼11000

gradient. Given this approximation, the new loss for a batch of images X taken from the second task is given by

$$\mathcal{L}(X) = \mathcal{L}_{\mathcal{BT}}(X) + \sum_i \frac{\lambda}{2} F_i \left(\theta_i - \theta_i^{T_1}\right)^2 \tag{3}$$

where $\mathcal{L}_{\mathcal{BT}}(X)$ is the BT loss (1) computed on the data from task $T2$, and λ weights the constraint on the previous task. If moving to a third task, we repeat the computation of the Fisher information matrix at the end of training for the second task and replace it. The CBT approach is summarized in Fig. 2, and the associated code is available online.[1] After the self-supervised pretraining, the network can be exploited for any downstream task of interest in EO. In particular, we explore in Section V, a fine-tuning for a semantic segmentation task.

IV. DATASETS

To perform the experiments, we build a novel dataset, which is a combination of three previously introduced datasets. Each contains images from a different source: airborne, satellite, and drone. As previously stated, the construction of a novel mixed dataset is crucial since the data are vastly heterogeneous, presenting almost non-overlapping domains, as shown in Fig. 1. In fact, the choice was dictated by the desire to demonstrate the effectiveness of the SSL on a challenging task (that is semantic segmentation), extending its validity even in the case of highly variable data, while most previous works focused on a single domain (see Section II). We briefly summarize next each dataset. Salient information is summed up in Table I.

A. Potsdam

The ISPRS Potsdam dataset [39] consists of 38 high-resolution aerial true orthophotos (TOP), with four available bands (near-infrared, red, green, and blue). Each image is 6000×6000 pixels, with a ground sample distance of 5 cm, ending up in covering 3.42 km^2. For our experiments, we took in consideration only the 38 RGB TOPs. These are annotated with pixel-level labels of six classes: background, impervious surfaces, cars, buildings, low vegetation, and trees. We used the eroded mask, and we selected 24 images for training, 13 for testing, and 1 for validation, without considering the background class, similarly to [72]. We cropped each image in 512×512 non-overlapping patches, ending up in 2640 images for training, 120 for validation, and 1680 for test.

1) Unmanned Aerial Vehicle Semantic Segmentation Dataset (UAVid): UAVid [70] consists in 42 video sequences, captured with 4 K high-resolution by an oblique point of view. UAVid is a challenging dataset due to the very high resolution of images, large-scale variation, and complexities in the scenes. The authors extracted ten labeled images per each sequence, ending up in 420 images with 3840×2160 pixels. The annotated classes are eight: building, road, static car, tree, low vegetation, human, moving car, and background clutter. The images are already divided into train, validation, and test, by the authors, however, the test segmentation maps have not yet been released. For this reason, we used a part (80%) of the validation set as test set in our experiments. Moreover, we cropped the images in 512×512 non-overlapping patches, ending up in $\sim$ 7500 images.

2) US3D: The US3D dataset [71] includes approximately 100 km^2 coverage for the United States cities of Jacksonville, Florida, and Omaha, Nebraska. Sources include incidental satellite images, airborne LiDAR, and feature annotations derived from LiDAR. The dataset is composed of 2783 images, 1024×1024, obtained from the WorldView-3 satellite: they are non-orthorectified and multiview. The images have eight bands, six of which are part of the visible spectrum, and two of the near-infrared. Semantic labels for the US3D dataset, derived automatically from Homeland Security Infrastructure Program, are five: ground, trees, water, building, and clutter. For our experiments, we considered only RGB bands and all the classes. Also for this dataset, we cropped the images in 512×512 non-overlapping patches, ending up in more than 11 000 images, randomly divided in train ($\sim$ 70%), validation ($\sim$ 10%), and test ($\sim$ 20%).

V. EXPERIMENTAL SETUP

For the training phase, a single Tesla V100-SXM2 32 GB GPU has been used. For the semantic segmentation task, we use UNet++ [19], with the squeeze and excitation strategy [73] and the softmax function as activation on the last layer. For the experiments on all the three datasets, we fix the batch size to 8, the number of epochs to 200, and the learning rate to 0.0001. Moreover, we used Adam as optimizer, Jaccard loss as the cost function, and the following set of augmentations: random horizontal flip, random geometric transformation (i.e., shifting, scaling, rotating), random Gaussian noise, random radiometric transformation (i.e., brightness, contrast, saturation variations). The mean intersection over union (mIoU), and F1-score (F1) are the selected evaluation metrics. Under these conditions, we test different pretraining strategies for UNet++:

1) *ImageNet*, ResNet50 pretrained on ImageNet in a supervised manner. It is used as a baseline;
2) *BT ImageNet*, ResNet50 pretrained on ImageNet data with BT strategy,[2] used as additional baseline;
3) *CBT*, an incrementally pretrained ResNet50 on the three datasets in this order: US3D, UAVid, Potsdam (with $\lambda = 10e-2$). The rest of parameters are as in [7]. This is our proposed model;
4) *BT*, a ResNet50 pretrained on the three datasets, taken together in only one step, being the upperbound baseline;

[1] [Online]. Available: https://github.com/VMarsocci/CBT

[2] [Online]. Available: Downloaded at https://github.com/facebookresearch/barlowtwins

TABLE II
ELAPSED TRAINING TIMES

Strategy	Step 1 (s)	Step 2 (s)	Step 3 (s)	Tot (h)
BT	41400	61500	88600	53.19
CBT	42000	36400	25100	28.75

CBT offers important advantages when data are provided in an incremental fashion.

5) *3-step BT*, a ResNet50 pretrained consequentially on the three datasets (US3D, UAVid, Potsdam) with a vanilla BT, without CL constraints. This experiment is meant to assess the emergence of catastrophic forgetting.

With the encoders so trained, we trained the semantic segmentation models in a supervised way, with different percentages (10%, 50%, 100%) of labeled data for the three datasets. Particularly, we run all the experiments three times, reporting the mean and the standard deviation of the resulting metrics. The results are shown and commented in the following Section VI.

VI. EXPERIMENTAL RESULTS

The results of the experiments on the downstream task are shown both in Fig. 3 and Tables III–V. As can be seen, the feature extractors that perform best are the ones obtained from the CBT and BT training on the three selected EO datasets. Precisely, these conformations outperform, with reference to mIoU, their counterpart trained with ImageNet supervised pretraining, respectively, by 3.39% and by 3.69% in average. Moreover, it is interesting to note that the performance of the models with the CBT feature extractor is only slightly lower (average drop of a negligible $\approx 0.3\%$, referring to mIoU) than that obtained with an encoder trained by means of BT, demonstrating how the proposed approach gives up only a slight part of optimal performance, against clear advantages in terms of computational efficiency and general versatility. In absolute terms, it is necessary to notice once again how self-supervision strategies lead to better results than exclusively supervised ones [21], [40], and, above all, how this is even more true when combining EO data from domains that are not homogeneous in terms of type of sensor, acquisition, resolution, and objects represented. In particular, in the next paragraphs, we will go in the depth of some specific evidences regarding: computational times (Section VI-A), UAVid experiments (Section VI-B), Potsdam experiments (Section VI-C), US3D experiments (Section VI-D), and catastrophic forgetting (Section VI-E).

A. Computational Times

As stated earlier, one of the best advantages of the proposed new method is the shorter computational time with a very limited performance drop in the case of incremental data availability. As already stressed, this situation is especially likely in the field of EO, where new data often arrives continuously [62], due to satellite revisit times, scheduled acquisition campaigns, and other variable parameters. In Table II, we can observe the results of the experiments. Concerning the traditional BT strategy, for the training of the three considered datasets, we simulate an incremental arrival of data as follows:

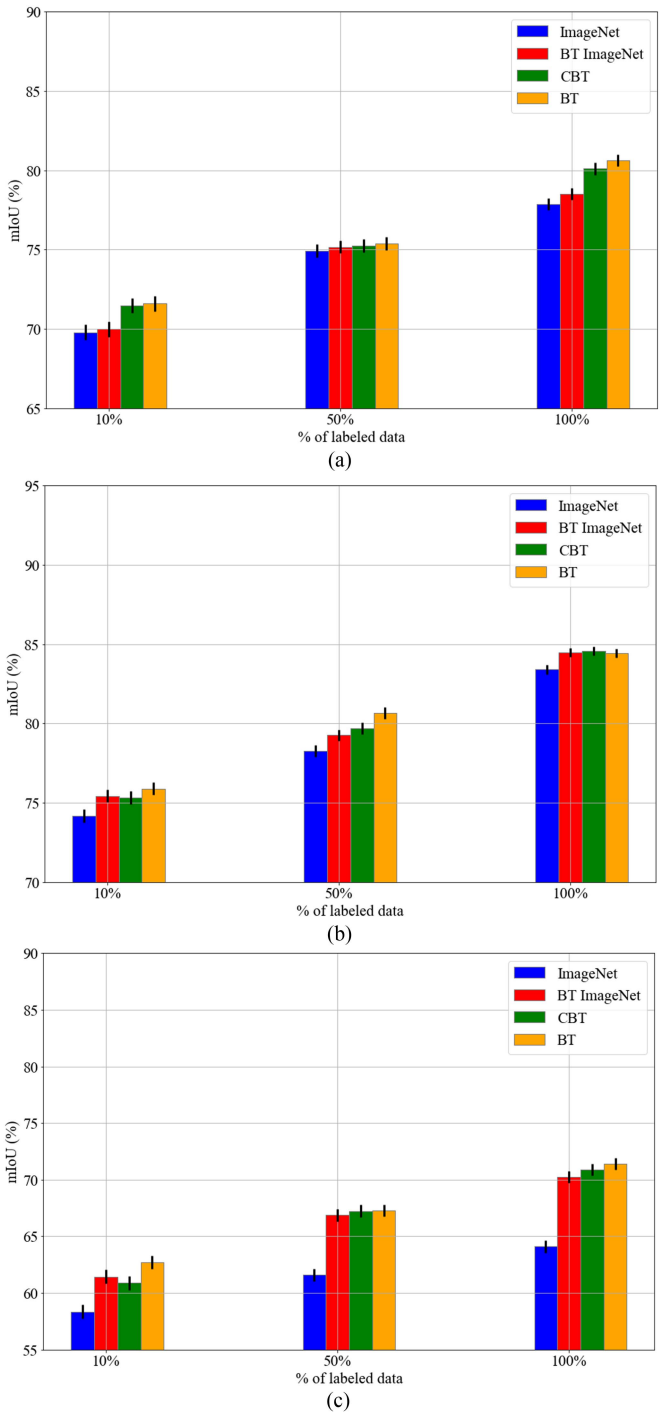

Fig. 3. mIoU metrics on experiments with an increasing amount of labeled data of, respectively, (a) UAVid, (b) US3D, and (c) Potsdam.

1) training of the only US3D;
2) joint training of US3D and UAVid;
3) training of the three datasets together.

This strategy is employed to create a salient feature extractor for all datasets, with the mean of avoiding catastrophic forgetting, in situations of incremental data availability. On the other hand, for CBT, step 1) is referred to training on US3D, step 2) on UAVid, and step 3) on Potsdam, as previously stated. We

TABLE III
UAVID RESULTS FOR DIFFERENT % OF TRAINING DATA

Encoder	%	mIoU	F1
ImageNet	10%	69.79 ± 0.48	80.32 ± 0.34
	50%	74.92 ± 0.42	84.60 ± 0.28
	100%	77.87 ± 0.37	86.67 ± 0.25
BT ImageNet	10%	69.99 ± 0.47	81.33 ± 0.32
	50%	75.17 ± 0.39	84.81 ± 0.28
	100%	78.51 ± 0.38	86.70 ± 0.26
CBT	10%	71.48 ± 0.47	81.50 ± 0.31
	50%	75.25 ± 0.42	84.73 ± 0.29
	100%	80.12 ± 0.39	88.42 ± 0.26
BT	10%	71.60 ± 0.48	81.64 ± 0.35
	50%	75.39 ± 0.41	84.85 ± 0.27
	100%	**80.64 ± 0.37**	**88.44 ± 0.25**

The highest score is marked in bold. The second highest is underlined. The tab reports the mean and the standard deviation of three experiments.

can easily affirm, observing Table II that our method could save nearly 50% of times, when the data are available incrementally. It is also interesting to state that, also in case of complete and immediate availability of all data, the computational times are comparable (28.75 h for CBT versus 24.61 h for BT, where the computational times of the latter consist of just the third step).

B. Unmanned Aerial Vehicle Semantic Segmentation Dataset

According to the results shown in Table III and represented in Figs. 3(a) and 4(a), when using a limited amount of data, the performance of the supervised pretrained encoder is inferior overall. Looking at Fig. 4(a), we can see that a better encoder, when 10% are used, leads to a more stable and effective training. On the other hand, the training with 50% of data is the most similar along the different pretrained encoders, with just $\sim 0.5\%$ mIoU gap between the worst result (74.92% mIoU, obtained with ImageNet encoder) and the best (75.39% mIoU, achieved with BT encoder) that it is almost negligible considering the standard deviations of the results. This trend can be mainly explained by what has been stated above with respect to the image domain. In fact, the images captured by drone are definitely more similar to close-range camera taken images, like ImageNet ones, than those from the other two RS datasets. This is mainly due to the point of view from which the images were captured. For Potsdam and US3D, the viewpoint is almost nadiral, while for UAVid it is oblique, more precisely the camera angle is set to around 45° to the vertical direction, at a flight height of about 50 m. As also stated by the authors [70], a non-nadiral view allows easier reconstruction of object geometry (i.e., volume, shape, etc.,), making the use of more sophisticated feature extractors less effective. These reasons favor the high performance of ImageNet pretrained models, especially with a limited number of labels. However, by increasing the number of labels, the models with encoders trained on the proposed datasets are able to match their features to the best conformation to solve the task with the best performance (80.64% mIoU with respect to 77.87% mIoU of the ImageNet encoder experiment). In addition, the fact that the pretraining on UAVid was the second of the three steps slightly affected the performance of the CBT pretraining strategy

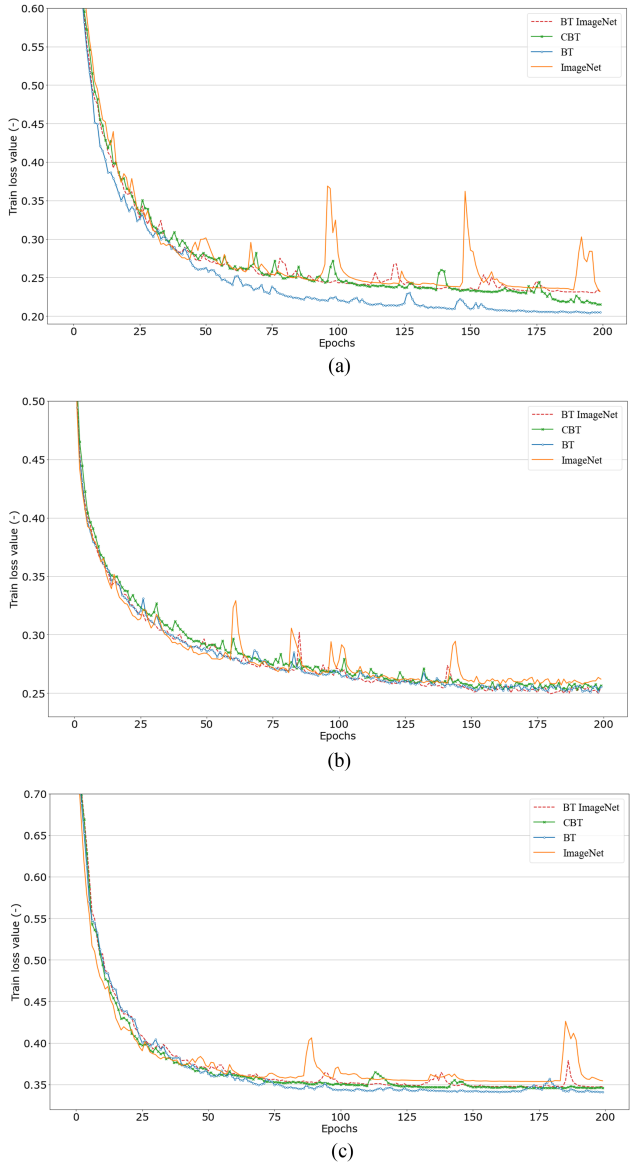

Fig. 4. Value of the loss on experiments with 10% labeled data of, respectively, (a) UAVid, (b) US3D, and (c) Potsdam.

(80.12% mIoU), with a very limited drop in performance ($\sim$ 0.5%).

Focusing on the qualitative results, reported in Fig. 5, we can confirm some considerations. First of all, we can see how some classes are difficult to be distinguished. Both in Fig. 5(a) and (b), especially when few labels are employed (i.e., 10%), building, road, and background clutter are confused, based on their similar radiometric information. Considering the differences among different pretrained strategies, ImageNet encoder does not work badly, especially if compared to the BT ImageNet one. On the other hand, the features learned on the three datasets are very beneficial for some classes, such as trees. In fact, this class is shared among the three datasets (e.g., see Fig. 6). For this reason, in Fig. 5(a), we can point out that CBT and BT better segment trees, when few labels are used.

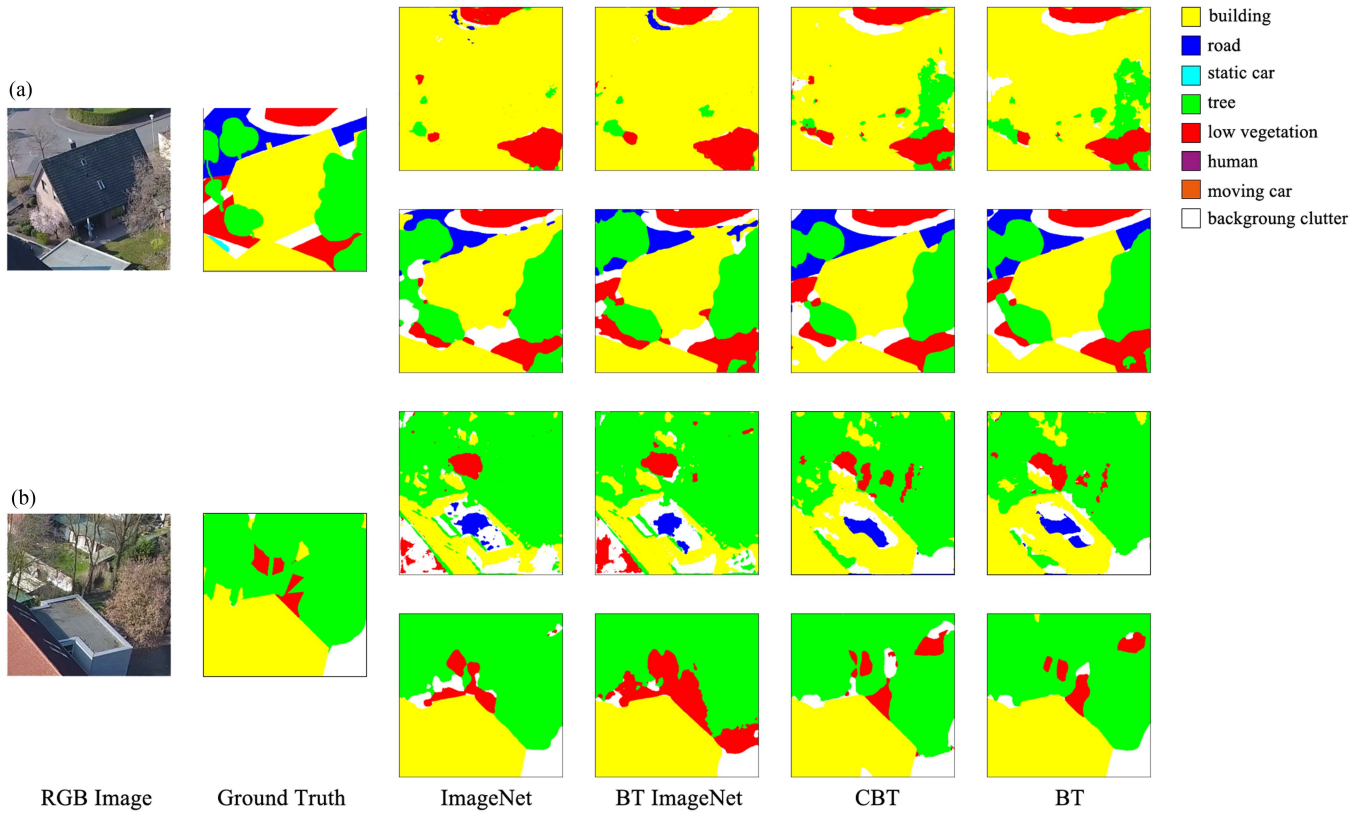

Fig. 5. Some results on the UAVid dataset, when using both 10% of (first row) and 100% (second row) of labeled data.

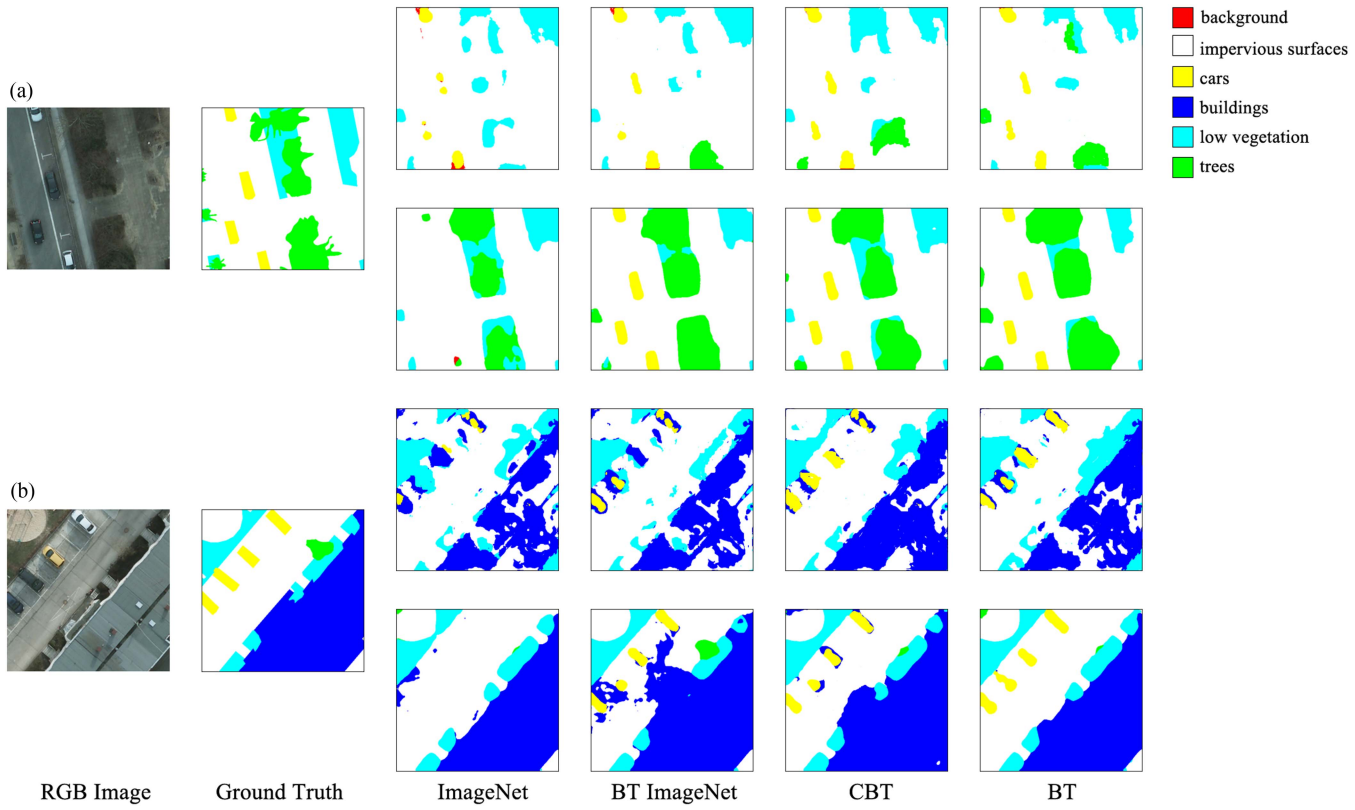

Fig. 6. Some results on the Potsdam dataset, when using both 10% of (first row) and 100% (second row) of labeled data.

TABLE IV
POTSDAM RESULTS FOR DIFFERENT % OF TRAINING DATA

Encoder	%	mIoU	F1
ImageNet	10%	58.36 ± 0.63	70.73 ± 0.50
	50%	61.59 ± 0.57	73.55 ± 0.44
	100%	64.12 ± 0.54	75.98 ± 0.40
BT ImageNet	10%	61.45 ± 0.59	73.25 ± 0.51
	50%	66.88 ± 0.55	77.41 ± 0.42
	100%	70.22 ± 0.52	79.42 ± 0.39
CBT	10%	60.90 ± 0.62	72.46 ± 0.52
	50%	67.24 ± 0.55	77.87 ± 0.44
	100%	70.90 ± 0.52	80.01 ± 0.39
BT	10%	62.72 ± 0.60	74.44 ± 0.51
	50%	67.29 ± 0.54	77.77 ± 0.43
	100%	**71.42 ± 0.52**	**80.63 ± 0.39**

The highest score is marked in bold. The second highest is underlined. The tab reports the mean and the standard deviation of three experiments.

TABLE V
US3D RESULTS FOR DIFFERENT % OF TRAINING DATA

Encoder	%	mIoU	F1
ImageNet	10%	74.18 ± 0.41	84.60 ± 0.26
	50%	78.25 ± 0.37	87.57 ± 0.14
	100%	83.41 ± 0.30	90.76 ± 0.12
BT ImageNet	10%	75.44 ± 0.39	85.38± 0.27
	50%	79.26 ± 0.35	87.89 ± 0.13
	100%	84.49 ± 0.28	91.27 ± 0.11
CBT	10%	75.31 ± 0.41	85.42 ± 0.27
	50%	79.70 ± 0.36	88.20 ± 0.14
	100%	**84.56 ± 0.28**	91.28 ± 0.12
BT	10%	75.89 ± 0.39	85.69 ± 0.25
	50%	80.67 ± 0.36	88.94 ± 0.13
	100%	84.43 ± 0.28	**91.30 ± 0.12**

The highest score is marked in bold. The second highest is underlined. The tab reports the mean and the standard deviation of three experiments.

C. Potsdam

As far as the results on Potsdam are concerned, in Table IV and Figs. 3(c) and 4(c), we see that the gap between results with self-supervised (71.42% mIoU) and supervised encoder (64.12% mIoU) is the largest among all experiments. This trend is true also for the experiments with a limited amount of training data, as the curves of Fig. 4(c) show. For example, with 10% of data, the gap between the ImageNet encoder (58.36% mIoU) and BT encoder (62.72% mIoU) is $\sim$4.4%. This can be explained by the fact that this dataset is the one with the least amount of data among the three available (see Table I). This insight is supported by the fact that the gap between performance with ImageNet encoders and performance with CBT and BT encoders is wider also for the other datasets when only 10% of the data is used (see also Fig. 3). Therefore, it is definitely the one that benefits the most from a more efficient encoder feature selection. This could be confirmed by the fact that the Potsdam domain is comparable with that of US3D, a very wide dataset, capable of improving the representations used for Potsdam semantic segmentation, confirming similar intuitions reached, for example, in [74].

Also in the qualitative results (see Fig. 6), we can observe that ImageNet works the worst among the pretrained encoders. Two are the main phenomena to be observed: ImageNet cannot detect some underrepresented classes (e.g., cars), as it poorly reconstructs the shapes. Compared to it, also BT ImageNet has poor reconstruction capabilities. On the other hand, both CBT and BT show good results. Also in this case, this is due to the fact that the proper RS finetuned features help the segmentation. For Potsdam, this is even more true for CBT, because of the order of training (US3D, UAVid, Potsdam). See, for example, the tree profile (similar to UAVid one) in Fig. 6(a) and the shape of the buildings (similar to US3D ones) in Fig. 6(b).

D. US3D

Also for US3D self-supervision leads to better results on downstream tasks. In this case, CBT performs best of all (CBT 84.56% versus ImageNet 83.41% mIoU) [see Table V and Figs. 3(b) and 4(b)]. Therefore, considering also the standard deviations of the final results, we observe that there are no significant differences in performance between the other encoders (BT ImageNet 84.49% versus CBT 84.56% versus BT 84.43% mIoU), since the US3D is a large dataset, composed of several images of the same area, captured from different points of view (i.e., multiview). This redundancy, working as data augmentation itself, facilitates the resolution of the task on this dataset, as once an efficient feature extractor is engaged, convergence is achieved quite effectively. This intuition is confirmed also by the training curves, showed in Fig. 4(b), where the training curves, except of some small instability for ImageNet one, follow a similar behavior. It is not surprising that similar conclusions are presented in [21], where self-supervision is applied on large datasets.

Looking at the qualitative results in Fig. 7, we can easily see that, when all the labels are available, the network reaches good performance, with any pretrained encoder. The situation changes when only 10% of labels are employed. Under this condition, ImageNet poorly identifies some underrepresented classes, as water and buildings. For US3D, we can affirm that a generical SSL approach, like BT ImageNet, could be considered enough for high performance. On the other hand, if we focus on some details, such as the shape of the inferred trees, we can see that CBT obtains slightly better performance, confirming the quantitative results (see Table V).

E. Overcoming Catastrophic Forgetting

CSSL, in addition to generically improving performance, has the main advantage of overcoming catastrophic forgetting. To illustrate this, we performed a series of experiments in which we trained a BT model in a sequential fashion, without introducing any CL strategy (called 3-step BT). Specifically, starting from a ResNet50 pretrained on ImageNet with BT, we performed three training steps, in which we used the model obtained in the previous step:

- BT on US3D dataset;
- BT on UAVid;
- BT on Potsdam.

Finally, we used the resulting ResNet50 as the backbone of UNet++ model, for the semantic segmentation downstream task. Fig. 8 and Table VI show the effectiveness of CBT as a strategy to overcome catastrophic forgetting, making possible to train a

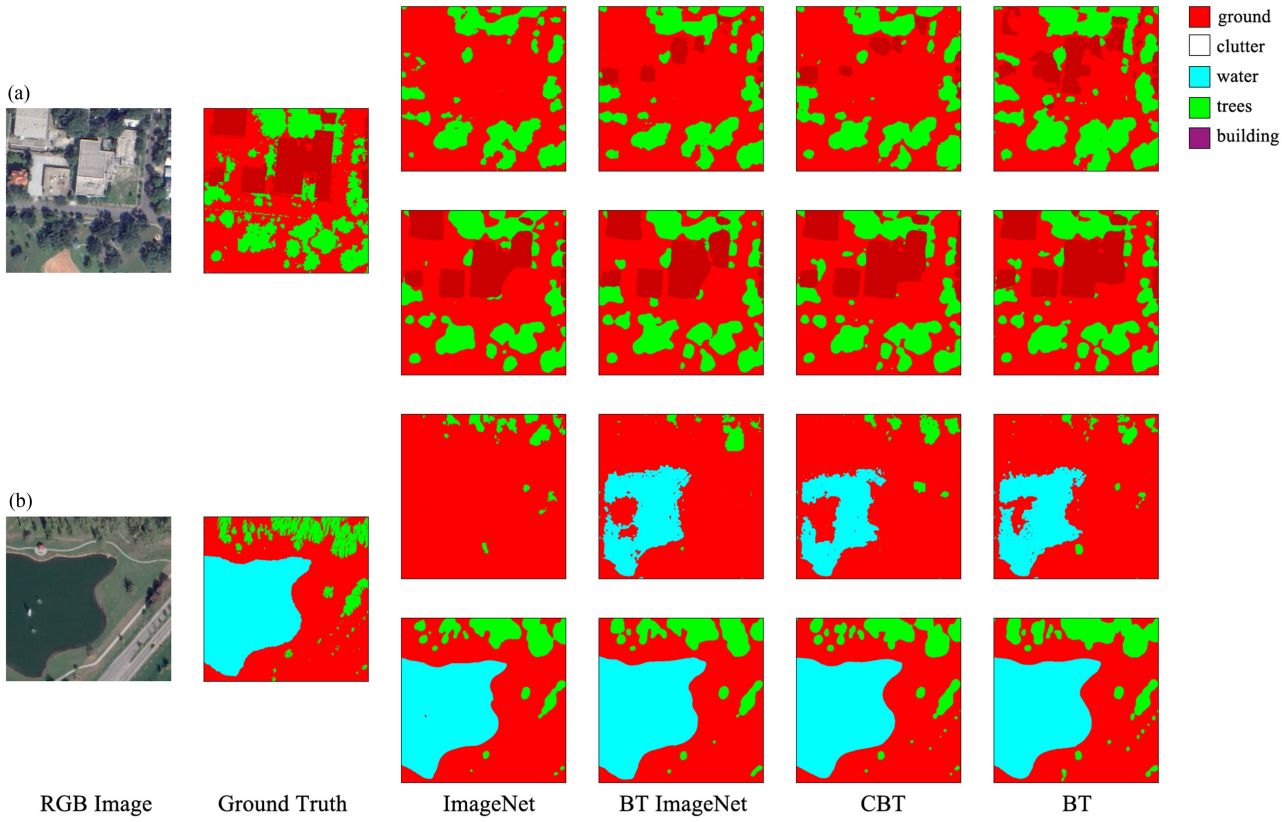

Fig. 7. Some results on the US3D dataset, when using both 10% of (first row) and 100% (second row) of labeled data.

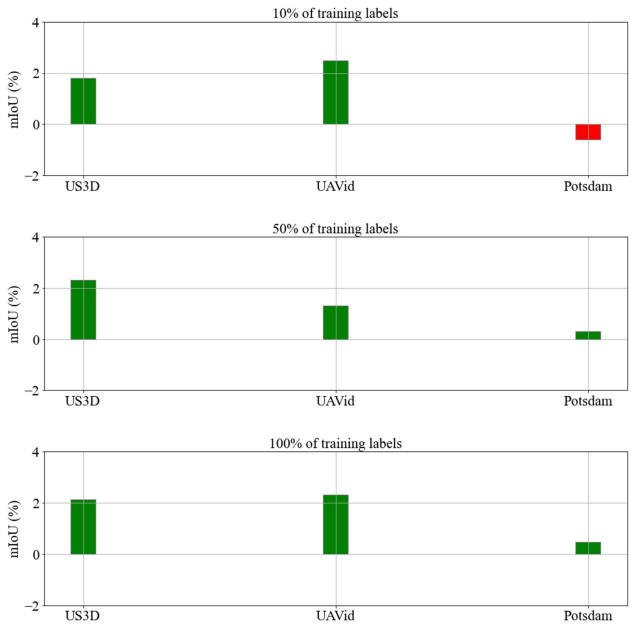

Fig. 8. Differences of mIoU among the experiments obtained with the encoder pretrained with the proposed CBT and 3-step BT, the catastrophic forgetting baseline (i.e., encoder pretrained with a vanilla BT sequentially trained on the three datasets).

TABLE VI
MIOU VALUES FOR EXPERIMENTS WITH DIFFERENT AMOUNTS OF LABELED DATA WITH THE ENCODER PRETRAINED WITH: I) CBT, II) A VANILLA BT TRAINED CONSECUTIVELY ON THE THREE DATASETS

Dataset	Encoder	mIoU (%)		
		10%	50%	100%
US3D	CBT	75.31	79.70	84.56
	BT	73.50	77.39	82.41
UAVid	CBT	71.48	75.25	80.12
	BT	68.98	73.93	77.81
Potsdam	CBT	60.90	67.24	70.90
	BT	61.51	66.94	70.42

powerful encoder, without the need of relying on all the data simultaneously. Particularly, we can affirm that constraining the parameters of the model pretrained on ImageNet is an easy and effective strategy to train the backbone. In fact, when it is not possible to rely on a vast amount of data specifically shaped for EO tasks, it is better to exploit the capabilities of models pretrained on extensive datasets [14], [15], like in this scenario.

Moreover, we can see in Fig. 8 and Table VI that once again UAVid is the dataset that, being more different from the others, suffers most from catastrophic forgetting (e.g., drop of $\sim 2.5\%$ when 10% of labels are used, $\sim 2\%$ when 100% of labels). In fact, as we have already observed, Potsdam and US3D have both nadiral views, making their characteristics more similar. For this very reason, the performance of 3-step BT on US3D is never excessively worse than the counterpart trained with CBT, even though the average performance drop (of $\sim 1.5\%$) is not negligible, being the first of the three dataset used. On the other hand, as one can expect, the performance on the Potsdam dataset, with the 3-step BT, are really similar to the one reached with

CBT. In fact, being the last dataset on which the algorithm is trained, most of the knowledge of the encoder came from this dataset. This is true especially when few data are used, where 3-step BT performance (61.51% mIoU) overcomes the CBT one (60.90% mIoU). In general, once again, given the required computational power and the overall performance, CBT seems the best solution to have consistent results on all the datasets.

VII. Conclusion

In this article, we have shown that the combination of CL and SSL offers an optimal compromise between performance and training efficiency and versatility for RS applications. In particular, we demonstrated a combined approach (CBT) leading to consistent performance in a novel combination of datasets with RS images that are heterogeneous in terms of sensors, resolution, acquisition, and scenes represented. Since the availability of unlabeled data is increasing at a great speed, and it is not possible for everyone to train repeatedly large models, a framework like CBT offers a potential solution. However, more work remains to be done. First, the validity of these results could be extended to new datasets and new tasks. Among the use of new datasets, we mention datasets containing multispectral images (i.e., not only with RGB bands). Second, other SSL and CL strategies can be combined into an effective and efficient framework.

References

[1] B. Zhang et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.

[2] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[3] M. Delange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[4] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[5] Y. Feng, X. Sun, W. Diao, J. Li, X. Gao, and K. Fu, "Continual learning with structured inheritance for semantic segmentation in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607017.

[6] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[7] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021.

[8] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[10] V. Stojnić and V. Risojević, "Evaluation of split-brain autoencoders for high-resolution remote sensing scene classification," in *Proc. Int. Symp. ELMAR*, 2018, pp. 67–70.

[11] Y. Wang et al., "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, 2022.

[12] C. Tao et al., "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, 2023.

[13] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 197–211.

[14] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlas: A large-scale, multi-task dataset for remote sensing image understanding," 2022, *arXiv:2211.15660*.

[15] M. Mendieta, B. Han, X. Shi, Y. Zhu, C. Chen, and M. Li, "GFM: Building geospatial foundation models via continual pretraining," 2023, *arXiv:2302.04476*.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[17] E. Fini, V. G. T. da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal, "Self-supervised models are continual learners," in *Proc IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

[18] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[19] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.

[20] N. A. A. Braham, L. Mou, J. Chanussot, J. Mairal, and X. X. Zhu, "Self supervised learning for few shot hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 267–270.

[21] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1182–1191.

[22] V. Marsocci, S. Scardapane, and N. Komodakis, "MARE: Self-supervised multi-attention REsu-Net for semantic segmentation in remote sensing," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3275.

[23] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, and P. Perez, "OBoW: Online bag-of-visual-words generation for self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6830–6840.

[24] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, 2021, Art. no. 2.

[25] N. Ammour, "Continual learning using data regeneration for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8012805.

[26] N. Ammour, Y. Bazi, H. Alhichri, and N. Alajlan, "Continual learning approach for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8000905.

[27] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "OpenEarthMap: A benchmark dataset for global high-resolution land cover mapping," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6254–6264.

[28] A. Toker et al., "DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21158–21167.

[29] L. Caccia and J. Pineau, "Special: Self-supervised pretraining for continual learning," *Continual Semi-Supervised Learning: First International Workshop, CSSL 2021, Virtual Event, August 1920, 2021, Revised Selected Papers*. Cham: Springer International Publishing, 2022.

[30] S. Purushwalkam, P. Morgado, and A. Gupta, "The challenges of continuous self-supervised learning," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 702–721.

[31] G. Lenczner et al., "Weakly-supervised continual learning for class-incremental segmentation," *IGARSS 2022-2022 IEEE Int. Geosci. Remote Sens. Symp.*, 2022.

[32] R. Peng, W. Zhao, K. Li, F. Ji, and C. Rong, "Continual contrastive learning for cross-dataset scene classification," *Remote Sens.*, vol. 14, no. 20, 2022, Art. no. 5105.

[33] A. S. Alakooz and N. Ammour, "A contrastive continual learning for the classification of remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 7902–7905.

[34] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "ClimaX: A foundation model for weather and climate," 2023, *arXiv:2301.10343*.

[35] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.

[36] Z. Zhang, X. Wang, X. Mei, C. Tao, and H. Li, "FALSE: False negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6518505.

[37] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5228011.

[38] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405516.
[39] F. Rottensteiner et al., "The ISPRs benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 1, no. 1, pp. 293–298, 2012.
[40] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10181–10190.
[41] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.
[42] 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 Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8004005.
[43] 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 Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.
[44] S. Vincenzi et al., "The color out of space: Learning self-supervised representations for earth observation imagery," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3034–3041.
[45] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.
[46] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, "Self-supervised representation learning for remote sensing image change detection based on temporal prediction," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1868.
[47] Y. Chen and L. Bruzzone, "Self-supervised change detection in multi-view remote sensing images," 2021, *arXiv:2103.05969*.
[48] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2021.
[49] X. Qian, T.-X. Jiang, and X.-L. Zhao, "SelfS2: Self-supervised transfer learning for sentinel-2 multispectral image super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 215–227, 2023.
[50] W. Huang, Y. Shi, Z. Xiong, Q. Wang, and X. X. Zhu, "Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 192–203, 2023.
[51] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6937–6956, Nov. 2014.
[52] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 371.
[53] S. Yun, H. Lee, J. Kim, and J. Shin, "Patch-level representation learning for self-supervised vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8354–8363.
[54] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021.
[55] Y. Liu et al., "Graph self-supervised learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5879–5900, Jun. 2023.
[56] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," 2022, *arXiv:2208.00173*.
[57] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, May 2022.
[58] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.
[59] Z. Zhu, P. Wang, W. Diao, J. Yang, H. Wang, and X. Sun, "Few-shot incremental learning with continual prototype calibration for remote sensing image fine-grained classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 210–227, 2023.
[60] Z. Zheng, X. Nie, and B. Zhang, "Fine-grained continual learning for SAR target recognition," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 2207–2210.
[61] K. Shen, X. Yang, S. Lolli, and G. Vivone, "A continual learning-guided training framework for pansharpening," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 45–57, 2023.
[62] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3524–3537, Sep. 2019.
[63] O. Tasar, Y. Tarabalka, and P. Alliez, "Continual learning for dense labeling of satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4943–4946.
[64] H. Li et al., "CLRS: Continual learning benchmark for remote sensing image scene classification," *Sensors*, vol. 20, no. 4, 2020, Art. no. 1226.
[65] S. D. Bhat, B. Banerjee, S. Chaudhuri, and A. Bhattacharya, "CILEA-NET: Curriculum-based incremental learning framework for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5879–5890, 2021.
[66] W. Liu, X. Nie, B. Zhang, and X. Sun, "Incremental learning with open-set recognition for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622916.
[67] A. Alqahtani and N. Ammour, "Continual learning using efficientnet and data generation for remote sensing image classification," in *Proc. 3rd Int. Conf. Distrib. Sens. Intell. Syst.*, 2022, pp. 222–228.
[68] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," 2023, *arXiv:2302.00487*.
[69] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, 2019.
[70] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 165, pp. 108–119, 2020.
[71] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1524–1532.
[72] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.
[73] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.
[74] C. J. Reed et al., "Self-supervised pretraining improves self-supervised pretraining," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2584–2594.

Valerio Marsocci received the B.Sc. and M.Sc. degrees in environmental engineering and the Second Level master's degree in big data in 2019 all from the Sapienza University of Rome, Rome, Italy, in 2016 and 2018.

After working as a Data Scientist in a startup, he enrolled in a Ph.D. program in Data Science (2019–2023) with the Sapienza University of Rome. During the Ph.D., he had two visiting periods with the University of Crete, Heraklion, Greece, and with Institut Géographique National, Paris, France.

Simone Scardapane (Member, IEEE) received the B.Sc. degree in computer engineering from Roma Tre University, Rome, Italy, in 2009 and the M.Sc. degree in artificial intelligence and robotics and the Ph.D. degree in information and communication technology from the Sapienza University of Rome, Rome, Italy, in 2011 and 2016, respectively.

He was a Software/Web Developer, for one year. He is currently an Assistant Professor with the Sapienza University of Rome. His research interests include distributed machine learning and adaptive audio processing.

Open Access funding provided by 'Universit? degli Studi di Roma "La Sapienza" 2' within the CRUI CARE Agreement