Multi-label Guided Soft Contrastive Learning for Efficient Earth Observation Pretraining

Yi Wang, *Student Member, IEEE*, Conrad M Albrecht, *Member, IEEE*, Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—Self-supervised pretraining on large-scale satellite data has raised great interest in building Earth observation (EO) foundation models. However, many important resources beyond pure satellite imagery, such as land-cover-land-use products that provide free global semantic information, as well as vision foundation models that hold strong knowledge of the natural world, are not widely studied. In this work, we show these free additional resources not only help resolve common contrastive learning bottlenecks, but also significantly boost the efficiency and effectiveness of EO pretraining.**

Specifically, we first propose soft contrastive learning that optimizes cross-scene soft similarity based on land-cover-generated multi-label supervision, naturally solving the issue of multiple positive samples and too strict positive matching in complex scenes. Second, we revisit and explore cross-domain continual pretraining for both multispectral and SAR imagery, building efficient EO foundation models from strongest vision models such as DINOv2. Adapting simple weight-initialization and Siamese masking strategies into our soft contrastive learning framework, we demonstrate impressive continual pretraining performance even when the input modalities are not aligned.

Without prohibitive training, we produce multispectral and SAR foundation models that achieve significantly better results in 10 out of 11 downstream tasks than most existing SOTA models. For example, our ResNet50/ViT-S achieve 84.8/85.0 linear probing mAP scores on BigEarthNet-10% which are better than most existing ViT-L models; under the same setting, our ViT-B sets a new record of 86.8 in multispectral, and 82.5 in SAR, the latter even better than many multispectral models. Dataset and models are available at https://github.com/zhu-xlab/softcon.

***Index Terms*—Remote sensing, Earth observation, foundation model, self-supervised learning, contrastive learning, continual pretraining, multispectral, SAR.**

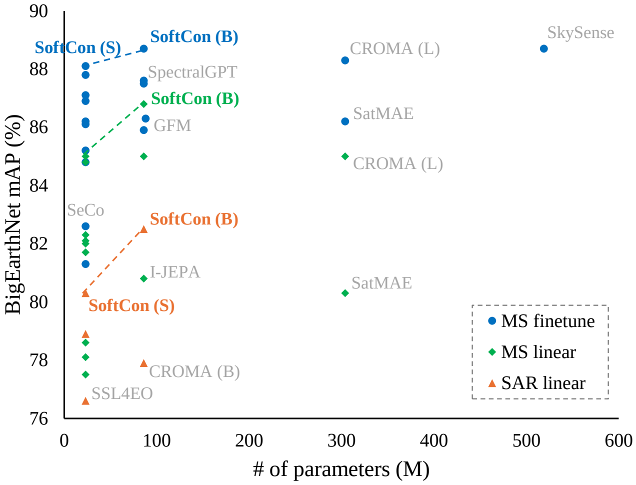


Fig. 1: A visual comparison of transfer learning performances on BigEarthNet-10%. SoftCon (ours) achieves SOTA results with lighter backbones on both linear probing and fine-tuning, and both multispectral and SAR. Our best multispectral linear result is comparable to best models' fine-tuning; our best SAR result outperforms many multispectral models. See Figure 5 for separated views of each setting.

I. Introduction

SELF-SUPERVISED learning has driven wide attention in pretraining Earth observation (EO) foundation models on large-scale satellite data [1]–[3]. While more and more efforts are spent on scaling up the data and model size with purely unsupervised pretraining, many other resources such as various land cover land use products tend to be overlooked. For example, ESA WorldCover [4], [5] provides the first global land cover maps for 2020 and 2021 at 10 m resolution, and Google Dynamic World [6] provides a continuous dataset of near-real-time land use land cover mapping. These dense products are highly correlated with commonly studied medium-resolution satellite imagery, and offer free semantic annotations with real-global coverage. Even though they are noisy at pixel-level due to semi-automatic creation process, they can be easily integrated into scene-level annotations with rather good quality. A similar concept has been demonstrated valid in supervised pretraining in GeoKR [7], where land cover products and geographical location are regarded as geographical knowledge to provide supervision. In this work, we will show the benefits of such auxiliary information in extending the popular contrastive self-supervised learning framework to build EO foundation models.

Contrastive learning with negative sampling such as SimCLR [8] and MoCo [9] have been shown robust and effective in EO pretraining [1], [10]. These methods pull together features of augmented views from the same anchor image (as positive sample), and push away features of other images (as negative samples). The negative samples are usually from a batch or a memory bank, and the model is trained to identify the positive sample from the negative ones. While such instance-discrimination-based methods learn good representations and have been widely used, they inevitably bear the risk of discriminating false negative samples. Specifically, in a large negative pool, there are likely samples that are not the

Y. Wang (yi4.wang@tum.de) and X. X. Zhu (xiaoxiang.zhu@tum.de) are with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM). X. X. Zhu is also with the Munich Center for Machine Learning. C. M. Albrecht (conrad.albrecht@dlr.de) is with the Remote Sensing Technology Institute, German Aerospace Center (DLR).

same as the anchor, but have very similar semantic information (e.g., belong to the same class). This issue is more significant in EO compared to natural images, as the Earth has a limited surface and the landscapes are usually very redundant.

Several methods have been proposed in computer vision to solve the false negative conflicts, of which one representative is supervised contrastive learning (SupCon) [11]. Leveraging image labels, SupCon defines positive samples as images belonging to the same semantic class. During training, samples from the same class are pulled together in the embedding space, while those from different classes are pushed away. Such a simple multi-positive design has been proven beneficial in natural images, and as we mentioned earlier, bears great potential to leverage the free land-use-land-cover annotations in EO. However, a satellite image usually contains more complex semantics than a single class, leading to at least multi-label annotations. Although one can define positive samples as perfect matching of each label component, this would result in two similar images with slightly different label distributions being forced apart (e.g. two neighboring urban scenes one with a small part of river and the other not).

To solve this problem, we extend SupCon to the multi-label scenario, proposing a novel soft contrastive learning method (we term as SoftCon) that takes into account the similarity of complex scenes. Specifically, we calculate cosine similarities of the normalized multi-label one-hot vectors across different scenes. Images with identical labels thus have the highest similarity scores, and images with similar labels have higher scores than images with very different labels. Then, we train the model to directly learn such cross-scene similarities by optimizing a soft contrastive loss on the cosine similarities of corresponding feature projections. In this way, semantically more similar images are pulled closer than semantically more dissimilar images, ending up with a smoothly distributed latent space. To prepare the training data, we match SSL4EO-S12 [10] images with Dynamic World [6] segmentation maps and integrate scene-level multi-label annotations, building a large-scale global multi-label classification dataset.

Meanwhile, another important resource that has huge potential in helping build EO foundation models lies in the general computer vision community: the vision foundation models. These models are exhaustively trained on huge amount of natural images, and have already gained strong knowledge of the visual world. Dating back to before the era of foundation models, ImageNet pretrained weights had been widely used and proved beneficial in many supervised Earth observation tasks. Similarly, they can also be used in unsupervised learning, leading to cross-domain continual pretraining. In this regard, recent works such as GFM [12] propose to build EO foundation models by distilling frozen ImageNet models. However, GFM-style training is limited to RGB images, restricting the flexibility to adapt to various EO sensors. Furthermore, the fast advances in computer vision have made available much stronger vision models than ImageNet supervised weights. To bridge this gap, we revisit the natural idea of simple weight initialization which has been preliminarily explored and verified effective in recent works [13], [14]. For unaligned input modalities, we simply leave the first layer's weights randomly initialized. In addition, to save hardware memory when continually training large Vision Transformers, we adopt Siamese masking inspired by masked Autoencoder [15]. Specifically, we randomly mask out a certain percentage of input patches on the trainable branch of the Siamese contrastive learning framework, and only send the remaining visible patches to the encoder. We show such simple but flexible continual pretraining strategies, when applied with strong vision foundation models such as DINOv2 [16], exhibit impressive effectiveness and efficiency even when the imaging sensors are completely different from the source domain.

Integrating the continual pretraining strategies into the soft contrastive learning framework, we efficiently train CNN and ViT foundation models that reach SOTA performances in 10 out of 11 downstream tasks. For example, our ResNet50/ViT-S (23M parameters, 100 epochs) achieve 84.8/85.0 linear probing mAP scores on BigEarthNet-10% which are better than most existing ViT-L models (300M parameters, 100-300 epochs) and comparable to CROMA and SkySense (≥600 epochs); under the same setting, our ViT-B (86M parameters) sets a new SOTA of 86.8 in multispectral, and 82.5 in SAR, the latter even better than most existing multispectral models.

In summary, our contributions are as follows:

- We explore the benefits of open resources beyond pure satellite imagery for efficient EO pretraining, producing multispectral and SAR foundation models that reach SOTA performances in 10 out of 11 downstream tasks.
- We propose soft contrastive learning that guides contrastive pretraining with land-cover-generated multi-label supervision.
- As a side product, we release a global multi-label scene classification dataset by matching noisy Dynamic World land cover maps with SSL4EO-S12 images and integrating good-quality multi-label annotations.
- We revisit cross-domain continual pretraining with simple weight initialization from strong vision foundation models and Siamese masking. We verify that even when the input modalities are not aligned, the strong knowledge can still be efficiently transferred to the target EO domain.

II. Related Work

A. Earth observation foundation models

Massive research has been conducted on the development of Earth observation (EO) foundation models. While there is also a line of supervised pretraining [17] on large-scale labeled datasets [18], [19], a majority of works tackle the technical adaptation of self-supervised pretraining techniques into EO domain. Early works focus on EO-specific data characteristics for contrastive view generation. For example, Tile2vec [20] proposed to pull together geospatially close tiles while pushing away far tiles. SeCo [21] proposed to use different seasons as augmented views for contrastive learning. CACo [22] proposed to perceive temporal changes with the spatiotemporal structure of remote sensing time series. Further, another group of works explores masked-image-modeling-based pretraining [23]. SatMAE [24] proposed temporal and spectral positional encoding for multispectral imagery and time series. RingMo [25] proposed less aggressive masking.

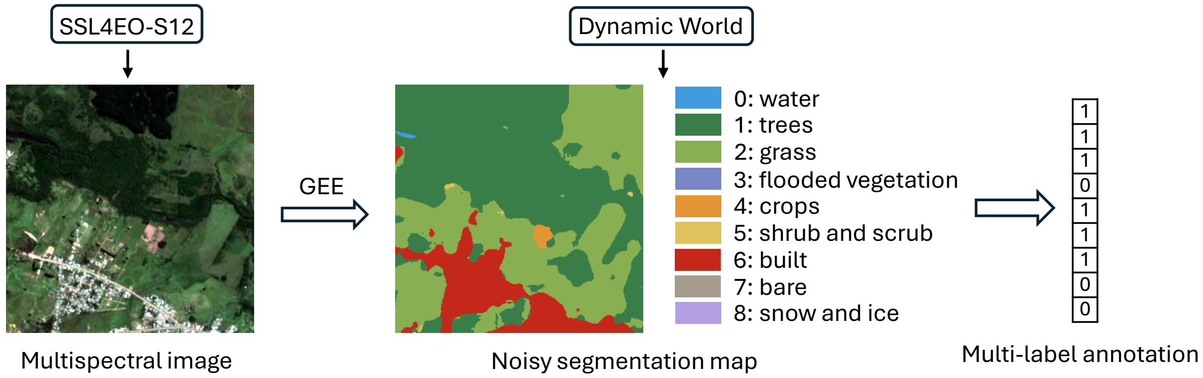


Fig. 2: The workflow of the multi-label dataset curation.

Scale-MAE [26] proposed GSD-based positional encoding and multi-scale reconstruction. SpectralGPT [27] proposed 3D masking to encode and reconstruct spectral data. FG-MAE [28] proposed to reconstruct remote sensing image features such as normalized difference indices. Going beyond a single modality and time stamp, CROMA [29], DeCUR [30], OFA-Net [31] and DOFA [32] investigated multi-modal multi-sensor pre-training, RemoteCLIP [33], SkyScript [34] and BITA [35] explored EO vision language pretraining, RingMo-sense [36], Presto [37] and Prithvi [38] studied EO time series, and SkySense [39] combined both modality and time sequence in a unified architecture, reaching SOTA performance in many downstream tasks. Meanwhile, another line of research targets the curation of EO pretraining datasets, such as SSL4EO-S12 [10] for Sentinel-1 and 2, SSL4EO-L [40] for Landsat series, and SatlasPretrain [41] for medium- and high-resolution satellite and aerial imagery with extensive annotations. While a general trend is to scale the data and model sizes with exhaustive training cost, important existing resources such as open annotations and vision foundation models tend to be overlooked. In this study, we share the conceptual insight with GeoKR [7] to use land-cover-generated multilabel annotations to guide EO pretraining, and with GFM [12] and DOFA to utilize vision models for continual pretraining.

B. Multi-positive contrastive learning

Contrastive learning beyond simple instance discrimination has been widely explored in computer vision. SupCon [11] proposed to use image labels for positive matching, extending contrastive learning to the fully-supervised setting. Almost at the same time, ML-CPC [42] proposed multi-label contrastive predictive coding, identifying multiple positive samples as a multi-label classification problem[1]. Further on, WCL [43] combined contrastive instance discrimination with SupCon by predicting pseudo weak labels. Sel-CL [44] extended SupCon to deal with noisy labels. All of these methods still deal with single-label datasets. To deal with more complex scenes, HiMulCon [45] presented a hierarchical representation learning framework that can leverage all available labels and preserve the hierarchical relationship between classes. MLS [46] proposed to assign multiple binary pseudo-labels for each input image by comparing its embeddings with those in two dictionaries, and training the model with binary cross-entropy loss. However, these two methods are still based on hard multi-positive matching, restricting the soft relationship between images with overlapped multi-label distributions. To tackle this issue, and to make the best use of real-world multi-label annotations, we propose soft contrastive learning that allows soft positive matching between different multi-label scenes.

C. Continual pretraining

Early works of continual pretraining were mainly developed in natural language processing to improve domain-specific large language models [47], [48]. In vision, existing works like [49] and [50] proposed hierarchical pretraining approaches for task adaptation, while not targeting task-agnostic representations. In remote sensing, CSPT [51] and TOV [52] proposed consecutive pretraining from natural images to remote sensing images, yet they are both limited to retraining on natural images for the source model. SpectralGPT [27] used a similar progressive pretraining pipeline between EO datasets to benefit from their unique advantages, yet it still requires first-stage pretraining from scratch. GFM [12] and DOFA [32] explored continual pretraining for generic EO foundation models by distilling knowledge from frozen vision backbones. However, it heavily relies on ImageNet weights whose prior knowledge is restricted compared to stronger vision foundation models. Meanwhile, GFM-style continual pretraining can only process RGB data, limiting the flexibility to adapt to various EO modalities such as multispectral imagery. To bridge this gap, we revisit and explore the simple cross-domain continual pretraining with weight initialization [13], [14] from strong vision foundation models such as DINOv2 [16]. For non-RGB imagery, we simply leave the input embedding layer randomly initialized. We show such a strategy, though naive, is both flexible and impressively effective in EO pretraining regardless of target sensors.

[1] We note that depending on the context, ”multi-label” in this paper may refer to two things: 1) the contrastive learning objective with multi-positive matching; 2) the input images are multi-label annotated.

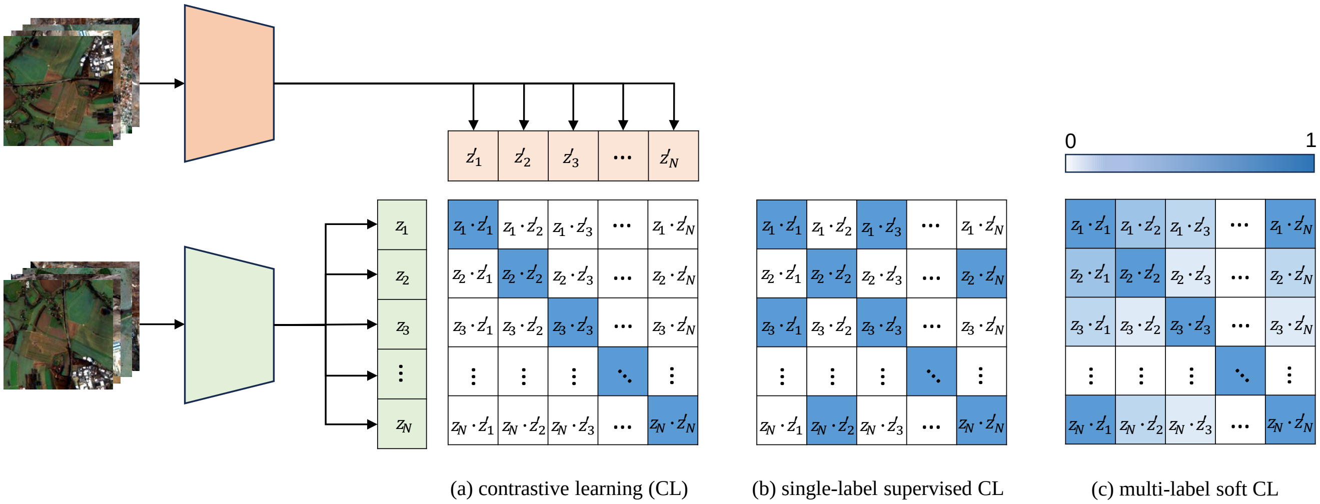


Fig. 3: Different contrastive learning designs. (a) The original contrastive learning performs strict instance discrimination where one anchor image has only one positive pair; (b) supervised contrastive learning allows multiple positive responses when images belong to the same class; (c) our proposed soft contrastive learning can effectively exploit multi-label annotations by assigning soft similarity scores to different pairs. SofCon is a more generic design that covers SupCon: when multi-label degrades to single-label, the soft similarities turn into binary scores, thus becoming multi-positive supervised contrastive learning.

III. Methodology

A. Building a large-scale multi-label dataset

We first build a global multi-label land-cover-land-use classification dataset by automatically matching multispectral/SAR imagery from a large-scale satellite dataset with open land cover products and integrating pixel-level labels to scene-level multi-label annotations.

We choose SSL4EO-S12 [10] as the source of satellite imagery, a multi-modal multi-temporal dataset specifically designed for large-scale self-supervised learning. The dataset consists of 4-seasonal Sentinel-1/2 SAR-optical images from 251,079 non-overlapping locations in the world, covering a wide range of geographical and temporal diversities. For accurate spatial and temporal matching of the scenes, we choose Dynamic World [6] to get the global land cover maps, which provide continuous land cover monitoring in 9 semantic classes. Both datasets are derived from Sentinel data stored in Google Earth Engine, and thus can be well aligned with the exact metainformation such as the acquisition time.

Figure 2 shows the general workflow of the curation of the multi-label dataset. Based on the geospatial coordinates and acquisition time, we match each SSL4EO-S12 L1C multispectral image with its corresponding Dynamic World land cover map in Google Earth Engine. Due to the effect of clouds, a few images do not have a corresponding segmentation map. This results in 247,377 locations on which there's a successful match for at least one season. Then, we gather the pixel labels into scene-level multi-labels for each image. As can be seen from the example scene in Figure 2, though the segmentation map is noisy, the scene-level semantic classes are rather accurate.

In total, we get 780,371 annotated multi-label images, each with size 264×264 in 10m resolution. We term this final dataset SSL4EO-S12-ML, and will release it for further research. Notably, SSL4EO-S12-ML can also be used as a large-scale multi-label benchmark dataset that covers the whole globe, complementing existing datasets like BigEarthNet [53]. For better reference, we provide a comprehensive dataset sheet and supervised benchmarking results in the appendix. In addition, we note that SSL4EO-S12-ML labels are derived from Sentinel-2 multispectral images, which are spatially well aligned with the corresponding Sentinel-1 SAR images, but temporally there can be shifts. Therefore, the SAR version is generally less accurate than the multispectral version. Nevertheless, it is enough to to guide our pretraining.

B. Multi-label guided soft contrastive learning

In this section, we introduce soft contrastive learning (SoftCon), which improves upon the original contrastive learning and supervised contrastive learning (SupCon) [11] to effectively utilize multi-label annotations.

Figure 3 provides a simplified illustration of the three different contrastive learning designs. Given N raw images, two batches of augmented views are generated and sent through the model to get the corresponding feature vectors $\{z_1, z_2, ..., z_N\}$ and $\{z'_1, z'_2, ..., z'_N\}$. In the original contrastive learning (Figure 3 a), strict instance discrimination is performed such that the cosine similarity of z_i and its augmented view z'_i is maximized, while the cosine similarities of z_i and all other features z'_j $(i \neq j)$ are minimized. This formulates an InfoNCE [54] loss which identifies one positive candidate from all the samples:

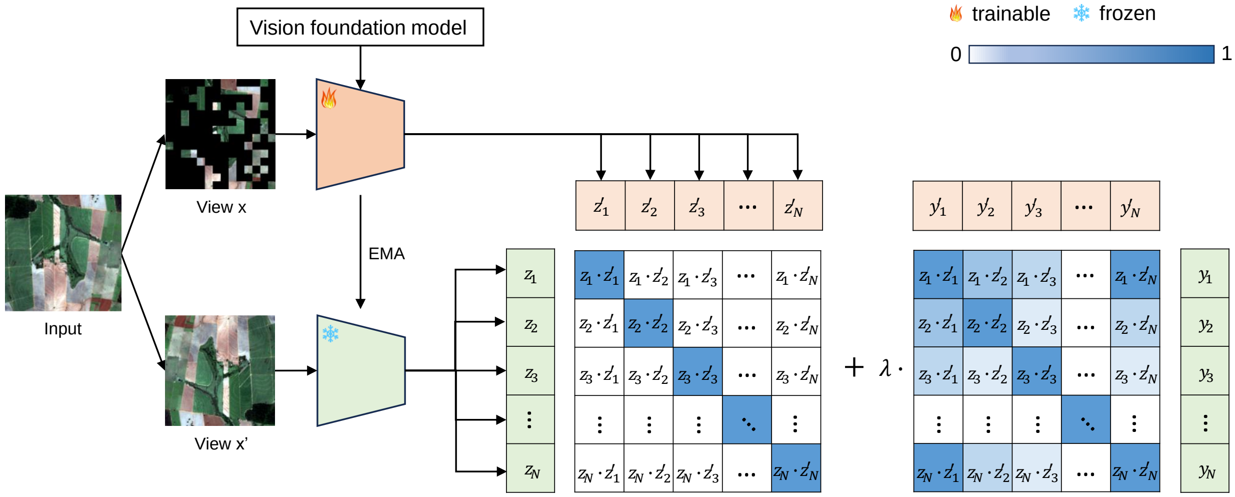

Vision foundation model
trainable
frozen
0
1
Input
View x
View x'
EMA
z'_1 z'_2 z'_3 ... z'_N
y'_1 y'_2 y'_3 ... y'_N
z_1 z_2 z_3 ⋮ z_N
y_1 y_2 y_3 ⋮ y_N
$z_1 \cdot z'_1$ $z_1 \cdot z'_2$ $z_1 \cdot z'_3$ ... $z_1 \cdot z'_N$
$z_2 \cdot z'_1$ $z_2 \cdot z'_2$ $z_2 \cdot z'_3$ ... $z_2 \cdot z'_N$
$z_3 \cdot z'_1$ $z_3 \cdot z'_2$ $z_3 \cdot z'_3$ ... $z_3 \cdot z'_N$
$z_N \cdot z'_1$ $z_N \cdot z'_2$ $z_N \cdot z'_3$ ... $z_N \cdot z'_N$
$+ \lambda \cdot$

Fig. 4: The general framework of SoftCon. Given a batch of input images, two batches of augmented views are parallelly sent through the two branches of a Siamese network. A similarity matrix is calculated based on the resulting two batches of feature vectors. A weighted sum of the contrastive and soft contrastive loss is optimized. For contrastive loss, this matrix should be close to Identity; for soft contrastive loss, this matrix should be close to the similarity matrix of the label vectors. We load vision foundation models for both the trainable (base encoder) and the frozen (momentum encoder) branches when initializing the model. During training, the weights of the momentum encoder are updated by exponential moving average (EMA) of the base encoder: $\theta_{\text{momentum}} \leftarrow m \cdot \theta_{\text{momentum}} + (1 - m) \cdot \theta_{\text{base}}$, where $m \in [0, 1)$ is a momentum coefficient. Siamese masking is only used with ViT backbones, where random patches of an image are masked out and only the visible patches are sent through the trainable branch.

$$\mathcal{L}^{\text{Contrast}} = -\sum_{i=1}^{N} \log \frac{\exp\left(z_i \cdot z'_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(z_i \cdot z'_j / \tau\right)} \tag{1}$$

where τ is softmax temperature. In practice, contrastive learning typically needs a large number of negative samples. Therefore, it inevitably faces the conflict that multiple images besides the augmented view can belong to the same class as the anchor image while the loss strictly pushes them apart.

SupCon tackles this conflict by introducing image labels to generalize contrastive learning to an arbitrary number of positives. As Figure 3 b shows, images belonging to the same class are pulled together, while images belonging to different classes are pushed away. Formally, SupCon accumulates multiple positive pairs while each item has a sole InfoNCE loss:

$$\mathcal{L}^{\text{SupCon}} = -\sum_{i=1}^{N} \frac{1}{N_p} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z'_p / \tau\right)}{\sum_{j=1}^{N} \exp\left(z_i \cdot z'_j / \tau\right)} \tag{2}$$

where $P(i) \equiv \{p \in \{1, 2, .., N\} : y_p = y_i\}$, y is the class label, and N_p is the total number of positive pairs for the anchor feature z_i. SupCon successfully alleviates the single-positive dilemma of the original contrastive learning. However, it only works with single-label imagery and the classes are mutually exclusive.

To effectively exploit multi-label annotations, we propose soft contrastive learning as is shown in Figure 3 (c). Samples with more similar label distributions are pulled closer than samples with more dissimilar labels, resulting in a soft positive matching. Specifically, we calculate the pair-wise cosine similarity matrix $Y \in \mathbb{R}^{N \times N}$ of the normalized multi-hot label vectors $\{y_1, y_2, ..., y_N\}$ and $\{y'_1, y'_2, ..., y'_N\}$. This is done through the dot product of the label vectors: $\mathbf{y}\mathbf{y}'^{\mathbf{T}}$. Similarly, we get the similarity matrix $X = \mathbf{z}\mathbf{z}'^{\mathbf{T}} \in \mathbb{R}^{N \times N}$ of the feature vectors $\{z_1, z_2, ..., z_N\}$ and $\{z'_1, z'_2, ..., z'_N\}$. Then, we optimize per-element binary cross entropy loss:

$$\begin{aligned}\mathcal{L}^{\text{SoftCon}} = -\sum_{i=1}^{N}\sum_{j=1}^{N} \Big(& Y_{ij} \cdot \log \sigma(X_{ij}) \\ & + (1 - Y_{ij}) \cdot \log\left(1 - \sigma(X_{ij})\right)\Big)\end{aligned} \tag{3}$$

where Y_{ij} is a soft score between 0 and 1, and $\sigma(\cdot)$ is the sigmoid function.

In practice, we follow MoCo-v2 [55] and MoCo-v3 [56] for the implementation of ResNet and ViT backbones, and combine the SoftCon loss with the Contrast loss as:

$$\mathcal{L} = \mathcal{L}^{\text{Contrast}} + \lambda \cdot \mathcal{L}^{\text{SoftCon}} \tag{4}$$

where λ is a weighting parameter. This combination is also similarly used by previous works in vision such as [46], [57]–[59]. However, our reason is conceptually different: our 9-class multi-label annotations are on a much coarser level than the real world, thus SoftCon alone may restrict the model's capacity to learn complex landscapes. Also, to prevent potential conflict optimization on the same feature embeddings, we use a separate projector for each objective.

C. Continual pretraining with Siamese masking

Finally, we introduce cross-domain continual pretraining into our framework to boost efficient Earth observation (EO) pretraining. Instead of sophisticated strategies like GFM [12] that are restricted by RGB input, we revisit the simple weight loading, initializing the backbone with strong vision foundation models. We use DINO [60] weights for ResNet backbones, and DINOv2 [16] weights for ViTs. As for the channel difference between natural RGB images and EO multispectral and SAR imagery, we simply let the input embedding layer remain randomly initialized. Our experiments empirically suggest that though naive, this strategy is both flexible and effective. In addition, to save hardware memory when continually training large vision Transformers, we adopt Siamese masking inspired by iBOT [61], MAE [15] and MSN [62]. Specifically, we randomly mask out a certain percentage of input patches on the trainable branch of the Siamese contrastive learning framework, and only send the remaining visible patches to the encoder. As the encoder only needs to process a portion of the full patches, both the memory and the training time can be reduced. Note that we do not conduct any reconstruction like MAE or iBOT, but rather view such masking as additional data augmentation. Different from MSN's non-negative online clustering, we verify it also works well in contrastive learning with negative sampling: with a reasonable masking ratio of about 20%, not only the efficiency, but also the performance can be improved. The full pretraining framework of SoftCon is illustrated in Figure 4.

IV. Implementation details

A. Pretraining

We pretrain SoftCon with ResNet [63] and ViT [64] backbones on the integrated multi-label dataset SSL4EO-S12-ML. We normalize the 16-bit multispectral images and the 32-bit SAR images to 8-bit with the mean and standard deviation provided in [10]. If there are multiple seasons for one scene, we randomly choose two for the base encoder and the momentum encoder, respectively. Data augmentations follow [10], including random crop (to the size 224×224), color jitter, greyscale, Gaussian blur, and random flip.

We adapt MoCo-v2 [55] for ResNet50 and MoCo-v3 [56] for ViT-small and ViT-base, with two separate projectors to get embeddings for the Contrast and the SoftCon loss, respectively. The weighting parameter λ in Equation (2) is 0.1. We set a queue size of 16384 for MoCo-v2, and a batch size of 1024 for both MoCo-v2 and MoCo-v3. The learning rate is warmed-up to 1.5e-4 for 10 epochs followed by cosine decay, and the optimizer is AdamW.

We load ResNet50 weights from DINO[2] [60] and ViT-S/14 and ViT-B/14 weights (without register) from DINOv2[3] [16], and conduct continual pretraining for 100 epochs. For ViTs, we randomly mask out 20% patches and send the remaining ones to the trainable encoder. Training is distributed in two nodes each with 4 NVIDIA A100 GPUs and takes 7-30 hours for different backbones. More details can be seen in the appendix.

B. Downstream tasks

We evaluate the pretrained backbones by linear probing and fine-tuning in 11 downstream tasks, including 6 land cover land use classification/segmentation datasets: BigEarthNet [53], BigEarthNet-SAR [65], EuroSAT [66], EuroSAT-SAR [28], fMoW-sentinel [24] and DFC2020 [67], one change detection dataset OSCD [68], and 4 multispectral datasets covering different applications from GEO-Bench [69]: m-so2sat, m-brick-kiln, m-cashew-plantation and m-SA-crop-type.

- BigEarthNet: a large-scale Sentinel-2 multi-label scene classification dataset covering 10 European countries. We use the version of 19 classes, and remove bad patches that are fully covered by seasonal snow or clouds. Following previous works [10], [21], [24], [29], [70], we train on 1% or 10% of the training split (31,166 images), and report micro mean average precision (mAP) on the full validation split (103,944 images).
- BigEarthNet-SAR: the paired Sentinel-1 SAR version of the BigEarthNet dataset. We train on 10% of the training split, and report mAP scores on the full validation split. The splits are aligned with the above multispectral version.
- EuroSAT: a 10-class scene classification dataset with 27,000 Sentinel-2 images collected from 34 European countries. Following [29], we use 16,200 training images and report overall accuracy on 5400 validation images.
- EuroSAT-SAR: the paired Sentinel-1 SAR version of the EuroSAT dataset.
- fMoW-sentinel: a large-scale scene classification dataset with 62 classes curated by matching fMoW [71] with Sentinel-2 images. Following [29], we use 10% of the training split (71,287 images) and report top-1 accuracy on the full validation split (84,939 images).
- DFC2020: a 10-class land cover semantic segmentation dataset that was originally collected for 2020 IEEE data fusion contest. We adjust the official test/validation data with 10m resolution labels for 5128 training and 986 testing images and report mean intersection over union (mIoU) scores.
- OSCD: a binary change detection dataset consisting of 24 pairs of multispectral images distributed worldwide. We use the official split: 14 training and 10 test pairs.
- m-so2sat: a subset of the 17-class So2Sat [72] dataset for local climate zone classification. We use the official split from GEO-Bench with 19992/986/986 train/val/test images and report test top-1 accuracies.
- m-brick-kiln: a subset of the brick-kiln [73] dataset for binary classification. We use the official split from GEO-Bench with 15063/999/999 train/val/test images and report test accuracies.
- m-cashew-plantation: a subset of the cashew-plantation [74] dataset for 7-class semantic segmentation. We use the official split from GEO-Bench with 1350/400/50 train/val/test images and report test mIoU scores.

[2] https://github.com/facebookresearch/dino

[3] https://github.com/facebookresearch/dinov2

TABLE I: Linear-probing / fine-tuning results on BigEarthNet-10% [53] and EuroSAT [66]. We report a comprehensive comparison with SOTA EO foundation models. MS/SAR/RGB represent data modalities. †: SoftCon starts from DINO/DINOv2 which were trained on ImageNet/LVD-142M. # indicates "the number of". *: SkySense employs a mixed architecture (in total 2B parameters) with ViT-L and Geo-context attention to encode MS images; 875K steps with batch size 240 roughly count to 1000 epochs. Left/right: linear/finetune. Best scores in **bold**.

	Pretrain dataset	# Pixels	Epochs	Backbone	# Param.	BE-10%	EuroSAT
Supervised	-	-	-	RN50	23M	83.4	98
MoCo-v2 [10], [55]	SSL4EO-S12 (MS)	70B	100	RN50	23M	82.1/86.2	98.0/99.1
DINO [10], [60]	SSL4EO-S12 (MS)	70B	100	RN50	23M	82.0/87.1	97.2/99.1
SeCo [21]	SeCo (MS)	70B	100	RN50	23M	78.6/82.6	-/93.1
CACo [22]	CACo (MS)	70B	100	RN50	23M	-/81.3	95.9/-
SoftCon (ours)	SSL4EO-S12 (MS)†	70B	100	RN50	23M	**84.8/87.8**	**98.6/99.3**
MoCo-v3 [10], [56]	SSL4EO-S12 (MS)	70B	100	ViT-S	23M	82.3/86.1	97.7/98.6
DINO [10], [60]	SSL4EO-S12 (MS)	70B	100	ViT-S	23M	81.7/86.9	97.7/99.0
MAE [10], [15]	SSL4EO-S12 (MS)	70B	100	ViT-S	23M	77.5/84.8	94.1/98.7
GFM [12]	GeoPile (RGB)	20B	100	Swin-B	88M	-/86.3	-
SpectralGPT [27]	fMoW+BigEarthNet (MS)	12B	300	ViT-B	86M	-/87.5	-/99.2
SatMAE [24]	fMoW (MS)	2.5B	200	ViT-L	304M	80.3/86.2	96.6/99.2
CROMA [29]	SSL4EO-S12 (MS,SAR)	140B	600	ViT-L	304M	85/88.3	**98.0/99.5**
SkySense [39]	SkySense (RGB,MS,SAR)	97T	1000*	ViT-L*	517M*	**-/88.7**	-
SoftCon (ours)	SSL4EO-S12 (MS)†	70B	100	ViT-S	23M	85/88.1	97.1/99.3
SoftCon (ours)	SSL4EO-S12 (MS)†	70B	100	ViT-B	86M	**86.8/88.7**	**98.0/99.5**

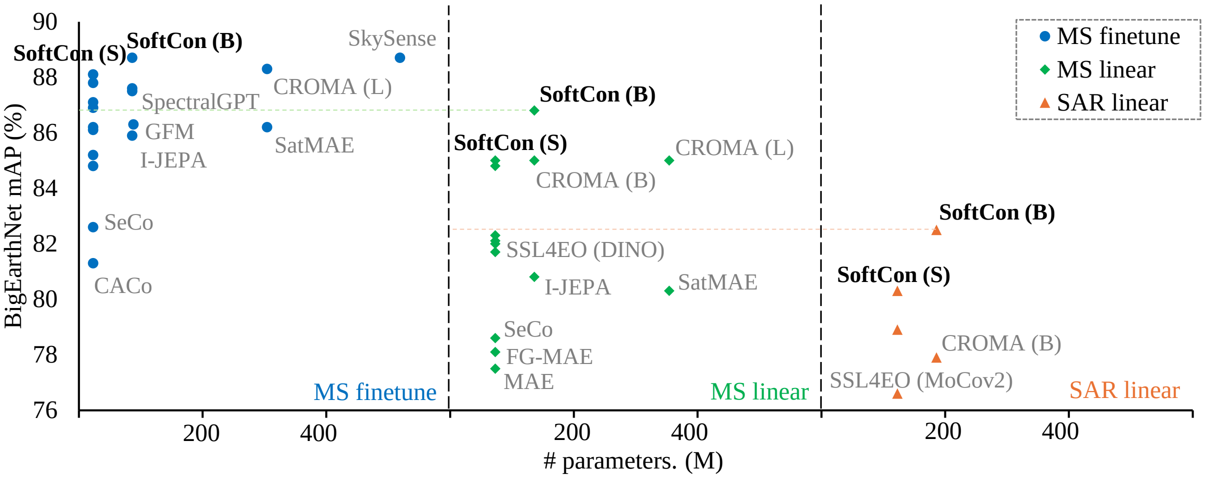


Fig. 5: A detailed comparison of transfer learning performances on BigEarthNet-10%. S/B/L represents ViT-small/base/large. SoftCon (ours) achieves SOTA results with lighter backbones on both linear probing and fine-tuning, and both multispectral and SAR. Our best multispectral linear result is better or comparable to many SOTA models' fine-tuning results; our best SAR result outperforms many multispectral models.

- m-SA-crop-type: a subset of the SA-crop-type [75] dataset for 10-class semantic segmentation. We use the official split from GEO-Bench with 3000/1000/1000 train/val/test images and report test mIoU scores.

We do a simple grid search for the learning rate for each dataset, using SGD or AdamW optimizer, and train each dataset for 30-100 epochs. We use a common input size 224×224 for all datasets. We freeze the encoder and train a linear layer or the decoder for all GEO-Bench datasets. See the appendix for hyperparameter details.

V. Results

A. Land cover classification

1) BigEarthNet and EuroSAT: We first report SoftCon results on BigEarthNet-10% [53] and EuroSAT [66] which are most commonly evaluated by SOTA EO foundation models. As can be seen from Table I, our models outperform most of the existing works in all scenarios. Specifically on BigEarthNet, our ResNet50 improves over other works with the same backbone by a large margin, with 2.7%/1.6% increase on linear/fine-tuning compared to the current best model, and 6.2%/5.2% increase compared to SeCo [21]. Our linear probing result for the first time outperforms fully supervised training from scratch. Notably, our small ResNet50 also outperforms many existing large ViT models in both linear and fine-tuning, verifying the effectiveness of ConvNet backbones.

For ViT backbones, our ViT-small is already comparable to the best existing models like SpectralGPT [27], CROMA [29] and SkySense [39]. Our ViT-base further pushes forward, achieving a new SOTA of 86.8% mAP on BE-10% with linear probing, and the same performance as SkySense in fine-tuning. More specifically, we achieve the same linear probing performance as CROMA [29] with single-modality pretraining (MS v.s. MS+SAR), much fewer model parameters (ViT-S v.s. ViT-L), and much shorter training epochs (100 v.s. 600). Similarly, we reach the fine-tuning performance of SkySense [39] with the above-mentioned advantages, plus much less pretraining data. For better visual comparison, we provide a scatter plot of various SOTA models' transfer learning performances w.r.t model size in Figure 5.

On EuroSAT [66], the best existing models can achieve a top-1 accuracy close to 99%. This means the dataset is becoming solvable with the fast technological development. As a result, it's less obvious to strictly compare the models' capacities. For example, our ResNet50 is even better than our ViT-base with much fewer parameters. Nevertheless, by comparing both linear and fine-tuning results, our models consistently outperform other works, achieving the same performance as CROMA [29] in both settings.

2) BigEarthNet-SAR and EuroSAT-SAR: Table II reports SoftCon linear probing results in the SAR modality on BigEarthNet-SAR [65] and EuroSAT-SAR [28] datasets. As we can see, our ResNet50 outperforms MoCo-v2 [55] by 2.3% on BigEarthNet-SAR with 10% labels, and by 4.7% on EuroSAT-SAR. Our ViT-S results are much better than MAE [15] and FG-MAE [28] with up to 11.6% improvement. This verifies again the advantage of contrastive learning over masked image model in producing out-of-the-box representations. Our ViT-B sets a new record of 81.4% on BigEarthNet-SAR, 3.5% better than CROMA [29]. **Notably, this score is already higher than many multispectral models** as compared to Table I, achieving a great breakthrough as it has always been very difficult for SAR to beat optical in cloud-free scenes. Our results confirm the great potential of SAR foundation models.

TABLE II: Linear probing results on BigEarthNet-SAR-10% [65] and EuroSAT-SAR [28].

	Backbone	BE-SAR-10%	EuroSAT-SAR
MoCo-v2 [10], [55]	RN50	76.6	82.4
SoftCon (ours)	RN50	**78.9**	**87.1**
MAE [10], [15]	ViT-S	69.8	79.3
FG-MAE [28]	ViT-S	71.7	80.7
CROMA [29]	ViT-B	77.9	87.5
SoftCon (ours)	ViT-S	80.3	87.1
SoftCon (ours)	ViT-B	**82.5**	**89.1**

3) fMoW-sentinel: We present linear probing and fine-tuning results on another more difficult land cover classification dataset fMoW-sentinel [24] in Table III. As the table shows, our ViT-small already achieves better performance than all existing models. Our ViT-base pushes further, with 4.8% better than the current best model CROMA [29] in linear probing, and 1.6% better in fine-tuning.

TABLE III: Linear probing/fine-tuning top-1 accuracy on fMoW-sentinel-10% [24].

	Backbone	Linear	Finetune
DINO [10], [60]	ViT-S/16	32.6	52.8
MAE [10], [15]	ViT-S/16	27.7	51.8
SatMAE [24]	ViT-S/16	35.2	57.2
I-JEPA [76]	ViT-S/16	32.4	53.5
CROMA [29]	ViT-L/8	39.2	59.0
SoftCon (ours)	ViT-S/14	39.9	59.7
SoftCon (ours)	ViT-B/14	**44.0**	**60.6**

B. Land cover segmentation

We report land cover semantic segmentation results on DFC2020 [67] as is shown in Table IV. We fine-tune DeepLabv3+ [77] for ResNet50 and UperNet [78] for ViT backbones. Promisingly, our ResNet50 outperforms MoCo-v2 [55] by 3%, and is also better than existing large ViTs. Our ViT-S outperforms CROMA [29] by 0.9%, which is further improved by our ViT-base with 0.3%. For visual comparison, we plot some example segmentation maps in Figure 6. As it shows, SoftCon captures more accurate and more fine-grained semantic information compared to SSL4EO-S12 [10].

TABLE IV: Fine-tuning mIoU results on DFC2020 [67].

	Backbone	mIoU
Supervised	RN50	42.9
MoCo-v2 [10], [55]	RN50	47.3
SoftCon (ours)	RN50	**50.3**
MAE [10], [15]	ViT-S	48.0
SatMAE [24]	ViT-L	44.1
CROMA [29]	ViT-L	49.8
SoftCon (ours)	ViT-S	50.7
SoftCon (ours)	ViT-B	**51.0**

C. Change detection

We report binary change detection results on OSCD [68] as is shown in Table Table V. We freeze the backbone and fine-tune a simple U-Net [79] for segmentation. The differences in feature maps between two timestamps are input to the network. SoftCon outperforms SeCo and SSL4EO in both recall and F1-score. The low precision score is due to the significant class unbalance, i.e.: predicting all pixels as unchanged would result in a good precision score. The combined F1-score highlights the superior performance of SoftCon.

TABLE V: Results with frozen encoders on OSCD [68].

	Backbone	Precision	Recall	F1
Rand. Init.	RN50	72.3	13.8	23.1
SeCo [21]	RN50	**74.9**	17.5	28.3
SSL4EO [10]	RN50	70.2	23.4	35.1
SoftCon (ours)	RN50	66.6	**29.2**	**40.6**

D. Other domain-specific applications

Finally, we evaluate SoftCon on four other domain-specific applications with the corresponding Sentinel-2 dataset collections from GEO-Bench [69]. These include two classification datasets m-so2sat and m-brick-kiln, and two segmentation datasets m-cashew-plantation and m-SA-crop-type. We conduct frozen-encoder training following [31], with a linear classifier for classification tasks and UperNet decoder for segmentation tasks. The results are shown in Table VI. We compare SoftCon with CROMA [29] and a recent work OFA-Net [31], and report the official fine-tuning (from timm [80]) results in [69] for reference. The table shows that SoftCon

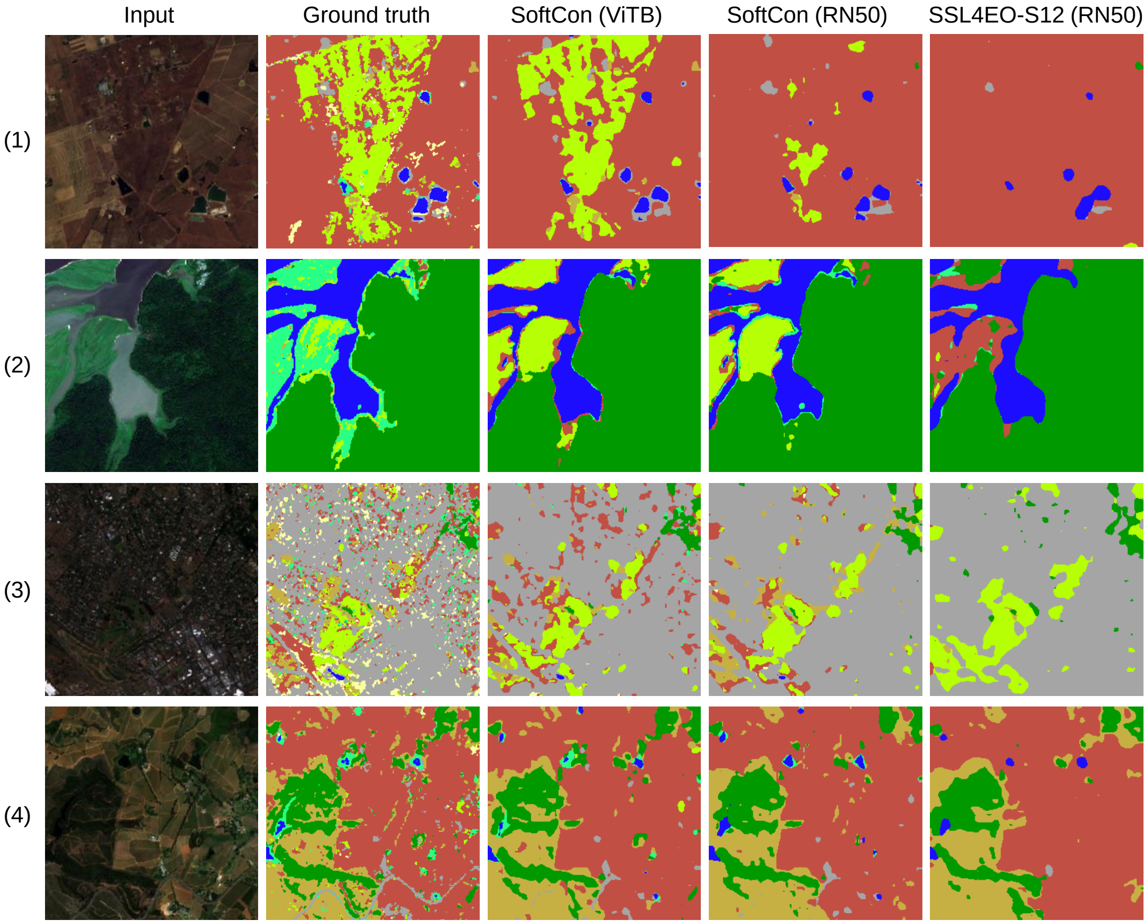


Fig. 6: Example segmentation maps on the DFC2020 [67] dataset.

TABLE VI: Transfer results with frozen encoders on four Sentinel-2 tasks in GEO-Bench [69]. We report top-1 accuracy/mIoU for classification/segmentation, respectively.

		m-so2sat	m-brick-kiln	m-cashew-plantation	m-SA-crop-type
GEO-Bench (FT)	RN50	52.8	98.7	44.7	29.9
CROMA [29]	ViT-B	49.2	91.0	-	31.4
OFA-Net [31]	ViT-B	46.0	91.3	37.4	**32.0**
SoftCon (ours)	ViT-S	49.9	92.6	44.4	31.5
SoftCon (ours)	ViT-B	**52.0**	**95.2**	**49.6**	31.5

significantly outperforms both CROMA and OFA-Net on three tasks, while only slightly worse than OFA-Net on m-SA-crop-type. Notably, our results with frozen-encoder outperform the official results with full fine-tuning on the two segmentation datasets, and only slightly worse on m-so2sat.

VI. Ablation and Discussion

For all ablation studies, we conduct linear probing experiments on BigEarthNet [53].

1) SoftCon loss: We ablate ResNet50 results on the soft contrastive loss in Equation (4) which are shown in Table VII. In line with our explanation in Section III-B, SoftCon alone is worse than Contrast alone since the 10 Dynamic World classes are too coarse-grained to fully represent the real-world semantics. However, combining the two losses provides significant benefits. When we degrade the SoftCon to SupCon in Equation (2), the performance drops as expected, while still better than contrastive loss alone. These results verify the effectiveness of using existing free annotations to boost Earth observation pretraining. Additionally, we ablate the weighting parameter λ in Equation (4) and suggest a best value of 0.1. Note all results in this ablation are without continual pretraining, which we will ablate next.

2) Continual pretraining: We ablate the continual pretraining strategy in Table VIII, which verifies the benefits of loading vision foundation models instead of pretraining from scratch. Also, we compare the performance of different vision models from ImageNet supervised weights to modern self-supervised weights. Interestingly, the stronger the vision model, the continual pretraining performance is also better. We report ImageNet top-1 linear scores of the corresponding models as a reference, which are well aligned with the cross-

TABLE VII: Ablation study on the SoftCon loss and the weighting parameter.

	BE-1%	BE-10%
Contrast only	78.9	82.1
SoftCon only	76.0	80.4
Contrast + SupCon	79.3	83.0
Contrast + SoftCon	**79.8**	**83.6**

	BE-1%	BE-10%
$\lambda = 0.01$	79.3	82.9
$\lambda = 0.1$	**79.8**	**83.6**
$\lambda = 0.5$	79.6	83.4
$\lambda = 1.0$	79.5	83.3

TABLE VIII: Ablation study on the source models of continual pretraining. *cont.* indicates continual pretraining.

	Backbone	BE-1%	BE-10%	ImageNet
w/o cont.	RN50	79.8	83.6	-
cont. (ImageNet)	RN50	80.2	83.9	-
cont. (MoCov3 [56])	RN50	80.9	84.2	74.6
cont. (DINO [60])	RN50	**81.4**	**84.8**	75.3
cont. (DINOv2 [16])	ViT-S	**82.6**	**85.0**	81.1

TABLE IX: Ablation study on the Siamese masking ratio.

	GPU memory	BE-1%	BE-10%
w/o masking	43G	82.3	84.5
masking (20%)	36G	**82.6**	**85.0**
masking (50%)	25G	81.7	84.4

TABLE X: Performance gains of different components under linear probing on BigEarthNet-10%. The explicitly introduced components are highlighted in **bold**.

Method	Backbone	Dataset	Performance gain
Contrast	RN50	SSL4EO-S12	-
(+ SupCon)	RN50	**+ multi-label**	+0.9
+ SoftCon	RN50	-	+0.6
+ cont.	RN50	(+ ImageNet)	+1.2
+ mask	ViT-S	-	+0.5

domain continual pretraining results.

In addition, we empirically find parameter-efficient fine-tuning (PEFT) techniques such as BitFit [81], prompt tuning [82] and LoRA [83], can be used for parameter-efficient continual pretraining, while not yet reaching a close performance as continually training all parameters. We report such preliminary results in the appendix.

3) Siamese masking: Finally, we ablate the Siamese masking strategy introduced mainly to boost training efficiency. As can be seen from Table IX, masking 50% of the patches can save almost half of the GPU memory without much performance drop. Interestingly, masking 20% of the patches can lead to even better results than no masking. This suggests that MAE-like random masking can be seen as an effective data augmentation strategy in contrastive learning. From another perspective, this design implicitly introduces the idea of masked image modeling that the model needs to know what information the masked patches have according to the unmasked branch, after which it knows the encoded features are similar. This could potentially also be one reason why the ideal masking ratio is much smaller compared to MAE, as the implicit masked image modeling plus the contrastive learning is a more challenging optimization task.

In summary, Table X provides an overview of the performance gains of each of our proposed components. It is worth noting that when introducing the multi-label annotation, the supervised contrastive learning method is implicitly introduced, which can be improved by our proposed soft contrastive learning as shown by the ablation in Table VII. Furthermore, when using the continual pretraining method, the vision dataset (e.g. ImageNet) is implicitly used for the vision model weights.

VII. Conclusion

This work revisits two important free resources beyond pure satellite imagery for efficient Earth observation pretraining: open land cover products and open vision foundation models. We build a large-scale multi-label dataset by matching an unsupervised pretraining dataset SSL4EO-S12 with Dynamic World land cover maps. To effectively utilize the multi-label annotations, we propose a novel soft contrastive learning method that allows soft matching between images with different label distributions. Meanwhile, we introduce strong vision models such as DINO and DINOv2 to a simple but flexible continual pretraining framework with Siamese masking. We efficiently train multispectral and SAR foundation models of both CNN and Transformer backbones that achieve SOTA performances in 10 out of 11 downstream tasks.

There are two main limitations of this work. First, our models target mainly medium-resolution data, while high-resolution images with more fine-grained semantic information remain to be explored. Second, retraining vision foundation models without any constraints inevitably makes the model forget knowledge from the source vision domain. Future work will explore more effective and flexible multimodal continual pretraining methods to build strong cross-domain cross-modal foundation models.

Acknowledgments

This work is jointly supported by the European Commission through the project “ThinkingEarth—Copernicus Foundation Models for a Thinking Earth” under the Horizon 2020 Research and Innovation program (Grant Agreement No. 101130544), by the Helmholtz Association through the Framework of HelmholtzAI, grant ID: ZT-I-PF-5-01 – Local Unit Munich Unit @Aeronautics, Space and Transport (MASTr), by the European Commission through the project ”EvoLand” under the Horizon 2020 Research and Innovation program (Grant Agreement No. 101082130), by the Excellence Strategy of the Federal Government and the Lander through the TUM Innovation Network EarthCare, by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab “AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001) and by Munich Center for Machine Learning. The compute is supported by the Helmholtz Association’s Initiative and Networking Fund on the HAICORE@FZJ partition.

References

[1] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *arXiv preprint arXiv:2206.13188*, 2022.

[2] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] X. X. Zhu, Z. Xiong, Y. Wang, A. J. Stewart, K. Heidler, Y. Wang, Z. Yuan, T. Dujardin, Q. Xu, and Y. Shi, "On the foundations of earth and climate foundation models," *arXiv preprint arXiv:2405.04285*, 2024.

[4] D. Zanaga, R. Van De Kerchove, W. De Keersmaecker, N. Souverijns, C. Brockmann, R. Quast, J. Wevers, A. Grosu, A. Paccini, S. Vergnaud, O. Cartus, M. Santoro, S. Fritz, I. Georgieva, M. Lesiv, S. Carter, M. Herold, L. Li, N.-E. Tsendbazar, F. Ramoino, and O. Arino, "Esa worldcover 10 m 2020 v100," Oct. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5571936

[5] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino, "Esa worldcover 10 m 2021 v200," Oct. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7254221

[6] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko *et al.*, "Dynamic world, near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, no. 1, p. 251, 2022.

[7] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[8] 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.

[9] 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*, 2020, pp. 9729–9738.

[10] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 3, pp. 98–106, 2023.

[11] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[12] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 806–16 816.

[13] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[16] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[17] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2022.

[18] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 14, pp. 4205–4230, 2021.

[19] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[20] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3967–3974.

[21] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[22] U. Mall, B. Hariharan, and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5261–5270.

[23] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022.

[24] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[25] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[26] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.

[27] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, X. Jia, A. Plaza *et al.*, "Spectralgpt: Spectral foundation model," *arXiv preprint arXiv:2311.07113*, 2023.

[28] Y. Wang, H. H. Hernández, C. M. Albrecht, and X. X. Zhu, "Feature guided masked autoencoder for self-supervised learning in remote sensing," *arXiv preprint arXiv:2310.18653*, 2023.

[29] A. Fuller, K. Millard, and J. Green, "Croma: Remote sensing representations with contrastive radar-optical masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decur: decoupling common & unique representations for multimodal self-supervision," *arXiv preprint arXiv:2309.05300*, 2023.

[31] Z. Xiong, Y. Wang, F. Zhang, and X. X. Zhu, "One for all: Toward unified foundation models for earth vision," *arXiv preprint arXiv:2401.07527*, 2024.

[32] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired foundation model for observing the earth crossing modalities," *arXiv preprint arXiv:2403.15356*, 2024.

[33] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[34] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "Skyscript: A large and semantically diverse vision-language dataset for remote sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5805–5813.

[35] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image-text alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[36] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang *et al.*, "Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[37] G. Tseng, I. Zvonkov, M. Purohit, D. Rolnick, and H. Kerner, "Lightweight, pre-trained transformers for remote sensing timeseries," *arXiv preprint arXiv:2304.14065*, 2023.

[38] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards *et al.*, "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.

[39] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," *arXiv preprint arXiv:2312.10115*, 2023.

[40] A. Stewart, N. Lehmann, I. Corley, Y. Wang, Y.-C. Chang, N. A. Ait Ali Braham, S. Sehgal, C. Robinson, and A. Banerjee, "Ssl4eo-l: Datasets and foundation models for landsat imagery," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[41] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, “Satlaspretrain: A large-scale dataset for remote sensing image understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 772–16 782.

[42] J. Song and S. Ermon, “Multi-label contrastive predictive coding,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 8161–8173, 2020.

[43] M. Zheng, F. Wang, S. You, C. Qian, C. Zhang, X. Wang, and C. Xu, “Weakly supervised contrastive learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 042–10 051.

[44] S. Li, X. Xia, S. Ge, and T. Liu, “Selective-supervised contrastive learning with noisy labels,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 316–325.

[45] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, “Use all the labels: A hierarchical multi-label contrastive learning framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 660–16 669.

[46] K. Zhu, M. Fu, and J. Wu, “Multi-label self-supervised learning with scene images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6694–6703.

[47] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” *arXiv preprint arXiv:2004.10964*, 2020.

[48] Z. Liu, G. I. Winata, and P. Fung, “Continual mixed-language pre-training for extremely low-resource neural machine translation,” *arXiv preprint arXiv:2105.03953*, 2021.

[49] A. Kalapos and B. Gyires-Tóth, “Self-supervised pretraining for 2d medical image segmentation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 472–484.

[50] C. J. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. Metzger *et al.*, “Self-supervised pretraining improves self-supervised pretraining,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2584–2594.

[51] T. Zhang, P. Gao, H. Dong, Y. Zhuang, G. Wang, W. Zhang, and H. Chen, “Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain,” *Remote Sensing*, vol. 14, no. 22, p. 5675, 2022.

[52] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, “Tov: The original vision model for optical remote sensing image understanding via self-supervised learning,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4916–4930, 2023.

[53] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, “Bigearthnet: A large-scale benchmark archive for remote sensing image understanding,” in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 5901–5904.

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

[55] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.

[56] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers. in 2021 ieee,” in *CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629.

[57] Y. Bai, X. Chen, A. Kirillov, A. Yuille, and A. C. Berg, “Point-level region contrast for object detection pre-training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 061–16 070.

[58] Z. Wang, Q. Li, G. Zhang, P. Wan, W. Zheng, N. Wang, M. Gong, and T. Liu, “Exploring set similarity for dense self-supervised representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 590–16 599.

[59] 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.

[60] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[61] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, “ibot: Image bert pre-training with online tokenizer,” *arXiv preprint arXiv:2111.07832*, 2021.

[62] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, “Masked siamese networks for label-efficient learning,” in *European Conference on Computer Vision*. Springer, 2022, pp. 456–473.

[63] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[65] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, “Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021.

[66] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[67] M. Schmitt, L. Hughes, P. Ghamisi, N. Yokoya, and R. Hänsch, “Ieee grss data fusion contest,” 2020.

[68] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, “Urban change detection for multispectral earth observation using convolutional neural networks,” in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. Ieee, 2018, pp. 2115–2118.

[69] A. Lacoste, N. Lehmann, P. Rodriguez, E. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin *et al.*, “Geobench: Toward foundation models for earth monitoring,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[70] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, “In-domain representation learning for remote sensing,” *arXiv preprint arXiv:1911.06721*, 2019.

[71] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, “Functional map of the world,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[72] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang *et al.*, “So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, 2020.

[73] J. Lee, N. R. Brooks, F. Tajwar, M. Burke, S. Ermon, D. B. Lobell, D. Biswas, and S. P. Luby, “Scalable deep learning to identify brick kilns and aid regulatory capacity,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, p. e2018863118, 2021.

[74] L. Yin, R. Ghosh, C. Lin, D. Hale, C. Weigl, J. Obarowski, J. Zhou, J. Till, X. Jia, N. You *et al.*, “Mapping smallholder cashew plantations to inform sustainable tree crop expansion in benin,” *Remote Sensing of Environment*, vol. 295, p. 113695, 2023.

[75] [Online]. Available: https://beta.source.coop/esa/fusion-competition/

[76] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[77] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[78] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[79] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[80] R. Wightman, “Pytorch image models,” https://github.com/rwightman/pytorch-image-models, 2019.

[81] E. B. Zaken, S. Ravfogel, and Y. Goldberg, “Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models,” *arXiv preprint arXiv:2106.10199*, 2021.

[82] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, “Visual prompt tuning,” in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[83] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

Appendix A
SSL4EO-S12-ML Dataset

A. General information

SSL4EO-S12-ML dataset is a large-scale multi-label land cover land use classification dataset. It consists of 780,371 multispectral Sentinel-2 images with size 264×264, divided into 247,377 non-overlapping scenes each with 1-4 multi-seasonal patches. Each image has a multi-label annotation from one or more categories in 9 land cover land use classes.

1) Data source: The Sentinel-2 images are from SSL4EO-S12 [10], a multi-modal multi-temporal dataset specifically designed for large-scale self-supervised learning. The dataset consists of 4-seasonal Sentinel-1/2 SAR-optical images from 251,079 non-overlapping locations in the world, covering a wide range of geographical and temporal diversities. The multi-label annotations are derived from Dynamic World [6], a near-real-time dataset that provides continuous pixel-level land cover monitoring in 9 semantic classes. The Dynamic World segmentation maps are automatically generated by algorithms developed on high-quality training data. We integrate the noisy maps into rather accurate scene-level classification labels.

2) Dataset curation: Based on the geospatial coordinates and acquisition time, we match each SSL4EO-S12 L1C multispectral image with its corresponding Dynamic World land cover map in Google Earth Engine. Due to the effect of clouds, a few images do not have a corresponding segmentation map, which are then dropped in our finall dataset. In total, there are 247,377 scenes with a successful match for at least one season. Then, we gather the pixel labels into scene-level multi-labels for each image, resulting in 780,371 labeled individual images. The workflow has ben shown in Figure 2.

B. Dataset statistics

1) Label number distribution: Figure 7 shows the distribution of the number of labels within each image. About 17% images contain a single label, while 70% images contain 4 or more labels.

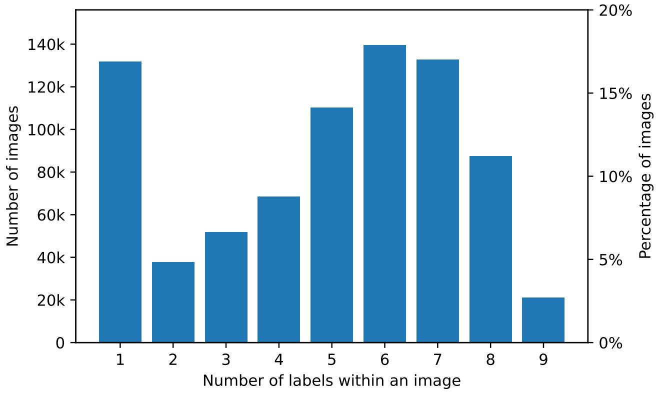


Fig. 7: The distribution of label numbers within each image.

2) Class distribution: Figure 8 shows the distribution of the number of images for each class. The number of images is well balanced in 7 common classes, while flooded vegetation and snow/ice are less represented.

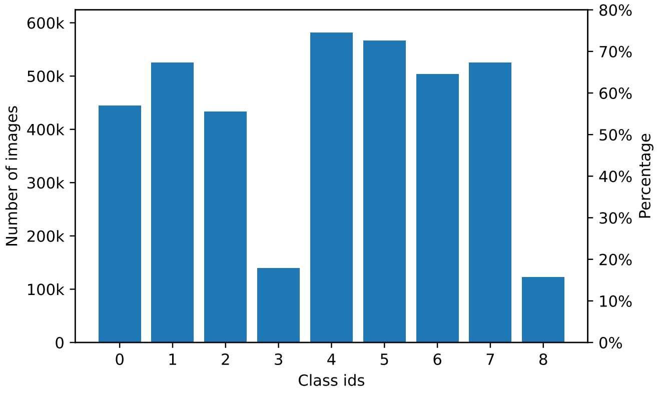


Fig. 8: The distribution of image numbers for different classes.

3) Season distribution: Figure 9 shows the distribution of the number of seasonal patches for each location. More than 40% locations have all 4 seasons, and more than 95% locations have at least 2 seasons.

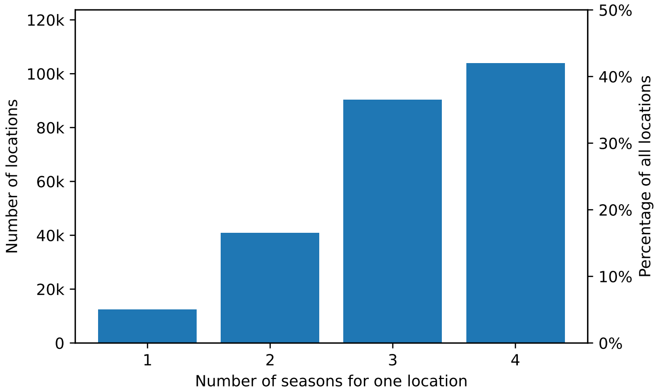
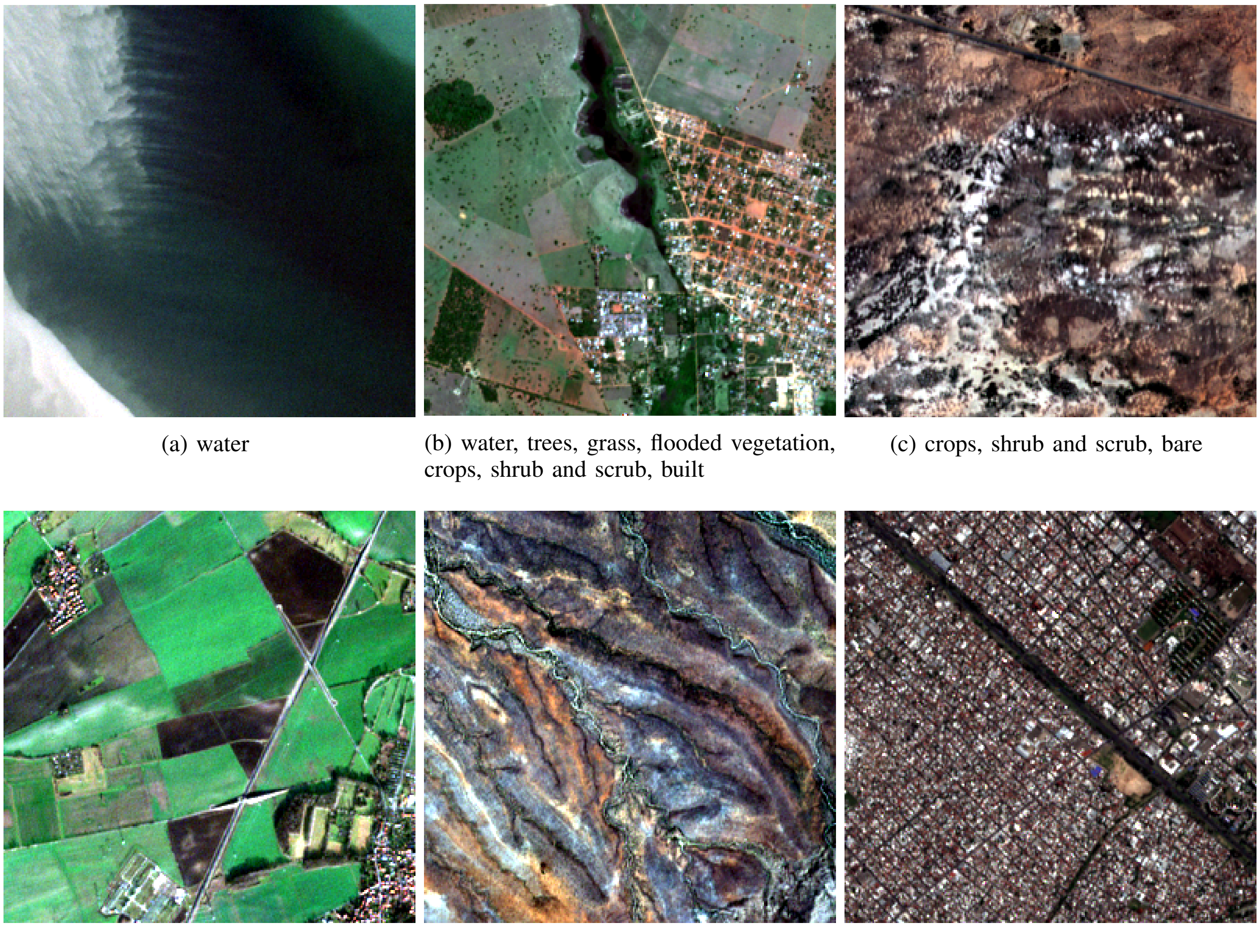


Fig. 9: The distribution of season numbers for each location.

C. Benchmark

We provide a preliminary benchmark on SSL4EO-S12-ML as a supervised multi-label classification task in Table XI. We split the dataset into 80% training data and 20% testing data according to the non-overlapping locations, and report micro and macro mAP on the test split. We use 10% of the training data and the full testing data for the benchmark. The table shows that this multi-label classification task is rather easy to solve, with a supervised micro mAP reaching 98.2%. This is similar to the training metric, indicating the balanced dataset split and rather good label quality. In addition, we also test the dataset as a downstream task to evaluate pretrained models. As is shown in the table, consistent performances as other popular datasets in the main paper are observed: 1) pretrained models improve upon random initialization; 2) SoftCon outperforms existing models such as SSL4EO. Therefore, SSL4EO-S12-ML can be considered as a global multi-label classification benchmark dataset that complements existing datasets such as BigEarthNet [53] which covers only Europe.



(a) water

(b) water, trees, grass, flooded vegetation, crops, shrub and scrub, built

(c) crops, shrub and scrub, bare

(d) water, trees, grass, crops, shrub and scrub, built, bare

(e) grass, crops, shrub and scrub, bare

(f) trees, grass, shrub and scrub, built, bare

Fig. 10: SSL4EO-S12-ML example images with multi-label annotations.

TABLE XI: Benchmark results of SSL4EO-S12-ML dataset. We use 10% of training data.

	mAP (micro)	mAP (macro)
rand. init.	93.7	84.4
supervised	98.2	94.7
SSL4EO (linear) [10]	95.8	88.7
SoftCon (linear)	96.5	90.2
SoftCon (fine-tune)	98.8	96.1

D. Data examples

Figure 10 provides some example images and corresponding multi-labels in the SSL4EO-S12-ML dataset covering different landscapes.

Appendix B
Parameter Efficient Continual Pretraining

Besides SoftCon weight initialization, we also explored other continual pretraining strategies, such as the adaptation of parameter efficient fine tuning (PEFT) techniques. PEFT is usually used in fine-tuning foundation models to specific downstream tasks with low cost. We studied the feasibility of adapting it to continual pretraining, termed parameter efficient continual pretraining (PECP), to build foundation models for a target domain such as Earth observation.

We study PECP with three representative PEFT techniques: bias-tuning like BiTFiT [81], visual prompt tuning [82], and low-rank adapter (LoRA) [83]. To fully evaluate the capacity of large foundation models, we transfer DINOv2 ViT-Large with about 300M parameters which is on par with commonly studied models in PEFT. To save computing costs and to rule out other factors, we use a simple MAE [15] for pretraining.

Table XII presents linear probing results on BigEarthNet with different pretraining strategies. Firstly, as the table shows, straightforward PEFT from DINOv2 does provide benefits compared to random frozen encoder. However, such benefits are rather limited compared to in-domain pretraining. Secondly, compared to pretraining from scratch, SoftCon-style continual pretraining offers significant improvement, especially when the in-domain data is limited in size: continual pretraining on 10% data is comparable to pretraining full data from scratch. Thirdly, all PECP strategies provide reasonable benefits with only a small fraction of parameters trainable, outperforming direct PEFT. Among them, LoRA performs slighter better than others. Similar to full continual pretraining, PECP also remains the performance when the size of pretraining data is restricted. Lastly, all the simple PECP strategies are not close to full continual pretraining, restricting the practical

TABLE XII: Linear probing mAP scores on BigEarthNet-10% with different pretraining strategies. We report pretraining with full SSL4EO-S12 data and 10% data. PEFT indicates directly transferring the DINOv2 model on BigEarthNet with PEFT techniques. *: SoftCon represents our SoftCon-style fully continual pretraining used in the main paper.

	pretrain module	# pretrain params.	pretrain 100%	pretrain 10%
rand. init.	-	-	64.4	-
supervised	-	-	74.7	-
PEFT (bias)	-	-	69.2	-
PEFT (lora)	-	-	69.2	-
pretrain from scratch	all	305M	79.2	72.8
full CP (SoftCon*)	all	305M	**81.0**	**78.1**
PECP (base)	patch embed	2.6M	70.6	69.1
PECP (bias)	patch embed + bias	2.9M	72.7	70.2
PECP (prompt)	patch embed + prompt	3.0M	74.0	72.1
PECP (lora)	patch embed + lora	5.8M	74.2	73.9

usage. This preliminary study motivates us to use full continual pretraining for SoftCon in the main papaer. However, with more research towards advanced designs, we believe PECP bears potential for future work, as the flexible adapters can pave the way towards unified cross-domain foundation models.

Appendix C
Implementation Details

A. Pretraining

We pretrain SoftCon with ResNet [63] and ViT [64] backbones on the proposed multi-label dataset SSL4EO-S12-ML, which consists of 247,377 scenes with 1-4 seasons. We normalize the 16-bit images to 8-bit with the mean and standard deviation provided in [10]. If there are multiple seasons for one scene, we randomly choose two for the base encoder and the momentum encoder, respectively. Data augmentations follow [10], including random crop (to the size 224×224), color jitter, greyscale, Gaussian blur, and random flip.

We adapt MoCo-v2 [55] for ResNet50 and MoCo-v3 [56] for ViT-S/14 and ViT-B/14, with two separate projectors to get embeddings for the Contrast and the SoftCon loss, respectively. Each projector consists of two linear layers and one ReLU activation function in between. The weighting parameter λ trading off the two losses is 0.1. We set a queue size of 16384 with a batch size of 256 for MoCo-v2, and a batch size of 1024 for MoCo-v3. For MoCo-v2, the learning rate starts from 0.03, followed by cosine decay to 0; for MoCo-v3, the learning rate is warmed-up to 1.5e-4 for 10 epochs, followed by cosine decay to 0. The optimizer is SGD for MoCo-v2 and AdamW for MoCo-v3. The softmax temperature is 0.2 for both MoCo-v2 and MoCo-v3.

We load ResNet50 weights from DINO [60] and ViT-S/14 and ViT-B/14 weights (without register) from DINOv2 [16], and conduct continual pretraining for 100 epochs. For ViTs, we randomly mask out 20% patches and send the remaining patches to the trainable encoder. All patches without masking are sent to the momentum encoder. Training is distributed in two nodes each with 4 NVIDIA A100 GPUs. Detailed compute and training time for different backbones are shown in Table XIII.

TABLE XIII: Compute and pretraining time.

Modality	Backbone	GPUs	Training time
MS	RN50	4xA100	21h
	ViT-S/14	4xA100	25h
	ViT-B/14	8xA100	15h
SAR	RN50	4xA100	7h
	ViT-S/14	4xA100	8h
	ViT-B/14	8xA100	7h

B. Downstream tasks

1) Downstream tasks: We evaluate the pretrained backbones by linear probing and fine-tuning in 8 downstream tasks, including 4 land cover land use classification/segmentation datasets: BigEarthNet [53], EuroSAT [66], fMoW-sentinel [24] and DFC2020 [67], and 4 multispectral datasets covering different applications from GEO-Bench [69]: m-so2sat, m-brick-kiln, m-cashew-plantation and m-SA-crop-type.

For classification datasets, we conduct linear probing and fine-tuning; for DFC2020, we fine-tune DeepLabv3+ [77] for ResNet backbone and UperNet [78] for ViT backbones; for GEO-Bench segmentation datasets, we freeze the encoder and train a UperNet decoder. For linear probing with ViT backbones, we pick either the output features from the last block, or the concatenation of features from the last 4 blocks following DINOv2 [16]. We do a simple grid search for the learning rate for each dataset. Specific hyperparameters for each dataset is summarized in the following tables.

TABLE XIV: DFC2020 fine-tuning hyperparameters.

	DFC2020 (ResNet)	DFC2020 (ViT)
Backbone	ResNet50	ViT-B/14, ViT-S/14
Input size	224x224	224x224
Augmentation	ResizedCrop (0.5,2.0), HorizontalFlip, VerticalFlip	ResizedCrop (0.5,2.0), HorizontalFlip, VerticalFlip
Batch size	16	16
Learning rate	1e-3	5e-3
LR schedule	poly	poly
Optimizer	SGD	AdamW
Weight decay	5.00E-04	5.00E-02
Warm-up	0	1000
Iter	20K	40K
Head	DeepLabv3+	UperNet

TABLE XV: Linear probing hyperparameters on BigEarthNet, EuroSAT and fMoW-S2.

	BigEarthNet	EuroSAT	fMoW-S2
Backbone	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14
Input size	224x224	224x224	224x224
Augmentation	ResizedCrop (0.8,1.0), HorizontalFlip	ResizedCrop (0.2,1.0), HorizontalFlip	ResizedCrop (0.2,1.0), HorizontalFlip
Batch size	256	256	1024
Learning rate	0.1	1e-3	4e-4
LR schedule	cos	step	cos
Optimizer	SGD	SGD	AdamW
Weight decay	0	0	0.01
Warm-up	0	0	0
Epoch	100	100	100
Feature block	4	1	4

TABLE XVI: Fine-tuning hyperparameters on BigEarthNet, EuroSAT and fMoW-S2.

	BigEarthNet	EuroSAT	fMoW-S2
Backbone	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14
Input size	224x224	224x224	224x224
Augmentation	ResizedCrop (0.8,1.0), HorizontalFlip, VerticalFlip	ResizedCrop (0.2,1.0), HorizontalFlip	ResizedCrop (0.2,1.0), HorizontalFlip, mixup&cutmix
Batch size	256	256	1024
Learning rate	1e-4	1e-3	4e-4
LR schedule	cos	step	cos
Optimizer	AdamW	SGD	AdamW
Weight decay	0.01	0	0.05
Warm-up	0	0	0
Epoch	50	100	50

TABLE XVII: ResNet linear probing (left) and fine-tuning (right) hyperparameters on BigEarthNet and EuroSAT.

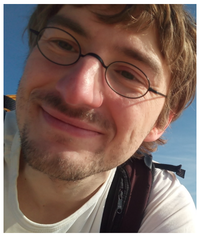
	BigEarthNet	EuroSAT
Backbone	RN50	RN50
Input size	224x224	224x224
Augmentation	ResizedCrop (0.8,1.0), HorizontalFlip	ResizedCrop (0.2,1.0), HorizontalFlip
Batch size	256	256
Learning rate	8	8
LR schedule	step	step
Optimizer	SGD	SGD
Weight decay	0	0
Warm-up	0	0
Epoch	100	100

	BigEarthNet	EuroSAT
Backbone	RN50	RN50
Input size	224x224	224x224
Augmentation	ResizedCrop (0.8,1.0), HorizontalFlip, VerticalFlip	ResizedCrop (0.2,1.0), HorizontalFlip
Batch size	256	256
Learning rate	1e-3	0.1
LR schedule	cos	step
Optimizer	AdamW	SGD
Weight decay	0.01	0
Warm-up	0	0
Epoch	50	100

TABLE XVIII: Transfer learning hyperparameters on the four GEO-Bench datasets.

	m-so2sat	m-brick-kiln	m-cashew-plantation	m-SA-crop-type
Backbone	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14	ViT-B/14, ViT-S/14
Input size	224x224	224x224	224x224	224x224
Augmentation	ResizedCrop (0.8,1.0), HorizontalFlip	ResizedCrop (0.8,1.0), HorizontalFlip	ResizedCrop (0.6,1.0), HorizontalFlip, VerticalFlip, Rotate	ResizedCrop (0.6,1.0), HorizontalFlip, VerticalFlip, Rotate
Batch size	256	256	64	64
Learning rate	1.0	10	0.01	0.01
LR schedule	cos	cos	cos	cos
Optimizer	LARS	LARS	AdamW	AdamW
Weight decay	0	0	0.01	0.01
Warm-up	0	0	3	3
Epoch	50	50	50	50

Yi Wang (S'21) received his B.E. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2018 and his M.Sc. degree in geomatics engineering from University of Stuttgart, Stuttgart, Germany, in 2021. He is pursuing his Ph.D. degree at the Technical University of Munich (TUM), Munich, Germany. From 2021 to 2024, he was a research associate at the Remote Sensing Technology Institute, German Aerospace Center (DLR). In 2020, he spent three months at the perception system group, Sony Corporation, Stuttgart, Germany. His research interests include self-supervised learning, weakly-supervised learning, and multimodal representations.



Conrad M Albrecht (M'17) is a researcher at the Earth Observation Center of the German Aerospace Center. Since April 2021, he is PI of the HelmholtzAI Young Investigator Group (YIG) "Large-Scale Data Mining in Earth Observation" (DM4EO). In July 2023, he was appointed a visiting associate professor with the Institute of Nasca at Yamagata University, Japan contributing to research in machine learning for the UNESCO World Heritage of the Nasca culture in Peru.

For over 6 years Conrad was a research scientist in the Physical Sciences department at the IBM T.J. Watson Research Center in Yorktown, NY, USA. While at the Institute for Theoretical Physics, he graduated in physics (International Max-Planck Research School for Quantum Dynamics in Physics, Chemistry and Biology) with an extra certification in computer science (Cluster- & Detector Management team at CERN, Switzerland). He received a corresponding Ph.D. degree from Heidelberg University, Germany in 2014 working on distributed computing to study physics at low temperatures.

Conrad's research agenda interconnect physical models and numerical analysis, employing Big Data technologies and machine learning through open-science research, https://conrad-m-albrecht.github.io. Conrad co-organized workshops at the IEEE BigData conference, IGARSS conferences, and the AAAS annual meeting. Conrad is home in Europe and the United States. Some of his transatlantic initiatives include: In 2023, he was invited by the Alexander-von-Humboldt Foundation to present at the German-American Frontiers of Engineering Symposium. Beginning in 2024, initiated by Helmholtz Imaging and Data Science, he is in close touch with the German Department at Princeton University for cross-atlantic undergraduate internships. For fall 2024, Conrad was invited Adjunct Professor of Earth and Environmental Engineering by Columbia University, NY, USA.

Xiao Xiang Zhu (S'10–M'12–SM'14–F'21) received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her "Habilitation" in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She is the Chair Professor for Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since May 2020, she is the PI and director of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond", Munich, Germany. Since October 2020, she also serves as a Director of the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, Zhu has been a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the head of the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport". Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. She is currently a visiting AI professor at ESA's Phi-lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu has been a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is a Fellow of the Academia Europaea (the Academy of Europe). She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020-2023) and Potsdam Institute for Climate Impact Research (PIK). She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition and served as the area editor responsible for special issues of IEEE Signal Processing Magazine (2021-2023). She is a Fellow of IEEE, AAIA, and ELLIS.