Cross-sensor self-supervised training and alignment for remote sensing

Valerio Marsocci, Nicolas Audebert

Cross-sensor self-supervised training and alignment for remote sensing

Valerio Marsocci, Nicolas Audebert

***Abstract*—Large-scale "foundation models" have gained traction as a way to leverage the vast amounts of unlabeled remote sensing data collected every day. However, due to the multiplicity of Earth Observation satellites, these models should learn "sensor agnostic" representations, that generalize across sensor characteristics with minimal fine-tuning. This is complicated by data availability, as low-resolution imagery, such as Sentinel-2 and Landsat-8 data, are available in large amounts, while very high-resolution aerial or satellite data is less common. To tackle these challenges, we introduce cross-sensor self-supervised training and alignment for remote sensing (X-STARS). We design a self-supervised training loss, the Multi-Sensor Alignment Dense loss (MSAD), to align representations across sensors, even with vastly different resolutions. Our X-STARS can be applied to train models from scratch, or to adapt large models pretrained on *e.g.* low-resolution EO data to new high-resolution sensors, in a continual pretraining framework. We collect and release MSC-France, a new multi-sensor dataset, on which we train our X-STARS models, then evaluated on seven downstream classification and segmentation tasks. We demonstrate that X-STARS outperform s the state-of-the-art by a significant margin with less data across various conditions of data availability and resolutions.**

***Index Terms*—self-supervised learning, remote sensing, multi-modality, pre-training**

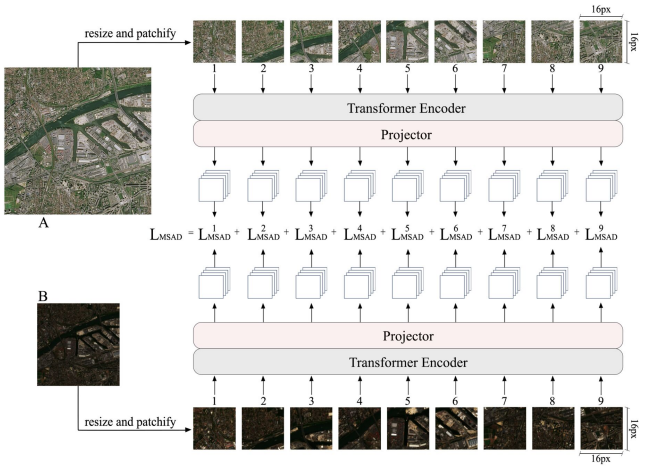

Fig. 1: Multi-sensor alignment dense (MSAD) loss aligns the features of the model applied on sensor B to the features of the model applied on sensor A. The alignment is applied patchwise. The images have different sizes and are resized before the application of the model.

I. Introduction

As computing and data are becoming more available, large-scale pretrained models are becoming more common for Earth Observation (EO) [1], [2]. Inspired by the "foundation models" that excel in natural language processing [3] and computer vision [4], large deep models have been tailored to EO data to consider the distinct challenges and opportunities of remote sensing (RS) and geospatial data [5]. One specificity of EO is that models should be *sensor aware* (or *sensor agnostic*). Indeed, RS data is sourced from numerous airborne and spaceborne sensors with different spatial resolutions, spectral bands and camera calibrations [1]. The ability to seamlessly adapt and exploit data from various sensors is crucial for ensuring consistent and reliable results, with models that generalize to new acquisitions [1]. It also alleviates resource constraints, as one should be able to efficiently fine-tune large pretrained models on a new sensor with minimal data, without retraining from scratch for every sensor. For these reasons, continual pretraining, *i.e.* adding pretraining stages to the optimization of generic models to integrate data, has gained a lot of traction for RS [6], [2]. By carefully training the model on new observations, better representations can be learned compared to full retraining, saving both time and resources.

One way to achieve sensor agnosticism is with multimodality. While most multimodal models deal with a combination of texts and images [7], EO offers many more possibilities: multisource and multispectral imaging, Synthetic Aperture Radar (SAR), vector geodata, ground-level pictures, etc. To date, most works in multimodal learning for RS focused on combining active (SAR) and passive (optical) sensors [8]. Yet, optical images from different sensors can contain very different information due to changes in resolution, wavelengths and altitude. Therefore, models trained on one optical sensor tend to poorly generalize when applied to another one. Furthermore, while a vast volume of Earth imagery is generated daily, access to open very high-resolution data remains limited. Most available datasets consist of mid to low-resolution images with a ground sampling distance (GSD) between 10 and 60 m, such as Sentinel-2 (10-60 m/px) or Landsat-8 (30 m/px). Adaptation techniques could allow for training large-scale models on low-resolution datasets and then adapting them on-demand to scarcer high-resolution data.

To this end, we introduce an algorithm for *Cross-sensor Self-supervised Training and Alignment of Remote Sensing data* (*X-STARS*). It can effectively fine-tune large-scale models trained on an optical sensor to another sensor, even with vastly different resolutions. X-STARS combines a contrastive self-supervised learning (SSL) with a novel sensor alignment objective: Multi-Sensor Alignment Dense loss (MSAD). The latter allows to capture the common semantics of local image patches independently from scale and sensors using knowledge distillation. Compared to masked image modeling (MIM) [9], [5], which requires large high-resolution datasets [2], X-STARS can deal with any image resolution. This allows us to improve

on models that have been trained on low-resolution images only, using a fraction of the high-resolution data used for MIM. X-STARS can be applied both as an end-to-end self-supervised for from-scratch pretraining objective, or as *a posteriori* fine-tuning loss for continual pretraining.

To train X-STARS, we collect Multi-Sensors Cities France (MSC-France), a new multimodal EO dataset that includes four sensors from low to very high-resolution imagery. We show that this approach outperforms existing pretrained models on downstream tasks across many different sensors and training setups (backbones, from-scratch training and continual pretraining). We establish a new state-of-the-art (SOTA) on various downstream tasks of classification and semantic segmentation of EO imagery. In summary, our contributions are:

- MSC-France, a novel dataset of ≈5000 image triplets covering most French cities with three different sensors (SPOT-6, Landsat-8, Sentinel-2), and a subset with several high-resolution pairs (SPOT-6, BDORTHO);
- Multi-Sensor Alignment Dense loss (MSAD), a cross-sensor contrastive loss that aligns representations learned from different sensors, even with different GSD;
- We train self-supervised models and improve performance on several downstream EO benchmarks with 10 to 100× fewer data compared to previous works and use continual pretraining to adapt pretrained models to new sensors, establishing new SOTA.

II. Related Works

A. Remote Sensing Self-Supervised Learning

RS is a fertile playground for SSL, as large amounts of (unlabeled) EO data are available worldwide. Many large-scale datasets have been published in the last few years, some labeled such as Dynamic EarthNet [10] for change detection, BigEarthNet for scene classification [11], OpenEarthMap for land cover mapping [12], and many others unlabeled [13], [9], [14]. SSL approaches from traditional computer vision have been adapted to address the specifics of EO [15], [16]. [17] adapts Online Bag-of-Words [18] to train without labels the backbone of a segmentation network. Contrastive learning is a staple of SSL for RS: SauMoCo [19] applies contrastive learning on spatially augmented views of an image; [20] train EO models with contrastive multiview coding [21] on three large datasets with both RGB and multispectral bands; [22] tunes the negative sampling strategy of contrastive learning to take into account the intra-class diversity of RS images; [23] adapts contrastive learning for semantic segmentation of EO images; Seasonal Contrast (SeCo) [13] uses contrastive learning by treating two images of the same area at different times as different views; CACo [24] extends this principle by incorporating more levels of seasonal differences. Other works use pretext tasks, such as inpainting [25] or reconstructing visible wavelengths from the other bands [26]. MIM is also popular: SatMAE [9] trains a masked autoencoder (MAE) [27], properly adapted to deal with multi-temporal data. Scale-MAE [5] deals with images of various resolutions using band filter decoding and scale-equivariant positional encoding.

While these approaches are effective in learning representations on unlabeled data, practical use remains limited to their poor generalization to sensors outside the training set. To this end, continual pretraining is a promising avenue. In particular, GFM [2] built upon the continual pretraining strategy from [6] to train geospatial "foundation models".

B. Multimodal learning

Multimodal learning, especially image and text, has been boosted by the release of large vision-language models such as CLIP [7]. The main challenge of multimodal learning is to learn *aligned* representations of the two modalities that preserve their common information, without discarding the modality-specific knowledge. In the following years, different works tried to tackle the limits with different strategies, such as softening the objective loss [28], [29], alternative losses [30] or combining the similarity loss with masking [31], [32]. RS has not escaped this trend [33], [34], [35], [36]. In RS, multimodality takes its root from data fusion and therefore is generally envisioned as mixing SAR and optical data [37], [8], [38] or optical data and vector geodata [39]. Recently, some works tried to use other combinations of data. For example, in [40], the authors propose a contrastive framework to compare street views and satellite images. However, even two different optical sensors can be considered as different modalities since their spectral and spatial characteristics can vary wildly. For example, [41] collects a high-resolution dataset from three different sensors (NAIP, Pleiades, SPOT-6) and argues for training a model with a combination of sensors for better generalization. In the same idea, [42] and [43] introduce datasets in which satellite images are enhanced with multi-view images, from different sources (*e.g.* airborne or street view). Yet, as multimodal data might not be available at training time, we design our approach so that it can be applied both to train from scratch or as a post-hoc adaption to fit an existing model on a new sensor.

C. Continual pretraining and knowledge distillation

Knowledge distillation is a popular framework to train models by using representations from a strong "teacher" model as learning targets for a weaker "student" model [44]. It has been used to reduce the size of a model by using a student architecture smaller than the teacher [45], to improve classification performance by softening the hard targets and extracting the "dark knowledge" from the teacher logits [46] or self-train models by aligning features from multiple views of the same observation [47], [48], [49]. In our work, we use knowledge distillation as a cross-modal alignment in the spirit of [50], *i.e.* as a way two align multiple views coming from *different sensors*, instead of different augmentations.

Using knowledge distillation to refine an existing model without supervision is akin to continual pretraining. Continual pretraining refers to the practice of adding new pretraining stages to an existing model, that aim to improve and/or specialize its representations on task-specific data. This was first introduced to fine-tune the abilities of large language models on specialized domains [51], [52], [53] and to fine-tune self-supervised vision models on medical imagery [54].

More broadly, continual pretraining has been found to improve the performance of self-supervised models on many tasks [6]. As RS data is available as large but unlabeled datasets, this has become a popular technique to specialize large-scale models for remote EO [55]. [56], [57] *e.g.* mix SSL and continual learning to integrate new observations, ending up in a continual pretraining framework to deal with non-stationary RS datasets. In this work, we show that knowledge distillation can be an effective tool for multimodal alignment as a continual pretraining objective.

III. SSL AND CROSS-SENSOR ALIGNMENT

Our goal is to design a self-supervised training scheme that can deal with multiple sensors, even with different resolutions. To do so, we divide the loss function in two:

- an off-the-shelf SSL objective, *e.g.* DINO [47],
- our novel scale-aware alignment loss (MSAD).

Given a multimodal dataset $\mathcal{D}$ and a Vision Transformer (ViT) parametrized by its weights θ, we minimize:

$$\mathcal{L}(\mathcal{D};\theta) = L_{\text{SSL}}(\mathcal{D};\theta) + \lambda \cdot L_{\text{MSAD}}(\mathcal{D};\theta) \tag{1}$$

where L_{SSL} is a contrastive self-supervised loss, L_{MSAD} is our Multi-Sensor Alignment Dense loss and $\lambda \geq 0$ is a weighting hyperparameter.

We design the $\mathcal{L}_{\text{MSAD}}$ as a contrastive loss that learns *invariant* representations across sensors, *i.e.* representations that are sensor agnostic. We collect a new multi-sensor dataset (MSC-France) of paired acquisitions, presented in Section IV. While MSAD can be used on top of any SSL objective, we use DINO in this work for its strong emerging representations [47]. To tackle the scarcity of very high-resolution RS data, we will show that our framework can be used to train large models on low-resolution data, and then adapt them later on to fewer high-resolution acquisitions.

A. Contrastive loss

Most contrastive losses pass two different augmentations of the same image to a student network S and a teacher network T that share their architecture. More precisely, multiple views x are generated, containing two global views and several local – *e.g.* cropped – views at different resolutions. This multi-scale approach is pivotal for our approach, based on cross-sensor alignment. All crops are passed through the student while only the global views are passed through the teacher, therefore encouraging "local-to-global" correspondences. The output of the teacher network is centered with a mean computed over the batch. Each network outputs a K-dimensional feature vector transformed into probability distributions P_s and P_t by a softmax with a temperature τ over the feature dimension:

$$P_{\theta_s}(x)^{(i)} = \frac{\exp\left(g_\theta(x)^{(i)}/\tau\right)}{\sum_{k=1}^{K}\exp\left(g_\theta(x)^{(k)}/\tau\right)}, \tag{2}$$

where g_θ is a network (*i.e.* S or T). With a fixed teacher, their similarity is then measured with a cross-entropy loss:

$$\min_{\theta s} -P_{\theta_t}(x)\log[P_{\theta_s}(x)] \tag{3}$$

Competitive SSL objectives, such as MIM, are ill-suited to our work as they require high-resolution images to be effectively trained on RS data [2]. Indeed, low-resolution patches are too easy to reconstruct, making the model perform poorly on downstream tasks, as we will show in Section V-B.

B. Multi-Sensor Alignment Dense loss

Consider a dataset of acquisition pairs from two different sensors A and B. We denote X_A and X_B images from these two sensors. Each image from one sensor (e.g., X_A) covers the exact same geographical area as the other sensor (e.g., X_B). We define $v^A = \psi_A(S(X_A))$ and $v^B = \psi_B(S(X_B))$ the representations – *i.e.* global token – obtained by passing the image through a Vision Transformer S and a projection ψ. In practice, ψ_A and ψ_B are two linear layers. The backbone S is trained using the contrastive InfoNCE loss so that v_A and v_B are invariant to their respective sensor:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}^A \cdot \log p_{ij}^A + y_{ij}^B \cdot \log p_{ij}^B \tag{4}$$

where y_{ij}^A and y_{ij}^B are defined as Kronecker Deltas δ_{ij}, and p_{ij}^A and p_{ij}^B are defined by:

$$p_{ij}^A = \frac{\exp\left(\text{sim}\left(v_i^A, v_j^B\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}\left(v_i^A, v_k^B\right)/\tau\right)} \tag{5}$$

and same for p_{ij}^B, with $\tau > 0$ a temperature parameter and sim the cosine similarity in feature space. However, using the global token from a ViT as the embedding, the alignment operates only on global information and does not consider local semantics. To integrate local information into the alignment, we instead use the patches from the last Transformer layer. Denoting $v[t]$ the T tokens from v^A and v^B, we sum the patchwise components of the loss on each token:

$$p_{ij}^A = \sum_{t=1}^{T}\frac{\exp\left(\text{sim}\left(v_i^A[t], v_j^B[t]\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}\left(v_i^A[t], v_k^B[t]\right)/\tau\right)} \tag{6}$$

where T is the number of tokens. By averaging over local tokens, the contrastive loss now compares not only the global semantics but also the local ones. This is especially important for EO, as similar scenes can be comprised of very different objects or patterns due to high intra-class diversity [58], [59].

A drawback of InfoNCE is that hard labeling only defines pairs as "positive" or "negative". Softening the targets has been shown to be useful in CLIP, *e.g.* when text and image are not perfectly aligned [60], [28]. This is especially important in EO, since acquisitions of the same area by different sensors are generally not simultaneous. Therefore, changes, either seasonal or structural, can happen. In addition, some unpaired acquisitions can look extremely similar, *e.g.* large forests with the same type of trees but in different areas. Therefore, we use softened targets to relax the similarity constraints and allow that some unpaired tokens of different sensors might be similar and vice-versa.

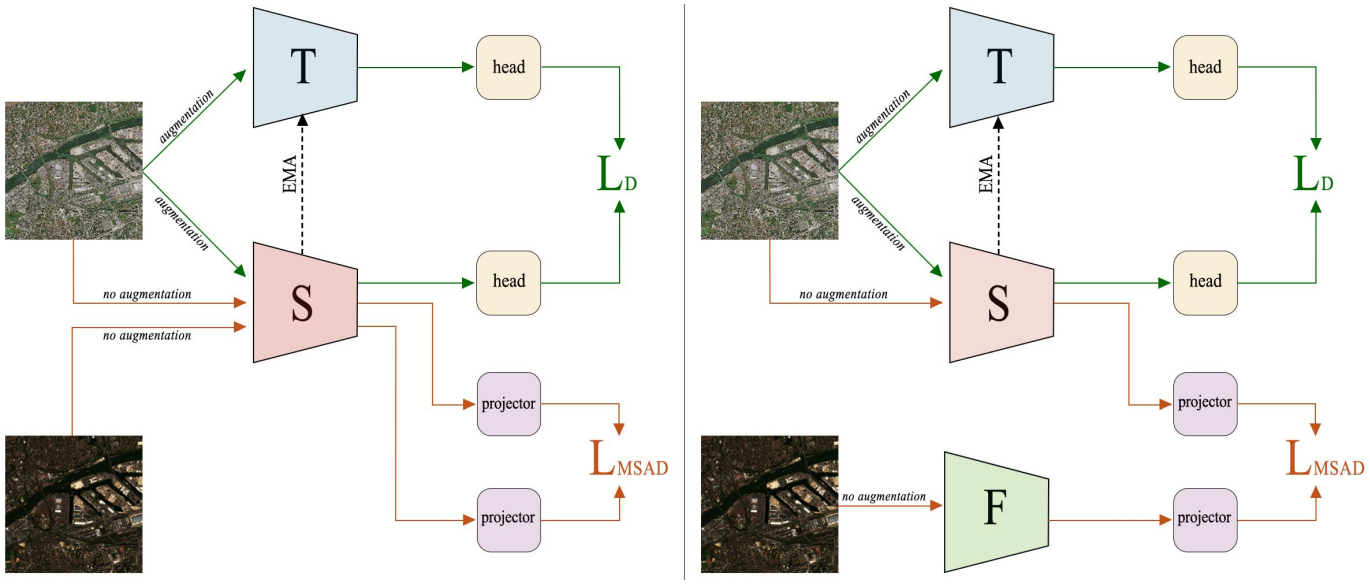

Fig. 2: Training strategies using X-STARS. On the left, pretraining from scratch adds the Multi-Sensor Alignment Dense loss to the standard DINO self-supervised training scheme. The teacher is updated by an exponential moving average (EMA) of the student. On the right, the continual pretraining approach uses a frozen backbone F to extract the features on which to apply the MSAD loss. The DINO student and teacher are initialized to the same weights as F.

In practice, the softened targets $\widetilde{\boldsymbol{y}}_i^l$ for the i-th pair can be formulated as:

$$\widetilde{\boldsymbol{y}}_i^l = (1-\alpha)\boldsymbol{y}_i^l + \alpha/(N-1), \tag{7}$$

where α is a fixed smoothing parameter.

Wrapping everything together, our final MSAD loss is:

$$\mathcal{L}_{\text{MSAD}} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\widetilde{y}_{ij}^A \cdot \log\left(p_{ij}Al\right) + \widetilde{y}_{ij}^B \cdot \log\left(p_{ij}^B\right)\right) \tag{8}$$

C. X-STARS

X-STARS can be used in two setups:

i) pretraining from scratch is the standard SSL setup. The backbone S is initialized and random and trained end-to-end using X-STARS by combining DINO and MSAD loss, as shown in Figure 2 (left) and detailed in Algorithm 1.

ii) Continual pretraining improves an existing large pretrained model on a new task through a secondary pretraining stage. In our case, we assume that a pretrained model – the domain teacher F – is available, pretrained on domain A. We initialize the student S and teacher T used by DINO to the same architecture and weights as F. However, F is frozen throughout the training process. We assume that, at least for a subset of images from domain A, there exists a paired dataset of images (X_A, X_B) of images respectively from pretraining sensor A and a new sensor B. We then train S and T as shown in Figure 2 (right). S and T receives the standard DINO loss on the images X_B of the new sensor B. In addition, F extracts the features l from the corresponding images X_A from the original sensor A. These images should have similar characteristics to those used to pretrain F. S is then also optimized through the MSAD loss computed between the representations of the student $v^A = \psi_A(S(X_A))$ and the frozen domain teacher $v^B = \psi_B(F(X_B))$.

Data: dataset $\mathcal{D} = \{(X_A, X_B)_{1\leq i\leq N}\}$
Input: $\lambda > 0$, ViT backbone S
Initialize at random S and projectors ψ_A, ψ_B;
Set T as an exponential moving average of S;
while *training* **do**
  Sample a pair $X_A, X_B \in \mathcal{D}$;
  /* DINO loss */
  for X *in* X_A, X_B **do**
    $X_1, X_2 \leftarrow \text{augment}(X)$; // aug. views
    $f_1^s, f_2^s \leftarrow S(X_1), S(X_2)$; // student feats
    $f_1^t, f_2^t \leftarrow T(X_1), T(X_2)$; // teacher feats
    $L_{\text{D}} \leftarrow \text{DINO}(f_1^s, f_2^s, f_1^t, f_2^t)$;
  end
  /* MSAD loss */
  $s, l \leftarrow \psi_S(S(X_A)), \psi_L(S(X_B))$; // projection
  $L_{\text{MSAD}} \leftarrow \text{MSAD}(s, l)$;
  /* Backpropagation */
  SGD on S for $L \leftarrow L_{\text{D}} + \lambda L_{\text{MSAD}}$;
end

Algorithm 1: Training of X-STARS from scratch.

IV. MSC-France dataset

To pretrain our model, we collected a multi-sensor dataset of aerial and satellite imagery over France. There were only three multi-optical sensor datasets in the literature: Contrastive Sensor Fusion (CSF) [41], AiRound [42] and FLAIR [43]. However, CSF contains few images, and mostly small patches from high-resolution images (SPOT-6, NAIP, Pléiades). Meanwhile, AiRound combines Sentinel-2 images with airborne data with

very different resolutions (0.3 m–4800 m GSD). Finally, FLAIR has a strong resolution gap between the two sensors (10 m GSD of Sentinel-2 vs 0.25 m GSD of BDORTHO). These reasons make these datasets impractical for our needs.

Multi-Sensor Cities France (MSC-France) aggregates RGB images from three sensors: Sentinel-2 (10 m/px), Landsat-8 (30 m/px) and SPOT-6 (1.5 m/px[1]). It contains 4496 triplets of Sentinel-2, Landsat-8 and SPOT-6 images. An example is shown in Figure 3. Specifically, Sentinel-2 images represent the level 2A reflectances values, processed by Sen2Cor. We considered only the second (blue), third (green) and fourth (red) bands, with a spatial resolution of 10 m/px. Landsat-8 images are captured by Landsat 8 OLI/TIRS sensors, represent the atmospherically corrected surface reflectance. As for Sentinel-2, we considered only the second (blue), third (green) and fourth (red) bands, with a spatial resolution of 30 m/px. SPOT-6 ©AIRBUS DS (2018) images are pansharpened from the 6 m/px multispectral to the 1.5 m/px panchromatic and orthorectified by IGN. Landsat-8 and Sentinel-2 were downloaded from Google Earth Engine. We filtered the images selecting only the on with $< 0.1\%$ cloud coverage, in the timeframe of summer of 2018. Then starting from the last available, we filled the areas covered with clouds with the precedent available info. As said, each triplet insists specifically on the same area, leading to different sizes among the different sensors:

- 2000×2000 for SPOT-6 (GSD 1.5 m/px);
- 300×300 for Sentinel-2 (GSD 10 m/px);
- 100×100 for Landsat-8 (GSD 30 m/px).

The dataset covers 12 major French cities and their suburbs (Paris, Nice, Toulouse, Bordeaux, Strasbourg, Lyon, Rennes, Lille, Marseille, Montpellier, Grenoble, and Nantes). This balances the urban-to-rural area ratio, while keeping a large diversity of backgrounds useful for downstream tasks. For example, the Toulouse region includes part of the Pyrenees mountains, while Nice also captures the seaside, and Paris imagery encompasses the city and its nearby forests. From MSC-France, we intentionally sample outside of over-represented but less distinctive areas, such as the sea and agricultural areas.

We introduce also another dataset: **Multi-View Île-de-France**. It consists of 45,200 pairs of aerial BDORTHO[2], with a GSD of 0.2 m/px, and satellite SPOT-6, with a GSD of 1.5 m/px. It covers all the Île-de-France, (*i.e.* the 75th, the 92nd, the 93rd, the 94th, and the 95th departments). Each BDORTHO image is 1250×1250 and each SPOT-6 is 167×167. In Figure 4, two examples are presented.

V. Experiments

We compare X-STARS to SOTA SSL models for EO data: Scale-MAE [5], SatMAE [9], a plain MAE [27] and SeCO [13]. We use the public weights for these models. Moreover, we also reported the results of SatMAE and Scale-MAE trained[3] on our MSC-France, to understand the impact of different data on the performance.

[1]Pansharpened from the 6 m/px multispectral image.
[2]https://geoservices.ign.fr/bdortho
[3]The hyperparameters are taken from the official GitHub repositories.

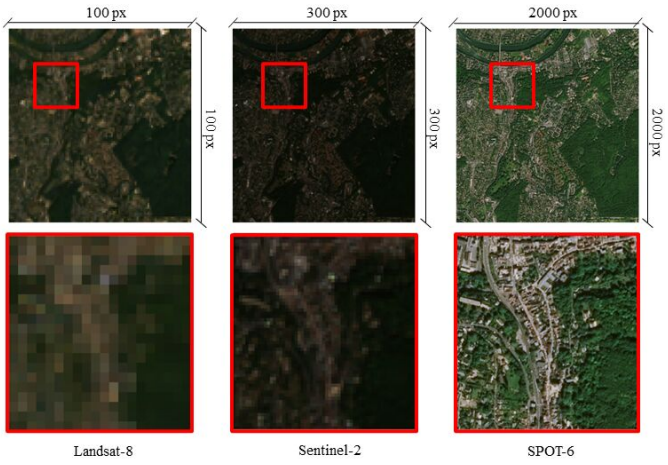

Fig. 3: A triplet from MSC-France dataset. On the first line the resized images are shown. On the second, we focus on a random area to show the different resolutions.

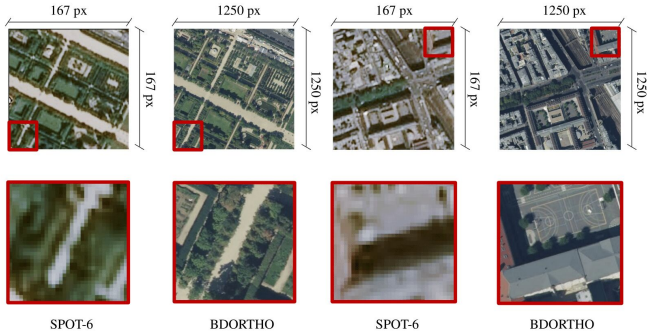

Fig. 4: Two pairs of the Multi-View Île-de-France. On the first line, the resized images are shown. On the second, we focus on a random area to show the different resolutions.

A. Experimental setup

For the "pretraining" setup of X-STARS, we train a ViT-L/16 and a ResNet50 backbone on 8 NVIDIA V100 GPUs for 800 epochs. We use only the DINO loss when one sensor is involved, or the full X-STARS loss with MSAD when two or more sensors are available. For the "continual pretraining" setup, the model is first pretrained, and then adapted using the MSAD loss for 400 additional epochs. We use a batch size of 32 images/GPU in all the experiments. All the input images are resized to 224×224. DINO hyperparameters are set as per [47]. We use a label smoothing $\alpha = 0.3$ and a weight $\lambda = 0.1$ for MSAD, obtained through cross-validation.

We evaluate the representations learned by the models for various downstream tasks for land cover scene classification and land cover semantic segmentation. We freeze the backbone and perform either a non-parametric k nearest-neighbor classification (k-NN), or a linear probing by training a linear classifier on top of the representations. This allows us to evaluate both whether semantically similar observations are grouped in the feature space, and whether the discriminative power of the representations is organized in a linear structure. For scene classification, we selected four datasets: EuroSAT [61],

TABLE I: Metrics (top-1 accuracy and mIoU) on seven downstream classification and segmentation tasks. C denotes Classification tasks, S denotes Semantic Segmentation. Results are reported for k-NN (average over multiple k)/linear probing. $A \rightarrow B$ denotes models trained using continual learning, with an alignment from sensor A to sensor B.

	Model	Encoder	Pretrain Data	Init Weights	EuroSAT (C)	Worldstrat (C)	UC-Merced (C)	CV-BrCT (C)	FLAIR (S)	SN-8 (S)	CLC (S)
From scratch	MAE [27]	V	ImageNet	random	90.7/89.9	68.9/68.1	75.1/65.7	71.8/62.0	47.1	67.9	19.4
	SeCo [13]	R	SeCo dataset	random	62.6/93.9	56.4/69.9	36.5/91.8	42.2/71.8	34.6	59.9	16.2
	SatMAE [9]	V	fMoW-Sentinel	random	91.1/94.6	68.8/69.4	82.4/80.0	69.9/64.9	46.7	67.8	20.2
	SatMAE	V	S2-LS	random	83.4/88.5	66.1/71.4	65.4/75.0	68.4/65.4	35.1	62.3	13.4
	Scale-MAE [5]	V	fMoW	random	92.9/93.9	**70.8**/70.4	82.2/84.9	75.7/71.4	51.9	68.4	20.1
	Scale-MAE	V	S2-LS	random	83.4/89.5	64.3/71.7	70.2/79.8	67.5/66.3	32.3	61.5	13.1
	Scale-MAE	V	S6	random	84.1/89.0	65.1/70.1	70.8/78.6	66.8/66.9	34.1	62.1	13.5
	X-STARS	R	S2-LS	random	77.3/94.6	57.6/69.4	72.2/89.1	54.4/70.3	36.1	62.1	18.5
	X-STARS	V	S2-LS	random	94.4/95.5	70.4/70.5	89.9/91.7	76.9/76.4	48.1	67.6	19.9
Continual	**X-STARS**	V	LS→S6	X-STARS S2-LS	95.0/**96.1**	69.8/70.7	90.3/92.5	77.4/76.8	47.5	68.1	20.5
	X-STARS	V	S2→S6	X-STARS S2-LS	**95.1**/95.3	69.8/69.5	91.9/91.9	78.2/76.6	45.3	68.1	19.7
	X-STARS	V	S6→S2	Scale-MAE	95.0/95.1	**70.8**/69.7	91.1/92.4	77.8/77.2	46.5	68.7	19.8
	X-STARS	V	S6→BDORTHO	Scale-MAE	91.9/94.3	69.4/**73.3**	**93.0/97.6**	**78.6/79.1**	**53.3**	**69.8**	**20.9**

Worldstrat [62], UC-Merced [63] and CV-BrCT [42]. We report the top-1 accuracy on the downstream datasets using k-NN (averaged over $k = 5, 10, 20, 50, 100, 200$) and linear probing. For the semantic segmentation task, for the ViT, we fine-tune a UperNet segmentation head [64] on top of our pretrained backbone. For the ResNet50, we fine-tune a decoder of a ResUNet, while freezing the pretrained encoder. We selected three datasets: SpaceNet8 [65], Chesapeake Land Cover (CLC) for which we use Landsat-8 only [66] and the subset of FLAIR [43] defined by [59]. For this task, we report mIoU. This choice of downstream tasks allows us to evaluate models on datasets with very different resolutions and sensor characteristics. For all datasets, we only consider the RGB bands. In addition to the benchmark results, we perform ablations studies for the alignment loss L_{MSAD} (Section V-C), impact of different input resolutions (Section V-D), combinations of input pairs (Section V-E), impact of the knowledge distillation continual pretraining (Section V-F), impact of model scale (Section V-G) and few-shot experiments (Section V-H). Experiments for these ablations are done with a ViT "tiny" backbone on 2 NVIDIA V100 GPUs and a batch size of 64/GPU. See the appendix for more details on the experimental setup.

B. Comparison with state-of-the-art

Table I reports results for existing self-supervised EO models and our approach.

a) Training from scratch: For classification, we observe that X-STARS outperforms existing approaches on nearly all downstream tasks when trained on Sentinel-2 and Landsat-8, both with k-NN and linear probing. On average, X-STARS trained from scratch outperforms SeCo and all MAE models (*e.g.* +1.3 % on EuroSAT and +7.7% on UC-Merced with respect to Scale-MAE). The multimodal alignment on low-resolution imagery (*i.e.* Sentinel-2 and Landsat-8) seems to learn effective representations, both for non-parametric k-NN and linear probing. In particular, X-STARS outperforms Scale-MAE on three of the four classification tasks, Worldstrat being the exception. This could be expected since Scale-MAE is trained on the fMoW [14] dataset, consisting mostly of WorldView and GeoEye imagery at <1 m/px GSD, which matches closely the 1.5 m/px GSD from the SPOT-6 sensor used for WorldStrat. In comparison, the Sentinel-2 data used to train our model has a 10 m/px GSD and X-STARS still manage a 70.4% k-NN accuracy vs. 70.8% for Scale-MAE. Note that this is despite MSC-France dataset containing only ≈5k patches, while Scale-MAE is trained on the 363k patches from fMoW [14] and SeCo on ≈200k images [13].

On segmentation tasks, X-STARS trained from scratch outperforms SeCo by a consistent $\approx +2\%$ and is competitive with SatMAE, only losing against Scale-MAE (-2.8% mIoU on FLAIR, -0.8% on SN-8, -0.2% on CLC).

Scale-MAE obtains the best accuracy on segmentation across models trained from scratch, thanks to their effective scale-invariant strategy on high-resolution images, *e.g.* mIoU of 51.9% vs 48.1% on the 0.25 m/px aerial images from FLAIR. Note however that the gap becomes smaller on lower resolution images, *i.e.* SN-8 (0.8 m/px GSD) and CLC (30 m/px GSD).

Moreover, to show that the X-STARS improvements are not due to the newly proposed dataset, we also reported some experiments with Scale-MAE and SatMAE[4] trained from scratch on MSC-France (*i.e.* Sentinel-2 and Landsat-8 pairs). For Scale-MAE we report also an experiment with only SPOT-6 images, to make it more similar to fMoW. As we can see from Table I, X-STARS outperforms both of the models. This, moreover, confirms that MIMs need either high-resolution or abundant data to work properly.

b) Continual pretraining: We also show that X-STARS can improve existing models using continual pretraining. Using the X-STARS S2-LS model, we adapt the model to SPOT-6 either using Landsat-8 or Sentinel-2 as a reference, improving the accuracy on most downstream tasks. However, we can also adapt an existing model, *e.g.* using Scale-MAE as a domain teacher. Since Scale-MAE was trained on the very high-resolution images from the fMoW dataset, we use the SPOT-6 (1.5 m/px) as the "source" domain for the adaptation.

Using Scale-MAE as initialization and adapting to Sentinel-2 images, X-STARS improves all the classification results (*e.g.* for k-NN, average top-1 accuracy 94.6% on EuroSAT and 70.8% on Worldstrat), as low-resolution features are more effective for high-level semantics. Meanwhile, adapting to the aerial images

[4]Without temporal or location metadata, *i.e.* similar to a vanilla MAE.

TABLE II: Accuracy on the validation test of downstream datasets with different configurations of the MSAD loss. PW = patchwise, LS = label smoothing.

MSAD	PW	LS	ES	WS	SN
			91.6	72.5	63.6
✓			91.9	72.0	64.1
✓	✓		92.4	71.6	64.6
✓		✓	85.6	70.2	61.9
✓	✓	✓	**92.7**	**73.1**	**64.8**

from BDORTHO, we observe a significant boost in accuracy on segmentation datasets (*e.g.* +1.4% mIoU on FLAIR), and also the classification task on the high-resolution datasets (*e.g.* +11% on UC-Merced, +3% on CV-BrCT). This demonstrates that X-STARS can be used to improve the performance of large-scale self-supervised EO models by adapting them to new sensors and finding representations better suited to downstream tasks.

C. Ablation on label smoothing and patchwise alignment

We report in Table II downstream accuracies on X-STARS trained on Sentinel-2 + Landsat-8 in various configurations of the MSAD alignment loss. All the experiments were conducted on the validation sets of three datasets (EuroSAT, Worldstrat and SpaceNet-8) so as not to bias the results on the test sets. First, we observe that using the MSAD alignment on the similarities computed on the global tokens already improves by $\approx 0.6\%$ model performance both in classification (EuroSAT and Worldstrat) and segmentation (SpaceNet-8).

Second, we observe an accuracy improvement brought by applying the alignment in a patchwise manner (PW). This shows the importance of learning to align tokens representing *local regions* and not only the global image representation. Third, we evaluate the impact of label smoothing (LS) in MSAD. Smoothing alone actually degrades performance compared to the baseline. However, when used in conjunction with patchwise alignment, smoothing improves performance. This is because image pairs have the same *global* semantics. Therefore, the smoothing adds noise to the learning process. However, semantics can be *locally different*, especially with images at different resolutions. In that case, smoothing the similarities helps model softer alignments, resulting in better downstream performance. The patchwise mechanism is useful for learning local invariants, and the smoothing for grasping high-level similarities among different objects (*e.g.* different forests or different crops).

D. Impact of different input resolutions

The results in Table III assess the impact of resizing data on model performance, considering variations in resolution between pretraining and downstream tasks. In Table I, the Sentinel-2 patches in MSC-France, at 300×300 with a GSD of 10 m/px, were resized to 224×224 for pretraining, resulting in a resolution of ≈13 m/px. In contrast, the downstream EuroSAT task utilized 64×64 patches, equivalent to a GSD of ≈3 m/px with 224×224 patches. Pretraining with a patch size the most similar to the original DINO yielded the best performance. This suggests looking for a trade-off between using the highest resolution and maintaining an appropriate patch size. Moreover, higher resolution generally leads to better performance, as shown also in [67]. The highest accuracy of 94.6% is achieved when both pretraining and linear evaluation input sizes are 224×224 (Table III). Interestingly, there appears to be no direct correlation between the GSD of pretraining and downstream tasks, despite using images from the same sensor.

TABLE III: Impact of different resolutions (i.e. patch size) for S2 images in X-STARS pretraining (i.e. MSC-France) and downstream task (i.e. EuroSAT)

PT	Lin Eval	Acc@1
224	224	**94,6**
	64	90,9
100	224	93,7
	64	87,3
300	224	92,9
	64	91,8

TABLE IV: Accuracy on downstream tasks after pretraining on different combinations of satellite images, with and without the MSAD.

Inp.	MSAD	Ep.	ES	WS
LS-S2		400	91.8	68.3
		800	93.3	68.7
	✓	400	93.4	**70.8**
S2-S6		400	92.2	68.8
		800	93.7	69.2
	✓	400	**93.8**	70.5
LS-S6		400	92.8	68.7
		800	93.6	68.7
	✓	400	93.7	69.5
S2-LS-S6		400	93.4	68.3
		800	**93.8**	68.4
	✓	400	**93.8**	68.8

E. Impact of training sensors

To investigate the impact of the alignment MSAD loss, we evaluate several models trained on different sensor combinations. Results are reported in Table IV. Without the MSAD, the model is trained alternatively with batches of one sensor or the other, with no specific alignment. We observe that, with MSAD, the model i) converges faster; ii) reaches higher downstream accuracy on average. However, the current implementation of X-STARS saturates when three or more sensors are trained together. Indeed, training on the S2-LS-S6 combination does not perform better than any other combination of the two. We hypothesize that this is due to the pairwise computation of the MSAD alignment. Indeed, in this setup, we compute all pairwise similarity pairs (*i.e.* S6-S2, LS-S2, S2-S6), which are then averaged over. This could allow for inconsistent similarities between pairs of sensors, slowing down the convergence overall. Nonetheless, multimodal alignment with MSAD consistently improves the baseline in all two sensor combinations.

F. Different continual pretraining data strategies

In this ablation, we evaluate the improvements brought by the "continual pretraining" strategy. We adapt an existing

TABLE V: Accuracy on downstream tasks after different combinations of adaptation through knowledge distillation.

PT	Adapt	ES	WS
LS	∅	86.0	64.2
LS	→ LS	91.5	68.0
LS	→ S2	92.6	69.1
LS	→ S6	92.3	69.3
S2	∅	86.2	64.1
S2	→ S2	92.2	68.2
S2	→ LS	93.2	68.4
S2	→ S6	92.7	69.4
S6	∅	82.2	63.5
S6	→ S6	91.7	68.4
S6	→ LS	92.8	**69.5**
S6	→ S2	**93.6**	69.1

pretrained model using the MSAD loss. Models are evaluated on two downstream tasks (EuroSAT and Worldstrat). Results are reported in Table V, where $A \rightarrow B$ means that the model is pretrained for 400 epochs on sensor A and then adapted for 400 epochs on sensor B. We also evaluate the $A \rightarrow A$ setting, which is standard "same sensor" knowledge distillation. Knowledge distillation in itself improves significantly the downstream model performance, even with the same sensor. However, in all cases, cross-sensor alignment improves the model even further. Generally speaking, when employing higher resolution images we can obtain slightly better results (*e.g.* 69.5% average top-1 accuracy for Worldstrat and 93.6% for EuroSAT). While adapting tends to improve performance on average, the gains are the most significant when the resolution of the downstream dataset matches the resolution of the images used for the adaptation. On average, EuroSAT results are improved by adapting on Sentinel-2, Worldstrat results are improved by adapting on SPOT-6, FLAIR results are improved by adapting on BDORTHO, etc. In conclusion, there is a trade-off between mixing features from different sensors and generality, that can be dealt with by choosing the appropriate sensor for the adaptation.

TABLE VI: Results for X-STARS, trained in a self-supervised way from scratch with different ViT backbone

Backbone	ES	WS	UC-M	CV-BrCT	FLAIR	SN8	Ches.
ViT-T/16	93.4	**70.5**	85.9	73.7	41.1	64.6	18.7
ViT-B/16	**95.2**	70.2	89.4	76.2	42.6	65.8	19.1
ViT-L/16	94.4	70.4	**89.9**	**76.9**	**48.4**	**67.6**	**19.9**

G. Model scale

To assess the effectiveness of X-STARS, we also tried to vary the size of the backbone, for the self-supervised training from scratch framework. The results are presented in Table VI. Among the results, we can see that in average, bigger transformers lead to better results. However, there are some interesting phenomena. Increasing the depth of the transformer (tiny/base vs. large) improves the performance in segmentation more clearly (ViT-L/16 gains +5.8% mIoU on FLAIR). On the other side, for classification tasks, we observe a clear correlation with dataset resolution. For datasets with lower resolution, even lighter transformers perform very robustly. In general, we can assert that our X-STARS reaches performance comparable with state-of-the-art results, also with a small model.

TABLE VII: Few-shot learning results on EuroSAT. The reported results are Top-1 accuracy (%)

Method/Data %	5%	10%	50%
Scale-MAE	75.2	79.7	86.7
X-STARS (pretraining)	88.1	91.4	**94.6**
X-STARS (continual)	**89.6**	**91.5**	94.5

H. Low-data regime experiments

To better assess the effectiveness of X-STARS, we also investigated the few-shot capabilities of our model. Specifically, we selected the best competitive model, *i.e.* Scale-MAE, and the best X-STARS under the two different approaches (*i.e.* continual and pretraining), and performed linear evaluation with an increasing number of labeled data available (*i.e.* 5%, 10%, 50%) for EuroSAT. The results, shown in Table VII, show the superiority of X-STARS.

VI. Conclusion

We introduced X-STARS, a cross-sensor alignment framework for RS. We designed MSAD, a dense contrastive loss that performs cross-sensor knowledge distillation to learn *sensor agnostic* representations of EO images, that also take into account *local semantics*. X-STARS achieves on-par results with existing self-supervised models such as ScaleMAE while using 10 to 100× less data. Moreover, X-STARS objective can be used for continual pretraining to improve existing models by adapting them to sensors that are more suited to the targeted downstream tasks. We establish new SOTA for SSL on multiple downstream classification and segmentation tasks, such as UC-Merced and FLAIR. Ablation studies demonstrate the benefits of the patchwise alignment and the multimodal loss, especially on segmentation tasks which are more challenging for self-supervised models than classification. X-STARS can be extended to integrate more challenging modalities, that so far weren't considered for their inherent greater complexity *e.g.* multispectral, SAR, thermal or even vector geodata. While the adaptation through continual pretraining is effective, efficiency could be improved by using better model adaptation such as low-rank adaptation as popularized for large language models [68]. This would allow for seamless transitions of representations from one sensor to another without the need for end-to-end fine-tuning.

Acknowledgements

The authors thank Devis Tuia for fruitful discussions on the design and scaling of self-supervised losses. This project was made possible thanks to the financial support of the ANR MAGE (ANR-22-CE23-0010) and Google for their donation under the Research Scholar program. This work was granted access to the HPC resources of IDRIS under the allocation AD011014518 made by GENCI.

REFERENCES

[1] G. Mai, C. Cundy, K. Choi, Y. Hu, N. Lao, and S. Ermon, "Towards a foundation model for geospatial artificial intelligence (vision paper)," in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '22, (New York, NY, USA), Association for Computing Machinery, 2022. 1

[2] 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 (ICCV)*, pp. 16806–16816, October 2023. 1, 2, 3

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023. 1

[4] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*, pp. 4651–4664, PMLR, 2021. 1

[5] 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 (ICCV)*, pp. 4088–4099, October 2023. 1, 2, 5, 6

[6] C. J. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. Metzger, K. Keutzer, and T. Darrell, "Self-supervised pretraining improves self-supervised pretraining," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2584–2594, January 2022. 1, 2, 3

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021. 1, 2

[8] S. Hafner, Y. Ban, and A. Nascetti, "Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data," *Remote Sensing of Environment*, vol. 280, p. 113192, 2022. 1, 2

[9] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022. 1, 2, 5, 6

[10] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, Ç. Şenaras, T. Davis, D. Cremers, *et al.*, "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21158–21167, 2022. 2

[11] 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*, pp. 5901–5904, IEEE, 2019. 2

[12] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023. 2

[13] 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*, pp. 9414–9423, 2021. 2, 5, 6

[14] 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*, pp. 6172–6180, 2018. 2, 6

[15] 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

[16] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *arXiv preprint arXiv:2211.08129*, 2022. 2

[17] V. Marsocci, S. Scardapane, and N. Komodakis, "MARE: Self-supervised multi-attention resu-net for semantic segmentation in remote sensing," *Remote Sensing*, vol. 13, no. 16, p. 3275, 2021. 2

[18] 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 *Proceedings IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6830–6840, June 2021. 2

[19] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2598–2610, 2020. 2

[20] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1182–1191, 2021. 2

[21] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. of the European Conference on Computer Vision (ECCV), Part XI 16*, pp. 776–794, Springer, 2020. 2

[22] 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 Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. 2

[23] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for earth observation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022. 2

[24] 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 (CVPR)*, pp. 5261–5270, June 2023. 2

[25] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, 2020. 2

[26] S. Vincenzi, A. Porrello, P. Buzzega, M. Cipriano, P. Fronte, R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara, "The color out of space: learning self-supervised representations for earth observation imagery," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3034–3041, IEEE, 2021. 2

[27] 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*, pp. 16000–16009, 2022. 2, 5, 6

[28] Y. Gao, J. Liu, Z. Xu, T. Wu, W. Liu, J. Yang, K. Li, and X. Sun, "Softclip: Softer cross-modal alignment makes clip stronger," 2023. 2, 3

[29] Y. Gao, J. Liu, Z. Xu, J. Zhang, K. Li, R. Ji, and C. Shen, "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining," *Advances in neural information processing systems*, vol. 35, pp. 35959–35970, 2022. 2

[30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *arXiv preprint arXiv:2303.15343*, 2023. 2

[31] Y. Yang, W. Huang, Y. Wei, H. Peng, X. Jiang, H. Jiang, F. Wei, Y. Wang, H. Hu, L. Qiu, *et al.*, "Attentive mask clip," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2771–2781, 2023. 2

[32] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, *et al.*, "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023. 2

[33] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *arXiv preprint arXiv:2306.11029*, 2023. 2

[34] C. Wen, Y. Hu, X. Li, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *arXiv preprint arXiv:2305.05726*, 2023. 2

[35] M. Singha, A. Jha, B. Solanki, S. Bose, and B. Banerjee, "Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2024–2034, June 2023. 2

[36] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," 2023. 2

[37] W. Li, Z. Niu, R. Shang, Y. Qin, L. Wang, and H. Chen, "High-resolution mapping of forest canopy height using machine learning by coupling icesat-2 lidar with sentinel-1, sentinel-2 and landsat-8 data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 92, p. 102163, 2020. 2

[38] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022. 2

[39] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1552–1560, 2017. 2

[40] F. Deuser, K. Habel, and N. Oswald, "Sample4geo: Hard negative sampling for cross-view geo-localisation," *arXiv preprint arXiv:2303.11851*, 2023. 2

[41] A. M. Swope, X. H. Rudelis, and K. T. Story, "Representation learning for remote sensing: An unsupervised sensor fusion approach," *arXiv preprint arXiv:2108.05094*, 2021. 2, 4

[42] G. Machado, E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama, and J. A. dos Santos, "Airound and cv-brct: Novel multiview datasets for scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 488–503, 2020. 2, 4, 6, 11

[43] A. Garioud, N. Gonthier, L. Landrieu, A. D. Wit, M. Valette, M. Poupée, S. Giordano, and B. Wattrelos, "FLAIR : a country-scale land cover semantic segmentation dataset from multi-source optical imagery," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 4, 6, 11

[44] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021. 2

[45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network.". 2

[46] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," in *European Conference on Computer Vision*, pp. 588–604, Springer, 2020. 2

[47] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021. 2, 3, 5

[48] 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. 2

[49] W.-C. Chen and W.-T. Chu, "Sssd: Self-supervised self distillation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2770–2777, 2023. 2

[50] F. M. Thoker and J. Gall, "Cross-Modal Knowledge Distillation for Action Recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 6–10, 2019. 2

[51] 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. 2

[52] 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. 2

[53] R. Han, X. Ren, and N. Peng, "Econet: Effective continual pretraining of language models for event temporal reasoning," *arXiv preprint arXiv:2012.15283*, 2020. 2

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

[55] M. Mendieta, B. Han, X. Shi, Y. Zhu, C. Chen, and M. Li, "Gfm: Building geospatial foundation models via continual pretraining," *arXiv preprint arXiv:2302.04476*, 2023. 3

[56] V. Marsocci and S. Scardapane, "Continual barlow twins: continual self-supervised learning for remote sensing semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023. 3

[57] H. Moieez, V. Marsocci, and S. Scardapane, "Continual self-supervised learning in earth observation with embedding regularization," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5029–5032, 2023. 3

[58] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 280–295, 2015. 3

[59] V. Marsocci, N. Gonthier, A. Garioud, S. Scardapane, and C. Mallet, "Geomultitasknet: Remote sensing unsupervised domain adaptation using geographical coordinates," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2075–2085, June 2023. 3, 6, 11

[60] Y. Gao, J. Liu, Z. Xu, J. Zhang, K. Li, R. Ji, and C. Shen, "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 35959–35970, Curran Associates, Inc., 2022. 3

[61] 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. 5, 11

[62] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25979–25991, 2022. 6, 11

[63] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010. 6, 11

[64] 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)*, pp. 418–434, 2018. 6

[65] R. Hänsch, J. Arndt, D. Lunga, M. Gibb, T. Pedelose, A. Boedihardjo, D. Petrie, and T. M. Bacastow, "Spacenet 8-the detection of flooded roads and buildings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1472–1480, 2022. 6, 11

[66] C. Robinson, L. Hou, K. Malkin, R. Soobitsky, J. Czawlytko, B. Dilkina, and N. Jojic, "Large scale high-resolution land cover mapping with multi-resolution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12726–12735, 2019. 6, 11

[67] I. Corley, C. Robinson, R. Dodhia, J. M. L. Ferres, and P. Najafirad, "Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters," 2023. 7

[68] 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. 8

APPENDIX A
DOWNSTREAM DATASET DETAILS

In Table VIII, we reported the main characteristics of the datasets used for the downstream tasks. We can note a high diversity in the number of classes, patch size, GSD and number of images. This shows the potentiality of X-STARS. For EuroSAT, that is made of Sentinel-2 MSIs, we used just the RGB images, and used the split already prepared in [61]. For Worldstrat, shaped originally for super-resolution, we selected only the Airbus SPOT6/7 images. We used the eight LCCS (land cover classes), provided in the metadata, as scene classification labels. Also, the stratified split is offered by the authors [62]. We used this dataset, because we needed a downstream task on SPOT images, to validate our ideas. For UC-Merced, we used the images as they are. We performed a custom split, given the absence of an official one and the variability of the usage. For CV-BrCT, we selected the subset with the aerial images. We discarded the ground images. We selected this dataset, due to its high variance in the resolution of the aerial images. We performed a custom split, for the same reasons as for UC-Merced. Concerning the segmentation datasets, for FLAIR, we followed the conventions utilized in [59]. We took a subset of 13 domains (10 for training and 3 for testing), and we trained the models on 256×256 patches. For SpaceNet8, we selected only the pre-flood images insisting on Louisana area. We used 256×256 patches and we adopted the split presented in [65]. For Chesapeake Land Use dataset, we needed a dataset made just of Landsat images. For this reason, among all the data available, we selected just the low-resolution Landsat data, consisting in 200×200 patches. The high-resolution were no used at all. This means that the dataset has really poor information, as shown by the average of the metrics ($\sim 20\%$ mIoU). We adopted the original split.

Dataset	Patch Size (px)	GSD (m)	N. Classes	N. Images	Task
EuroSAT [61]	64	10	10	27,000	C
Worldstrat [62]	1054	1.5	7	3,823	C
UC-Merced [63]	256	0.3	21	2,100	C
CV-BrCT [42]	500	0.3-4800	9	24,000	C
FLAIR [43]	256	0.2	14	21,396	S
SpaceNet8 [65]	256	0.5	3	14,975	S
Chesapeake [66]	200	30	15	731	S

TABLE VIII: Dowsntream datasets' info. C stands for Classification, S for Semantic Segmentation

APPENDIX B
DOWNSTREAM TRAINING PROCESS DETAILS

Both for linear probing and semantic segmentation, we used a single NVIDIA RTX 6000 GPU. For linear probing, we trained the linear layer for 100 epochs, with a learning rate of 0.01. We used a batch size of 32.

For semantic segmentation, we trained an UperNet decoder, freezing the backbone, for 100 epochs. An early stopping stops the training after 25 epochs of patience. We fixed 32 as the batch size and 0.01 as the learning rate.

Finally, for segmentation, we show some results for SpaceNet8 (Figure 5)and FLAIR (Figure 6), to show the robustness of the adapted features of X-STARS, w.r.t. Scale-MAE.

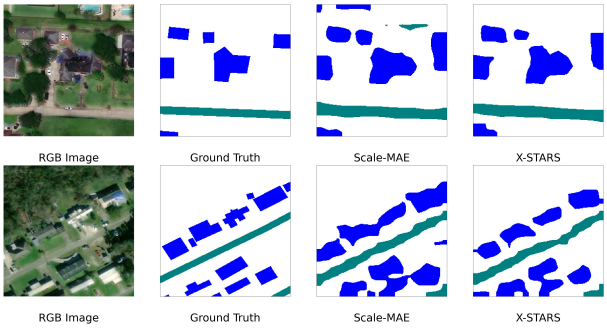

Fig. 5: Few examples from SpaceNet8 predictions.

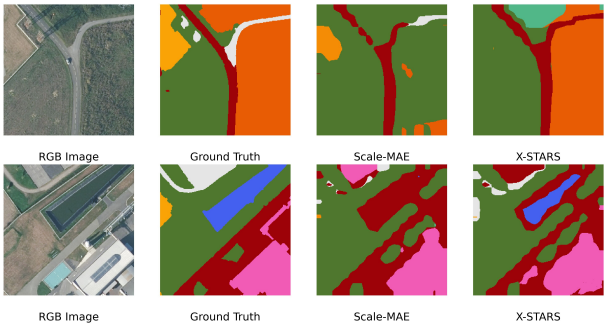

Fig. 6: Few examples from FLAIR predictions.