Paving the way toward foundation models for irregular and unaligned Satellite Image Time Series

Iris Dumeur, Silvia Valero, Jordi Inglada

Paving the way toward foundation models for irregular and unaligned Satellite Image Time Series

Iris Dumeur, Silvia Valero, Jordi Inglada

***Abstract*—Although recently several foundation models for satellite remote sensing imagery have been proposed, they fail to address major challenges of operational applications. Indeed, representations that do not take into account the spectral, spatial and temporal dimensions of the data as well as the irregular or unaligned temporal sampling are of little use for most real world applications. As a consequence, we address existing shortcomings in the design of foundation models for remote sensing data. In particular, we propose an ALIgned Sits Encoder (ALISE), a novel approach that leverages the spatial, spectral, and temporal dimensions of irregular and unaligned SITS, while producing aligned latent representations. ALISE provides easy-to-use fixed-size SITS representations which preserve the spatial resolution of the input SITS required for most mapping tasks. Moreover, to learn informative representations of SITS, we investigate the integration of instance discrimination losses within a masked auto-encoding pre-training task, utilizing a multi-view framework. The model is pre-trained on a custom-built Sentinel-2 multi-year SITS unlabeled dataset. The genericity of the provided representations is assessed on three downstream tasks: crop segmentation, land cover segmentation, and an unsupervised crop change detection task. The results suggest that the use of ALISE's aligned representations is significantly more effective than previous SSL methods for linear probing segmentation tasks. Additionally, the experiments show the interest of using ALISE representations for unsupervised change detection. Lastly, the impact of the pre-training hyper-parameters and the proposed method for aligning irregular and unaligned time series are examined in detail. The code, the pre-trained model as well as the datasets are released at https://src.koda.cnrs.fr/iris.dumeur/alise.**

***Index Terms*—Satellite Image Time series (SITS), Foundation Model, Self-Supervised Learning, Representation Learning, Multi-task Self-Supervised Learning**

I. Introduction

Over the past decade, numerous Earth observation satellite missions have been launched to monitor the effects of climate change on land surfaces. In particular, missions combining high spectral and spatial resolutions with high temporal revisit, such as Sentinel-2 (S2) [1], enable an exhaustive and systematic capture of land surfaces. The acquired data correspond to Satellite Image Time Series (SITS), which are 4D objects encompassing spatial, spectral, and temporal dimensions. By nature, SITS serve as essential inputs for generating frequently large scale land cover segmentation maps required by the climate and geoscience community. Currently, numerous Deep Learning (DL) studies have been proposed to exploit SITS for producing high-resolution maps characterizing land surfaces. For instance, several works have addressed the production of crop semantic segmentation maps [2], [3], [4]. Nevertheless, a major limitation of supervised DL technologies is their need for large amounts of labeled data. Existing supervised methods are therefore ill-suited to the production of large-scale maps, covering several time periods and/or the entire globe. In addition, DL methods applied to SITS are task-specific and thus not adapted to solving the multiple tasks required for Earth monitoring. Therefore, we propose to investigate the production of SITS representations that can be used with minimal training time and a small amount of labeled data for a wide range of Earth monitoring applications.

Building a model capable of providing multi-task representations is in line with the recent DL paradigm known as "foundation model" [5]. In remote sensing (RS), foundation models (FM) have been recently theorized as a solution to avoid the training of numerous task-specific models, exploiting the synergy of multimodal SITS, as well as mitigating the need for costly and time-consuming data annotation [6]. To do so, the large neural network architectures of these FM need to be trained on large unlabeled datasets using self-supervised learning (SSL) strategies. FM are then supposed to have learned a generic representation relevants for numerous downstream tasks. Training a RS FM on a large unlabeled dataset requires taking into account the specificities of SITS, which are among others **irregular** and **unaligned**. More precisely, irregularity refers to variable temporal gaps within the same time series, caused by missing acquisitions due, for example, to inappropriate atmospheric conditions (haze, fog or clouds) or sensor problems. Unalignment corresponds to different acquisition dates when comparing at least two time series. Even for using a single sensor, SITS acquired in different geographical areas are unaligned due to orbital phase shifts. Providing task-agnostic SITS representations, which take into account SITS specificities should also simplify the exploitation of satellite data. While recent studies have proposed SSL methodologies for SITS [7], [8], [9], [10], the existing methods do not produce: **easy-to-use (i)**, **informative (ii)** and **generic (iii)** SITS representations. These three characteristics are further detailed below.

(i) Providing easy-to-use representations of irregular and unaligned SITS. FM should provide representations which serve as basis for the geosciences and climate communities. As a consequence, we consider that easy-to-use representations: are relevant without model fine-tuning, are aligned and of fixed size, and preserve the spatial resolution of the SITS. No existing SITS representation encoder [11], [12], [7], [8]

I. Dumeur, S. Valero, J. Inglada are with Univ Toulouse 3 Paul Sabatier, Univ Toulouse, CNES/IRD/CNRS/INRAe, CESBIO, Toulouse, France (e-mail: iris.dumeur@univ-tlse3.fr,silvia.valero-valbuena@iut-tlse3.fr, jordi.inglada@cesbio.eu).

This work is supported by the DeepChange project under the grant agreement ANR-DeepChange CE23

matches those criteria. For instance, most current methods [11], [12], [7], [10] provide representations of SITS whose temporal dimensions match those of the input SITS. As a result, these methods provide unaligned representations of variable temporal size. Therefore, such representations can not be directly used in traditional machine learning methods for regression or classification tasks. Another study [9] proposes to collapse the temporal dimension of the latent representation during training. Nevertheless, the proposed latent representation has also a lower spatial resolution than the input, unsuitable for full resolution mapping.

(ii) Learning informative representations from unlabeled data. The SSL pre-training strategy used to train a FM should yield meaningful high-level SITS representations. Masked Auto-Encoders (MAE) have been frequently employed for SITS pre-training due to their ease of implementation [11], [12], [8], [7]. SITS encoders are trained to reconstruct parts of the input SITS. Nevertheless, the ability of MAE to extract high-level features has been questioned [13]. Indeed, it is hypothesized that being trained to predict in a low-level semantic space (pixel level), could reduce the ability to learn more complex and abstract SITS representations. In addition, other SSL training techniques compute a loss at the level of the latent space, which is supposed to be of a higher semantic level than the input data. For example, instance discrimination strategies are multi-view SSL techniques designed to maximize the similarity between representations of two views from the same input data. Views of the input data must be generated in such a way as to preserve the semantic information needed to solve downstream tasks. While artificial data augmentation are often employed to generate views in computer vision, such techniques are not relevant for SITS. Apart from SkySense [10], instance discrimination remains largely unexplored in SITS because it requires domain specific view generation, and often needs large batch sizes. To mitigate the disadvantages of both types of SSL strategies, recent works propose hybrid approaches [14], [15]. These methods combine various SSL strategies, such as instance discrimination with MAE, to learn more informative representations. Except for [9], these hybrid approaches are not applied to SITS representation learning.

(iii) Providing and assessing generic representations. The genericity of a SITS representation refers to the ability of a model to perform well on two important scenarios: first, in a variety of geographic and temporal configurations, and second, in a variety of downstream tasks. Under this purpose, it is crucial to pre-train FM on large-scale data-sets covering diverse geographical and temporal configurations, which avoids learning representations only relevant to specific areas and periods. Nevertheless, there is currently a significant lack of well-designed large-scale unlabeled datasets for SITS. Existing datasets [16], [17], [18] do not provide enough temporal acquisitions or correspond to restricted geographical and temporal configurations [19]. In addition, the production of RS FM is also limited by the lack of downstream reference tasks that are representative of the needs of the geosciences and climate communities. While several FM [7], [17] in RS are evaluated on scene-classification tasks, most real-world RS applications necessitate high spatial resolution semantic maps. Despite some growth, downstream labeled segmentation datasets for SITS remain scarce, limiting the evaluation of FM in the production of generic spectro-spatio-temporal representations of SITS.

In view of the above (i), (ii), (iii), we propose an aligned SITS encoder (ALISE) as a further step towards the development of a FM for SITS. As the numerous criteria mentioned above already require substantial research, ALISE was designed to process data from just one sensor (S2). Although ALISE is not a FM it addresses the production of easy-to-use, informative and generic representations. First, ALISE furnishes aligned and fixed-size representations, leveraging the spatial, spectral and temporal dimensions of multi-year SITS, which are easy-to-use (i). The resulting ALISE representations also preserve the spatial resolution of the input SITS, which is considered as crucial for downstream segmentation tasks. Secondly, hybrid SSL strategies are investigated to obtain informative SITS representations (ii). Notably, we have studied the possibility of integrating an SSL instance discrimination strategy alongside an MAE task. We propose a cross-view reconstruction task, where views are subseries of the original SITS which are each composed of different acquisitions. We also investigate whether integrating additional instance discrimination losses leads to more informative latent representations. These losses enforce invariance between the SITS views representations and decorrelate latent variables. Thirdly, generic representations are sought by using a new pre-training unlabeled open-source dataset and evaluating the model on three downstream tasks (iii). In particular, the three distinct downstream tasks are: crop segmentation (PASTIS [3]), dense land cover segmentation (MultiSenGE [20]), change detection (with the specially designed *CropRot* dataset [21]). This novel labeled dataset was built to enhance the benchmark of downstream tasks for the assessment of FM. On the two segmentation downstream tasks, we train a single linear layer to perform pixel level classification. The quality of ALISE's representations is evaluated by using them in frozen (linear-probing) and fine-tuning configurations. Finally, the change detection task is performed without any additional learning step. Change maps are generated by measuring the distance between a pair of aligned SITS representations obtained from ALISE. In addition, this paper details: ALISE's performance under a labeled data scarcity scenario, a qualitative assessment of the proposed temporal alignment method, and an extensive study of the influence of the view generation protocol and instance discrimination losses.

Our contributions can be summarized as follows:

- We introduce ALISE, a novel SITS encoder that provides aligned representations of SITS at high spatial resolution.
- We explore a new multi-view SSL task specifically designed for SITS.
- We provide two novel datasets: a large scale unlabeled pre-training multi-year European S2 dataset [22] and a labeled downstream crop change detection dataset [21].
- We achieve state-of-the-art performance on linear probing segmentation tasks [7], [8].

The code[1] and the pre-training [22] and change detection [21] datasets are available. The remainder of the paper is organized as follows. First, section II presents the current state-of-the-art in terms of SSL strategies for SITS, methods for aligning irregular and unaligned SITS and pre-training datasets. Next, section III corresponds to our proposed methodology. Moreover, the explanation of the experimental setup is detailed in section IV and the results in section V. Finally conclusions are drawn in section VI.

II. Related works

In this section we review works related to the production of informative (subsection II-A, subsection II-B), ready-to-use (subsection II-C) and generic (subsection II-D) representations. More specifically, this section presents an overview of the two main categories of SSL, which are: MAE applied in most cases with SITS (subsection II-A), and instance discrimination SSL strategies which have rarely been explored in the context of SITS (subsection II-B). Subsequently, methodologies for the generation of aligned SITS representations are presented (subsection II-C), followed by an overview of existing S2 large-scale pre-training datasets (subsection II-D).

A. Masked auto-encoders for time series

Masked auto-encoders (MAE) were popularized thanks to the great performance obtained by BERT [23]. The MAE strategy involves corrupting multiple elements (tokens) of the input sequence and training the model to reconstruct them. This SSL strategy was soon extended to other domains such as time series [24], [25], [26] or image processing [27]. In the field of RS, a variety of studies employed MAE in the analysis of mono-date satellite images [28], [29] or satellite video [30]. For SITS, the proposed MAE use either a temporal masking strategy with a temporal transformer or a spatio-temporal masking strategy with a Vision Transformer (ViT). As detailed in [8], existing spatio-temporal masking strategies, such as the SatMAE [17] and Prithvi [18], process exclusively SITS composed of three acquisitions. The temporal approaches, on the other hand, are often applied on fully-temporal architectures [11], [7] or with narrow spatial context [12]. To the best of our knowledge, the sole spatio-spectro-temporal architecture adapted to SITS and pre-trained as an MAE is U-BARN [8]. Furthermore, MAE methodologies adapted to time series (not specifically SITS) differ on two important points.

First, they differ on how they handle corrupted tokens. Inspired by the original BERT [23], methods with temporal masking inject the corrupted tokens directly into the SITS encoder. Unfortunately, this input data corruption introduces a distribution shift between pre-training and downstream tasks, as the latter have no input corruption step. Spatio-temporal methods [18], [17] address this last limitation by using an asymmetric encoder-decoder architecture. Inspired by MAE in vision [27], the corrupted tokens are not fed to the encoder. Instead, the latter are concatenated to the input representations and processed solely by the decoder using a self-attention mechanism. In addition, outside RS studies, recent MAE for time series such as [25], [26] also employ an asymmetric architecture. Instead of a self-attention mechanism, a lightweight decoder that performs cross-attention between corrupted tokens and the latent representation, is employed.

Secondly, MAE on time series differ in the employed masking pattern. While retrieving a masked word in natural language processing (NLP) requires a holistic understanding of the sentence, neighboring data points in time series are often highly correlated. Therefore, several studies [26], [25], [24] advocate splitting the time series into non-overlapping temporal sub-series before masking. Therefore, the masking strategy is applied at the sub-series level to force the model to reconstruct local variations. However, this methodology is not directly applicable to irregular SITS, where each sub-series would represent different temporal scales. Existing methods on SITS, thus often mask random time steps [8], [12], ignoring the potential redundancy of the temporal information. Sole [7] considers consecutive acquisition masking.

Consequently, unlike several previous studies on SITS [11], [12], [8], [17], we propose a temporal masking strategy, where the corrupted tokens are not processed by the SITS encoder. We also consider masking successive acquisitions, and the reconstruction task utilizes a lightweight decoder with cross-attention.

B. Self-supervised instance discrimination methods

The use of MAE in vision has nevertheless been criticized for focusing on learning local relationships within an input sample, instead of modeling the relationship between samples [31]. Besides, while MAE techniques are easy to implement, they can produce representations that are generally of a lower semantic level than instance discrimination techniques [32].

As per [33], we consider instance discrimination as a subset of SSL, where a siamese network (composed of two branches) is trained to produce similar representations of two views of the same data. The views correspond to alternative ways of observing the input data. Different views can be, for example, signals from different sensors or artificially augmented versions of the input data. In addition, views are expected to preserve the input semantic information needed for downstream tasks. While computer vision studies often employ a set of benchmark data augmentation techniques, determining relevant view generation on SITS is challenging. Nevertheless, several studies have tackled view generation methods on single RS image analysis. For instance, [34] have shown the interest of using domain-adapted augmentation on RS images. These methods propose to consider two images taken at the same location but at different times as two views of the same input data. Other works exploit the multi-modality and consider images from different sensors (optical and radar pairs) as two different views [35], [36].

Multi-view SSL techniques can be divided into four categories: contrastive [37], clustering [38], [39], distillation [40], [41], [42] and redundancy reduction [43], [44], [45]. These approaches differ in their strategies to prevent representation collapse, a scenario in which the encoder always predicts the same representation regardless of the input.

[1] https://src.koda.cnrs.fr/iris.dumeur/alise

Contrastive learning [37] and its variants for segmentation tasks [46] heavily rely on negative pair sampling (i.e. finding pairs of samples representing different semantics). Efficient negative pair sampling is challenging for SITS because pixels from different SITS may still represent the same classes. In the absence of labels, it is difficult to identify negative examples that are not trivial and are actually beneficial to learning. Consequently, for pixel-level SITS classification, a contrastive loss is often used in a semi-supervised framework where labels help generating relevant negative pairs [47]. SSL clustering methods, on the other hand, do not require sampling of negative pairs. Instead, they use a clustering algorithm dedicated to the generation of pseudo-labels (also called prototypes). To do so, strong assumptions about the batch distribution are required. For example, [39] assumes that all examples in a batch are evenly distributed among the prototypes. Distillation-based techniques [40], [41], [42], on the other hand, require complex training tricks such as momentum-encoder or stop-gradient. Lastly, compared to other instance discrimination SSL frameworks, the implementation of redundancy reduction techniques [43], [44], [45] is straightforward. These strategies prevent informational collapse by decorrelating every pair of variables of the embedded latent representation. The VicReg approach [44] proposes the use of three losses: the invariance loss, which enforces similarity between the embedded latent representations of the two views; the variance loss, which maintains the variance of the embedded variables above a threshold; and the co-variance loss, which intends to decorrelate the variables of each embedded view. Furthermore, a modified version of VicReg, named VicRegL [45], has been adjusted for downstream segmentation tasks. In this latter method the three previous losses are also calculated at the pixel level. Consequently, due to its simplicity, the combination of VicReg losses with a cross-reconstruction multi-view task is explored in this paper.

C. Processing irregular and unaligned SITS

For downstream tasks, SITS representations provided by pre-trained models must be of fixed-size, in order to be injected into classical lightweight classifiers including Random Forest [48], TempCNN [4], Recurrent Neural Networks [49], Sparse Variational Gaussian Process (SVGP) [50]. Additionally, those representations should be "aligned", enabling the direct comparison of the features of two different samples.

The generation of aligned SITS prior to their use in a machine learning (ML) model has been addressed in several works. A common practice is to perform a linear interpolation of the annual time series into a common temporal grid [48]. While this method is efficient for annual SITS classification, it might remove fine-grain information or introduce noise. Besides, cloud masks are required to perform the interpolation, and the definition of reference dates, composing the common temporal grid, is challenging. Another common pre-processing method is the generation of *composite* images, which aims to summarize valid acquisitions over a temporal extent. For example, the authors of Presto [7] suggest using time series with monthly information to fuse multi-sensor SITS. In this case, optical SITS monthly information corresponds to the least cloudy image, while a median of SAR acquisitions over a month is used. Unfortunately, this temporal pre-processing strategy to prevent irregular temporal feature representations has several shortcomings. First, the proposed monthly optical sampling protocol does not ensure that each pixel of the image has a clear acquisition. Second, the monthly sampling protocol induces an important loss of temporal information.

Furthermore, the latter two methods are not flexible and do not adjust to the target task. For this reason "data-driven alignment strategies" have been proposed. For instance, [51] proposes the learning of an attention-based interpolation (mTAN) [52], which is trained end-to-end along a Gaussian Process classifier (SVGP). In particular, the interpolation weights are computed thanks to a scaled dot product between embedded acquisition dates and embedded reference dates. The temporal embedding of the dates is performed with learnable embedding functions [53]. Besides, [51] shows that the attention-based interpolation framework outperforms SVGP fed with linearly interpolated data. Although this mechanism is more flexible and relevant for a classifier requiring aligned SITS such as SVGP, novel attention based encoder architectures can now process irregular and unaligned SITS [54]. With these latter networks, a temporal pre-processing will result in an unnecessary loss of information. As a consequence, aligning SITS after their encoding by an attention-based encoder, seems a more appropriate approach. A second limitation of mTAN is that the interpolation weights are calculated exclusively with the acquisition dates, without consideration of the content of the time series.

Recently, in computer vision, advanced DL architectures such as the Perceiver I/O [55] have emerged to map arbitrary input sequences onto a fixed-size aligned latent space using a flexible cross-attention querying mechanism. Specifically, the projection of the input sequence is determined by a scaled dot product between the input sequence and a learnable array, denoted *learnable queries*. The resulting length of the output aligned sequence is determined by the number of learned queries. This mechanism has been applied to SITS with a single learnable query in [9] to provide SITS representations with a collapsed temporal dimension. However, the use of this mechanism with a larger number of queries and its analysis has not yet been investigated.

As a result, we propose a flexible query mechanism to align irregular and unaligned SITS after their encoding by an attention based encoder [56]. We also investigate how the information is stored in the aligned latent representations.

D. Sentinel-2 pre-training data-sets

As the number of attempts to construct a RS FM increases, a multitude of pre-training data-sets have been constructed. Table I provides an overview of existing large scale pre-training datasets employed to pre-train SITS encoders. Ideally, a RS FM should be trained on multi-spectral data covering multiple geographic and temporal configurations. The pre-training data should also include SITS with numerous acquisitions to allow learning complex temporal features. Moreover, a SITS FM is

expected to learn to ignore cloudy pixels. As a consequence, invalid pixels should not be removed from the pre-training input data. Finally, to train a spatio-spectro-temporal network, it is necessary to collect SITS with a large spatial extent.

With the exception of the Prithvi data-set, all the proposed data-sets provide at least the ten S2 bands (10m-20m resolution) bands. In addition, except for FR-S2, the data-sets show remarkable geographical variability. However, none of the existing data-sets fits all the aforementioned criteria to pre-train a large-scale SITS FM. Indeed, most of these data-sets [16], [18], [7] do not provide long enough time series to capture the complex temporal dynamics of the Earth's surface. Additionally, several pre-training data-sets employ a severe cloud filtering of the information (SSL4EO-S12, Presto and Clay data-sets). As a consequence, this filtering prevents the FM from identifying cloudy pixels as irrelevant. This could decrease performance on downstream tasks where the cloud mask information may not be available. In contrast, a pre-training data-set like FR-S2 applies less strict cloud filtering, and provides validity mask. Besides, it is presumed that a RS FM pre-training dataset should be balanced [6]. Nevertheless, some existing datasets, such as SSL4EO-S12 [16], focus on urban areas. The scarcity of natural landscapes in such datasets could be an obstacle to learning complex temporal dynamics.

Consequently, we introduce a novel large-scale European S2 SITS pre-training data-set named MMDC-EU, which is detailed in subsubsection IV-A1.

III. Method

The method consists in pre-training an ALIgned SITS Encoder, ALISE, which produces aligned representations for multi-year irregular and unaligned SITS. The details of the ALISE architecture are presented in the next section, followed by the description of the multi-view self-supervised pre-training strategy. Specifically, the proposed SSL method studies the combination of two types of losses: a generative cross-reconstruction loss and instance discrimination losses computed in the latent space.

A. ALISE: ALigned SITS Encoder

ALISE harnesses the spectral, spatial, and temporal dimensions of irregular and unaligned input time series $X \in \mathbb{R}^{(b_s,t,c,h,w)}$, where b_s, t, c are respectively the batch, temporal, and spectral dimensions and h, w the spatial dimensions. Although t may vary for each SITS, ALISE generates a latent representation $Y \in \mathbb{R}^{(b_s,n_q,d_{model},h,w)}$ of fixed dimensions, where d_{model} and n_q are the channel and temporal sizes.

As illustrated in Figure 1, ALISE is composed of two main blocks. First, a Spatial, Spectral and Temporal Encoder (SSTE), noted Ψ in Equation (1), which corresponds to the U-BARN architecture detailed in [8]. As the original U-BARN was initially designed to handle annual SITS, the positional encoding (PE) in ALISE has been modified to process multi-year SITS. Specifically, the temporal information provided to

[2]Harmonized LandSat Sentinel 2 data-set

[3]https://clay-foundation.github.io/model/release-notes/data_sampling.html

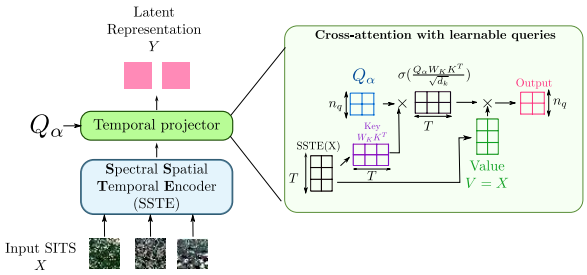

Figure 1. Overall description of ALISE architecture. The input time series X is first processed by the spectral spatial temporal encoder (SSTE) [8]. The obtained intermediate representations are then processed by a temporal projector. The temporal projector corresponds to a cross-attention mechanism with learnable queries Q_α. For visual clarity, the cross-attention is represented for one attention head.

ALISE is not the Day of Year (DoY). Instead, δ_t, which is the difference in days between the image acquisition date and a given reference date (03/03/2014), is employed.

$$O^h = \sigma\left(\frac{Q_{\alpha^h} {W_1^h}^T \Psi(X)^T}{\sqrt{d_{model/H}}}\right)\Psi(X) \tag{1}$$

Besides, a temporal projector processes the irregular and unaligned outputs of the SSTE, $\Psi(X)$, to generate aligned SITS representations. The projection is performed by a multi-head temporal cross-attention mechanism between learnable queries and $\Psi(X)$. The scaled dot product of the cross-attention mechanism on a head h is detailed in Equation (1) with Q_α^h the learnable queries, X the input time series, d_{model} the number of features in $\Psi(X)$, H the number of heads, and σ the softmax function. The attention product is fully temporal, thus $Q_\alpha^h \in \mathbb{R}^{(n_q,d_{model/H})}$, $W_1^h \in \mathbb{R}^{(d_{model},d_{model/H})}$ and $\Psi(X) \in \mathbb{R}^{(t,d_{model})}$. As detailed in Equation 2 , all the outputs of each head $O^h \in \mathbb{R}^{n_q,d_{model}/H}$ are then concatenated along the feature dimension and processed by a Multi-Layer Perceptron (MLP), which generates the latent representation Y.

$$Y = \text{MLP}(\text{concat}_h\left(O^1, \ldots, O^H\right)) \tag{2}$$

The temporal dimension of the latent representation Y is determined by the number of learnable queries (n_q). It must be noted that the temporal projector does not shrink the spatial dimension of $\Psi(X)$, meaning that each pixel of the SITS is represented by d_{model} features along n_q temporal features. It is crucial to understand that in the resulting latent representations the notion of "time" is not preserved. In other words, the time series is folded in a way that does not preserve the notion of order or distance in the time axis. As a result, the latent representation is not a time series; rather, it is a stack of temporal features. Therefore, we refer to the n_q vectors in the aligned latent representation as the "latent temporal features".

B. Multi-view pre-training task

As detailed in Figure 2, the multi-view SSL pre-training task, combines a cross-reconstruction loss with two additional losses computed on the embedded latent representations. As

Table I
DESCRIPTION OF S2 SITS UNLABELED TRAINING DATA-SETS USED TO PRE-TRAIN LARGE SCALE SITS MODELS. TEMPORAL EXTENT REFERS TO THE LONGEST TIME INTERVAL EXISTING IN THE DATA-SET. THE NUMBER OF DATES CORRESPONDS TO THE EXACT OR AVERAGE (∼) NUMBER OF DATES IN THE SITS. THE ROI SIZE CORRESPONDS THE SPATIAL DIMENSIONS OF THE IMAGES WITHIN THE SITS. "AVAILABLE" INDICATES WHETHER THE DATA-SET IS AVAILABLE FOR DOWNLOAD. "CLOUD FILTER" INDICATES WHETHER CLOUDY IMAGES HAVE BEEN REMOVED. WHEN THE PERCENTAGE IS SHOWN, IT SPECIFIES THE MAXIMUM PERCENTAGE OF CLOUDS ACCEPTED IN AN IMAGE. "?" IS EMPLOYED WHEN THE INFORMATION IS NOT GIVEN IN THE CORRESPONDING ARTICLE.

Data-Set Name	Data	Temporal Extent	Number of dates	Geographical extent	Roi size	Available	Cloud filter
SSL4EO-S12 [16]	S2-L2A 13 bands	2020	4, (1/season)	Worldwide	264 × 264	✓	yes, ≤ 10 %
FR-S2 (U-BARN) [8]	S2-L2A 10 bands + valid mask	2018-2020	∼ 179	France	1024 × 1024	✓	yes, ≤ 30 %
Presto [7]	S2-L2A 10 bands	2020-2021	24 (1/month)	Worldwide	1 × 1	✗	yes
Prithvi [18]	NASA HLS^{2}V2 L30	?	?	USA	64 × 64	✗	yes
SkySense [10]	S2 L2A 10 bands	?	∼ 65	Wordlwide	64 × ×64	✗	yes, ≤ 1 %
Clay[3]	S2 L2A 10 bands	2018-2023	8, (1/quarter)	Worldwide	224 × 224	✓	yes
MMDC EU (ours)	S2 L2A 10 bands + valid mask	2017-2020	∼ 354	Europe	512 × 512	✓	no

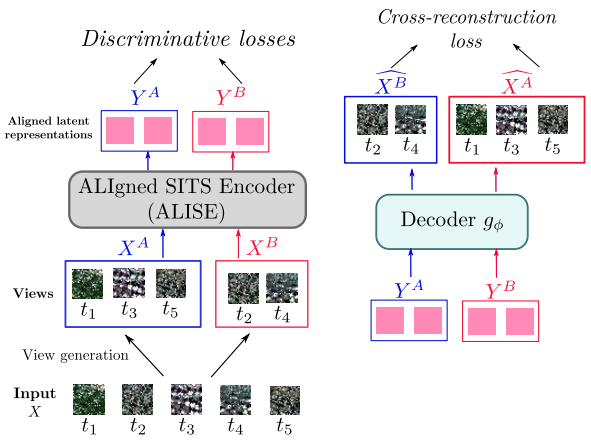

Figure 2. Description of the proposed multi-view SSL strategy. Given an input time series X two views are generated: X^A and X^B. Each view is processed independently by ALISE which generates their respective aligned latent representations Y^A and Y^B. A decoder g_ϕ is trained to reconstruct one view using the latent representation of the other. Additional discriminative losses are computed on the latent representation.

detailed in Equation (3), the total SSL loss, corresponds to the weighted sum of three terms L_{inv}, L_{cov} and L_{rec} respectively the invariance, covariance and reconstruction losses.

$$L = w_{inv}L_{inv} + w_{cov}L_{cov} + w_{rec}L_{rec} \tag{3}$$

1) View generation: After the view generation phase, ALISE encodes each of the two views, resulting in two aligned representations. The latter are used to compute the invariance and covariance losses. In the cross-reconstruction loss, the representation of each view is used to reconstruct the other view. The objective of the view generation protocol is to provide views that preserve semantic content. For SITS, the generation process aims to create views that maintain the pixel information of the observed landscape. Consequently, the two views, X^A and X^B, represent a time series at the same location but with different acquisition times. The view generation process starts by selecting N adjacent acquisitions among an irregular and multi-year SITS. As detailed in Equation (4), this latter time series is divided into n_w non-overlapping temporal windows, each composed of t_w dates. Given that SITS are irregular, each temporal window may represent a different temporal span.

$$X = \bigcup_{i=0}^{n_w-1} \{X_j \mid i \times t_w \leq j < (i+1) \times t_w\} \tag{4}$$

Finally, to ensure that the two views cover nearly identical periods, every other temporal window is used to construct respectively X^A (Equation (5a)) and X^B (Equation (5b)). Therefore, t_w corresponds to the number of consecutive dates that the model has to predict during the training process. We posit that increasing t_w complexifies the cross-reconstruction task as more temporal variations have to be retrieved by the model. This generation approach ensures that the views are temporally intertwined: $X^A \cup X^B = X$ and $X^A \cap X^B = \emptyset$ and provides a parameter t_w which controls the difficulty of the pre-training task.

$$X^A = \bigcup_{i=0}^{\frac{n_w}{2}-1} \{X_j \mid 2 \times i \times t_w \leq j < (2 \times i + 1) \times t_w\} \tag{5a}$$

$$X^B = \bigcup_{i=0}^{\frac{n_w}{2}-1} \{X_j \mid (2 \times i + 1) \times t_w \leq j < (2 \times i + 2) \times t_w\} \tag{5b}$$

2) Discriminative losses: As illustrated in Figure 2, the augmented views X^A, X^B are independently encoded by ALISE. The aligned latent representations Y^A and Y^B are

then processed into embeddings Z^A, Z^B by a projector. The projector aims to eliminate the information by which the two representations differ. Specifically, the projector operates exclusively on the channel dimensions: $\pi_\omega : \mathbb{R}^{(d_{model})} \rightarrow \mathbb{R}^{(d_{emb})}$. In other words, pixel-level latent vectors of each n_q query are independently processed by the projector. As shown in Figure 3, the proposed projector consists of one fully connected layer followed by batch normalization and ReLU, and a second linear layer. It is assumed that the choice of the projector's architecture affects the computation of the covariance loss. However, no empirical benefits were found from using a deeper or wider projector architecture for our considered downstream tasks.

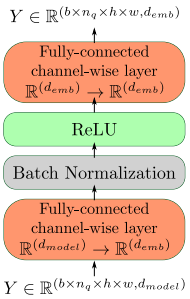

Figure 3. Description of the projector architecture.

We denote $\mathbf{z}^k_{(b,n,i,j)} \in \mathbb{R}^{d_{emb}}$ the pixel-level embedded vector of Z^k located at the spatial position (i,j) for the n^{th} query and b^{th} batch position. We propose to compute the invariance and covariance losses on the embeddings Z^A and Z^B.

First, the invariance loss maximizes the similarity between the embedded vectors $\mathbf{z}^A$ and $\mathbf{z}^B$ (see Equation (6)). As X^A and X^B have distinct acquisition dates but cover the same time-period, L_{inv} aims at learning representations which are invariant to the acquisition dates.

$$L_{inv}(Z^A, Z^B) = \frac{1}{b_s \times n_q \times h \times w} \sum_{(b,n,i,j)} \|\mathbf{z}^A_{b,n,i,j} - \mathbf{z}^B_{b,n,i,j}\|_2^2 \tag{6}$$

Second, we also investigate whether the covariance loss allows learning better representations. The covariance loss decorrelates the different d_{emb} features. The total covariance loss, Equation (8), corresponds to the sum of the covariance losses computed for each embedding Z^k. For centered embeddings $Z \in \mathbb{R}^{(b_s\times n_q\times h\times w, d_{emb})}$ the covariance loss aims to minimize the off-diagonal values of the co-variance matrix $C(Z)$ in Equation (7). In other words, the covariance matrix of the d_{emb} variables, is estimated on a batch composed of $b_s \times n_q \times h \times w$ samples. In Appendix B, we discuss how these discriminative losses are related to the VicRegL [45] losses.

$$l_{cov}(Z) = \frac{1}{d_{emb}} \sum_{i\neq j} [C(Z)]_{i,j}^2 \tag{7}$$

$$L_{cov} = l_{cov}(Z^A) + l_{cov}(Z^B) \tag{8}$$

3) Cross reconstruction task: As depicted in Figure 4, the latent representations Y^A, Y^B are employed in a cross-reconstruction task. A specific fully-temporal decoder using a cross-attention mechanism followed by a fully-connected layer is trained to recover one view X^B (resp. X^A) from the latent representation Y^A (resp. Y^B) of the other view X^A (resp. X^B). The fully-connected layer operates exclusively on the channel dimension of each pixel of the images, to recover the S2 bands from the d_{model} features.

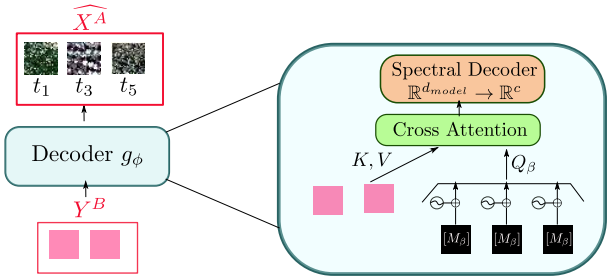

Figure 4. Description of the lightweight decoder employed for the cross-reconstruction task.

As proposed in [25] the cross-attention mechanism exploits $Q_\beta \in \mathbb{R}^{(t_w\times n_w, d_{model})}$ which specifies the dates to be reconstructed. As detailed in Equation (9), Q_β corresponds to the sum of a shared learnable masked token $M_\beta \in \mathbb{R}^{(d_{model})}$ with the temporal positional encoding[4] of the date to reconstruct. Additionally, as described in Equation (10), the latent representation Y^k with $k \in \{A, B\}$, is used to construct the keys $Y^k W_2$ and the values Y^k.

$$Q_\beta = [M_\beta + PE(\delta_{t_i})]_{1\leq i\leq t_w\times n_w} \tag{9}$$

$$\text{Cross Attention}(Q_\beta, Y^k) = \sigma\left(\frac{Q_\beta W_1 W_2^T Y^{k^T}}{\sqrt{d_{model}}}\right) Y^k \tag{10}$$

Finally, the quality of the reconstruction is assessed by using the classical Mean Square Error. As described in Equation (11), the reconstruction loss is the average of the reconstruction losses of each view.

$$L_{rec} = \frac{1}{2}[l_{rec}(X^A, Y^B) + l_{rec}(X^B, Y^A)] \tag{11}$$

Following the approach of [8], pixels with invalid measurements due to the acquisition conditions (e.g. cloudy and out of swath pixels) are ignored in the reconstruction loss. As detailed in Equation (12) M_t^{valid} represents the boolean validity mask, n_t^{valid} represents the number of clear pixels, $T = \frac{n_w \times t_w}{2}$ the number of acquisitions in a view, and $\odot$ is the Hadamard product.

$$l_{rec}(X^k, Y^l) = \frac{1}{T} \sum_{t=1}^{T} \frac{M_t^{valid}}{n_t^{valid}} \odot ||X_t^k - g_\phi(Y^l)_t||_2^2 \tag{12}$$

The validity mask is only used in the cross-reconstruction loss and is not included in the input data injected to ALISE. Therefore, no validity masks are required for downstream tasks.

[4]The temporal positional encoding used is the same as the one employed in ALISE.

C. Implementation details

To pre-train ALISE, the cosine annealing scheduler with warm restarts [57] was employed with T_0=2, and maximum learning rate of 1e-3. To generate the different views from a multiyear SITS, 60 consecutive dates were randomly selected among the 4 years of data. Within our unlabeled data-set, 60 consecutive acquisitions can extend over a maximum of four years of data and a minimum of four months. To increase the diversity of the training data, the selection of the t consecutive dates used in the view generation is random for each SITS and changes at each epoch. Besides, ALISE architectural hyper-parameters are also detailed in Appendix C. The pre-trainings tasks were conducted on a single Tesla V100 GPU for 260 epochs. The values of the pre-training hyper-parameters are shown in Table II and their choices are explained in subsection V-D. The pre-trained model with the lowest loss on the pre-training validation set is selected for downstream task assessment.

Table II
DEFAULT HYPER-PARAMETERS FOR PRE-TRAINING ALISE.

t_w	n_q	batch size	d_{model}	d_{emb}	w_{rec}	w_{inv}	w_{cov}	H
2	10	2	64	128	1	1	0	2

IV. EXPERIMENTAL SETUP

First, the four S2 L2A data-sets used in our different experiments are presented: the novel unlabeled large scale data-set (MMDC-EU) used for pre-training ALISE and the three downstream labeled data-sets (PASTIS, MultiSenGE and the novel *CropRot*). Secondly, the implementation details of our two types of downstream tasks setup (semantic segmentation and change detection) as well as the corresponding competitive works are described.

A. Data-Sets

ALISE is pre-trained on a large scale multi-year European data-set. Besides, three labeled data-sets are used to assess the quality of the pre-trained SITS encoders. The geographical distribution of the different used data-sets is presented in Figure 5. For these data-sets, only the four 10 m and the six 20 m resolution bands of S2 are used. The 20 m resolution bands are re-sampled onto the 10 m resolution grid by bi-cubic interpolation. Similarly to [8], a robust data normalization is applied on S2 L2A reflectances. Due to GPU memory limitations, ALISE is trained to process SITS with a spatial dimension of 64×64 pixels (at 10 m resolution). If the used data-set provides larger images, a random crop [5] (resp. center crop) is operated during training (resp. validation/testing) steps.

[5]https://pytorch.org/vision/main/generated/torchvision.transforms.RandomCrop.html

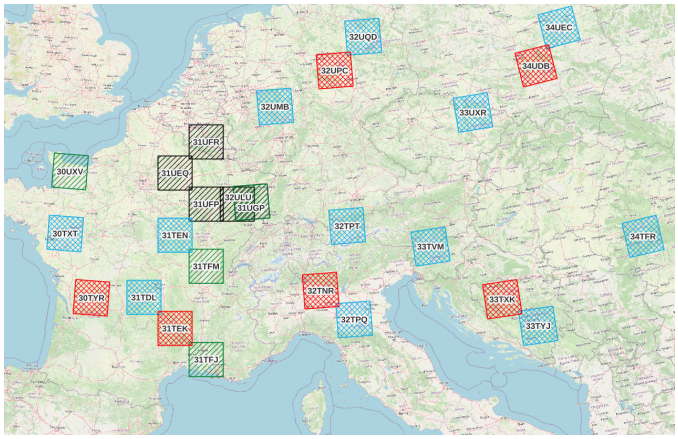

Figure 5. Geographical distributions of the different tiles composing the data-sets. The unlabeled pre-training data-set is composed of multi-year SITS selected within the blue and red boxes for the training and validation data-set respectively. MultiSenGE labeled data are selected in the area delineated by the black boxes. The PASTIS as well as CropRot data-sets are within the green boxes.

1) MMDC-EU: We have constructed an unlabeled, multi-year, multimodal SITS data-set spanning Europe. This multimodal datacube is designated as MMDC-EU. In practice, this data-set is composed of the following data: the S2 L2A product, the Sentinel-1 (S1) ascending and descending acquisitions, ECMWF AGERA5[6] weather variables, and the Copernicus 30 digital elevation model (DEM). Each SITS of each modality is spatially re-sampled onto the S2 grid. The data cube was downloaded with the openEO platform[7]. The code[8] used to create the multimodal data-set is provided for reference, allowing for potential future expansion. As this paper proposes a model that processes exclusively S2 SITS, we focus on the description of the pre-training data for this sole modality. Multi-year S2 SITS from January 2017 to December 2020 were built using all the available acquisitions. The data is split into training and validation sets with respectively 1920 and 180 SITS of spatial dimension 64×64 pixels. The downloaded S2 SITS correspond to data processed by Sen2Cor [58]. The validity mask employed in the cross-reconstruction task is built thanks to the information provided by SLC and CLM layers[9]. Specifically, as shown in Figure 5, the pre-training data-set gathers data from 18 S2 tiles. To build the training data-set, 10 smaller regions of interest (ROIs) of size 512×512 pixels are randomly selected from each of the 12 training tiles. The disjoint validation data-set is composed of the remaining 6 S2 tiles, from which 30 ROIs of size 128×128 pixels are randomly drawn.

2) PASTIS crop segmentation: The PASTIS data-set [3] provides labels for 18 crop classes from the French Land Parcel information System. The SITS considered in our experiments are collected from January to December 2019. The complete data-set contains 2433 SITS and it is divided into

[6]https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-agrometeorological-indicators?tab=overview
[7]https://openeo.cloud/
[8]https://gitlab.cesbio.omp.eu/dumeuri/openeo_datasets.git
[9]https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/

5 stratified folds. In line with [8], the segmentation task is performed exclusively on known crop classes. Background and void classes are ignored. The competitive method Presto requires cloud masks. As these data are not available in the original PASTIS data-set, the raw S2 L2A and their cloud masks were downloaded from the Sentinel hub collection [10]. These S2 data also pre-processed by Sen2Cor are used to assess not exclusively Presto but all models.

3) MultiSenGE land cover segmentation: MultiSenGE [20] is a dense land cover labeled data-set for eastern France in 2020. It is composed of 5 urban classes and 9 natural classes. This data-set is composed solely of images with less than 10% cloud cover and no cloud mask is provided. The resulting SITS are composed of 3 to 14 acquisitions. In contrast to PASTIS, MultiSenGE provides dense labels. In this data-set, we selected 4145 SITS with a spatial dimension of 256×256 pixels. A random split is performed to divide the data-set into training (60%), validation (16%) and test (24%). The class distribution is detailed in Appendix D. Lastly, in opposition to the two previous data-sets MultiSenGE data are pre-processed with MAJA [59] instead of Sen2Cor.

4) CropRot Crop change detection: This paper presents a new data-set for detecting abrupt changes between two SITS. Unlike the identification of changes in a time series (break detection), which may be solved by signal-based methods, the proposed task requires a more advanced semantic understanding. More specifically, thanks to the labels provided by *RPGExplorer Crop successions* [60], *CropRot* identifies crop rotations between two consecutive years in France. In particular, the RPGExplorer database provides crop sequence labels based on the RPG (*Registre Parcellaire Graphique*)[11]. Within a sequence (e.g. 2015-2020), parcels are unified (each parcel has a unique identifier).

For this data-set, the following classes were selected based on the RPG labels: rapeseed, cereals, proteaginous, soybean, sunflower, maize, rice, tubers and grassland. These classes categorize vegetation based on its physiological characteristics and can be identified using RS data. Pixels that are not part of these crops for the two years 2019 and 2020 are considered as background. Then, the label *change* is assigned to pixels that have a different label between 2019 and 2020. Each data-set sample includes S2 L2A SITS for 2019 and 2020, with their corresponding labels. The label tensor has three channels containing crop labels for 2019, 2020, and change label. In our proposed downstream task, change detection is performed while ignoring background pixels. The SITS were built using the SITS spatial extent from PASTIS where sufficient labels from the RPGExplorer were available. Due to this specific selection, the crop classes proteaginous, soybean and tuber do not appear in our data-set. Nevertheless, the code used to build this labeled data-set is published[12], enabling it to be extended to other regions of France and to other years. These missing classes might be integrated in an augmented version of the data-set. The change matrix between 2019 and 2020 is presented in Appendix A.

B. Evaluation Protocol

We propose two ways of exploiting ALISE representations, detailed in Figure 6, corresponding to the two types of downstream tasks: semantic segmentation and change detection.

1) Semantic segmentation tasks: As detailed in Figure 6, we classify each pixel-level latent vector by using a single linear layer in both segmentation tasks. Noting the pixel-level latent vector as $\mathbf{y}_{(b,h,w)} \in \mathbb{R}^{(d_{model} \times n_q)}$, the unnormalized logits for each class k at the pixel level can be written as: $\mathbf{c}_{(b,h,w)} = \mathbf{y}_{(b,h,w)}A + \mathbf{b}$ where $A \in \mathbb{R}^{(d_{model} \times n_q, k)}$ and $\mathbf{b} \in \mathbb{R}^k$. The classical cross-entropy loss function is used for training[13]. The latent representations are generated by a pre-trained ALISE whose weights are frozen in linear probing or updated during fine-tuning. We denote the fine-tuning and linear probing configurations as ALISE$^{\text{FT}}$ and ALISE$^{\text{FR}}$ respectively, while the fully supervised model is denoted as ALISE$^{\text{FS}}$. During the downstream tasks, ALISE as well as competitive models are trained with ADAM optimizer, a learning rate of 1e-4 and ReduceLROnPlateau scheduler with a patience of 10 epochs and a decay rate of 0.05.

2) Change detection: As detailed in Equation (13), and illustrated in Figure 6, change detection between two SITS X^1, X^2 is performed at a pixel level. For each pixel located at location (h, w), the mean square error between the corresponding latent vectors $Y^1_{(\cdot,\cdot,h,w)}$ and $Y^2_{(\cdot,\cdot,h,w)}$ is calculated. The resulting distance value serves as change detection criterion.

$$d(Y^1_{(\cdot,\cdot,h,w)}, Y^2_{(\cdot,\cdot,h,w)}) = \frac{1}{n_q \times d_{model}} \sum_{n,d} (y^1_{n,d,h,w} - y^2_{n,d,h,w})^2 \tag{13}$$

C. Competitive methods

As mentioned above, ALISE is evaluated on two different types of downstream tasks: semantic segmentation and change detection. Consequently, in this section we detail the competing works associated with these two types of tasks.

1) SITS segmentation concurrent works: ALISE can be exploited in the downstream task in three ways : fully-supervised (FS), fine-tuned (FT) and frozen (FR). Therefore, ALISE is compared with the following concurrent works: U-TAE [14] [3] (FS), Presto [15] [7] (FT, FR), U-BARN [8](FS, FT, FR) and U-BARN-GF (FR). This last approach is a variant of U-BARN providing, as ALISE, fixed dimensional SITS representations.

(a) U-TAE is a fully supervised architecture composed of a Unet network with a lightweight temporal attention mechanism located at the bottleneck.
(b) Presto is a lightweight temporal SITS encoder pre-trained as an MAE. It takes as input a monthly synthesis. The authors suggest selecting the least cloudy scene of each month. In contrast, as usually operated in RS, we train

[10] https://www.sentinel-hub.com/

[11] https://artificialisation.developpement-durable.gouv.fr/bases-donnees/registre-parcellaire-graphique

[12] https://src.koda.cnrs.fr/iris.dumeur/modcix

[13] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

[14] https://github.com/VSainteuf/utae-paps

[15] https://github.com/nasaharvest/presto (commit 5486fd5)

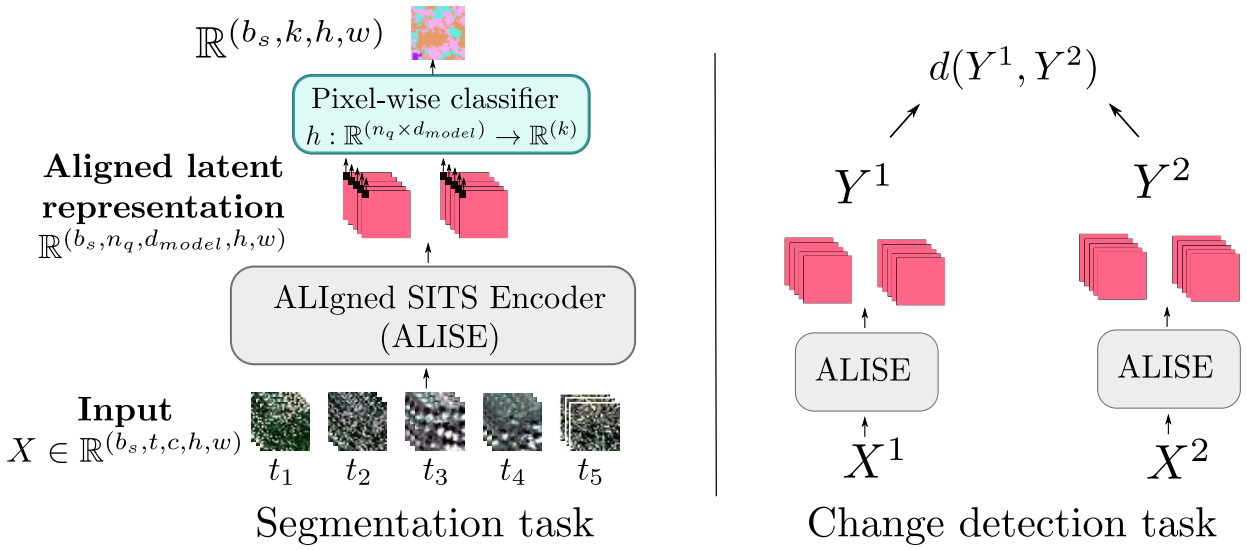

Figure 6. The two types of downstream tasks considered. Left: segmentation task framework. A single fully-connected layer projects, for each pixel of the latent representation Y, the $n_q \times d_{model}$ features into a vector of size $\mathbb{R}^k$ with k the number of classes. Right: change detection task between two SITS X^1 and X^2. The mean square error is computed between the two aligned latent representations Y^1 and Y^2.

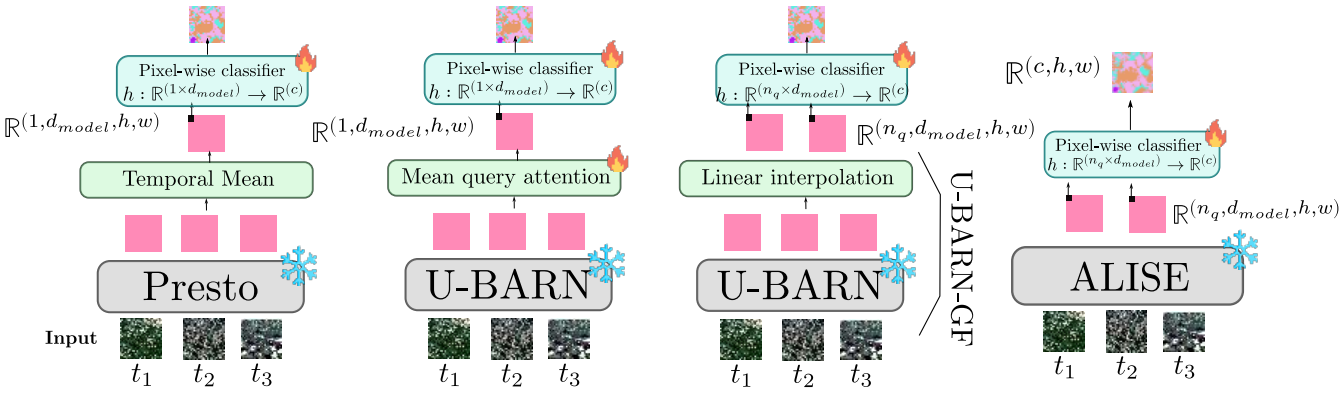

Figure 7. Comparison of Presto, ALISE,U-BARN and U-BARN-GF when frozen for the semantic segmentation task.

Presto with SITS composed of the median value of each band among the cloud-free acquisitions of each month. As suggested in [7], to exploit the latent representations provided by Presto a temporal mean is performed.

(c) U-BARN [8] is a spatio-spectro-temporal SITS encoder pre-trained as an MAE. As U-BARN does not encode SITS into a fixed size latent representation, the shallow classifier with a mean query attention mechanism proposed in [8] is considered here. Compared to the original implementation, we have modified the positional encoding so that U-BARN can process multi-year SITS. Besides, we have pre-trained U-BARN on MMDC-EU with the same pre-training configuration as ALISE. We call these SSL models U-BARN$^{\text{FT}}$ and U-BARN$^{\text{FR}}$ to denote the fine-tuning and frozen configurations.

(d) U-BARN-GF. To assess the effectiveness of the proposed learnable temporal projector, we compare ALISE with an encoder composed of U-BARN followed by a linear interpolation layer, denoted U-BARN-GF. The irregular and unaligned representations from U-BARN are projected into $n_q = 10$ regularly spaced reference dates in the temporal extent of the downstream task. The resulting aligned representations are then processed through a single fully connected layer as done with ALISE. We denote U-BARN-GF$^{\text{FR}}$ the configuration where U-BARN is pre-trained and frozen during the downstream task. If the learnable projector is effective, we expect ALISE$^{\text{FR}}$ to outperform U-BARN-GF$^{\text{FR}}$.

2) Change detection baseline: As indicated in Figure 6, we propose to exploit ALISE representations without additional learning steps to perform change detection. To establish a fair comparison, ALISE is also compared to an unsupervised change detection strategy. In the proposed competitive work, the input SITS are interpolated onto a fixed annual common time grid using a linear interpolation (gap-filling) method. Specifically we interpolate the SITS valid acquisitions onto a regular temporal grid with a period of 5 days. The distance map is computed between the interpolated raw SITS.

V. Experiments

This section evaluates the representations provided by the pre-trained ALISE on three downstream tasks and compares them to competitive methods. First, we present a detailed analysis of ALISE's performance in both fine-tuned and frozen configurations for the two segmentation data-sets (PASTIS and MultiSenGE). We also examine the effectiveness of the pre-training under a severe label scarcity scenario. Additionally, ALISE representations are assessed on an unsupervised change detection task with the CropRot data-set. Next, we provide an extensive discussion on the influence of several pre-training parameters (t_w, n_q, w_{rec}, w_{inv}, w_{cov}). Lastly, we propose a qualitative assessment of the role of the learnable queries in the temporal projector.

A. Segmentation tasks results

The segmentation performances of ALISE either frozen, fine-tuned or fully-supervised on both labeled data-sets are compared here to competitive works. The two downstream data-sets differ on two main points. Firstly, in the MultiSenGE data-set, semantic labeling is dense, whereas in PASTIS, all pixels not belonging to a known crop are not classified. Consequently, we assume that spatial context must be better taken into account to succeed in the MultiSenGE task than in PASTIS. However, we assume that to distinguish between the 18 PASTIS crop classes compared to the 14 land cover classes of MultiSenGE, more complex temporal features are required. Table III presents the averaged F1 score, the overall accuracy (OA) and the mean intersection over Union (MIoU) on the PASTIS and MultiSenGE segmentation data-sets. For each segmentation task, detailed F1 scores per class are displayed on Table IV and Table V, respectively. To better understand the performance differences between ALISE and Presto the confusion matrix is shown in Figure 8.

First, although this paper does not focus on the construction of a novel fully supervised framework for SITS, ALISE^{FS} architecture achieves performances consistent with current SOTA (U-TAE, U-BARN^{FS}). Next, Table III demonstrates that ALISE^{FR} outperforms the existing models such as Presto^{FR} and U-BARN^{FR} on the PASTIS data-set. Besides, compared to the PASTIS segmentation task, performances are lower on MultiSenGE. This may be explained by the fact that this last data-set is highly imbalanced, and minority classes decrease the macro-averaged mIoU and F1 scores as shown in Table V. Furthermore, the fine-tuned configuration (ALISE^{FT}) does not significantly outperform the fully-supervised approach (ALISE^{FS}). This finding is nonetheless consistent with the previous study [8] conducted on U-BARN. Differences between ALISE and the other two competitive works are illustrated in Figure 7 and further detailed below.

1) ALISE vs U-BARN: Segmentation metrics detailed in Table III, show that ALISE^{FR} outperforms U-BARN^{FR}: (+9% F1), (+3% 0A) and (+9% mIoU) on the PASTIS data-set and (+3% F1), (+1%OA) and (+1% on mIoU) on MultiSenGE. Remarkably, ALISE significantly outperforms U-BARN in linear probing, while having a shallower classifier (no learnable mean query) and smaller latent representations (10 latent temporal features). We also observe on Table IV, that the boost of performance may vary depending on the PASTIS crop classes. Several classes such as winter tricitale (+25%), sunflower (+14%), sorghum (+19%) and mixed cereal (+15%) exhibit stronger boost of performances compared to the other classes. The overall gain of performance can be explained by the differences between ALISE and U-BARN. ALISE differs from U-BARN in two main aspects: (i) its encoder provides fixed-size, aligned representations, and (ii) the pre-training strategy is different. As detailed in subsection III-A, ALISE corresponds to the U-BARN architecture on top of which is placed a temporal projector. Experiments detailed in subsection V-D show that ALISE's pre-training is primarily driven by its cross-reconstruction task, which is close to U-BARN's MAE pre-training. Therefore, we believe that the improvement in performance when freezing the pre-trained SITS encoder is largely due to the inclusion of the temporal projector in ALISE. Furthermore, to ensure that these improved results are not caused by the use of the shallow classifier architecture combined with ALISE, ALISE^{FR} performances are compared with U-BARN-GF. Similarly to ALISE, U-BARN-GF aligned features are injected into a fully-connected layer which operate on the spectro-spatio-temporal features. We observe in Table III that U-BARN^{FR} and U-BARN-GF^{FR} have close performances and that ALISE^{FR} outperforms U-BARN-GF^{FR} on both data-sets. This result demonstrates the effectiveness of using a learnable temporal projector over a linear interpolation strategy. Furthermore, by design, ALISE provides more relevant representations in the frozen configuration than U-BARN.

2) ALISE vs Presto: We observe that ALISE^{FR} outperforms both frozen and fine-tuned Presto configurations, by 41.5% and 13,6%, respectively. To study more deeply the obtained performances, we compare the confusion matrices obtained by ALISE^{FT} and Presto^{FT} in Figure 8. We observe that ALISE and Presto show similar confusions between classes: winter triticale and soft winter wheat, leguminous fodder and meadow, corn and sorghum. Nevertheless, the confusion values between classes obtained by ALISE are each time lower than values reached by Presto. We posit that this result can be explained by several factors. Firstly, Presto is a lightweight spectro-temporal architecture that does not take spatial context into account. Such a design may not be relevant to segmentation tasks. Additionally, due to the required under-sampling protocol (Presto exploits monthly synthesis instead of all available acquisitions), it may miss key temporal information in comparison to ALISE. Furthermore, the implemented temporal positional encoding in the released code[15] raises questions. Traditionally, in the classical transformer model, the positional encoding is added or concatenated to the input data along the channel dimension. However, from our understanding of the code, in the proposed implementation, the positional encoding is concatenated along the temporal dimension.

B. Label scarcity scenario

To assess the behavior of the model in a severe label scarcity scenario, a reduced version of the PASTIS data-set has been created. Following the approach of [8], we have

Table III
F1 SCORE AVERAGED PER CLASS ON PASTIS AND MULTISENGE DATA-SETS. THE MEAN OF THE F1 SCORES ARE OBTAINED ON PASTIS' 5 FOLDS. ON THE MULTISENGE DATA-SET, TWO TRAININGS ARE CONDUCTED WITH DIFFERENT SEEDS. EACH COLOR CORRESPONDS TO A PRE-TRAINING CONFIGURATION, AND THE HIGHEST SCORE WITHIN A CONFIGURATION IS UNDERLINED. AS NO CLOUD MASKS ARE PROVIDED ON MULTISENGE, PRESTO CAN'T BE ASSESSED ON THIS SEGMENTATION TASK. THE NUMBER OF TRAINABLE PARAMETERS ARE ESTIMATED ON THE PASTIS TASK.

Name	Pre-training Data-Set	Trainable parame-ters	PASTIS F1	PASTIS OA	PASTIS mIoU	MSenGE F1	MSenGE OA	MSenGE mIoU
ALISE^{FR}	MMDC-EU	12.2K	0.68 ± 0.02	0.85 ± 0.01	0.56 ± 0.02	0.17 ± 0.00	0.57 ± 0.00	0.11 ± 0.00
ALISE^{FT}	MMDC-EU	1.1M	**0.81** ± 0.02	**0.91** ± 0.01	**0.70** ± 0.02	**0.23** ± 0.00	**0.63** ± 0.00	**0.16** ± 0.00
ALISE^{FS}	✗	1.1M	0.80 ± 0.01	**0.91** ± 0.01	0.69 ± 0.01	0.21 ± 0.00	0.62 ± 0.01	0.15 ± 0.00
Presto^{FR}	worlwide	2.5K	0.27 ± 0.01	0.62 ± 0.00	0.19 ± 0.01	✗	✗	✗
Presto^{FT}	worlwide	404K	0.55 ± 0.02	0.76 ± 0.01	0.41 ± 0.02	✗	✗	✗
U-BARN-GF^{FR}	MMDC-EU	12.2K	0.60 ± 0.01	0.82 ± 0.01	0.48 ± 0.01	0.12 ± 0.00	0.55 ± 0.00	0.09 ± 0.00
U-BARN^{FR}	MMDC-EU	13.8K	0.59 ± 0.03	0.82 ± 0.01	0.47 ± 0.03	0.14 ± 0.00	0.56 ± 0.00	0.10 ± 0.00
U-BARN^{FT}	MMDC-EU	1.1M	**0.81** ± 0.02	**0.91** ± 0.01	**0.70** ± 0.02	**0.23** ± 0.01	0.62 ± 0.00	**0.16** ± 0.01
U-BARN^{FS}	✗	1.1M	0.80 ± 0.01	0.90 ± 0.01	0.69 ± 0.02	**0.23** ± 0.01	0.62 ± 0.00	**0.16** ± 0.01
U-TAE	✗	1.1M	**0.81** ± 0.02	**0.91** ± 0.01	**0.70** ± 0.03	0.15 ± 0.02	0.61 ± 0.01	0.11 ± 0.01

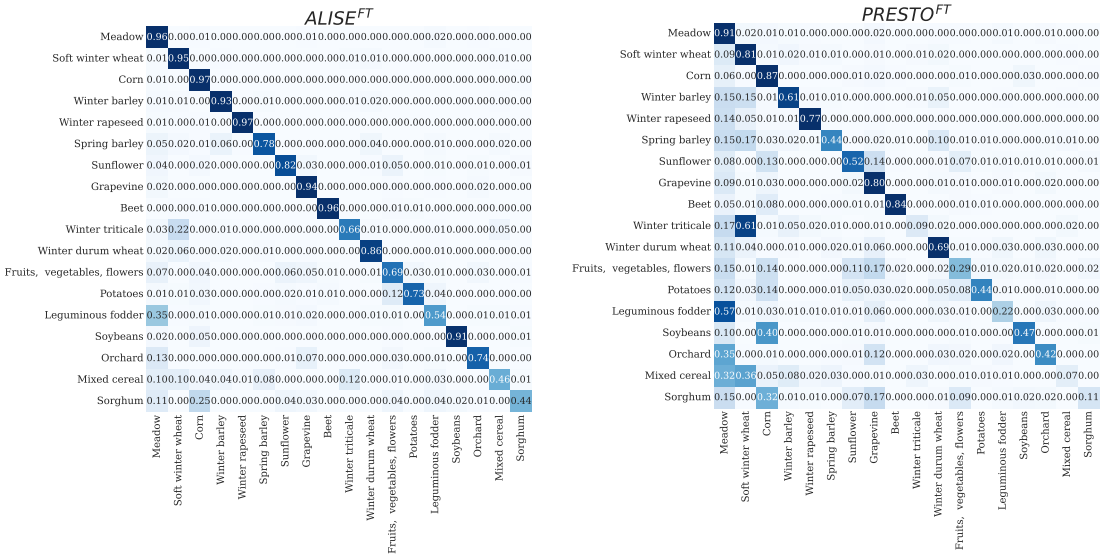

Figure 8. Confusion matrices on the PASTIS crop segmentation data-set obtained after fine-tuning. For each confusion matrix, rows correspond to true labels and columns to predictions. The matrices are normalized per row. On the left obtained with ALISE^{FT} and on the right Presto^{FT}.

used five smaller data-sets, each composed of 30 SITS for each PASTIS fold. Therefore, the results shown in Table VI correspond to the averaged macro F1 score across 25 trials. Under severe label scarcity, the fine-tuned model outperforms the fully-supervised framework by 12.5%. Interestingly, the frozen ALISE also outperforms its fully-supervised configuration by 9.7%. Given its reduced number of pre-trainable parameters compared to fully-supervised and fine-tuned approaches, ALISE^{FR} is an ideal candidate for scenarios with limited labeled data.

C. Change detection task

To quantitatively compare the change detection performances on the proposed *CropRot* dataset, we compute the area under the receiver operating characteristic curve (AUC) score. This score is calculated on the distance map computed between the representations of SITS from two different years. Table VII shows that better change detection results are obtained by comparing SITS representations encoded by ALISE. Besides, in contrast to the linear interpolation on the raw input SITS, ALISE change detection framework does not require cloud mask information. Furthermore, our results demonstrate that ALISE's representations are relevant for change detection, even though its positional encoding information is absolute and not relative to a year (day of the year). In other words, ALISE can still learn SITS invariance between different years, even if the positional encoding differs for each of the two years being compared. In addition, a qualitative analysis of the change detection maps is performed. Given two annual irregular and unaligned SITS from 2019 and 2020, Figure 9 illustrates the obtained change maps. In this example SITS, the number of available spring acquisitions is greater in 2020 than in 2019. Besides, Figure 9 shows class variability even when there is no change. These variations can be caused by different agricultural practices, meteorological events, and different acquisition dates. In Figure 9, this intra-class variability is observed when

Table IV
F1 SCORE PER CLASS ON PASTIS DATASET FOR EACH TRAINING CONFIGURATION. MEAN AND STANDARD DEVIATION OF THE F1 SCORE OBTAINED USING K-FOLD TRAINING ARE DETAILED.

	ALISE^{FR}	**ALISE^{FT}**	**ALISE^{FS}**	**Presto^{FR}**	**Presto^{FT}**	**U-BARN^{FR}**	**U-BARN^{FT}**	**U-BARN^{FS}**	**U-TAE**
Meadow	0.91 ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01	0.75 ± 0.01	0.85 ± 0.01	0.90 ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01
Soft winter wheat	0.89 ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01	0.65 ± 0.03	0.79 ± 0.02	0.86 ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01	**0.94** ± 0.01
Corn	0.93 ± 0.01	**0.96** ± 0.01	**0.96** ± 0.00	0.71 ± 0.01	0.85 ± 0.01	0.91 ± 0.01	**0.96** ± 0.01	**0.96** ± 0.01	0.96 ± 0.01
Winter barley	0.82 ± 0.02	**0.92** ± 0.01	**0.92** ± 0.02	0.21 ± 0.04	0.67 ± 0.03	0.77 ± 0.04	**0.92** ± 0.02	**0.92** ± 0.02	**0.92** ± 0.01
Winter rapeseed	0.91 ± 0.01	**0.96** ± 0.01	**0.96** ± 0.01	0.49 ± 0.04	0.81 ± 0.02	0.87 ± 0.04	**0.96** ± 0.01	**0.96** ± 0.01	**0.96** ± 0.01
Spring barley	0.67 ± 0.08	**0.79** ± 0.05	0.77 ± 0.06	0.01 ± 0.01	0.49 ± 0.07	0.61 ± 0.09	0.78 ± 0.06	0.76 ± 0.06	0.77 ± 0.05
Sunflower	0.71 ± 0.03	**0.83** ± 0.01	**0.83** ± 0.02	0.18 ± 0.03	0.56 ± 0.05	0.57 ± 0.07	**0.83** ± 0.04	0.81 ± 0.03	**0.83** ± 0.05
Grapevine	0.82 ± 0.02	**0.92** ± 0.01	0.91 ± 0.01	0.57 ± 0.06	0.71 ± 0.05	0.77 ± 0.03	0.91 ± 0.01	0.91 ± 0.00	0.91 ± 0.01
Beet	0.93 ± 0.02	**0.96** ± 0.01	0.95 ± 0.02	0.45 ± 0.10	0.85 ± 0.02	0.89 ± 0.01	**0.96** ± 0.01	**0.96** ± 0.02	**0.96** ± 0.01
Winter triticale	0.41 ± 0.06	0.70 ± 0.06	0.66 ± 0.06	0.00 ± 0.00	0.15 ± 0.05	0.16 ± 0.06	0.69 ± 0.04	0.65 ± 0.07	**0.73** ± 0.04
Winter durum wheat	0.74 ± 0.03	**0.83** ± 0.03	0.81 ± 0.03	0.41 ± 0.02	0.64 ± 0.03	0.69 ± 0.01	0.82 ± 0.03	0.80 ± 0.02	0.82 ± 0.04
Fruits/veg/flowers	0.49 ± 0.07	0.70 ± 0.03	0.65 ± 0.03	0.04 ± 0.02	0.34 ± 0.10	0.35 ± 0.05	**0.71** ± 0.03	0.65 ± 0.03	0.69 ± 0.06
Potatoes	0.65 ± 0.07	**0.76** ± 0.08	0.70 ± 0.05	0.10 ± 0.06	0.52 ± 0.09	0.55 ± 0.09	0.74 ± 0.03	0.70 ± 0.06	0.71 ± 0.09
Leguminous fodder	0.44 ± 0.08	0.60 ± 0.07	**0.64** ± 0.07	0.10 ± 0.02	0.30 ± 0.05	0.35 ± 0.11	0.63 ± 0.07	0.63 ± 0.09	0.61 ± 0.08
Soybeans	0.82 ± 0.04	0.92 ± 0.02	**0.93** ± 0.02	0.01 ± 0.02	0.54 ± 0.06	0.71 ± 0.07	**0.93** ± 0.02	0.92 ± 0.02	0.92 ± 0.03
Orchard	0.61 ± 0.05	0.77 ± 0.04	0.77 ± 0.04	0.12 ± 0.04	0.47 ± 0.07	0.55 ± 0.04	0.77 ± 0.05	0.77 ± 0.04	**0.79** ± 0.04
Mixed cereal	0.23 ± 0.08	0.53 ± 0.07	0.53 ± 0.05	0.00 ± 0.00	0.11 ± 0.04	0.08 ± 0.05	0.54 ± 0.07	0.54 ± 0.05	**0.55** ± 0.06
Sorghum	0.29 ± 0.08	0.50 ± 0.11	0.50 ± 0.09	0.00 ± 0.00	0.18 ± 0.08	0.10 ± 0.09	0.52 ± 0.11	0.49 ± 0.12	**0.53** ± 0.13

Table V
F1 SCORE PER CLASS ON THE MULTISENGE SEGMENTATION TASK FOR EACH CONFIGURATION. TWO TRAININGS ARE PERFORMED FOR EACH CONFIGURATION.

	ALISE^{FS}	**ALISE^{FT}**	**ALISE^{FR}**	**U-BARN^{FR}**	**U-BARN^{FS}**	**U-BARN^{FT}**	**U-TAE**
Dense Built-Up	**0.04** ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Sparse Built-Up	0.11 ± 0.02	0.13 ± 0.00	0.02 ± 0.01	0.01 ± 0.00	0.14 ± 0.02	0.14 ± 0.02	**0.15** ± 0.01
Specialized Built-Up Areas	0.13 ± 0.00	**0.31** ± 0.01	0.07 ± 0.00	0.02 ± 0.00	0.27 ± 0.03	0.27 ± 0.09	0.02 ± 0.02
Specialized but Vegetative Areas	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Large Scale Networks	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Arable Lands	0.67 ± 0.01	**0.68** ± 0.00	0.62 ± 0.00	0.62 ± 0.00	**0.68** ± 0.00	**0.68** ± 0.00	0.67 ± 0.01
Vineyards	0.46 ± 0.04	**0.51** ± 0.01	0.32 ± 0.00	0.21 ± 0.04	0.47 ± 0.04	0.48 ± 0.02	0.19 ± 0.27
Orchards	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Grasslands	0.39 ± 0.04	0.39 ± 0.00	0.33 ± 0.01	0.31 ± 0.02	0.40 ± 0.01	**0.41** ± 0.01	0.40 ± 0.00
Groces,Hegdes	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Forest	0.72 ± 0.01	**0.73** ± 0.00	0.67 ± 0.00	0.66 ± 0.00	**0.73** ± 0.00	**0.73** ± 0.00	0.71 ± 0.02
Open Spaces,Mineral	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Wetlands	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table VI
MACRO-AVERAGED F1 SCORES OBTAINED ON PASTIS ON LABELED DATA SCARCITY SCENARIO. EACH PASTIS FOLD IS COMPOSED OF 30 LABELED SITS.

Model	F1
ALISE^{FR}	0.44 ± 0.01
ALISE^{FT}	**0.47** ± 0.04
ALISE^{FS}	0.34 ± 0.06

Table VII
AREA UNDER THE ROC CURVE METRIC ON CROPROT.

Input Data	AUC
ALISE representations	**0.91**
raw interpolated SITS	0.88

looking at the fields located at the center bottom of the SITS (red circle). Although the crop class of this field has not changed between 2019 and 2020, we can visually observe important differences between 14/05/2019 and 18/05/2020 which are supposed to be close acquisitions. Nevertheless, the distance map shown in Figure 9 is not affected by such input intra-class variability. In addition, compared to the distance map obtained from interpolated raw SITS, the distance on ALISE representations better distinguishes modified crops from unchanged ones.

D. Co-influence of $t_w, w_{inv}, w_{cov}, w_{rec}$

t_w is a hyper-parameter involved in the view generation process and detailed in subsubsection III-B1. More precisely, t_w is the number of acquisitions contained in the temporal window used to build each view. Increasing t_w is supposed to create greater discrepancies between views therefore impacting both the discriminative and the cross-reconstruction losses. As t_w is the number of consecutive acquisitions, when t_w is increased, the cross-reconstruction task is no longer a simple interpolation task. Therefore, we aim to assess the co-influence of the view generation protocol (controlled by t_w) and the losses weights. Different pre-training configurations evaluating the impact of the four parameters $(t_w, w_{inv}, w_{cov}, w_{rec})$ have been performed. For each configuration, results obtained on the five PASTIS folds are averaged by considering four different pre-trained models with different seeds. Only the loss weights and the t_w parameter are evaluated, all the rest of hyper-parameters are fixed for the rest of pre-trained model configurations. The covariance weight value is set to

Figure 9. Visualization of a change map obtained on the change detection data-set with the pre-trained ALISE. The top and bottom rows represent a portion of the S2 SITS along with their crop classes for 2019 and 2020, respectively. These SITS portions have similar index position within their SITS. In the crop label maps, dark blue represents the background class. To the right, the distance maps computed from the aligned representations from ALISE and the Gap-Filling methods are shown. The same scale is used in the colorbar of the distance maps. Pixels that belong to the background class are masked. On the far right, the label change map is represented with, in white the background class, in blue the *no change* label, and in yellow the *change* label.

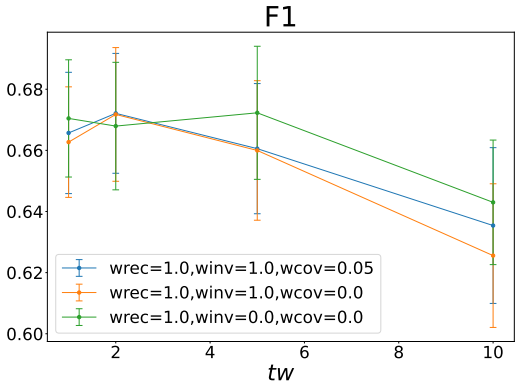

Figure 10. Segmentation task performances on PASTIS linear probing as a function of t_w. In all these experiments $n_q = 10$, 4 pre-trainings were conducted and their performances on 5 of PASTIS folds were evaluated.

0.05 to reproduce the balance between the invariance and covariance losses suggested in VicReg [44]. The influence of t_w is analyzed by studying the macro averaged F1 score before providing a more precise analysis per crop class.

1) Macro averaged F1 score: Figure 10 displays the linear probing performances as a function of t_w. This figure shows that the additional invariance latent loss (L_{inv}) significantly degrades the linear probing performances for t_w greater than 2, (the orange and blue curves are lower than the green one). We assume that the invariance loss might constrain too much the total loss when pre-training data is composed of large dissimilar views. For $t_w = 2$, there seems to be a slight improvement in the linear probing performances when invariance loss is incorporated. Lastly, these experiments do not show any benefit from using the covariance loss (L_{cov}) in addition to the invariance loss. There are several possible explanations for this outcome. First, the large memory size of SITS limits the batch size, therefore experiments have been conducted with a batch size equal to 2. Although we use $b \times n_q \times h \times w$ samples to estimate the covariance matrix, these samples are correlated. In the original VicReg implementation [44], the covariance was estimated across 2048 samples, each corresponding to a different image. Second, the covariance loss in VicReg plays a crucial role in preventing information collapse. In our framework, the cross-reconstruction loss prevents collapsing, making the covariance loss less important during pre-training. Third, the projector architecture may impact the computation of the covariance loss. Therefore, more experiments studying the impact of the batch size and the projector could be necessary. Lastly, the green curve in Figure 10 depicts the influence of t_w when solely the reconstruction loss is applied. In this case, the downstream segmentation performance is impacted also by t_w. With large temporal windows ($t_w = 10$), the pre-training reconstruction task may become too complex, which prevents the model from learning informative SITS representations. Surprisingly, with smaller values of $t_w \leq 5$, no major differences are observed. This could be explained by the fact that, unlike regular time series processing, t_w does not control the temporal extent that is reconstructed. Inherent important temporal gaps in S2 SITS might provide a complex pre-training task even with $t_w = 1$.

2) F1 score per class: We propose a more in-depth analysis of the effect of t_w and the pre-training loss weights in Figure 11. Notably, similar to the previous experiment, the F1 score for each PASTIS crop class is plotted as a function of t_w. Different behaviors are observed depending on the crop classes. For many crop classes, there is a decrease in the F1 score when t_w increases. However, some crop classes such as meadow, corn, spring barley, grapevine, fruits, vegetables & flowers, potatoes, leguminous fodder, and orchard are unaffected by t_w. With the exception of grassland, maize and spring barley, we hypothesize that the lack of t_w effect for these classes may be linked to the fact that they are either permanent (grapevine, orchards) or greenhouse. Interestingly, the soybean class exhibits an outlier behavior, with an increase in F1 score as t_w increases. This experiment demonstrates that the influence of pre-training conditions differs depending on the target class.

E. Impact of n_q

For practical purposes, it is relevant to reduce the size of the latent representation (n_q) while preserving the downstream tasks performances. Figure 12 plots the segmentation performances on the PASTIS data-set as a function of n_q. For each configuration, one pre-training was conducted, and the performances were assessed on one out of the five available PASTIS experiments. We observe that increasing the value of n_q improves downstream task F1 score. This can be attributed to two factors. Firstly, a greater value of n_q means a larger latent space, and therefore potentially more information contained within it. Secondly, it is assumed that a smaller value of n_q makes the cross-reconstruction task more difficult due to the compression performed in the proposed temporal projector.

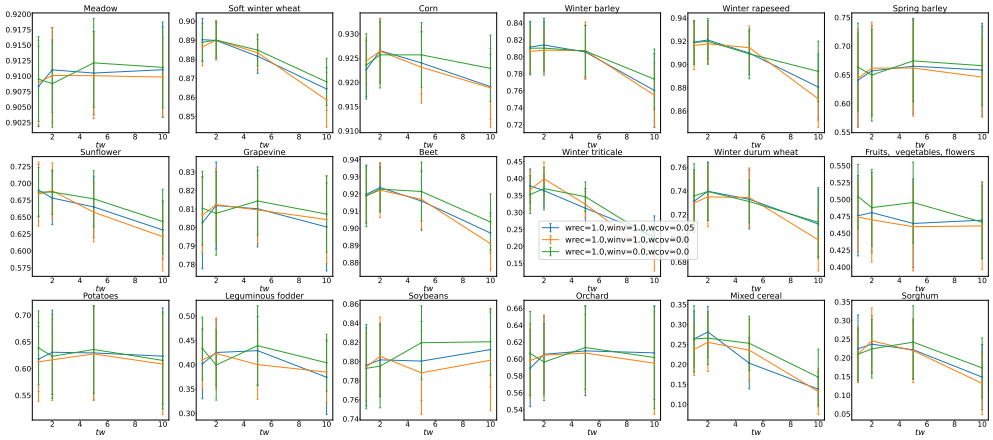

Figure 11. F1 score per class on PASTIS linear probing as a function of t_w. In all these experiments $n_q = 10$. Results are averaged over 4 pre-trained models for all 5 PASTIS folds.

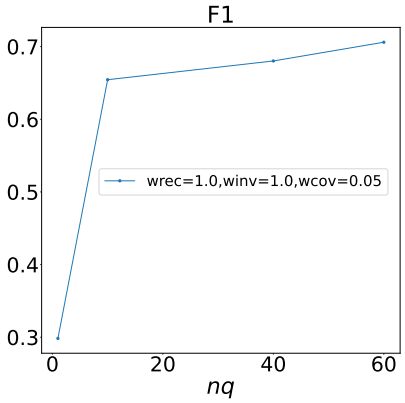

Figure 12. Segmentation task performances in linear probing configuration on the PASTIS data-set as a function of n_q. In these experiments, $t_w = 5$. For each configuration, one pre-training is conducted and the downstream task is performed on one out of five PASTIS experiments.

This compression could penalize the cross-reconstruction pre-training task. Nevertheless, these comparisons with different possible values of n_q may not be totally fair. Indeed, a higher value of n_q results in a larger classifier during linear probing, which may have a positive impact on downstream task performances.

F. Qualitative analysis of the temporal projector

To gain a better understanding of the information encoded by ALISE, we conduct a qualitative analysis of the latent representations. For this purpose, we propose to study how latent information is used by the self-attention mechanism of the decoder during the reconstruction process. We note the pixel-level latent temporal vector, indexed by i, as $\mathbf{y}_{(\mathbf{i},.)}$. To understand the importance of each latent temporal feature during the reconstruction process, the attention weights of each decoder head are displayed in Figure 13. For improved visualization, the reconstruction decoder is asked to reconstruct in Figure 13, a regularly sampled time series from 2017-01-01 to 2021-07-19 with a step of 10 days. The reconstruction temporal grid considered here corresponds to the years observed during pre-training (2017-2020) as well as years outside the temporal extent of the pre-training data-set (2021). In an attention matrix, a high attention score at a given row (indexed by i) and column (corresponding to a date d_j) indicates the importance of latent temporal feature $\mathbf{y}_{(\mathbf{i},.)}$ for the reconstruction of the date d_j. For each latent temporal feature (row), high attention scores (bright color) are often observed on specific narrow intervals, while attention weights are low outside of them. A notable finding is that for a latent temporal feature, these intervals are often separated by approximately one year. Although no annual periodicity is explicitly given as input to the decoder, dates to reconstruct spaced of 365 days exhibit similar high attention score on the same latent temporal feature. Furthermore, it is worth noting these annual periodic patterns are also observed for reconstructed dates not included in the pre-training (year 2021), suggesting that the model might have forecasting (extrapolation) abilities.

The influence of the latent temporal features on the reconstruction is also illustrated in Figure 14. Specifically, the reconstruction is conducted with either all latent temporal features, a single latent temporal feature, or a triplet of features. The top row depicts random acquisition dates within the input SITS, while the next rows display the reconstruction of the decoder of the SITS on dates ranging from 2017-01-01 to 2021-11-01, with acquisitions spaced by 70-day intervals. Consequently, the image acquisitions of the first row are not temporally aligned with the rows below. Nevertheless, the predictions generated by the model using all latent temporal features (second row) and the input SITS (first row) are often consistent. For example, the acquisitions marked by the blue, green, and orange rectangles are temporally close to some of the input predictions. These latter acquisitions show coherent reconstructions, which highlights that the use of the temporal alignment projector does not lead to a significant loss of information. The following three rows in Figure 14

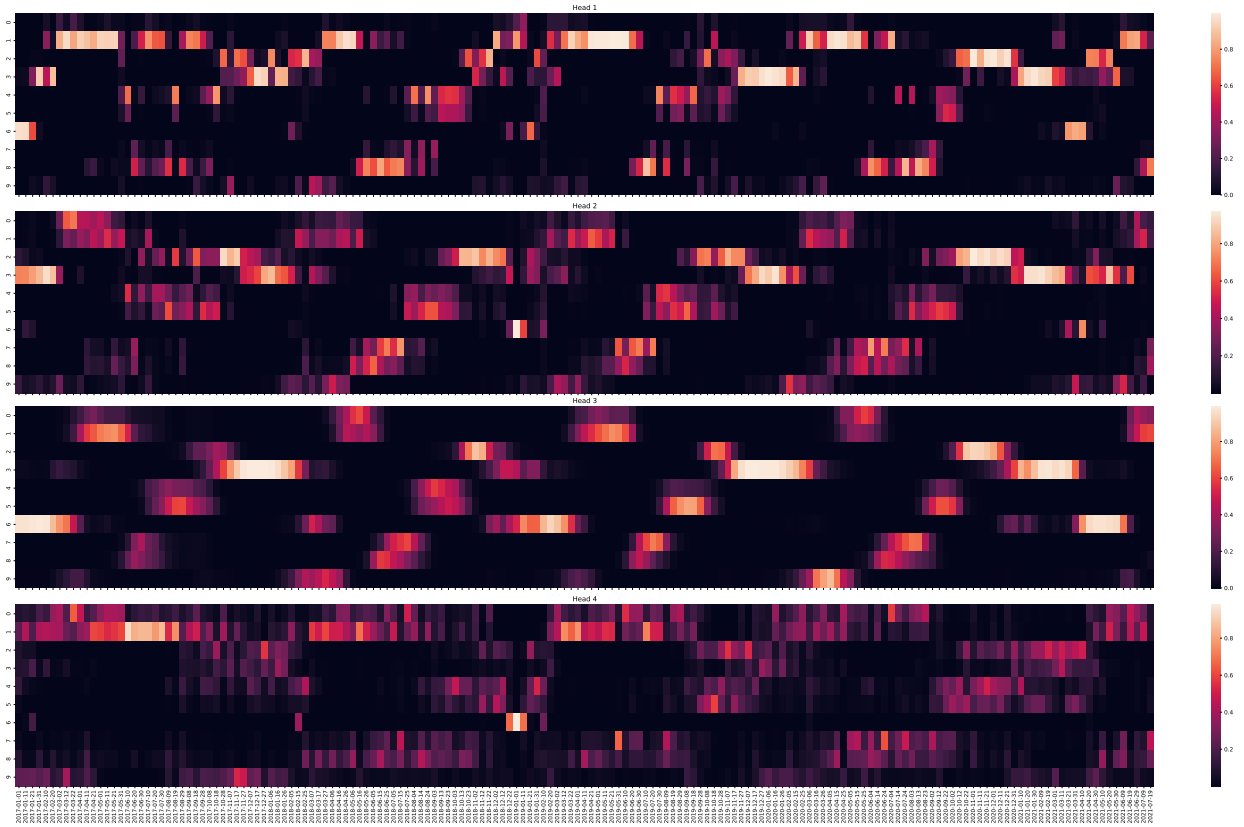

Figure 13. Attention weights of the reconstruction decoder when reconstructing a time series from 2017-01-01 to 2021-07-19. Attention matrices plotted correspond to the average attention matrix obtained for each pixel of the SITS. Each matrix from top to bottom corresponds to a different head. On each score matrix, the column corresponds to the latent temporal feature, while the row corresponds to the dates to reconstruct.

illustrate reconstruction results obtained when a single latent temporal feature is used during the decoding stage, $\mathbf{y}_{(\mathbf{9},.)}$, $\mathbf{y}_{(\mathbf{8},.)}$ and $\mathbf{y}_{(\mathbf{5},.)}$, respectively. For each row, it can be observed that the model invariably generates the same image, revealing that the temporal dynamics are not reconstructed. Conversely, the temporal dynamic can be observed in the reconstructions obtained from only three latent temporal features (last row). For this case, the contribution of each latent temporal feature to the image reconstruction process is easily recognizable. In the proposed example, the reconstruction obtained from $\mathbf{y}_{(\mathbf{9},.)}$ highly contributes to the reconstruction of winter images shown on the last row. The latent temporal features $\mathbf{y}_{(\mathbf{8},.)}$ and $\mathbf{y}_{(\mathbf{5},.)}$ seem to intervene for May to July and July to November, respectively. In this last row, the majority of the images appear to be highly similar to the reconstruction obtained from one single latent temporal feature. Nevertheless, a few acquisitions seem to be the obtained through the mix of information from different latent temporal features. For instance for 2017-05-21, it seems that the latent temporal features $\mathbf{y}_{(\mathbf{8},.)}$ and $\mathbf{y}_{(\mathbf{9},.)}$ are merged. Areas showing this combination are marked by the red arrows in Figure 14. From these observations we may consider the aligned representations as a novel basis for input SITS representation. From this point of view, latent temporal feature data would serve as prototype and the attention weights in the decoder would fulfill the role of membership degree.

VI. Conclusion

This article discusses the notable challenges involved in learning to represent satellite image time series (SITS). In particular, our work paves the way for the construction of a FM for land surface monitoring using Earth observation.

This paper proposes a new SITS encoder named ALISE, which exploits spatial, spectral and temporal dimensions and generates aligned and fixed-size representations of irregular and unaligned multi-year SITS. ALISE is pre-trained using a new multi-view hybrid SSL pre-training task that combines MAE loss with instance discrimination losses. In addition, ALISE pre-training data-set (MMDC-EU) is a custom-built large-scale multi-year unlabeled dataset. The quality of ALISE's representation has been assessed on three downstream data-sets: PASTIS (crop segmentation), MultiSenGE (dense land cover segmentation) and the novel CropRot (crop change detection). Our results demonstrate the significant progress made with regard to the three representation characteristics considered: **easy to use**, **informative** and **generic**.

Firstly, ALISE provides aligned, fixed-size representations that preserve the spatial resolution of the input data. Consequently, ALISE representations can be easily exploited by a shallow classifier. In this paper, a single fully connected layer was used to perform two segmentation tasks (crop mapping and dense land cover). Results have also shown that pre-trained and frozen ALISE outperforms the fully supervised approach when labeled data are scarce. The remarkable per-

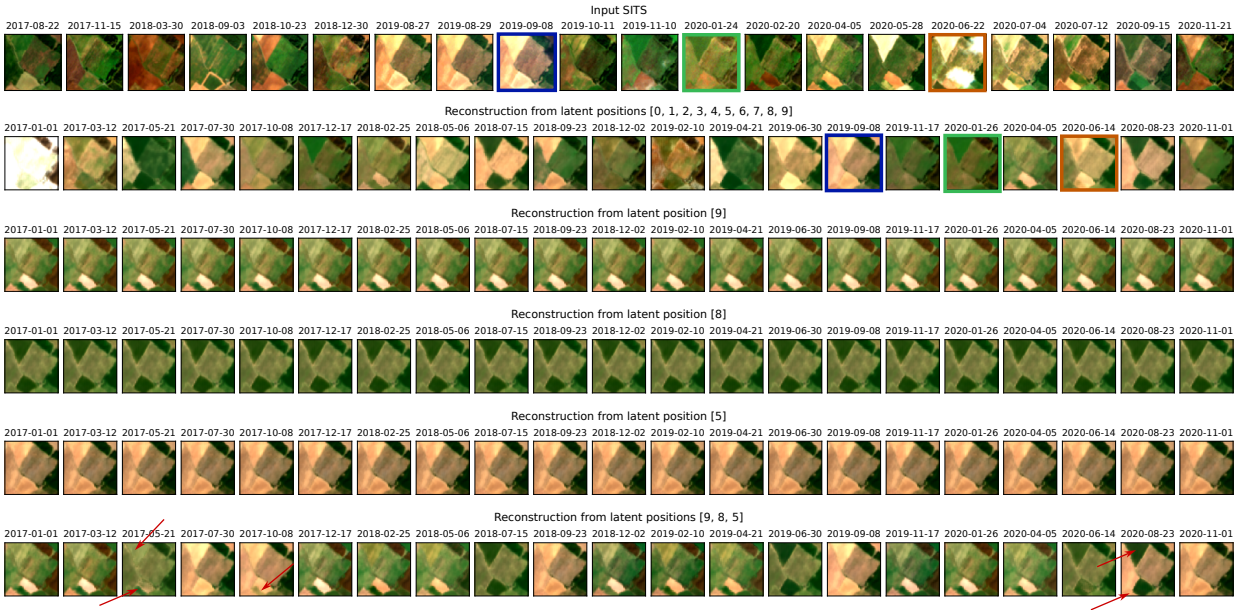

Figure 14. Reconstructions obtained from the decoder given different latent temporal features configurations. First row: random dates from the input SITS. Then reconstruction obtained by using: all latent temporal features (second row), latent temporal feature n°9 (third row), latent temporal feature n°8 (fourth row), latent temporal feature n°5 (fifth row), latent temporal features 9,8 and 5 (bottom row).

formance of frozen and pre-trained ALISE indicates that an important step has been taken towards the creation of **ready-to-use** SITS representations. The production of aligned, fixed-size representations has been achieved through the use of a learnable query-based cross-attention mechanism. We have also provided the first qualitative study of the aligned latent temporal features obtained through this latter mechanism. It appears that each latent pseudo-date summarizes a specific part of the input SITS. To reconstruct a SITS, the pre-training decoder successfully recovers the annual periodicity of the SITS, whereas our temporal encoding does not rely on the day of the year.

Secondly, the quality of ALISE representations was compared with existing competitive works. Results obtained by pre-trained and frozen ALISE outperform Presto [7] and U-BARN [8] for both semantic segmentation tasks. As a result, ALISE's representations may be considered more **informative** than other existing works. In addition, we have studied in depth our proposed hybrid SSL approach. Notably, the impact of the view generation method and the contribution of each loss has also been investigated. Our results show that most of the pre-training is driven by the cross-reconstruction task. Nevertheless, depending on how the view generation is performed, which is strongly influenced by t_w, the invariance loss may or may not improve performances. This leads us to think that other view generation protocols could be investigated. Besides our experiments did not reveal a significant contribution from the covariance loss. These unexpected findings also highlight the important challenges that remain in applying ideas from the wider computer vision field to the specificities of SITS (temporal dynamics, physics of the measure, etc.).

Thirdly, the **genericity** of the representations was assessed on three proposed downstream tasks. In addition to the great performances obtained in both segmentation tasks, our results demonstrate that the proposed aligned SITS representations can be used for downstream unsupervised change detection tasks. We also consider the proposed novel crop change detection data-set named *CropRot*, as an important contribution to assess future FM on SITS. Besides, to learn **generic** representations, ALISE was pre-trained on a new large-scale and multi-year data-set. Nevertheless, the geographical diversity of the pre-training and downstream data-sets could be improved, since pre-training and downstream data only includes European geographical areas. The development of a scalable method, trained and evaluated on numerous geographical configurations, remains unexplored here.

It should also be noted, however, that this article does not address certain aspects. For instance, ALISE memory consumption is quadratic with the temporal size of the input SITS. Therefore, lightweight architectures based on learnable queries [61], [62], [63] could be considered. Additionally, the construction of a global pre-training data-set as well as the study of the incorporation of thermal encoding [64] to improve spatial scalability are worthy of interest. Lastly, a major remaining challenge in developing FM is the processing of multi-sensor data. For example, combining S1 data with S2 data is beneficial when optical data are unavailable due to unsuitable weather conditions. Furthermore, using different modalities in a multi-view SSL protocol is promising, and we might observe a greater contribution from instance discrimination losses in this context.

REFERENCES

[1] F. Spoto, O. Sy, P. Laberinti, P. Martimort, V. Fernandez, O. Colin, B. Hoersch, and A. Meygret, "Overview Of Sentinel-2," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 1707–1710.

[2] V. S. F. Garnot and L. Landrieu, *Lightweight Temporal Self-attention for Classifying Satellite Images Time Series*, ser. Advanced Analytics

and Learning on Temporal Data. Springer International Publishing, 2020, pp. 171–181. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-65742-0_12

[3] ——, "Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10–17.

[4] C. Pelletier, G. Webb, and F. Petitjean, "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019. [Online]. Available: http://dx.doi.org/10.3390/rs11050523

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," 2022. [Online]. Available: https://arxiv.org/abs/2108.07258

[6] 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*, May 2024.

[7] G. Tseng, I. Zvonkov, M. Purohit, D. Rolnick, and H. R. Kerner, "Lightweight, Pre-trained Transformers for Remote Sensing Timeseries," *ArXiv*, vol. abs/2304.14065, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258352331

[8] I. Dumeur, S. Valero, and J. Inglada, "Self-Supervised Spatio-Temporal Representation Learning of Satellite Image Time Series," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–18, 2024. [Online]. Available: http://dx.doi.org/10.1109/JSTARS.2024.3358066

[9] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, "OmniSat: Self-Supervised Modality Fusion for Earth Observation," 2024.

[10] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 672–27 683.

[11] Y. Yuan and L. Lin, "Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 474–487, 2021. [Online]. Available: http://dx.doi.org/10.1109/JSTARS.2020.3036602

[12] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z.-G. Zhou, "SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102651, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243421003585

[13] 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*, ser. Lecture Notes in Computer Science. Springer Nature Switzerland, 2022, pp. 456–473. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-19821-2_26

[14] H. Wang, X. Guo, Z. Deng, and Y. Lu, "Rethinking Minimal Sufficient Representation in Contrastive Learning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 020–16 029, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247446649

[15] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised Learning from a Multi-view Perspective," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=-bdp_8Itjwp

[16] 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. [Online]. Available: http://dx.doi.org/10.1109/MGRS.2023.3281651

[17] 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. [Online]. Available: https://openreview.net/forum?id=WBhqzpF6KYH

[18] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, K. Ankur, M. Ramasubramanian, I. Gurung, S. Khallaghi, H. S. Li, M. Cecil, M. Ahmadi, F. Kordi, H. Alemohammad, M. Maskey, R. Ganti, K. Weldemariam, and R. Ramachandran, "Foundation Models for Generalist Geospatial Artificial Intelligence," *Preprint Available on arxiv:2310.18660*, 10 2023.

[19] I. Dumeur, S. Valero, and J. Inglada, "Unlabeled Sentinel 2 time series dataset : Self-Supervised Spatio-Temporal Representation Learning of Satellite Image Time Series," May 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7891924

[20] R. Wenger, A. Puissant, J. Weber, L. Idoumghar, and G. Forestier, "Multisenge: a Multimodal and Multitemporal Benchmark Dataset for Land Use/land Cover Remote Sensing Applications," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-3-2022, pp. 635–640, 2022. [Online]. Available: http://dx.doi.org/10.5194/isprs-annals-V-3-2022-635-2022

[21] I. Dumeur, S. Valero, and J. Inglada, "CropRot," Sep. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13832554

[22] ——, "MMDC Europe (full dataset)," Sep. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13790323

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 10 2018. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[24] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in *The Eleventh International Conference on Learning Representations* , 2023. [Online]. Available: https://openreview.net/forum?id=Jbdc0vTOcol

[25] H. Liu, J. Gan, X. Fan, Y. Zhang, C. Luo, J. Zhang, G. Jiang, Y. Qian, C. Zhao, H. Ma, and Z. Guo, "PT-Tuning: Bridging the Gap between Time Series Masked Reconstruction and Forecasting via Prompt Token Tuning," 2023.

[26] M. Cheng, Q. Liu, Z. Liu, H. Zhang, R. Zhang, and E. Chen, "TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders," 2023. [Online]. Available: https://arxiv.org/pdf/2303.00320.pdf

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988.

[28] 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, 2023.

[29] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "Spectralgpt: Spectral Remote Sensing Foundation Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2024.3362475

[30] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang, C. Chen, J. Yu, X. Sun, and K. Fu, "Ringmo-Sense: Remote Sensing Foundation Model for Spatiotemporal Prediction Via Spatiotemporal Evolution Disentangling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–21, 2023. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2023.3316166

[31] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[32] 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.

[33] C. Tao, X. Zhu, W. Su, G. Huang, B. Li, J. Zhou, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2132–2141.

[34] O. Manas, A. Lacoste, X. G. i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2021. [Online]. Available: http://dx.doi.org/10.1109/ICCV48922.2021.00928

[35] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7797–7808, 2022.

[36] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 213–247, 2022.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[38] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, *Deep Clustering for Unsupervised Learning of Visual Features*, ser. Computer Vision - ECCV 2018. Springer International Publishing, 2018, pp. 139–156. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-01264-9_9

[39] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[40] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," in *ICCV 2021 - International Conference on Computer Vision*, Virtual, France, Oct. 2021, pp. 1–21. [Online]. Available: https://hal.science/hal-03323359

[41] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

[42] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 745–15 753.

[43] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-Supervised Learning via Redundancy Reduction," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 12 310–12 320. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[44] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR*, 2022.

[45] ——, "VICRegL: Self-Supervised Learning of Local Visual Features," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=ePZsWeGJXyp

[46] P. O. Pinheiro, A. Almahairi, R. Y. Benmalek, F. Golemo, and A. Courville, "Unsupervised learning of dense visual representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[47] Y. Yuan, L. Lin, Z.-G. Zhou, H. Jiang, and Q. Liu, "Bridging Optical and Sar Satellite Image Time Series Via Contrastive Feature Extraction for Crop Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 222–232, 2023. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2022.11.020

[48] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational High Resolution Land Cover Map Production At the Country Scale Using Satellite Image Time Series," *Remote Sensing*, vol. 9, no. 1, p. 95, 2017. [Online]. Available: http://dx.doi.org/10.3390/rs9010095

[49] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1685–1689, 2017.

[50] V. Bellet, M. Fauvel, and J. Inglada, "Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–21, 2023.

[51] V. Bellet, M. Fauvel, J. Inglada, and J. Michel, "End-to-End Learning for Land Cover Classification Using Irregular and Unaligned SITS by Combining Attention-Based Interpolation With Sparse Variational Gaussian Processes," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 2980–2994, 2024.

[52] S. N. Shukla and B. Marlin, "Multi-Time Attention Networks for Irregularly Sampled Time Series," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=4c0J6lwQ4_

[53] S. M. Kazemi, R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Poupart, and M. Brubaker, "Time2Vec: Learning a Vector Representation of Time," 2020. [Online]. Available: https://openreview.net/forum?id=rklklCVYvB

[54] M. Rußwurm and M. Körner, "Self-Attention for Raw Optical Satellite Time Series Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, 2020. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2020.06.006

[55] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=fILj7WpI-g

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[57] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[58] J. Louis, V. Debaecker, B. Pflug, M. Main-Knorn, J. Bieniarz, U. Mueller-Wilm, E. Cadau, and F. Gascon, "Sentinel-2 Sen2Cor: L2A processor for users," in *Proceedings living planet symposium 2016*. Spacebooks Online, 2016, pp. 1–8.

[59] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure," *Remote Sensing*, vol. 11, no. 4, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/4/433

[60] F. Levavasseur, P. Martin, C. Bouty, A. Barbottin, V. Bretagnolle, O. Thérond, O. Scheurer, and N. Piskiewicz, "RPG Explorer: A new tool to ease the analysis of agricultural landscape dynamics with the land parcel identification system," *Computers and Electronics in Agriculture*, vol. 127, pp. 541–552, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.compag.2016.07.015

[61] C. Yang, J. Xu, S. D. Mello, E. J. Crowley, and X. Wang, "GPVit: A high resolution non-hierarchical vision transformer with group propagation," in *The Eleventh International Conference on Learning Representations* , 2023. [Online]. Available: https://openreview.net/forum?id=IowKt5rYWsK

[62] X. Cai, Y. Bi, P. N. Nicholl, and R. Sterritt, "Revisiting the Encoding of Satellite Image Time Series," in *34th British Machine Vision Conference 2022, BMVC 2022, Aberdeen, UK, November 20-24, 2023*. BMVA Press, 2023, pp. 402–404. [Online]. Available: http://proceedings.bmvc2023.org/402/

[63] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General Perception with Iterative Attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 4651–4664. [Online]. Available: https://proceedings.mlr.press/v139/jaegle21a.html

[64] J. Nyborg, C. Pelletier, and I. Assent, "Generalized Classification of Satellite Image Time Series With Thermal Positional Encoding," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2022, pp. 1392–1402.

APPENDIX

ACKNOWLEDGMENTS

This work was funded by the ANR-JCJC DeepChange project under Grant Agreement 20-CE23-0003 and by the EU Horizon Europe EvoLand project under grant agreement 101082130. This work was partly performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014912) This work was partly performed using HPC resources from CNES Computing Center.

A. CropRot additional information

The change matrix between 2019 and 2020 is represented by Figure 15. As expected, the rate of change depends on the considered class. We observe important rotations between cereal and corn, while grassland mostly remains unchanged.

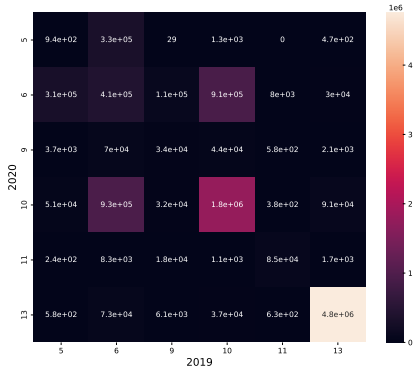

Figure 15. Change matrix between years 2019 and 2020 on the crop classes. Classes correspondence is {5: rapeseed, 6: cereal, 9 : sunflower, 10 : corn, 11 : rice, 13: grassland}

B. Comparison with VicRegL

The proposed discriminative losses are similar to those of VicRegL [45]. However, three notable modifications have been introduced. Firstly, unlike VicRegL, the invariance loss does not require any matching functions to realign the pixels of both views since geometric augmentation is not performed. In our case, each embedded vector at the pixel level is compared with the embedded vector of the other view at the same spatial position. Secondly, the large SITS size strongly constrains the batch size, which differs from the larger batch values of VicRegL. In VicRegL, the covariance loss is computed for each pixel of the latent representation using the b samples of the batch. The final local covariance loss is the sum over the spatial dimensions $h \times w$ of the pixel-level losses. Instead of estimating a covariance for each pixel, our covariance loss is estimated for the d_{emb} variables using $b \times h \times w$ samples. Thirdly, the variance loss is not considered in our approach. If the variance was estimated by considering $b \times h \times w$ samples, keeping the variance of each variable above a threshold would enforce a strong variability between pixels that might come from the same image. This loss could then deteriorate the spatial consistency of the representation.

C. ALISE architecture

1) Other architecture hyper-parameters:

(a) U-BARN Table VIII and Table IX describe the architectural hyper-parameters of the spatio-spectro-temporal encoder.

Table VIII
HYPER-PARAMETERS OF THE ARCHITECTURE OF THE UNET ENCODER, WITH B AND T RESPECTIVELY THE BATCH AND TEMPORAL DIMENSIONS. THE *down block* ARCHITECTURE IS DETAILED IN [8]

Block Name	Input dimensions	Output dimensions
Input Convolution	(B*T,64,64,10)	(B*T,64,64,64)
Down Block 1	(B*T,64,64,64)	(B*T,32,32,64)
Down Block 2	(B*T,32,32,64)	(B*T,16,16,64)
Down Block 3	(B*T,16,16,64)	(B*T,8,8,128)

Table IX
ARCHITECTURAL HYPER-PARAMETERS OF THE TRANSFORMER IN U-BARN

N_{layers}	N_{head}	$attn_{dropout}$	dropout	d_{model}	d_{hidden}
3	4	0.1	0.1	64	128

(b) Temporal projector.
The temporal projector is composed of a lightweight multi-head cross-attention mechanism with two heads. Inspired by the attention mechanism proposed in [2], the channels of the input embeddings are distributed among the heads.

D. MultiSenGE data distribution