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 real/operational applications. Indeed, embeddings that don't 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 uses. As a consequence, 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. Unlike SSL models currently available for SITS, ALISE incorporates a flexible query mechanism to project the SITS into a common and learned temporal projection space. Additionally, thanks to a multi-view framework, we explore integration of instance discrimination along a masked autoencoding task to SITS. The quality of the produced representation is assessed through three downstream tasks: crop segmentation (PASTIS), land cover segmentation (MultiSenGE), and a novel crop change detection dataset. Furthermore, the change detection task is performed without supervision. The results suggest that the use of aligned representations is more effective than previous SSL methods for linear probing segmentation tasks. Additionally, the experiments show that ALISE representations are suitable for change detection. Lastly, the code and 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, a number of satellite missions have been launched with the objective of monitoring the changes induced by climate change. To detect these shifts, missions such as Sentinel-2 [1] provide multi-spectral land surface imagery with a high temporal revisit. These data, which can be exploited in the form of Satellite Image Time Series (SITS), provide crucial information for Earth monitoring tasks such as land use classification, agricultural management, climate change or disaster monitoring [2], [3], [4], [5]. However, these applications often lack labeled data, which hinders the development of scalable methods that cover a wide range of temporal and geographical configurations. Therefore, pre-trained foundation models are a promising solution to significantly reduce the need for labeled data in these applications. Thanks to their self-supervised pre-training, these models can learn from vast unlabeled datasets. However, despite their potential and the abundance of open-source satellite data, pre-trained remote sensing foundation models with SITS remain largely unexplored. In this paper, we address three key obstacles to constructing remote sensing foundation models designed to generating **easy-to-use** and **meaningful** SITS representations.

First, due to the critical role of temporal signals in Earth monitoring, remote sensing foundation models must take into account the specificities of SITS. These time series often have varying acquisition dates and revisit frequencies, leading to unalignment and irregularity, respectively. We posit that existing methods [6], [7], [8], [9] do not produce SITS representations that are user-friendly for geoscientists. We propose that to ensure usability, the pre-trained model should require no further training for downstream tasks (remain frozen), and the latent representation should be aligned and of fixed dimension. In contrast, current methods generate SITS representations with temporal dimensions matching those of the input SITS, resulting in non-aligned representations of variable temporal size.

Second, the pre-training strategy used to train a foundation model should yield meaningful SITS representations. Masked auto-encoders have been frequently employed for SITS pre-training due to their ease of implementation [7], [6], [8], [9]. However, these strategies predict in a low-semantic space, which can limit the extraction of high-level semantic features in the representations [10]. In contrast, other self-supervised learning (SSL) techniques propose to perform the self-supervision at the latent space level. 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 while avoiding representation collapse. Instance discrimination remains largely unexplored in SITS because it requires aligned representations, specific domain data augmentation, and often benefits from large batch sizes. Moreover, recent researches [11], [12] suggest combining various SSL strategies, such as instance discrimination with masked auto-encoders, to learn more meaningful representations.

Third, while several foundation models [9], [13] in remote sensing are evaluated on classification tasks, most remote sensing applications necessitate high spatial resolution semantic maps. Despite some growth, there remains a scarcity of downstream labeled segmentation datasets for SITS, limiting the assessment of foundation models in producing meaningful spectro-spatio-temporal SITS representations.

Given the above challenges, we propose an Aligned SITS Encoder (ALISE) as a new step toward developing a foundation model for SITS. Our approach addresses the previously

I. Dumeur, S. Valero, J. Inglada are with CESBIO, Université de Toulouse, CNES/CNRS/INRAe/IRD/UT3, 31000 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

mentioned obstacles. First, the proposed network leverages the spatial, spectral, and temporal dimensions of multi-year SITS while providing aligned and fixed-dimensional representations. Next, we explore integrating an instance discrimination SSL strategy alongside a masked autoencoding task, using domain-adapted view generation for SITS. Additionally, we have constructed a novel labeled dataset to enhance the benchmark of downstream tasks for foundation model assessment. This new dataset, *RotCrop*, identifies changes that occurred between two annual SITS.

Specifically, in our SSL framework we propose a cross-reconstruction task where each view is reconstructed using the latent representation of the others. We also investigate whether integrating additional instance discrimination latent losses improves the aligned latent representations. These losses enforce invariance between the SITS views representations and decorrelate latent variables. Besides, as remote sensing applications require high spatial resolution semantic maps, ALISE representations preserve the spatial resolution of the input SITS. The quality of ALISE's representations is evaluated by exploiting them in three distinct downstream tasks: crop segmentation (PASTIS [14]), land cover segmentation (MultiSenGE [15]), and a novel crop change detection *CropRot*. On the two segmentation downstream tasks, we train a single linear layer to perform pixel level classification. We evaluate the quality of ALISE's representations by using them in 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 two aligned SITS representations from ALISE.

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 present a new multi-view SSL task specifically designed for SITS.
- We propose two novel datasets: an unlabeled multi-year European Sentinel-2 dataset and a labeled crop change detection dataset.
- We achieve state-of-the-art performance on linear probing segmentation tasks [9], [8].

Additionally, we assess ALISE pre-training under a labeled data scarcity scenario and conduct an extensive study on the influence of view generation and instance discrimination loss. Upon acceptance, we will release the code, as well as the pre-training and change detection datasets.

II. Related works

A. Masked auto-encoder on SITS

Masked auto-encoders (AE) with Transformer architecture [16] were popularized thanks to the great performance obtained by BERT [17] in NLP. The masked AE strategy involves corrupting several elements (tokens) of the input sequence and training the model to recover these corrupted tokens. For SITS, masked auto-encoders employ either a temporal masking strategy or a spatio-temporal masking strategy, depending on whether a temporal transformer or Vision Transformer (ViT) is used, respectively. On one hand, in fully temporal masking strategies, the Transformer processes pixel-level time series and is trained to recover corrupted acquisitions. The models are either fully-temporal such as SITS-BERT [7] and Presto [9], or spatio-temporal such as SITS-Former [6] and U-BARN [8]. In these two latter configurations, the Transformer backbone is merged with a spectral spatial preprocessing. The Transformer processes pixel-level time series of pseudo spectral-spatial features. On the other hand, motivated by the success of masked auto-encoders with ViT, other works such as SatMAE [13] and Prithvi [18] propose fully-attentional spatio-temporal masking for SITS. In these methods, each input image of the SITS is divided into small patches, and the pre-training involves reconstructing these masked patches. However, this spatio-temporal attention limits the input size, leading SatMAE and Prithvi to process SITS with only three temporal acquisitions [8].

These two families of methods also differ in how they handle corrupted tokens. Inspired by the original BERT [17], methods with temporal masking provide the corrupted tokens directly to the SITS encoder. In contrast, spatio-temporal methods, inspired by masked auto-encoders in vision [19], use an asymmetric encoder-decoder architecture. Here, the corrupted tokens are not fed to the encoder but are concatenated to the input representations and processed solely by the decoder using a self-attention mechanism. Additionally, recent masked auto-encoders for regular time series such as [20], [21] employ a lightweight decoder that performs cross-attention between corrupted tokens and the latent representation.

Another interesting idea from regular time series processing is the proposed masking pattern. While retrieving a masked word in NLP requires a holistic understanding of the sentence, neighboring data points in time series or image processing are highly correlated. Therefore, several studies [21], [20], [22] advocate splitting the time series into non-overlapping temporal sub-series before model processing and applying the masking strategy 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.

Consequently, unlike several previous studies on SITS [7], [6], [8], [13], our approach masks successive acquisitions, and the reconstruction task utilizes a lightweight decoder with cross-attention.

B. Instance discrimination self-supervised learning

As per [23], we consider instance discrimination as a subset of SSL, where a siamese network is trained to produce similar representations of two views of the same data. These multi-view SSL techniques can be divided into four categories: contrastive [24], clustering [25], [26], distillation [27], [28], [29] and redundancy reduction [30], [31], [32]. These approaches differ in their strategies to prevent representation collapse. First, contrastive learning [24] and its variants for segmentation tasks [33] heavily rely on negative pair sampling. Efficient negative pair sampling is challenging for SITS because pixels from different SITS may still represent the same classes. Consequently, for pixel-level SITS classification,

contrastive loss is often used in a semi-supervised framework where labels help generate relevant negative pairs [34].

Compared to contrastive, clustering or distillation based SSL frameworks, the implementation of redundancy reduction techniques [30], [31], [32] is straightforward. These strategies prevent informational collapse by decorrelating every pair of variables of the embedded latent representation. VicReg [31], in particular, does not impose the branches' symmetry or asymmetry, batch-wise and feature-wise normalization, vector quantization, or predictor module. The VicReg 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 [32], has been adjusted for downstream segmentation tasks, where the three previous losses are also calculated at the pixel level.

Lastly, these techniques require that the generated views preserve the semantic information necessary for downstream tasks. Consequently, augmentations developed for vision tasks, such as color jittering or crop, are unsuitable for SITS.

Consequently, due to its simplicity, we integrate VicReg losses alongside a cross-reconstruction task in this paper. Additionally, we propose a view generation frameworks adapted specifically for SITS.

III. Method

Our method, depicted in Figure 1, consists in the pre-training of an ALIgned SITS Encoder, ALISE, which produces aligned representations for multi-year irregular and unaligned SITS. The details of ALISE architecture are presented in the next section, followed by the description of the multi-view self-supervised learning framework.

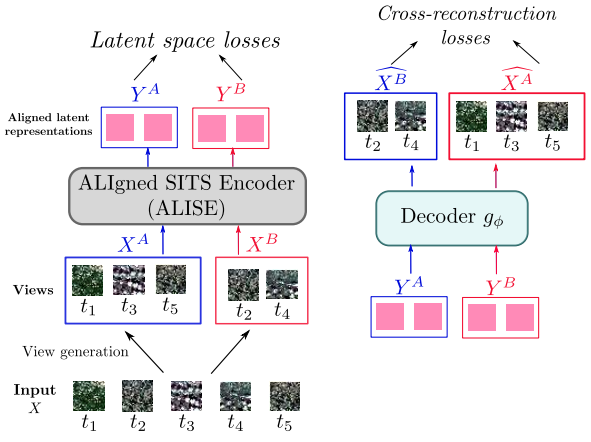

Figure 1. Description of the proposed multi-view SSL learning strategy. Given an input time series X two views are generated : X^A and X^B. Each view is processed independently by ALISE which generates the respective aligned latent representations Y^A and Y^B. A decoder g_ϕ is trained to reconstruct one view using the latent representation of the other. Additional discriminative latent space losses can be computed on the latent representation.

A. ALISE: Aligned SITS representation 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 dimension, where d_{model} and n_q are the channel and temporal sizes of the latent representation. As illustrated in Figure 2, ALISE

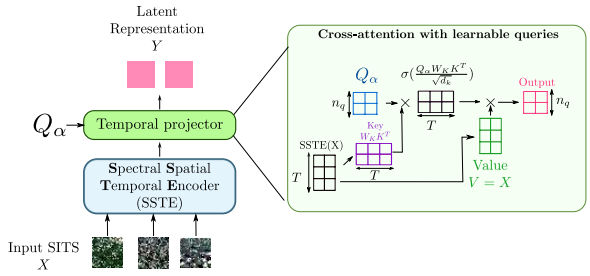

Figure 2. Overall description of ALISE architecture. The input time series X is first processed by the spectral spatial temporal encoder (SSTE). The obtained intermediate representations is then processed by a temporal projector. The temporal projector corresponds to a cross-attention mechanism with learnable queries Q_α.

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 in ALISE has been modified to process multi-year SITS. Specifically, the temporal information provided to ALISE is not the Day of Year (DoY), but δ_t, the difference in days between the image acquisition date and a given reference date (03/03/2014).

Second, to generate aligned SITS representations, a temporal projector processes the irregular and unaligned output of the SSTE, $\Psi(X)$. Specifically, based on the Perceiver I/O mechanism proposed in [35], the temporal projector consists in a temporal cross-attention mechanism between learnable queries and $\Psi(X)$ to project $\Psi(X)$ into a common temporal projection. The scaled dot product of the cross-attention is detailed in Equation (1) with Q_α the learnable queries, X the input time series, d_{model} the number of features in $\Psi(X)$ and σ the softmax function. The attention product is fully temporal, thus $Q_\alpha \in \mathbb{R}^{(n_q,d_{model})}$, $W_1 \in \mathbb{R}^{(d_{model},d_{model})}$ and $\Psi(X) \in \mathbb{R}^{(t,d_{model})}$. The temporal dimension of the latent representation Y is determined by the number n_q of learnable queries. Besides, 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 positions.

$$Y = \sigma\left(\frac{Q_\alpha W_1^T \Psi(X)^T}{\sqrt{d_{model}}}\right)\Psi(X) \quad (1)$$

B. Multi-view pre-training task

The multi-view SSL task, detailed in Figure 1, combines a cross-reconstruction loss with additional losses computed

on the embedded latent representations. As detailed in Equation (2), 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, described in the following sections.

$$L = w_{inv}L_{inv} + w_{cov}L_{cov} + w_{rec}L_{rec} \tag{2}$$

1) View generation: The view generation protocol is driven by the need to generate views that preserve semantic meaning. For SITS, we aim to create views that maintain the pixel information of the observed Earth's surface. Consequently, we construct two views, X^A and X^B, representing the same location but with different acquisition times. Specifically, first, N adjacent acquisitions are selected among an irregular and multi-year SITS. As detailed in Equation (3), this latter time series is divided along n_w non-overlapping temporal windows composed of t_w dates. Given that SITS are irregular, each subseries may represent a different temporal scale.

$$X = \bigcup_{i=0}^{n_w-1} \{X_j \mid i \times t_w \leq j < (i+1) \times t_w\} \tag{3}$$

Finally, to ensure that the two views cover nearly identical periods, every other sub-series is used to construct respectively X^A (Equation (4a)) and X^B (Equation (4b)). Therefore, t_w corresponds to the number of consecutive dates that the model is trained to reconstruct. We posit that increasing t_w complexifies the cross-reconstruction task as more variations should 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 pretraining task.

$$X^A = \bigcup_{i=0}^{\frac{n_w}{2}-1} \{X_j \mid 2 \times i \times t_w \leq j < (2 \times i + 1) \times t_w\} \tag{4a}$$

$$X^B = \bigcup_{i=0}^{\frac{n_w}{2}-1} \{X_j \mid (2 \times i + 1) \times t_w \leq j < (2 \times i + 2) \times t_w\} \tag{4b}$$

2) Latent space losses: As illustrated in Figure 1, 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 in order 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. 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 (5)) . 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.

$$L_{inv}(Z^A, Z^B) = \frac{1}{b_s \times n_q \times h \times w} \sum_{(b,n,i,j)} \|\mathbf{z}^A_{b,n,i,j} - \mathbf{z}^B_{b,n,i,j}\|_2^2 \tag{5}$$

Second, we also investigate whether the covariance loss allows learning better representations. The covariance loss decorrelates the d_{emb} different features. The total covariance loss, Equation (7), corresponds to the sum of the covariance loss 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 (6). 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 subsection VII-D, we discuss how these latent losses are related to the VicRegL [32] losses.

$$l_{cov}(Z) = \frac{1}{d_{emb}} \sum_{i \neq j} [C(Z)]^2_{i,j} \tag{6}$$

$$L_{cov} = l_{cov}(Z^A) + l_{cov}(Z^B) \tag{7}$$

3) Cross reconstruction loss: As depicted in Figure 3, the latent representations Y^A, Y^B are also employed in a cross-reconstruction task. A specific fully-temporal decoder using a

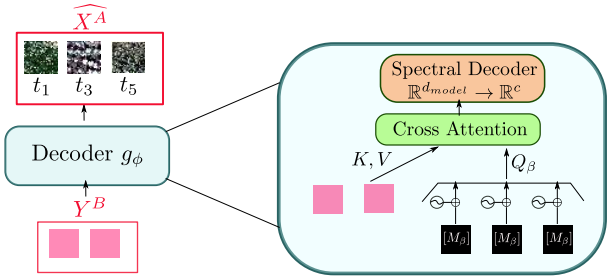

Figure 3. Description of the lightweight decoder employed for the cross-reconstruction task.

cross-attention mechanism followed by a fully-connected layer is trained to recover the latent representation of one view from the other. The fully-connected layer operates exclusively on the channel dimension of each pixel of the images, to recover the Sentinel-2 bands from the d_{model} features. As proposed in [20] 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 (8), Q_β corresponds to the sum of a shared learnable masked token $M_\beta \in \mathbb{R}^{(d_{model})}$ with the temporal positional encoding[1] of the acquisition to reconstruct. Additionally, as described in Equation (9), 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.

$$Q_\beta = [M_\beta + PE(\delta_{t_i})]_{1 \leq i \leq t_w \times n_w} \tag{8}$$

$$\text{Cross Attention}(Q_\beta, Y^k) = \sigma\left(\frac{Q_\beta W_1 W_2^T Y^{k^T}}{\sqrt{d_{model}}}\right) Y^k \tag{9}$$

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

Table I
DEFAULT HYPER-PARAMETERS FOR PRE-TRAINING ALISE. MATERIALS.

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

Finally, the quality of the reconstruction is assessed by using the classical Mean Square Error. As described in Equation (10), the reconstruction loss is the average of the reconstruction losses of each view.

$$L_{rec} = \frac{1}{2}[l_{rec}(X^A, Y^B) + l_{rec}(X^B, Y^A)] \tag{10}$$

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 (11) M_t^{valid} represents the boolean validity mask and n_t^{valid} represents the number of clear pixels.

$$l_{rec}(X^k, Y^l) = \frac{2}{n_w \times t_w} \sum_{t \in \frac{n_w \times t_w}{2}} \frac{M_t^{valid}}{n_t^{valid}} \odot ||X^k - g_\phi(Y^l)||_2^2 \tag{11}$$

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.

C. Implementation details

To pre-train ALISE, the cosine annealing scheduler with warm restarts [36] 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 dataset, 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 consecutive dates used in the view generation is random for each SITS and changes at each epoch. The pre-trainings were conducted on a single Tesla V100 GPU for 260 epochs. The pre-training value of the pre-training hyperparameters employed Table I are justified in subsection V-D.

IV. EXPERIMENTAL SETUP

First, the four Sentinel 2 L2A data-sets used in our different experiments are presented: the novel unlabeled large scale data-set used for pre-training U-BARN and the three downstream labeled data-sets (PASTIS, MultiSenGE and the novel *RotCrop*). Secondly, the implementation details of our two type of downstream tasks setup (segmentation and change detection) as well as the corresponding competitive works are described.

A. Datasets

ALISE is pre-trained on a large scale multi-year European dataset. Besides, three labeled datasets are used to assess the quality of the pre-training. The geographical distribution of the different datasets used is presented in Figure 4. For these data-sets, only the four 10 m and the six 20 m resolution bands of S2 are used. The 20m resolution bands are resampled 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 limitation, ALISE is trained to process SITS with a spatial dimension of 64 × 64. If the used dataset provides larger images, a random crop [2] (resp. center crop) is operated during training (resp. validation/testing) steps.

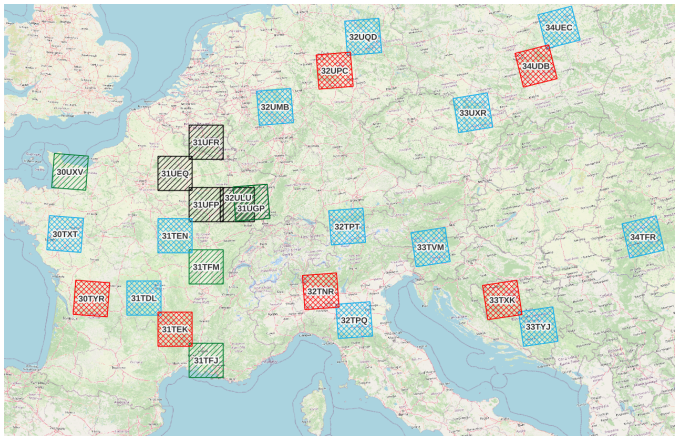

Figure 4. Geographical distributions of the different tiles composing the datasets. The unlabeled pre-training dataset is composed of multi-year SITS selected within the blue and red boxes for the training and validation dataset respectively. MultiSenGE labeled data are selected in the area delineated by the black boxes. The PASTIS dataset as well as the change detection dataset are within the green boxes.

1) European unlabeled pre-training dataset: We have built an unlabeled dataset composed of multi-year S2 SITS acquired from January 2017 to December 2020. It is divided into training and validation sets with respectively 1920 and 180 SITS of spatial dimension 64×64. The downloaded S2 SITS correspond to data processed by Sen2cor[3]. The validity mask employed in the cross-reconstruction task is built thanks to the information provided by SLC and CLM layers[4]. Specifically, as shown in Figure 4, the pre-training dataset gathers data from 18 S2 tiles. To build the training dataset, 10 smaller regions of interest (ROIs) of size 512×512 are randomly selected from each tile. The disjoint validation dataset is composed of the remaining 6 S2 tiles, from which 30 ROIs of size 128×128 are randomly drawn. The pre-trained model with the lowest loss on the validation set is selected for downstream task assessment.

2) PASTIS crop segmentation: The PASTIS dataset [14] 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 dataset contains 2433 SITS and it is divided into 5 stratified folds. In line with [8], the segmentation task is performed exclusively on known crop classes. Background and void class are ignored. The competitive method Presto requires cloud masks. As these data are not available in the original PASTIS dataset, the raw Sentinel-2 L2A and their

[2]https://pytorch.org/vision/main/generated/torchvision.transforms.RandomCrop.html
[3]https://step.esa.int/main/snap-supported-plugins/sen2cor/
[4]https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/

cloud masks were obtained from the Sentinelhub collection [5]. These S2 data also preprocessed by Sen2cor are used to assess not exclusively Presto but all models.

3) MultiSenGE land cover segmentation: MultiSenGE [15] is a dense land cover labeled dataset for eastern France in 2020. It is composed of 5 urban classes and 9 natural classes. We selected 4145 SITS with a spatial dimension of 256×256. Only images with less than 10% cloud cover were selected [15] and no cloud masks are provided. SITS are composed of 3 to 14 acquisitions. In contrast to PASTIS, MultiSenGE provides dense labels. A random split is performed to divide the dataset into training (60%), validation (16%) and test (24%). Lastly, in opposition to the two previous datasets MultiSenGE data are preprocessed with Theia and not Sen2cor.

4) RotCrop Crop change detection: This paper introduces a novel dataset for change detection. The dataset was generated using labels provided by RPGExplorer [37]. For this dataset, the following classes were selected based on the RPG (*Registre Parcellaire Graphique*)[6] 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 remote sensing data. Pixels that are not part of these crops for the two years 2019 and 2020 considered as background. Then, the label *change* is assigned to pixels that have a different label between 2019 and 2020. Each dataset sample includes Sentinel L2A SITS for 2019 and 2020, along 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 dataset. Nevertheless, once accepted, the code used to build this labeled data set will be published, 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 dataset. Lastly, the change matrix between 2019 and 2020 is detailed in subsection VII-C.

B. Evaluation Protocol

1) Downstream segmentation tasks: As detailed in Figure 5, we classify the pixel-level latent vector thanks to 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[7]. The latent representations are generated by a pre-trained ALISE whose weights are frozen in linear probing and updated during fine-tuning. We denote the fine-tuning and linear probing configurations as ALISE$^{\text{FT}}$ and ALISE$^{\text{LP}}$ respectively, while the fully supervised model is denoted 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 10 of and a decay rate of 0.05.

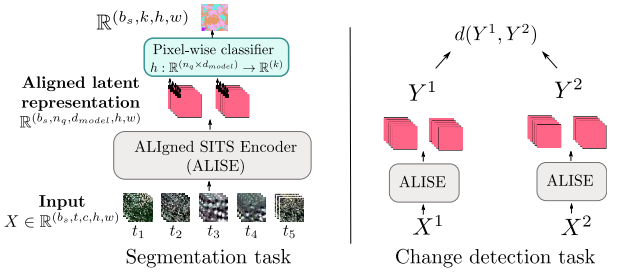

Figure 5. The two types of downstream tasks considered. Left: segmentation task framework. A single fully-connected layer projects, for each pixel of the latent representation Y, the $n_q \times d_{model}$ features into a vector of size $\mathbb{R}^k$ with k the number of classes. Right: change detection task between two SITS X^1 and X^2. The Euclidean distance is computed between the two aligned latent representations Y^1 and Y^2.

2) Change detection: As detailed in Equation (12), and illustrated in Figure 5, change detection between two SITS X^1, X^2 is performed by computing the mean square error between two representations which is averaged along the channel and pseudo-temporal dimensions.

$$d(Y^1, Y^2) = \frac{1}{n_q \times d_{model}} \sum_{n,d} ||\mathbf{y}^1_{n,d,h,w} - \mathbf{y}^2_{n,d,h,w}||_2^2 \quad (12)$$

C. Competitive methods

1) SITS segmentation concurrent works.: We compare the ALISE architecture with two fully supervised baselines, UTAE[8] [14] and U-BARN$^{\text{FS}}$[8]. The representation from the pre-trained ALISE is compared to two other masked AE SSL frameworks for the segmentation tasks.

1) Presto. In Presto, to process irregular SITS from different sensors, the time series are aligned on a common temporal grid corresponding to the least cloudy scene of each month. This sampling protocol does not ensure that each pixel of the image has a clear acquisition. Therefore, as usually operated in remote sensing, we train Presto with SITS composed of the median value of each band among the cloud-free acquisitions of each month. To exploit the latent representations provided by Presto a temporal mean is performed [9].
2) U-BARN. 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 (SC) 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 our European unlabeled dataset 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.

[5] https://www.sentinel-hub.com/
[6] https://github.com/nasaharvest/presto
[7] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
[8] https://github.com/VSainteuf/utae-paps

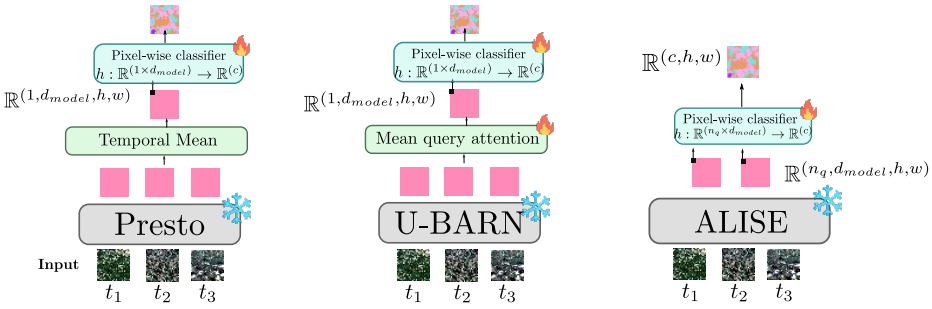

Figure 6. Comparison of Presto, ALISE and U-BARN in linear probing semantic segmentation task.

2) Change detection baseline.: Since there are no learning steps during the change detection task, we compare ALISE with a framework that also requires no learning. We propose to re-interpolate the SITS on a fixed annual common temporal grid using a linear gap-filling method. Specifically we re-interpolate the SITS valid acquisitions on a regular temporal grid with a period of 5 days. The distance map is computed between the re-interpolated raw SITS. We call this framework GF.

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 tasks (PASTIS and MultiSenGE). We also examine the effectiveness of the pre-training under a scenario of severe labeled data scarcity. Next, we provide an extensive discussion on the influence of various pre-training parameters (t_w, n_q, w_{rec}, w_{inv}, w_{cov}).

A. Segmentation tasks results

Table II presents the averaged F1 on the PASTIS and MultiSenGE segmentation tasks. Additional metrics are given in the supplementary materials

First, although this paper does not focus on the construction of a novel fully supervised framework for SITS, FS architectures achieve performances consistent with current SOTA (U-TAE). Then, we observe that, in linear probing, ALISE$^{\text{LP}}$ outperforms the previous frameworks Presto$^{\text{LP}}$ and U-BARN$^{\text{FR}}$ by respectively 41,5% and 8,8% on PASTIS dataset. Differences between ALISE and the two competitive works is illustrated in Figure 6 and further detailed below.

1) ALISE vs U-BARN: ALISE significantly outperforms U-BARN in linear probing while having a shallower classifier and a smaller latent representation. These results 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. First, as detailed in subsection III-A, ALISE corresponds to the U-BARN architecture on top of which we have 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 masked AE pre-training. Therefore, we believe the improvement in performance when freezing the pre-trained SITS encoder is largely due to the inclusion of the temporal projector in ALISE. This finding also aligns with the observation that fine-tuning results are similar between ALISE and U-BARN. Typically, pre-training is expected to have a significant impact on fine-tuning results. Consequently, we posit that the performance boost observed with ALISE is due to the pre-training of the temporal projector, which, unlike U-BARN furnish aligned representations. This aligned representations can then be used by a single fully connected layer, without performing a temporal compression as in the U-BARN's shallow classifier.

2) ALISE vs Presto: We observe that ALISE$^{\text{LP}}$ outperforms both frozen and fine-tuned configuration of ALISE, by 41.5% and 13,6%, respectively. This unexpected low performance of Presto may be due to several factors. Firstly, Presto is a lightweight fully-temporal architecture, which may not be relevant for segmentation tasks. Additionally, due to the required under-sampling protocol (Presto exploits monthly synthesis instead of all available acquisitions), it may miss important temporal variations in comparison to ALISE. Furthermore, the proposed temporal positional encoding in [8] raises questions. In the Transformer model, the positional encoding is usually added or concatenated to the input along the channel dimension. However, from our understanding of the code, in the proposed implementation, the positional encoding is concatenated along the temporal dimension. We do not fully understand the relevance of this choice for the attention mechanism in the Transformer.

B. Label scarcity scenario

To assess the model's behavior under a severe data scarcity scenario, a smaller version of the PASTIS dataset has been created. Following the approach in [8], five smaller datasets, each composed of 30 SITS, are created each PASTIS fold. Therefore, the results shown in Table III correspond to the averaged macro F1 score across 25 trials. Under severe data 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

Table II
F1 SCORE AVERAGED PER CLASS ON PASTIS AND MULTISENGE DOWNSTREAM TASKS. THE MEAN OF THE F1 SCORE ARE OBTAINED ON PASTIS 5 EXPERIMENTS. ON THE MULTISENGE DATASET, TWO TRAININGS ARE CONDUCTED WITH DIFFERENT SEED. EACH COLOR CORRESPONDS TO A PRE-TRAINING CONFIGURATION, AND THE HIGHEST SCORE WITHIN A CONFIGURATION IS UNDERLINED. AS NO CLOUD MASK ARE PROVIDED ON MULTISENGE, PRESTO CAN'T BE ASSESSED ON THIS SEGMENTATION TASK. THE NUMBER OF TRAINABLE PARAMETERS ARE ESTIMATED ON PASTIS TASK.

	Pre-training dataset	Trainable parameters	PASTIS F1	MultiSenGE F1
ALISE^{FT}	multi-year European dataset	1.1M	80.8 ± 1.6	**23.3** ± 0.2
ALISE^{FS}	✗	1.1M	79.9 ± 1.0	21.5 ± 0.4
ALISE^{LP}	multi-year European dataset	12.2K	68.2 ± 2.2	17.0 ± 0.1
PRESTO^{FT}	worlwide	404K	54.6 ± 1.9	✗
PRESTO^{LP}	worlwide	2.5K	26.7 ± 1.0	✗
U-BARN^{FT}	multi-year European dataset	1.1M	**80.9** ± 1.7	23.0 ± 0.8
U-BARN^{FS}	✗	1.1M	79.5 ± 1.3	22.7 ± 0.9
U-BARN^{FR}	multi-year European dataset	13.8K	59.4 ± 2.8	14.3 ± 0.3
U-TAE	✗	1.1M	**80.9** ± 2.4	15.3 ± 1.7

its reduced number of pre-trainable parameters compared to fully-supervised and fine-tuned approaches, ALISE^{LP} is an ideal candidate for scenarios with limited labeled data.

Table III
MACRO-AVERAGED F1 SCORE OBTAINED ON PASTIS WITH LABELED DATA SCARCITY. EACH PASTIS FOLD IS COMPOSED OF 30 LABELED SITS.

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

C. Change detection task

To evaluate the relevance of the frozen ALISE representations for change detection, we compare the Area Under the ROC Curve (AUC) score[9] of the distance map computed between the representations of SITS from two different years. The AUC on the novel crop change detection dataset is presented in Table IV. As expected, ALISE provides representations that are relevant for change detection. Besides, in contrast to the Gap-filling method, U-BARN do not require cloud mask information. Furthermore, even though we do not provide information on the annual periodicity of the SITS in the temporal encoding, ALISE can still learn the invariance of SITS between different years. For a qualitative analysis of the change detection task, see subsection VII-A.

D. Influence of t_w

Increasing t_w is assumed to have a dual effect: increasing the difficulty of the reconstruction task while creating more discrepancy between views. Therefore, we aim to assess the co-influence of the view generation protocol (controlled by t_w) and the losses weights. Therefore, we detail here the result obtained by conducting four pre-trainings with different seed for each $(t_w, w_{inv}, w_{cov}, w_{rec})$ configuration and assessing each of them on five PASTIS fold. Between all these pre-trained models, only the losses weights and t_w vary. All other hyper-parameters are fixed. We set the covariance weight

[9]https://torchmetrics.readthedocs.io/en/v0.8.2/classification/auc.html

Table IV
AREA UNDER THE ROC CURVE METRIC ON ROTCROP.

Model	AUC
ALISE	**0.91**
GF	0.88

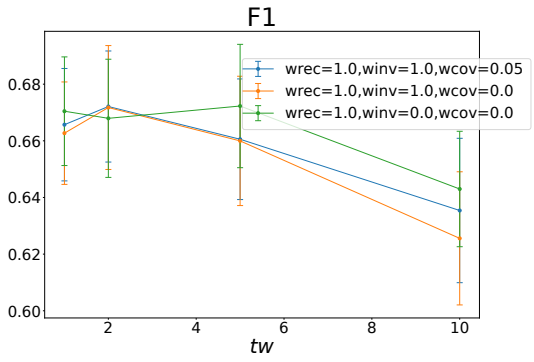

Figure 7. Segmentation task performances on PASTIS linear probing as a function of t_w. In all these experiments $n_q = 10$, 4 pre-trainings were conducted and their performances on 5 of PASTIS folds experiments were evaluated.

values to 0.05 to reproduce the balance between the invariance and covariance losses indicated in VicReg [31]. We first analyze the influence of t_w by studying the macro averaged F1 score before furnishing a more precise analysis per crop class.

1) Macro averaged F1 score: Figure 7 illustrates the linear probing performances as a function of t_w. First, we observe that with t_w greater than 2, the additional invariance latent loss significantly degrade the linear probing performances (the orange and blue curves are lower than the green one in Figure 7). We assumed that adding the invariance loss when the views are too different might constrain too much reconstruction task and prevent it from learning meaningful representations. At $t_w = 2$, there seems to be a slight improvement in the linear probing performances when employing the invariance compared to without it. Lastly, these experiments do not show any benefit from using the covariance loss 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. Experiments have been conducted with a batch size of 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, 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 crucial during pre-training. Third, more research combining a larger batch dimension with a different projector size should be performed.

Lastly, the green curve in Figure 7 depicts the influence of t_w during the sole cross-reconstruction task. We observe that when only the cross-reconstruction loss is applied, the downstream segmentation performance is impacted by t_w. With large temporal windows ($t_w = 10$), the reconstruction task may become too difficult during pre-training, preventing the model from learning meaningful 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 interval that is reconstructed. There might be some randomness even with $t_w = 1$, which still presents a complex masked auto-encoder task.

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 8. 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 with an increase in t_w. However, some crop classes such as meadow, corn, spring barley, grapevine, fruits, vegetables & flowers, potatoes, leguminous fodder, and orchard are unaffected by t_w. Apart from meadow, corn, and spring barley, we hypothesize that the lack of effect of t_w for these classes is due to the fact that they may correspond to greenhouse crops. Interestingly, the soybeans class exhibits an outlier behavior, with an increase in F1 score as t_w increases. Although we cannot explain all the results, this experiment demonstrates that the influence of pre-training conditions differ depending on the target class.

E. Impact of n_q

For practical purposes, it is relevant to reduce n_q while preserving the downstream tasks performances. Figure 9 plots the segmentation performances on the PASTIS dataset as a function of n_q. For each configuration, one pre-training was done, and the performance was assessed on one out of the five available PASTIS experiments. We observe that increasing the value of n_q improves downstream task performance. This can be attributed to two factors. Firstly, a larger value of n_q results in a larger classifier during linear probing. Secondly, it is assumed that a smaller value of n_q makes the cross-reconstruction task more challenging due to temporal compression in the temporal projector. This effect could penalize the cross-reconstruction pre-training task. The second hypothesis is reinforced by a second experiment. We observe a strong drop of performances on experiments with $w_{inv} = 0, w_{cov} = 0$ between $n_q = 10$ and $n_q = 1$. Additionally, unlike when $n_q = 10$, when operating strong temporal compression ($n_q = 1$) there is a significant improvement in segmentation performances when the invariance loss is used. Furthermore, Table V studies the impact of additional latent losses in a high-temporal compression configuration ($n_q = 1$) for two different values of t_w. While as observed in Figure 7, at $t_w = 5$ and $n_q = 10$ latent losses degrade the linear probing segmentation task, we observe conversely with $n_q = 1$ and $t_w = 5$ a 9.6% gain in F1 score when using latent losses. Nevertheless, unexpectedly, for $t_w = 2$ and $n_q = 1$, we do not obtain a similar trend. Given that our findings are based on one pre-training, these results should consequently be interpreted with caution, and further experiments could be conducted to extract a more meaningful trend. Nevertheless, this result underlines that the correct balance between the losses weights and t_w also heavily depends on n_q. Each of these parameters (t_w, n_q, w_{rec}, w_{inv}, w_{cov}) influences pre-training in its own way, and understanding the interaction between them remains a challenge. A discussion on the effect of the batch size and d_{emb} is also proposed in subsection VII-B.

Table V
STUDY OF THE INFLUENCE OF THE ADDITIONAL LATENT LOSSES WITH A STRONG TEMPORAL COMPRESSION FRAMEWORK, WITH t_w EITHER EQUALS TO 2 OR 5. FOR EACH EXPERIMENT, ONE PRE-TRAINING SESSION HAS BEEN CONDUCTED, AND RESULTS ARE COMPUTED ON ONE PASTIS EXPERIMENT. THE UNDERLINED SCORE INDICATES THE BEST F1 SCORE AT A SPECIFIC t_w VALUE.

t_w	w_{inv}	w_{cov}	F1
2	0.00	0.00	**33.5**
	1.00	0.00	22.1
	1.00	0.05	27.7
5	0.00	0.00	21.3
	1.00	0.00	30.9
	1.00	0.05	29.8

VI. CONCLUSION

This paper paves the way toward the construction of remote sensing foundation model. Our proposed method, named ALISE, leverages the spatial, spectral, and temporal dimensions and generates aligned and fixed dimensional representations of irregular and unaligned multi-year SITS. In the novel devised a multi-view SSL pre-training task, we have explored the contribution of instance discrimination SSL approaches to MAE on SITS. First, it has been demonstrated that in linear probing on crop segmentation and land cover segmentation downstream tasks ALISE outperforms competitive methods Presto [9] and U-BARN [8]. Therefore, in contrast to existing works, we provide aligned representations which are meaningful and easy-to-use. Indeed, due to their fixed-dimension, these representations could be used by traditional machine learning algorithms, which are often employed by geoscientists to address numerous earth monitoring tasks. Additionally, we have proposed a novel crop change detection downstream task, named *CropRot* to assess foundation model on SITS. Our results demonstrate that the proposed aligned SITS representations can be used for downstream change detection

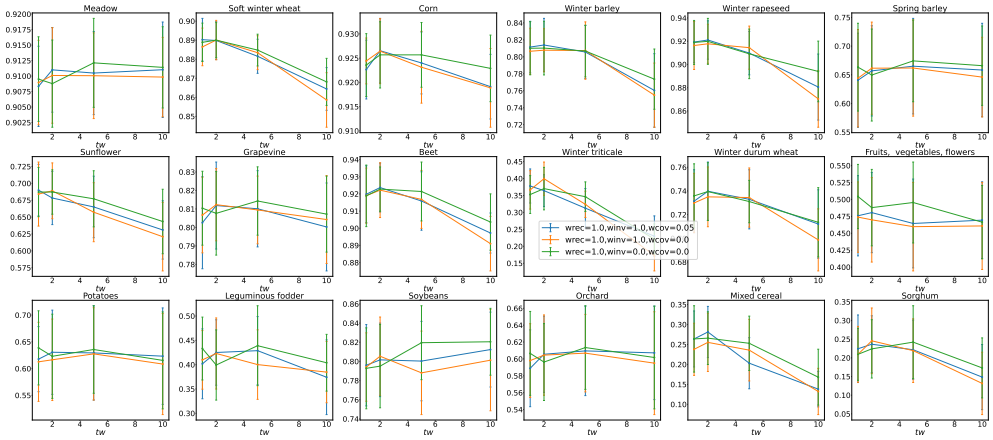

Figure 8. F1 score per class on PASTIS linear probing as a function of t_w. In all these experiments $n_q = 10$, 4 pre-trainings were conducted and their performances on 5 of PASTIS folds experiments were evaluated.

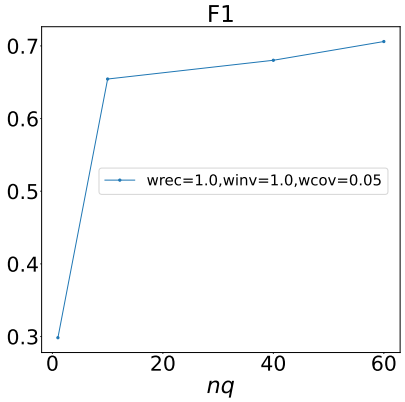

Figure 9. Segmentation task performances in linear probing configurations on the PASTIS dataset as a function of n_q. In these experiments, $t_w = 5$. For each configuration, one pre-training is conducted, and the downstream task is performed on one out of five PASTIS experiments.

tasks without the need for additional supervised training. In this exploratory application of instance discrimination task to SITS, the impact of the view generation method and the contribution of each loss were also 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, the invariance loss may or may not improve performance. This leads us to think that other view generation protocol could be investigated. Besides our experiments did not reveal a significant contribution from the covariance loss. These unexpected findings also highlights the important challenges that remain in applying ideas from wider computer vision field to the specificities of SITS (temporal dynamics, physics of the measure, etc.). Further aspects remain untouched in this article. For instance, the influence of the decoder architecture on the cross-reconstruction task should be investigated. In addition, ALISE memory consumption is quadratically related to the temporal size of the input. Therefore, lightweight architectures based on learnable queries [38], [39], [35] could be considered. Additionally, ALISE is pre-trained and evaluated solely on data from Europe. To develop a scalable method for various temporal and geographical configurations, building worldwide pre-training dataset as well as investigating the incorporation of day of year temporal encoding or thermal encoding [40] is of interest. Lastly, a major remaining challenge in developing foundation models is the processing of multi-sensor data. For example, combining Sentinel-1 data with Sentinel-2 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 loss 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] L. Zeng, B. D. Wardlow, D. Xiang, S. Hu, and D. Li, "A review of vegetation phenological metrics extraction using time-series, multispectral satellite data," *Remote Sensing of Environment*, vol. 237, p. 111511, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425719305309

[3] J. Segarra, M. L. Buchaillot, J. L. Araus, and S. C. Kefauver, "Remote sensing for precision agriculture: Sentinel-2 improved features and applications," *Agronomy*, vol. 10, no. 5, 2020. [Online]. Available: https://www.mdpi.com/2073-4395/10/5/641

[4] J. Verbesselt, R. Hyndman, A. Zeileis, and D. Culvenor, "Phenological change detection while accounting for abrupt and gradual trends in satellite image time series," *Remote Sensing of Environment*, vol. 114, no. 12, pp. 2970–2980, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425710002336

[5] Y. Ban, P. Zhang, A. Nascetti, A. R. Bevington, and M. A. Wulder, "Near real-time wildfire progression monitoring with sentinel-1 sar time series and deep learning," *Scientific Reports*, vol. 10, no. 1, p. 1322, 2020. [Online]. Available: http://dx.doi.org/10.1038/s41598-019-56967-x

[6] 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

[7] 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

[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. 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2021, pp. 4852–4861. [Online]. Available: http://dx.doi.org/10.1109/ICCV48922.2021.00483

[15] 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

[16] 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

[17] 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

[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] 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.

[20] 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.

[21] M. Cheng, Q. Liu, Z. Liu, H. Zhang, R. Zhang, and E. Chen, "Timemae: Self-supervised representations of time series with decoupled masked autoencodesr," 2023. [Online]. Available: https://arxiv.org/pdf/2303.00320.pdf

[22] 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

[23] 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.

[24] 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

[25] 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

[26] 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.

[27] 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

[28] 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

[29] 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.

[30] 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

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

[32] ——, "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

[33] 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.

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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/

[40] 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.

[41] 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

VII. Appendix

A. Change detection qualitative analysis

Given two annual irregular and unaligned SITS from 2019 and 2020, Figure 10 illustrates the change map produced. In this example, the 2020 SITS is composed of more dates during spring than the 2019 SITS. When performing change detection on SITS, different agricultural practices, meteorological events, and different acquisition dates cause important intra-class variablity. We observe in Figure 10 this intra-class variability when looking at the fields located at the center bottom of the SITS. Although the crop class of this field has not changed between 2019 and 2020, we can visually observe important differences between the 14/05/2019 and the 18/05/2020 which are supposed to be close acquisitions. Nevertheless, the distance map shown in Figure 10 does not suffer from such intra-class variability. Additionally, compared to the gap-filling method, ALISE is able to better distinguish changed crops from unchanged ones.

B. Influence of the batch size and d_{emb}

As the batch dimension interferes in the estimation of the covariance loss, we have studied the influence of the batch size during the pre-training task (see Table VI). In contrast to other experiments, the experiment with $b_s = 8$ is conducted on an NVIDIA-A100 GPU. On the linear probing performances, we see no significative improvement when increasing the batch size in the pre-training. As working with a batch size larger than 2 requires more memory resources, all other experiments were carried with $b_s = 2$. The VicReg paper suggests using a projector architecture with d_{emb} greater than d_{model}. Table VII presents the F1 score obtained given various values of d_{emb}. We observe that too small d_{emb} values degrades the linear probing performances. Nevertheless, between d_{emb} equal to 64 or 128 no important differences are found.

Table VI
PERFORMANCES ON LINEAR PROBING ON THE PASTIS SEGMENTATION DATASET FOR VARIOUS BATCH SIZES. IN THESE EXPERIMENTS, ONE PRE-TRAINING IS CONDUCTED, AND THE RESULTS ARE ASSESSED ON ONE PASTIS EXPERIMENT. THE OTHER PRE-TRAINING HYPER-PARAMETERS ARE $t_w = 2$, $w_{rec} = 1$, $w_{cov} = 0.05$, $w_{inv} = 1$, D_{EMB}=64.

b_s	F1
2	68.7
8	**69.0**

Table VII
PERFORMANCES ON LINEAR PROBING IN ONE OF THE PASTIS SEGMENTATION TASK EXPERIMENTS. ONE PRE-TRAINING SESSION IS PERFORMED AND ASSESSED ON ONE PASTIS EXPERIMENTS. WE COMPARE THE QUALITY OF THE PRE-TRAINING FOR VARIOUS VALUES OF d_{emb}. THE PRE-TRAINING HYPER-PARAMETERS ARE $t_w = 5$, $w_{rec} = 1$, $w_{inv} = 1.0$, $w_{cov} = 0.05$, $d_{model} = 64$, $n_q = 10$.

d_{emb}	F1
4	66.4
64	**68.3**
128	68.0

For both parameters b_s and d_{emb} no major impact on the linear probing segmentation task was observed. We did not conduct more extensive hyper-parameter search for d_{emb} and b_s. Nevertheless, we consider that those parameters should be explored in future works, leading maybe to an improved contribution of the covariance loss in the quality of the representations.

C. RotCrop additional information

The change matrix between 2019 and 2020 is represented by Figure 11. As expected, the rate of change depends on the class considered. We observe important rotations between cereal and corn, while grassland mostly remain unchanged.

D. Comparison with VicRegL

The proposed latent space losses are similar to those of VicRegL [32]. 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 important 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.

E. ALISE architecture

1) Projector architecture: The latent representations Y are encoded into embeddings Z using a projector. The proposed

Figure 10. Visualization of a change map obtained on the change detection dataset 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.

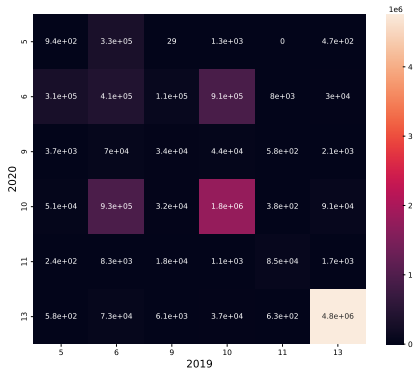

Figure 11. Change matrix between years 2019 and 2020 on the crop classes. Classes correspondance is {5: rapeseed, 6: cereal, 9 : sunflower, 10 : corn, 11 : rice, 13: grassland}

projector architecture is based on the VicReg implementation [31]. However, unlike the VicReg projector, which comprises two fully connected layers with batch normalization and ReLU, followed by a third linear layer, we employ a shallower architecture. As shown in Figure 12, 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 have been found from using a deeper or larger projector architecture for our considered downstream tasks. Further experiments should be conducted to study optimal projector architectures.

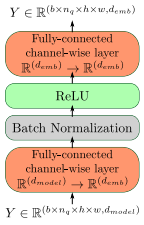

Figure 12. Description of the considered projector architecture.

2) Other architecture hyper-parameters:

1) U-BARN Table VIII and Table IX describe the architectural hyper-parameters of the spatio-spetro-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

2) 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 [41], the channels of the input embeddings are distributed among the heads.