Proceedings of TerraBytes: Towards global datasets and models for Earth Observation Workshop at the 42nd International Conference on Machine Learning 1–15

Resampling Augmentation for Time Series Contrastive Learning: Application to Remote Sensing

Antoine Saget ANTOINESAGET19@GMAIL.COM
Baptiste Lafabregue
Pierre Gançarski
ICube, University of Strasbourg, Strasbourg, France

Antoine Cornuéjols
AgroParisTech, Paris, France

Abstract

Given the abundance of unlabeled Satellite Image Time Series (SITS) and the scarcity of labeled data, contrastive self-supervised pretraining emerges as a natural tool to leverage this vast quantity of unlabeled data. However, designing effective data augmentations for contrastive learning remains challenging for time series. We introduce a novel resampling-based augmentation strategy that generates positive pairs by temporally upsampling time series and extracting disjoint subsequences while preserving temporal coverage. We validate our approach on multiple agricultural classification benchmarks using Sentinel-2 imagery, showing that it outperforms common alternatives such as jittering, resizing, and masking. Further, we achieve state-of-the-art performance on the S2-Agri100 dataset without employing spatial information or temporal encodings, surpassing more complex mask-based SSL frameworks. Our method offers a simple, yet effective, contrastive learning augmentation for remote sensing time series.

Keywords: contrastive learning, time series, remote sensing, data augmentation, cropland classification, self-supervised learning

1. Introduction

Every five days, the Sentinel-2 satellite constellation (Drusch et al., 2012; Gascon et al., 2017) captures multispectral images of Earth's entire surface at 10-meter resolution, generating an unprecedented amount of data of our planet's changing landscapes. However, a major challenge lies in the scarcity of labeled data. While large volumes of raw Satellite Image Time Series (SITS) data are available, labeling them is costly, time-consuming, expert-dependent, and often domain-specific. Consequently, only a fraction of this data is leveraged in supervised frameworks, leaving the rest unused.

Self-Supervised Learning (SSL) methods offer a promising solution to exploit these large, unlabeled SITS datasets. By learning meaningful representations from unlabeled data, SSL can achieve higher accuracy on downstream tasks with fewer labeled samples (Henaff, 2020). Generally, SSL methods can be divided into two main categories: *contrastive methods*, which rely on bringing closer together pairs of similar samples (generally created through data augmentation) in representation space, and *generative* or mask-based methods, which focus on reconstructing missing (artificially masked) information in the data.

Masking strategies have been extensively studied for time series data. Although the performance achieved is interesting, these methods can be difficult to implement for spatio-temporal data. They have a high computational cost, and the definition of tokens for transformer-based models is highly dependent on the dataset used. In addition, masking strategies have better performance when combined with contrastive methods (Cheng et al., 2023). Hence, further study of contrastive methods is needed. However, contrastive learning methods for SITS remain under-explored due to the difficulty of designing robust augmentations for time series data. Indeed, a key component of contrastive methods is the formation of positive sample pairs: two views of the same underlying instance that should be mapped close together in the representation space. In computer vision, standard data augmentations such as cropping, rotation, and color jittering have been well studied (Chen et al., 2020a,b). Yet for time series data, including SITS, designing similarly effective augmentations is less straightforward and remains an active research topic (Liu et al., 2024; Yuan et al., 2025). In this paper, we make three main contributions:

First, we introduce a novel, resampling-based, augmentation technique for time series contrastive learning. This straightforward approach generates two views of a time series by temporally upsampling the original sequence, then extracting two disjoint subsequences from it while maintaining temporal coverage across the full temporal range.

Second, we experimentally show that our resampling augmentation outperforms traditional time series augmentations (namely jittering, resizing, and masking) for contrastive learning on satellite image time series data. Furthermore, despite its simplicity and without relying on spatial information or temporal positional encodings, our approach achieves state-of-the-art performance on the S2-Agri100 dataset (Garnot et al., 2020; Yuan et al., 2022) for satellite image time series classification of agricultural fields.

Third, we investigate the impact of pretraining data distribution in SITS. We show that pretraining on unlabeled data from the target domain (S2-Agri100 dataset) rather than a different domain (SITS-Former dataset) enables a simple logistic regression to outperform state-of-the-art models trained on the SITS-Former dataset. Also, we observe a minimal performance difference between full finetuning and linear evaluation, suggesting that feature quality plays a greater role than classifier complexity, and that collecting large quantities of unlabeled data from the target domain can be as valuable as obtaining small quantities of labels.

The remainder of this paper is organized as follows: Section 2 reviews related work in contrastive learning, self-supervised learning for remote sensing, and time series augmentations. Section 3 details our resampling augmentation technique. Section 4 describes the experimental setup, including datasets, model architectures, and training protocols. Section 5 presents results: (1) comparison of contrastive frameworks, (2) label efficiency across datasets, (3) benchmarking on S2-Agri100, and (4) the effect of pretraining data distribution. Finally, Section 6 discusses limitations and future directions.

Code for models, training, evaluation, and datasets preprocessing is available at[1],[2].

2. Related Works

2.1. Contrastive Learning Frameworks

The core idea of contrastive self-supervised learning is to bring the representations of

1. https://github.com/antoinesaget/ts_ssl
2. https://github.com/antoinesaget/sits_dl_preprocess

similar samples (positive pairs) closer together in the embedding space. These positive pairs are typically created by applying different data augmentations to the same input sample. However, focusing solely on making positive pairs similar can lead to representation collapse, where the model maps all inputs to the same representation. Different frameworks address this challenge in distinct ways.

SimCLR (Chen et al., 2020a) uses in-batch negatives, treating other samples within the batch as negative examples and pushing them apart in representation space. MoCo (He et al., 2019) extends this approach with a memory bank to include more negative examples and using a momentum-updated encoder to generate consistent representations.

However, these methods can suffer from false negatives when samples from the same class are mistakenly pushed apart. BYOL (Grill et al., 2020) avoids negative examples entirely and prevents collapse using a momentum encoder and asymmetric branches (prediction head in one branch, none in the other). SimSiam (Chen and He, 2021) simplifies BYOL by showing that the momentum encoder is not necessary. VICReg (Bardes et al., 2021) directly prevents collapse through variance and covariance regularization terms in the loss. Other approaches include SwAV (Caron et al., 2020), which uses online clustering, and Barlow Twins (Zbontar et al., 2021), which maximizes the independence of features.

2.2. Self-Supervised Learning in Remote Sensing

Recent advances in Self-Supervised Learning (SSL) for remote sensing have largely focused on developing Remote Sensing Foundation Models (RSFMs). These models all contribute towards the ideal of universal representations applicable across any satellite sensors, spatial scales, geographical locations, temporal resolutions, and downstream tasks. Two main approaches have emerged: masked modeling and contrastive learning.

Masked modeling approaches, inspired by the success of masked autoencoders (MAE) (He et al., 2022) in computer vision, have been widely adopted. SatMAE (Cong et al., 2022) and Prithvi (Jakubik et al., 2023) use temporal and spectral encodings alongside traditional spatial encodings to handle the multi-modal nature of satellite data. Scale-MAE (Reed et al., 2022) contributes towards scale invariance by separately reconstructing low and high-frequency components of masked regions.

These spatio-temporal MAE approaches face an inherent computational challenge: the cubic growth in the number of tokens (width $\times$ height $\times$ time) added to the quadratic growth of self-attention in transformer models with respect to the number of tokens. This requires a trade-off, and most models prioritize spatial coverage at the expense of temporal depth (e.g., SatMAE (Cong et al., 2022) is limited to 3 timesteps). Presto (Tseng et al., 2023) takes the opposite approach by focusing exclusively on the temporal dimension without spatial context, enabling it to process much longer time series. Trained on a large-scale worldwide dataset of 20M time series combining Sentinel-1 SAR, Sentinel-2 multispectral, ERA meteorological data, and more, it treats each pixel independently and applies temporal and spectral masking on sequences of 12 timesteps covering 12 months, demonstrating competitive performance even against methods that leverage spatial information.

Contrastive learning offers an alternative approach. Seasonal Contrast (SeCo) (Manas et al., 2021) trains on large-scale Sentinel-2 imagery using three simultaneous objectives with separate projection heads from a shared embedding space: one head learns in-

variance to standard image augmentations (random cropping, color jittering, flipping), another learns invariance to seasonal changes by bringing closer images of the same location at different times, and a third combines both types of invariance. While this results in time-aware representations, the model cannot directly process time series as input. SSL4EO-S12 (Wang et al., 2022) extends this work to multi-modal data (Sentinel-1 and Sentinel-2) while evaluating various contrastive frameworks (MoCo, DINO, MAE).

Recent work has focused on improving the universality of these models. DOFA (Wang et al., 2023) generates dynamic weights to adapt to unseen sensors, trained on a diverse dataset spanning Sentinel-2 multispectral, Sentinel-1 SAR, EnMAP hyperspectral, and high-resolution aerial imagery. SkySense (Guo et al., 2024) employs multi-granularity contrastive learning to create embeddings effective at pixel, object, and image scales. Its training data combines high-resolution WorldView-3/4 imagery with temporal sequences from Sentinel-1/2. Despite being one of the largest RSFM to date (25,000 NVIDIA A100 GPU hours) with the longest time series support, SkySense notably does not incorporate temporal augmentations during contrastive learning.

2.3. Time Series Augmentations

While masked modeling approaches do not require data augmentations, contrastive methods traditionally rely on spatial augmentations like cropping, rotation, and color jittering that are not directly applicable to time series data. The diversity of time series data —from satellite observations to electrocardiograms and stock prices— makes designing universal augmentations particularly challenging due to their varying characteristics, sampling frequencies, and lengths.

Liu et al. (2024) provide a comprehensive analysis of time series augmentations for contrastive learning, evaluating eight common transformations: jittering (adding random noise), scaling (multiplying by a random factor), flipping (reversing values), permutation (shuffling segments), resizing (temporal interpolation), time masking (zeroing random timesteps), frequency masking (filtering frequency components), and time neighboring (selecting adjacent windows). They identify which augmentations are most effective for different types of time series based on properties such as seasonality, trend, and noise levels. More recently Yuan et al. (2025) extensively study data augmentations for pixelwise satellite image time series, including a different yet similar resampling-based augmentation. However, their study does not cover contrastive learning.

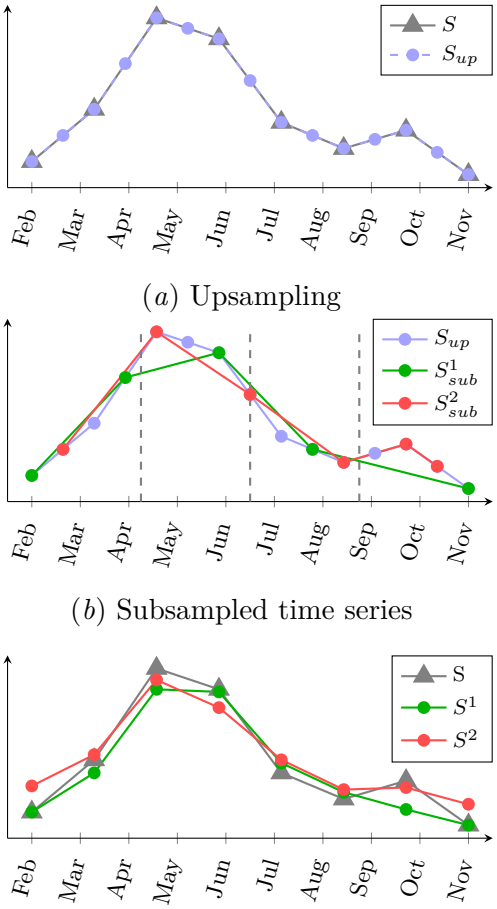
3. Resampling augmentation

Given an input time series $\boldsymbol{S} = \{s_1, \ldots, s_T\} \in \mathbb{R}^{T \times C}$, where T is the number of timesteps with index $\mathcal{X} = \{1, ..., T\}$ and C is the number of channels, we perform the temporal resampling augmentation in three steps.

First (Figure 1(*a*)), we upsample the original time series to T_{up} timesteps (typically $T_{up} = 2 \times T$) using linear interpolation:

$$\boldsymbol{S}_{up} = f_{linear}(\boldsymbol{S}) \in \mathbb{R}^{T_{up} \times C}$$

Second (Figure 1(*b*)), we sample two subsequences $\boldsymbol{S}^1_{sub}$ (resp. $\boldsymbol{S}^2_{sub}$) with indices $\mathcal{X}^1_{sub}$ (resp. $\mathcal{X}^2_{sub}$) containing T^1_{int} (resp. T^2_{int}) timesteps from $\boldsymbol{S}_{up}$ (typically $T^i_{int} = T/2$). With $T_{up} = 2 \times T$ and $T_{int} = T/2$, this means that each subsequence samples a quarter of the timesteps from the upsampled series. This empirically resulted in sufficiently different views to provide a meaningful learning signal, while remaining similar enough to preserve semantic meaning.



(*a*) Upsampling

(*b*) Subsampled time series

(*c*) Interpolated and aligned time series

Figure 1: Example visualization of the resampling augmentation process.

The sampling strategy follows two constraints:

- The subsequences use distinct timesteps: $\mathcal{X}^1_{sub} \cap \mathcal{X}^2_{sub} = \emptyset$
- Each subsequence samples uniformly at least $\lfloor T^i_{int}/4 \rfloor$ timesteps within each quarter of $\boldsymbol{S}_{up}$. With $j \in \{1, .., 4\}$:

$$|\mathcal{X}^i_{sub} \cap \{(j-1)\frac{T_{up}}{4}, ..., j\frac{T_{up}}{4}\}| \geq \lfloor \frac{T^i_{int}}{4} \rfloor$$

This structured sampling constraint ensures that both sequences maintain complete temporal coverage of the original signal and prevent large temporal gaps.

Third (see Figure 1(*c*)), we resample both sequences to match the original temporal resolution:

$$\boldsymbol{S^i} = f_{resample}(\boldsymbol{S^i_{sub}}) \in \mathbb{R}^{T \times C}$$

The function $f_{resample}$ transforms a subsequence $\boldsymbol{S^i_{sub}}$ with timestamps $\mathcal{X}^i_{sub}$ into a time series aligned with the original timestamps $\mathcal{X}$ through two steps:

1. First, we linearly rescale the timestamps $\mathcal{X}^i_{sub}$ to span the full range $[1, T]$, preserving their relative spacing. This maps the subsequence onto the same temporal range as the original series.

2. Then, since the rescaled timestamps generally don't align with the original timestamps $\mathcal{X}$, we use linear interpolation to compute values at exactly the timestamps $\mathcal{X}$, ensuring the resulting time series has both the same temporal resolution and temporal alignment as the input.

This third step is optional, as unaligned time series of different length can actually provide a stronger learning signal, however all experiments carried in this paper include this step.

This augmentation results in two distinct but similar time series that preserve the overall temporal structure, length, and alignment while introducing controlled variations, making them suitable as positive pairs for contrastive learning. This view generation process bears resemblance to the one proposed in ALISE (Dumeur et al., 2024), where views are generated by taking non-overlapping temporal blocks from the original time series.

Table 1: Datasets characteristics for self-supervised pretraining and/or supervised downstream task evaluation. G is the number of time series per sample, T is the number of timesteps per time series, C is the number of channels per timestep.

	Sample shape			Pretraining	Downstream task evaluation	
Dataset	**G**	**T**	**C**	**Nb samples**	**Nb Classes**	**Nb samples/class**
`FranceCrops`	100	60	12	∼5.8M	20	5–100
`FranceCrops CVdL`	100	60	12	–	20	5–100
`PASTIS`	100	60	10	∼85k	18	5–100
`SITS-Former`	25	24	10	∼1.6M	–	
`S2-Agri100`	25	24	10	∼120k	15	100

4. Experimental Setup

4.1. Datasets

Unlike recent works aiming to build general-purpose remote sensing foundation models (Tseng et al., 2023; Jakubik et al., 2023; Reed et al., 2022; Cong et al., 2022; Guo et al., 2024; Wang et al., 2023, 2022; Manas et al., 2021), we focus on task-specific pretraining. While our resampling augmentation could be used in a more flexible setting with varying data shapes and temporal resolutions, in this paper we focus on demonstrating its effectiveness in a more limited setting. We pretrain each model on unlabeled data that matches the characteristics (i.e. data shape, source, location) of its target downstream task. Table 1 details the pretraining and downstream datasets used in our experiments.

`FranceCrops` (Saget et al., 2024) is a large-scale Sentinel-2 time series dataset for agricultural parcel classification in metropolitan France. Each sample consists of 100 pixel time series sampled within a crop field's geometric bounds, with 60 temporally aligned timesteps spanning February-November 2022 across all 12 Sentinel-2 L2A bands. It is split into an unlabeled contrastive learning set (≈4M samples) and labeled sets (train/val/test) containing 20 selected common crop types. A separate dataset for the Centre-Val de Loire region follows the same structure with a different subset of 20 classes, enabling evaluation of geographical generalization.

`PASTIS` (Garnot and Landrieu, 2021) is a similar agricultural parcel classification dataset but at a smaller scale with 85k samples and 18 crop types. Unlike FranceCrops, raw PASTIS samples contain varying numbers of unaligned time series per parcel. We preprocess the dataset following the same procedure as FranceCrops to obtain aligned time series of equal length resulting in 100, 60 timesteps time series per sample. While PASTIS also offers versions with full imagery for spatio-temporal models and Sentinel-1 data, we only use the Sentinel-2 pixel-set version in this study.

`SITS-Former` (Yuan et al., 2022) consists of ≈1.66M unlabeled Sentinel-2 time series of 24 timesteps across 10 channels sampled from California's Central Valley during 2018-2019. Each sample is a 5×5 pixel patch extracted at regular intervals from cloud-filtered (<10%) Level-2A images. We process each 5×5 patch as a set of independent pixel time series, disregarding spatial relationships between pixels.

`S2-Agri100` (Yuan et al., 2022) is a variant of the S2-Agri dataset (Garnot et al., 2020) for crop type classification, sharing similar characteristics with the SITS-Former dataset

(5×5 pixel patches, 24 timesteps, 10 channels, cloud-filtered Sentinel-2 bands) but located in southern France spanning January-October 2017. The dataset contains ≈175k test samples from a 12,100km^2 area, with 100 samples per class in both training and validation sets. Except for the final experiment in Section 5.4, this dataset is only used for downstream evaluation.

4.2. Architectures

Figure 2 describes our architecture. Following a standard contrastive learning architecture (Chen and He, 2021), our model consists of an encoder network followed by a projection head. The encoder maps the input time series to a representation space we use for downstream tasks, while the projection head further transforms these representations for optimization of the contrastive loss. In Section 5.1 we compare performance on SimCLR (Chen et al., 2020a), MoCo (He et al., 2019), BYOL (Grill et al., 2020) and VICReg (Bardes et al., 2021) frameworks. We refer readers to SimSiam (Chen and He, 2021) for diagrams comparing different contrastive learning architectures.

We use a ResNet encoder adapted for time series (Wang et al., 2016), configured with 256 filters in the first convolutional layer. The encoder outputs 512-dimensional embeddings. In all considered datasets, each sample consists of a set of multiple time series. Therefore, we reuse the approach from Saget et al. (2024) to aggregate multiple time series embeddings from the same sample into one. This process begins by randomly selecting G (for group) pixel time series from each sample during the forward pass. Each of the G series is then processed independently by the shared ResNet blocks and global pooling, yielding G embeddings of dimension 512 per sample. An extra adaptive average pooling layer (appended after the ResNet encoder output) aggregates them into a single 512-dimensional vector per sample. This aggregation layer fuses multiple time series embeddings into one by averaging along the group dimension. Empirically, setting $G = 4$ captures intra-sample time series variability without over-smoothing discriminative features (see Saget et al. (2024)). Note that for datasets without multiple time series per sample (when $G = 1$), the aggregation layer is equivalent to the identity operation and can be discarded.

We use a 2-layer Multi-Layer Perceptron (MLP) projection head with a hidden dimension of 512 and an output dimension of 128.

Preliminary experiments with other encoders (transformer encoder, other CNN architectures, Garnot et al. (2020)) did not show significantly different results. Our hypothesis is that the performance bottleneck is currently in the contrastive learning objective or the views generation process, not in the encoder design. However, this remains to be investigated further.

4.3. Training and evaluation protocol

Pretraining: Models are trained using Stochastic Gradient Descent (SGD) with 5e-4 weight decay and 0.9 momentum. For frameworks using momentum encoders (MoCo and BYOL), the target model weights are updated with a momentum of 0.996. We employ a one-cycle learning rate policy starting at 2e-3, increasing to 5e-2 for the first 20% of training, then decreasing to 5e-5. Training runs for 50k steps with a batch size of 1024. The best checkpoint is selected based on performance on a small validation dataset, evaluated every 1k steps.

Downstream Evaluation: We evaluate the pretrained encoder in two ways:

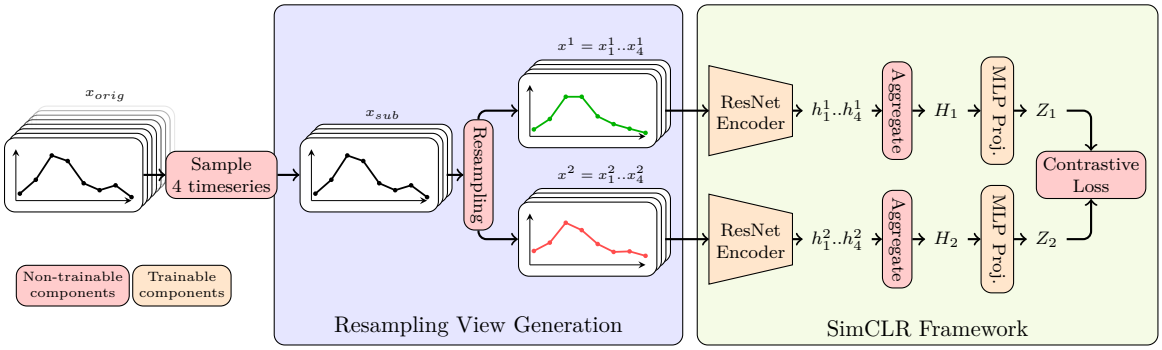


Figure 2: Self-Supervised Contrastive pretraining framework. For each sample x_{orig} in a batch, 4 time series are randomly sampled from it into x_{sub}. The resampling augmentation is then applied to create two views: x^1 and x^2. Each of the 4 time series per view per sample is individually fed to a ResNet encoder, producing 4 embeddings per view per sample into a representation space h that will be used for downstream tasks. For each view, these 4 embeddings are then aggregated by averaging into a single representation (H_1 and H_2) and further projected (Z_1 and Z_2) by an MLP on which the contrastive loss is computed. The diagram illustrates the SimCLR framework but any contrastive learning framework that relies on positive pairs can be used.

- **Linear Evaluation:** A logistic regression (max_iter: 2000, tol: 1e-5, C: 1.0) is trained on frozen encoder features.

- **Finetuning:** A 2-layer MLP with a hidden dimension of 256, ReLU activation, and 20% dropout is added to the encoder. The MLP is trained alone for 10 epochs (encoder frozen), then the full model is finetuned for 100 epochs. The MLP learning rate is 1e-3 and the encoder learning rate is 2e-5 with 5e-4 weight decay.

We train and evaluate on a single NVIDIA RTX 4090 GPU. Pretraining takes approximately 6 hours, finetuning takes 2-3 minutes, and linear evaluation takes a few seconds.

As datasets include multiple time series per sample, the final prediction of a sample is obtained through majority voting over individual time series predictions.

5. Results

5.1. Contrastive learning framework comparison

Table 2 shows our experimental results comparing the resampling augmentation with SimCLR, MoCo, BYOL, and VICReg contrastive learning frameworks on the FranceCrops dataset against jittering, resizing, and time masking. VICReg achieved the best results with 72% accuracy when using resampling augmentation. Our resampling augmentation outperforms tested augmentations across all frameworks tested.

However, we observed important differences in training stability between frameworks. Notably, representation collapse occurs more frequently with time series data compared to images. We hypothesize that this is due to the lower dimensionality of time series data, making it easier for models to find trivial solutions that map all inputs to the same representation. While VICReg

achieves the highest accuracy, we found it occasionally diverged on other datasets and required careful hyperparameter tuning. In contrast, SimCLR showed more stable training behavior despite achieving slightly lower accuracy (69%). For this reason, we use SimCLR for the remainder of our experiments (except for FranceCrops where VICReg is stable), prioritizing reliable training over marginal performance gains.

5.2. Label Efficiency

To evaluate how well our approach performs with limited labeled data, we conducted experiments across three datasets with varying numbers of labeled samples per class, from very few (5) to relatively many (100). Results in Table 3 show that our resampling augmentation consistently outperforms all other approaches across all sample sizes and datasets. The improvement is most significant with few labels (5-20 samples), achieving up to 23 percentage points over raw features on FranceCrops (44% to 67% with 5 samples). Importantly, this performance advantage is maintained even as more labeled data becomes available, showing that our method can effectively leverage additional supervision without compromising its representational power.

In contrast, simpler augmentations like jittering show limited scalability. While beneficial with few labels (5-10 samples), it performs worse than raw features with more labels, suggesting it oversimplifies representations, making them unable to capture the full complexity of the data when more supervision is available.

Time masking performs second-best. Like our resampling approach, it creates views by removing information from the original time series. This connection is particularly interesting as it bridges the gap between contrastive learning and masking methods: both rely on learning from incomplete temporal information, but while masking methods reconstruct the missing values, our approach learns to match representations of different incomplete views.

Despite many similarities with FranceCrops, we observe lower performance on PASTIS, likely due to differences in cloud filtering, a more challenging set of crop types, and a smaller pretraining dataset.

5.3. Image Time Series Task - S2-Agri100 Results

We evaluate our model's ability to generalize across different geographical regions by pretraining on SITS-Former's California data and testing on French agricultural parcels from the S2-Agri100 dataset. Our approach does not leverage any spatial information beyond the assumption that pixels within the same patch share the same label (Saget et al., 2024). Following Tseng et al. (2023), we use 100 parcels per class for training, a validation set for early stopping, and the remaining parcels for testing.

Table 4 shows the corresponding performance after finetuning models that were first pretrained in a self-supervised manner on the SITS-Former dataset. Despite our model's simpler architecture, which ignores spatial information and temporal positions, it achieves superior performance. Following Tseng et al. (2023), we also report in Table 5 results when training each model from random initialization (without pretraining) using only the S2-Agri100 training set. In this scenario, our model performs worse than SITS-Former. This highlights the importance of contrastive pretraining and shows that our bare architecture (ResNet encoder + MLP projection head) might not be the most appropriate for the task.

Table 2: Mean test accuracy (%) of logistic regressions trained on features from encoders trained with different contrastive learning frameworks. Results averaged over 20 runs with different training sets on FranceCrops with 10 labeled samples per class. All standard deviations are ≤ 1. A logistic regression on raw data achieves 52% accuracy.

	SSL Framework			
Aug.	***SimCLR***	***BYOL***	***VICReg***	***MoCo***
Jittering	49	48	53	48
Resizing	54	56	65	57
Time Masking	61	62	67	63
Resampling	**69**	**69**	**72**	**68**

5.4. Impact of Pretraining Data Distribution

In this experiment, we investigate the impact of pretraining data distribution. We compare pretraining on data from the same distribution as the target task versus pretraining on data from a different distribution. While we used the SITS-Former dataset (California) for pretraining and S2-Agri100 (France) for evaluation in our previous experiment, we both pretrain and evaluate on S2-Agri100 in this experiment.

The S2-Agri100 train split is too small (1500 samples) for pretraining, so we pretrain on 70% of the large S2-Agri100 test split in an unsupervised manner and evaluate on the remaining 30% (over 40,000 samples). We still use the standard S2-Agri100 training split for finetuning and linear evaluation.

Table 6 shows that both approaches significantly outperform all previous methods, including our model pretrained on SITS-Former. This improvement should not be attributed to our specific model architecture, similar gains might be observed if other models like Presto were pretrained in the same way. Rather, these results illustrate two general principles: first, the importance of pretraining on data from the same distribution as the target task, and second, as shown by the minimal gap between logistic regression (74.30% OA) and full finetuning (76.84% OA), most of the performance comes from the quality of the learned features rather than the complexity of the supervised classifier.

These results have important implications for practical applications. First, they highlight that matching the distribution between pretraining data and target task is crucial for optimal performance. Second, they suggest that feature learning can be effectively decoupled from classification: strong features can be learned from unlabeled data, while a simple classifier trained on these features with limited labeled data can achieve excellent performance. In other words, collecting large quantities of unlabeled data from the target domain can be as valuable as obtaining small quantities of labels.

6. Conclusion

Our approach significantly reduces the need for labeled data across all tested datasets and outperforms other traditional augmentations like jittering, masking, and resizing. With just 5 labeled samples per class, our method achieves performance comparable to training the same model on raw features with 20-50 samples per class, representing a 4-10x reduction in required labeled data.

The effectiveness of our resampling augmentation stems from its ability to create

Table 3: Mean Test Accuracy (%) of Logistic Regressions trained on features extracted by a SimCLR encoder (VICReg for FranceCrops). Results averaged over 20 runs with different training sets. All standard deviations are ≤ 1 unless specified in parentheses.

(a) FranceCrops

Aug. Strategy	Samples per Class				
	5	***10***	***20***	***50***	***100***
Raw Features	44	52	61	70	76
Jittering	49	53	59	65	69
Resizing	60	65	69	74	77
T. Masking	62	67	73	78	81
Resampling	**67**	**72**	**76**	**80**	**83**

(b) FranceCrops Centre-val de Loire

Aug. Strategy	Samples per Class				
	5	***10***	***20***	***50***	***100***
Raw Features	49	56	64	74	79
Jittering	52	58	63	68	71
Resizing	59	63	68	73	76
T. Masking	62	68	72	77	80
Resampling	**65**	**71**	**75**	**79**	**82**

(c) Pastis

Aug. Strategy	Samples per Class				
	5	***10***	***20***	***50***	***100***
Raw Features	24	28	32	37	40
Jittering	23	26	29	32	33
Resizing	26	29	33	36	37
T. Masking	37	41	45	47	48
Resampling	**38**	**42**	**46**	**49**	**50**

meaningful positive pairs while preserving temporal structure. Notably, it requires only two hyperparameters (T_{up} and T_{sub}) that we set to natural values ($T_{up} = 2 \times T$ and $T_{sub} = T/2$) and did not optimize.

However, our approach has limitations. It requires time series with a high temporal sampling rate relative to the frequency of meaningful events. This assumption holds well for remote sensing data, where Sentinel-2's 5-day revisit time captures most agricultural and land cover changes that typically occur over weeks or months. For datasets with rare or high-frequency events, the subsampling might lose critical information.

Table 4: Results on the S2-Agri100 dataset after self-supervised pretraining on SITS-Former and finetuning on S2-Agri100. We report Overall Accuracy (OA), Kappa Cohen score (K) and macro-F1 score following Tseng et al. (2023). Averages over three runs.

Model	M Params	OA	K	F1
SITS-Former	2.5	67	56	43
Presto	0.4	69	58	40
Ours	8.2	**70**	**60**	**44**

Table 5: Results on the S2-Agri100 dataset when training from random initialization (no pretraining).

Model	M Params	OA	K	F1
SITS-Former	2.5	**65**	**55**	**42**
Presto	0.4	46	35	27
Ours	8.2	62	52	40

Our experiments on S2-Agri100 showed that features learned from unlabeled data can be more important for performance than an advanced classifier. The minimal gap between logistic regression and full finetuning performance suggests that when domain-specific unlabeled data is available, strong results can be achieved with simple linear classifiers on pretrained features.

This study suggests several directions for future work:

- Evaluating the resampling augmentation in standard supervised learning settings, beyond its current use in contrastive learning.
- Exploring the applicability of our augmentation to sequential data from other domains beyond remote sensing, with

Table 6: Results when unsupervised pretraining on S2-Agri100 versus pretraining on SITS-Former illustrate the importance of learning from data with similar distribution as the target task. 🔥 indicates full finetuning, ❄️ indicates linear evaluation (logistic regression on frozen features). Average over three runs.

Model	Pretrain Data	SFT?	OA	K	F1
SITS-Former	SITS-Former	🔥	67	56	43
Presto		🔥	69	58	40
Ours		🔥	70	60	44
Ours	S2-Agri100	🔥	**77**	**68**	**49**
Ours		❄️	74	65	48

comparable and different temporal patterns.

- Extending our method to a Remote Sensing Foundation Model framework supporting variable-length sequences, non-uniform sampling, multiple modalities, Earth-wide pretraining, and diverse downstream applications beyond crop classification.

- Incorporating spatial context instead of purely temporal analysis of individual pixels or pixel-sets.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning by improving data augmentation strategies for self-supervised contrastive learning on remote sensing time series. Our resampling augmentation contributes towards foundation models that can leverage vast amounts of unlabeled satellite image time series while significantly reducing the need for labeled data. By enhancing label efficiency, this approach can lower the barrier to entry for resource-constrained practitioners and promote broader use of Earth observation for environmental monitoring, agriculture, and natural disaster response.

Acknowledgments

We thank the French National Research Agency (ANR) for funding through the ArtIC and HERELLES projects. We also express our gratitude to the European Space Agency (ESA) and the Copernicus program for making Sentinel-2 data freely accessible to the scientific community. Finally, we acknowledge the use of Claude 3.7/4 Sonnet (Anthropic), GPT-4o/o4-mini/o3 (OpenAI) and Gemini-2.5-Pro (Google) large language models to assist writing code, and edit this article for grammar checking, polishing, and formatting.

References

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A sim-

ple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*, 2023.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

Iris Dumeur, Silvia Valero, and Jordi Inglada. Paving the way toward foundation models for irregular and unaligned satellite image time series, 2024.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12325–12334, 2020.

Ferran Gascon, Catherine Bouzinac, Olivier Thépaut, Mathieu Jung, Benjamin Francesconi, Jérôme Louis, Vincent Lonjou, Bruno Lafrance, Stéphane Massera, Angélique Gaudel-Vacaresse, et al. Copernicus sentinel-2a calibration and products validation status. *Remote Sensing*, 9(6): 584, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arxiv e-prints, art. *arXiv preprint arXiv:1911.05722*, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *CoRR*, 2023.

Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Guidelines for augmentation selection in contrastive learning for time series classification. *arXiv preprint arXiv:2407.09336*, 2024.

Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.

C Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, S Candido, M UyttenDAele, and T Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. 2023 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 4065–4076, 2022.

Antoine Saget, Baptiste Lafabregue, Antoine Cornuéjols, and Pierre Gançarski. Learning from few labeled time series with segment-based self-supervised learning: application to remote-sensing. In *Proceedings of SPAICE2024: The First Joint European Space Agency/IAA Conference on AI in and for Space*, pages 275–279, 2024.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Y Wang, NAA Braham, Z Xiong, C Liu, CM Albrecht, XX Zhu, et al. Ssl4eo-s12: a large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. arxiv. *arXiv preprint arXiv:2211.07044*, 10, 2022.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: a strong baseline. corr abs/1611.06455 (2016). *arXiv preprint arXiv:1611.06455*, 2016.

Yuan Yuan, Lei Lin, Qingshan Liu, Renlong Hang, and Zeng-Guang 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*, 106:102651, 2022.

Yuan Yuan, Lei Lin, Qi Xin, Zeng-Guang Zhou, and Qingshan Liu. An empirical study on data augmentation for pixelwise satellite image time-series classification and cross-year adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 5172–5188, 2025. doi: 10.1109/JSTARS.2025.3527017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.