Using Barlow Twins to Create Representations From Cloud-Corrupted Remote Sensing Time Series

Madeline C Lisaius, *Student Member, IEEE*, Andrew Blake, *Fellow, IEEE*, Srinivasan Keshav, *Fellow, IEEE*, and Clement Atzberger

***Abstract*—Satellite-based monitoring is a key tool for supporting global food security and natural resource management but is challenged by cloud corruption and lack of labeled training data. To address these issues, self-supervised learning (SSL) techniques have been developed that first learn representations from almost limitless available unlabeled data, before using labeled samples for a specific downstream task. As the learned representations detect, integrate, and compress information in the dataset in a fully unsupervised manner, the downstream tasks require only small labeled datasets. In this study, we present spectral–temporal Barlow Twins (ST-BT), a new pixelwise SSL architecture that generates useful representations designed to be invariant to extensive cloudiness. We demonstrate that ST-BT representations enable stable and high F1 scores on the downstream task of crop classification even with cloud cover reaching 50% of available dates and using only a few labeled samples. The ST-BT representations achieve maximum F1 scores of 0.94 and 0.90 on the two benchmark classification datasets used. These results indicate that ST-BT can create useful representations of pixelwise multispectral Sentinel-2 timeseries despite cloud corruption.**

***Index Terms*—Crops, remote sensing, self-supervised learning, time series analysis.**

I. Introduction

SATELLITE-BASED remote sensing offers low-cost, global land surface data that allow monitoring of agricultural landscapes and natural resources with spatial resolution ranging from decimeters to kilometers [1]. In the agricultural context, remote sensing, with its high revisit frequency, can inform important agricultural decisions, including crop yield estimates, crop area measurements, and disease and pest tracking, which are at the foundation of global food security.

The huge potential of remote sensing for agricultural applications is still underutilized [2] largely for two reasons. First, the presence of clouds interrupts the otherwise regular sensing of land surface spectral signatures, "corrupting" the data. Many existing classification algorithms are intolerant of heterogeneous temporal sampling, making them unusable for cloud-corrupted data without complex preprocessing steps for corruption management. Second, labeling remotely sensed data can be costly and sparse, which impedes supervised machine learning approaches.

The main existing approaches for handling corrupted data are 1) modifying the data to minimize the impact of corruption and 2) building models invariant to corruption. Data modification is more common and includes hand-selecting data [3], applying gap-filling [1], [4], [5], fusing sensors [6], and deriving temporal metrics [7]. Corruption invariance is less common and was historically solved using curve fitting approaches [8]. In newer approaches, corruption invariance is handled using ML approaches [9], [10], [11].

As demonstrated in the fields of computer vision and speech recognition, self-supervised learning (SSL) has the potential to address data corruption, while the subsequent downstream (e.g., classification) tasks require fewer labels as compared to models trained directly on the original data [12], [13]. SSL methods work by extracting meaningful representations of input data by optimizing a surrogate objective [14]. The extracted representations are leveraged in downstream tasks using only a small amount of labeled data. Seen through a compressed sensing and redundancy reduction framework, SSL aims to create broadly useful, informationally dense representations. Versions of SSL for image data have been used since the 1990s [15] but have only gained popularity in recent years. Recent work applying SSL to remote sensing contexts has found that SSL methods can outperform supervised models [16].

Different approaches, such as RankMe [17], have been developed to assess the quality of the derived representations. In the remote sensing context, a good candidate downstream task to help understand the usefulness of SSL-derived representations is crop-type classification. For one, precise crop-type classifications are essential for agricultural statistics, production forecasts, and food security issues. Second, crops are often grown in cloudy areas and have unstable crop-type specific spectral–temporal fingerprints impeding their identification across seasons and ecoregions. This traditionally requires intensive and costly labeling procedures, which SSL approaches could partly alleviate.

As we show in Section II, proposed SSL methods for time series analysis are complex, do not use the full range of spectral information, or use spatial contextual data (making them unsuitable for small fields). To address this research gap, we demonstrate the use of a novel self-supervised machine

Manuscript received 16 February 2024; revised 7 May 2024 and 1 July 2024; accepted 7 July 2024. Date of publication 10 July 2024; date of current version 5 August 2024. This work was supported in part by UK Research and Innovation (UKRI) and in part by Mantle Labs. *(Corresponding author: Madeline C Lisaius.)*

Madeline C Lisaius and Srinivasan Keshav are with the University of Cambridge, CB3 0FD Cambridge, U.K. (e-mail: mcl66@cam.ac.uk; sk818@cam.ac.uk).

Andrew Blake is with the Clare Hall, University of Cambridge, CB3 9AL Cambridge, U.K., and also with Mantle Labs, W1J 5RL London, U.K.

Clement Atzberger is with Mantle Labs, W1J 5RL London, U.K.

Digital Object Identifier 10.1109/JSTARS.2024.3426044

learning approach for learning information-rich representations: spectral–temporal Barlow Twins (ST-BT). We demonstrate the quality of the derived representations in the context of crop type classification. Specifically, we show that on two publicly available remote sensing datasets, our method can reach an F1 score of up to 0.94, and that this F1 score is stable even with cloudiness reaching 50% of available dates.

II. Related Work

There is little literature on SSL applied to crop classification using spectral time series data [18].

Yuan and Lin [19] proposed an SSL method (SITS-BERT) for satellite time series representation generation for crop-type classification based on bidirectional encoder representations from a BERT architecture [20]. They prepare spectral–temporal data using Sentinel-2 remote sensing images from the Central Valley of California, United States, but restrict analysis to cloud-free images. SITS-BERT creates representations by learning to fill artificially noised time series observations. The pretrained SITS-BERT plus a shallow neural network (NN) are afterward finetuned using a small, labeled dataset to leverage representations for landcover classification. The greatest limitation of this work is that, while the authors report an up to 3.5% boost in representation performance with their finetuning compared to without finetuning, the published code does not show a similar increase.

Yuan et al. [21] also proposed SITS-Former, a method for learning useful representations from patch (spatial–spectral–temporal) Sentinel-2 time series data. As in SITS-BERT [19], only cloud-free images are used. SITS-Former learns to create representations by filling artificially masked subpatches. As in Yuan and Lin's work [19], the authors assumed that the learning task of identifying masked or artificially corrupted data is sufficient for building representations that capture the spatial–spectral–temporal structure of data. The biggest challenge with that work is that a 5×5 pixel patch is required to learn from or classify each pixel making it unsuitable for small fields. In addition, it remains unclear to what extent this method is successful for patches containing several (crop-type) classes, as the authors used for evaluation only patches with at least 50% of pixels matching the class of the center target pixel.

Wang et al. [22] proposed a crop classification oriented SSL framework. Pixelwise time series Sentinel-2 imagery in red, green, and blue, and NIR spectral bands are formed into 3-D tensors. These tensors are passed into the Sim-SCAN SSL model, a combination of SimCLR [23] and semantic clustering [24]. Sim-SCAN works by learning to create useful representations from clustering and denoising positive and negative sample pairs as a pretext task. The authors find that this method performs similarly or slightly better than ResNet18, although it takes three times longer to run per sample. The greatest drawbacks to this approach are its run time, the use of limited spectral bands, and the need for negative samples.

Zheng et al. [25] presented a model for remote sensing data for use in a diverse array of tasks, named SkySense. It incorporates SAR and multispectral data across spectral, spatial, and temporal dimensions. Cloudy images are omitted from the multispectral datasets used using sensor-provided data. They propose a learning task of multigranularity contrastive learning, which allows the model to generate useful representations. After pretraining, SkySense is evaluated as a foundation model on a variety of downstream tasks with finetuning, including crop-type classification. The main drawback of SkySense is its architectural complexity.

Our review of prior work suggests that SSL approaches for EO time series analysis are still in its infancy, and that the challenge of crop classification can potentially benefit from useful representations generated with SSL.

III. Methods and Data

A. Barlow Twins (BT)

In this work, we propose a novel SSL approach. Our approach is based on BT ([26]), an SSL architecture that generates representations from multispectral time series (see Fig. 1). BT learns by passing two augmentations of spectral–temporal data through an encoder, which produces two representations, and then a projector, which creates embeddings from the representations. The augmentations are sparse random temporal samples of the spectral signatures of a given pixel location but excluding cloudy observations. The loss function pushes the model to generate representations of the two distorted inputs more closely in the multidimensional space, building invariance to the distortions in the augmentation. The loss also encourages redundancy reduction.

To leverage the rich spectral and temporal information provided by modern satellite sensors, such as Sentinel-2, and to avoid spatial convolutions, we use 2-D arrays of spectral signatures for all datetimes as inputs to the BT instead of the classical 3-band image data used in the original BT. We label this data structure as "d-pixel" (see Fig. 1). In a d-pixel, data are organized as an array with spectral bands as columns and datetimes d as rows.

Given the relative simplicity of the 2-D d-pixel compared to a RGB image, we use an architecturally simpler encoder; the original BT uses a ResNet-50 network. We use a four-layer fully connected NN with rectified linear units following the first three layers (see Table II).

B. Data Augmentations

To build ST-BT invariance to cloudy dates when using the d-pixel structure, sparse random temporal sampling is used: two sets of randomly selected spectral observations are sampled from the available cloud-free datetimes in the original d-pixel, each representing a sparse subset of the original; this corresponds to selecting a random subset of rows of the d-pixel (see Fig. 1). These two subsets of cloud-free spectral temporal data create augmentations of the original d-pixel, forcing the algorithm to learn the intrinsic spectral–temporal structure of the observed pixel. Through the use of sparse temporal sampling as the augmentation for d-pixels, invariance to irregular datetime

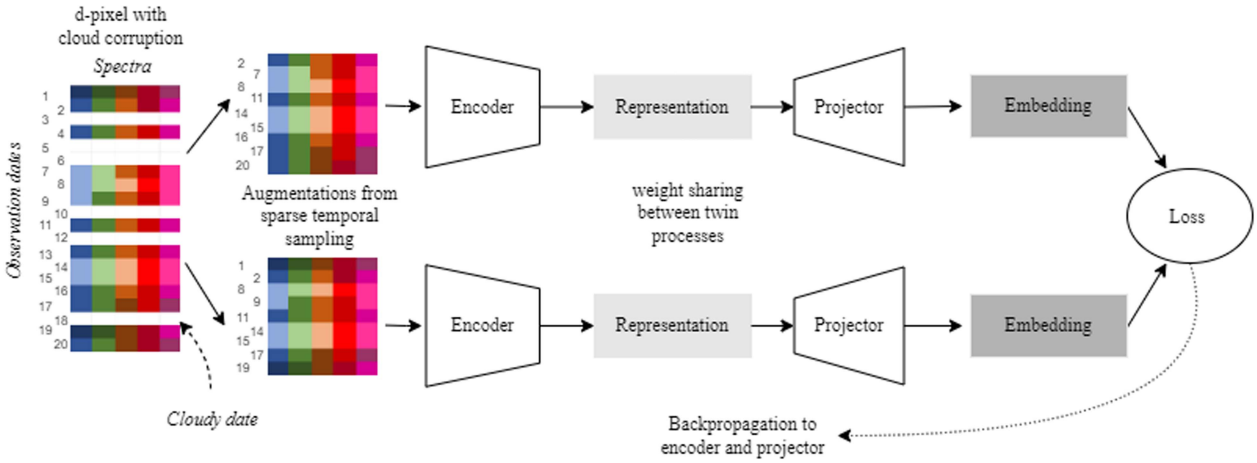

Fig. 1. ST-BT architecture. The corrupted d-pixel and two extracted augmentations are on the left, with datetimes in rows and spectral bands in columns. ST-BT generates representations from d-pixels instead of images. (Visualization adapted from Zbontar et al. [26].)

TABLE I
DATASET DETAILS

	D1: Pelletier et al. (2019) [27]	D2: Yuan et al. (2022) [21]
Region	Victoria, Australia	Central Valley, California, USA
Sensor	Sentinel-2	Sentinel-2
Number of samples	1600 labeled	1.66 million unlabeled, 145k labeled
Number of classes	8	13
Cloud Treatment	Cloudy days gap-filled	Cloudy days masked
Number of cloud-free observations	73	Between 7 and 43
SSL training	Random 50% of labeled data for SSL training (labels discarded)	Unlabeled data
Downstream crop classification	7.5% of labeled data for downstream training, and 42.5% for evaluation	1%–20% labeled data used for downstream task training, and remaining for evaluation height

availability is created, essential for the representation learning from cloud corrupted time series.

C. ST-BT Loss

ST-BT pushes together the two embeddings through the loss term, which is directly adopted from the original BT and defined as

$$\text{Loss} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{1}$$

where C is the cross-correlation matrix calculated between embeddings. An example of the C_{ij} calculation is

$$C_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}}. \tag{2}$$

In each batch b, as indexed by z_b, each pair of embedding dimensions is indexed by i and j.

The first term of the loss function is considered the "invariance term" and is calculated between corresponding dimensions of the two embeddings. This term is minimized when the sum of the cross-correlation for each corresponding embedding dimension approaches one, pushing the model to generate embeddings that are more similar in each epoch. Meanwhile, the second term of the loss is considered the "redundancy reduction term," and is calculated between noncorresponding dimensions of the embedding pair, for all pairs in the batch. This term is minimized when the sum of all cross-correlations between noncorresponding embedding dimensions approaches zero, which encourages the model to reduce redundancy in the embeddings. The redundancy reduction term is weighted with the λ parameter. Once the model is trained, representations are expected to cluster by similarities in a multidimensional space, and compresses the core spectral time series into a meaning-dense form. It can therefore be seen as an implementation of compressive sensing and redundancy reduction for diverse downstream applications.

IV. RESULTS

A. Datasets

We test ST-BT representation creating using two publicly available datasets: D1 and D2 (see Table I). In both cases, ST-BT is first trained without any labels. Only after completion of the SSL training, labels are used to assess the quality of the representations within a crop-type classification task.

The first dataset (D1) is from Pelletier et al. [27]. It contains 1600 pixelwise, labeled, and class-balanced, 10-band Sentinel-2 spectra, across 73 dates in Australia for 2017–2018. One spectral signature is recorded every five days, and eight classes are distinguished. Cloudy dates are gap-filled by using a linear temporal interpolation [27]. This dataset is ideal for experimenting with

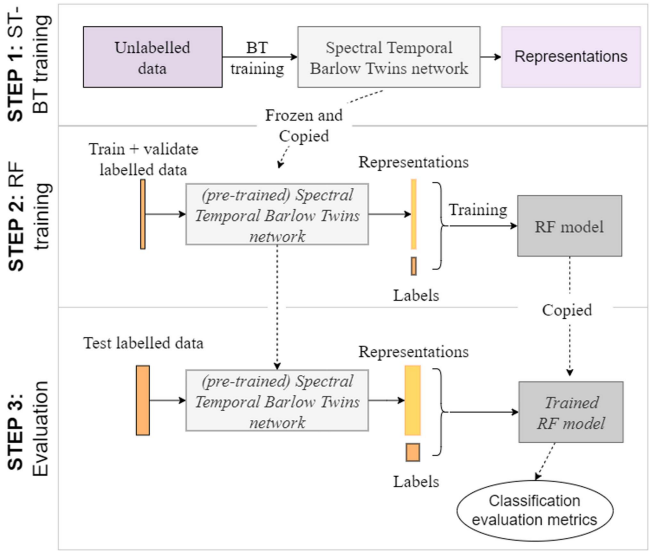

Fig. 2. Downstream classification pipeline. Other results may be calculated from the unlabeled representations.

model architectures and for sensitivity analysis with respect to (artificial) cloud corruption.

The second dataset (D2) is from Yuan et al. [21]. The pixelwise dataset comprises 1.66 million unlabeled samples and 145 000 labeled samples, each of a single geographic point with ten Sentinel-2 spectral bands. The time series extends across one year in the Central Valley of California (2018–2019) and includes 13 classes. Cloudy data are masked. This dataset is appropriate as a larger, more challenging dataset for benchmarking work.

B. Evaluation Details

SSL approaches are trained on unlabeled data, which are abundant. Task-oriented labeled dataset are required only for evaluation tasks. We first train ST-BT using unlabeled data (see Fig. 2). The data used for this step are specified in Table I. Using the two benchmark datasets (D1 and D2), we investigate different use cases and conditions.

1) Impact of cloudiness on classification accuracy (D1).
2) Impact of the number of sparse temporal samples per augmentation on classification F1 score (D1).
3) Stability of the derived representations for different numbers of sparse temporal samples (D1).
4) Suitability of representations for supervised classification and comparison against composite-based baseline classifier (D1) and SITS-BERT (D2).
5) Benefits in using more labeled samples (D2).

As a baseline for comparison, a random forest (RF) is trained and evaluated with classical seasonal composites [4], [28]. Here, three-month season windows are composited, pixelwise and bandwise. For occasional cases where no valid (cloud-free) observations are available within a season, the neighboring composited seasonal values are used, prioritizing the preceding season, and if this is not feasible, the subsequent season. The four seasonal composites for each pixel are flattened and used to train the random forest for classifying unseen data. The validation set is used to optimize the RF hyperparameter "number of trees."

TABLE II
SETTINGS AND HYPERPARAMETERS ASSUMED UNLESS OTHERWISE NOTED

	Setting or Hyperparameter	Value(s)
ST-BT	No. sparse temp. samples/augmentation	15
	No. days emulated cloud cover	0
	No. pairs of augmentations/d-pixel	15
	Batch size	128
	Encoder architecture	3-layer fully connected NN
	Encoder output dimension	128
	Projector architecture	3-layer fully connected NN
	Projector output dimension	128
	Lambda (loss function)	5e-3
	Trainable Parameters	3.3 million
	Activation function	ReLU
	Learning rate	1e-4
	No. epochs for SSL training	300
RF (both)	% Labeled samp. for training	10
	No. trees considered	[100, 200, 300, 400, 500]
	Criterion	Gini impurity
	RF depth	Unconstrained

For cases where we emulate cloudiness impacts on ST-BT, we mask a fixed number of randomly selected dates by removing random rows from the d-pixels. We then generate representations of the labeled d-pixels using the trained ST-BT model. To assess the importance of the number of sparse temporal samples, we create two datasets with zero and 50% cloudiness, and calculate the F-score using 5, 10, 15, and 25 sparse temporal samples, respectively.

Parameters for ST-BT, downstream methods, and baseline are provided in Table II. Parameters for the downstream RF and the baseline RF are the same.

To evaluate the representation stability, we randomly sample unlabeled data with different numbers of sparse temporal samples (5 and 25) and calculate the coefficient of variation (CV) over five augmentations for 0% and 50% cloudiness. In evaluating the representations using the downstream task of crop classification, we divide the labeled representations into train–test–validate sets and use an RF classifier for the supervised training (parameters for the downstream RF and the baseline RF are the same). Using this off-the-shelf method allows us to evaluate the usefulness of the representations without modification, and directly compare to the RF baseline. The randomly sampled training set is used for RF training and the validation set is used to select the number of trees. Classifications of each d-pixel are based on the majority vote across the classifications of all representations. We evaluate performance using the standard F1, balanced accuracy and overall accuracy metrics on the test set. Parameters for ST-BT, downstream methods, and baseline are provided in Table II.

C. Performance on Dataset D1

All results presented are for ST-BT pretrained on 50% of D1 data (labels removed).

1) Impact of Cloudiness: We find that ST-BT representations lead to a consistently higher F1 score than the baseline composites even as the number of emulated cloudy days increases from 0% to 50% (see Fig. 3). The Kolmogorov–Smirnov (KS) test

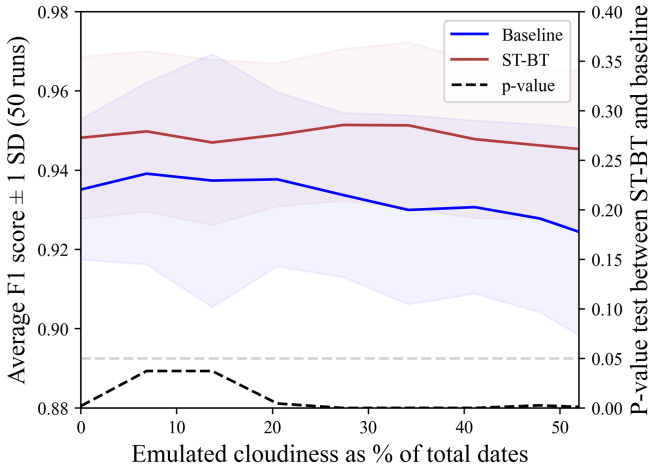

Fig. 3. ST-BT consistently outperforms the baseline even with 50% cloudy days. ST-BT performance on D1 with RF used for downstream classification. F1 scores ±1 standard deviation (SD) from 50 runs are shown with varying percentages of cloudy days. The broken black line shows the *p*-value of the KS significance test between ST-BT and the baseline F1 scores (right *y*-axis). The broken gray line shows the significance bound of 0.05.

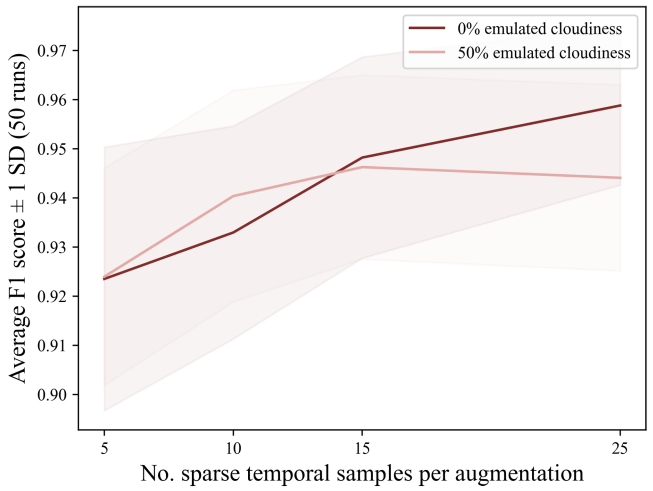

Fig. 4. ST-BT has higher F1 scores with increasing number of sparse temporal samples. This effect saturates when the number of samples approaches the number of available noncloudy dates. The average ST-BT F1 score ±1 SD for 50 runs is shown for two cases: No cloudiness and 50% emulated cloudiness, as the number of sparse samples used per augmentation varies.

confirms the statistical significance between the two classifications; the KS test offers a nonparametric comparison of the cumulative distributions of two datasets with the null hypothesis that both came from the same distribution. Notably, the ST-BT representations achieve a maximum F1 score of 0.94, given that the baseline has an already high F1 score of 0.93.

2) Impact of Number of Sparse Temporal Samples: For both cloud-free conditions and 50% cloudiness (see Fig. 4), the performance of ST-BT representations increases with increasing number of temporal samples used per augmentation. The F1 scores for ST-BT representations derived from data with 0% and 50% cloudiness, respectively, are not statistically significantly different (using the KS test), demonstrating that ST-BT is stable

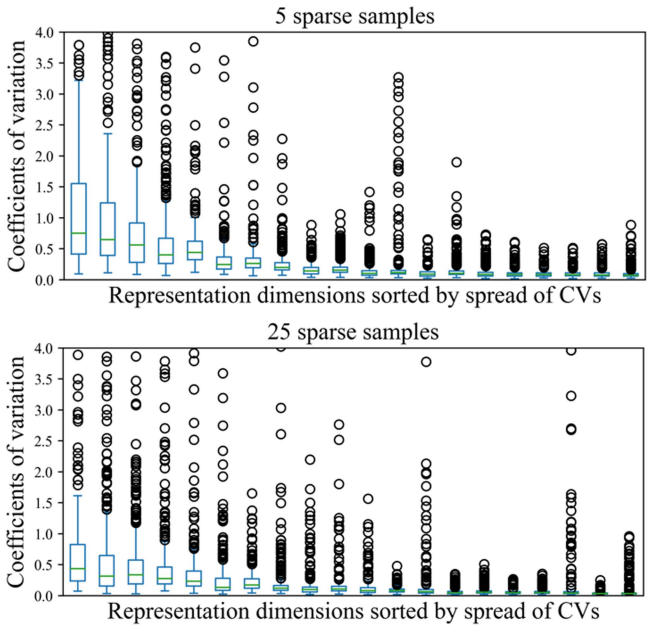

Fig. 5. Stability of representations increases when more sparse temporal samples are used per augmentation. Boxplots of the coefficient of variation for each dimension of 20-dimension representations from all d-pixels of D1 ordered by magnitude, comparing representations from five sparse temporal samples per augmentation (top) with 25 sparse temporal samples per augmentation (bottom).

even under highly cloudy conditions. ST-BT representations stop benefiting from increasing number of sparse temporal samples once the number of sparse temporal samples used in each augmentation approaches the number of available (noncloudy) dates. When the number of sparse temporal samples approaches the number of cloud free observations, the augmentations are not sufficiently different, inhibiting the learning process.

3) Representation Stability: With a greater number of sparse temporal samples per augmentation—and if cloudiness conditions permit—the stability of the output representations increases. We quantify representation stability by the coefficient of variation (CV) calculated for each of the 20 representations per d-pixel in D1, for five runs of ST-BT. The CV permits quantification and comparison of the stability of representations independent of the downstream task. The CV of each representation dimension for all d-pixels is visualized in Fig. 5.

With a larger number of sparse temporal samples used per augmentation, the representations across all dimensions have less dispersion on average, and are thus more stable. Given that intra-d-pixel variability remains and that the number of sparse temporal samples is limited by the overall cloudiness conditions in a study area, this motivates aggregating more than one pair of augmentations per d-pixel. Overall, our results for D1 demonstrate that the ST-BT representations are effective in crop-type classification despite using cloud corrupted data for training.

D. Performance on Dataset D2

D2 is a much more challenging dataset, with more class overlap. The dataset allows us to study the impact of the quantity

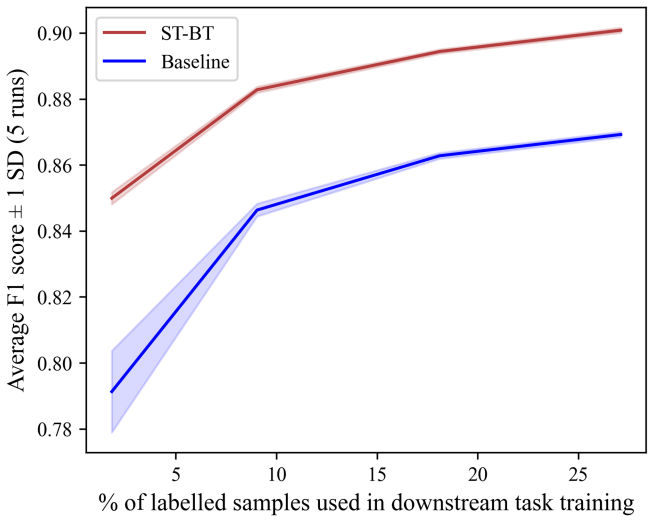

Fig. 6. ST-BT always has higher F1 scores than the baseline, regardless of downstream labeled training samples used on D2. F1 scores ±1 SD across five runs with increasing number of labeled samples used for downstream task training, and the greatest improvement happens as the % of labeled samples used increases to ∼10%.

TABLE III
OVERALL ACCURACY BY % OF LABELED TRAINING SAMPLES USED FOR DOWNSTREAM TRAINING (D2)

	% of Labeled Samples			
	∼2%	∼9%	∼18%	∼27%
Baseline	0.789	0.854	0.872	0.882
ST-BT	0.855	**0.884**	**0.896**	**0.902**
SITS-BERT (reported)	**0.879**	—	—	—

of labeled samples on the F1 score. All results presented are for ST-BT pretrained on unlabeled D2 data. The downstream tasks splits the labeled data into train, test, and validation sets.

1) Impact of Number of Labeled Samples and Downstream Method: We find that the F1 score of ST-BT representations increases as the number of training samples used for the downstream training increases, and remains statistically significantly above the baseline composites (see Fig. 6).

With about 9% of the labeled data used for downstream task training, ST-BT outperforms the reported performance of SITS-BERT [19], which is the closest related work (see Table III). We are unable to evaluate SITS-BERT performance when trained with more than the reported 2600 labeled samples (2% of labeled data) because their work is unreproducible.

V. DISCUSSION AND CONCLUSION

Our work addresses the potential usefulness of representations derived from multispectral Earth Observation timeseries using the example downstream task of crop type classification. The research specifically takes into account the presence of clouds and the lack of adequate quantities of labeled data.

Using a relatively simple architecture, ST-BT is a promising SSL approach to creating a foundation model for environmental remote sensing. Our algorithm is relatively insensitive to extensive cloudiness and works on pixel level without any spatial convolution. When the derived representations are used for classification purposes, the method requires fewer labels compared to the baseline RF classifier based on seasonal composites to reach the same classification accuracy.

We show that ST-BT offers a best-in-class approach to generating representations for crop-type classification with cloud-corrupted time series data. Using a sufficient number of sparse temporal samples, within the limits of available cloud-free data, it creates stable representations and has high F1 scores even with cloud cover reaching 50% of available dates, achieving a maximum F1 score of 0.94 on D1 and 0.90 on D2.

Our work can be extended in several ways. First, a fine-tuning step could further boost the performance of the representations on downstream classification. In this article, we have frozen ST-BT after pretraining and only optimized the downstream classifier in downstream task training. Second, we only used a single augmentation type—sparse temporal sampling—for ST-BT to learn invariance to missing data due to cloud corruption. Other augmentations could be tested to explore the extent to which the model might learn additional types of invariance and further compress information implicit in the data. Third, future work could explore the impact of leaving cloud corruption in the data entirely to understand the extent to which ST-BT can learn to identify and ignore corruption. Fourth, ST-BT, as evident from its naming of *spectral–temporal*, does not consider spatial contextual data. The impact of spatial data can be analyzed for model performance, especially when considering varying field sizes. Finally, additional methods for evaluating SSL separately from downstream task performance can be explored.

REFERENCES

[1] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, Jan. 2020, Art. no. 111402.
[2] C. Atzberger, "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs," *Remote Sens.*, vol. 5, no. 2, pp. 949–981, Feb. 2013.
[3] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042609.
[4] J. Inglada et al., "Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery," *Remote Sens.*, vol. 7, no. 9, pp. 12356–12379, Sep. 2015.
[5] Z. Zhu and C. E. Woodcock, "Continuous change detection and classification of land cover using all available Landsat data," *Remote Sens. Environ.*, vol. 144, pp. 152–171, Mar. 2014.
[6] L. Martinez-Ferrer et al., "Quantifying uncertainty in high resolution biophysical variable retrieval with machine learning," *Remote Sens. Environ.*, vol. 280, Oct. 2022, Art. no. 113199.
[7] H. Müller, P. Rufin, P. Griffiths, A. J. B. Siqueira, and P. Hostert, "Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian Savanna landscape," *Remote Sens. Environ.*, vol. 156, pp. 490–499, Jan. 2015.
[8] G. D. Badhwar, "Classification of corn and soybeans using multitemporal thematic mapper data," *Remote Sens. Environ.*, vol. 16, no. 2, pp. 175–181, Oct. 1984.
[9] C. S. Agastya, S. Ghebremusse, I. Anderson, C. Reed, H. Vahabi, and A. Todeschini, "Self-supervised contrastive learning for irrigation detection in satellite imagery," in *Proc. Tackling Climate Change with Mach. Learn. Workshop, ICML*, 2021, *arXiv:2108.05484*.
[10] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[11] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefevre, and M. Korner, "BreizhCrops: A time series dataset for crop type mapping," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2-2020, pp. 1545–1551, May 2020.
[12] R. Rani, J. Sahoo, S. Bellamkonda, S. Kumar, and S. K. Pippal, "Role of artificial intelligence in agriculture: An analysis and advancements with focus on plant diseases," *IEEE Access*, vol. 11, pp. 137999–138019, 2023.
[13] J. Tian, J. Lei, J. Zhang, W. Xie, and Y. Li, "SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image," *IEEE Trans. Geoscience Remote Sensing*, vol. 62, 2024, Art. no. 5613213, doi: 10.1109/TGRS.2024.3371481.
[14] R. Shwartz-Ziv and Y. LeCun, "To compress or not to compress- self-supervised learning and information theory: A review," *Entropy*, vol. 26, no. 3, 2024, Art. no. 252.
[15] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.
[16] J. Wang et al., "Crop specific inversion of PROSAIL to retrieve green area index (GAI) from several decametric satellites using a Bayesian framework," *Remote Sens. Environ.*, vol. 278, Sep. 2022, Art. no. 113085.
[17] Q. Garrido, R. Balestriero, L. Najman, and Y. Lecun, "RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank," in *Proc. 40th Int. Conf. Mach. Learn.*, Jul. 2023, pp. 10929–10974, iSSN: 2640-3498.
[18] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Sep. 2022.
[19] Y. Yuan and L. Lin, "Self-supervised pre-training of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2021.
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, May 2019, pp. 4171–4186.
[21] 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," *Int. J. Appl. Earth Observation Geoinformation*, vol. 106, Feb. 2022, Art. no. 102651.
[22] H. Wang et al., "CC-SSL: A self-supervised learning framework for crop classification with few labeled samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8704–8718, 2022.
[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *PMLR*, Nov. 2020, pp. 1597–1607.
[24] A. Kuhn, S. Ducasse, and T. Gîrba, "Semantic clustering: Identifying topics in source code," *Inf. Softw. Technol.*, vol. 49, no. 3, pp. 230–243, Mar. 2007.
[25] W.-J. Zheng, X.-L. Zhao, Y.-B. Zheng, J. Lin, L. Zhuang, and T.-Z. Huang, "Spatial-spectral-temporal connective tensor network decomposition for thick cloud removal," *ISPRS J. Photogrammetry Remote Sens.*, vol. 199, pp. 182–194, May 2023.
[26] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-supervised learning via redundancy reduction," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jun. 2021, pp. 12310–12320.
[27] C. Pelletier, G. I. Webb, and F. Petitjean, "Deep learning for the classification of Sentinel-2 image time series," in *Proc. Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2019, pp. 461–464.
[28] P. Defourny et al., "Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world," *Remote Sens. Environ.*, vol. 221, pp. 551–568, Feb. 2019.

Madeline C Lisaius (Student Member, IEEE) received the B.S. and M.S. degrees in Earth Systems with a focus on environmental spatial statistics and remote sensing from Stanford University, Stanford, CA, USA, in 2018 and 2019, respectively, and received M.Res. degree in environmental data science from University of Cambridge, Cambridge, U.K., in 2022. She is currently working toward the Ph.D. degree focused on SSL for agriculture and justice with the Department of Computer Science and Technology, University of Cambridge.

Her research interests include food security and environmental justice, remote sensing, and machine learning.

Andrew Blake (Fellow, IEEE) received the Ph.D. degree from the University of Edinburgh, in 1983.

He has been the Director of Microsoft Research in Europe from 2010 to 2015, inaugural Director with Alan Turing Institute from 2015 to 2018, and Chair with Samsung AI, Cambridge, U.K. from 2018 to 2022. Recently he has been consulting for Samsung, Siemens, and the U.K. Stock Exchange. He is a pioneer in the development of the theory and algorithms that make it possible for computers to behave as seeing machines. He is currently a Consultant in artificial intelligence, Scientific Adviser to the FiveAI autonomous driving company, recently acquired by Bosch, and Chief Scientific Adviser at Mantle Labs.

Dr. Blake was the recipient of the FREng and FRS fellowships.

Srinivasan Keshav (Fellow, IEEE) received the Ph.D. degree in computer science from the University of California, Berkeley, CA, USA, in 1991.

He was employed with AT&T Bell Labs and Cornell University. Most recently, he was a Professor with the University of Waterloo, Waterloo, ON, Canada. He is the Robert Sansom Professor of computer science with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K. His research focuses at the intersection of computer science and sustainability.

Dr. Keshav is a Fellow of the Royal Society of Canada, Association for Computing Machinery, and a Distinguished Alumnus of the Indian Institute of Technology, New Delhi, India.

Clement Atzberger received the Ph.D. degree in crop growth modeling and remote sensing data assimilation from Trier University, Trier, Germany, in 1997.

He occupied several academic positions as Associate to Full Professor in Germany, the Netherlands, Italy, and Austria, and is the Co-Founder and Research Director with Mantle Labs, London, U.K. His research interests include time-series analysis, forward and inverse radiative transfer modeling, crop growth modeling and data assimilation, imaging spectroscopy, and machine learning.