remote sensing

MDPI

Article

Self-Supervised Learning across the Spectrum

Jayanth Shenoy *, Xingjian Davis Zhang, Bill Tao, Shlok Mehrotra, Rem Yang, Han Zhao and Deepak Vasisht

University of Illinois Urbana-Champaign, Champaign, IL 61801, USA; xdzhang2@illinois.edu (X.D.Z.); yutao4@illinois.edu (B.T.); shlokm2@illinois.edu (S.M.); remyang@mit.edu (R.Y.); hanzhao@illinois.edu (H.Z.); deepakv@illinois.edu (D.V.)

* Correspondence: jshenoy2@illinois.edu

Abstract: Satellite image time series (SITS) segmentation is crucial for many applications, like environmental monitoring, land cover mapping, and agricultural crop type classification. However, training models for SITS segmentation remains a challenging task due to the lack of abundant training data, which requires fine-grained annotation. We propose S4, a new self-supervised pretraining approach that significantly reduces the requirement for labeled training data by utilizing two key insights of satellite imagery: (a) Satellites capture images in different parts of the spectrum, such as radio frequencies and visible frequencies. (b) Satellite imagery is geo-registered, allowing for fine-grained spatial alignment. We use these insights to formulate pretraining tasks in S4. To the best of our knowledge, S4 is the ***first*** multimodal and temporal approach for SITS segmentation. S4's novelty stems from leveraging multiple properties required for SITS self-supervision: (1) multiple modalities, (2) temporal information, and (3) pixel-level feature extraction. We also curate m2s2-SITS, a large-scale dataset of unlabeled, spatially aligned, multimodal, and geographic-specific SITS that serves as representative pretraining data for S4. Finally, we evaluate S4 on multiple SITS segmentation datasets and demonstrate its efficacy against competing baselines while using limited labeled data. Through a series of extensive comparisons and ablation studies, we demonstrate S4's ability as an effective feature extractor for downstream semantic segmentation.

Keywords: SITS; foundational models; self-supervised learning; multimodal

Citation: Shenoy, J.; Zhang, X.D.; Tao, B.; Mehrotra, S.; Yang, R.; Zhao, H.; Vasisht, D. Self-Supervised Learning across the Spectrum. *Remote Sens.* **2024**, *16*, 3470. https://doi.org/10.3390/rs16183470

Academic Editors: Claudio Piciarelli and Dino Ienco

Received: 17 July 2024
Revised: 24 August 2024
Accepted: 3 September 2024
Published: 19 September 2024

1. Introduction

In recent years, many organizations [1–3] have launched large satellite constellations for Earth observation. These constellations regularly capture high-resolution Earth imagery that is critical for measuring climate change [4,5], responding to humanitarian crises [6], precision agriculture [7], and natural resources management [8]. Specifically, satellites with multiple visits over a given location on Earth provide unique insights into complex spatial and temporal patterns [9–11] at such locations, unlike single satellite images. These satellite image time series (SITS) (as shown in Figure 1), for example, can provide greater insights into how crops on a farm grow over time, what types of crops are growing, or when the crops are ready to be harvested. SITS is also more robust to temporary disruptions such as cloud cover that may occur in single images. Due to its key advantages and environmental implications, SITS semantic segmentation has become a task of critical importance and has widespread use in many Earth-sensing applications, such as deforestation monitoring [12], urban planning [13], and agriculture crop type classification [14].

However, training segmentation models for SITS requires collecting large amounts of labeled data, requiring laborious manual annotation from domain experts [15]. This is especially challenging for semantic segmentation which requires pixel-level annotations. Moreover, many satellite images use nonoptical channels [16] beyond the standard RGB wavelength, making them difficult to interpret for humans.

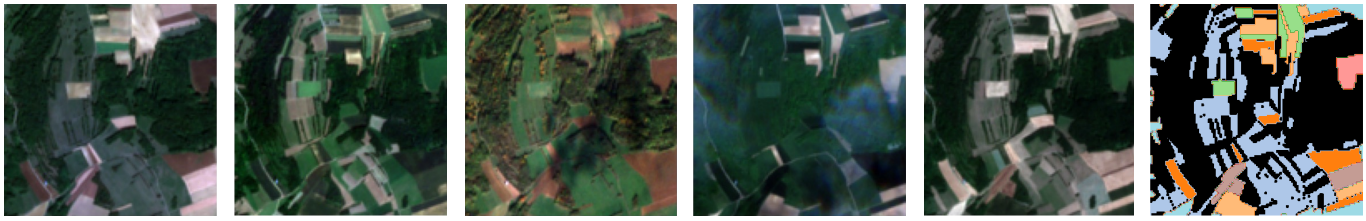

Figure 1. Optical images in one SITS captured at different points in time over the same location. The rightmost image is the segmentation mask corresponding to this spatial location. The different images illustrate the significant temporal variation that occurs during crop growth.

We propose S4, a novel *self-supervised approach* for semantic segmentation of SITS that eliminates the need for large amounts of labeled data. We observe that while labeling requires human effort, unlabeled data are abundant because satellites continuously orbit the Earth and collect data. Our key insight is that we can leverage this unlabeled data by utilizing two properties unique to SITS:

- **Multimodal Imagery:** Different satellites (or different sensors on the same satellite) collect images in different parts of the electromagnetic spectrum (e.g., RGB, radar). We can use such multimodal images for cross-modal self-supervision.
- **Spatial Alignment and Geographic Location:** Satellite images are geo-referenced, i.e., each pixel has a geographic coordinate (latitude and longitude) associated with it. This allows for spatial alignment between data collected in different parts of the spectrum.

Given the unique properties of SITS, S4 exploits the abundant unlabeled satellite data through cross-modal self-supervision. Specifically, we use different data modalities for a given location to learn informative intermediate representations *without any labeled data*. Using unlabeled SITS, we can pretrain representative SITS encoders that perform effectively on downstream SITS segmentation. We achieve this by pretraining SITS segmentation models through two auxiliary tasks:

- **Cross-Modal Reconstruction Network:** We design a new cross-modal SITS reconstruction network that attempts to reconstruct imagery in one modality (e.g., radar) from the corresponding imagery in another modality (e.g., optical). Our reconstruction network encourages the encoder networks to learn meaningful intermediate representations for pixel-wise tasks by leveraging the structured spatial alignment in satellite image data.
- **MMST Contrastive Learning:** We formulate a novel multimodal, spatiotemporal (MMST) contrastive learning framework for SITS. We train one encoder for each modality (e.g., for radar and optical imagery) and align the intermediate representations using a contrastive loss. Our contrastive loss operates along both the space and time dimensions of the feature space to align multimodal SITS. Intuitively, our loss helps negate the impact of temporary noise (such as cloud cover) that is visible in only one of the input images.

We also design a **temporal resampling strategy** to reduce temporal misalignment between modalities (Section 5) as a preprocessing step. Our temporal preprocessing strategy leverages **timestamp metadata** from the satellite imagery to provide a course-grained time alignment of images across multiple modalities. After temporal preprocessing, the coarsely aligned multimodal SITS is then fed to our encoder network. We jointly train two encoders, one for each modality, using our auxiliary constrastive learning and reconstruction tasks as defined above. Consequently, our encoders generate informative intermediate feature representations appropriate for downstream semantic segmentation. The auxiliary tasks solely rely on unlabeled data. The intermediate representations will be fed into a task-specific decoder network for segmentation. Given our pretraining tasks, the downstream decoder network needs only a small amount of labeled data for training. We visualize our proposed model in Figure 2.

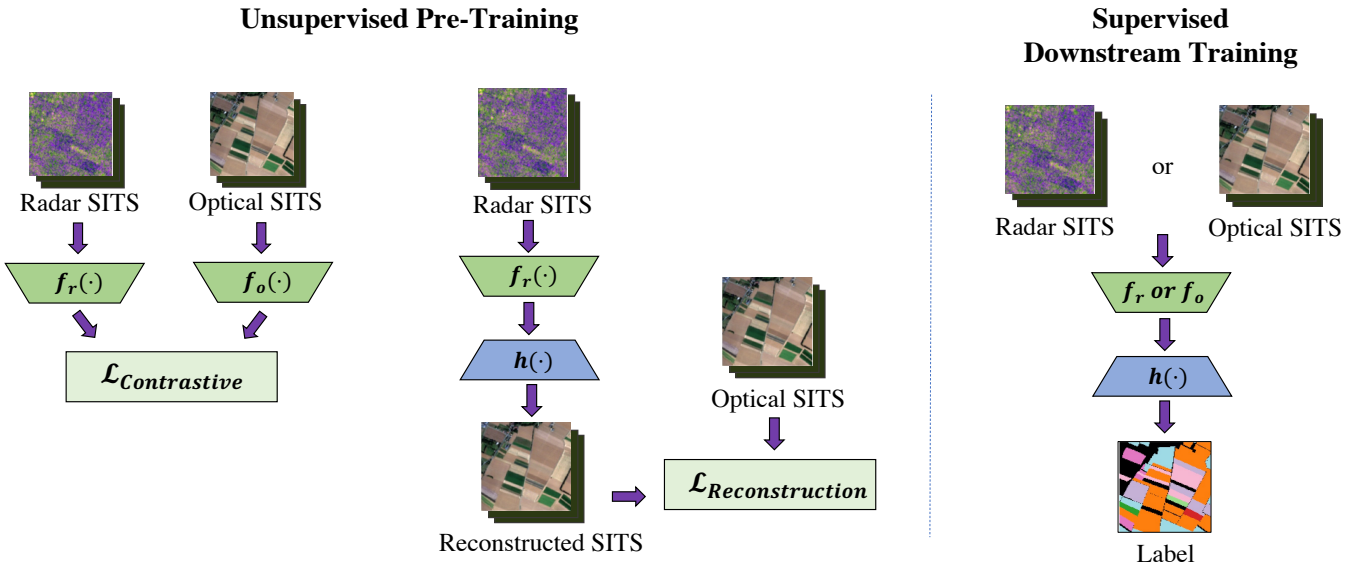

Figure 2. Overview of S4. S4 takes in temporally preprocessed multimodal time series data. During pretraining, radar-optical SITS pairs flow through the network and our proposed MMST contrastive loss and cross-modal reconstructive loss operate on their encodings. After pretraining, a small amount of labeled data are used to fine-tune the model for SITS segmentation.

We evaluate S4's performance on two satellite image datasets: PASTIS and Africa Crop Type Mapping segmentation tasks. To demonstrate the efficacy of aligned satellite imagery and showcase the opportunity for self-supervised pretraining for SITS segmentation, we collected m2s2-SITS, two large-scale unlabeled but modality-aligned datasets of satellite images corresponding to the same regions of our labeled datasets. In our evaluation, we pretrain our model using our curated dataset m2s2-SITS and fine-tune the models on the existing datasets with segmentation labels. We compare against multiple self-supervised remote sensing baselines including SatMAE (masked autoencoder), SeCo (temporal contrastive learning), GSSL (geographical and temporal contrastive learning), and CaCo (change aware sampling and contrastive learning). Additionally, we compare against a custom-designed multimodal baseline based on naive variants of S4 and conduct detailed ablation studies across various influences on our model, such as geographic data usage, cloud cover robustness, input modalities, and differing loss functions. Experiments demonstrate that S4 outperforms competing self-supervised baselines for segmentation, especially in the case where the number of labeled data is relatively small. As a result, S4 takes a first step towards self-supervised SITS segmentation through novel techniques that reliably leverage multimodal and spatially aligned imagery. In summary, this paper makes the following contributions:

- We propose S4, a self-supervised training method for SITS semantic segmentation that considers the unique structural characteristics of satellite data, such as multiple modalities, spatial alignment, and temporal change through novel cross-modal reconstruction and contrastive learning frameworks.
- We release m2s2-SITS, a large dataset of spatially-aligned, multimodal SITS to aid in self-supervised pretraining.
- We evaluate S4 on multiple SITS datasets and benchmark our approach against other self-supervised approaches commonly used on satellite imagery. Our results demonstrate the effectiveness of S4 through significant improvement over prior state-of-the-art methods on downstream SITS semantic segmentation.

2. Design Motivation and Advantages of S4

Our design solves multiple challenges unique to SITS segmentation. First, S4 can significantly reduce the need for labeled training data by exploiting abundant unlabeled images. Second, some image modalities such as optical images are obstructed by clouds, leading to missing or incorrect information (see Figure 3). In fact, around 75% of the Earth's surface is covered by clouds at any given point in time [17–19]. Through pixel-wise alignment of radar and optical image encoders, S4 effectively leverages the rich information provided by radar in cloudy settings (since radar passes through clouds). As a result, S4 pretrains powerful encoders for both modalities that can reduce the negative impact of cloud cover on model performance. Third, for each modality, the reflectance value of each pixel is different. Because of this, certain Earth surface characteristics may be clearly visible in one modality, but not necessarily as visible in the other. S4 solves this through its cross-modal reconstruction network which is able to infer the presence of vegetation (for example) in one modality based on the patterns learned from the other. Intuitively, this enables the model to learn to understand the signatures or indicators that would correspond to certain features across modalities. Lastly, although satellite images are easily aligned spatially, it is difficult to align them temporally, since different imaging modalities are often on separate satellites [20]. As these separate satellites have different orbital patterns, they do not simultaneously pass over the same location at the exact same instant. S4 resolves the temporal heterogeneity of satellite data through temporal alignment across modalities—first through its coarse-grained preprocessing strategy and second through its fine-grained contrastive learning framework.

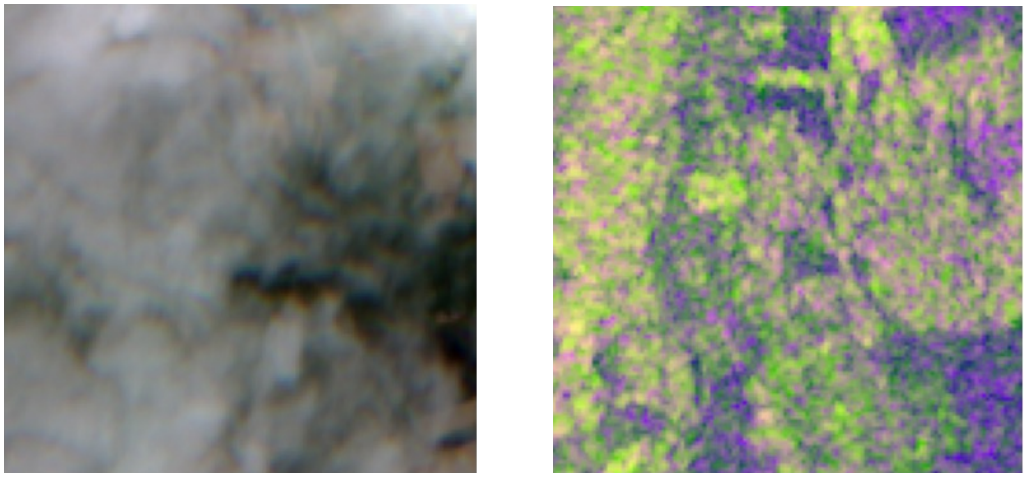

Figure 3. Multimodal images captured on the same day: while the optical image (**left**) is occluded by clouds, the radar image (**right**) is not affected.

Importantly, S4 delivers *single-modality inference*. Single-modality inference is crucial due to two real-world constraints. First, satellites capturing images of different modalities may be operated by different entities. In fact, 95% of Earth Observation satellites are equipped with only a single sensing modality [21]. Thus, while multimodal training data may be available through aggregating public datasets, such data are almost always not available at inference time [20]. Second, requiring both modalities during inference increases the delay of decision-making in response to critical events (e.g., floods and fires), since multiple modalities can be offset in time by several hours to days (depending on satellite orbits) [22]. Hence, while we leverage multimodal data at training time, we limit ourselves to a single modality inference.

3. Related Work

To the best of our knowledge, we present the first self-supervised approach for semantic segmentation of multimodal satellite image time series. We discuss related work below.

Learning with Satellite Imagery Prior work on satellite imagery can be characterized as (a) single-image or (b) SITS. Single-image methods [23–25], although more extensively studied, are unable to effectively gain insights into many environmental sensing applications that typically evolve over time, such as crop mapping and disaster monitoring [26,27]. Most of the prior work in self-supervised learning for satellite imagery are single-image unimodal techniques [28,29] that cannot leverage the multitemporal and multimodal structure of SITS data. Therefore, there has been a growing number of recent supervised efforts that leverage SITS, which better captures the complex characteristics that evolve over time in many environmental sensing tasks. These efforts have designed SITS-based models for a variety of downstream Earth observation tasks, such as image classification [26], super-resolution [30], and segmentation [9]. Each of the cited methods focuses on developing space–time encoding for effective feature extraction.

Although the state-of-the-art SITS-based techniques yield vast improvements over single-image methods for a variety of tasks, they mainly rely on unimodal, optical satellite imaging. However, optical imagery is not robust under low visibility conditions (e.g., due to rain, night, or clouds), making it difficult to obtain such data in time-sensitive settings [31]. S4 extracts insights even from nonoptical SITS during training, making it significantly more practical in these challenging conditions.

More recently, self-supervised methods have been explored for satellite imagery that aims to provide downstream benefits on a variety of different satellite imaging tasks [28,29,32,33]. Although these methods demonstrate some promise, they all provide only uni-modal solutions for self-supervision and do not leverage the spatial alignment between modalities as S4 does. Other solutions that do provide a multimodal solution to self-supervision are often only monotemporal and are incapable of performing any SITS-related tasks since they only operate on a single image [34]. Additionally, many prior works in satellite image self-supervision [29,32,34] often break down the structure of the image pixels in the feature space by flattening the image. As a result, although this may be beneficial for some tasks like classification, it is problematic for semantic segmentation which requires a deeper level of spatial context from pixel-wise features in the image. Unlike prior works in satellite image self-supervision, S4 is specifically designed for the task of self-supervision of SITS by leveraging the *ALL* key characteristics of SITS data (1) multiple modalities, (2) temporal alignment, and (3) pixel-wise feature extraction.

Learning with Multiple Modalities Many modern satellites are equipped with nonoptical sensing modalities [3,35]. Computer vision in nonoptical imaging modalities, such as radar, has been explored much less than optical imaging modalities. This is due to radar images being difficult to interpret by humans compared with optical images, making it harder to acquire labeled data. Most prior works focus on exploring radar images using unsupervised techniques [36–38]. These techniques do not generalize well to different events and often exhibit limited performance. Prior work on multimodal satellite imagery has also explored the reconstruction of obscured or cloudy optical images by leveraging aligned nonoptical radar images [19,39,40]. These multimodal reconstruction models tend to provide a more accurate optical reconstruction than prior unimodal methods like image in-painting [41,42], demonstrating the potential utility of nonoptical multimodal learning. More recently, there have been efforts to try and incorporate multiple modalities for SITS [27,43,44]. Such efforts typically focus on designing fusion techniques for modalities along with reliable spatiotemporal encodings to improve performance. S4 distinguishes itself from these methods by providing a training method that requires significantly less labeled data and only a single SITS modality at inference time.

Self-Supervised Learning Self-supervised learning for visual representations has gained prominence within the last few years [45–49]. One of the most recent notable self-supervised methods has been contrastive representation learning, which attempts to align similar pairs of images as a pretraining task to help with downstream model performance [50–52]. Although prior work mainly focuses on instance-level contrastive learning for downstream classification, recent works explore pixel-level contrastive learning

techniques, which provide better transfer to segmentation tasks [53–56]. For the majority of these contrastive learning approaches, positive pixel pairs are assigned either by using corresponding pixels with the same label or through corresponding pixels from different augmented views of the same image. In contrast, S4 leverages the spatial alignment between different satellite modalities and associates positive pairs through corresponding pixels in different modalities.

Semantic Segmentation of SITS Many prior methods have found success in using UNet-based architectures [57] for encoding representations helpful for satellite image segmentation [9,44,58]. More recent efforts specific to SITS have also designed multitemporal and multimodal fusion schemes using convolutional encoders [27,43]. The advantage of S4 is that we require only a single modality of SITS during inference time, whereas every prior multimodal method requires both. S4 also incorporates a novel self-supervised approach that significantly reduces the need for labeled data.

4. Problem Setup: Satellite Imaging with Multiple Modalities

Our work is situated in the emerging context where different satellite constellations capture Earth imagery in different frequency bands. We seek to extract spatiotemporal insights from these data. A majority of the satellites capture optical images that passively monitor the reflections of sunlight off the Earth's surface. These optical images are often multispectral, including imaging bands outside the standard visible red, green, and blue channels. However, a key disadvantage of such imagery is that optical satellite images are often occluded by clouds (Figure 3) and are easily obscured in low-lighting conditions, such as night and fog [18,19,39,40].

Some satellites are equipped with radar imaging that works by actively transmitting pulses of radio waves and measuring the reflectance of these radio pulses to produce radar images. These radio waves utilize a longer wavelength than optical images and are typically better at monitoring certain aspects of the surface, such as moisture and topology. However, the resolution of radar imagery is lower than that of optical images. Satellites are typically equipped with either optical or radar imaging modalities, *but not both* [20]. Therefore, images in optical and radar SITS cannot be perfectly aligned in time.

Each image captured by satellites is georeferenced, i.e., we can extract per-pixel geographic coordinates. This allows us to spatially align images even when captured on different satellites. However, leveraging the temporal aspect of SITS data poses some challenges. First, images in SITS, unlike videos, are not taken at regular intervals. Images are taken over a location only when a satellite orbits over that location, meaning the time between images in SITS is irregular based on the satellite's orbit. Second, for multimodal SITS, different sensing modalities are often located on different satellites, meaning that images of SITS of different modalities are not only unaligned in time, but they can also result in time series of vastly different lengths.

S4's primary task is to use the ample amount of unlabeled imagery collected by satellites for cross-modal self-supervision. Our formulation builds upon the key idea of *pixel-level semantic consistency between multimodal images captured over the same location at roughly the same time*. We propose a new training objective that encourages the similarity of corresponding space–time features across modalities while maximizing the distance between the features corresponding to either different locations or different times. Though different modalities have certain distinctions like differing spectral ranges, the *semantic representation of the underlying scene should be agnostic to both wavelength and noise*, e.g., cloud cover (for optical) or capture angle (for radar), and thus S4 can be used to achieve a course-grained alignment of the multimodal, multitemporal features beneficial for self-supervised learning. By leveraging these natural structural characteristics of SITS, our approach extracts a more informative representation that limits the impact of modality-specific noise.

Notation and Setup We consider the respective radar and optical image modalities, $\mathcal{X}_r \subset \mathbb{R}^{(T_1 \times C_1 \times H \times W)}$ and $\mathcal{X}_o \subset \mathbb{R}^{(T_2 \times C_2 \times H \times W)}$, where T_i, C_i, H, and W are the number of images in the time series, number of image channels, image height, and image

width dimensions, respectively. During training, we assume access to N SITS pairs $\{(x_o^{l_n}, x_r^{l_n})\}_{n=1}^{N} \in (\mathcal{X}_o \times \mathcal{X}_r)^N$, where l_i corresponds to the location where the SITS was captured. Although we have N total SITS pairs, we assume that only K of these N image pairs ($K \ll N$) have segmentation labels: $\{(x_o^{l_k}, x_r^{l_k}, y^{l_k})\}_{k=1}^{K}$, where the label $y^{l_k} \in \mathbb{N}^{H \times W}$ maps each pixel location to a given class.

5. Method

Figure 2 provides an overview of S4. At a high level, S4 operates in three stages:

- **Pretraining:** During the pretraining stage, S4 leverages abundant unlabeled data by jointly optimizing the proposed pixel-wise multimodal contrastive loss $\mathcal{L}_c$ (Section 5.1) and reconstruction loss $\mathcal{L}_r$ (Section 5.2):

$$\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_r, \tag{1}$$

  where λ is a hyperparameter controlling the relative weight between the two loss terms. Neither of the above two losses require labels.
- **Downstream Training:** In this stage, our network is fine-tuned on the K SITS pairs with labels for downstream segmentation by further appending a segmentation module over the learned features (Section 5.3).
- **Inference:** In the final stage, S4 predicts a single segmentation map per different location from the SITS of a *single* modality (either radar or optical).

Time Series Interpolation A key challenge for S4 is that satellites visit the same location at different times, leading to temporal mismatch across modalities. Higher temporal mismatch across images causes more semantic mismatch in the underlying representation. To avoid this problem, we introduce a preprocessing strategy to coarsely align the temporal dimension between differing modalities. Our preprocessing strategy leverages **temporal metadata** from satellite imagery. This preprocessing step is necessary to ensure finer-grained spatial and time alignment through the rest of the training process. Recall that we are given as input $x_r^{l_n} \in \mathbb{R}^{(T_1 \times C_1 \times H \times W)}$ and $x_o^{l_n} \in \mathbb{R}^{(T_2 \times C_2 \times H \times W)}$, where $T_1 \neq T_2$ in general. We determine which SITS modality has fewer time frames: let $T_{min} = \min(T_1, T_2)$ and define $x_{min}^{l_n} := x_r^{l_n}$ if $T_{min} = T_1$ and $x_{min}^{l_n} := x_o^{l_n}$ otherwise. The time series $x_{min}^{l_n}$ remains unchanged. To make the other modality's SITS the same length, we adopt nearest-timestamp interpolation: for each image $x_{min_i}^{l_n} \in x_{min}^{l_n}$, we find the image in the corresponding time series of the opposite modality that was captured at the time closest to $x_{min_i}^{l_n}$. The result of our interpolation strategy results in N SITS pairs $\{(x_o^{l_n}, x_r^{l_n})\}_{n=1}^{N}$ where both modalities' time series each contain T_{min} images coarsely aligned in time.

Encoder Design The first part of S4 consists of an encoder network that takes the spatially aligned optical and radar SITS, $x_o^{l_n}$ and $x_r^{l_n}$, as input. The encoder consists of four convolution layers, of which the first two are input-specific based on modality. Let y_r, y_o denote the first two layers (used for the radar and optical domains, resp.) and y_c denote the last two layers. The encoder for the radar and optical domains are $f_r = y_c \circ y_r$ and $f_o = y_c \circ y_o$. We use the outputs of f_r, f_o as the features passed to the rest of the network. The encoders $f_r(\cdot)$ and $f_o(\cdot)$ use a 3D U-Net backbone architecture [59] consisting of convolution, batchnorm [60], and max-pooling layers with leaky ReLU activations. The 3D operations are applied along both the temporal dimension and the spatial dimensions. This architecture has been used as a state-of-the-art benchmark for a wide variety of prior work in SITS segmentation tasks [9,44,58] and offers a relatively simple design that is comparable in performance to other state-of-the-art SITS segmentation architectures that use a separate sequential technique to handle the temporal dimension.

5.1. Multimodal Space–Time Contrastive Learning

Our approach builds upon the key idea of *semantic scene consistency between varying satellite modalities that are captured over the same space and at the same time*. Therefore, our

encoder should map image pixels captured over similar space–time to similar representations; while encoding random noncorresponding pixels to differing representations. We incorporate this intuition in our training scheme through contrastive learning. Inspired by recent successes of pixel-wise contrastive learning [56], we propose a pixel-wise contrastive loss that preserves the spatiotemporal structure of our representations for better transfer to downstream pixel-level tasks like semantic segmentation. Figure 4 outlines this approach.

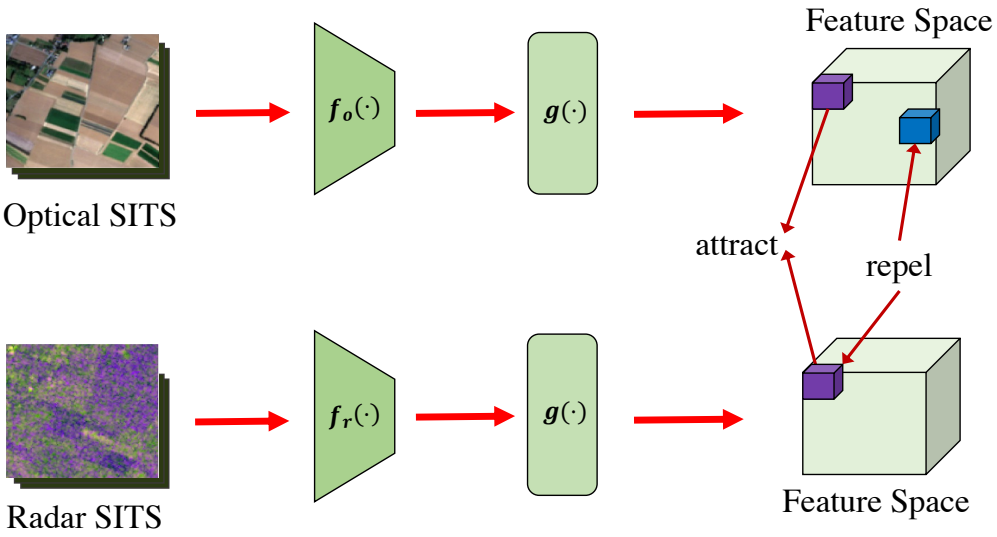

Figure 4. Multimodal Space–Time Contrastive Learning for SITS. Our approach operates on the encoded SITS feature maps. Corresponding space–time pixels on the feature map are denoted as positive pairs that the contrastive loss tries to align. Noncorresponding pixel pairs are negative and repelled by the loss.

Prior work [28,29,50] on contrastive learning for images often use single image views and perform a variety of data augmentations (e.g., crop, rotate, blur) on a single view. Different augmentations that correspond to the same view are often correlated together as a positive pair for the loss function. However, in the case of satellite images, we benefit from the availability of multimodal data and omit the augmentation step. Each modality captures the same view of Earth at different wavelengths and can be used as a different transformation.

Similar to prior work [50,56], we implement a projection head network $g(\cdot)$ that maps the output of $f_r(\cdot)$ and $f_o(\cdot)$ to the latent space where the contrastive loss is applied. The projection head consists of two successive $1 \times 1 \times 1$ 3D convolution layers with batch normalization and LeakyReLU activation. Note that $g(\cdot)$ is only used during contrastive pretraining and not while training the reconstruction network or the downstream segmentation network. The output of $g(\cdot)$ is a feature map of the encoded SITS with compressed spatiotemporal dimensions.

We assign positive pairs as pixels in the feature space with the *same spatial and temporal dimensions*, across different modalities. Pixel pairs with different space or time dimensions are considered negative pairs in our loss, since they correspond to different semantics. We opt to use the InfoNCE loss [61] as our contrastive loss function:

$$\mathcal{L}_c = -\log\left(\frac{e^{sim(z_i,z_j)/\tau}}{e^{sim(z_i,z_j)/\tau} + \sum_{z_n \in Z} e^{sim(z_i,z_n)/\tau}}\right) \tag{2}$$

Positive pairs z_i and z_j are corresponding space–time pixels in the feature map representations of opposite modalities. Z is the set of all negative feature map pairs with

anchor pixel z_i in the opposite modality. More broadly, Z consists of the features that were captured at different spatial locations or different times from z_i. The cosine similarity function is defined as $sim(u, v) = u^T v / \|u\| \|v\|$. The temperature hyperparameter τ is set to 0.5 by default. The loss is first averaged over all pixels in the first modality's feature map; then, we compute the loss averaged across all pixels in the second modality feature map as anchor pixels. Finally, we average the loss across both modalities together to compute the final contrastive loss per sample in the batch.

5.2. Cross-Modality Reconstruction Network

Although acquiring semantic segmentation labels for SITS is challenging, an advantage of SITS is that images can be easily aligned spatially. To leverage the spatial alignment between multiple modalities, we design a reconstruction network that infers the SITS of one modality given the other. By learning to reconstruct SITS from other modalities as an auxiliary task, the reconstruction network is able to learn representative features for the input modality that are helpful for the downstream segmentation task.

Our reconstruction network uses encoder f_{in} (either f_r or f_o depending on the inference modality) and decoder h. The network takes as input a SITS from one modality (denoted $x_{in}^{l_n}$, which has C_{in} channels) and attempts to reconstruct the corresponding SITS of the other modality (denoted $x_{out}^{l_n}$, which has C_{out} channels). The output of our reconstruction network is the estimated reconstruction of the SITS of the second modality: $\hat{x}_{out}^{l_n} = h(f_{in}(x_{in}^{l_n}))$. We define the loss for our reconstruction network as the mean absolute error ($L1$ loss) between the original and the reconstructed time series as expressed in the equation below:

$$\mathcal{L}_r = \frac{\|\hat{x}_{out}^{l_n} - x_{out}^{l_n}\|_1}{T_{min} \cdot C_{out} \cdot H \cdot W} \tag{3}$$

5.3. Downstream Training

Finally, after pretraining our network with spatially aligned modalities, we fine-tuned the network on a small number of labeled samples. We use the same encoder $f_{in}(\cdot)$ and decoder $h(\cdot)$ networks used during pretraining. However, we modify the number of channels of the decoder's final convolution layer for the relevant segmentation map output. We carry out the downstream training using standard cross-entropy loss.

Generalizing to Other Temporal Encoders A key advantage of S4 is that it can be easily extended to other types of SITS segmentation architectures that may encode the temporal dimension differently. Such architectures may use convolutional layers to encode the spatial dimensions and a temporal model, such as LSTM/RNN [9,62], to handle the temporal dimensions. In these cases, S4 can first be used to train the convolutional spatial encoders of the network. During downstream training, the temporal encoder can be added to the network and trained using multimodal features extracted from the spatial encoders.

6. Experiments and Results

In this section, we describe experiments conducted to evaluate S4. We train all self-supervised models in two phases. First, we pretrain all models for 100 epochs. For pretraining, the models are trained using m2s2-SITS, our curated geographic-specific, pretrain datasets. The pretrain datasets consist only of images and do not have annotated labels. In the second stage, we fine-tune the network for the downstream segmentation task for 50 epochs using the datasets with annotated labels. For optical imagery, we train using only using the RGB channels to be consistent and fair to prior work in self-supervised models for remote sensing [28,29,33].

6.1. Curated Pretrain Datasets

We demonstrate the efficacy of multimodal self-supervised pretraining by gathering a large unlabeled dataset of aligned optical and radar SITS. Although labeling satellite images is difficult, there is an abundant amount of unlabeled multimodal satellite SITS.

Our main motivation for curating this dataset is to illustrate how geographically specific aligned multimodal satellite data are easy to acquire, allowing for greater opportunities to benefit from pretraining. Given the geo-tagged characteristic of satellite images, we can also collect data from geographically specific locations and study how the geographic location of images from certain regions can have an impact on the performance of downstream segmentation. Given that our pretrain data come from the same geographic location as the fine-tuned sets, our curated dataset is suitable and a geographically representative set of imagery for pretraining.

Motivation: We collected our own pretraining dataset because there is no large-scale SITS dataset available in which radar and optical satellite imagery are spatially aligned, i.e., all images of the same location have the same number of pixels, and the same pixel in all images corresponds to the exact same geographic coordinate. We curate this dataset by collecting and aligning Sentinel-1 (radar) and Sentinel-2 (optical) images. Furthermore, a constantly shifting satellite orbit requires stitching multiple different images that each capture a given location partially. This pixel-level alignment of m2s2-SITS is crucial for our self-supervised model which requires pixel-wise contrastive and reconstruction loss in pretraining.

Curation: Images in our dataset were collected from Sentinel 1 and 2 satellites and were aligned using the Microsoft FarmVibes SpaceEye workflow [63]. m2s2-SITS consists of satellite imagery from randomly sampled geographic locations within France and South Sudan, where the fine-tune datasets (PASTIS-R and Africa Crop Type Mapping, respectively) are captured from, and the time period of m2s2-SITS is approximately one year. We ensure that the images of m2s2-SITS are taken at least a year prior to the images in the fine-tuned datasets to prevent the chance of duplicates. Although in our evaluation we use only 3 RGB bands to ensure a fair comparison with baseline approaches, our dataset contains the full 12-band multispectral imagery from optical S2 and the 2 polarizations from radar S1. Specifically, we collect a pretraining dataset over France that contains 5314 time series with a total of 731 k Sentinel 1 images and 90 k Sentinel 2 images. After pretraining on this dataset, the models are fine-tuned on the PASTIS-R dataset. Our Africa pretrain dataset contains 5941 time series, with 193 k Sentinel 1 images and 70 k Sentinel 2 images. After pretraining on this dataset, the models are fine-tuned on the Africa Crop Type Mapping dataset. We plan to release our custom-curated pretrain datasets.

6.2. Fine-Tuned Datasets

PASTIS-R: The PASTIS-R [43] agricultural dataset contains 2433 optical and radar SITS from the ESA's Sentinel 1 and 2 satellites. Each SITS contains between 38 and 61 images taken between September 2018 and November 2019. The dataset provides an annotated semantic segmentation map for each of 2433 spatial locations, where every pixel is given a semantic label from one of 20 different crop type classes. Many optical images are partially occluded by clouds. Note that we only consider the semantic segmentation labels from this dataset and DO NOT perform parcel classification experiments (as carried out in the original paper), as semantic segmentation is a strictly more challenging task. Table 1 shows the semantic segmentation classes for the labels of this dataset.

Africa Crop Type Mapping: The Africa Crop Type Mapping dataset [14] contains multi-modal SITS over various regions in Africa. Ground truth labels in this dataset were collected for 4 classes in 2017. Table 2 shows the semantic meanings of the labels for the Africa Crop Type Mapping dataset. For our experiments, we used 837 fields in the South Sudan partition.

Table 1. Table showing class names and the number of parcels in the PASTIS-R dataset [43].

Class Name	Number of Parcels
Meadow	31,292
Soft winter wheat	8206
Corn	13,123
Winter barley	2766
Winter rapeseed	1769
Spring barley	908
Sunflower	1355
Grapevine	10,640
Beet	871
Winter triticale	1208
Winter durum wheat	1704
Fruits, vegetables, flowers	2619
Potatoes	551
Leguminous fodder	3174
Soybeans	1212
Orchard	2998
Mixed cereal	848
Sorghum	707
Void label	35,924

Table 2. Table showing class names and the number of parcels in the Africa Crop Type Mapping South Sudan dataset [14].

Class Name	Number of Parcels
Groundnut	59
Rice	75
Maize	84
Sorghum	619

6.3. Implementation Details

Preprocessing: We preprocess data using mean-std standardization using values from the fine-tuned dataset. We preprocess both the pretrain and fine-tuned datasets. For optical images, we use only the RGB channels in our experimentation rather than the 10–12 multispectral channels to be consistent with the prior self-supervised baseline approaches that were designed for 3-channel images [33].

Training: Across experiments, we set λ (the joint hyperparameter weighting $\mathcal{L}_c$ and $\mathcal{L}_r$) to be 10^{-2} when pretraining. Additionally, we split our original datasets into the train, validation, and test splits. For PASTIS-R, we use folds 1, 2, and 3 for training, fold 4 for validation, and fold 5 for testing as specified by the authors in [43]. For Africa Crop Type Mapping, we use the original partitions specified by the curators of the dataset for training, validation, and testing [58]. We run each segmentation model on the validation set after every epoch during training; at test time, we evaluate using the model checkpoint that attains the highest validation IoU. We train all models on a 2 $\times$ NVIDIA A100 GPUs. Pretraining takes approximately 3 h and fine-tuning takes around 1.5 h. We use the Adam optimizer [64] with a learning rate of 10^{-3}. In our evaluation of the Africa Crop Type dataset, due to high class imbalance and irrelevance of predicting the background class, we ignore the background class when computing the mIoU score.

Model Architecture: We follow the implementation of 3D U-Net provided by *Garnot and Landrieu* [9]. The **encoder** takes as input either an optical or radar image. In total, the encoder consists of a total of five 3D convolution layers. In each 3D convolution layer, stride and padding of 1 are used. The input dimension of the first convolution block is modified to either three or ten based on whether the input image is radar or optical. Three-dimensional batchnorm followed by a leaky ReLU activation function is used after all convolution layers. When training on the PASTIS-R crop segmentation dataset, 3D max-pooling layers are used after the 2nd and 4th convolution layers. The max-pooling layers use a kernel size of 2, stride of 2, and padding of 0. The multimodal fusion model uses this backbone as well.

The **decoder** consists of a total of four 3D convolution layers. The first two convolution layers are followed by 3D transposed convolution layers. The transposed convolution layers use a stride size of 2 and padding of 0. Each of these layers is also followed by a 3D batchnorm and leaky ReLU activation function. The output dimensions of the final convolution layer are modified to match the number of classes for the corresponding segmentation task. In the case of using the architecture for the reconstruction task, the final number of output features is equivalent to the number of input features to the encoder. The network architectures also utilize *skip connections* between the encoder and decoder.

For fine-tuning, given that our time series can be of variable length, we use the collate function provided by the original dataset. In pretraining, we fix the time series to the length of the 90th percentile length of the time series in the dataset. For shorter time series than our fixed length, we repeat the last image in the time series to ensure all time series are of the same length. For longer time series, we clip the last few samples to make the time series match the fixed length.

6.4. Evaluation Metric

We chose to use mIoU as the main evaluation metric for evaluating our model's segmentation performance. Many prior works on satellite image segmentation [65,66] use this metric as opposed to overall segmentation accuracy since it is more robust to datasets with high class imbalance. For example, in a dataset that consists mostly of background class labels, a model's accuracy can be high by simply overpredicting the background class. However, mIoU is a much more strict metric that gives equal importance to all the predicted classes. Formally, mIoU can be defined by the following equation:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{4}$$

where TP, FP, and FN denote the number of true positive, false positive, and false negative pixels, respectively. N is the total number of classes.

6.5. Baselines

We benchmark S4 against several competing self-supervised baselines. To the best of our knowledge, we are the first self-supervised approach that leverages multimodal imagery for SITS.

SatMAE: SatMAE (2022) [32] is a SOTA Masked AutoEncoder-based vision transformer architecture. designed specifically for multitemporal satellite imagery. We implement SatMAE from the original codebase provided by the authors. Given that our work pertains to SITS, we use the multitemporal variant of SatMAE designed for SITS. We pretrain these models for 100 epochs. For segmentation, we similarly use a transpose convolutional neural network as a decoder that is trained during the fine-tuning stage. The authors of SatMAE also use a convolutional decoder when performing such experiments for downstream segmentation. We tuned the hyperparameters of the baseline to achieve the best possible performance on our two datasets during the evaluation.

SeCo, CaCo, and GSSL: We compare against modern prior self-supervised work for remote sensing that uses single-modal contrastive loss. We compare against SeCo

(2021) [28], CaCo (2023) [33], and GSSL (2021) [29], all of which use contrastive learning to align single-modal image pairs. In their original papers, these baselines use a single image ResNet-NN network to contrast scenes of different timestamps.

We implement SeCo, CaCo, and GSSL from the original code base provided by the authors. All models use the MoCo-V2 architecture [67] with ResNet [68] backbone. We pretrain these models for 100 epochs. In the original implementation, a 2D UNet is used as the fine-tuned network. We found that the performance of this implementation was limited in our scenario since it is incapable of training on the entire SITS. Considering such limitations, our fine-tuning implementation for these baselines involves feeding each image in the SITS through the pretrained ResNet encoder and collecting the encoded feature maps of all of them. As a result, instead of one single feature map at each skip connection, we now obtain a time series of feature maps. We then pass the encoded image time series through a pixel-wise ConvLSTM decoder network to reduce the temporal dimension for fair comparison, before feeding the reduced single feature map into the upsampling part of the UNet-2D structure to achieve the final semantic segmentation predictions. We tuned the hyperparameters of the baselines to achieve the best possible performance on our two datasets during the evaluation.

Self-Supervised Multimodal Fusion: This approach is a naive self-supervised approach of leveraging multimodal data for SITS segmentation. Let $x^{l_n}_{m_1}$ denote the SITS modality we have access to at inference time. We first pretrain a network $r(\cdot)$ that, given $x^{l_n}_{m_1}$, learns to reconstruct the SITS of the other modality $x^{l_n}_{m_2}$ (using the loss in Equation (3)). Then, we train a separate network that takes as input the concatenation of $x^{l_n}_{m_1}$ and $r(x^{l_n}_{m_1})$. Using this network, we produce the segmentation label using the PASTIS early fusion technique [43]. During inference, we similarly generate the SITS of the missing modality using the reconstruction network and perform segmentation on the original and generated modalities.

6.6. Quantitative Segmentation Results

We first examine the segmentation performance quantitatively using only a few labels for downstream training.

Results on PASTIS-R: Table 3 reports the mIoU on the PASTIS-R test set using both 100% and 10% of the labeled dataset. S4 outperforms all competing baselines across the board for both optical and radar inference experiments. We observe less relative improvement in the radar inference experiments due to radar being a low-resolution modality that provides less information than nonoccluded optical images. We also observe greater improvement when the amount of labeled data provided is lower. Finally, although the self-supervised fusion technique leverages multimodal, temporal data and largely outperforms all other baselines, S4 provides greater performance gain through its sophisticated cross-modal contrastive and reconstruction framework.

Table 3. Segmentation results on the PASTIS-R and Africa Crop Type Datasets (mIoU).

Dataset	**PASTIS-R**				**Africa Crop Type**			
Method	**Radar 100%**	**Optical 100%**	**Radar 10%**	**Optical 10%**	**Radar 100%**	**Optical 100%**	**Radar 10%**	**Optical 10%**
SatMAE	37.9	36.2	11.1	28.2	12.4	11.9	3.30	7.06
SeCo	43.3	23.9	27.2	12.4	9.05	9.05	2.85	2.58
GSSL	41.8	21.9	30.9	13.6	18.8	4.36	5.67	8.04
CaCO	42.7	23.4	24.9	16.4	9.05	9.05	2.78	2.82
MMF [1]	53.2	48.3	36.0	27.4	9.08	18.1	9.02	18.5
Sup. [2]	51.6	52.9	22.8	17.1	7.08	6.94	2.58	2.69
S4 (Ours)	**54.6**	**53.4**	**36.5**	**33.7**	**21.2**	**31.2**	**10.2**	**24.4**

[1] MMF stands for multi-modal fusion baseline. [2] Sup. stands for supervised baseline.

Results on Africa Crop Type Mapping: In Table 3, we report segmentation results on the Africa Crop Type Mapping test set. This dataset is more challenging due to multiple reasons. First, the pretrain dataset contains less temporal information due to sparse image collection by the Sentinel satellites. Therefore, we see lower mIoU values for this dataset. However, S4 continues to significantly outperform all baselines for different modalities. Without S4's self-supervised multimodal approach, the mIoU drops for both single and multimodel baselines.

We also provide quantitative results on a strong supervised baseline [62] specifically designed for SITS semantic segmentation and crop type mapping. Since the baseline is supervised, it is only trained on labeled data from the fine-tuned set in our experiments. Naturally, the supervised model performs well when given lots of labeled data but performs very poorly in limited labeled scenarios. We observe, however, that S4 still manages to outperform the supervised baseline in all cases, demonstrating its efficacy in extracting useful information from the pretraining dataset.

6.7. Ablation Study

We provide ablation studies using the PASTIS-R test dataset.

Loss Ablation: We measure the individual contribution of different losses used in S4. Table 4 reports the results using the PASTIS-R dataset with both radar and optical inference. In both scenarios, the benefits of jointly optimizing the contrastive and reconstruction losses are higher as the number of labels increases. This demonstrates S4's ability to provide both temporal and spatial alignment benefits in pretraining to improve downstream model performance.

Table 4. Loss Ablation Results.

Ablation	Inference Modality	mIoU
$\mathcal{L}_c$	Optical	52.6
$\mathcal{L}_r$	Optical	52.5
$\mathcal{L}_c$	Radar	53.6
$\mathcal{L}_r$	Radar	53.3
$\mathcal{L}_r + \mathcal{L}_c$	Optical	53.4
$\mathcal{L}_r + \mathcal{L}_c$	Radar	54.6

Modality Ablation: We measure the effect of multiple modalities in Table 5. We compare against a unimodal variant of S4, where our proposed contrastive objective operates over an optical SITS and the same optical SITS with random augmentations, similar to how contrastive loss is used in prior work. We find significant gains in performance when the radar modality is added during training. This holds true for both scenarios of inference modality.

Table 5. Modality Ablation Results.

Ablation	Inference Modality	mIoU
Single Modal	Optical	51.3
Single Modal	Radar	53.8
S4 (Multimodal)	Optical	53.4
S4 (Multimodal)	Radar	54.6

Geographical Ablation: We report ablation results on the PASTIS dataset by pretraining on our curated, unlabeled Africa dataset. Table 6 reports our results for SL (Same Location Pretraining) and DL (Different Location Pretraining). In this setup, we fix our pretraining sets to have the same number of SITS samples for fair comparison. Although we see a dip in performance due to different geographical pretraining locations, which

is less representative of the fine-tuned dataset, the performance drop is limited and still performs well compared with other self-supervised approaches even when they use SL pretraining. This demonstrates the utility of S4 even on data where the geographic location is unknown.

Table 6. Geographical Ablation Results.

Ablation	Inference Modality	mIoU
SL [1] 10%	Optical	33.7
SL 10%	Radar	36.5
SL 100%	Optical	53.4
SL 100%	Radar	54.6
DL [2] 10%	Optical	29.5
DL 10%	Radar	33.4
DL 100%	Optical	48.7
DL 100%	Radar	51.9

[1] short for Same Location pretraining. [2] short for Different Location Pretraining.

6.8. Qualitative Evaluation

In Figure 5, we plot an example of segmentation results of S4 from the test set of the PASTIS-R dataset from models trained with 100% of the labels. Images from the first two rows show model results with optical inference. Images from the last two rows show model results with radar inference. Qualitatively, we can visualize our model's ability to benefit from supervision, as we can segment hard class labels such as the light green ones with very few training examples.

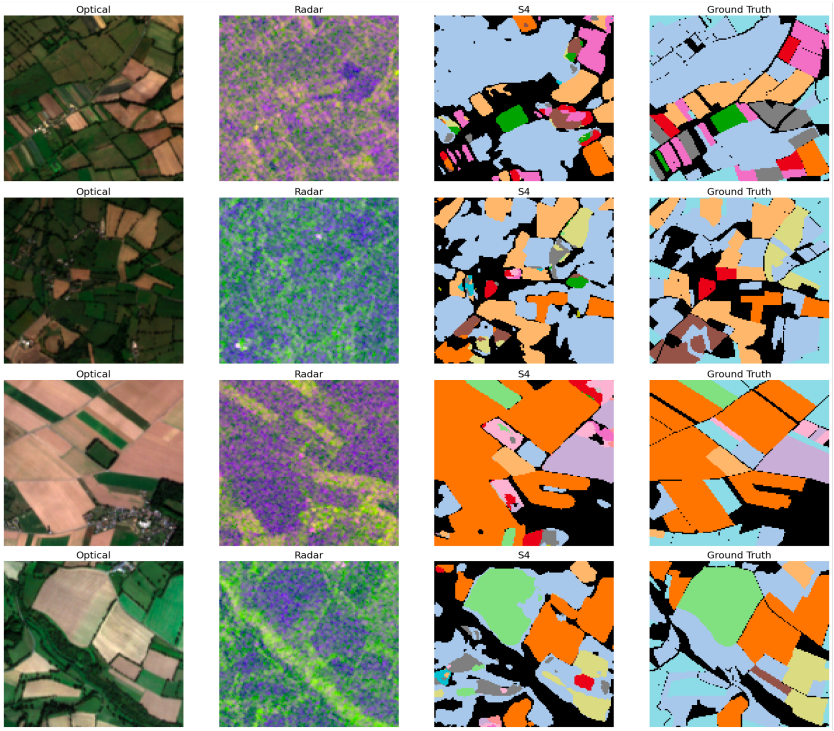

Figure 5. Qualitative results on optical inference. Each row represents a different sample or geographic location from the PASTIS-R dataset for S4's evaluation. The first column (**leftmost**) is a single optical image from the optical SITS. The second column is a single radar image from the radar SITS. The third column is the prediction from S4. The fourth column (**rightmost**) is the ground truth segmentation map.

In Figure 6, we plot example segmentation results of S4 from the test set of the Africa Crop Type Mapping dataset from models trained with 100% of the labels. Just as in Figure 5, the rows show different samples and the columns show the model inputs, outputs, and labels. Note that in this set of visualizations, we omit the background class due to the high background label class imbalance of the dataset. On this dataset, we can examine S4 can largely identify the correct class for most of the pixels associated with the relevant agricultural parcel.

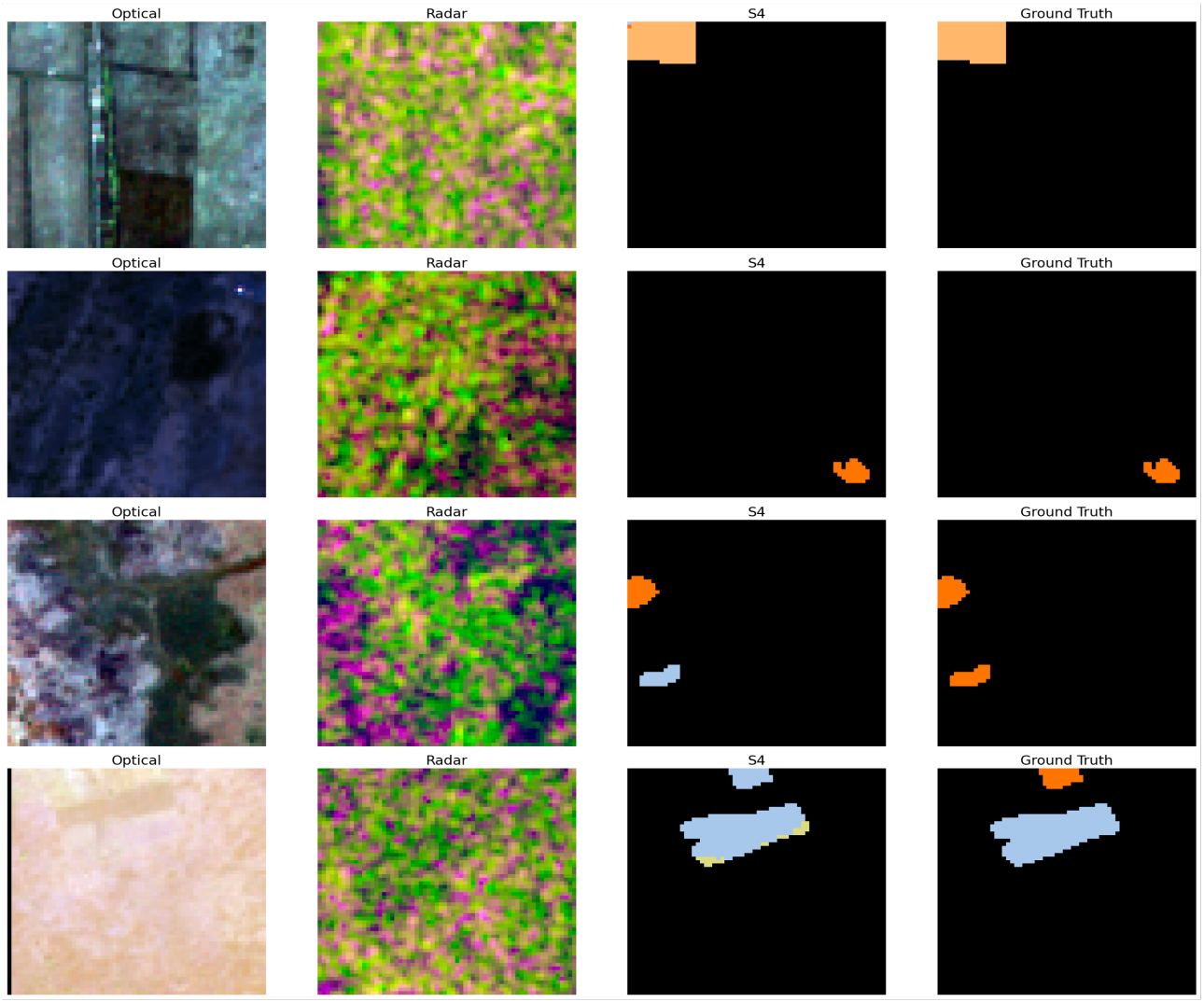

Figure 6. Qualitative results on optical and radar inference. Each row represents a different sample or geographic location from the Africa Crop Type Mapping dataset for S4's evaluation. The first column (**leftmost**) is a single optical image from the optical SITS. The second column is a single radar image from the radar SITS. The third column is the prediction from S4. The fourth column (**rightmost**) is the ground truth segmentation map.

6.9. Reconstruction Visualization Results

We plot visuals of the reconstruction network from S4. Figure 7 visualizes the S4 radar image reconstruction when using optical images as input. Figure 8 visualizes the S4 optical image reconstruction when using the radar images as input. S4 can effectively reconstruct the key shapes in the scene shared across modalities, illustrating its potential effectiveness as a feature extractor. In general, radar imagery is typically stronger in sensing high moisture surfaces such as bodies of water, which may otherwise appear as shadows or clouds in optical images. Some of these key shapes, including the river in Figure 8,

which is a high moisture surface, are able to be successfully extracted in the reconstruction likely due to multimodal training on radar data. Although the contours and shapes of our reconstructions are accurate, there is a significant difference in the color map. The difference in the magnitude of colors stems from the satellite being at different heights, affecting the magnitude of the SAR measurements. As a result, the model is trained on SAR measurements with dramatically varying and noisy magnitudes, creating a discrepancy in the color maps.

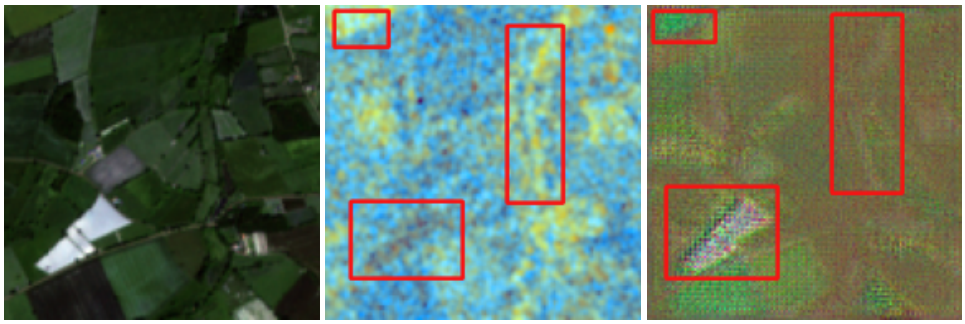

Figure 7. Optical to radar reconstruction of S4 (optical input, radar ground truth, radar reconstruction). Red boxes indicate similar features between reconstruction and output modality images.

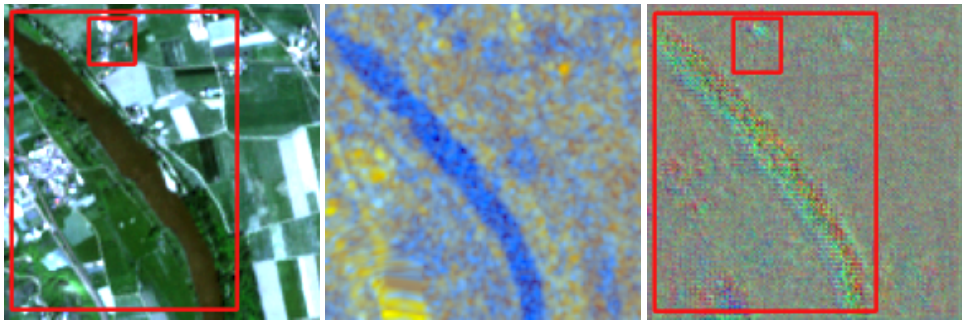

Figure 8. Radar to optical reconstruction of S4 (optical ground truth, radar input, optical reconstruction). Red boxes indicate similar features between reconstruction and output modality images.

6.10. Robustness to Cloud Cover

We analyze S4's ability to tackle the challenge of cloud cover during inference on optical images. We start by dividing the optical SITS in the PASTIS test set into different groups based on the amount of cloudy pixels they contain. We obtain a cloud mask for every image in the PASTIS test set using the S2Cloudless algorithm [69]. We compute the cloud cover ratio as the number of clouded pixels to total pixels in the SITS. After grouping every SITS by cloud cover ratio, we compute the mIoU. Figure 9 reports the mIoU gain of S4 over the CaCo baseline for 100% labels (the scenario when our model has the highest relative improvement). The mIoU improvement is the mean improvement over CaCo baseline. The results illustrate that our approach provides greater gains in the presence of clouds. The radar data can help guide the model to make better predictions on partially clouded data, since the radar images provide insights into how the model can "see through the clouds". The improvement drops for SITS where the cloud cover ratio is greater than 25%, where many images are mostly or even fully occluded by clouds (and contain little useful information).

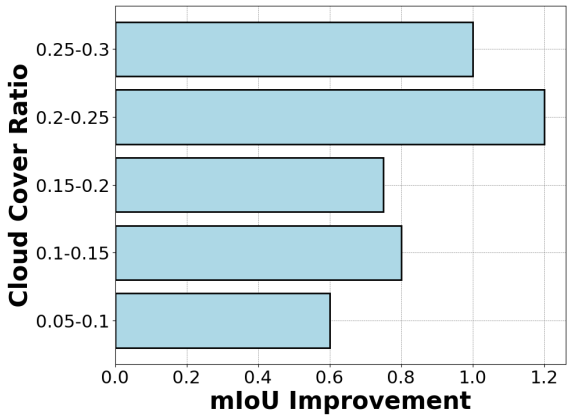

Figure 9. Updated graph cloud cover robustness prediction on optical image inference over PASTIS-R test set.

7. Discussion and Limitations

We discuss some of the limitations of S4 based on the results of our evaluation.

High cloud cover scenarios As illustrated by Figure 9, S4 does provide improved robustness against cloud cover in optical images due to the incorporation of radar data. However, the method still faces performance challenges when cloud cover exceeds 25%. In such cases, the optical information is so heavily occluded that it provides so little useful information, impacting the overall performance of the model. Potential avenues for future work to solve this solution include the incorporation of historical time series data (years before the crop growth phase). Additionally, training on synthetic data generation or painted cloud images may also be useful in addressing the challenges in training with very highly cloud-covered images.

Modalities other than radar or optical Although we explore the benefits of exploring multimodal self-supervision specifically with radar and optical imagining modalities, satellite imaging has ample additional modalities such as infrared and hyperspectral, which we do not explore in this work. We envision that S4's reconstructive and contrastive frameworks can be extended to support these additional modalities by using multiple modality-specific encoders.

Adaptive fusion techniques In our study, we observed that different modalities excel in specific conditions. For instance, optical images perform well when there is low cloud cover, while radar images are more effective in high-moisture environments, such as near lakes and rivers. Currently, our model, S4, treats all multimodal inputs with equal importance during training and segmentation. However, based on our observations, we believe that enhancing S4 to incorporate adaptive multimodal fusion—where the model dynamically adjusts the weight given to each modality based on the specific conditions—could significantly improve overall performance.

Applications of S4 S4's ability to provide accurate segmentation results with limited labeled data makes it beneficial for a variety of different environmental segmentations beyond the agricultural crop segmentation studies discussed above. Some of these applications include land cover classification and ecosystem monitoring, as well as early detection of natural disaster events such as wildfires, mudslides, etc. S4 can be generalized to all these different applications specifically because the amount of actual labeled data needed for its training can be very small.

8. Conclusions

In this paper, we introduced S4, a multimodal self-supervised training framework for satellite image time series segmentation. S4's design can be characterized by (1) multimodal learning, (2) temporal alignment, and (3) a pixel-wise feature space. To enable improved self-supervision for SITS, S4 proposes the following:

- Novel joint pixel-wise space–time contrastive learning;
- Reconstruction loss for multimodal satellite imagery;
- A SITS preprocessing strategy to temporally align SITS across modalities.

We also curate M2S2-SITS, a new multimodal SITS dataset that enables our new geographic ablations and highlights the greater opportunities to benefit from multimodal SITS pretraining. Using our datasets, we demonstrate how S4 can outperform a variety of other self-supervised baselines on the downstream task of semantic segmentation, and we conduct detailed ablations to better illustrate the robustness of our model in specific situations, such as cloud cover and geographic diversity. We envision that S4 will unlock the potential of using satellite imagery for emerging Earth-scale applications like climate monitoring and precision agriculture by reducing the requirement for large, labeled datasets.

Author Contributions: Conceptualization, J.S. and D.V.; methodology, H.Z.; software, R.Y.; validation, S.M., X.D.Z. and B.T.; formal analysis, X.D.Z.; investigation, B.T.; resources, D.V. and H.Z.; data curation, X.D.Z.; writing—original draft preparation, J.S.; writing—review and editing, H.Z.; visualization, J.S.; supervision, D.V.; project administration, J.S.; funding acquisition, D.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSF grant number 2237474, Cisco Systems Inc., IBM-IL Discovery Accelerator Institute, and CloudBank [70] under NSF grant 1925001. Jayanth Shenoy is funded by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE – 1746047.

Data Availability Statement: Link to name's pretrain datasets can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CC0GD0.

Acknowledgments: We thank all members of the Illinois WoW Lab for their input and feedback on multiple versions of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Escher, A. Inside Planet Labs' New Satellite Manufacturing Site. TechCrunch. 2018. Available online: https://techcrunch.com/2018/09/14/inside-planet-labs-new-satellite-manufacturing-site/ (accessed on 1 May 2023).
2. Spire Global Inc. Available online: https://spire.com/ (accessed on 12 November 2022).
3. Castelletti, D.; Farquharson, G.; Stringham, C.; Duersch, M.; Eddy, D. Capella Space First Operational SAR Satellite. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021.
4. Bamber, J. Five Revealing Satellite Images Show How Fast Our Planet Is Changing. Available online: https://www.weforum.org/agenda/2021/06/this-is-why-satellites-are-so-vital-for-protecting-the-health-of-our-planet/ (accessed on 1 May 2023).
5. Macaulay, T. AI Detects Plastics in the Oceans by Analyzing Satellite Images. The Next Web. 2020. Available online: https://thenextweb.com/news/ai-detects-plastics-in-the-oceans-by-analyzing-satellite-images (accessed on 2 October 2023).
6. Mueller, H.; Groeger, A.; Jonathan, H.; Matranga, A.; Serrat, J. Monitoring war destruction from space using machine learning. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2025400118. [CrossRef] [PubMed]
7. Wensley, S. The Power of Satellite Imagery in Agriculture & Farming, 2022. Available online: https://farmtogether.com/learn/blog/the-power-of-satellite-imagery-in-agriculture (accessed on 7 December 2023).
8. Sexton, J. Managing the World's Natural Resources with Earth Observation. 2022. Available online: https://aws.amazon.com/blogs/publicsector/managing-worlds-natural-resources-earth-observation/ (accessed on 19 May 2023).
9. Garnot, V.S.F.; Landrieu, L. Panoptic Segmentation of Satellite Image Time Series With Convolutional Temporal Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
10. Oehmcke, S.; Chen, T.H.K.; Prishchepov, A.V.; Gieseke, F. Creating Cloud-Free Satellite Imagery from Image Time Series with Deep Learning. In Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Seattle, WA, USA, 3 November 2020.
11. Mall, U.; Hariharan, B.; Bala, K. Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery. In Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, Virtual, 28 November 2022.

12. Karaman, K.; Sainte Fare Garnot, V.; Wegner, J.D. Deforestation Detection in the Amazon with Sentinel-1 SAR Image Time Series. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Cairo, Egypt, 2–7 September 2023; pp. 835–842.
13. Adorno, B.V.; Körting, T.S.; Amaral, S. Contribution of time-series data cubes to classify urban vegetation types by remote sensing. *Urban For. Urban Green.* **2023**, *79*, 127817. [CrossRef]
14. Rustowicz, R.M.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019.
15. Bianchetti, R.A. In Sea of Satellite Images, Experts' Eyes Still Needed. 2021. Available online: https://theconversation.com/in-sea-of-satellite-images-experts-eyes-still-needed-53192 (accessed on 5 March 2022).
16. DiBiase, D. Multispectral Imaging from Space. 2020. Available online: https://www.e-education.psu.edu/natureofgeoinfo/node/1899 (accessed on 16 July 2024).
17. Wylie, D.P.; Menzel, W.P. Two Years of Cloud Cover Statistics Using VAS. *J. Clim.* **1989**, *2*, 380–392. [CrossRef]
18. Stubenrauch, C.J.; Rossow, W.B.; Kinne, S.; Ackerman, S.; Cesana, G.; Chepfer, H.; Girolamo, L.D.; Getzewich, B.; Guignard, A.; Heidinger, A.; et al. Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1031–1049. [CrossRef]
19. Zhao, M.; Olsen, P.; Chandra, R. Seeing Through Clouds in Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4704616. [CrossRef]
20. Ziemnicki, P. Optics or Radars? What Is Better for the Earth Observation Purposes? Defence24. 2018. Available online: https://defence24.com/technology/optics-or-radars-what-is-better-for-the-earth-observation-purposes (accessed on 15 March 2022).
21. Union of Concerned Scientists Satellite Database. Union of Concerned Scientists Satellite Database. 2023. Available online: https://www.ucsusa.org/resources/satellite-database (accessed on 12 November 2023).
22. Vasisht, D.; Shenoy, J.; Chandra, R. L2D2: Low Latency Distributed Downlink for LEO Satellites. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference, Virtual Event, 23–27 August 2021.
23. Cattoi, A.; Bruzzone, L.; Haensch, R. Transcoding-based pre-training of semantic segmentation networks for PolSAR images. In Proceedings of the European Conference on Synthetic Aperture Radar, Leipzig, Germany, 25–27 July 2022.
24. Gulyanon, S.; Limprasert, W.; Songmuang, P.; Kongkachandra, R. Data Generation for Satellite Image Classification Using Self-Supervised Representation Learning. *arXiv* **2022**, arXiv:2205.14418.
25. Zhang, M.; Singh, H.; Chok, L.; Chunara, R. Segmenting Across Places: The Need for Fair Transfer Learning With Satellite Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 18–24 June 2022.
26. Garnot, V.S.F.; Landrieu, L.; Giordano, S.; Chehata, N. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020.
27. Rudner, T.G.J.; Rußwurm, M.; Fil, J.; Pelich, R.; Bischke, B.; Kopačková, V.; Biliński, P. Multi3Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
28. Manas, O.; Lacoste, A.; GiroiNieto, X.; Vazquez, D.; Rodriguez, P. Seasonal Contrast: Unsupervised Pre-Training From Uncurated Remote Sensing Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
29. Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-Aware Self-Supervised Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
30. He, Y.; Wang, D.; Lai, N.; Zhang, W.; Meng, C.; Burke, M.; Lobell, D.B.; Ermon, S. Spatial-Temporal Super-Resolution of Satellite Imagery via Conditional Pixel Synthesis. In Proceedings of the Neural Information Processing Systems, Virtual, 6–14 December 2021.
31. Ramsauer, J. Radar vs. Optical: Optimising Satellite Use in Land Cover Classification. 2020. Available online: https://ecologyforthemasses.com/2020/05/27/radar-vs-optical-optimising-satellite-use-in-land-cover-classification/ (accessed on 12 June 2023).
32. Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; Ermon, S. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 Novemeber–9 December 2022; Volume 35.
33. Mall, U.; Hariharan, B.; Bala, K. Change-Aware Sampling and Contrastive Learning for Satellite Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
34. Fuller, A.; Millard, K.; Green, J.R. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Curran Associates Inc.: Red Hook, NY, USA, 2024.
35. Sentinel Missions. 2021. Available online: https://sentinel.esa.int/web/sentinel/missions (accessed on 1 May 2023).

36. Shang, R.; Liu, M.; Lin, J.; Feng, J.; Li, Y.; Stolkin, R.; Jiao, L. SAR Image Segmentation Based on Constrained Smoothing and Hierarchical Label Correction. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5102216. [CrossRef]
37. Poodanchi, M.; Akbarizadeh, G.; Sobhanifar, E.; Ansari-Asl, K. SAR image segmentation using morphological thresholding. In Proceedings of the 2014 6th Conference on Information and Knowledge Technology (IKT), Shahrood, Iran, 27–29 May 2014.
38. Galland, F.; Nicolas, J.M.; Sportouche, H.; Roche, M.; Tupin, F.; Refregier, P. Unsupervised Synthetic Aperture Radar Image Segmentation Using Fisher Distributions. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2966–2972. [CrossRef]
39. Ebel, P.; Meraner, A.; Schmitt, M.; Zhu, X.X. Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5866–5878. [CrossRef]
40. Wang, L.; Xu, X.; Yu, Y.; Yang, R.; Gui, R.; Xu, Z.; Pu, F. SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks. *IEEE Access* **2019**, *7*, 129136–129149. [CrossRef]
41. Zhang, Y.; Zhao, C.; Wu, Y.; Luo, J. Remote sensing image cloud removal by deep image prior with a multitemporal constraint. *Opt. Contin.* **2022**, *1*, 215–226. [CrossRef]
42. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
43. Sainte Fare Garnot, V.; Landrieu, L.; Chehata, N. Multi-Modal Temporal Attention Models for Crop Mapping from Satellite Time Series. *ISPRS J. Photogramm. Remote Sens.* **2021**, *187*, 294–305. [CrossRef]
44. Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.; Hoderlein, A.P.; Şenaras, C.; Davis, T.; Cremers, D.; et al. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
45. Wang, Y.; Zhuo, W.; Li, Y.; Wang, Z.; Ju, Q.; Zhu, W. Fully Self-Supervised Learning for Semantic Segmentation. *arXiv* **2022**, arXiv:2202.11981.
46. Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; Rueckert, D. Self-Supervision with Superpixels: Training Few-shot Medical Image Segmentation without Annotation. *arXiv* **2020**, arXiv:2007.09886.
47. Agastya, C.; Ghebremusse, S.; Anderson, I.; Reed, C.; Vahabi, H.; Todeschini, A. Self-supervised Contrastive Learning for Irrigation Detection in Satellite Imagery. *CoRR* **2021**.
48. Zhang, T.; Qiu, C.; Ke, W.; Süsstrunk, S.; Salzmann, M. Leverage Your Local and Global Representations: A New Self-Supervised Learning Strategy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
49. Zou, Y.; Zhang, Z.; Zhang, H.; Li, C.L.; Bian, X.; Huang, J.B.; Pfister, T. PseudoSeg: Designing Pseudo Labels for Semantic Segmentation. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
50. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, 2020, Proceedings of Machine Learning Research, Virtual, 13–18 July 2020.
51. Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.K.; Li, X.; Guan, C. Self-supervised Contrastive Representation Learning for Semi-supervised Time-Series Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *12*, 15604–15618. [CrossRef]
52. Wu, H.; Wang, X. Contrastive Learning of Image Representations With Cross-Video Cycle-Consistency. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
53. Xie, Z.; Lin, Y.; Zhang, Z.; Cao, Y.; Lin, S.; Hu, H. Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
54. Zhong, Y.; Yuan, B.; Wu, H.; Yuan, Z.; Peng, J.; Wang, Y.X. Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
55. Liu, S.; Zhi, S.; Johns, E.; Davison, A. Bootstrapping Semantic Segmentation with Regional Contrast. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
56. Chaitanya, K.; Erdil, E.; Karani, N.; Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020.
57. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer International Publishing: Cham, Switzerland, 2015.
58. Rustowicz, R.M.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019.
59. Cicek, O.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger Olaf, E.S.; Joskowicz, L.; Sabuncu, M.R.; Unal, G.; Wells, W. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016, Athens, Greece, 17–21 October 2016.
60. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, Lille, France, 6–11 July 2015.
61. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *CoRR* **2018**.

62. Chamorro Martinez, J.A.; Cué La Rosa, L.E.; Feitosa, R.Q.; Sanches, I.D.; Nigri Happ, P. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 188–201. [CrossRef]
63. Microsoft. FarmVibes.AI: An AI Platform for Agriculture. GitHub Repository. 2023. Available online: https://github.com/microsoft/farmvibes-ai (accessed on 10 October 2023).
64. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
65. Shaar, F.; Yılmaz, A.; Topcu, A.E.; Alzoubi, Y.I. Remote Sensing Image Segmentation for Aircraft Recognition Using U-Net as Deep Learning Architecture. *Appl. Sci.* **2024**, *14*, 2639 [CrossRef]
66. Zhang, W.; Zhang, H.; Zhao, Z.; Tang, P.; Zhang, Z. Attention to Both Global and Local Features: A Novel Temporal Encoder for Satellite Image Time Series Classification. *Remote Sens.* **2023**, *15*, 618. [CrossRef]
67. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv* **2020**, arXiv:2003.04297.
68. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
69. Zupanc, A. Improving Cloud Detection with Machine Learning. 2019. Available online: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13 (accessed on 3 August 2022).
70. Norman, M.; Kellen, V.; Smallen, S.; DeMeulle, B.; Strande, S.; Lazowska, E.; Alterman, N.; Fatland, R.; Stone, S.; Tan, A.; et al. CloudBank: Managed Services to Simplify Cloud Access for Computer Science Research and Education. In Proceedings of the Practice and Experience in Advanced Research Computing, Boston, MA, USA, 18–22 July 2021.