Self-supervised Vision Transformers for Land-cover Segmentation and Classification

Linus Scheibenreif* Joëlle Hanna* Michael Mommert Damian Borth
AIML Lab, School of Computer Science, University of St.Gallen
Rosenbergstrasse 30, St. Gallen, Switzerland
{firstname}.{lastname}@unisg.ch

Abstract

Transformer models have recently approached or even surpassed the performance of ConvNets on computer vision tasks like classification and segmentation. To a large degree, these successes have been enabled by the use of large-scale labelled image datasets for supervised pre-training. This poses a significant challenge for the adaption of vision Transformers to domains where datasets with millions of labelled samples are not available. In this work, we bridge the gap between ConvNets and Transformers for Earth observation by self-supervised pre-training on large-scale unlabelled remote sensing data[1]*. We show that self-supervised pre-training yields latent task-agnostic representations that can be utilized for both land cover classification and segmentation tasks, where they significantly outperform the fully supervised baselines. Additionally, we find that subsequent fine-tuning of Transformers for specific downstream tasks performs on-par with commonly used ConvNet architectures. An ablation study further illustrates that the labelled dataset size can be reduced to one-tenth after self-supervised pre-training while still maintaining the performance of the fully supervised approach.*

1. Introduction

The identification of land-cover characteristics from satellite imagery is one of the key objectives in the remote sensing domain. With an increasing number of Earth observation satellites in orbit, the amount of available remote sensing data is steadily growing. This abundance of data makes it possible to address a wide range of problems in Earth observation with statistical learning methods that benefit from large datasets. However, while raw satellite data is available in large quantities today, land cover labels are comparatively scarce and must be obtained through a manual annotation process that is prohibitively expensive for more than a small fraction of the existing satellite data.

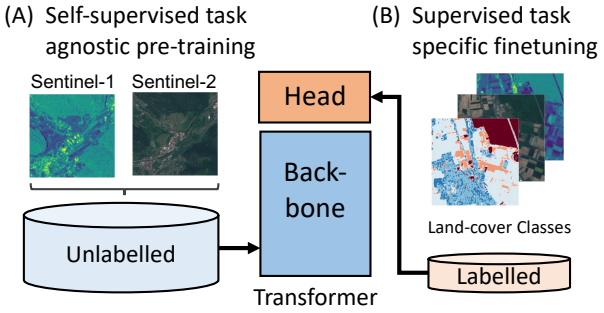


Figure 1. We propose to use large datasets of unlabelled remote sensing data for self-supervised pre-training of vision Transformers. After self-supervised training of the backbone (A), the model and task-specific head can be fine-tuned on much smaller labelled datasets for different downstream tasks (B).

In parallel to the general computer vision (CV) community, deep learning techniques such as Convolutional Neural Networks (ConvNets) have become the state-of-the-art tools for a range of tasks in remote sensing over the last years [41]. More recently, variations of the Transformer architecture [35], which was originally devised for sequential data and has led to breakthroughs in natural language processing, have matched ConvNet performance on important CV benchmark tasks like ImageNet classification [10].

In this work, we propose to combine Transformer-based computer vision approaches with self-supervised learning (SSL) for the remote sensing domain. This enables us to pre-train Transformer models on large amounts of unlabelled satellite imagery with a contrastive self-supervised training setup tailored to multi-modal remote sensing data. This facilitates training of large and data intensive Transformer models when only a small amount of labelled data is available. Additionally, it allows us to use the same backbone model for tasks such as classification or segmentation by changing only the model head. We utilize

*Both authors contributed equally to this work
[1] https://github.com/HSG-AIML/SSLTransformerRS

the shifting-window vision Transformer architecture (Swin Transformer) [17] with a contrastive data fusion SSL strategy [25] and evaluate this pipeline with single-label classification, multi-label classification and semantic segmentation (i.e. pixel-wise classification) as downstream tasks.

We summarize the contributions of this work as follows:

- We show that vision Transformers combined with self-supervised pre-training provide an effective approach for deep learning in the remote sensing domain, surpassing ConvNet performance in some settings.
- We show that latent representations derived through self-supervised pre-training and subsequent supervised fine-tuning are task agnostic and can be utilized for both land cover classification *and* segmentation.
- Our work further illustrates that SSL in combination with vision Transformers or ConvNets can yield large performance gains (up to +30% over supervised baselines) across different downstream tasks when fine-tuned with labelled data.
- In an ablation study on fine-tuning self-supervised models with different amounts of labelled data we demonstrate that fully supervised approaches can be outperformed with as little as 10% of labelled data through SSL.

2. Related Work

2.1. Self-supervised Learning

Self-supervised learning is a branch of machine learning that aims to learn data representations from unlabelled datasets. The literature on self-supervised deep learning rapidly expanded in the last years, following the success of methods like Word2Vec [19] or GloVe [22] in Natural Language Processing (NLP). Consequently, similar approaches were also adopted in the vision domain. Most SSL methods for images either employ pretext tasks or the principle of contrastive learning. **Pre-text based** methods utilize inherent properties of data samples to construct prediction tasks for training. These tasks include the prediction of future states from earlier states in sequential data (e.g. audio [34] or text [38]), as well as colorization of artificially gray-scaled images [40], jigsaw tasks [20] or rotation prediction [12] in image data. The second popular SSL strategy, **contrastive learning**, trains neural networks to learn the relationships between different data points by distinguishing among them. Fundamentally, this approach aims to structure the latent space such that embeddings of similar samples are close together, while those of dissimilar samples are far apart [13]. Different techniques for contrastive learning on image data have been proposed [6,14,21,32,37] and recently even surpassed the performance of supervised training for ImageNet classification [7].

The remote sensing community has adapted SSL techniques to learn meaningful representations of satellite imagery in multiple works. Pretext tasks like inpainting and the prediction of relative positions for image patches have been utilized with different satellite datasets and compared to contrastive estimation [30]. The authors of Seasonal Contrast [18] obtain positive samples for contrastive learning from satellite images of the same locations at different points in time together with augmented data points. Additionally, the data is mapped into multiple embedding subspaces, which results in representations with invariances with respect to different transformations. Work based on the momentum contrast SSL technique [14] also utilizes satellite imagery of given locations at different points in time as temporal positives in contrastive learning, but combines it with location classification in a multi-task framework [1]. The Contrastive Multiview Coding [32] framework for SSL has also been adapted to remote sensing data [27,29]. These works explored the potential of multi-spectral imaging data in SSL with different band and sensor combinations, as well as cross-dataset transfer of pre-trained models. A different strategy for self-supervised pre-training specific to Transformer models is proposed in [39]. This approach exploits the temporal structure of satellite imagery and frames the prediction of artificially corrupted observations in a satellite image timeseries as pretext task. Most relevant to our work, [8] propose the use of a UNet-like architecture to obtain pixel-wise representations of multi-modal remote sensing data through contrastive learning. Similarly, [24] combines three different unsupervised loss functions, including a contrastive loss, on multi-modal remote sensing data to pre-train a change detection model. Our work builds on the multi-modal SSL approach from [25] which utilizes image pairs from different satellite instruments as positive pairs. However, we move beyond ConvNets and the proposed vision Transformer backbone enables our technique to learn task agnostic representations for classification and segmentation downstream tasks in a self-supervised fashion.

2.2. Vision Transformers

Transformers in NLP Transformers have revolutionized the field of Natural Language Processing, being the state of the art for several NLP tasks [3, 9], and slowly replacing RNN-based models. Unlike RNNs, Transformers use attention mechanisms that allow them to process sequential data without necessarily following the order of the sequence, and capture long-range dependencies between tokens in a sequence (e.g. words in a sentence).

Transformers in vision In computer vision, however, convolutional architectures remain dominant. Inspired by the successes of Transformers in NLP, several works [2,31] attempt to combine ConvNet-like architectures with atten-

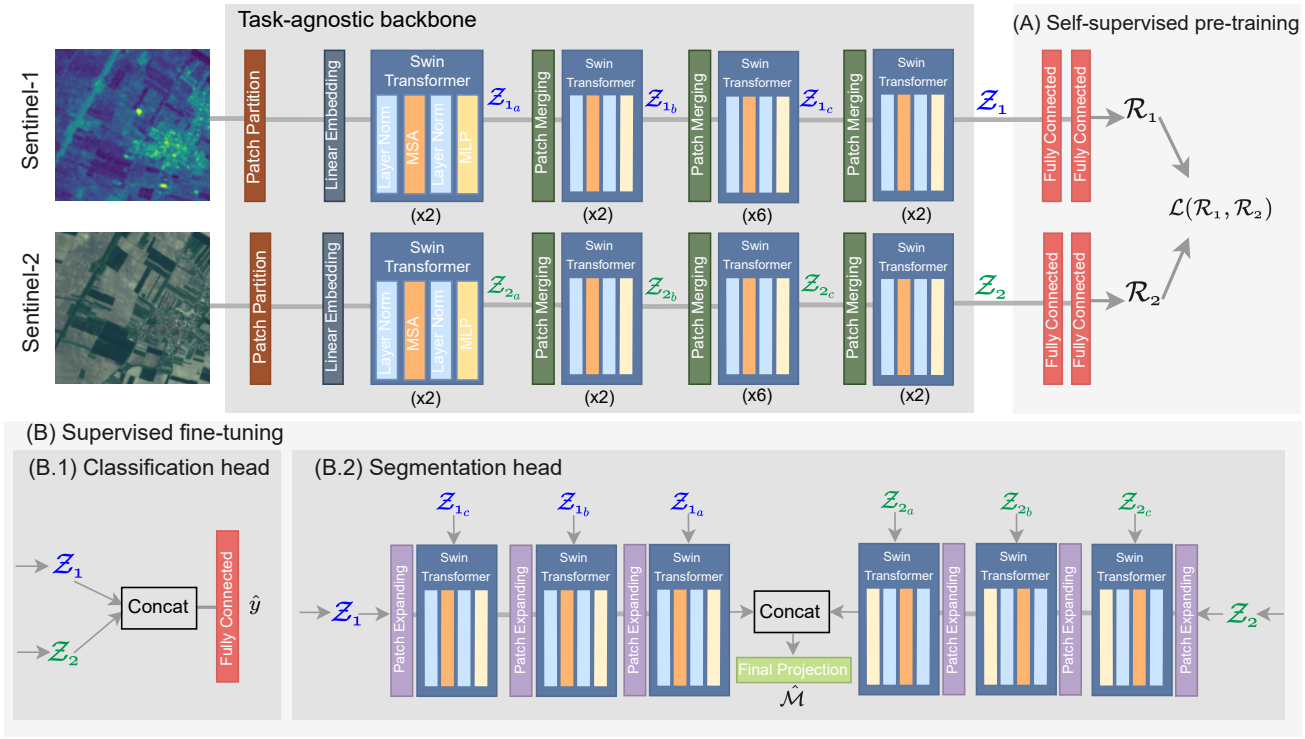


Figure 2. Network architecture for our proposed method. The training is performed in two stages. First, for Sentinel-1 and Sentinel-2 input pairs, we train a unique backbone consisting of two streams of Swin Transformers (Section 3.2), using a self-supervised contrastive approach (A) (Section 3.1). Then, for the supervised learning of both tasks (B), the two outputs of the backbone ($\mathcal{Z}_1, \mathcal{Z}_2$) are fed into the classification head (B.1) and the segmentation head (B.2). Intermediate representations ($\mathcal{Z}_{1_i}$ and $\mathcal{Z}_{2_i}$) are also used for the segmentation head. The final projection layer of the segmentation head consists of an up-sampling layer followed by a 1x1 convolutional layer.

tion mechanisms. Moreover, ConvNet-Transformer hybrid models began to emerge, using convolutions for the backbone, and appending a Transformer for the task head [5]. Vision Transformer (ViT) [10] is the first to replace convolutions entirely and proposes to apply a standard Transformer directly to images, with as little modification as possible, by dividing an image into patches and treating these patches the same as tokens (words) in an NLP application. This Transformer applies self-attention on a global receptive field, and has a quadratic computational complexity to the number of token. After being pre-trained on a large-scale labelled dataset, ViT obtained competitive results on ImageNet, but has some limitations on dense pixel-level predictions (e.g., semantic segmentation), failing to capture the fine details due to its fixed patch size. For these reasons, the Swin Transformer [17], a variant of the vision Transformer, proposes a hierarchical way of processing the image, with the goal of achieving scale-invariant representation. It uses the same concept of dividing the image into patches, but groups non-overlapping patches into windows and applies self-attention within each window. A shifted-window scheme is used to allow for cross-window attention connection, which provides a better global representation. The Swin Transformer achieves a better speed-accuracy tradeoff compared to other architectures of the same complexity, for many downstream tasks such as image classification and object detection. We use the Swin Transformer for single-label and multi-label classification.

Semantic segmentation with Transformers Semantic segmentation consists of classifying each pixel of an image into a label. This prediction task requires modeling the interactions between pixels to generate refined representations, which is not straightforward using Transformers. Recently, [36] proposed a pure attention-based model for semantic segmentation and introduced the position-aware axial attention layer that propagates information densely and efficiently along the axes of height and width sequentially. While this work follows a ConvNet-like design by gradually reducing the spatial dimension of feature maps, others have proposed complete encoder-decoder architectures based on Transformers [28]. Here, we do not propose a new segmentation network, instead we study the advantages of using a task-agnostic representation obtained by self-supervised pre-training of Swin Transformers with multimodal inputs, which greatly improves the segmentation task.

3. Methods

Figure 2 illustrates our overall approach, which we detail in the following.

3.1. Self-supervised Learning

In this work we propose the use of contrastive SSL for pre-training of Transformer models on remote sensing data. A key property of remote sensing data is that data obtained by a multitude of sensors aboard different satellites close in time may be available for the same location. This property can be exploited to generate multiple views of the same scene in an augmentation free manner. The resulting SSL strategy uses satellite imagery from different sensors for the same location as positive image pairs and images from other sensors and locations as negative samples [25]. This approach enables contrastive SSL without the use of strong random augmentations and with dedicated encoders for each modality (i.e., no weight-sharing), contrary to standard practice in SSL methods for natural images [6]. The contrastive loss is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}, \quad (1)$$

where $\mathcal{R}_i$ and $\mathcal{R}_j$ are representations of a positive pair
$\mathcal{R}_k$ are negative (contrastive) representations
$\text{sim}(\cdot, \cdot)$ is the dot product
$\mathbb{1}$ is the indicator function
τ is a so-called temperature parameter

The vector representations $\mathcal{R}$ are obtained by passing the samples through the backbone models (see Figure 2).

3.2. Swin Transformers

Swin Transformers [17] are vision Transformers designed as backbones for all kinds of visual tasks, such as image classification, object detection, semantic segmentation. They owe their success to their scale invariance property, which allows them to be used for both high-level and dense predictions. Their strength results from their shifted window approach: a window contains non-overlapping squared patches, and self-attention is calculated locally, within each window, before shiting. As a result, the computational complexity is reduced compared to the standard transformer. The backbone (encoder) is made up of 4 building blocks, where each block is constructed by connecting a patch merging layer and several Swin Transformer blocks. A Swin Transformer block is composed of Multi-head Self-Attention (MSA), followed by a 2-layer Multi-Layer Perceptron (MLP). A Layer Norm (LN) is applied before each MSA and MLP. The first Swin Transformer block uses a standard window partitioning configuration to locally compute self-attention from uniformly separated windows. The next one adopts a window configuration shifted by a certain pixel offset relative to the previous layer, and so on.
Regarding the classification task, we concatenate the two representations coming out of the backbone (see Figure 2), and feed them into a fully connected layer. We detail in the next section the architecture used for the segmentation task.

3.2.1 SwinUNet

We use SwinUNet [4] as the basic framework for the semantic segmentation task. The architecture is similar to a UNet [23]; it consists of an encoder, bottleneck, decoder and skip connections. While in a traditional UNet the encoder and decoder are symmetric blocks of convolutional and max-pooling layers, in the SwinUNet they are symmetric blocks of Swin Transformer (see Section 3.2). To generate the hierarchical representations of the features, the Swin Transformer blocks are preceded by a patch merging (downsampling) step in the contracting path and a patch expanding (upsampling) step in the expansive path. Features extracted after blocks of the same spatial dimension are merged via skip connections. Finally, to restore the initial resolution (height and width) of the input, an upsampling operation if performed on the last patch expansion layer.
Dual SwinUNet: We propose two separate SwinUNet network streams with identical architectures (see previous description) to process pairs of Sentinel-1 and Sentinel-2 data in parallel, in a dual stream concept (see Figure 2). The features extracted after the last patch expansion layer of each stream are concatenated along the channel dimension and fed into the last pointwise convolution (1x1 kernel) to produce the segmentation predictions at the pixel-level. The purpose of the dual SwinUNet is to capture the information belonging to each of the two inputs first, before merging the two representations for the final decision.

4. Data

This work applies SSL and vision Transformers on paired satellite data from the Sentinel-1 and Sentinel-2 missions of the European Space Agency's Copernicus program.
Sentinel-1 is a satellite mission for Earth observation with Synthetic Aperture Radar (SAR) [33]. It provides medium resolution (∼10m) C-band SAR measurements with dual polarisation, which enables data acquisition during night or through cloud cover. The two Sentinel-1 satellites are in sun-synchronous orbits with a 12 day repeat cycle. This work utilizes VV and VH polarized data from the main Interferometric Wide-swath mode.
Sentinel-2 is a constellation of two sun-synchronous satellites for optical Earth observation at medium resolution [11]. The on-board instrument provides multi-spectral observations in the visible, near- and short-wave infrared in 13 bands with up to 10m pixel resolution. The two Sentinel-

Table 1. Results for single- and multi-label classification downstream tasks with ResNet50 and Swin Transformer backbones. S1 and S2 models are trained solely on data from Sentinel-1 or Sentinel-2 without data fusion. EarlyF. and LateF. perform Sentinel-1/2 data fusion at the model input or embedding level. FT corresponds to fine-tuning the pre-trained self-supervised model for the downstream task, whereas "Frozen" models only train task-specific heads. Accuracies are reported with their standard deviations from 5 runs. The best performing model for each class is highlighted in bold. The frequency of each class in the training set is noted in parentheses in the Class column, reflecting class imbalances.

		ResNet50						**Swin Transformer**					
		Baselines				SSL		Baselines				SSL	
	Class	S1	S2	EarlyF.	LateF.	FT	Frozen	S1	S2	EarlyF.	LateF.	FT	Frozen
Single-label (Accuracy %)	Forest (8%)	34 ± 4	13 ± 4	12 ± 3	15 ± 3	$\mathbf{65 \pm 8}$	34 ± 2	8 ± 1	19 ± 3	30 ± 2	30 ± 3	17 ± 6	35 ± 2
	Shrubl. (4%)	24 ± 2	32 ± 5	31 ± 5	34 ± 1	56 ± 11	$\mathbf{73 \pm 1}$	9 ± 2	41 ± 4	42 ± 1	46 ± 2	47 ± 9	57 ± 2
	Grassl. (10%)	$\mathbf{10 \pm 1}$	4 ± 2	7 ± 5	7 ± 3	9 ± 6	2 ± 1	1 ± 0	1 ± 0	0 ± 0	3 ± 1	7 ± 5	5 ± 1
	Wetl. (18%)	35 ± 6	21 ± 5	21 ± 8	14 ± 3	15 ± 8	60 ± 2	44 ± 4	10 ± 3	2 ± 1	10 ± 3	22 ± 8	$\mathbf{65 \pm 4}$
	Cropl. (16%)	47 ± 4	30 ± 3	35 ± 5	$\mathbf{58 \pm 3}$	45 ± 8	51 ± 0	20 ± 2	30 ± 1	33 ± 1	39 ± 4	55 ± 10	54 ± 2
	Urban (6%)	82 ± 4	74 ± 7	88 ± 4	82 ± 3	$\mathbf{95 \pm 1}$	98 ± 0	85 ± 1	54 ± 2	89 ± 2	88 ± 2	94 ± 2	93 ± 1
	Barren (2%)	29 ± 5	26 ± 4	26 ± 4	27 ± 4	39 ± 3	37 ± 2	33 ± 6	40 ± 4	35 ± 4	35 ± 2	48 ± 4	$\mathbf{50 \pm 2}$
	Water (36%)	96 ± 2	91 ± 9	93 ± 7	96 ± 1	99 ± 1	95 ± 0	97 ± 0	78 ± 3	97 ± 0	97 ± 0	$\mathbf{99 \pm 0}$	98 ± 0
	Overall	54 ± 2	42 ± 2	45 ± 1	52 ± 1	$\mathbf{67 \pm 2}$	60 ± 1	40 ± 1	40 ± 2	52 ± 1	53 ± 1	55 ± 3	60 ± 1
	Average	43 ± 1	36 ± 2	39 ± 1	42 ± 1	53 ± 1	56 ± 1	37 ± 1	34 ± 1	41 ± 1	44 ± 0	49 ± 2	$\mathbf{57 \pm 1}$
Multi-label (F1 Score %)	Forest (20%)	63 ± 2	59 ± 7	67 ± 6	73 ± 4	$\mathbf{79 \pm 2}$	65 ± 2	18 ± 2	55 ± 2	69 ± 2	48 ± 3	65 ± 7	69 ± 2
	Shrubl. (8%)	24 ± 2	32 ± 3	35 ± 2	35 ± 1	$\mathbf{40 \pm 3}$	32 ± 1	13 ± 2	32 ± 2	32 ± 1	29 ± 1	38 ± 3	39 ± 0
	Grassl. (27%)	33 ± 8	$\mathbf{53 \pm 5}$	49 ± 4	43 ± 4	47 ± 6	11 ± 3	14 ± 4	51 ± 2	39 ± 7	50 ± 2	18 ± 6	40 ± 3
	Wetl. (35%)	18 ± 2	9 ± 2	10 ± 1	12 ± 2	24 ± 2	16 ± 2	18 ± 1	9 ± 1	9 ± 1	10 ± 1	23 ± 3	$\mathbf{27 \pm 1}$
	Cropl. (23%)	63 ± 1	58 ± 3	60 ± 1	63 ± 1	$\mathbf{70 \pm 3}$	64 ± 1	46 ± 3	56 ± 1	56 ± 2	61 ± 1	69 ± 1	65 ± 1
	Urban (10%)	70 ± 3	55 ± 3	61 ± 2	73 ± 1	80 ± 2	79 ± 1	69 ± 0	51 ± 1	65 ± 2	70 ± 1	77 ± 1	$\mathbf{83 \pm 1}$
	Barren (6%)	27 ± 3	24 ± 2	22 ± 3	22 ± 2	$\mathbf{34 \pm 3}$	25 ± 0	14 ± 2	22 ± 2	25 ± 1	26 ± 2	33 ± 4	32 ± 2
	Water (43%)	95 ± 0	89 ± 2	96 ± 3	96 ± 0	$\mathbf{97 \pm 0}$	93 ± 0	95 ± 0	72 ± 1	94 ± 0	95 ± 0	$\mathbf{97 \pm 0}$	96 ± 1
	Overall	56 ± 2	56 ± 2	59 ± 2	61 ± 1	$\mathbf{67 \pm 1}$	60 ± 0	42 ± 1	51 ± 2	58 ± 1	56 ± 1	60 ± 1	62 ± 1
	Average	49 ± 2	47 ± 2	50 ± 1	52 ± 1	$\mathbf{59 \pm 1}$	48 ± 1	36 ± 1	43 ± 0	49 ± 1	49 ± 0	53 ± 1	56 ± 1

Table 2. Results for segmentation downstream tasks with Swin Transformer backbone. S1 and S2 models are trained solely on data from Sentinel-1 or Sentinel-2 without data fusion. EarlyF. and LateF. perform Sentinel-1/2 data fusion at the model input or embedding level. FT corresponds to fine-tuning the pre-trained self-supervised model for the downstream task, whereas "Frozen" models only train task-specific heads. Per-class accuracies and mean Intersection over Union are reported with their standard deviations from 5 runs. The best performing model for each class is highlighted in bold. Per-class pixel-wise distribution in our training set is mentioned next to each class.

	Baselines								**SwinUNet SSL**					
	UNet				SwinUNet				FT			Frozen		
Class	S1	S2	EarlyF.	LateF.	S1	S2	EarlyF.	LateF.	S1	S2	LateF	S1	S2	LateF.
Forest (9%)	78 ± 0	74 ± 2	80 ± 1	81 ± 1	68 ± 1	72 ± 1	78 ± 1	81 ± 0	78 ± 1	67 ± 0	62 ± 2	70 ± 2	74 ± 1	$\mathbf{84 \pm 2}$
Shrubl. (5%)	17 ± 1	24 ± 1	20 ± 1	22 ± 2	4 ± 1	22 ± 2	23 ± 1	27 ± 0	13 ± 2	40 ± 2	$\mathbf{48 \pm 1}$	14 ± 1	20 ± 2	24 ± 1
Grassl. (12%)	25 ± 2	28 ± 0	38 ± 3	34 ± 2	9 ± 1	19 ± 0	$\mathbf{38 \pm 0}$	14 ± 2	17 ± 0	6 ± 1	6 ± 0	18 ± 2	19 ± 1	23 ± 3
Wetl. (18%)	5 ± 0	4 ± 1	4 ± 0	6 ± 0	5 ± 0	6 ± 3	7 ± 3	3 ± 0	6 ± 2	11 ± 3	$\mathbf{16 \pm 1}$	8 ± 0	6 ± 0	8 ± 1
Cropl. (13%)	57 ± 4	47 ± 1	49 ± 2	50 ± 1	37 ± 2	39 ± 2	44 ± 2	48 ± 2	53 ± 0	$\mathbf{60 \pm 0}$	52 ± 2	51 ± 2	48 ± 2	47 ± 0
Urban (5%)	55 ± 1	47 ± 1	51 ± 2	48 ± 0	37 ± 0	46 ± 1	54 ± 1	57 ± 1	60 ± 1	74 ± 1	$\mathbf{82 \pm 0}$	65 ± 1	58 ± 1	62 ± 2
Barren (3%)	19 ± 1	23 ± 2	$\mathbf{28 \pm 1}$	27 ± 2	16 ± 2	20 ± 0	22 ± 0	18 ± 0	22 ± 2	19 ± 2	36 ± 1	36 ± 1	32 ± 0	$\mathbf{39 \pm 1}$
Water (35%)	97 ± 0	93 ± 2	98 ± 1	98 ± 0	96 ± 1	94 ± 2	98 ± 1	96 ± 2	98 ± 3	99 ± 0	$\mathbf{99 \pm 0}$	98 ± 0	96 ± 1	98 ± 2
Overall	57 ± 0	56 ± 1	57 ± 2	58 ± 1	47 ± 0	53 ± 2	59 ± 1	60 ± 1	52 ± 1	57 ± 2	$\mathbf{63 \pm 0}$	59 ± 0	59 ± 0	62 ± 2
Average	43 ± 1	43 ± 1	46 ± 1	45 ± 1	33 ± 2	39 ± 1	44 ± 2	43 ± 1	43 ± 2	46 ± 1	$\mathbf{51 \pm 1}$	44 ± 2	44 ± 0	48 ± 2
mIoU	32 ± 1	31 ± 2	32 ± 1	31 ± 3	24 ± 3	28 ± 2	32 ± 0	33 ± 1	29 ± 2	35 ± 0	$\mathbf{37 \pm 1}$	31 ± 2	32 ± 1	35 ± 1

2 satellites achieve a revisit rate of 5 days at the equator.

SEN12MS The SEN12MS dataset [26] is a large-scale collection of spatially aligned observation pairs from Sentinel-1 and Sentinel-2. The dataset contains 180,662 observations and covers different geographical areas around the world. All Sentinel-1/2 image pairs are obtained in the same season and pre-processed to a harmonized resolution of 10m for all bands. This work utilizes SEN12MS for self-supervised

pre-training without access to any labels.

DFC2020 The DFC2020 dataset is an extension to SEN12MS constructed for the IEEE GRSS Data Fusion Contest 2020 [16]. This dataset consists of a validation and test sets with 986 and 5,128 paired Sentinel-1/2 observations, respectively. In addition to the satellite imagery, DFC2020 also provides dense (i.e., pixel-level) land-cover annotations for the classes `Forest`, `Shrubland`, `Grassland`, `Wetland`, `Cropland`, `Urban/Built-up`, `Barren` and `Water`. We use the DFC2020 dataset to evaluate the downstream tasks of single- and multi-label classification and semantic segmentation. In this work we use the validation split for training, as in Track 2 of the Contest. We note that this data is highly unbalanced (see Tables 1 and 2). Furthermore, we follow a different objective than the Data Fusion Contest [16] by utilizing its dataset to investigate the use of SSL and vision Transformer models in the remote sensing domain.

5. Experiments and Results

We perform extensive experiments to assess the performance of vision Transformers on three different downstream tasks based on the DFC2020 dataset. The Transformer architectures are compared against different baselines, including commonly used ResNet50 ConvNet models [15]. In particular, we focus on the benefits of SSL and subsequent fine-tuning over training from scratch to leverage the large vision Transformer models on small labelled remote sensing datasets.

5.1. Baselines

Classification Baselines: We use four different data settings for each classification baseline model architecture: Only Sentinel-1 input data, only Sentinel-2 input data, **early Sentinel-1/2 fusion** through concatenation across channel dimension at the data input stage, and **late fusion** by concatenating feature maps derived from Sentinel-1/2 inputs with distinct model backbones before the final classification layer. Besides Swin Transformers, we use ResNet50 as baseline model architecture as it comprises a comparable number of parameters to the Transformer (Swin-t). These models are trained from scratch on the *validation* split of the DFC2020 dataset. Results evaluated on the *test* split are reported in Table 1. Both architectures result in moderately good performance for single-label classification despite the small training dataset. The Swin Transformer achieves the best average accuracy in the late-fusion setting (44$\pm$0 percentage points), while the ResNet50 slightly outperforms the Transformer in early-fusion and uni-modal settings. For multi-label classification (see Table 1 bottom half) the late fusion approach yields the highest F1 Scores for both backbone architectures (52$\pm$1 and 49$\pm$0 percentage points for ResNet50 and Swin Transformer, respectively).

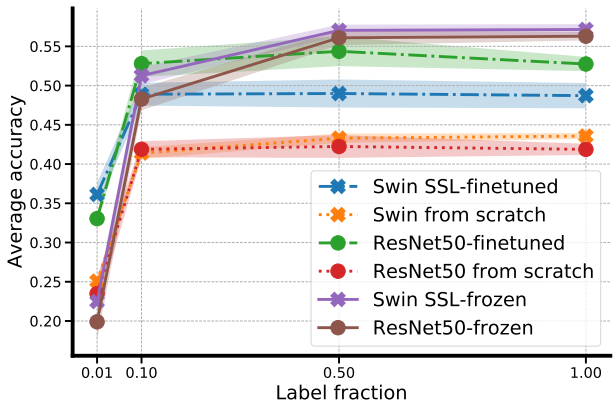
Segmentation Baselines: We use two baseline models to compare the self-supervised model we present for the semantic segmentation task, i) a standard UNet architecture [23] and ii) a SwinUNet (see Section 3.2.1). We train both these baselines from scratch, following the same experimental setup described above. Note that for the late fusion experiment, we use a Dual SwinUNet model (see Section 3.2.1). Results are reported in terms of pixel-wise accuracy and mean Intersection over Union (mIoU) in Table 2. We note that UNet achieves higher pixel accuracy than SwinUNet in both uni-modal and multi-modal configurations, with the best average accuracy reaching 46 ± 1.

5.2. SSL Pre-training with Fine-tuning

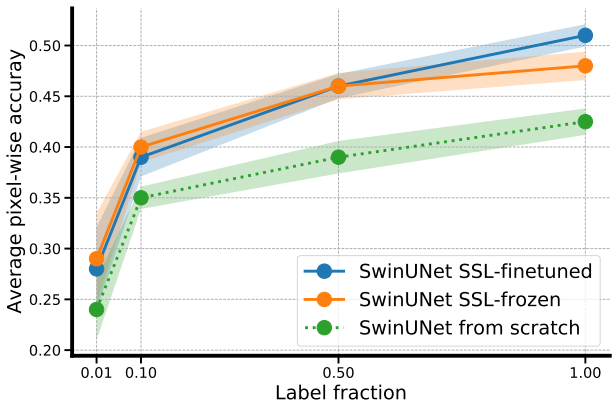
After self-supervised training of the respective model backbone on the SEN12MS dataset ($\sim$3 days on a NVIDIA Tesla V100 GPU), all model parameters (backbone and task-specific head) are fine-tuned for the downstream task of interest. For single-label **classification**, we find that self-supervised pre-training with subsequent fine-tuning strongly outperforms all baseline models (see Table 1). The average accuracy score increases by 10 and 5 percentage points for ResNet50 and Swin Transformer, respectively. This corresponds to a relative increase of 23.3% and 11.4% over the best baselines. We find similar results in the multi-label setting, with relative increases in F1 Scores of 13.5% and 8.2% over the best baselines with the fine-tuned self-supervised ResNet50 and Swin Transformer models. **Segmentation** results are shown in Table 2 for uni-modal (Sentinel-1, Sentinel-2) and multi-modal (late-fusion, see Figure 2) inputs. For the multi-modal case, we note that the average pixel accuracy of our fine-tuned model increased by 8 (+19% relative increase) and 6 (+13%) percentage points, compared to the SwinUNet and the UNet trained from scratch, respectively.

5.3. SSL Pre-training with Frozen Backbone

We investigate to what degree feature maps produced by models trained in a self-supervised manner encode relevant information for land-cover classification and segmentation downstream tasks. To test this, we freeze all parameters of the model backbones and only train the parameters of randomly initialized classification or segmentation heads for each task. Evidently, our SSL strategy extracts meaningful features for land-cover classification. Training a single-label classification head on top of the frozen ResNet50 backbone yields strong performance, and even surpasses the best ResNet50 baseline model by 13 percentage points average accuracy (+30.2%). The frozen Swin Transformer model even yields the best single-label classification performance of all approaches presented in this work with an average accuracy of 57$\pm$1 (+29.5% over best Swin Transformer baseline). The pre-trained and frozen Swin



(a) Swin Transformer and ResNet50 models pre-trained with SSL strongly outperform training from scratch on the classification downstream task.



(b) SwinUNets with pre-training outperform training from scratch when freezing *or* fine-tuning the backbone.

Figure 3. (a) Classification and (b) segmentation results for training on different fractions of labelled data.

Transformer yields similar performance on the multi-label downstream task (+14% average F1 Score over best baseline), even though the frozen ResNet50 fails to outperform the best multi-label classification baseline (-7.7% average F1 Score). This indicates that the self-supervised Transformer model learns more meaningful representations that encode sufficient information to extract multiple class labels with a small classification head.

We observe largely similar behavior for the segmentation task; training only the segmentation head surpasses our two baselines (UNet and SwinUNet) by 5 (+12%) and 3 (+7%) percentage points, respectively, but performs no better than fine-tuning.

5.4. Label Fraction Experiments

We investigate the degree to which SSL can offset the problem of small labelled training datasets. To that end the models are trained with subsets consisting of 50%, 10% and 1% of our training data (corresponding to ∼500, 100 and 10 observations). This results in strongly reduced performance when using only 1% of data (see Figure 3). However, the fine-tuned self-supervised models significantly outperform both the self-supervised models with frozen backbone, and the baseline models trained from random initializations (36 vs. 25 average accuracy points for the fine-tuned and baseline Swin Transformer, respectively). With only 10% of the labelled data, all self-supervised modles outperform the best supervised baselines trained on the entire dataset. The performance rapidly increases with the amount of available labelled training data for all models.

5.5. Implementation Details

We perform extensive experiments across different model backbone architectures, data fusion strategies and downstream tasks. To limit computational cost, hyperparameters of the task-specific fine-tuning experiments are fixed to sensible values a-priori, rather than tuned for every individual experimental setting. The batch size is set to 32, learning rate to $3 \cdot 10^{-6}$ and the number of training epochs to 200. This approach also makes it possible to utilize the full DFC2020 validation set (986 observations with dense land-cover labels) for training as we do not require a validation set for hyperparameter tuning.

Table 3. Comparison of different methods for segmentation. All models are trained using multi-modal input, *SSL-ft* means we are fine-tuning the pre-trained model

	UNet	SwinUNet	SwinUNet SSL-ft.	Ensemble
Avg. Accuracy	0.45	0.43	0.51	**0.53**
Avg. IoU	0.31	0.33	0.37	**0.39**

6. Discussion

Our work highlights the benefits of pre-training Swin Transformer backbones with a contrastive learning approach and subsequently fine-tuning them for different downstream tasks. Following this protocol we observe a significant improvement in performance for each of our downstream tasks over standard fully supervised training.

In the **classification task**, the self-supervised fine-tuned ResNet outperforms the SSL-fine-tuned Swin Transformer in average accuracy by a small margin. This could be explained by the comparatively higher maturity of the ResNet architecture over Transformers in computer vision, leading to better default parameter configurations. Moreover, we observe that the SSL-Swin Transformer with frozen backbone performs better than the SSL-ResNet with frozen backbone, indicating that the Transformer model manages to learn more informative representations for our downstream tasks. This advantage is particularly apparent in multi-label classification, further illustrating that the Swin Transformer manages to extract informative features

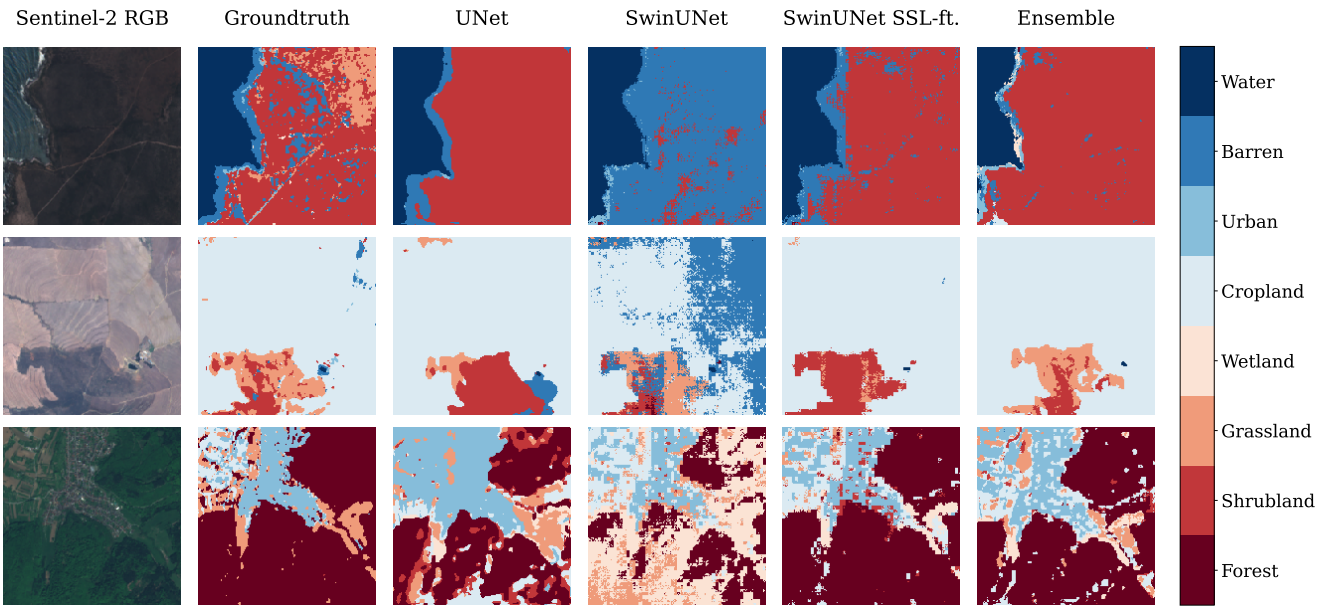


Figure 4. Qualitative comparison of results for 3 different regions. Results from left to right: Sentinel-2 true color (RGB), DFC groundtruth, UNet trained from scratch on fusion of both inputs, SwinUNet trained from scratch on both inputs, SwinUNet fine-tuned on both inputs, and finally an ensemble model of both UNet and SwinUNet (see Section 6).

through self-supervised pre-training. For single label classification, the frozen backbone models performed better than finetuning all parameters, which we attribute to the hyperparameter choice. For the **segmentation task**, we arrive at the same conclusion as above: self-supervised pre-training considerably boosts performances. Nevertheless, we note that the SSL-Swin Transformer with frozen backbone does not perform better than the SSL-fine-tuned one. This may be a result of the segmentation head architecture, which uses skip connections to merge the multi-scale characteristics of the encoder with the upsampled characteristics of the decoder. Therefore, to achieve the best performance, encoder and decoder parameters should be updated simultaneously. The importance of our approach for **data-efficient learning** is further underlined by the results of the label fraction experiment. Across all downstream tasks, our pre-trained and fine-tuned models perform on par with models trained from scratch with as little as 10% of the labelled data. This SSL approach thus opens a path to learning based on very small datasets ($\sim$100 samples), enabling data-efficient applications. In a **qualitative comparison**, we show in Figure 4 some segmentation results. We first observe that the UNet method produces smoother segmentation masks compared to the SwinUNet. On the other hand, the segmentation masks produced by the SwinUNet are much more detailed and accurate. These observations motivated the idea of trying an ensemble of these two methods where we take the average of the predictions of the two models, before computing the final prediction (see Figure 4). Across all our downstream tasks, we note that classes like Grassland or Wetland are commonly misclassified on some images. This is most likely due to our limited and very unbalanced training set (see Tables 1 and 2). Overall, the ensemble model works best, both visually, giving smooth and detailed results, and numerically (see Table 3). This pushes us to explore this direction in future works, with the aim of improving land cover segmentation, taking into account the advantages of each of these methods.

7. Conclusion

This work introduced a self-supervised pre-trained Swin Transformer for land cover classification and segmentation using a contrastive learning approach as illustrated in Figure 2. The training is done in two stages; first, self-supervised training of one unique backbone is performed on a large unlabeled dataset, second, supervised fine-tuning of this backbone is performed on a small, labeled dataset, for two separate downstream tasks. Experimental results on the test set validate our proposed method over training various different baseline models trained from scratch. Our self-supervised approach yields consistently higher performance across different downstream tasks, with particularly strong improvements in the low-data regime. Furthermore, our work illustrates the utility of Transformer models for Earth observation without the need for large labelled datasets.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 2

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019. 2

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *ArXiv*, abs/2105.05537, 2021. 4

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020. 3

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 2, 4

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2

[8] Yuxing Chen and Lorenzo Bruzzone. Self-supervised sar-optical data fusion of sentinel-1/-2 images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 1, 3

[11] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 4

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Michael Schmitt; Lloyd Hughes; Pedram Ghamisi; Naoto Yokoya; Ronny Hänsch. 2020 ieee grss data fusion contest, 2019. 6

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 4

[18] Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 6

[24] Sudipan Saha, Patrick Ebel, and Xiao Xiang Zhu. Self-supervised multisensor change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 2

[25] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive self-supervised data fusion for satellite imagery. In *International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 2, 4

[26] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 5

[27] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021. 2

[28] Robin A. M. Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7252, 2021. 3

[29] Aidan M Swope, Xander H Rudelis, and Kyle T Story. Representation learning for remote sensing: An unsupervised sensor fusion approach. *arXiv preprint arXiv:2108.05094*, 2021. 2

[30] Chao Tao, Ji Qi, Weipeng Lu, Hao Wang, and Haifeng Li. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geoscience and Remote Sensing Letters*, 2020. 2

[31] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, and Hong Liu. Attention-guided cnn for image denoising. *Neural networks : the official journal of the International Neural Network Society*, 124:117–129, 2020. 2

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2

[33] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012. 4

[34] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 125:2, 2016. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[36] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Loddon Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 3

[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[38] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 2

[39] Yuan Yuan and Lei Lin. Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:474–487, 2020. 2

[40] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[41] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 1