CONTRASTIVE SELF-SUPERVISED DATA FUSION FOR SATELLITE IMAGERY

Linus Scheibenreif*, Michael Mommert, Damian Borth

AIML Lab, School of Computer Science, University of St. Gallen
(linus.scheibenreif, michael.mommert, damian.borth)@unisg.ch

KEY WORDS: Self-supervised Learning, Data Fusion, Remote Sensing

ABSTRACT:

Self-supervised learning has great potential for the remote sensing domain, where unlabelled observations are abundant, but labels are hard to obtain. This work leverages unlabelled multi-modal remote sensing data for augmentation-free contrastive self-supervised learning. Deep neural network models are trained to maximize the similarity of latent representations obtained with different sensing techniques from the same location, while distinguishing them from other locations. We showcase this idea with two self-supervised data fusion methods and compare against standard supervised and self-supervised learning approaches on a land-cover classification task. Our results show that contrastive data fusion is a powerful self-supervised technique to train image encoders that are capable of producing meaningful representations: Simple linear probing performs on par with fully supervised approaches and fine-tuning with as little as 10% of the labelled data results in higher accuracy than supervised training on the entire dataset.

1. INTRODUCTION

Increasing numbers of Earth-orbiting satellites produce large quantities of remote sensing data every day. The analysis of this data with machine-learning (ML) techniques is of great interest for many applications in Earth Observation, such as land use monitoring or change detection (Zhu et al., 2017). However, many of the most frequently used ML algorithms for these tasks are supervised, and thus depend on the availability of high-quality labels for each observation (Scheibenreif et al., 2021). Obtaining the labels is typically a laborious process, involving expensive human expert annotators. This leaves the vast majority of available remote sensing data unlabelled, and therefore out of reach for supervised ML algorithms. Our work targets all applications of ML in the remote sensing domain where it is possible to obtain small amounts of labelled data at reasonable expense, but not in quantities that are sufficient to train large neural network models. In such scenarios, a combination of self-supervised pre-training and subsequent supervised fine-tuning makes it possible to leverage large unlabelled datasets in conjunction with a small amount of labelled observations. In particular, contrastive self-supervised learning (SSL) recently emerged as a powerful way of fitting deep neural network models on unlabelled datasets and to obtain strong performance on related down-stream tasks (Jaiswal et al., 2021). The central idea of contrastive SSL is to compare and distinguish samples from one instance with samples from other instances (Wu et al., 2018). This incentivizes the model to learn meaningful features that are constant for different observations of the same instance, but vary across instances. In existing literature, multiple observations of one instance (*e.g.*, an image) are typically obtained by applying strong random augmentations to the original sample. The models are trained to match augmented versions of the same image and thus learn to become invariant to the applied augmentations (Chen et al., 2020a). The direct application of standard contrastive SSL methods for natural images to remote sensing data is not straightforward given the different data characteristics (Ayush et al., 2021). Commonly used augmentation techniques (*e.g.*, changing image hue or saturation) might not be well-defined for non-RGB remote sensing modalities, or introduce undesirable invariances in the resulting model.

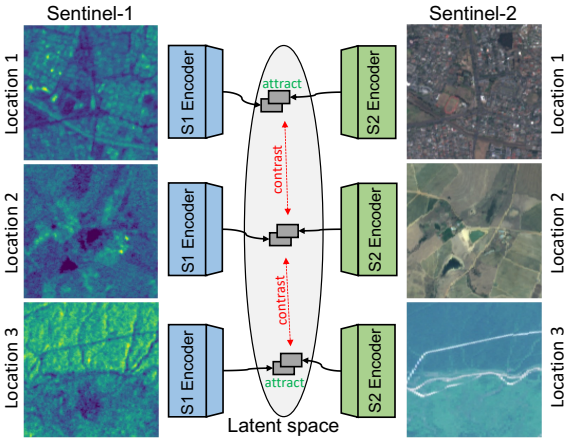

Figure 1. Contrastive SSL with multi-modal remote sensing data. Convolutional encoders are trained to maximize the similarity of latent representations between Sentinel-1/2 images of the same location (positive pairs) while distinguishing them from other scenes (negative samples).

The contributions of our work are as follows:

- We propose an augmentation-free variant of contrastive SSL. Same-instance (*i.e.*, positive) samples are obtained from near-in-time imagery of the same scene by satellites with different sensing techniques (see Fig. 1).
- Our approach exploits the geo-location information of remote sensing data to match observations between sensors. This enables the model to jointly learn representations of data from multiple sources, thus performing data fusion without supervision.
- We show that this approach yields significant improvements on down-stream classification tasks, particularly when only small amounts of labelled data are available.

* Corresponding author

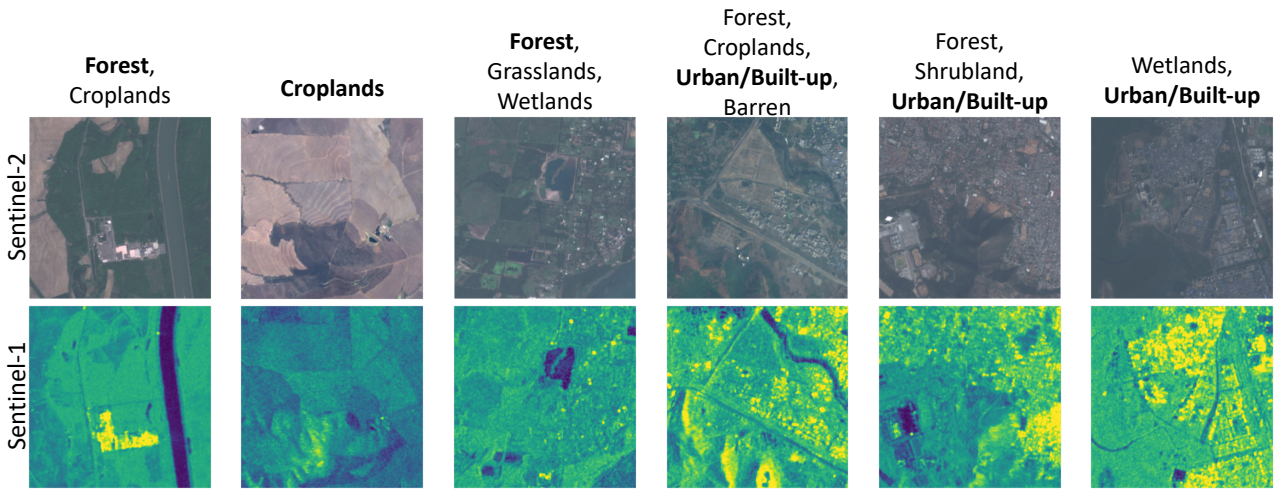

Figure 2. Sample of Sentinel-1/2 image pairs (RGB bands of Sentinel-2 and VV polarization of Sentinel-1) with multi-labels obtained from the DFC2020 dataset. The multi-label classes are given with the majority class in bold font.

2. RELATED WORK

The idea of contrastive learning (Hadsell et al., 2006) has recently received a lot of attention in the computer vision literature (Wu et al., 2018, Oord et al., 2018, Chen et al., 2020a, He et al., 2020, Tian et al., 2020) and subsequently produced methods that achieved stronger ImageNet classification results than supervised training (Chen et al., 2020b). Following this success, contrastive SSL was also adopted in the remote sensing domain. Recently, the geo-information of remote sensing data was exploited to collect images of the same scene at different points in time as *temporal positives* and contrasted against images from other locations (Ayush et al., 2021). In combination with an auxiliary geo-location classification task, this approach results in improved performance across a number of downstream classification and segmentation problems. The Contrastive Predictive Coding method (Oord et al., 2018) has been adapted to remote sensing data by drawing positive pairs as different crops from a satellite image, resulting in improved downstream classification accuracy over ImageNet pre-training, particularly in the multi-spectral case (Stojnic and Risojevic, 2021). In Seasonal Contrast, a two-step procedure for contrastive SSL on remote sensing data has been proposed (Mañas et al., 2021). First, a representative, unlabelled dataset is purpose-built and then a SSL model based on MoCo-v2 (Chen et al., 2020c) with multiple embedding subspaces is trained with augmented and temporal positives. Besides classification and segmentation, contrastive SSL has also recently been applied to tackle problems that are more specific to the remote sensing domain like change detection (Chen and Bruzzone, 2021a, Saha et al., 2021) or data fusion (Chen and Bruzzone, 2021b). Most similar to our work, self-supervised learning has been shown to enable change detection with pre- and post-change satellite images of different modalities (Saha et al., 2021). This technique differs from our approach in the loss function, which combines deep clustering, temporal consistency and contrastive losses, the backbone architecture which starts to share latent representations of different image modalities before a common convolutional layer, and the change detection down-stream task. Another recent related work on self-supervised data fusion with multi-modal satellite data (Chen and Bruzzone, 2021b) utilizes model architectures based on ResUnet blocks and operates on pixel-level representations, distinguishing it from our work.

3. DATA

This work uses multi-modal satellite data from the Sentinel-1 and Sentinel-2 satellites of the European Space Agency's Copernicus Program. Spatially aligned image pairs are obtained from the SEN12MS dataset (Schmitt et al., 2019). See Fig. 2 for representative samples.

Sentinel-1 The Sentinel-1 mission consists of two polar-orbiting satellites with C-band synthetic aperture radar (SAR) devices (Torres et al., 2012). Sentinel-1 provides SAR imaging at up to 5m resolution with dual polarization and revisit times of about 1 week, even in cloudy conditions. We use the VV and VH polarizations of the ground-range-detected Sentinel-1 products in interferometric wide swath mode (10m resolution).

Sentinel-2 Sentinel-2 consists of two polar-orbiting satellites that provide multi-spectral imagery covering the visible, near infrared, and short-wave infrared wavelengths with a ∼5 day revisit rate and up to 10m resolution (Drusch et al., 2012).

SEN12MS SEN12MS is a large scale dataset of spatially aligned Sentinel-1/2 images (180,662 paired observations obtained in the same season), which we use in this work (Schmitt et al., 2019). The resolution of all bands for both modalities is pre-processed to 10m. SEN12MS also contains MODIS land-cover information, which is not utilized here due to its low resolution (500m). Instead, we use a dataset of Sentinel-1/2 observations published by the IEEE GRSS for the Data Fusion Contest 2020 (**DFC2020**) with high-fidelity dense land-cover annotations for model evaluation in a classification task (Yokoya et al., 2020). DFC2020 provides a split into test and validation sets of 5,128 and 986 observations, respectively. The land-cover labels are available on a pixel basis and cover the 8 classes: `Forest`, `Shrubland`, `Grassland`, `Wetland`, `Cropland`, `Urban/Built-up`, `Barren` and `Water`. To create classification targets, we aggregate the dense labels of each scene by selecting the majority class. For multi-label classification, each scene is labelled with all classes that cover more than 10% of the image, following the approach by (Schmitt and Wu, 2021) (see Fig. 2). We utilize VV and VH polarizations of Sentinel-1, and all 13 spectral bands of Sentinel-2.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2022-705-2022 | © Author(s) 2022. CC BY 4.0 License.

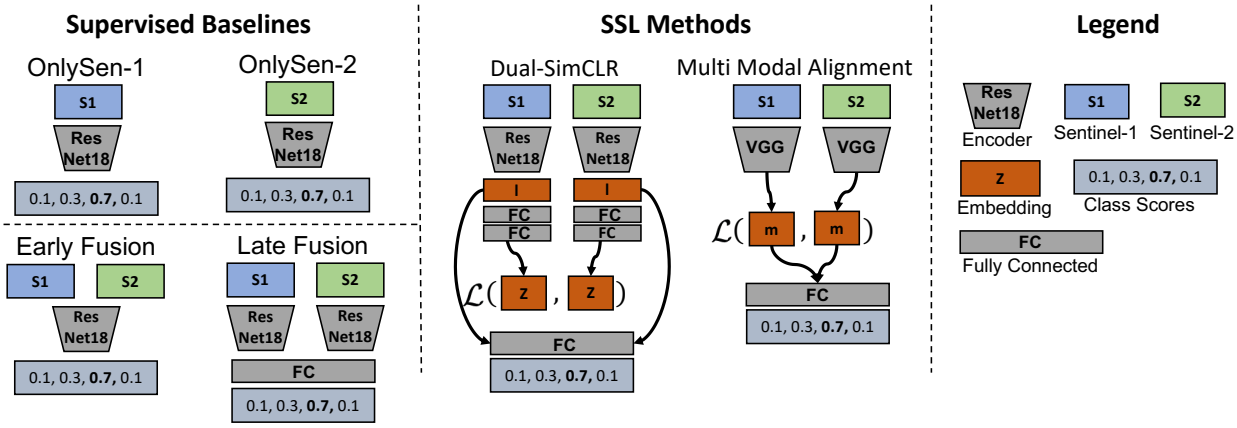

Figure 3. Overview of the model architectures for supervised (left) and self-supervised (middle) land-cover classification approaches including fully-connected classification heads.

4. METHODS

We address the problem of learning meaningful data representations from spatially aligned multi-modal satellite imagery in a self-supervised fashion. This paper presents two approaches that extend recent advances in SSL of natural (Chen et al., 2020a) and medical (Windsor et al., 2021) image representations to the remote sensing domain. These approaches use contrastive SSL, which tasks image encoders to map multiple views of an instance close together in latent space, while maintaining distance to other instances (see Fig. 1). Multiple views are typically not available in natural images, thus necessitating the use of random augmentations to simulate them. In medical imaging, multiple views are available when multiple imaging techniques are applied on the same the subject. Similarly, in the remote sensing domain geo-location information facilitates the collection of multiple views per scene from different sensors, which we leverage in this work. The resulting models are tested with single- and multi-label land cover classification as down-stream tasks.

4.1 Supervised Baselines

As point of comparison for the self-supervised methods presented in this paper, we provide the results of 4 supervised learning strategies. The two single-source methods **OnlySen-1** and **OnlySen-2** are based on either Sentinel-1 or Sentinel-2 data alone and consist of a ResNet18 network (He et al., 2016) with adapted number of input channels (2 for OnlySen-1, 13 for OnlySen-2). For the **EarlyFusion** data fusion approach, the Sentinel-1/2 inputs are concatenated across the channel dimension and processed by a ResNet18 to estimate the land-cover class for the given scene. Finally, the **LateFusion** model consists of two ResNet18 encoders with adapted input layers for the Sentinel-1 and Sentinel-2 inputs. The resulting embeddings are concatenated before the ResNets' fully connected layers and then processed by a single linear classification layer (see Fig. 3).

4.2 SSL Approach 1: D-SimCLR

The Simple framework for Contrastive Learning of visual Representation (SimCLR) is a commonly used approach for contrastive SSL with image data (Chen et al., 2020a). SimCLR defines a contrastive loss in the latent space to maximize the similarity of augmented versions of the same data sample (see Eqn. 1). In the classical setup, the model is a siamese neural network (Bromley et al., 1993) with weight sharing that consists of a convolutional encoder $f(\cdot)$ followed by a non-linear multi-layer perceptron (MLP) $g(\cdot)$. During training, each sample of the mini-batch $\mathbf{x}_i$ is randomly augmented twice to create two visually different views of the same data point. The loss for the positive pair i, j (augmented versions of the same image) over a batch of $2N$ augmented samples is computed as:

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

where $\text{sim}(\cdot,\cdot)$ is the dot product
$\mathbb{1}$ is the indicator function
τ is a so-called temperature parameter

The latent vectors $\mathbf{z}_i$ result from a pass of sample $\mathbf{x}_i$ through the encoder and the subsequent projection MLP, *i.e.*, $\mathbf{z}_i = g(f(\mathbf{x}_i))$. We implement this standard formulation of **SimCLR** on the RGB bands of Sentinel-2 images, following the original method as closely as possible. The positive pairs are created by randomly sampling strong augmentations such as `ColorJitter`, `Flipping`, `Grayscaling` and `GaussianBlur` for each view.

Extension: D-SimCLR To adjust the SimCLR approach for data fusion with satellite imagery, we omit the random augmentations in favor of spatially aligned images from different sensors. To account for the potentially large difference between the two data modalities (*e.g.*, SAR for Sentinel-1 and multi-spectral imagery for Sentinel-2), we replace the weight sharing with dedicated encoders $f_{s1}(\cdot)$, $f_{s2}(\cdot)$ and projection MLPs $g_{s1}(\cdot)$, $g_{s2}(\cdot)$ for the different sources (called Dual-SimCLR, or **D-SimCLR**). The latent vectors of different views therefore depend on distinct model components. To calculate the contrastive loss (Eqn. 1), latent vectors of the positive pair i, j (Sentinel-1/2 images of the same scene) and other elements of the mini-batch k, are computed as:

$$\mathbf{z}_i = g_{s1}(f_{s1}(x_i)) \quad (2)$$

$$\mathbf{z}_j = g_{s2}(f_{s2}(x_j)) \quad (3)$$

$$\mathbf{z}_k = \begin{cases} g_{s1}(f_{s1}(x_k)) & \text{if } \mathbf{x}_k \in \{\text{Sentinel-1}\} \\ g_{s2}(f_{s2}(x_k)) & \text{otherwise,} \end{cases} \quad (4)$$

Table 1. Accuracy (%) for single-label classification on the test set (*i.e.*, DFC2020 validation split). Supervised methods OnlySen-1, OnlySen-2, EarlyFusion and LateFusion are trained from scratch, SSL methods SimCLR, D-SimCLR and MMA are fine-tuned on labelled data. OA indicates the overall accuracy. Dashed line separates supervised and self-supervised methods. Values are averaged over 5 runs with different random seeds. We note that individual runs of the supervised methods specialize on different classes (resulting in high standard error across runs) but converge to similar average performance.

Accuracy (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	**Average**	**OA**
OnlySen-1	80 ± 15	57 ± 2	18 ± 17	0 ± 0	75 ± 10	67 ± 9	58 ± 2	97 ± 2	57 ± 3	62 ± 1
OnlySen-2	43 ± 26	78 ± 12	45 ± 29	11 ± 6	59 ± 9	62 ± 5	61 ± 18	96 ± 6	57 ± 6	62 ± 5
EarlyFusion	60 ± 12	66 ± 37	62 ± 8	1 ± 1	66 ± 10	73 ± 6	66 ± 18	99 ± 0	62 ± 4	66 ± 2
LateFusion	62 ± 23	76 ± 14	51 ± 18	1 ± 2	64 ± 11	71 ± 5	75 ± 9	100 ± 1	62 ± 4	65 ± 3
SimCLR (RGB)	11 ± 12	69 ± 13	45 ± 14	3 ± 3	66 ± 22	26 ± 23	77 ± 14	99 ± 1	49 ± 3	58 ± 4
D-SimCLR	78 ± 11	84 ± 6	62 ± 10	10 ± 6	63 ± 3	84 ± 4	82 ± 7	99 ± 0	$\mathbf{70 \pm 2}$	$\mathbf{70 \pm 1}$
MMA	68 ± 17	89 ± 5	53 ± 13	8 ± 9	71 ± 7	80 ± 6	81 ± 7	100 ± 0	69 ± 2	69 ± 1

where $\{\text{Sentinel-1}\}$ represents the set of all observations from Sentinel-1 in the training data

In our experiments, both encoders are identical ResNet18 networks with adjusted input layers for the Sentinel-1/2 bands. The projection heads are MLPs with two fully connected layers and ReLU activation functions that map to a latent dimensionality of 128. For downstream land-cover classification, the vectors $f_{s1}(x_i^{s1})$ and $f_{s2}(x_i^{s2})$ are concatenated and processed by a linear layer to obtain classification scores (see Fig. 3).

4.3 SSL Approach 2: Multi Modal Alignment

In medical imaging, contrastive SSL has been used to align whole body scans of a subject obtained with different scan modalities for the purpose of unsupervised cross-modal scan registration (Windsor et al., 2021). The contrastive learning procedure is defined as a matching problem where the model tries to maximize the similarity of latent representations derived from scans of one subject, while distinguishing it from those of other subjects. Multi Modal Alignment (**MMA**) uses two spatial encoders $f_{\text{vgg}}(\cdot)$ with identical architecture inspired by the VGG network to compute spatial feature maps (see Fig. 3 and (Windsor et al., 2021)). A correlation map for the scans is computed as the 2D convolution of the two feature maps over each other: $\mathbf{C}_{i,j} = \mathbf{z}_i * \mathbf{z}_j$, with $\mathbf{z}_i = f_{\text{vgg}}(\mathbf{x}_i)$. The contrastive loss then follows Eqn. 1 with $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ defined as the maximum value of the correlation map $\mathbf{C}_{i,j}$. Unlike SimCLR, this method computes the similarity at the level of 2D feature maps rather than between vectors, which retains spatial information at the embedding level. Additionally, this method omits the projection heads. We adapt this approach by replacing medical data (*i.e.*, whole body scans with different modalities) with remote sensing data from different sensing techniques. The matching problem thus tasks the model to match scenes rather than individuals across modalities. For evaluation with land-cover classification, the encoders' feature maps are average-pooled, concatenated and then passed to a linear classification layer.

5. EXPERIMENTS

The supervised baseline models are trained on the `test` split of the DFC2020 dataset (see Section 3) with the Adam optimizer and cross entropy loss function. Similarly, the SSL models with classification head are fine-tuned on the DFC2020 dataset after self-supervised training on SEN12MS. The targets are single- and multi-label land cover classes at the scene-level. We use an image size of 128×128 pixel, cut at random locations from the native 256×256 pixel images. To mitigate the unbalanced class distribution (*e.g.*, 1600 instances of `Forest`, but only 99 of `Barren`), we oversample rare classes during training by drawing multiple 128×128 pixel crops at random locations from the original images. This results in a dataset of 10,393 observations with approximately uniform class distribution. This dataset is randomly divided into training and validation splits which contain 80% and 20% of the data, respectively (resulting in 4102 unique training samples). We tune hyperparameters (batch size, learning rate, number of training epochs) with random search based on the performance on the validation split. Model performance is evaluated on the `validation` split of the DFC2020 dataset (*i.e.*, the test set in our work). We again use 128×128 pixel crops but evaluate the entire images by drawing 4 non-overlapping 128×128 pixel crops in a sliding window fashion from the original data, resulting in 3,944 images.

Evaluation Metrics The classification models are evaluated based on accuracy (see Eqn. 5) for single label classification, and F1-Score (see Eqn. 8) for multi-label classification. We provide class-wise metrics for the 8 land-cover classes, the average of class-wise values, and the overall average across all samples (*i.e.*, without first aggregating by class). This ensures fair evaluation despite the unbalanced class distribution in the DFC2020 validation set. We report the arithmetic mean and standard deviation over 5 runs with different random seeds for each metric.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where TP, TN, FP, FN are the numbers of true/false positive/negative predictions

5.1 Supervised Setup

The 4 supervised baseline models are based on single modalities (OnlySen-1, OnlySen-2) and different data fusion approaches (EarlyFusion, LateFusion). We train these methods for up to 200 epochs for single- and multi-label classification.

Table 2. F1-Score (%) for multi-label classification on the test set (*i.e.*, DFC2020 validation split). Supervised methods OnlySen-1, OnlySen-2, EarlyFusion and LateFusion are trained from scratch, SSL methods SimCLR, D-SimCLR and MMA are fine-tuned on labelled data. O-F1 indicates overall F1-Score. Dashed line separates supervised and self-supervised methods. See notes for Table 1.

F1 Score (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	**Average**	**O-F1**
OnlySen-1	69 ± 2	46 ± 6	29 ± 5	8 ± 8	68 ± 7	81 ± 3	60 ± 8	96 ± 1	57 ± 2	62 ± 2
OnlySen-2	37 ± 20	51 ± 14	43 ± 20	23 ± 18	76 ± 2	79 ± 6	63 ± 10	94 ± 2	58 ± 3	63 ± 2
EarlyFusion	48 ± 10	53 ± 7	45 ± 13	13 ± 11	69 ± 5	84 ± 4	71 ± 4	94 ± 1	60 ± 3	62 ± 3
LateFusion	56 ± 6	45 ± 11	33 ± 9	18 ± 24	64 ± 3	69 ± 16	53 ± 15	96 ± 1	54 ± 7	61 ± 5
SimCLR (RGB)	3 ± 4	49 ± 11	24 ± 16	10 ± 8	63 ± 24	40 ± 36	49 ± 15	73 ± 6	39 ± 10	49 ± 6
D-SimCLR	62 ± 2	61 ± 3	53 ± 7	31 ± 2	72 ± 3	87 ± 0	77 ± 1	96 ± 1	$\mathbf{67 \pm 1}$	$\mathbf{69 \pm 1}$
MMA	58 ± 5	57 ± 5	35 ± 8	10 ± 6	77 ± 3	89 ± 1	73 ± 5	97 ± 0	62 ± 2	66 ± 1

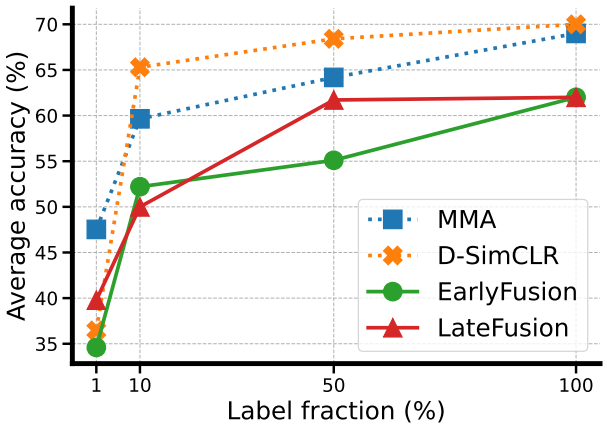

Figure 4. Average accuracy at different label fractions for supervised EarlyFusion/LateFusion, and fine-tuned self-supervised D-SimCLR/MMA methods.

Single-label For single-label classification, OnlySen-1 and OnlySen-2 achieve average (overall) accuracies of 57±3% (62±1%) and 57±6% (62±5%), respectively. In the data fusion setting, these results improve to 62±4% (66±2%) for EarlyFusion and 62±4% (65±3%) for LateFusion (see Table 1). Closer inspection of class-wise accuracies reveals that each of the unimodal approaches struggles with some classes (*e.g.*, `Grassland` for OnlySen-1 and `Forest` for OnlySen-2), while EarlyFusion and LateFusion achieve good accuracy on any class that can be detected well from at least one of the modalities, thus illustrating the central advantage of data fusion. We also note that individual runs with different random seeds "specialize" for different classes (*i.e.*, reach high accuracy), leading to high class-specific standard deviations across runs.

Multi-label We obtain similar results for all 4 baseline methods in the multi-label case. The average F1-Scores for OnlySen-1 and OnlySen-2 are are 57±2%, and 58±3%, respectively. The EarlyFusion approach results in 60±3% and LateFusion in 54±7% average accuracy (see Table 2).

Label fraction We also evaluate the influence of dataset size (*i.e.*, number of labelled observations) on model performance. To that end, the EarlyFusion and LateFusion models are trained for single-label classification on random subsets of the DFC2020 `test` split comprising 1%, 10% or 50% of the original dataset (corresponding to about 80, 800 and 4,000 samples). We find moderate performance differences between training on 50% or 100% of the data, however accuracy is greatly reduced when using 10% or 1% (-12 and -22 average accuracy points for LateFusion) of labelled samples (see Fig. 4).

5.2 Self-Supervised Setup

The self-supervised models are trained on the SEN12MS dataset without access to land-cover labels. Standard SimCLR utilizes the RGB channels of Sentinel-2 images with batch-size 768 and learning rate of $3 \cdot 10^{-5}$ for 100 epochs. For D-SimCLR, we use a batch-size of 128, learning rate of $3 \cdot 10^{-5}$ and temperature value of 0.07 for 50 epochs. MMA is trained with a learning rate of 10^{-5} for 100 epochs while the batch size and temperature are 128 and 0.005, respectively. To evaluate the quality of resulting image encoder models, we add a classification head consisting of a single linear layer to the pre-trained models. The performance is then evaluated by fine-tuning them for single-label and multi-label land-cover classification with labelled samples of the DFC2020 dataset.

Single-label Fine-tuning for single-label classification results in average (overall) accuracies of 70±2% (70±1%) for D-SimCLR and 69±2% (69±1%) for MMA (see Table 1). The SimCLR baseline performs significantly worse at 49±3% average accuracy (58±4% overall accuracy). The contrastive data fusion SSL models thus strongly outperform standard SimCLR (+21 and +20 average accuracy points) and the supervised baselines (+8 and +7 for D-SimCLR and MMA over LateFusion).

Multi-label Fine-tuning the pre-trained SSL models for multi-label classification yields F1-Scores of 39±10% for SimCLR, 67±1% for D-SimCLR and 62±2% for MMA (see Table 2). The corresponding overall F1-Scores are 49±6%, 69±1% and 66±1%. As for the single-label classification, D-SimCLR performs strongest among the SSL methods and also increases the average F1-Score by +7 points over the best supervised approach (EarlyFusion).

Label fraction We investigate the degree to which a lack of labelled samples can be offset by self-supervised pre-training of the image encoders. To that end, the SSL models are fine-tuned with varying amounts of labelled data (see Section 5.1). This reveals strong performance of the SSL approaches at any label fraction, but particularly when little labelled data is available (see Fig. 4). Using only 10% of labels, D-SimCLR still outperforms the strongest supervised approach (LateFusion) trained on 100% of the data by +3 average accuracy points.

Table 3. Average accuracy and F1-Score (%) of linear probe on single- and multi-label classification on the test data.

Accuracy/F1-Score (%)	Single-label	Multi-label
D-SimCLR	$\mathbf{59 \pm 6}$	$\mathbf{60 \pm 0}$
MMA	57 ± 1	56 ± 1

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2022-705-2022 | © Author(s) 2022. CC BY 4.0 License.

Linear-probe We evaluate the quality of the image representations obtained from the self-supervised encoders by linear probing on the DFC2020 land-cover classification problem. To that end, the self-supervised embeddings are fixed and we train only the parameters of a linear layer for single-label classification. This setting allows us to assess how well the SSL methods encode the samples into linearly separable land-cover groups in the latent space. In this simple linear probing setup, D-SimCLR still performs on par with the supervised methods (trained from scratch) and achieves single-label and multi-label accuracies of 59±6% and 60±0%, respectively (see Table 3). Qualitative visual inspection of the latent spaces of our supervised baselines and the SSL methods with t-SNE (Van der Maaten and Hinton, 2008) reveals that MMA and D-SimCLR structure the latent space by land-cover classes to a similar degree as the supervised approaches (see Fig. 5).

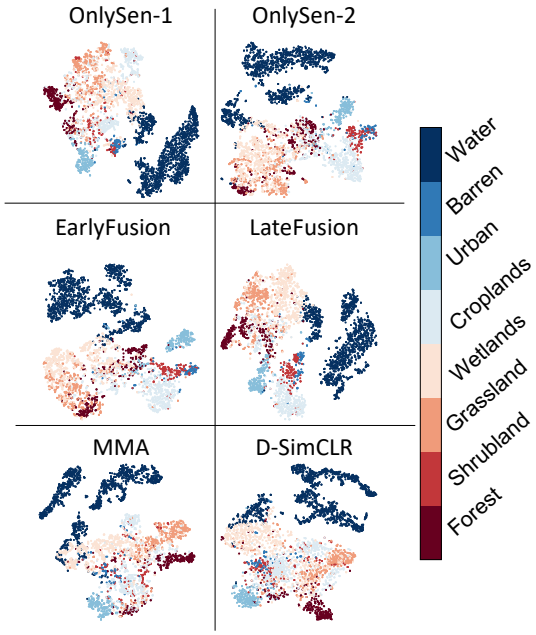

Figure 5. Visualization of the latent space of supervised baseline methods (top uni-modal, middle data fusion) and self-supervised methods (bottom), generated with t-SNE. Samples are colored by land-cover class.

5.3 Cross-dataset Evaluation

To assess if the power of self-supervised pre-training transfers across datasets, we fine-tune MMA and D-SimCLR models which were trained on SEN12MS for land cover classification on the EuroSAT dataset (Helber et al., 2019). EuroSAT consists of 27,000 Sentinel-2 images with land-cover labels from 10 classes. We randomly split the dataset into train (60%), validation (20%), and test (20%) sets. The OnlySen-2 model is used as supervised baseline. After 20 training epochs, this yields an average classification accuracy of 91±1% on the test set. After self-supervised pre-training on SEN12MS, we fine-tune MMA and D-SimCLR for the EuroSAT classification task. To that end, the Sentinel-1 backbones are dropped from the models. We find that the pre-trained models converge fast to strong land-cover classifiers and achieve average accuracies of 97±1% (MMA) and 95±0% (D-SimCLR) after 20 training epochs, outperforming the supervised baseline.

6. DISCUSSION

Our experiments with contrastive self-supervised data fusion focus on two aspects: (1) We fine-tune image encoders that were pre-trained on an unlabelled dataset in a self-supervised fashion. This reveals strong performance on the land-cover classification downstream task in both the single- and multi-label setting. Fine-tuning the D-SimCLR method results in better classification accuracy than any of the supervised baseline approaches. This illustrates that contrasting multi-modal satellite imagery is a useful target in SSL that effectively learns Sentinel-1/2 data characteristics from unlabelled datasets. This insight is particularly valuable when training labels are scarce. As our experiments with varying label fractions reveal, the self-supervised pre-training strategy consistently outperforms supervised training, even when few labelled observations are available. (2) We use linear probing to evaluate the self-supervised image embeddings. Here we find that linear classification of frozen D-SimCLR embeddings can provide better accuracy than standard supervised training. This result further establishes the utility of SSL for data fusion. Across our experiments, D-SimCLR consistently outperforms the MMA method. MMA was initially designed to preserve spatial information from the input image in the embedding space. This property is useful for tasks like scan registration or dense prediction, but is not properly utilized in our single- and multi-label classification problems, which might explain the performance difference to D-SimCLR. The results of the original SimCLR approach also were not competitive with our D-SimCLR. Putatively, this is due to a reduced amount of spectral information in the RGB input data and the problem that strong augmentations as suggested in the original paper result in hardly recognizable scenes when applied to remote sensing data. The D-SimCLR method on the other hand is tailored to remote sensing data and bypasses the challenges of the original approach by leveraging multi-modal satellite data.

7. CONCLUSION

This work investigated the idea of leveraging the geo-location information of remote sensing data for contrastive self-supervised data fusion. We present two techniques that utilize multi-modal remote sensing data of the same location (but from different satellites) as positive pairs in contrastive SSL. Both techniques produce meaningful representations of Sentinel-1 and Sentinel-2 images, as illustrated by linear probing. Fine-tuning the contrastive SSL models strongly outperforms standard supervised data fusion approaches for both single-lablel and multi-label classification. With only 10% of labels, D-SimCLR performs better than any of the supervised approaches trained on the entire dataset. These results demonstrate the potential of SSL for data fusion. Future work could extend the presented methods to dense prediction tasks, or investigate the utility of incorporating additional satellite data modalities.

ACKNOWLEDGEMENTS

We thank the ESA Copernicus Programme and the organizers of the IEEE GRSS Data Fusion Contest 2020 for providing the data used in this work and the anonymous reviewers for helpful comments.

REFERENCES

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S., 2021. Geography-aware self-supervised learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10181–10190.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R., 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04), 669–688.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. *ICML*, PMLR, 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G., 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Chen, X., Fan, H., Girshick, R., He, K., 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, Y., Bruzzone, L., 2021a. Self-supervised Remote Sensing Images Change Detection at Pixel-level. *arXiv preprint arXiv:2105.08501*.

Chen, Y., Bruzzone, L., 2021b. Self-supervised SAR-optical Data Fusion of Sentinel-1/-2 Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25–36.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2, IEEE, 1735–1742.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2217–2226.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., Makedon, F., 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2.

Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., Rodriguez, P., 2021. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. *arXiv preprint arXiv:2103.16607*.

Oord, A. v. d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Saha, S., Ebel, P., Zhu, X. X., 2021. Self-supervised multisensor change detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Scheibenreif, L., Mommert, M., Borth, D., 2021. Estimation of Air Pollution with Remote Sensing Data: Revealing Greenhouse Gas Emissions from Space. *ICML 2021 Workshop Tackling Climate Change with Machine Learning*.

Schmitt, M., Hughes, L., Qiu, C., Zhu, X., 2019. SEN12MS–A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. *arXiv preprint arXiv:1906.07789*.

Schmitt, M., Wu, Y., 2021. Remote Sensing Image Classification with the SEN12MS Dataset. *arXiv preprint arXiv:2104.00704*.

Stojnic, V., Risojevic, V., 2021. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1182–1191.

Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, Springer, 776–794.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M. et al., 2012. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9–24.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A., 2021. Self-supervised Multi-modal Alignment for Whole Body Medical Imaging. *MICCAI*, Springer, 90–101.

Wu, Z., Xiong, Y., Yu, S. X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Yokoya, N., Ghamisi, P., Hänsch, R., Schmitt, M., 2020. Report on the 2020 IEEE GRSS data fusion contest-global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(4), 134–137.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.