Multi-Modal Multi-Objective Contrastive Learning for Sentinel-1/2 Imagery

Jonathan Prexl Michael Schmitt
{jonathan.prexl, michael.schmitt}@unibw.de
Department of Aerospace Engineering
University of the Bundeswehr Munich, Germany

Abstract

The field of spaceborne Earth observation offers, due to constant monitoring of the Earth's surface, a huge amount of unlabeled data. At the same time, for many applications, there still exists a shortage of high-quality labelled datasets. This is one of the major bottlenecks for progress in developing globally applicable deep learning models for analysing the dynamics of our planet from space. In recent years self-supervised representation learning revealed itself to state a very powerful way of incorporating unlabeled data into the typical supervised machine learning workflow. Still, many questions on how to adapt commonly used approaches to domain-specific properties of Earth observation data remain. In this work, we introduce and study approaches to incorporate multi-modal Earth observation data into a contrastive self-supervised learning framework by forcing inter- and intra-modality similarity in the loss function. Further, we introduce a batch-sampling strategy that leverages the geo-coding of the imagery in order to obtain harder negative pairs for the contrastive learning problem. We show through extensive experiments that various domain-specific downstream problems are benefitting from the above-mentioned contributions.

1. Introduction

The number of spaceborne Earth observation (EO) missions with open-data politics has grown spectacularly in recent years. The newfound availability of this data is a significant driver for both industry stakeholders as well as researchers within various disciplines. This data source enables analytical studies on a large spatial scale at high temporal revisit times. Hence, monitoring of large-scale agricultural areas, changes in the vegetation cycle across multiple ecosystems, or the mapping of the ever-growing urban expansion all belong to the set of questions which drastically benefit from the availability of this data. Already today machine learning is an indispensable part of analyzing EO data. Advances in computer vision also translated to the field of EO and therefore machine learning algorithms represent the state of the art in many applications [21]. However, one limiting factor for many potential applications is the absence of high-quality ground truth (GT) labels to train models - since in many cases - good GT data goes hand in hand with elaborate human annotation or measuring campaigns. Existing, already collected GT data also might suffer from local bias and cannot be used to train globally applicable models since changes in the data distribution, due to differences in regional characteristics, are not manifested in the training process. On the other side, the field of satellite data has the unique property that all data is drawn from a finite (spatial) distribution, which can be fully accessed even with multiple modalities.

Given the above-stated two actualities, incorporating the huge amount of available unlabeled EO data into data analytics workflows holds huge potential for the field of EO. This is usually implemented by the class of self-supervised learning algorithms. Here a so-called *pretext task* on the unlabeled data is formulated - where the target variable of the learning problem can be derived by simple and deterministic manipulations of the input data. Networks that are successfully trained on a well-chosen *pretext task* are shown to also perform better, finetuned on the actual problem later on.

The number of potential strategies to successfully pretrain models and therefore learn descriptive model weights on unlabeled data has risen in recent years. One very successful approach to this is given by the class of contrastive learning algorithms, which are representing the state of the art for many classical computer vision problems. Already, many researchers working in the field of EO exploited this kind of pretraining strategy in order to enhance the performance on single downstream tasks [11]. Still, many approaches currently existing are working on a single modality, hence they are possibly not capitalizing on the full typically available information. Even though some multi-modal approaches exist (compare Sec. 2), they have not been fully harmonized in order to drive the understanding of general applicable representation learning strategies for EO data.

Understanding the underlying mechanism behind all different approaches and the role multi-modal data could play in this context, would enable researchers to develop foundation models and therefore potentially drive the quality of derived products and applications. Even if fully foundational models might be far in the future, both pre-training for specific tasks as well as in a more general sense requires a detailed understanding of all underlying mechanisms.

Our main contributions within this work can be stated as follows:

- We define the similarity concept for the contrastive learning problem on different levels to enforce intra- and inter-modality similarity of the learned representations.
- We introduce a new batch-sampling method by choosing spatially close patches for each batch in order to obtain harder negative examples for the contrastive learning problem.
- We perform an extensive comparison between all introduced models and sampling strategies over four different downstream problems that are representing common tasks in the remote sensing domain.

2. Related Work

Early works defined the pretext task for the unsupervised derivation of visual features from imagery through a series of different approaches. These span from predicting the relative location of extracted image patches [3], solving jigsaw puzzles [13] or predicting a random rotation artificially applied to an image [10]. In all those cases labels for pretext tasks can be easily created by simple image manipulations on the raw data, hence they are suitable for a self-supervised learning framework. In all approaches, the basic thought is, that in order to solve the task, a complete understanding of the scene is necessary. Features learned while solving the pretext task can therefore potentially also be descriptive for other problems and therefore help downstream applications. The design of those pretext tasks is critical. In the case of solving the jigsaw puzzles [13], the authors showed that if neighbouring extracted patches have a smooth transition into each other, the features close to the edge are already sufficient to solve the problem, hence no general features and understanding of the underlying data will be learned. Shortcomings in the design of the pretext tasks are representing one of the most critical problems and are referred to as a so-called *shortcut* to the pretext task. Other approaches are given by the set of generative tasks where some aspect of the data will be hidden from the model, whereas the reconstruction of this information is forming the pretext task. Prominent examples are colourizing of grey-scale images [24] or in-painting hidden parts of an image [14]. Since the features learned during the self-supervised pretraining are application agnostic and do not necessarily translate to be descriptive for all downstream applications, further generalizability can be archived by combing pretext tasks in a multitask setting [4], where the authors showed that combining multiple self-supervised signals can help the generalizability on downstream applications.

One newer approach to redefine the problem of self-supervised representation learning for visual features was set by the framework of SimCLR [1] and its successor SimCLRv2 [2] which fall in the class of *contrastive learning methods*. Here a set of augmentations $\mathcal{A}$ (e.g. Cropping, Rotation, Color-Augmentations) is applied to an image patch twice in order to create two *views* of a given image. The self-supervised signal is then defined by mapping those two views into the identical point in latent space (forcing similar features) while preserving the distance to all other images in the batch (negative examples). Impressively the authors of [2] were able to outperform the fully supervised *ImageNet* [17] baseline by only fine-tuning their pre-trained network on 10% of the available labels.

Given the fact that the network is explicitly asked to be invariant to the set of augmentations $\mathcal{A}$, those augmentations are representing the most critical part in the design of a contrastive learning framework. While introducing *SimCLR* [1] the authors stated that a significant strength of augmentations is necessary to avoid the shortcut problems, analogue to the previously described methods. Being invariant to specific augmentations can - depending on the downstream tasks - also be counterproductive. In [23] the authors reported that certain transformations are reducing the accuracy for different kinds of downstream applications if the invariance to the augmentation is suppressing the learning of features that are relevant to solve the corresponding problem. Besides the potentially unwanted invariance to augmentations, another pitfall of contrastive representation learning frameworks can occur if the span of the representation vectors does not cover the full latent space as described in [9]. We refer to that phenomenon as *dimensional collapse* from here on.

Besides the artificial generation of two views, given an input image, contrastive examples can also be introduced by forcing the similarity of features over different modalities, e.g. text and images. In [25] the authors stated that in order to fully capitalize on the information included in all modalities similarity on multiple levels must be preserved. We will build on this idea while translating this concept to the field of EO later on.

To this day, self-supervised pretraining has also received attention from within the remote sensing domain. Early works including the study of image in-painting and predicting relative patch locations [20] to define a self-supervised signal showed that pretraining will heavily support remote sensing specific downstream tasks especially while working with limited labelled data. Further researchers made use of the unique characteristics of remote sensing imagery for the novel and domain-specific formulation of a contrastive loss problem. Examples include introducing seasonal changes [12], as an additional and alternative way to produce views without applying artificial augmentations to the input data. Further [8, 18] made use of the multi-modal nature of the observations spectra. Here the views are, equal to [12], produced without applying argumentations, by considering images from the optical and synthetic-aperture-radar (SAR) domain from the same location, as a positive pair for the constative loss function.

3. Method

3.1. Architecture Design

Earth observation data differs in a few key aspects from classical object-centred imagery. This fact can be explicitly exploited for the development of domain-specific self-supervised learning algorithms. Two of the major differences are the inherent geo-referencing of the observations and the fact that one usually has access to subsequent measurements of sensors with the same or different modalities. Previously [18] exploited this inherent data property to develop a so-called *augmentation-free* contrastive learning algorithm. Here, measurements of the *Sentinel constellation* with identical locations captured by different sensor modalities x_{S1} and x_{S2} served as the positive pair in the contrastive loss function. This holds the advantage that the (sometimes unwanted) property, described in the previous section, of the augmentation-invariance, can be suppressed. However, we argue that this leads to another intrinsic problem. Given two views of the same scene acquired with different sensor modalities, the encoders can only focus on information which is present in both views, since both inputs are aimed to be mapped into the identical location in the latent space. From here on we refer to this as inter-modality information. Further, the network has to learn to be invariant to all information that only occurs in one of the two views, e.g. spectral information in the x_{S2} case or SAR-specific features in the x_{S1} case, since this information cant be used to align the vectors in latent space. We refer to this kind of feature as intra-modality information.

To visually support this hypothesis Fig. 1 shows an example of the same scene captured by *Sentinel-1* and *Sentinel-2*, respectively, which was used in [18], and also will represent the two modalities studied throughout this work. Even when [18] is showing promising results while focusing on inter-modality views only, an increasing performance can be expected by solving the above-mentioned shortcoming.

In order to approach this issue we propose ***Intra-** and **I**nter-modality SimCLR* (*IaI-SimCLR*) a multi-objective setup where representations are forced to fulfil both, i.e. an inter-modality similarity as well as an intra-modality similarity, to ensure the information that is exclusively present in one of the two images, is preserved. For that, we are building on top of [18] where augmentation-free views represented by different modalities $\mathbf{x}_{S1} \in \mathbb{R}^{2\times W\times H}$ and $\mathbf{x}_{S2} \in \mathbb{R}^{10\times W\times H}$ are processed by separate encoders $\mathbf{h}_{S1} = f^{S1}(\mathbf{x}_{S1})$ and $\mathbf{h}_{S2} = f^{S2}(\mathbf{x}_{S2})$ and further projected into a lower-dimensional latent space by two separate projection heads $\mathbf{z}_{S1} = g^{S1}_{inter}(\mathbf{h}_{S1})$ and $\mathbf{z}_{S2} = g^{S2}_{inter}(\mathbf{h}_{S2})$. We denote individual samples that form the positive pairs (same location different modality) of the batches $\mathbf{z}_{S1}$ and $\mathbf{z}_{S2}$ as $\mathbf{z}_i$ and $\mathbf{z}_j$. From here the *NT-Xent* loss [1] can be calculated along all positive pairs of a mini-batch, as:

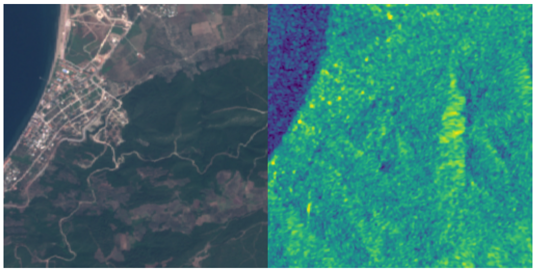

Figure 1. An exemplary *Sentinel-2* and *Sentinel-1* measurement over the identical location. Our hypothesis is that matching the two modalities into the same point in latent space will suppress intra-modality information e.g. the spectral information in S2.

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

where N denotes the batch size, sim defines the cosine similarity, τ the temperature parameter and $\mathbf{z}_k$ negative samples from either modality. Here the operator $\mathbb{1}_{k\neq i}$ evaluates to 1 if $k \neq i$ and 0 otherwise. We extend this approach by two additional losses that are forcing the intra-modality similarity with respect to applied transformations. For that, we introduce two augmented versions (per modality) of the original view which will be further denoted by $\mathbf{x}'_{S1}$, $\mathbf{x}''_{S1}$ and $\mathbf{x}'_{S2}$, $\mathbf{x}''_{S2}$, respectively. In the following, we denote all definitions that are analogues for the processing of the two modalities with the subscript $_{S12}$ which stands for either modality. Introducing two additional projection heads for the intra-modality similarity g^{S12}_{intra}, the representations can be calculated as $\mathbf{z}'_{S12} = g^{S12}_{intra}(f^{S12}(\mathbf{x}'_{S12}))$ and $\mathbf{z}''_{S12} = g^{S12}_{intra}(f^{S12}(\mathbf{x}''_{S12}))$, analogue to the inter-modality case. With, $\mathbf{z}_i$ and $\mathbf{z}_j$ representing positive pairs from the batches $\mathbf{z}'_{S12}$ and $\mathbf{z}''_{S12}$ the intra-modality loss can be calculated with Eq. (1) and will be denoted as $\mathcal{L}^{S12\,intra}_{i,j}$. Similar to [1] the choice of applied augmentations and their corresponding strength is one of the most critical hyper-parameters when working with artificially generated views.

Type	Probability	Applied to
Crop	100%	S1+S2
Flip	50%	S1+S2
Grey-scale	10%	S2
Color Augmentation	80%	None or S2
Gaussian Blur	30%	S1+S2

Table 1. The augmentations that form the set $\mathcal{A}$ and their corresponding event probabilities.

An overview of the used augmentations for each modality, and their respective probabilities, can be seen in Tab. 1. One of the key findings of [1] is the potential shortcut that occurs when augmentations changing the colour distribution of the imagery $a_{color} \in \mathcal{A}$ are not present. Since the effect of this augmentation on multi-spectral remote sensing imagery is not trivial, we subdivide all evaluations of *IaI-SimCLR* into ones with active and non-active colour augmentation, in order to empirically study the impact of a_{color}, as an element of $\mathcal{A}$.

All three losses are equally weighted throughout our study, even though we expect potentially to further fine-tune representations towards a specific task when weighing appropriately. We, therefore, calculate the final loss as

$$\mathcal{L}^{total} = \mathcal{L}^{inter} + \mathcal{L}^{S1\,intra} + \mathcal{L}^{S2\,intra} \tag{2}$$

This way it is ensured that f^{S12} cannot suppress information which is only present in one of the two modalities. A graphical illustration of the proposed method and all corresponding combinations of the loss formulation we study can be seen in Fig. 2.

3.2. Batch Sampling Strategies

As the core objective of any contrastive learning problem is to split apart positive - similar - from negative - dissimilar - examples, the variation of the shown imagery has a large effect on the complexity of the problem. Design decisions that are introducing a shortcut possibility of self-supervised pretraining will hence heavily affect the resulting performance on downstream tasks. Satellite imagery has the unique property that the similarity of examples can be easily enforced by choosing spatially close examples. We study the effect of picking images randomly and via *local batch sampling* (LBS), which stands in contrast to *(spatially) random batch sampling* (RBS). A visual comparison between these two approaches can be seen in Fig. 3. Throughout this work, we subdivide and compare all results produced by *IaI-SimCLR* and the corresponding baselines into trained on RBS and LBS, respectively.

4. Implementation Details, Data and Evaluation Protocol

4.1. Self-Supervised pretraining

We perform the training of the encoders f^{S1} and f^{S2} on multi-modal imagery of the *SEN12MS* dataset [19], represented by measurements of the *Sentinel-1* and the $10\,m$ and $20\,m$ GSD bands of *Sentinel-2*. Throughout this study, we perform all operations on patches of the size $W = H = 256$ and use *ResNet18* [6] architecture for both encoders f^{S12} as well as two linear layers with respective activation for the four projection heads $\mathbf{h}^{S12}_{inter}$ and $\mathbf{h}^{S12}_{intra}$. The feature size for the latent space after the encoding step is $\mathbf{h}_{S12}, \mathbf{h}'_{S12}, \mathbf{h}''_{S12} \in \mathbb{R}^{512}$, before the loss will be calculated on the reduced latent space $\mathbf{z}_{S12}, \mathbf{z}'_{S12}, \mathbf{z}''_{S12} \in \mathbb{R}^{128}$. For evaluation purposes, we mainly test the linear separability of features on the concatenated latent space $\mathbf{h}_{S1+S2} = \mathbf{h}_{S1} \oplus \mathbf{h}_{S2}$. We train the setup with all combinations of losses to study the effect of forcing similarity of the gained representations on multiple levels. All training has been carried out with the Adam optimizer (initial learning rate of 10^{-4}), a batch size of 128 images over 100 epochs (no significant benefit could be found for longer pretraining).

4.2. Evaluation Protocol

For the sake of creating meaningful representations that can be applied in a wide range of remote sensing-specific downstream applications, it is of highest importance to test their respective quality on a wide range of tasks that represents common applications in the field of Earth observation. In the following, we introduce the downstream tasks used for evaluating our method:

Land-cover SEN12MS-DFC. We define a seven-class classification problem based on the imagery of the *SEN12MS-DFC* [15] dataset. We reduce the problem to *Forest*, *Shrubland*, *Grassland*, *Wetland*, *Cropland*, *Urban* and *Barren*. We decided to drop all patches with *Water* as its main land cover class ($> 60\%$) since the corresponding accuracy was found to be $\gg 90\%$, whereas the overall evaluation, therefore, gets less sensitive to changes in the other classes. Overall, this results in 2808 training and 1204 validation samples, respectively.

Land-cover SEN12MS-TP-EWC. The second land cover downstream task is based on the *SEN12MS-TP* [16] dataset. Here we use the ESA-World Cover classification scheme [22] resulting in an eleven-class classification problem. After deriving the main class for a given image patch we restrict the number of samples to 12000 for training and 4000 for validation, respectively.

For both above mentioned land cover problems, we do have matching observations of both *Sentinel-1* and *Sentinel-2*, respectively. The percentage of the main land cover class is above 30% in both cases. Since both datasets are un-

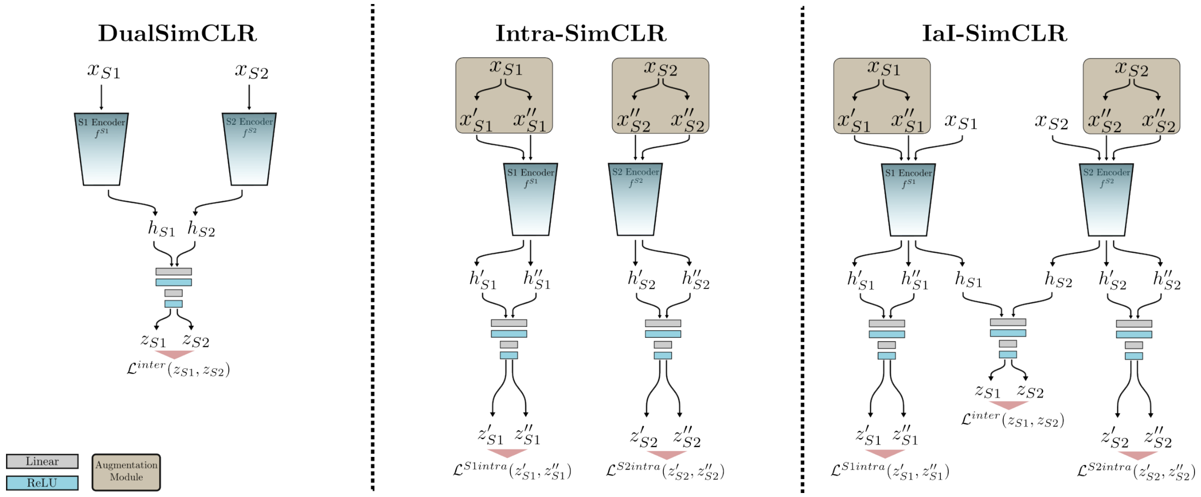

Figure 2. A graphical overview of the proposed experiment setup, comparing the three models *Dual-SimCLR* [18], *Intra-SimCLR* and *IaI-SimCLR*. *Intra-SimCLR* is the original *SimCLR* [1] approach, slightly adapted to the multi-modality setup.

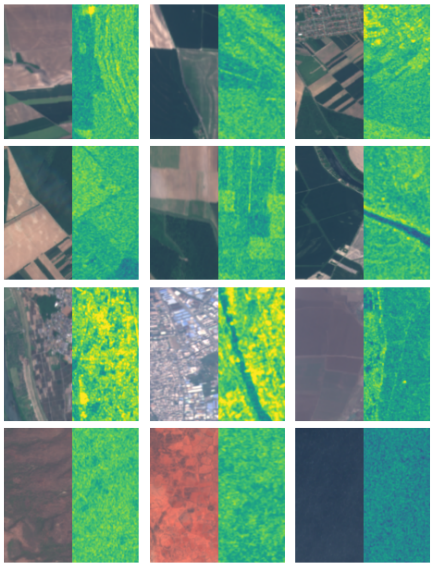

Figure 3. Example images (S2 left half, S1 right half) drawn with the two different batch sampling methods. The top two rows of images were drawn via LBS, bottom two rows of images were drawn by RBS (compare Sec. 3.2).

balanced in their class distribution, we oversampled classes that are less common, during training.

Crop type mapping. Besides the land cover classification problem, we also apply all models on the *Canadian-Crop* dataset [7]. In contrast to the previous two downstream tasks we only have access to the optical modality, and therefore only test the capability of f^{S2}. Since this dataset is heavily biased towards certain crop types we restrict the classes to *Pasture*, *Orchard*, *Potato*, *Soybean*, *Mixedwood*, *Barley*, *Oat* and *Corn*.

All three above-mentioned downstream tasks have been solved by using *Cross-Entropy Loss* optimized with Adam optimizer with an initial learning rate of 10^{-3}, a batch size of 32 images, and trained for 200 epochs. We rank results by the best validation accuracy (Acc), calculated every 5 epochs.

Biomass estimation. The last downstream test is given by the estimation of the total biomass within a given image patch. For that, we use again the above-mentioned *SEN12MS-TP* dataset [16] and search for intersection biomass information y_{BM} of the space-born LIDAR *GEDI-L4A* [5] mission. Formulated as a regression problem one can solve the mapping of $\mathbf{h}_{S1+S2} \rightarrow y_{BM}$ via linear regression, which holds a significant performance advantage and could also be done during the self-supervised training. Assuming this downstream task does hold significant meaning, this would allow for real-time inspection of the self-supervised training process. It must be stated that following this approach the temporal correlation between images and labels is not necessarily given. Nevertheless, we argue that, within some bounds, a general estimation of the total biomass in a given scene is still reasonable. We rank the results for each pre-trained model by the mean absolute error (MAE) on the validation set.

5. Results of Experiments

In this section, we showcase extensive experiments to compare the above-described method *IaI-SimCLR* with *DualSimCLR* and *Intra-SimCLR*, while also varying the type of applied augmentations. Further, we study the effect of the two above-mentioned batch-sampling methods on the performance of the linear evaluation pre-trained networks, with respect to the downstream problem. We compare the performance across multiple benchmark downstream tasks as well as the respective influence of each modality on the final prediction.

5.1. Model Performance

First, we want to guide the reader through Fig. 4 where we show the results of our empirical study, divided into the model, batch sampling strategy and downstream task, all as a function of the length of self-supervised pre-training. Some general observations - first model-independent - can be stated as follows:

Batch Sampling Strategy. We found that - independent of the used method - the difficulty of the pretext task problem is one of the key aspects while trying to learn descriptive features across downstream applications. Models trained on RBS tend to lose their descriptive quality with an increasing number of epochs, while models trained via LBS show a more stable behaviour (compare Fig. 4 left and the right column). Most pronounced is the case for *DualSimCLR* where in the case of RBS we found the model drastically drops in performance after ≈ 40 epochs along all four downstream tasks. Here, changing the type of batch sampling strategy enables the model to stay competitive with all other approaches and - in the case of biomass prediction - represents the best-performing strategy. Similar, but less pronounced behaviour can be observed across almost all five model settings. A possible explanation could be given by the occurrence of the dimensional collapse, similar as reported in [9]. Still, comparing the maximum achieved metric score along each downstream task and ignoring the trend of the performance, we did not find LBS to consistently outperform RBS. We will come back to the consequences of that observation later in this section.

The effect of the color augmentation. While working with multi-spectral optical Earth observation data, a lot of the information is encoded in the relative position of the band values. The effects of changing those values, by introducing a_{color} during the augmentation step, might have non-trivial effects. Comparing the two models (*Intra-SimCLR* and *IaI-SimCLR*) in terms of whether color augmentation a_{color} is part of $\mathcal{A}$, reveals that for the case of LBS, it is throughout better to have it non-active. Further, on all runs trained via the RBS approach, this is still the preferable setting, since it outperforms models trained with active a_{color} in most cases. A possible explanation is given by the fact that for LBS the color distribution of image patches within a mini-batch are already similar, due to the close spatial origin. The biggest difference can be observed in the case of the *Canadian-Crops* downstream tasks, which is in line with the spectral preserving requirement needed for crop-type mapping since the pure geometrical features do not necessarily differ for different crop classes.

With those two model-overarching observations in mind, we compare the different approaches to define the contrastive loss with respect to the modalities of Fig. 4.

Individual Model Performance. Comparing the performance of the individual model across all downstream applications, we can not determine any approach which performs best on all downstream tasks. Still, Fig. 4 reveals that the best performing approach is in most cases either *IaI-SimCLR* or *Intra-SimCLR*, both without a_{color} being active.

To enable the reader to put the absolute performance values for each downstream application into perspective, we also report the fully supervised baseline (same architecture all weights trainable) as well as a fine-tuning on a network with random initialized weights (same architecture, last layer trainable) in Tab. 2. Generally, the model performance lies between the two baselines and in some cases even exceeds the fully supervised benchmark. Still, the objective of this study is to investigate the relative performance of the approaches and hence absolute performance is only given for the sake of completeness.

5.2. Modality Contribution

While having two separate encoders f^{S1}, f^{S2} one can raise the question of how much the individual representations $\mathbf{h}_{S1}$ and $\mathbf{h}_{S2}$ do contribute to the final prediction. Especially in a time-critical scenario where - during the application - only one of the two is present, a balanced information content between both latent spaces is desirable. Let $\mathcal{P}(\cdot)$ be the operator that gives the linear probing accuracy for a given representation $\mathbf{h}_{S12}$. We compare the information content by solving the classification problem of the *SEN12MS-TP-EWC* dataset in Tab. 3 on models trained for 100 epochs. For all scenarios, we find the highest descriptive properties for samples from the combined latent space $\mathbf{h}_{S1+S2}$. The interesting behaviour reveals itself while looking at the descriptive power of $\mathbf{h}_{S1}$ and $\mathbf{h}_{S2}$ as a function of the sampling method. Tab. 3 reveals that for models that make use of $\mathcal{L}^{S2\,intra}$ RBS leads to the situation that the descriptive level of $\mathcal{P}(\mathbf{h}_{S1}) > \mathcal{P}(\mathbf{h}_{S2})$. When changing the samples strategy to LBS this behaviour switches to $\mathcal{P}(\mathbf{h}_{S1}) < \mathcal{P}(\mathbf{h}_{S2})$, which indicated that LBS can be used to suppress the potential color-related shortcut when not having a_{color} present.

5.3. Mixed Strategy Approach

While we earlier demonstrated that in the case of RBS models tend to lose their discriminative property during training, while in some cases still represent the best-performing model for early checkpoints, a combination of the two approaches seems promising. In order to test this hypothesis we changed the batch-sampling method after 30 epochs of pretraining from RBS to LBS and continued for additional 70 epochs. The performance as a function of the

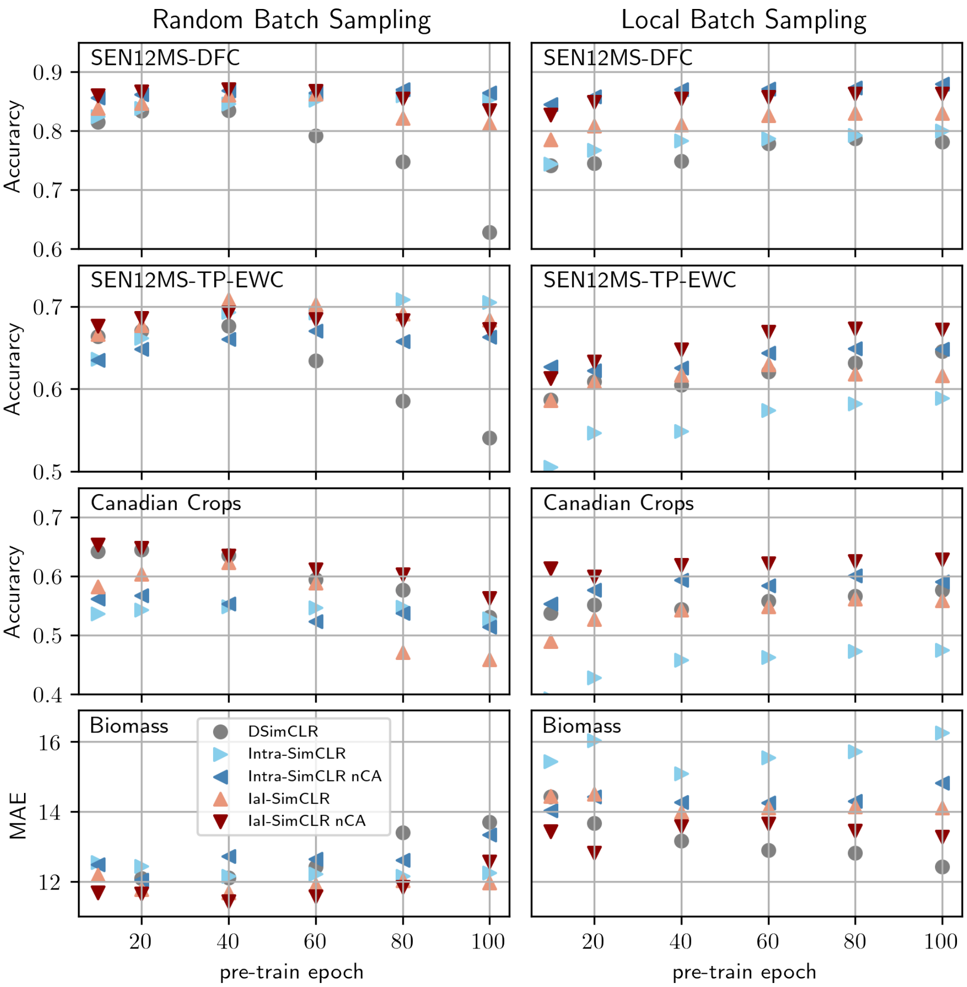

Figure 4. The performance of the models *IaI-SimCLR* with *DualSimCLR* and *Intra-SimCLR* with active- and non-active (nCA) color augmentation a_{color} over four different downstream tasks (rows) described in 4.2. We subdivided models further into trained with random batch sampling (left column) and local batch sampling (right column). Models within one family that only differ in their respective colour augmentation are indicated by the same color (red, blue) and only differ in shade (light/dark).

Task	Random Features	Fully Supervised	Best Linear Eval.
SEN12MS-DFC	0.68 Acc	0.86 Acc	0.89 Acc
SEN12MS-TP-EWC	0.51 Acc	0.62 Acc	0.71 Acc
Canadian Crops	0.58 Acc	0.76 Acc	0.66 Acc
Biomass	22.27 MAE	11.24 MAE	11.4 MAE

Table 2. Baseline for the four downstream applications with fully supervised training and linear probing on a frozen network initiated with random weights. The best-obtained results from the linear evaluation across all models from Fig. 4 are given for comparison.

pre-training epoch can be seen in Fig. 5. Here we can observe a near-monotone positive trend and a clear outperformance of *IaI-SimCLR* for two of the four downstream tasks, which underlines the significance of mixed strategies training approaches.

6. Discussion

Adapting self-supervised methods developed for object-centred imagery, to a new domain, always poses questions about the handling of special domain characteristics i.e. multi-spectral imagery or multiple modalities. This work

Model	Sampling	$\mathcal{P}(\mathbf{h}_{S1})$	$\mathcal{P}(\mathbf{h}_{S2})$	$\mathcal{P}(\mathbf{h}_{S1+S2})$
Dual-SimCLR	RBS	0.449	**0.486**	0.541
	LBS	0.481	**0.537**	0.646
Intra-SimCLR nCA	RBS	**0.574**	0.471	0.662
	LBS	0.465	**0.527**	0.648
IaI-SimCLR nCA	RBS	**0.548**	0.528	0.672
	LBS	0.479	**0.560**	0.671

Table 3. The performance of the three models (without a_{color}) on the *SEN12MS-TP-EWC* classification problem as a function of the sampling strategy evaluated on the individual representations $\mathbf{h}_{S1}$, $\mathbf{h}_{S2}$ and $\mathbf{h}_{S1+S2}$.

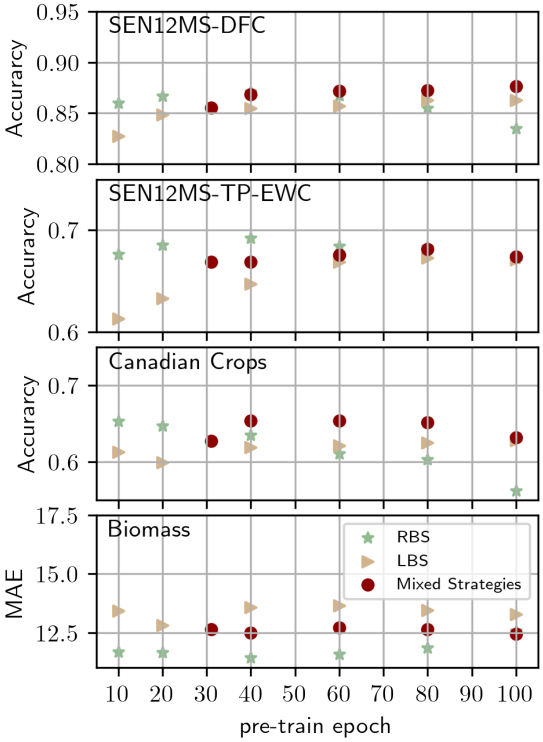

Figure 5. The mixed strategy approach (red) compared to RBS (light green) and LBS (brown).

outlines strategic approaches on how to incorporate inter-modality information while, at the same time, preserving the intra-modality properties of the data. In the following we want to discuss some additional observations of our study:

- For the downstream task based on *SEN12MS-DFC* the best-achieved accuracy approaches 90%, therefore the room for further improvement is intrinsically limited. Here, the data uncertainty of the dataset could suppress all potential performance gains achieved by better representations of the data. This thought is also supported by the fact that on the second land cover task, *SEN12MS-TP-EWC*, our pretraining leads to significant outperformance of the fully supervised benchmark.
- Comparing the absolute performance on the *Canadian-Crops* downstream task with the randomly initialized baseline (compare Tab. 2 and Fig. 4), we can observe that outperformance only happens for *IaI-SimCLR* and *Intra-SimCLR* without a_{color}. At this point, the reader should be reminded that this downstream task is performed on data from $\mathbf{x}_{S2}$ exclusively, therefore a better performance can be expected by changing the set of augmentations $\mathcal{A}$ or enhancing the weight of $\mathcal{L}^{S2\,intra}$ within the total training loss.
- Regarding the mixed strategy approach described in Sec. 5.3, we expect better results by slowly increasing the amount of locally drawn batches during training, in contrast to a discrete and sudden change of batch sampling strategies.

7. Conclusion

We introduced *IaI-SimCLR* and studied different approaches how to define the contrastive learning problem within and across different modalities. We found a strong dependency of the performance for different approaches on how data for individual batches is sampled and showed that common methods might be subject to a dimensionality collapse when not appropriately designed. We compared the performance over multiple downstream applications and show that benchmarking the performance over the training length of the self-supervised pre-training helps in order to gain an understanding of the dynamics.

8. Acknowledgement

We want to thank the team of Prof. Dr. Damian Borth, in particular Prof. Dr. Michael Mommert, Linus Scheibenreif and Joelle Hanna, for fruitful discussions.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pages 1597–1607. PMLR, 2020. 2, 3, 4, 5

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Proc. NeurIPS*, 33:22243–22255, 2020. 2

[3] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proc. ICCV*, pages 1422–1430, 2015. 2

[4] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proc. ICCV*, pages 2051–2060, 2017. 2

[5] RO Dubayah, J Armston, JR Kellner, L Duncanson, SP Healey, PL Patterson, S Hancock, H Tang, J Bruening, MA Hofton, et al. Gedi l4a footprint level aboveground biomass density, 2022. 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 4

[7] Amanda A Boatswain Jacques, Abdoulaye Baniré Diallo, and Etienne Lord. Towards the creation of a canadian land-use dataset for agricultural land classification. In *42nd Canadian Symposium on Remote Sensing*, 2021. 5

[8] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv:2209.02329*, 2022. 3

[9] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv:2110.09348*, 2021. 2, 6

[10] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *Proc. ICLR*, 2018. 2

[11] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. Self-supervised pre-training enhances change detection in Sentinel-2 imagery. In *Proc. ICPR*, pages 578–590. Springer, 2021. 1

[12] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proc. CVPR*, pages 9414–9423, 2021. 3

[13] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. ECCV*, pages 69–84. Springer, 2016. 2

[14] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, pages 2536–2544, 2016. 2

[15] Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021. 4

[16] Thomas Rossberg and Michael Schmitt. Towards a global model for ndvi estimation from Sentinel-1 SAR backscatter. In *Proc. EUSAR*, pages 245–248, 2022. 4, 5

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2

[18] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:705–711, 2022. 3, 5

[19] 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. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, pages 153–160, 2019. 4

[20] 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*, 19:1558–0571, 2020. 2

[21] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020. 1

[22] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, W De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. Esa worldcover 10 m 2021 v200. 2022. 4

[23] Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proc. CVPR*, pages 16650–16659, 2022. 2

[24] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. ECCV*, pages 649–666, 2016. 2

[25] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proc. CVPR*, pages 1450–1459, 2021. 2