Multimodal contrastive learning for remote sensing tasks

Umangi Jain, Alex Wilson, Varun Gulshan
Google Research
jainumangi@google.com

Abstract

Most popular contrastive-loss based methods use multiple views of the same image by applying contrived augmentations on the image to create positive pairs and contrast them with negative examples. While there have been some attempts to capture a richer set of deformations in the positive samples, in this work, we explore a promising alternative to generating positive examples for remote sensing data within the contrastive learning framework. Images captured from different sensors at the same location and nearby timestamps can be thought of as strongly augmented instances of the same scene, thus removing the need to explore and tune a set of hand crafted strong augmentations. In this paper, we propose a simple dual-encoder framework, which is pre-trained on a large unlabeled dataset ($\sim 1M$) of Sentinel-1 (radar) and Sentinel-2 (optical) image pairs. We test the embeddings on two remote sensing downstream tasks: flood segmentation and land cover mapping, and empirically show that embeddings learnt from this technique outperform the conventional technique of collecting positive examples via aggressive data augmentations. We also show that our approach facilitates each modality to learn features which are more clearly discriminable in the other modality.

1 Introduction

Recently, self-supervised learning (SSL) techniques have seen tremendous success as a way to pre-train supervised models. Popular self-supervised frameworks for computer vision tasks, including SimCLR [10], MoCo [20], MoCo-v2 [11], Barlow twins [38], BYOL [19], learn representations by imposing invariance to several image augmentations. Many successful SSL techniques proposed in the past few years use a contrastive learning framework, where the pretext task is based on instance discrimination [35], which treats every instance of an image as a separate class. The positive examples for each class (instance, in this case) are gathered by applying augmentations on each image. Commonly used hand-crafted augmentations include random cropping, gaussian blurring, color jitter, and color drop. Some works have also proposed more elaborate means of collecting positive examples and capturing a richer set of deformations [7] [34] [24] [33] [36]. These SSL techniques have also worked well for many remote sensing tasks. While ImageNet initialization is a strong baseline for remote-sensing applications like scene classification [30], pre-training on unlabeled satellite imagery using SSL techniques provides a further improvement [29].

Another line of work explores contrastive SSL techniques specific to remote sensing domain. Geo-CLR [36] generates positive pairs by utilizing the geolocation metadata available in seafloor imagery and gathering images which are physically close as positives. However, the distance within which an image is considered a positive instance has to be tuned based on the downstream task. GASSL [4] and SeCo [25] leverage spatially aligned images over time to create temporal positive pairs. While these methods improve performance, they still rely on artificial augmentations with many hyper-parameters.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Remote sensing offers another unique possibility of applying multimodal learning on images captured from different sources. Unlike RGB camera images (e.g., ImageNet [15], CoCo [12]), images obtained from different satellite constellations acquire different types of information about the scene: Multispectral, LiDAR, Hyperspectral, Synthetic Aperture Radar (SAR), all capture surface information differently and contain complementary information. Sheehan *et al.* [31] and Heidler *et al.* [22] use additional metadata in the form of geo-tagged Wikipedia articles and audio recordings, respectively, to boost performance. Multimodal/multiview self-supervised learning has also been explored in [13] [8] [23]. However, the scale of pre-training in these works is small, and often the learned representations are not general purpose and limited to a single application.

We propose a multimodal framework for learning representations by using data from two different remote sensing satellites. Our hypothesis is that images from different remote sensors, captured at the same geolocation and close by timestamps, provide better positive examples for contrastive learning than what is obtained using the hand-crafted augmentation techniques. It also allows each modality to learn features which are more clearly visible/discriminable in the other modality. Images captured from different remote sensing sensors can be thought of as naturally occurring strong augmentations of the same scene. We adapt the SimCLR framework and replace the synthetic augmentations by cross-modal positive pairs. Our main contributions include a dual-encoder multimodal framework for remote sensing applications which leverages the naturally occurring augmentations obtained from different remote sensors capturing the same scene. It is pre-trained on a large dataset of $\sim 1M$ paired Sentinel-1 [32] and Sentinel-2 [17] imagery. We test the quality of these embeddings on two publicly available downstream tasks of semantic segmentation: flood segmentation and land cover mapping.

2 Datasets

Multi-satellite unlabeled pre-training dataset: We extract paired Sentinel-1 (also referred to as SAR in this paper) and Sentinel-2 pre-training data using Google Earth Engine [18]. Sentinel-1 and Sentinel-2 satellites acquire imagery using very different mechanisms. SAR imagery captures signals irrespective of the weather conditions, clouds, or darkness, as opposed to optical imagery which shows high variation depending upon prevailing cloud cover. However, SAR images are not as easy to interpret for non-expert humans and doesn't discriminate well between certain land cover types compared to optical imagery. We collect a dataset of 1,087,502 image pairs. Details about data sampling and visualizations are in Appendix A.2.

Downstream labeled datasets: We evaluate trained embeddings on two publicly available datasets:

Sen1Floods11 [5]: We use this dataset for flood segmentation. It consists of hand-labeled Sentinel-1 images of flood scenes and the segmentation task is to demarcate flooded regions.

Dynamic World [6]: This dataset is used for land cover segmentation (into 9 classes) using Sentinel-2 imagery. We also augment this dataset by joining every Sentinel-2 image with a corresponding Sentinel-1 image to allow for models to be trained and evaluated on SAR images that are robust to changing weather and lighting conditions. This dataset is used to set up two downstream tasks, one that uses only Sentinel-1 images as inputs and the other only Sentinel-2 images. The two datasets are referred to as *Dynamic World (Sentinel-1)* and *Dynamic World (Sentinel-2)*, respectively.

Appendix A.3 further details image normalization, dataset size, splits, and label quality.

3 Multimodal pre-training

We adapt the SimCLR contrastive learning framework for multimodal pre-training by constructing positive pairs from different satellite collections, which act as natural augmentations to each other. We take a pair of such images and apply only spatial augmentations to them independently by taking a crop and resizing. Applying spatial augmentations is required to avoid the network from just learning the local edge features. Figure 1 illustrates our proposed method.

These paired images are passed through a dual-encoder architecture and a contrastive loss is applied to map their embeddings closer in feature space and away from the non-matching pairs. The multimodal contrastive loss is defined per batch of images. Consider a batch of N such image pairs, with N images coming from Sentinel-1 $\{s_{1k}, k \in [1, ..., N]\}$ and another N from Sentinel-2 $\{s_{2k}, k \in [1, ..., N]\}$. Each s_{1i} has a corresponding positive example s_{2i} and the remaining $2(N-1)$ images are considered

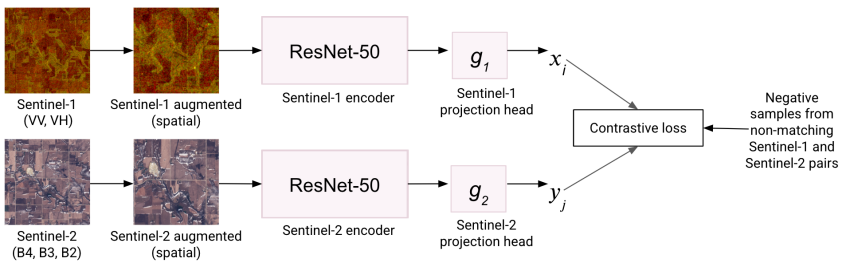


Figure 1: Overview of our framework. Sentinel-1 and Sentinel-2 are jointly mapped in the same space and the representations can be used for both Sentinel-1 or Sentinel-2 based downstream tasks.

as negative examples for this pair. Sentinel-1 and Sentinel-2 images are encoded with different encoder network $f_1(.)$ and $f_2(.)$ to generate feature representations h_{1i} and h_{2i}, respectively. These feature representations are passed through a non-linear projection head, $g_1(.)$ and $g_2(.)$ to produce embeddings x_i and y_i, respectively (with $x_i = g_1(f_1(s_{1i}))$ and $y_i = g_2(f_2(s_{2i}))$). For a positive pair i, we define multimodal contrastive loss per batch as: $l_i = l_{ixy} + l_{iyx}$, where l_{iab} is defined as:

$$l_{iab} = -log \frac{exp(sim(a_i, b_i)/\tau)}{\Sigma_{k=1}^{N} \mathbb{1}_{[k \neq i]} exp(sim(a_i, a_k)/\tau) + \Sigma_{k=1}^{N} exp(sim(a_i, b_k)/\tau)} \tag{1}$$

$sim(.)$ denotes the cosine similarity function between two normalized vector embeddings. The same encoder networks $f_1(.)$ and $f_2(.)$ are also used as the encoder networks for the downstream tasks to allow transfer of these learned representations.

Baselines: We compare multimodal pre-trained models against Random, ImageNet, and SimCLR [10] initializations by fine-tuning on downstream tasks. We follow the SimCLR augmentations proposed in [29] which focused on remote sensing applications, and makes for a stronger baseline.

Both SimCLR and Multimodal models are trained using the same hyperparameters, similar to [10]. We use Deeplabv3+ encoder-decoder architecture [9] for segmentation tasks, with the same encoder as used for pre-training. Training details for pre-training and fine-tuning are specified in Appendix B.

4 Experimental evaluation

The performance on Sen1Floods11 and Dynamic World dataset is reported in pixel-wise IoU of the water class and overall classification accuracy, respectively. For understanding the impact of transferred representations in label scarce settings, we sub-sample multiple sets of the labeled dataset at 1%, 10%, and 100%. For each dataset, the learning rate sweep is run on the entire training dataset and the same learning rate is used for the sub-sampling experiments. Our model performs better than SimCLR across both the tasks and on all the sub-sampling splits as shown in Table 1 and Table 2.

Sen1Floods11: Despite being pre-trained on large datasets, both ImageNet and SimCLR do not significantly outperform the random initialization. Multimodal consistently outperforms ImageNet initialization, with +3.5% and +3.1% improvement in the absolute IoU value (for water class) on 10% and 100% of the training data, respectively.

Table 1: Water IoU on Sen1Floods11 test set. 1% experiment is not conducted as the training set has only 252 examples.

Checkpoint	10% split	100% split
Random	55.22 ± 5.29	66.42 ± 0.25
ImageNet	54.33 ± 2.84	65.56 ± 0.47
SimCLR	55.35 ± 4.95	66.40 ± 0.21
Multimodal	$\mathbf{57.89 \pm 5.65}$	$\mathbf{68.71 \pm 0.29}$

Dynamic World (Sentinel-1): SimCLR initialization provides huge gain over training from ImageNet initialization. Multimodal performs the best, providing a +1 to +1.3% increase in the classification accuracy over SimCLR. The results on 1% and 10% experiments show that in-domain multimodal pre-training can give huge gains [14] [29], providing +8.9% and +5.7% absolute classification accuracy improvement, respectively, over ImageNet initialization, making the learning extremely data efficient.

Dynamic World (Sentinel-2): We observe a similar trend when using Sentinel-2 images as inputs. Multimodal pre-training gives +2.2% and +1.3% absolute improvement on classification accuracy over

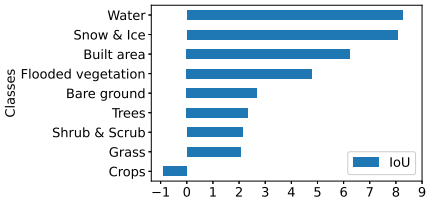
SimCLR for 1% and 10% split, respectively. The performance seems to saturate when using all the labeled data, with SimCLR and multimodal performing almost equally well. The overall land cover segmentation performance on Dynamic World (Sentinel-2) is higher than Dynamic World (Sentinel-1) as Sentinel-2 bands are more discriminative for land cover classes.

Table 2: Classification accuracy on Dynamic World test set using Sentinel-1 (top) and Sentinel-2 (bottom).

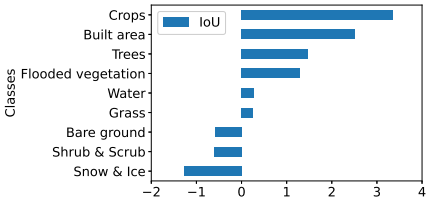
Checkpoint	1% split	10% split	100% split
Random	47.49 ± 1.31	55.81 ± 0.91	61.86 ± 0.48
ImageNet	50.71 ± 1.52	59.00 ± 0.43	65.93 ± 0.20
SimCLR	58.49 ± 0.42	63.70 ± 0.43	67.45 ± 0.18
Multimodal	$\mathbf{59.59 \pm 0.96}$	$\mathbf{64.73 \pm 0.39}$	$\mathbf{68.72 \pm 0.55}$
Random	49.96 ± 0.66	62.90 ± 0.92	71.32 ± 0.26
ImageNet	56.71 ± 1.45	69.00 ± 0.46	73.02 ± 0.32
SimCLR	65.87 ± 1.08	71.67 ± 0.65	$\mathbf{74.96 \pm 0.24}$
Multimodal	$\mathbf{68.07 \pm 1.02}$	$\mathbf{72.93 \pm 0.26}$	74.95 ± 0.18

Learning complimentary information across modalities: We compare the performance of multimodal checkpoint with SimCLR for each class in the Dynamic World (Sentinel-1) and Dynamic World (Sentinel-2) datasets to understand how each class is affected. We report the difference (averaged on 5 sets of 1% training dataset) in pixel-wise IoU for each class when fine-tuned using our multimodal checkpoint, compared to SimCLR checkpoint. Our findings are summarized in Figure 2. We hypothesize that multimodal contrastive learning enables the network to learn complimentary information from different modalities, i.e., it allows each modality to learn features which are more clearly distinguishable in the other modality, while also retaining the features from the original modality.

Consider fine-tuning on Dynamic World (Sentinel-2) dataset, the multimodal checkpoint outperforms SimCLR checkpoint for 8 out of the 9 Dynamic World classes. Highest gain is observed in water class by +8.26%, followed by snow and ice with a +8.05% absolute IoU gain, and built area by +6.24%. Active sensors, like SAR, are known to be good at identifying very smooth surfaces (calm water and smooth ice) and very rough surfaces (man-made buildings) [3] [26] [16] [28]. In line with this, we observe that water, ice/snow, and built area classes observe maximum gains with multimodal pre-training. Compared to SimCLR pre-training on just Sentinel-2 images, multimodal training also incorporates the discriminative capabilities of Sentinel-1 during pre-training.



(a) Dynamic World (Sentinel-2)



(b) Dynamic World (Sentinel-1)

Figure 2: Delta change in absolute per-class IoU with Multimodal initialization over SimCLR initialization on Dynamic World (a) Sentinel-2 (b) Sentinel-1.

We observe a similar pattern when fine-tuning on Dynamic World (Sentinel-1). The multimodal initialization improves over SimCLR in 6 out of the 9 classes. Crops and built area show improved absolute IoU by +3.36% and +2.52%, respectively. Optical imagery is the preferred data source for agricultural crop classification [27], as multispectral optical imagery can measure and monitor the growth, stage type, and crop health. Multimodal learning can leverage this discriminative characteristic of Sentinel-2 to enhance the Sentinel-1 embeddings. These results align with our hypothesis that multimodal learning improves the representations for both the modalities as each modality also learns complimentary information from the other modality.

5 Conclusion

We present a simple method of leveraging abundant amounts of unlabeled data across different input modalities that remote sensing satellites offer. Our method avoids selecting and tuning hand-crafted augmentations for satellite images and outperforms the traditional baselines on two remote sensing tasks. A similar training architecture could also be applied to other modality combinations outside of the two satellites we explored in this paper. Other input types, like temperature, precipitation, elevation, geo-tagged articles, or audio, all contain rich features and can be explored in future work.

References

[1] Sentinel-1 sar grd: C-band synthetic aperture radar ground range detected, log scaling. `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD`. Accessed: 2021-08-01.

[2] Sentinel-2 msi: Multispectral instrument, level-1c. `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2`. Accessed: 2021-08-01.

[3] N Anusha and B Bharathi. Flood detection and flood mapping using multi-temporal synthetic aperture radar and optical data. *The Egyptian Journal of Remote Sensing and Space Science*, 23(2):207–219, 2020.

[4] 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.

[5] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020.

[6] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):1–17, 2022.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[8] Keumgang Cha, Junghoon Seo, and Yeji Choi. Contrastive multiview coding with electro-optics for sar semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[13] Yuxing Chen and Lorenzo Bruzzone. Self-supervised change detection in multi-view remote sensing images. *arXiv preprint arXiv:2103.05969*, 2021.

[14] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Wolfgang Dierking. Sea ice monitoring by synthetic aperture radar. *Oceanography*, 26(2):100–111, 2013.

[17] 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.

[18] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[20] 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.

[21] 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.

[22] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. Self-supervised audiovisual representation learning for remote sensing data. *arXiv preprint arXiv:2108.00688*, 2021.

[23] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Self-supervised learning for invariant representations from multi-spectral and sar images. *arXiv preprint arXiv:2205.02049*, 2022.

[24] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[25] Oscar Manas, Alexandre Lacoste, Xavier Giró-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.

[26] Sandro Martinis and Christoph Rieke. Backscatter analysis using multi-temporal and multi-frequency sar data in the context of flood mapping at river saale, germany. *Remote Sensing*, 7(6):7732–7752, 2015.

[27] Heather McNairn, Catherine Champagne, Jiali Shang, Delmar Holmstrom, and Gordon Reichert. Integration of optical and synthetic aperture radar (sar) imagery for delivering operational annual crop inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(5):434–449, 2009.

[28] Alberto Moreira. Synthetic aperture radar (sar): Principles and applications. `https://earth.esa.int/documents/10174/642943/6-LTC2013-SAR-Moreira.pdf`, 2013. Accessed: 2021-08-01.

[29] Chaitanya Patel, Shashank Sharma, and Varun Gulshan. Evaluating self and semi-supervised methods for remote sensing segmentation tasks. *arXiv preprint arXiv:2111.10079*, 2021.

[30] Rafael Pires de Lima and Kurt Marfurt. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing*, 12(1):86, 2019.

[31] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2698–2706, 2019.

[32] 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.

[33] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34:16238–16250, 2021.

[34] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12586–12595, 2021.

[35] 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.

[36] Takaki Yamada, Adam Prügel-Bennett, Stefan B Williams, Oscar Pizarro, and Blair Thornton. Geoclr: Georeference contrastive learning for efficient seafloor image interpretation. *arXiv preprint arXiv:2108.06421*, 2021.

[37] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[38] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

A Datasets

A.1 Satellite sources

Sentinel-1: Sentinel-1 [1] [32] satellite constellation provides data from its Synthetic Aperture Radar (SAR) instrument, which is an active data collection sensor. It emits microwave radiation in the C-band (5.4GHz) which gets reflected from Earth after interacting with the surface and the bounced signal is recorded to characterize the surface properties. The images are pre-processed by thermal noise removal, radiometric calibration, and terrain correction and the pixel values of the exported images are in decibels. We collect Vertical Transmit-Vertical Receive (VV) and Vertical Transmit-Horizontal Receive (VH) bands from Sentinel-1 at 10m resolution.

Sentinel-2: Sentinel-2 [2] [17] is a satellite constellation that acquires multispectral images at high resolution. It works passively by collecting light reflected from the surface of the Earth. We use Sentinel-2 Level 1C product which represents Top of Atmosphere (TOA) reflectance values. There are 13 spectral bands in this constellation, out of which we only use the RGB bands (B4, B3, and B2 respectively) captured at 10m resolution.

As mentioned in Section 2, the mechanism through which the two sources acquire imagery is very different, as visualized in Figure 3.

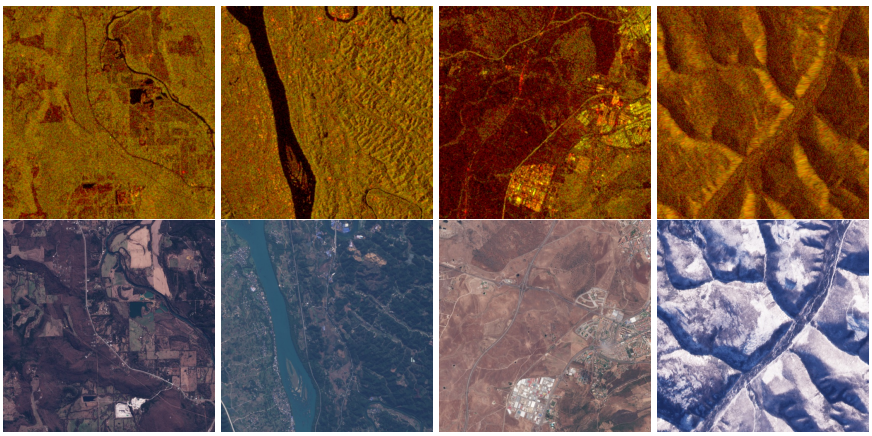


Figure 3: Randomly sampled Sentinel-1 (top) and Sentinel-2 (bottom) image pairs sampled from the same location and close by timestamp. An additional channel of zeros is concatenated to the Sentinel-1 bands for visualization.

A.2 Unlabeled dataset sampling

We obtain Sentinel-1 and Sentinel-2 image pairs by generating IID samples of latitude, longitude from the global land mass, and IID samples of timestamp values collected over a period of 5 years from 31st December, 2016 to 31st December, 2021. We exclude Greenland and Antarctica from the global landmass as it might be difficult for contrastive loss to discriminate between homogenous images. An image pair is collected if there is a Sentinel-1 and a Sentinel-2 image available at the specified location and is within 30 days (in the past) of the specified timestamp. In cases where there are multiple images in the 30 days window, we choose the image which is closest to the specified timestamp. We apply cloud filtering to remove images with more than 15% cloud coverage in Sentinel-2. This is done because cloud covered images obscure semantic information and do not contain features needed for learning in a contrastive learning framework. Sentinel-1 SAR images are, however, not affected by cloud cover and do not need this filter. The total number of images collected for pre-training are 1,087,502 with an image size of $512 \times 512 \times 2$ for Sentinel-1 and $512 \times 512 \times 3$ for Sentinel-2. The regions from where the data is sampled are highlighted in Figure 4.

Figure 4: Regions highlighted inside the red polygons are used to generate random lat-lon values, which are used to export unlabeled Sentinel-1 and Sentinel-2 paired imagery.

A.3 Downstream labeled datasets

Sen1Floods11 [5]: Sen1Floods11 SAR images of flood scenes are collected from 11 flooding events across 6 continents. The authors released 4,831 images which were labeled using simple thresholding models, yielding noisy weak labels. A small subset of 446 images were hand corrected by experts, and we use only this subset for our experiments to avoid training data quality issues, and also test the effectiveness of representation learning in a data scarce setting. The authors provide an IID partition of the data comprising of 252 train images, 89 validation images, and 90 test images.

Dynamic World [6]: We use the train and test dataset that has been labeled by human annotators. The number of Sentinel-2 training examples in the publicly available dataset is 22,906 and 409 samples in the test set. Sentinel-2 images in the Dynamic World dataset consist of 9 spectral bands. We augment this dataset by joining every Sentinel-2 image with a corresponding Sentinel-1 image (VV and VH bands). The joining criteria used is that a Sentinel-1 image should be available at the exact same location and within 3 months of the Sentinel-2 image. If this criteria cannot be met, we discard that example. Upon the join with Sentinel-1 images, 22,891 train images and 407 test images are obtained. We create an IID split of the train dataset into roughly 80:20 train and validation samples (as a separate validation set is not provided explicitly in the data). For the Sentinel-2 images, we only use the RGB bands as inputs.

Table 3 summarises the key attributes of these downstream labeled datasets.

Table 3: Summary of key attributes for both the datasets used for evaluation.

	Sen1Floods11	Dynamic World (Sentinel-1)	Dynamic World (Sentinel-2)
Source	Sentinel-1	Sentinel-1	Sentinel-2
Bands	VV and VH	VV and VH	B4, B3, and B2 (RGB)
Resolution	10m	10m	10m
Image size	512x512	510x510	510x510
Label classes	2 (water, no water)	9 (water, trees, grass, flooded vegetation, crops, shrub and scrub, built area, bare ground, snow and ice)	9 (water, trees, grass, flooded vegetation, crops, shrub and scrub, built area, bare ground, snow and ice)
No. train images	252	18,293	18,293
No. validation images	89	4,598	4,598
No. test images	90	407	407
Train regions	11 flooding events from 6 continents	Global	Global

A.4 Image normalization

Data from both the satellite sources is normalized into a consistent range during data pre-processing. As discussed above, raw Sentinel-1 images are in decibels (dB) scale. We clip these images to a fixed range ([-20dB, 5dB]) and scale it linearly to pixel values between [0, 255]. Sentinel-2 images represent scaled TOA reflectance values. For Sentinel-2, we use a logarithm-based nonlinear scaling method as in [6]. This is done because cloudy pixels in Sentinel-2 have large reflectance values compared to the non-cloudy pixels and a linear scaling would result in a smaller range for the non-cloudy pixels. The same image normalization is applied on the unlabeled pre-training datasets and downstream datasets.

B Training details

We compare the performance of the multimodal pre-trained models against all the baselines mentioned in Section 3 for each dataset.

Pre-training: ResNet-50 [21] is used as the encoder architecture, for both SimCLR as well as our dual-encoder multimodal pre-training. The first 7x7 convolutional layer in the architecture is replaced with two 3x3 convolutional layers, as done in DeepLabv3+ [9]. Our pre-training setup is similar to [10]. We train on a batch size of 4096, weight decay of 10^{-4}, using the Layer-wise Adaptive Rate Scaling (LARS) optimizer [37] with momentum 0.9. The initial learning rate is set to 0.48 with a cosine learning rate decay schedule (we reduce the learning rate by a factor of 10 compared to the default SimCLR training as the default one was high for training on this dataset, resulting in NaNs during training). The models are trained on 256×256 crop sizes till 160k steps. The temperature value for contrastive loss is kept constant as 0.1. The output from Resnet-50 is passed through a projection head giving 128 dimensional embeddings which are normalized before passing to the loss function. For the multimodal model, the image encoder of the modality that comprises the downstream task is transferred for fine-tuning.

Fine-tuning: We use Deeplabv3+ encoder-decoder architecture [9] for segmentation tasks, with the same encoder as used for pre-training. We optimize for the cross entropy loss. For ImageNet, SimCLR, and Multimodal pre-training, the initialization is done only for the encoder, and the atrous convolution and decoder layers are trained from scratch for all models. The weights are optimized on the train split, hyperparameter and checkpoint selection is done on the validation split and the final evaluation is done on the test split. We use a batch size of 64 with 321×321 image size for training. Atrous rates are set to (3, 6, 9) and the weight decay is kept at 10^{-6} for all experiments. The optimizer used is momentum with the momentum parameter set to 0.9. Polynomial schedule is used for the learning rate, starting from an initial value and decaying till zero with power 0.9. We do a sweep over the learning rate values, starting from 10^{-1} to 10^{-4}, varying by a factor of 10^{-1}.

For the sub-sampling experiments, we sample 5 sets of the training data at 1% and 10%. These samples are chosen only once, are non-overlapping, and fixed for all experiments. For 100%, we repeat the experiment on the entire dataset thrice. The learning rate sweep is run only on the entire training dataset. A summary of the training details is given in Table 4.

Table 4: Training details of fine-tuning on the downstream datasets.

	Sen1Floods11	Dynamic World (Sentinel-1)	Dynamic World (Sentinel-2)
Encoder	ResNet-50	ResNet-50	ResNet-50
Evaluation Metric	Mean IoU of Water class	Classification accuracy	Classification accuracy
Number of train steps	20,000	100,000	100,000
Sub-sampling experiment	3 sets of 100% 5 sets of 10%	3 sets of 100% 5 sets of 10% 5 sets of 1%	3 sets of 100% 5 sets of 10% 5 sets of 1%

C Results on validation set

The hyper-parameter and checkpoint selection is done on the validation set of the labeled data for each dataset. The results are detailed in Table 5.

Table 5: Results on the validation set of Sen1Floods11 and Dynamic World dataset. The numbers are aggregated mean and standard deviation of respective metrics.

Dataset	Checkpoint	Learning rate	1% split	10% split	100% split
Sen1Floods11 (IoU water)	Random	0.1		51.04 ± 5.38	63.50 ± 1.03
	ImageNet	0.001		54.66 ± 3.29	64.21 ± 1.66
	SimCLR	0.001		51.24 ± 4.35	64.30 ± 0.36
	Multimodal	0.01		$\mathbf{55.83 \pm 3.5}$	$\mathbf{66.89 \pm 0.15}$
Dynamic World (Sentinel-1) (Classification accuracy)	Random	0.01	51.67 ± 0.88	59.99 ± 0.80	65.39 ± 0.47
	ImageNet	0.01	55.02 ± 0.95	63.50 ± 0.45	69.32 ± 0.16
	SimCLR	0.001	61.75 ± 0.60	67.30 ± 0.25	70.31 ± 0.16
	Multimodal	0.001	$\mathbf{63.26 \pm 0.61}$	$\mathbf{68.49 \pm 0.23}$	$\mathbf{71.26 \pm 0.06}$
Dynamic World (Sentinel-2) (Classification accuracy)	Random	0.01	56.39 ± 1.44	67.56 ± 0.47	73.92 ± 0.18
	ImageNet	0.001	62.06 ± 1.33	71.88 ± 0.36	75.06 ± 0.30
	SimCLR	0.001	69.06 ± 1.30	74.56 ± 0.28	77.00 ± 0.13
	Multimodal	0.001	$\mathbf{70.74 \pm 1.11}$	$\mathbf{74.98 \pm 0.18}$	$\mathbf{77.14 \pm 0.17}$

D Additional experiments

D.1 Improving SimCLR augmentations

We explore further optimizations over the SAR image SimCLR augmentations used in [29] (these include random flips, color jitter, and Gaussian blur to the image crops). We reduce the intensity and probability of applying color jitter and the other parameters are kept the same.

The list of augmentations applied on SAR images include:

- Random horizontal flip with probability 0.5
- Random vertical flip with probability 0.5
- Color distortion (jitter) with a probability of 0.5 and strength 5
- Random Gaussian blur with a probability of 0.5 and strength 4

Table 6: Results on the test set of Sen1Floods11 and Dynamic world (Sentinel-1) dataset with improved SAR augmentations.

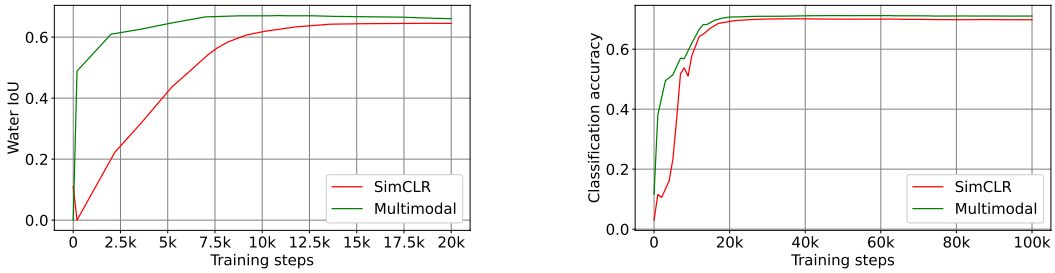
Dataset	Checkpoint	Learning rate	1% split	10% split	100% split
Sen1Floods11 (IoU water)	SimCLR	0.001		55.35 ± 4.95	66.40 ± 0.21
	SimCLR (improved augmentation)	0.01		56.55 ± 4.37	67.83 ± 0.42
	Multimodal	0.01		$\mathbf{57.89 \pm 5.65}$	$\mathbf{68.71 \pm 0.29}$
Dynamic world (Sentinel-1) (Classification accuracy)	SimCLR	0.001	58.49 ± 0.42	63.70 ± 0.43	67.45 ± 0.18
	SimCLR (improved augmentation)	0.001	57.86 ± 1.08	64.14 ± 0.07	67.92 ± 0.42
	Multimodal	0.001	$\mathbf{59.59 \pm 0.96}$	$\mathbf{64.73 \pm 0.39}$	$\mathbf{68.72 \pm 0.55}$

More details about the definition and implementation of these operations can be found in [10]. Table 6 shows the results with this augmentation for Sen1Floods11 and Dynamic World (Sentinel-1) dataset. With this improved pre-trained Sentinel-1 SimCLR model, we observe a +1.2% and +1.4% gain in absolute IoU for water class on 10% and 100% split for Sen1Floods11 dataset compared to the SimCLR checkpoint described in [29]. For the Dynamic World (Sentinel-1) dataset, the change in absolute classification accuracy for 1%, 10%, and 100% split is -0.6%, +0.4%, +0.5%, respectively.

While this makes for a stronger baseline for SimCLR on Sentinel-1 images, our multimodal model still outperforms SimCLR on both the tasks. It shows that exploring optimal parameters for these hand chosen augmentations (operation, strength, and probability) is a computationally expensive task and influences the performance of SimCLR heavily. Our model eliminates the requirement of looking for optimal augmentation hyperparameters and leverages images from different sensors instead to provide a more effective set of augmentations.

D.2 Training speed

We observe that fine-tuning with Multimodal pre-trained checkpoint reaches peak performance in nearly ~ 2.4x less steps (averaged over multiple runs) compared to SimCLR on Sen1Floods11 dataset and comparably on Dynamic World. Figure 5 compares validation curves for both Multimodal and SimCLR. All curves correspond to training with 100% training data.



(a) Sen1Floods11

(b) Dynamic World (Sentinel-1)

Figure 5: Validation curves comparing fine-tuning using Multimodal and SimCLR pre-trained models on (a) Sen1Floods11 (b) Dynamic World (Sentinel-1) dataset.