Exploring Generalisability of Self-Distillation with No Labels for SAR-Based Vegetation Prediction

Laura Martínez-Ferrer
Universitat de València, Spain
laura.martinez-ferrer@uv.es

Anna Jungbluth
European Space Agency, Climate Office, UK
anna.jungbluth@esa.int

Joseph A. Gallego-Mejia
Universidad Nacional de Colombia, Colombia
jagallegom@unal.edu.co

Matt Allen
University of Cambridge, UK
mja78@cam.ac.uk

Francisco Dorr
Independent, Argentina
fran.dorr@gmail.com

Freddie Kalaitzis
University of Oxford, UK
freddie.kalaitzis@cs.ox.ac.uk

Raúl Ramos-Pollán
Universidad de Antioquia, Colombia
raul.ramos@udea.edu.co

Abstract

In this work we pre-train a DINO-ViT based model using two Synthetic Aperture Radar datasets (S1GRD or GSSIC) across three regions (China, Conus, Europe). We fine-tune the models on smaller labeled datasets to predict vegetation percentage, and empirically study the connection between the embedding space of the models and their ability to generalize across diverse geographic regions and to unseen data. For S1GRD, embedding spaces of different regions are clearly separated, while GSSIC's overlaps. Positional patterns remain during fine-tuning, and greater distances in embeddings often result in higher errors for unfamiliar regions. With this, our work increases our understanding of generalizability for self-supervised models applied to remote sensing.

1 Introduction

Earth observation (EO) and remote sensing techniques have witnessed significant advancements in recent years, allowing us to characterize our planet with ever increasing resolution and accuracy. Using EO data, we can e.g. characterize global vegetation in a consistent manner, and link vegetation changes to the complex feedback mechanisms of climate change[1]. Synthetic Aperture Radar (SAR) data [2] has proven immensely valuable for this task, since SAR sensors are sensitive to the dielectric and geometric characteristics of plants, and provide consistent, all-weather, day-and-night monitoring capabilities. However, making sense of the wealth of (unlabelled) SAR data requires advanced computational techniques such as machine learning (ML) and specifically self-supervised learning (SSL) which has revolutionized the training of neural networks without the need for labeled data. Self-supervised learning with SAR images is a new research direction and has not been widely explored [3], and can therefore greatly enhance the scalability and cost-effectiveness of SAR-based vegetation estimation [4].

Workshop on Distribution Shifts: New Frontiers with Foundation Models at NeurIPS 2023.

In this study, we investigate the application of self-**di**stillation with **no** labels (DINO)[5] to SAR-based vegetation estimation. We specifically focus on understanding how the embedding space of the SSL pre-trained model empirically links to generalizability across diverse geographic regions (in our case Europe, China, the Continental United States (Conus), and South America), and when fine-tuning with limited labelled data. By assessing model performances on areas outside of the training set, we uncover valuable insights into the models' robustness and scalability. More specifically, we show that shared embedding spaces across geographic regions improve the generalizability of pre-trained models to previously unseen regions. With this, our work significantly advances our understanding of the applicability and reliability of SSL techniques for remote sensing, with profound implications for environmental science and land management, where labelled datasets might be limited.

2 Data

Areas of Interest & Data Splits. Our dataset includes four regions of interest (AOIs): Europe, China, the Continental United States (Conus), and South America. For each region, we created ML-ready datasets of SAR images and associated vegetation percentage. Among the AOIs, SSL pre-training was performed on three areas (Europe, Conus, and China), while downstream fine-tuning was performed on each AOI separately. We partitioned all data into tiles measuring 448×448 pixels, using `geetiles`[1] and `sartiles`[2] for pre-processing. These tiles were further divided into training (60%), validation (20%), and test (20%) sets based on geographic bands to minimize data leakage across contiguous tiles (see Figure 4 in the SI). To keep the data volume manageable, we only used data from 2020 in this work.

SAR Data. We processed two data diversities from Sentinel-1: (1) SAR coherence from the Global Seasonal Sentinel-1 Interferometric Coherence and Backscatter (GSSIC) dataset [6] and (2) SAR amplitude from the Sentinel-1 Level 1 Ground Range Detected (S1GRD) dataset. While the amplitude characterizes the intensity of the back-scattered signal in vertical (V) or horizontal (H) polarisation, coherence measures the similarity of two complex SAR images taken at different times, i.e. using both intensity and phase information. For GSSIC, we used seasonal averages (spring, summer, autumn, winter) for all possible 12 and 24 timedeltas available to calculate SAR coherence using VV[3] polarization, totalling 8 channels. For S1GRD we used seasonal averages for polarizations VV and VH[4] including the logarithmic difference between them (VV-VH), resulting in 12 channels.

Vegetation Data. After pre-training on unlabeled SAR data, we fine-tuned our models for vegetation estimation, using the Terra MODIS Vegetation Continuous Fields (VCF) product. For this, we calculated the yearly mean vegetation percentage per SAR tile as labels for the supervised fine-tuning. The distribution of labels for each AOI is shown in Figure 6 in the SI.

3 Self-Supervised Learning & Fine-tuning using DINO

Self-**di**stillation with **no** labels (DINO) [5] is an SSL method based on a student-teacher framework, where the student network aims to emulate the representations generated by the teacher. While the teacher network has access to the full input, the student only sees cropped components of the image to learn context from limited information. The weights of the teacher are updated using an exponential moving average (ema) from the student, and the student and teacher embeddings are matched using a cross-entropy loss. In this work, we adapted the original DINO architecture for SAR data. We used ViTs as the student and teacher backbones to make use of the ViTs attention mechanism for model fine-tuning (see Figure 5 in the SI). Due to SAR specifics, we only employed horizontal flipping and cropping as transformations during pre-training. Figure 1 shows a diagram of our ML pipeline. We use either GSSIC or S1GRD as unlabeled input to DINO during pre-training. Afterwards, we use the attention maps generated by the pre-trained student backbone as the input to a linear decoder to predict the mean vegetation percentage for each SAR tile. While we also performed experiments where the weights of the backbone are frozen, the best results were obtained when both the backbone and the decoder were trained during fine-tuning. To rationalize model performance and investigate

[1] `https://github.com/rramosp/geetiles`

[2] `https://github.com/rramosp/sartiles`

[3] I.e. For pulses sent and received in vertical polarization.

[4] I.e. For pulses sent in vertical and received in horizontal polarization.

generalizability, we analyzed the embeddings of the pre-trained or fine-tuned backbone (labelled with pink stars in Figure 1).

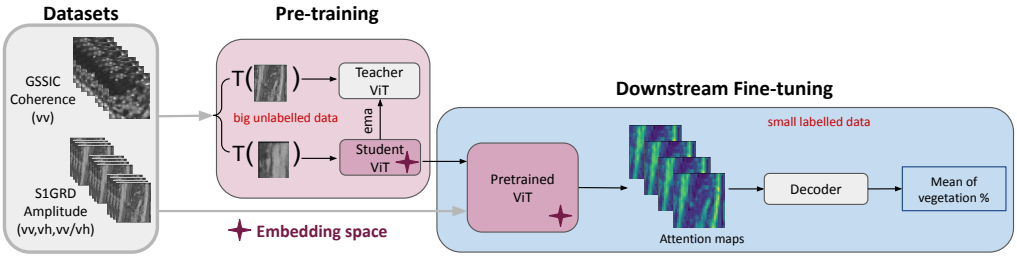

Figure 1: Diagram of the ML pipeline used in this work. GSSIC (8 channels) or S1GRD (12 channels) SAR data is used as an unlabelled input to DINO. After pre-training, the attention maps generated by the student backbone are used as input to a linear decoder to predict mean vegetation percentage. The pink stars show the steps in the pipelines that were used to visualize model embeddings.

4 Results

Benefits of Model Pre-training. For both S1GRD or GSSIC, our model was pre-trained in China, Conus, and Europe, totalling ∼650K input tiles. After pre-training, the model was fine-tuned on each AOI separately. Table 2 in the SI shows the root mean square error (RMSE[%]) in predicting mean vegetation percentage when training the model from scratch or using the pre-trained backbone. In all cases, supervised fine-tuning was performed using only 1% of available data in each AOI, to investigate the performance of SSL when labelled datasets are limited. With more data, the RMSE[%] reduces slightly (see Figure 6 in the SI). We generally observe lower RMSE[%] with pre-training, and therefore focus the remaining discussion on the pre-trained models.

Empirical Investigation of Embedding Spaces & Model Generalizability. The generalizability of SSL models hinges on meaningful embedding spaces, which should capture essential and distinguishing data features for effective generalization to new, unseen data. This requires thoughtful design of pretext tasks and model architectures. Figure 2 shows a 2D TSNE reduction of the embeddings from each AOI after pre-training DINO using S1GRD (2a) or GSSIC (2b). The positional embedding of each tile is colored with the logarithm of the mean vegetation percentage. Importantly, the vegetation labels are not used to compute the embeddings or the 2D TSNE position, and are completely unseen by the model at this stage. The embeddings of the pre-trained models show clear positional patterns, and similar values of vegetation are clustered together. In the case of S1GRD, all AOIs are clearly separated, while the embedding space of GSSIC overlaps regions in a continuous label distribution. We hypothesize that the complexity of features contained in the different SAR input data contributes to the observed differences. S1GRD utilizes amplitude data from multiple polarizations, while our GSSIC input relies only on VV polarization coherence. In high density vegetation regions, the forest structure sensitivity of coherence leads to low values [7], this may hinder the model's coherence-vegetation relationship learning when predicting it. In addition, the S1GRD dataset has higher resolution (448x448 pixels/tile compared to ∼ 50x60 pixels/tile for GSSIC) and it provides more information about local features. Therefore S1GRD likely has more data diversity than GSSIC, resulting in lower RMSE[%] values (see Table 2 in SI).

To empirically assess generalizability, we visualize the embedding space after model fine-tuning and investigate whether model performance on previously unseen regions can be predicted from looking at the embedding space alone. More specifically, we plot the 2D TSNE reduction of the embedding space after fine-tuning on 1% of data in Europe, Conus, or China, and investigate how well vegetation in South America (which has not been seen) is predicted. In Figure 3 the first (third) column shows the embeddings and prediction error for 2000 example tiles in the respective fine-tuning datasets for S1GRD (GSSIC), while the second (fourth) column shows the embeddings and prediction errors when performing inference of the fine-tuned models on 2000 tiles in South America. Similar plots for inference on the training regions are included in the Supplementary Information, but not discussed here. Looking at Figure 3, we observe that after fine-tuning using S1GRD, the embedding spaces of the different AOIs remain separated. Generally, prediction errors

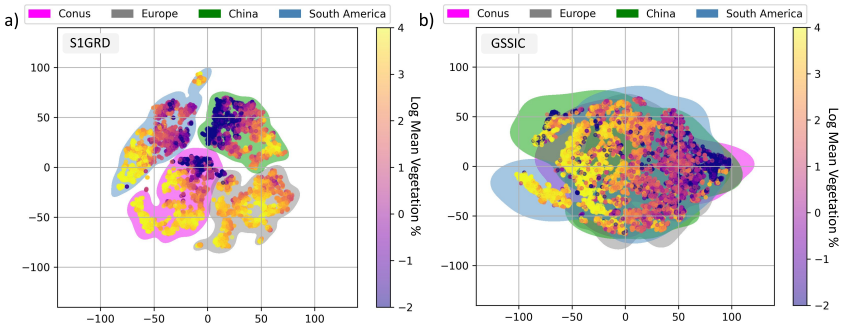

Figure 2: Comparison of the embedding space of the pre-trained DINO model for a) S1GRD, and b) GSSIC, for tiles from Conus (pink), Europe (grey), China (green), and South America (blue). The embedding space was visualized using the class token and via TSNE dimensionality reduction. For each area, we show the embedding of 2000 tiles from the test set (dots), colored with each tile's mean vegetation percentage.

are high (yellow dots), for tiles that are far from the embedding space of the fine-tuning AOI. We quantified embedding separation using Sliced Wasserstein Distance (SWD), which is a mathematical metric used for comparing high-dimensional distributions [8]. All SWD values were calculated for 10 seeds and 10000 projections. The lowest SWD and RMSE[%] for predicting vegetation in South America is observed when using the model fine-tuned on Conus. For the model fine-tuned on Europe or China, both the SWD and RMSE[%] values are high, although no direct correlation between them was observed. In the case of GSSIC, the embedding spaces of all AOIs still show large overlaps after fine-tuning. We empirically observe that prediction errors in South America match those in the fine-tuning AOIs for shared embeddings. In other words, good predictions in the model fine-tuned on e.g. Conus translate to low errors for South America in the same embedding space. We observe that prediction errors increase if tiles in South America fall further outside the embedding space of the fine-tuning AOI. For Europe-fine-tuned model, South America's embedding space barely overlaps with the few tiles in Europe that the model is able to predict well, resulting in the highest RMSE for South American inference. For both input modalities, the embedding space of Europe shows the largest SWD to the embedding space of South America. Looking at the distribution of vegetation across the AOIs, we hypothesize that the stark difference in vegetation found in Europe compared to the other regions leads to the poor prediction performance when applying the model fine-tuned in Europe to other regions.

5 Conclusions

To summarize, we pre-trained DINO on unlabelled SAR data, and later fine-tuned our models on small datasets across diverse geographic regions to predict mean vegetation percentage. We find that pre-training leads to lower errors compared to training from scratch. This highlights the potential of SSL for remote sensing applications, where labeled datasets are limited. Furthermore, we visualized the embedding spaces of the pre-trained and fine-tuned models to draw conclusions about their generalizability. We empirically find that the complexity of the data could lead to larger distances in embedding spaces (measured via the SWD) and therefore cause larger prediction errors for unseen regions as is the case of S1GRD dataset. However, it is important to acknowledge the limitations of this work, and further research is needed to analyze additional datasets. The exploration should involve finding a metric to characterize the complexity and diversity of the data, thereby strengthening the support for our hypothesis. With this, our work presents a first step towards creating fully generalizable models for Earth Observation.

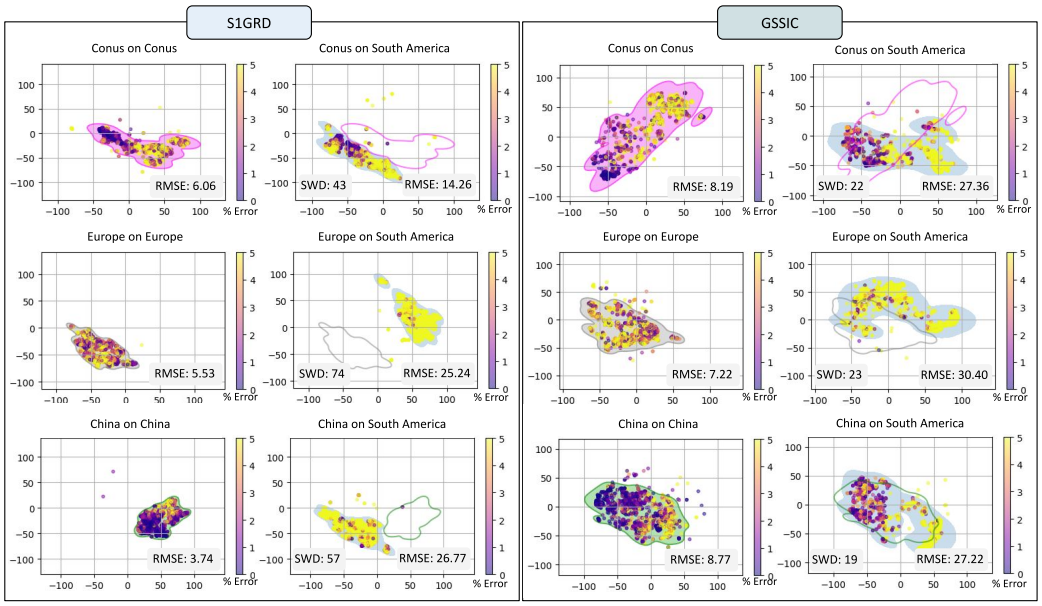

Figure 3: Embedding space and prediction errors (in %) for models pre-trained on S1GRD or GSSIC. The first (third) column shows the embeddings and prediction error for 2000 example tiles in the respective fine-tuning areas for S1GRD (GSSIC), while the second (fourth) column shows the embeddings and prediction errors when performing inference of the fine-tuned models on 2000 tiles in South America. The Sliced Wasserstein Distance (SWD) of the embedding distributions and the root mean square error (RMSE[%]) for inference on the entire test is included.

6 Acknowledgements

This work has been enabled by Frontier Development Lab Europe (`https://fdleurope.org`) a public / private partnership between the European Space Agency (ESA), Trillium Technologies, the University of Oxford and leaders in commercial AI supported by Google Cloud and Nvidia, developing open science for all Humankind. L.M-F. was supported by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). M. J. A. was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks [EP/S022961/1], and additionally by Trinity Hall, Cambridge. We are also indebted to Nicolas Longépé, Carlos López-Martínez, Fabio A. González Osorio, Samuel Bancroft, Emma Hatton, Alison Lowndes, Alistair Francis, Ioanna Bouri and the rest of reviewers during 2023 FDL-Europe sprint.

References

[1] Susan L Ustin and Elizabeth M Middleton. Current and near-term advances in earth observation for ecological applications. *Ecological Processes*, 10:1–57, 2021.

[2] Chang an LIU, Zhong xin CHEN, Yun SHAO, Jin song CHEN, Tuya Hasi, and Hai zhu PAN. Research advances of sar remote sensing for agriculture applications: A review. *Journal of Integrative Agriculture*, 18(3):506–525, 2019.

[3] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.

[4] Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, AS Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Josef Kellndorfer, Oliver Cartus, Marco Lavalle, Christophe Magnard, Pietro Milillo, Shadi Oveisgharan, Batu Osmanoglu, Paul A. Rosen, and Urs Wegmüller. Global seasonal sentinel-1 interferometric coherence and backscatter data set. *Scientific Data*, 9(1), March 2022.

[7] S.M. Ghosh and M.D. Behera. Aboveground biomass estimates of tropical mangrove forest using sentinel-1 sar coherence data - the superiority of deep learning over a semi-empirical model. *Computers & Geosciences*, 150:104737, 2021.

[8] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K. Rohde. Generalized sliced wasserstein distances. 2019.

A Supplementary Information

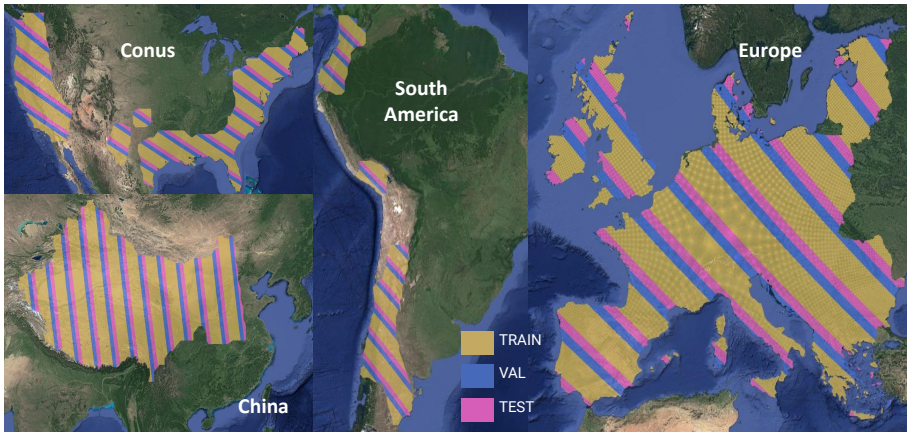

Figure 4: Data splits for Europe, Conus, China, and South America. In total there are 167K image tiles for Conus, 285K tiles for China, 200K tiles for Europe, and 83K tiles for South America.

Table 1: Parameters used for pre-training DINO on unlabelled S1GRD or GSSIC datasets.

	GSSIC	S1GRD
Model Architecture	ViT_Base	ViT_Base
Learning Rate	0.00001	0.000001
Teacher Temperature	0.04	0.001
Student Temperature	0.1	0.03
Warm-up Teacher Temperature	0.04	0.01
Warm-up Teacher Temperature Epochs	10	5
Center Momentum	0.90	0.99

Table 2: Root mean square error (RMSE[%]) for predicting mean vegetation percentage using S1GRD or GSSIC. The model was either trained from scratch or using our pre-trained backbone with 1% of available data in Europe, China, Conus, or South America.

S1GRD				
	Europe	China	Conus	South America
From scratch	5.50	3.39	5.58	**10.14**
With pre-training	**5.31**	**3.12**	**5.30**	11.73
GSSIC				
	Europe	China	Conus	South America
From scratch	**6.33**	7.29	8.29	11.09
With pre-training	6.78	**6.81**	**7.54**	**10.77**

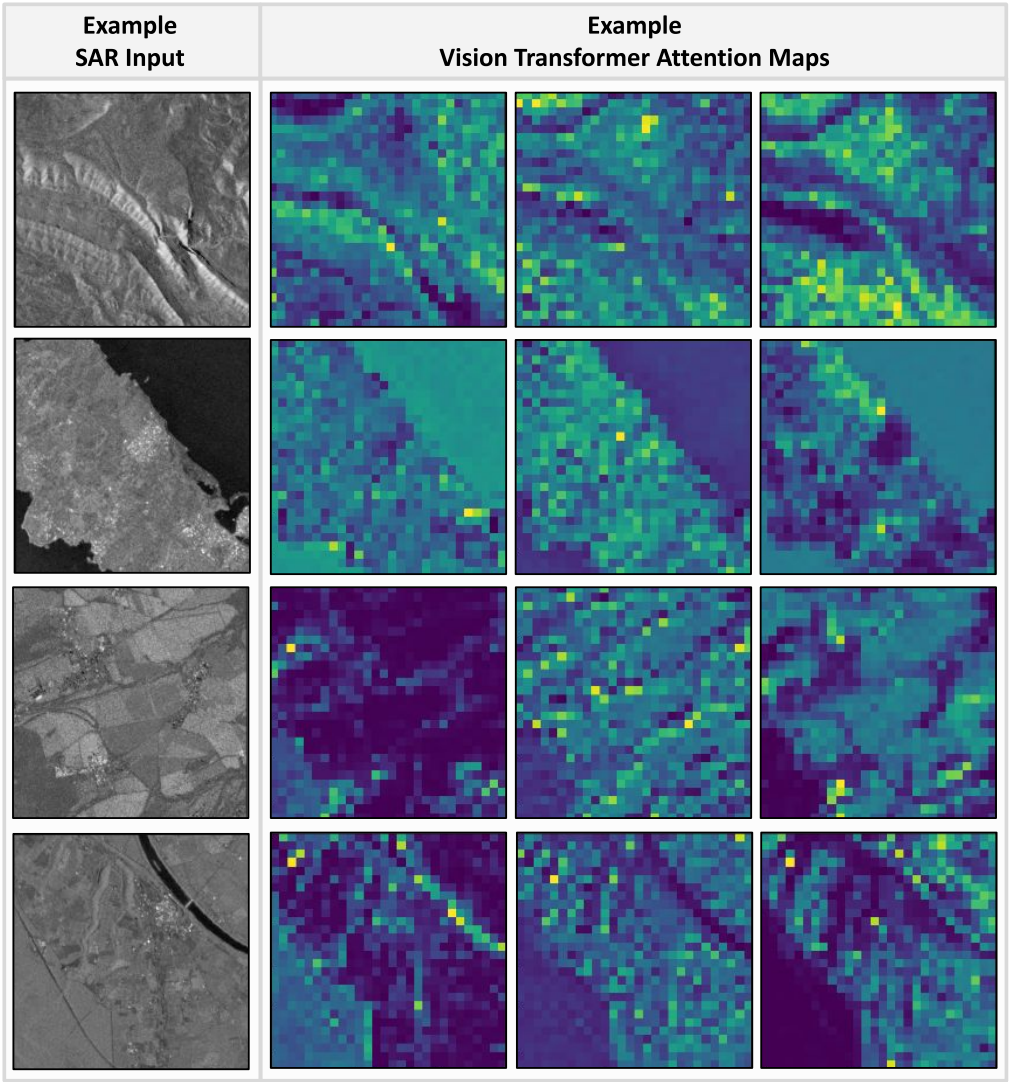

Figure 5: Example S1GRD SAR input images and ViT attention maps. The ViT used during this work have 12 attention heads that pay attention to different features in the input data.

Table 3: Root mean square error (RMSE[%]) values when fine-tuning (columns) and performing inference (rows) in each AOI. All fine-tuning was done using only 1% of available data, and all inferencing was performed on the entire test set. The values in purple refer to the models trained on S1GRD, and the values in teal refer to the models trained on GSSIC. For completeness, we also fine-tuned a model on $\sim 1\%$ of data sampled from the combined area of Europe, Conus, and China, shown in the last row.

	Conus	Europe	China	South America
Conus	6.06 / 8.19	9.53 / 15.56	22.05 / 15.00	14.26 / 27.36
Europe	18.29 / 13.80	5.53 / 7.22	20.93 / 19.70	25.24 / 30.40
China	27.93 / 31.90	20.71 / 25.34	3.74 / 8.77	26.77 / 27.22
South America	21.12 / 19.88	16.36 / 21.10	17.18 / 12.60	12.72 / 11.87
Conus/Europe/China	6.33 / 11.33	5.77 / 9.71	4.38 / 10.60	18.95 / 14.49

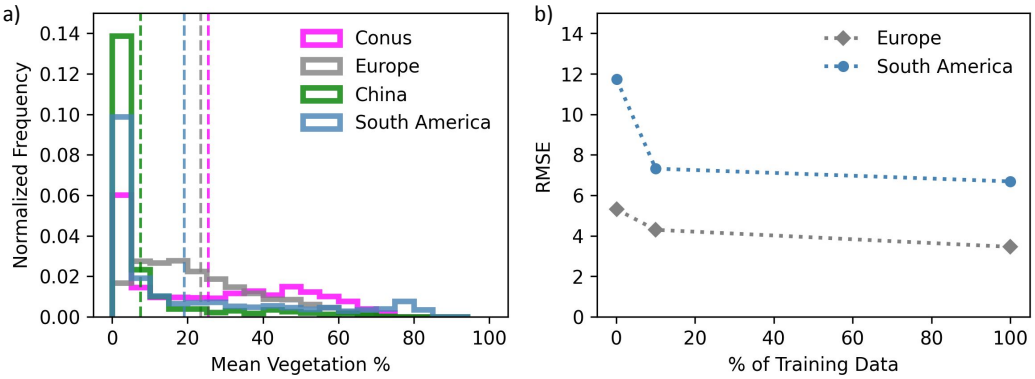

Figure 6: a) Histogram distributions of the mean vegetation percentage labels for each AOI, and b) root mean square error (RMSE[%]) in predicting mean vegetation percentage after fine-tuning on Europe (grey) or South America (blue) using 1%, 10% or 100% of available data. Analyzing the distributions of vegetation percentage in different AOIs can offer insights into the distribution learned by the model during the fine-tuning process. We observe that for Europe, which was part of the pre-training set, the improvement in RMSE[%] when fine-tuning on 100% compared to 1% of data is very small. Even for South America, which was not part of the pre-training set, good model performances are obtained for small data fine-tuning.

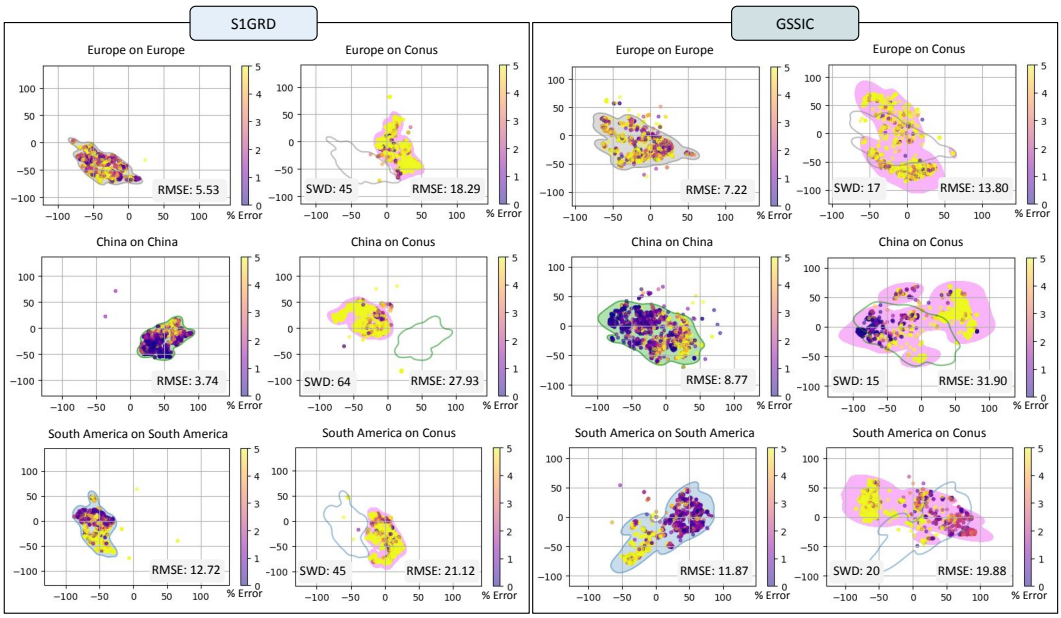

Figure 7: Embedding space and prediction errors (in %) for models pre-trained on S1GRD or GSSIC. The first (third) column shows the embeddings and prediction error for 2000 example tiles in the respective fine-tuning areas for S1GRD (GSSIC), while the second (fourth) column shows the embeddings and prediction errors when performing inference of the fine-tuned models on 2000 tiles in Conus. The Sliced Wasserstein Distance (SWD) of the embedding distributions and the root mean square error (RMSE[%]) for inference on the entire test is included.

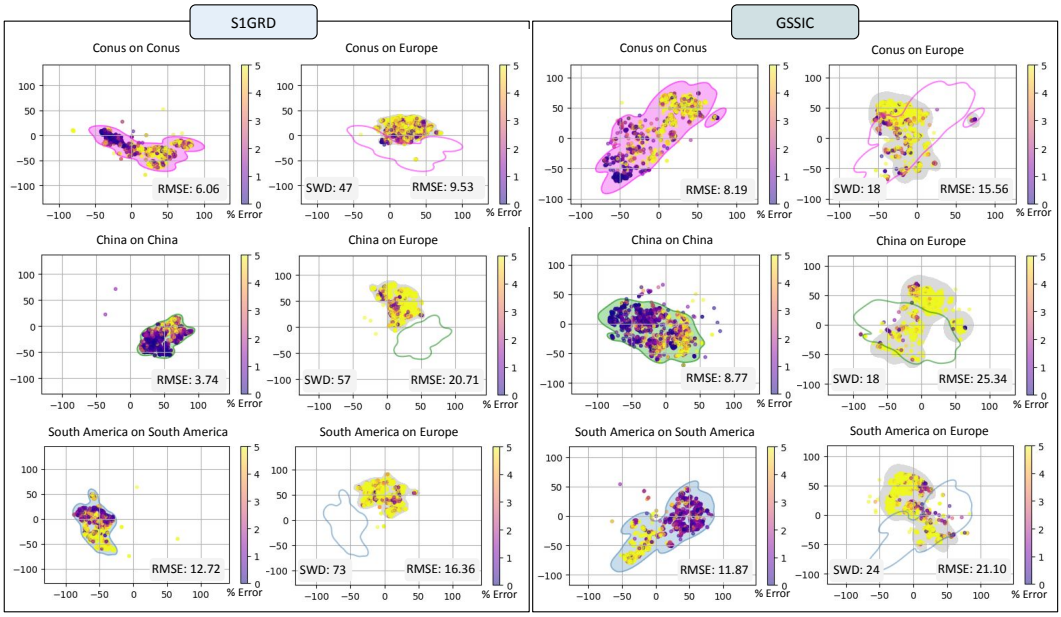

Figure 8: Embedding space and prediction errors (in %) for models pre-trained on S1GRD or GSSIC. The first (third) column shows the embeddings and prediction error for 2000 example tiles in the respective fine-tuning areas for S1GRD (GSSIC), while the second (fourth) column shows the embeddings and prediction errors when performing inference of the fine-tuned models on 2000 tiles in Europe. The Sliced Wasserstein Distance (SWD) of the embedding distributions and the root mean square error (RMSE[%]) for inference on the entire test is included.

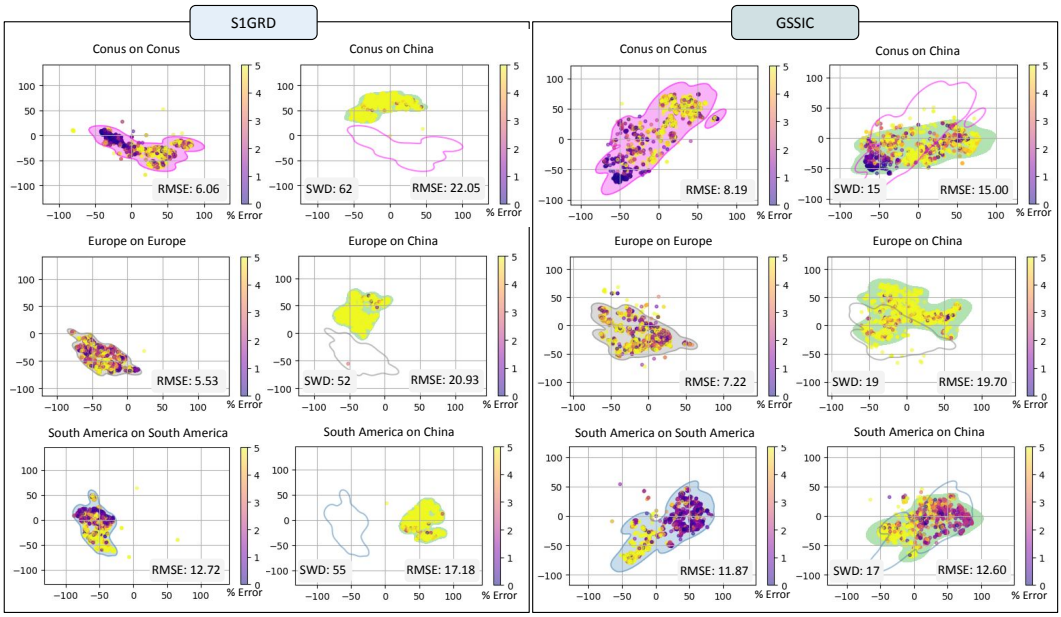

Figure 9: Embedding space and prediction errors (in %) for models pre-trained on S1GRD or GSSIC. The first (third) column shows the embeddings and prediction error for 2000 example tiles in the respective fine-tuning areas for S1GRD (GSSIC), while the second (fourth) column shows the embeddings and prediction errors when performing inference of the fine-tuned models on 2000 tiles in China. The Sliced Wasserstein Distance (SWD) of the embedding distributions and the root mean square error (RMSE[%]) for inference on the entire test is included.