Exploring DINO: Emergent Properties and Limitations for Synthetic Aperture Radar Imagery

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

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

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

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

Self-supervised learning (SSL) models have recently demonstrated remarkable performance across various tasks, including image segmentation. This study delves into the emergent characteristics of the Self-Distillation with No Labels (DINO) algorithm and its application to Synthetic Aperture Radar (SAR) imagery. We pre-train a vision transformer (ViT)-based DINO model using unlabeled SAR data, and later fine-tune the model to predict high-resolution land cover maps. We rigorously evaluate the utility of attention maps generated by the ViT backbone and compare them with the model's token embedding space. We observe a small improvement in model performance with pre-training compared to training from scratch and discuss the limitations and opportunities of SSL for remote sensing and land cover segmentation. Beyond small performance increases, we show that ViT attention maps hold great intrinsic value for remote sensing, and could provide useful inputs to other algorithms. With this, our work lays the groundwork for bigger and better SSL models for Earth Observation.

1 Introduction

Self-supervised learning (SSL) models have gained attention in recent years for their ability to learn from unlabeled data and their knowledge transferability to unseen domains. State-of-the-art SSL models, such as Self-**Di**stillation with **No** Labels (DINO) [1], Contrastive Language-Image Pretraining (CLIP) [2], and Masked Autoencoders (MAE) [3], have shown remarkable performance and generalizability in several domains. For example, DINO algorithms using vision transformer (ViT) [4] backbones have shown great performance on segmentation tasks without requiring labels [1]. This knowledge transferability could improve the characterization results of earth observation (EO) and remote sensing of our changing planet. In addition to optical imagery, Synthetic Aperture Radar (SAR) data, collected via active satellites, has proven immensely valuable, due to its all-weather, day-and-night monitoring capabilities. SAR imagery can be used for a variety of tasks, such as land cover segmentation, vegetation estimation, and flood detection. Leveraging the information contained in the wealth of SAR data greatly benefits from computational techniques like machine learning

ICBINB: I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models at NeurIPS 2023.

(ML). However, using supervised ML algorithms is challenging, due to the large amounts of labeled data required [5]. Here, SSL has shown great promise (e.g. [6]).

In this work, we perform SSL on SAR imagery using ViT-based DINO models. After pre-training, we fine-tuned the model backbones to create high-resolution land cover maps. We showcase the models' ability to detect emergent features and discuss DINO's strengths and limitations when applied in this context. More specifically, we compare the performance of using ViT attention maps or token embeddings for model fine-tuning. In addition, we extensively discuss how pre-training and fine-tuning on small datasets compares to training the same architectures, or simpler supervised models, from scratch. Beyond comparing model performances, we showcase the intrinsic value and emerging segmentation of ViT attention maps when applied to large amounts of unlabelled SAR data. With this, our work highlights the opportunities and limitations of SSL for EO and lays the groundwork for future efforts aiming to create planetary-scale foundation models for remote sensing applications.

2 Previous Work

Foundational models, in which a pretext task is performed, have shown enormous advantages in problems with small or sparsely labeled data sets [7]. The attention mechanism used by most foundational models has shown remarkable properties not only in large language models but in computer vision [3, 2, 1]. The potential usefulness of these foundational models could be carried over to remote sensing, where early attempts have shown promising results [8, 9, 10]. However, most of the work on remote sensing was trained and tested with optical data, unlike the present work, where we pre-trained the models with Synthetic Aperture Radar (SAR) imagery.

3 Model Architecture

In this section, we explain the pre-training task model and the top-down task model for the foundational DINO model. The pretraining task model is pre-trained using only unlabeled datasets and is subsequently fit using a top-down dataset where the dataset is assumed to be smaller compared to the pretraining dataset.

3.1 Pre-training Task Model

Figure 1 shows a diagram of the ML pipeline employed in this work. The DINO model used for pre-training is structured as a teacher-student network, where the weights of the teacher are updated using an exponential moving average (ema) of the student. To preserve SAR specifics, we only perform horizontal flipping and cropping as transformations during pre-training. While many architectures can be used for the student and teacher networks, we use ViTs for their attention mechanism. This attention mechanism yields a token embedding space, and attention maps as a by-product. The attention maps, produced by each head of the ViT, capture distinct features and patterns in the unlabelled SAR data. The output of the final block of the attention mechanism is referred to as the class token. In our experiments, we omit the class token and utilize all other tokens to generate an embedding space. Both the attention maps and token embeddings can be used as input to other ML architectures.

3.2 Downstream Task Decoders

After pre-training, we use the student backbone to fine-tune the model for land cover segmentation. We either generate ViT attention maps as input to a simple U-Net [11], or decode the token embedding space of the ViT backbone using a convolutional decoder. The pixel-level classification of the model is compared to the ground truth land cover maps using a cross-entropy loss. We used two different approaches to decode the embeddings created by DINO. Firstly, we use the ViT attention maps as input to a simple U-Net. Our U-Net consists of a series of down-conversions (containing 3 double convolution blocks (Conv2d, Batch Norm, Relu, Conv2d, Batch Norm, Relu) with max pooling), and a series of up-conversions (containing 4 blocks of bilinear upsampling and double convolutions), followed by a final 2D convolution to predict the 11 distinct classes of the ESAWC maps. The same U-Net is used for our baseline experiments. Secondly, we use the ViT token embedding space as

encoding for predicting high-resolution land cover maps. In this case, the decoder is a series of convolutional upsampling blocks (Conv2d, Batch Norm, Relu, bilinear upsampling), to get from the token embeddings to the 448x448 pixels ESAWC predictions.

4 Experimental Evaluation

Three experiments were performed and are shown in this section. All experiments were tested on the ESA World Cover segmentation dataset. The first experiment uses a small pre-trained ViT-based DINO model and tests it on the ESA's dataset. The second experiment uses a larger pre-trained ViT-based DINO model and tests it on a larger portion of the ESA's dataset. The third experiment is a comparison of a pre-trained U-Net ViT-based DINO model with a supervised U-Net model.

4.1 Dataset

We prepared ML-ready datasets of SAR imagery and high-resolution land cover maps for three areas of interest (AOIs): Europe, China, and the Continental United States (CONUS). SSL pre-training was either performed just in Europe or all three AOIs. All models were fine-tuned in Europe, using small labeled datasets. The data was partitioned into tiles (448×448 pixels), using `geetiles`[1] and `sartiles`[2], and all tiles were divided into training (60%), validation (20%), and test (20%) sets based on geographic bands (see Figure 2 in the SI). To keep the data volume manageable, we only used data from 2020. We processed SAR amplitude from the Sentinel-1 Level 1 Ground Range Detected (S1GRD) dataset. SAR amplitude measures the intensity of vertically (V) or horizontally (H) polarized pulses and back-scatter. As input to our model, we calculated and stacked seasonal averages of VV, VH, and the logarithmic difference (VV-VH), totaling 12 channels. To prepare our labeled dataset for model fine-tuning, we used the ESA's World Cover (ESAWC) dataset. ESAWC provides global land cover maps at 10 m resolution and contains 11 distinct land cover classes.

4.2 Experiment 1: Benefits & Limitations of Small Pre-trained ViT-based DINO model

To investigate the impact of SSL for downstream land cover segmentation, we pre-trained two models using the DINO architecture; a model with ViT Tiny backbones (3 attention heads, 6.1 million parameters) was trained on Europe alone, while a model with ViT Base backbones (12 attention

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

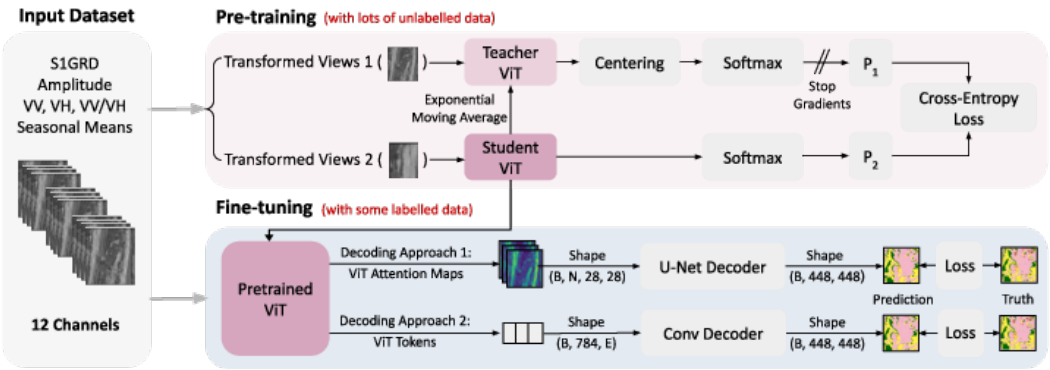

Figure 1: Schematic diagram of the ML pipeline employed in this work. We use Synthetic Aperture Radar (SAR) amplitude as input to DINO. The images undergo transformations involving multiple scaling and cropping operations before being passed to the student-teacher network. The goal of the student-teacher network is to learn the same embedding space. Centering of the teacher network is imperative to prevent model collapse. Following pre-training, the student ViT is used as the backbone for fine-tuning. Here, we compare two different approaches: we either use (1) the ViT attention maps as an input to a U-Net, or (2) the ViT token embedding space and a convolutional decoder to create high-resolution land cover maps.

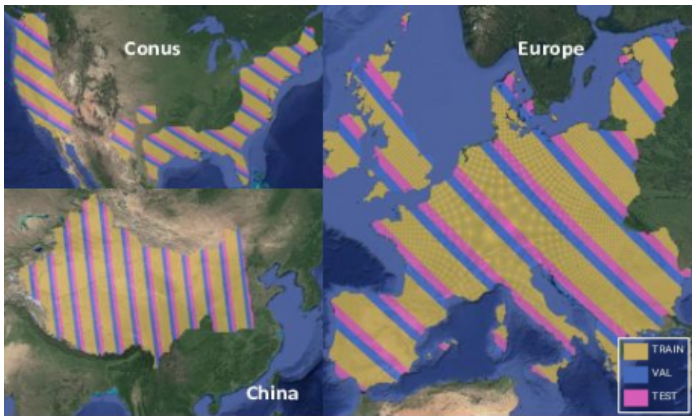

Figure 2: Data splits for CONUS, China, and Europe. There are 167K image tiles for CONUS, 285K tiles for China, and 200K tiles for Europe. We divided the data into training (60%), validation (20%), and test (20%) sets based on geographic bands to minimize data leakage across contiguous tiles.

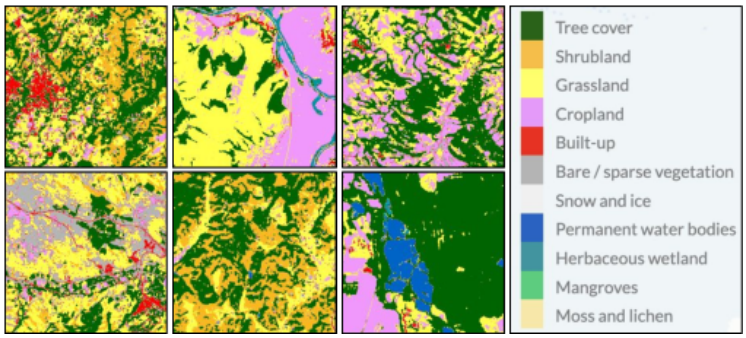

Figure 3: Example ESA World Cover maps. Each map contains pixel-level classification across 11 distinct classes. The maps are derived from Sentinel 1 and Sentinel 2 imagery. Each tile measures 448x448 pixels, with a spatial resolution of 10 meters.

heads, 88 million parameters) was trained on data from Europe, CONUS, and China. All ViTs encoded information in patches of size 16x16 pixels, leading to attention maps of 28x28 pixels.

Throughout our fine-tuning experiments, we optimized the binary cross-entropy between the model outputs and ground truth ESAWC maps and reported the mean intersection over the union (MIOU) as the performance metric. Table 1 (top) presents the results of land cover segmentation when using attention maps or token embeddings as the encoding. The model was either trained from scratch or using the ViT backbone pre-trained in Europe. In order to investigate model performances when labeled datasets are limited, we fine-tuned all models on 0.1%, 1%, or 10% of available data in Europe.

Results and discussions Using attention maps, we observe a 4% (3.5%) improvement in MIOU with the pre-trained ViT compared to training from scratch on 0.1% (1%) of data. However, with 10% of available data, we no longer see a performance improvement with pre-training. In contrast, when using the token embedding space as encoding, we observe a more substantial improvement of 9% (9.5%) in MIOU with pre-training compared to training from scratch on 0.1% (1%) of data. Similar to the attention map approach, this difference disappears when using 10% of the fine-tuning data.

Notably, models with a frozen backbone, where the ViT component is not updated during fine-tuning, consistently yield inferior results. Unsurprisingly, performances increase when more data is used for fine-tuning. The token embeddings generally lead to better performances than when attention maps are used as encoding. However, the maximum 9.5% improvement in MIOU remains relatively modest, begging the question whether the added computational cost of pre-training is worth this small improvement.

4.3 Experiment 2: Large foundational pre-trained ViT-based DINO model

We next explore whether larger ViTs pre-trained on data from diverse geographic regions improve the separability between pre-training and training from scratch. Table 1 (bottom) presents the results obtained when pre-training a larger ViT-based DINO model on S1GRD images from Europe, CONUS, and China.

Results and discussions Notably, with just 0.1% of fine-tuning data, training from scratch yields better results than using the pre-trained backbone for both encoding approaches. The pre-trained ViT with attention map encoding demonstrates a 4% (2%) improvement over training from scratch when fine-tuning on 0.1% (10%) of data. In contrast, the pre-trained model with token embeddings exhibits a more substantial 14% (10%) improvement over training from scratch with 1% (10%) of data.

We hypothesize that using a larger ViT may require a larger fine-tuning dataset to fully leverage the information learned during pre-training. In line with the results presented for pre-training on Europe, freezing the pre-trained backbone of the multi-region model consistently leads to inferior results.

4.4 Experiment 3: Pretrain Foundational model against Supervised model

In order to fully evaluate the performance of our models, we compare our approach to a supervised U-Net baseline, where instead of using the attention maps or token embeddings as input, we directly translate the 12-channel S1GRD SAR imagery to ESAWC maps.

Results and discussions Table 4.4 shows that adding attention maps as an additional input further improves the MIOU, leading to the overall best results presented here. Interestingly, for small dataset sizes (<10%) a simple U-Net baseline leads to much better results. With 10% of fine-tuning data, equally good results are achieved for the U-Net baseline and our larger ViT using token embeddings. We therefore hypothesize that the benefits of pre-training become more apparent when larger pre-training and fine-tuning datasets are used.

Table 1: Comparison of using attention maps or token embeddings for predicting ESAWC maps, when pre-training on only Europe (top table; ViT Tiny) or Europe, CONUS, and China (bottom table; ViT base). The mean intersection over the union (MIOU) was calculated across the entire test set. All models were fine-tuned on 0.1%, 1%, or 10% of training data, for 100 - 200 epochs.

Europe	Attention Maps			Tokens		
	0.1%	1%	10%	0.1%	1%	10%
From scratch	0.278	0.348	**0.398**	0.248	0.377	0.430
With pre-training (backbone frozen)	0.160	0.209	0.249	0.245	0.317	0.403
With pre-training (backbone not frozen)	**0.290**	**0.360**	0.396	**0.275**	**0.395**	**0.431**

Europe, CONUS, China	Attention Maps			Tokens		
	0.1%	1%	10%	0.1%	1%	10%
From scratch	**0.310**	0.387	0.449	**0.293**	0.354	0.476
With pre-training (backbone frozen)	0.205	0.324	0.369	0.254	0.338	0.436
With pre-training (backbone not frozen)	0.267	**0.403**	**0.460**	0.286	**0.403**	**0.477**

Table 2: Comparison of a supervised U-Net baseline, to a U-Net with attention maps as an additional input. The reported metric is the intersection of the union (MIOU). The U-net baseline was trained from scratch on 12 channels of S1GRD (VV, VH, VV/VH for each season) to predict ESAWC maps. This is compared to a U-Net trained on 15 channels, i.e, 12 channels of S1GRD, plus 3 attention maps from a vision transformer. The vision transformer was trained on 100% of training data in Europe.

	U-Net (only S1GRD)			U-Net (S1GRD + Attention Maps)		
	0.1%	1%	10%	0.1%	1%	10%
From scratch	0.376	0.412	0.475	**0.380**	0.426	**0.480**
With pre-training (backbone frozen)	-	-	-	0.371	0.430	0.474
With pre-training (backbone not frozen)	-	-	-	0.378	**0.432**	0.477

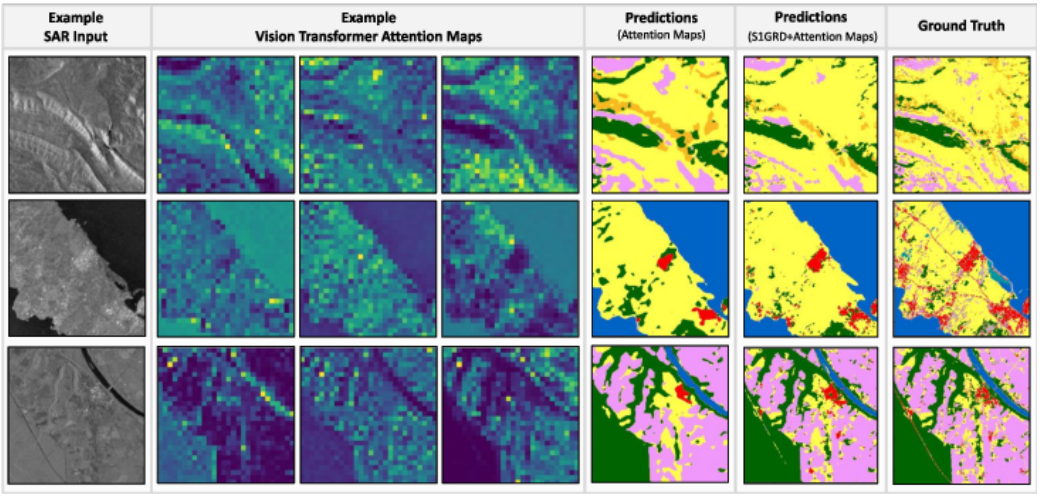

Figure 4: Example SAR input (column 1), ViT attention maps (columns 2-4), land cover predictions (columns 5-6), and ground truth ESAWC maps. The predictions were either generated using only attention maps, or S1GRD + attention maps as input to a U-Net. Both models were pre-trained and fine-tuned on 10% of data in Europe, without freezing the backbone. For all land cover maps, the different colors correspond to different classes, as indicated in the SI.

5 Conclusions

In this manuscript, we have drawn several noteworthy conclusions on the use of SSL for remote sensing and land cover segmentation. Firstly, we found that pre-training ViT-based DINO models leads to small improvements in downstream task performance compared to training from scratch, especially when using the ViT tokens embedding space as encoding. Using only ViT attention maps as input for fine-tuning shows a limited ability to produce details in high-resolution land cover maps. Here, the choice of ViT patch size emerges as a crucial factor, with smaller patch sizes leading to higher-resolution attention maps and potentially better results for tasks that require great spatial detail. Nonetheless, attention maps could provide great value as inputs for other ML algorithms. Notably, while a supervised U-Net baseline generally outperformed our ViT-based encoding-decoding architectures, the integration of attention maps as an additional input yields promising results, indicating the potential for synergistic effects. While further experimentation into SSL for EO is needed, particularly for planetary-scale foundational models, we see attention maps as a useful tool for achieving fully explainable and generalizable models, especially when applying learned knowledge to diverse downstream tasks with limited labeled data.

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] 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.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Cheng Zhong, Yue Liu, Peng Gao, Wenlong Chen, Hui Li, Yong Hou, Tuohuti Nuremanguli, and Haijian Ma. Landslide mapping with remote sensing: challenges and opportunities. *International Journal of Remote Sensing*, 41(4):1555–1581, 2020.

[6] Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets sar: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):143–172, 2021.

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[8] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022.

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

[10] 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.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

7 Supplementary Information

Table 3: Overview of the training parameters used for DINO pre-training using ViT tiny or ViT base.

Parameter	ViT Tiny	ViT Base
Center Momentum	0.99	0.99
Embedding Dimension	192	768
Number of Heads	3	12
Learning Rate	0.000001	0.000001
Student Temperature	0.03	0.03
Teacher Temperature	0.001	0.001
Warm-up Teacher Temperature	0.01	0.01
Warm-up Teacher Epochs	5	5
Model Parameters	6.1 M	88.8 M

Intrinsic Value of Attention Maps.

To visually inspect the results, Figure 4 shows example SAR inputs, ViT attention maps, model predictions, and ground truth ESAWC maps. Column 4 shows the predictions using just the attention maps as input, while column 5 shows the results for a model trained with S1GRD + attention maps.

Firstly, we observe that the ViT attention maps clearly capture features and patterns in the SAR data. Secondly, we see that predictions using S1GRD + attention maps show greater detail than if only attention maps are used as input to the U-Net. All attention maps span 28x28 pixels, which is likely too small to predict high-resolution land cover maps with great detail. While the ViT patch size could in principle be reduced to obtain higher-resolution attention maps, this would greatly increase the required memory during training.

However, even at low resolutions, the ViT attention maps clearly contain valuable information derived from the SAR input data. Figure 5 shows how well features and patterns detected by the ViT match satellite imagery (e.g. Sentinel 2) and derived data products for a variety of downstream tasks (e.g. vegetation, biomass, or built area estimation). Consequently, we believe that attention maps are useful inputs to other ML algorithms, and will prove valuable tools for model explanability and generalizability in the context of SSL or EO.

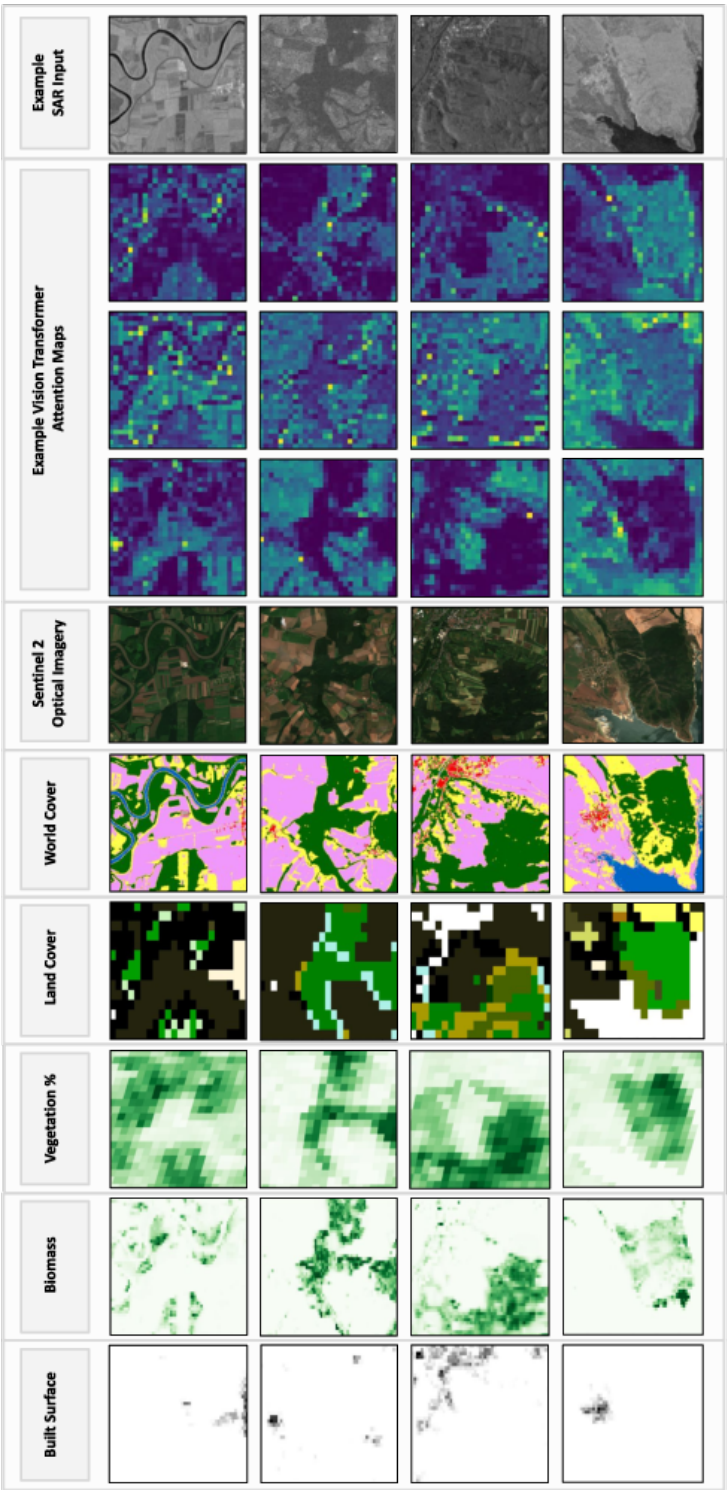

Figure 5: Comparison of ViT attention maps and various satellite imagery and derived data products. The SAR input (row 1) was used to pre-train the ViT. Looking at 3 example attention maps (rows 2-4) we clearly see emergent segmentation. Each attention head focuses on different features and patterns in the data. When comparing these features to other types of data, such as Sentinel 2 optical imagery (row 5), ESA World Cover maps (row 6), ESA CCI Medium Resolution Land Cover maps (row 7), estimations of vegetation percentage (row 8), biomass (row 9) or built surface (row 10), we clearly see that many of the features detected by the ViT translate to prominent features in the various downstream datasets.