Large Scale Masked Autoencoding for Reducing Label Requirements on SAR Data

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

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

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

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

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

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

Satellite-based remote sensing is instrumental in the monitoring and mitigation of the effects of anthropogenic climate change. Large scale, high resolution data derived from these sensors can be used to inform intervention and policy decision making, but the timeliness and accuracy of these interventions is limited by use of optical data, which cannot operate at night and is affected by adverse weather conditions. Synthetic Aperture Radar (SAR) offers a robust alternative to optical data, but its associated complexities limit the scope of labelled data generation for traditional deep learning. In this work, we apply a self-supervised pretraining scheme, masked autoencoding, to SAR amplitude data covering 8.7% of the Earth's land surface area, and tune the pretrained weights on two downstream tasks crucial to monitoring climate change - vegetation cover prediction and land cover classification. We show that the use of this pretraining scheme reduces labelling requirements for the downstream tasks by more than an order of magnitude, and that this pretraining generalises geographically, with the performance gain increasing when tuned downstream on regions outside the pretraining set. Our findings significantly advance climate change mitigation by facilitating the development of task and region-specific SAR models, allowing local communities and organizations to deploy tailored solutions for rapid, accurate monitoring of climate change effects.

1 Introduction

Satellite remote sensing has fundamentally changed the way we address climate change, offering large scale, high resolution data for applications such as forest mapping [1][2], wildfire monitoring [3] and flood detection [4]. Data from such tasks is crucial to address problems caused by climate change, but is restricted by the limitations of optical sensing. The inability of these sensors to operate at night, through cloud cover and without atmospheric interference mean that optical data is inappropriate

for time sensitive tasks such as natural disaster management [4]. Synthetic Aperture Radar (SAR) overcomes these limitations, providing more consistent, all-weather, day-night monitoring capabilities. These enhanced capabilities are invaluable for timely intervention in situations including extreme weather [5], natural disasters [6], rapid ecological shifts [7], and deforestation [8], all of which have implications for climate change.

While SAR's robust capabilities offer a promising avenue for overcoming the challenges associated with optical sensors, it comes with its own set of complexities. The technical demands associated with processing SAR data, including aspects like coherence estimation and interferogram formation, make it challenging to apply conventional machine learning techniques. Such hurdles limit the ease of generating labeled data for supervised learning, limiting SAR's effectiveness in automated analysis.

Self-supervised learning offers the advantage of learning directly from the input data without requiring ground truth labels. Methodologies such as those based on contrastive learning [9], masked image modelling [10] and knowledge distillation [11] have achieved remarkable successes on RGB image data, yielding significant improvements in various tasks such as image classification, segmentation, and object detection [12][13], reducing dependency on labelled data. Despite these advancements, the application of self-supervised learning to SAR data remains relatively unexplored [14]. Approaches based on data fusion with other remote sensing data sources such as RGB satellite or aerial imagery exist [15][16], but although these approaches can exploit the specifics of SAR data, they may not be robust to the absence of usable RGB data at night or under cloud cover. A small number of methods operating solely on SAR data exist [17][18][19], but have not yet clearly shown the geographic or temporal generalisabilty often lacking in remote sensing models [20]. Applying self-supervised learning directly to the large amounts of available unlabelled SAR data would allow practitioners to circumvent the limitations posed by the absence of reliable RGB data at night or in cloudy conditions - improving accuracy and response time in areas such as disaster management and environmental monitoring. Moreover, the use of large-scale, geographically diverse data with a model large enough to accommodate it has the potential to overcome the generalisability issues that often plague remote sensing models, presenting a robust alternative solely based on SAR data.

In this work, we take a self-supervised pretraining scheme - masked autoencoding [10] - that has been proven effective on curated RGB imagery, and apply it to polarimetric SAR data on a large set of data covering 8.7% of the Earth's land surface. We finetune the pretrained model on two downstream tasks - vegetation cover prediction (per-image regression) and land cover classification (semantic segmentation). We show that, in all cases, pretraining improved performance on downstream tasks. We also show that models initialized with pretrained weights still outperform their randomly initialized counterparts when using substantially fewer labels. We show that the pretrained model generalised well to regions that were not seen in the pretraining set.

2   Methods

2.1   Data

Split   Our data comprises four areas of interest (AOIs) - China, the Continental United States (CONUS), Europe and South America. Of these AOIs, three comprise the pretraining set (Europe, CONUS, China). For each AOI, we divide imagery and labels into tiles of size 4480m×4480m and split the resulting tiles using geographic bands into train, validation and test sets, to avoid data leakage on contiguous tiles as much as possible (Appendix A.1). We used data from 2020 exclusively in this work to avoid the computational expense of preprocessing datasets from multiple years. Since the Earth is finite, it is feasible to pretrain a model on the entire planet, so future work should focus whether our approach also generalises temporally.

Input Data   For all tasks, we derived input data from ESA Sentinel-1 Level 1 Ground Range Detected SAR (S1GRD) amplitude data, tiled from Google Earth Engine using `geetiles`[1] We used seasonal averages (spring, summer, autumn, winter) in two acquisition modes and their logarithmic difference (VV, VH, VV-VH) as input, totalling 12 channels. The resolution of S1GRD imagery is approximately 10m/pixel.

---

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

Task Labels **Vegetation cover** percentage labels were obtained from the Terra MODIS Vegetation Continuous Fields product (MODISVEG), available in Google Earth Engine as MOD44B.006[2]. The resolution of the MODISVEG product is approximately 250m/pixel. We predicted the mean value of vegetation cover within each tile (percentage area covered by vegetation). **Land cover classification** labels were obtained from the ESA World Cover product (ESAWC), also from Google Earth Engine[3]. The resolution of the ESAWC dataset is approximately 10m/pixel, and spans 11 land cover classes. We report segmentation accuracy using mean intersection-over-union (mIoU). For both tasks, we evaluated downstream performance on one region within (Europe) and one outside (South America) the pretraining set. In all cases we trained one model from scratch and one with an encoder pretrained using masked autoencoding. We do not compare results directly to other work on the same datasets to avoid conflating performance differences due to the methods and architectures we chose when developing our model with those due to differences in input data type or SAR preprocessing differences.

2.2 Models

For pretraining we used a masked autoencoder with a ViT-B [21] encoder followed by a reconstruction decoder based on [10]. We motivate our model selection by the observation that the original masked autoencoder behaves reasonably with minimal data augmentation [10]. Selecting appropriate data augmentations for SAR data introduces additional complexity compared to RGB data - for example, rotation or flipping may introduce invariance to information specific to each polarisation of the instrument. A contrastive approach based on two different SAR modes [9] - for example, polarimetry with two different polarisations or polarimetry and another mode such as coherence - may similarly neglect information specific to each mode. We therefore choose to omit comparison to additional methods in this short work, although future comparison remains of interest.

We applied two modifications to the model - we use 12 channels, as described in Section 2.1, and reduce the patch size from 32 to 16 - the same size in pixels as per the original implementation, but smaller relative to our input image size of 448×448, therefore resulting in a longer sequence input to the transformer encoder. We motivate this change with the observation that distant pixels in remote sensing imagery are less likely to be correlated than distant pixels in a curated photograph. See Appendices A.2 and A.3 for qualitative reconstruction results and hyperparameter details.

For the MODISVEG task, we replaced the reconstruction decoder with a regression head comprising 1D convolutions, of output dimension 196, in both the sequence and hidden dimensions, followed by 3 fully connected layers of sizes {512, 256, 128}. We use ReLU [22] activation functions between hidden layers and ELU [23] before the regression output.

For the ESAWC task, we followed SETR-PUP [24]. We increase the number of decoder layers compared to the original implementation to maintain a maximum upsampling of $2\times$ per layer.

3 Results

ESAWC Quantitative results for the ESAWC task are presented in Table 1, and qualitative results in Figure 1. In the fully supervised cases, both with and without pretraining and in the regions within (Europe) and outside (South America) the pretraining set, the model was skillful in classifying land cover (Best mIoU Europe: 0.533, South America: 0.426). For a fixed number of labels, in all cases, performance was improved by pretraining the encoder using masked autoencoding. The effect of pretraining on downstream performance increased when the downstream task inputs were from outside the pretraining set - using the South American data downstream, the model using the pretrained encoder and 10% of the labelled data (mIoU: 0.399) outperformed the randomly initialised, fully supervised model trained with 100% of available data (mIoU: 0.393). When performing the same ablation on the European data, the pretrained model using 10% of the downstream labels (mIoU: 0.508) did not outperform the randomly initialised, fully supervised model (mIoU: 0.522).

MODISVEG Quantitative results for the MODISVEG task are presented in Table 2, and correlation plots in Figure 2. In all fully supervised cases the model was skillful in predicting mean vegetation

---

[2] https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD44B
[3] https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v200

Table 1: **ESA World Cover (ESAWC) segmentation accuracy reported as mIoU**: European data is in the pretraining set, and South American data is unseen before the downstream task. Best results for each region in **bold**.

	Europe		South America	
	Scratch	Pretrained	Scratch	Pretrained
0.1%	0.345	0.405	0.214	0.267
1%	0.438	0.475	0.290	0.360
10%	0.486	0.508	0.339	0.399
100%	0.522	**0.533**	0.393	**0.426**
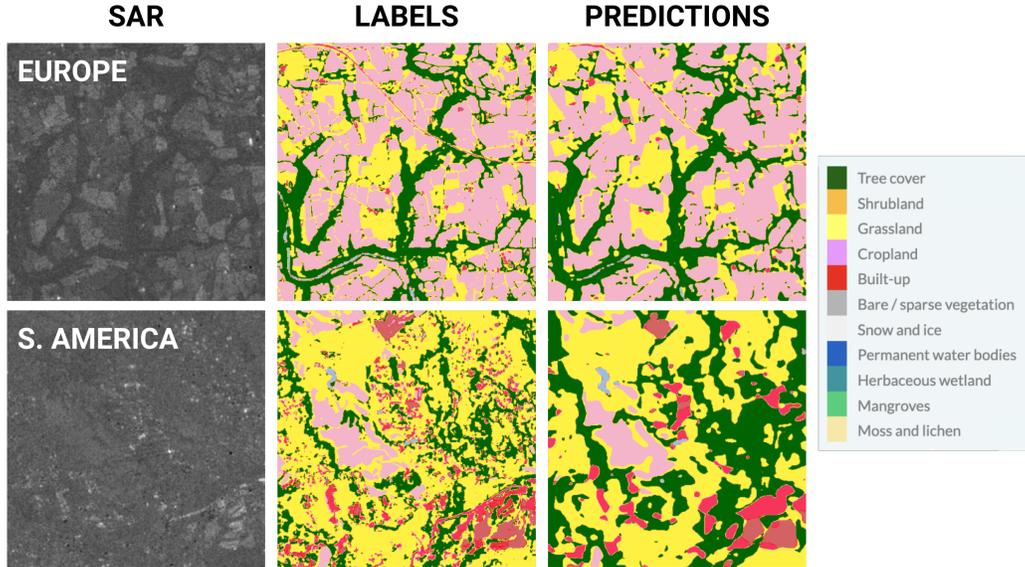
Figure 1: **Qualitative results for ESAWC land cover classification:** Land cover classification for ESAWC on data from Europe (top row) and South America (bottom row).

cover percentage, and pretraining using masked autoencoding improved performance in all cases. The effect of pretraining was very strong for this task - the pretrained model needed an order of magnitude less data than the randomly initialised model to achieve the same or greater performance for all data percentages in both regions. Again, the effect of pretraining increased in the region outside of the pretraining set (South America), with the pretrained model tuned using 1% of the task labels (RMSE 8.390 Veg %) outperforming the fully supervised randomly initialised model (RMSE 8.883 Veg %). For the model tuned on the European data, the pretrained model using 10% of the task labels (RMSE 3.282 Veg %) outperformed the fully supervised, randomly initialised model (RMSE 3.749 Veg %),

Table 2: **Terra MODIS Vegetation (MODISVEG) prediction reported as RMSE (mean vegetation cover %):** European data is in the pretraining set, and South American data is unseen before the downstream task. Best results for each region in **bold**.

	Europe		South America	
	Scratch	Pretrained	Scratch	Pretrained
0.1%	7.187	5.138	14.326	10.902
1%	5.671	4.082	12.415	8.390
10%	4.171	3.282	10.116	7.256
100%	3.749	**3.032**	8.883	**5.943**

although the model tuned with 1% of the task labels did not (RMSE 4.082 Veg %). It is unclear if the improved label efficiency for regions outside the training set is due to geographic diversity or due to the encoder being trained on a larger combined set of pretraining and training tiles.
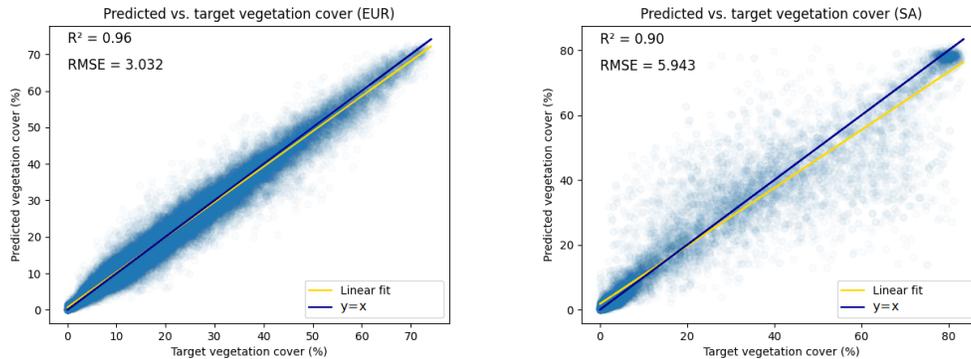

Figure 2: **Correlation plots for MODISVEG prediction finetuning the pretrained model with 100% of the labelled data:** Europe (left column) and South America (right column). Linear fits obtained by ordinary least squares (OLS).

4   Conclusions

Satellite remote sensing with Synthetic Aperture Radar (SAR) offers significant advantages over optical sensors, notably the ability to operate in all-weather conditions and during both day and night. These capabilities are essential for timely responses in climate change mitigation and natural disaster management. Processing and labelling this data, however, is subject to substantially more complexity. In this context, we showed that self-supervised pretraining on SAR data using masked autoencoding dramatically reduces the label requirements for effective performance in downstream tasks. The benefits of pretraining were particularly pronounced for geographic regions not seen during pretraining. By reducing label requirements and improving geographic generalisability, our work enables the application of deep learning to SAR for all-weather, day-night monitoring - significantly improving our capability to address climate change on a near-real-time basis. This enhanced monitoring frequency is crucial during extreme weather events, natural disasters, and rapid ecological changes, allowing for more timely intervention and mitigation strategies.

References

[1] Markus Immitzer, Francesco Vuolo, and Clement Atzberger. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sensing*, 8(3):166, February 2016. ISSN 2072-4292. doi: 10.3390/rs8030166. URL `http://www.mdpi.com/2072-4292/8/3/166`.

[2] Anders U. Waldeland, Øivind Due Trier, and Arnt-Børre Salberg. Forest mapping and monitoring in Africa using Sentinel-2 data and deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 111:102840, July 2022. ISSN 1569-8432. doi: 10.1016/j.jag.2022.102840. URL https://www.sciencedirect.com/science/article/pii/S1569843222000425.

[3] Xikun Hu, Yifang Ban, and Andrea Nascetti. Sentinel-2 MSI data for active fire detection in major fire-prone biomes: A multi-criteria approach. *International Journal of Applied Earth Observation and Geoinformation*, 101:102347, September 2021. ISSN 1569-8432. doi: 10.1016/j.jag.2021.102347. URL https://www.sciencedirect.com/science/article/pii/S0303243421000544.

[4] Angelica Tarpanelli, Alessandro C. Mondini, and Stefania Camici. Effectiveness of Sentinel-1 and Sentinel-2 for flood detection assessment in Europe. *Natural Hazards and Earth System Sciences*, 22 (8):2473–2489, August 2022. ISSN 1684-9981. doi: 10.5194/nhess-22-2473-2022. URL https://nhess.copernicus.org/articles/22/2473/2022/.

[5] Friederike E.L. Otto, Geert Jan Van Oldenborgh, Jonathan M. Eden, Peter A. Stott, David J. Karoly, and Myles R. Allen. The Attribution Question. *Nature Climate Change*, 6(9):813–816, August 2016. ISSN 1758-678X. doi: 10.1038/nclimate3089.

[6] Ben Clarke, Friederike Otto, Rupert Stuart-Smith, and Luke Harrington. Extreme Weather Impacts of Climate Change: An Attribution Perspective. *Environmental Research: Climate*, 1(1):012001, June 2022. ISSN 2752-5295. doi: 10.1088/2752-5295/ac6e7d. URL https://dx.doi.org/10.1088/2752-5295/ac6e7d. Publisher: IOP Publishing.

[7] Craig D. Allen, David D. Breshears, and Nate G. McDowell. On Underestimation of Global Vulnerability to Tree Mortality and Forest Die-Off from Hotter Drought in the Anthropocene. *Ecosphere*, 6(8):art129, 2015. ISSN 2150-8925. doi: 10.1890/ES15-00203.1. URL https://onlinelibrary.wiley.com/doi/abs/10.1890/ES15-00203.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/ES15-00203.1.

[8] Fanny Moffette, Jennifer Alix-Garcia, Katherine Shea, and Amy H. Pickens. The Impact of Near-Real-Time Deforestation Alerts Across the Tropics. *Nature Climate Change*, 11(2):172–178, February 2021. ISSN 1758-6798. doi: 10.1038/s41558-020-00956-w. URL https://www.nature.com/articles/s41558-020-00956-w. Number: 2 Publisher: Nature Publishing Group.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs].

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

[12] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks, August 2022. URL http://arxiv.org/abs/2208.10442. arXiv:2208.10442 [cs].

[13] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models, June 2022. URL http://arxiv.org/abs/2205.01917. arXiv:2205.01917 [cs].

[14] 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*, September 2022. doi: 10.48550/arXiv.2206.13188. URL http://arxiv.org/abs/2206.13188. arXiv:2206.13188 [cs].

[15] Yuxing Chen and Lorenzo Bruzzone. Self-Supervised SAR-Optical Data Fusion of Sentinel-1/-2 Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. ISSN 1558-0644. doi: 10.1109/TGRS.2021.3128072. URL https://ieeexplore.ieee.org/abstract/document/9614157?casa_token=IFz7EwnWRncAAAAA:IWgKysklytWiT4jG_SQjA_TPbBj8W8vh7BARKqg_evLBYdfptu3cLAVpFkp1rRWL7e3ccRF8. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[16] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. ISSN 1558-0644. doi: 10.1109/TGRS.2022.3194732. URL `https://ieeexplore.ieee.org/abstract/document/9844015?casa_token=WT6lAEysCCMAAAAA:JogMRpJAY1TME0QIJ4bBRvdAcejCCM7kIZ8v7WmEF2Ikj1n4h8XapQksh1GNbp-ZGZdUUDb9`. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[17] Bo Ren, Yangyang Zhao, Biao Hou, Jocelyn Chanussot, and Licheng Jiao. A Mutual Information-Based Self-Supervised Learning Model for PolSAR Land Cover Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9224–9237, November 2021. ISSN 1558-0644. doi: 10.1109/TGRS.2020.3048967. URL `https://ieeexplore.ieee.org/abstract/document/9329052?casa_token=UUF6qpE-r0QAAAAA:mceBByxyEM_behWfPXyKv9oZ2z-vtlX30ruUuVy2QupiQcl9-Rlea-ACcY69DLArLnGbjGx5`. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[18] Zaidao Wen, Zhunga Liu, Shuai Zhang, and Quan Pan. Rotation Awareness Based Self-Supervised Learning for SAR Target Recognition with Limited Training Samples. *IEEE Transactions on Image Processing*, 30:7266–7279, 2021. ISSN 1941-0042. doi: 10.1109/TIP.2021.3104179. URL `https://ieeexplore.ieee.org/abstract/document/9515580?casa_token=BWO3M9mYZXoAAAAA:i-CkOsGLldD0a1HbPcHzLhreO_QUKvsZIdI7n8zhp76j-XqTIJ3QxoglHI8_4QJMp8EC00-F`. Conference Name: IEEE Transactions on Image Processing.

[19] Yanjie Xu, Hao Sun, Jin Chen, Lin Lei, Kefeng Ji, and Gangyao Kuang. Adversarial Self-Supervised Learning for Robust SAR Target Recognition. *Remote Sensing*, 13(20):4158, January 2021. ISSN 2072-4292. doi: 10.3390/rs13204158. URL `https://www.mdpi.com/2072-4292/13/20/4158`. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

[20] Anastasiia Safonova, Gohar Ghazaryan, Stefan Stiller, Magdalena Main-Knorn, Claas Nendel, and Masahiro Ryo. Ten Deep Learning Techniques to Address Small Data Problems with Remote Sensing. *International Journal of Applied Earth Observation and Geoinformation*, 125:103569, December 2023. ISSN 1569-8432. doi: 10.1016/j.jag.2023.103569. URL `https://www.sciencedirect.com/science/article/pii/S156984322300393X`.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, June 2021. URL `http://arxiv.org/abs/2010.11929`. arXiv: 2010.11929.

[22] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU), February 2019. URL `http://arxiv.org/abs/1803.08375`. arXiv:1803.08375 [cs, stat].

[23] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), February 2016. URL `http://arxiv.org/abs/1511.07289`. arXiv:1511.07289 [cs].

[24] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, July 2021. URL `http://arxiv.org/abs/2012.15840`. arXiv:2012.15840 [cs].

A   Supplementary Material

A.1   Data Split

We used repeated geographic bands to define our training, validation and test sets. These bands can be seen for the four AOIs in Figure A.1. Coverage was determined by intersection with the coverage of the ARIA S1 GUNW dataset[4], which was not used in this work. This approach minimises data leakage compared with a fully randomised split, while also reducing the train-test distribution shift that would occur when using one geographically contiguous band for each set. Data was split into the training, validation and test sets at a 60:20:20 ratio. A total of 737,050 tiles were generated, spanning an area of $1.4793 \times 10^7 \text{km}^2$. See Table A.1 for a breakdown by AOI.
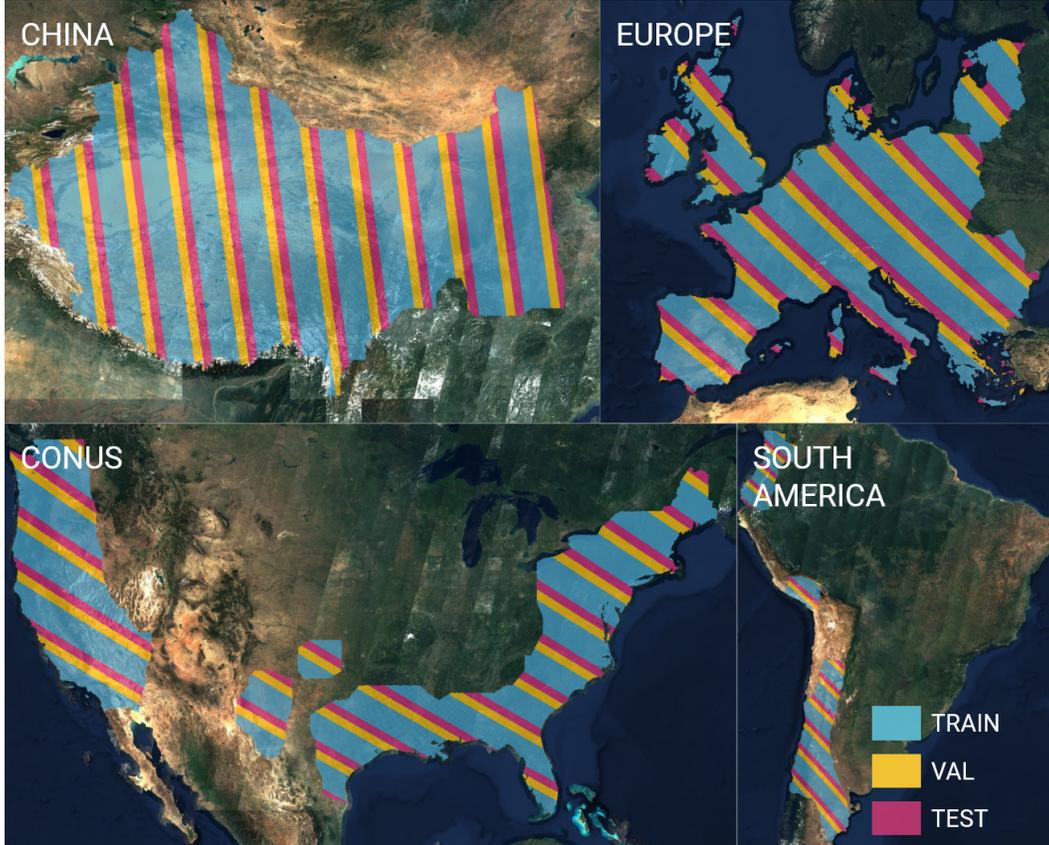

Figure A.1: **Data split**: Geographic bands for training/validation/test sets at a ratio of 60:20:20.

Table A.1: **Statistics for S1GRD dataset tiles:**. Regions in the pretraining set are shown in **bold**. Regions used in downstream tasks are shown in *italics*.

AOI	Total No. Tiles	Area (km^2)	% Earth's land surface	Size (GB)
China	**285402**	**5.728×10^6**	**3.7%**	**1740**
Conus	**167403**	**3.360×10^6**	**2.3%**	**1003**
Europe	***200489***	***4.024×10^6***	**2.7%**	***1228***
South America	*83756*	*1.681×10^6*	1.1%	*502*
Pretrain	**653294**	**12.112×10^6**	**8.7%**	**3971**
Total	737050	14.793×10^6	9.8%	4473

---

[4]`https://asf.alaska.edu/data-sets/derived-data-sets/sentinel-1-interferograms/`

A.2 SAR Reconstructions

Reconstructions by the masked autoencoder of SAR data masked during pretraining can be seen in Figure A.2. Note that the explicit aim of pretraining is to learn input features, not to obtain high reconstruction accuracy. The model largely predicts low-frequency features, as in [10].
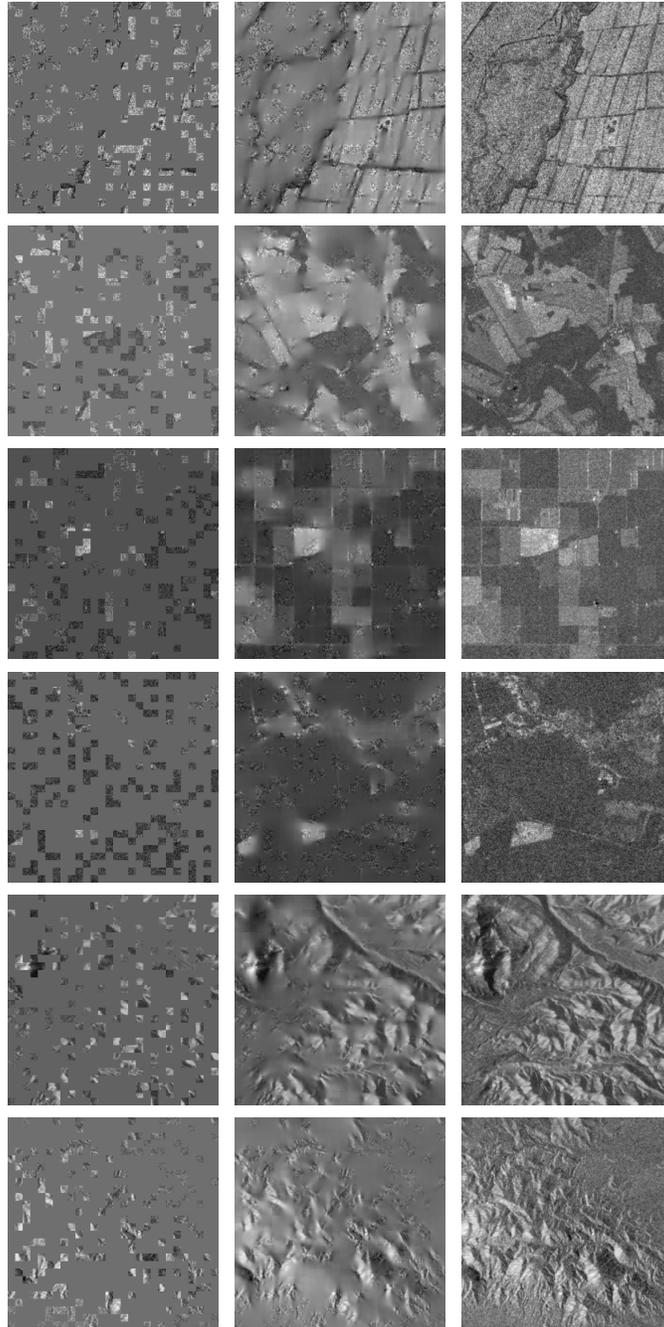

Figure A.2: **Masked SAR Reconstruction:** Masked autoencoder-based reconstruction of SAR amplitude imagery from the validation set. Within each row, we show the masked image (left), reconstruction (centre) and original image (right). A masking ratio of 0.75 was applied to patches of size 16×16 on images of size 448×448.

A.3 Training Details

Hyperparameter details for MAE-based pretraining and task finetuning can be seen in Table A.2. The patch size was halved relative to the size of the image compare to [10], based on the intuition that distant pixels in remote sensing imagery are less likely to be correlated than distant pixels in curated photographs. This intuition appeared to be evidenced by linear probe performance on the validation data for vegetation prediction, although we do not report the results from the linear probe as it was not used on all tasks or regions. Beyond this, the probe was not used to make further design decisions. Increasing or decreasing the learning rate by an order of magnitude did not improve convergence on the validation data. No significant hyperparameter tuning was undertaken beyond these two decisions, as the computational expense of performing an equal level of tuning for all tasks and regions was too high.

Table A.2: **Hyperparameter details** for MAE pretraining and downstream training on MODISVEG and ESAWC.

	MAE Pretraining	**MODISVEG**	**ESAWC**
Encoder	ViT-B [21]	ViT-B [21]	ViT-B [21]
(Params)	(88.8M)	(88.8M)	(88.8M)
Decoder	Reconstruction [10]	1×1 Conv + FC	SETR-PUP [24]
(Params)	(5.5M)	(Section 2.2; 19.7M)	(3.0M)
Loss Function	MSE	RMSE	Cross Entropy
Input Image Size	448×448	448×448	448×448
Output Image Size	448×448	Scalar	448×448
Patch Size	16×16	16×16	16×16
Masking Type	Random	N/A	N/A
Masking Ratio	0.75	N/A	N/A
Optimiser	AdamW	AdamW	AdamW
Learning Rate	1.00E-04	1.00E-04	1.00E-04
No. Epochs (Pretrain)	75	N/A	N/A
No. Epochs (0.1%)	N/A	75	100
No. Epochs (1%)	N/A	50	75
No. Epochs (10%)	N/A	50	50
No. Epochs (100%)	N/A	25	35

A.4 Training Histories

Training histories for MODISVEG and ESAWC can be seen in Figures A.3 and A.4 respectively. We did not tune the optimiser extensively, beyond changing the learning rate to achieve a reasonable rate of convergence. Note that the model was not evaluated before the first epoch (Epoch 0), so datapoints indicated at Epoch 0 are after one epoch of training.
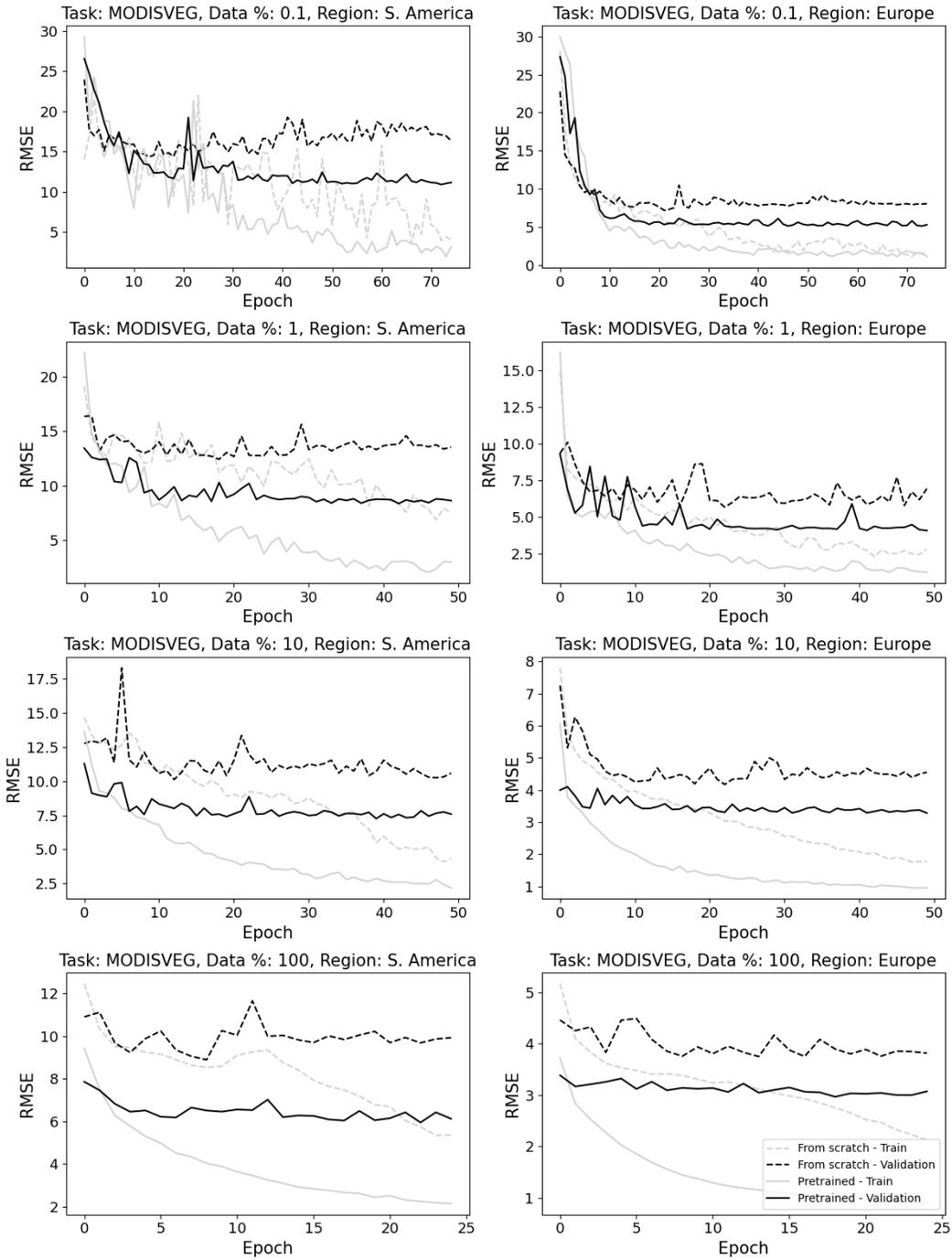
Figure A.3: Training histories for Terra MODIS Vegetation prediction. **Datapoints indicated at Epoch 0 are after one epoch of training. Models were not evaluated before the first epoch.**
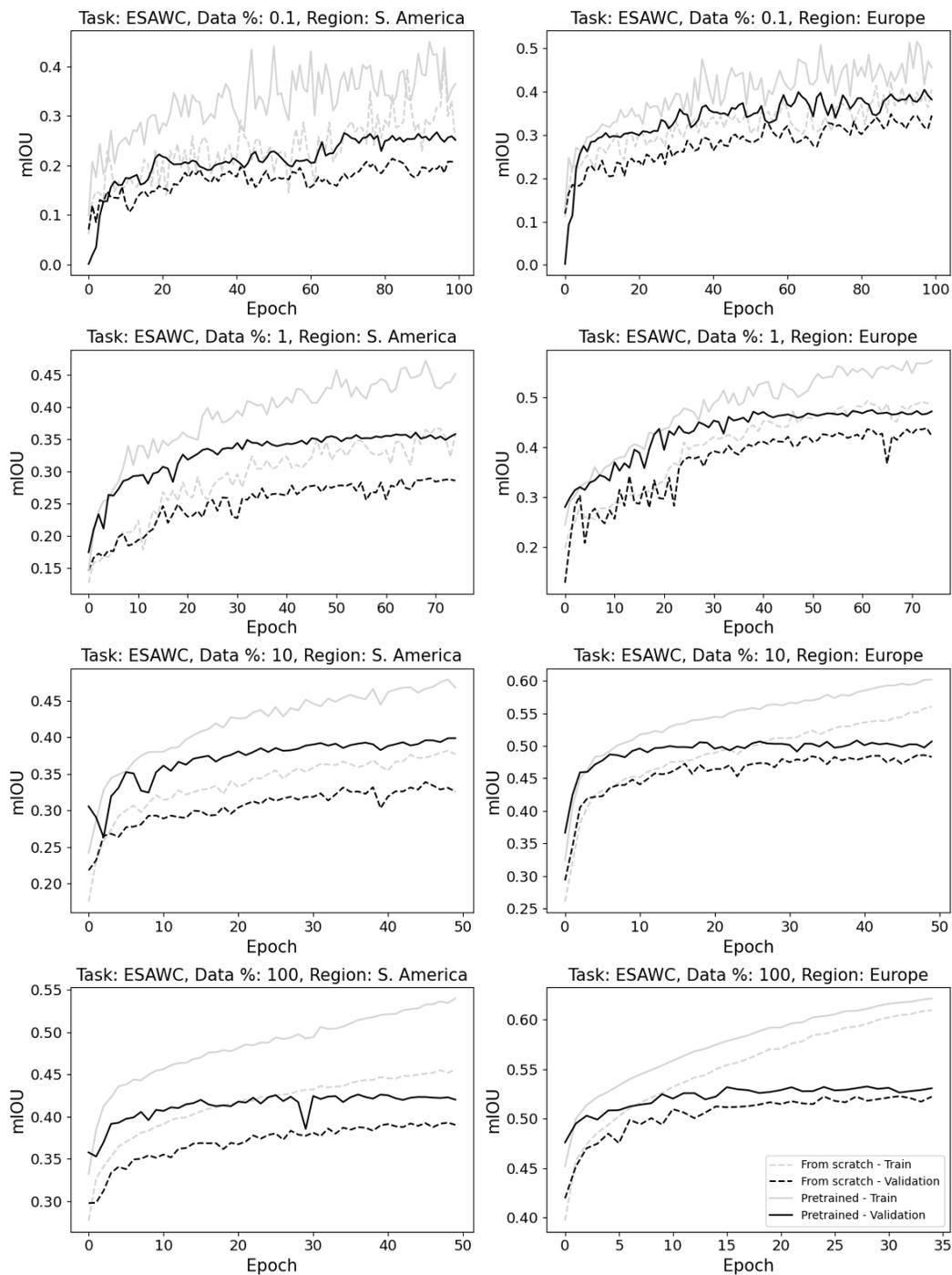
Figure A.4: Training histories for ESA World Cover prediction. **Datapoints indicated at Epoch 0 are after one epoch of training. Models were not evaluated before the first epoch.**