Brewing Stronger Features: Dual-Teacher Distillation for Multispectral Earth Observation

Supplementary Material

A. Pretraining

In this section, we will expand on the pretraining methodology and details introduced in Section 3. We will first provide more insight into our pretraining methodology, followed by details on hyperparameters to aid in reproducibility.

A.1. Contrastive self-distillation

The full loss for the *coding rate regularizer* [21] described in Section 3.1 can be formulated as:

$$\mathcal{L}_{\text{CR}} = -\frac{p+NB}{pNB}\,\frac{1}{V}\sum_{i=1}^{V}\log\det\left(\boldsymbol{I}_p + \frac{p}{BN\varepsilon}\,z_i^\top z_i\right), \tag{1}$$

where $z_i^\top z_i$ represents the covariance matrix, and $\log\det$ is calculated using the Cholesky expansion, i.e.,

$$\log\det(A) = 2\sum_{i=1}^{p}\log L_{ii}. \tag{2}$$

Here, L_{ii} are diagonal elements of the matrix L that satisfies the Cholesky decomposition $A = LL^\top$. B represents the batch size, while N is the number of used GPUs. p is the dimension of the features z after projection (256 for MS learning). Since $\mathcal{L}_{\text{CR}}$ is calculated only on the global student and teacher views, $V = 2$. Finally, ε is a small constant used for balancing, here 0.05. The first factor $\frac{p+NB}{pNB}$ is a heuristic to balance the loss and can be adjusted accordingly.

Locality degrades in the contrastive-only baseline because DINO-style objectives emphasize global invariances without preserving local structure. Patch-token reconstruction (as in DINOv3) could address this, but it lacks one of DEO's key benefits: distillation of semantic priors from a general VFM. In addition, DEO implicitly restores locality through patch-token distillation, without the complexity of DINOv3's reconstruction pipeline, thereby providing a more complete and efficient solution.

A.2. Hyperparameters

In Table 1 we provide details about the hyperparameters used during pretraining. In Table 2 we provide details about the sizes of network elements.

Parameter	Value
n	2
m	10
α_1	1
α_2	0.5
α_3	0.5
γ	1
EMA	0.996
Cos scheduler WD	0.04
Base LR	0.0005
Warmup epochs	10

Table 1. Details of pretraining hyperparameters.

Element	Value
Swin	
Input patch size	4
Embedding dim.	128
Windows size	12
Projection head	
Hidden dim.	2048
Bottleneck dim. (MS)	256
Bottleneck dim. (Optical)	1024

Table 2. Network element sizes.

A.3. Distillation training data

DINOv2 [12] introduces an automatic data curation pipeline to build a diverse dataset consisting of 142 million images. It emphasizes curation, deduplication, and filtering to remove near-duplicates, low-quality content, and domain biases.

DINOv3 [17] builds upon the previous work of automatic data curation introduced in DINOv2, by introducing the larger LVD-1689M dataset. It is a large, web-curated dataset containing 1.689 billion images and providing a balanced representation of data available on the internet.

RADIOv2.5 [7] was not trained on any particular dataset, but is trained using agglomerative multi-teacher distillation. It distills from DINOv2 [12], CLIP [13], and SAM [9], combining their features and thereby distilling the datasets they were originally trained on.

B. Evaluation

In this section, we provide additional details on the datasets used for evaluation, along with a more detailed description of our evaluation methodology.

Segmentation

GEO-Bench [10]	In paper	Image Size	# Classes	Train	Val	Test	# Bands	RGB res	Sensors
m-pv4ger-seg	GB-pv	320×320	2	3000	403	403	3	0.1	RGB
m-chesapeake-landcover	GB-chesa.	256×256	7	3000	1000	1000	4	1.0	RGBN
m-cashew-plantation	GB-cas.	256×256	7	1350	400	50	13	10.0	Sentinel-2
m-SA-crop-type	GB-SA-c.	256×256	10	3000	1000	1000	13	10.0	Sentinel-2
m-nz-cattle	GB-cattle	500×500	2	524	66	65	3	0.1	RGB
Others									
SpaceNetv1 [18]	SN	224×224	2	5000	1000	1000	3	0.5	DigitalGlobe WorldView 2
Sen1Floods11 [2]	S1F11	512×512	3	252	89	90	13	10.0	Sentinel-2
PASTIS [6]	PASTIS	128×128	20	1455	482	496	10	10.0	Sentinel-2

Table 3. Details for segmentation datasets used in the paper for evaluation.

Classification

GEO-Bench [10]	In paper	Image Size	# Classes	Train	Val	Test	# Bands	RGB res	Sensors
m-bigearthnet	GB-ben	120×120	43	20000	1000	1000	12	10.0	Sentinel-2
m-so2sat	GB-s2s	32×32	17	19992	986	986	18	1.0	Sen.-2 + Sen.-1
m-eurosat	GB-es	64×64	10	2000	1000	1000	13	10.0	Sentinel-2

Table 4. Details for classification datasets used in the paper for evaluation.

Change detection

GEO-Bench [10]	In paper	Image Size	# Classes	Train	Val	Test	# Bands	RGB res	Sensors
LEVIR-CD [3]	LEVIR	256×256	2	7120	1024	2048	3	0.5	Google Earth satellite
OSCD [4]	OSCD	96×96	2	827	-	385	10	10.0	Sentinel 2

Table 5. Details for change detection datasets used in the paper for evaluation.

B.1. Datasets

Semantic segmentation. For semantic segmentation, we use a mixture of established benchmarks in the form of GEO-Bench [10] and standalone datasets. Details are provided in Table 3.

Classification. We use three multispectral classification datasets from the GEO-Bench [10] benchmark. m-bigearthnet is a multi-label classification task, while m-so2sat and m-eurosat are single-label classification tasks. We provide details in Table 4.

Change detection. We use the optical LEVIR [3] and multispectral OCSD [4] change detection datasets. Details are provided in Table 5.

B.2. Evaluation details

For all experiments, we use a batch size of 64 and fine-tune for 50 epochs using a learning rate. We provide other details in the following section.

Semantic segmentation. For all methods, we fine-tune an UPerNet [22] segmentation head on top of a frozen backbone. We extract features from four stages of the backbone: for ViT-B backbones, these are stages 3, 5, 8, and 11; for ViT-L, stages 7, 11, 15, and 23; and for Swin-based backbones, we take the four Swin stages before pooling. For ViT backbones, the latest stage is downsampled, while the first and second stages are upsampled four times and 2 times, respectively. UPerNet details are provided in Table 6.

Classification. We extract the last layer features from a frozen backbone and train a simple linear layer on top of them. For ViT backbones, we extract the last-layer class token; for Swin backbones, we pool the last-layer features to simulate a class token.

Change detection. For all evaluated methods, we first perform backbone feature fusion of a pair of images using a simple element-wise subtraction. We then pass them to a UPerNet [22], except for ViT-based methods, where we use the UNet [16] decoder, as it performs better. Following related work [15, 19], we also train the backbone for change detection. We extract features from four stages for all models: for ViT-B backbones, stages 3, 5, 7, and 11; for ViT-L, stages 7, 11, 15, and 23; and for Swin-based backbones, the four Swin stages before pooling. UPerNet details are provided in Table 6, and for UNet, we use the same setup as in [15].

Parameter	Value
Pool scales	1, 2, 3, 6
Hidden size	512

Table 6. Details of the UPerNet segmentation head.

C. Method details

We implement each evaluated method using its official repository. We use consistent learning rates per method, except for SatDiFuser [8], for which we use the provided learning rates. Learning rates are presented in Table 7. Official repositories of evaluated methods are listed in the following:

- DINOv2 [12]: https://github.com/facebookresearch/dinov2
- DINOv3 [17]: https://github.com/facebookresearch/dinov3
- Scale-MAE [14]: https://github.com/bair-climate-initiative/scale-mae
- GFM [11]: https://github.com/mmendiet/GFM
- SatDiFuser [8]: https://github.com/yurujaja/SatDiFuser
- CROMA [5]: https://github.com/antofuller/CROMA/tree/main
- TerraFM [1]: https://github.com/mbzuai-oryx/TerraFM/blob/master/terrafm.py
- Copernicus-FM [20]: https://github.com/zhu-xlab/Copernicus-FM

	LR	
Dataset	Other methods	SatDiFuser
m-pv4ger-seg	10^{-4}	10^{-2}
m-chesapeake-landcover	10^{-4}	10^{-2}
m-cashew-plantation	10^{-2}	10^{-2}
m-SA-crop-type	10^{-4}	10^{-2}
m-nz-cattle	10^{-4}	10^{-3}
SpaceNetv1	10^{-4}	10^{-2}
Sen1Floods11	10^{-4}	10^{-2}
PASTIS	10^{-1}	10^{-2}
m-bigearthnet	10^{-3}	10^{-2}
m-so2sat	10^{-3}	10^{-4}
m-eurosat	10^{-2}	10^{-2}
LEVIR-CD	10^{-4}	10^{-4}
OSCD	10^{-4}	10^{-4}

Table 7. Learning rates for datasets and methods.

D. Additional experiments

In this section, we provide additional experiments to support and ablate various claims from the paper.

D.1. Latent space alignment

We conduct a quantitative Central Kernel Alignment analysis to provide additional support for the superior latent-space alignment of our method with DINOv3 compared to MIM-based methods. On optical inputs, DEO aligns more closely with DINOv3 than Copernicus-FM (0.65 vs. 0.49), supporting our claim that objective-level compatibility with contrastive self-distillation enables effective optical knowledge transfer. On MS inputs, alignment with DINOv3 is lower (0.15 vs. 0.50), *reflecting the intended integration of additional spectral information beyond RGB*. This behavior is consistent with the use of the CR loss, which promotes integration of MS information and improves MS performance. Replacing DEO's contrastive objective with MIM (all else fixed) reduces both alignment and downstream performance (Table 8).

D.2. Coding rate experiments

We perform experiments without the CR loss (Table 8). The loss is applied to the MS branch; therefore, its omission significantly reduces MS performance. Total collapse is still avoided because the VFM teacher serves as an additional regularizer.

D.3. Coding rate experiments

The pretraining experiments with artificial shift (Table 8) show that our model is robust to misalignment.

D.4. ViT model

We train a ViT-S (patch size 8) with a parameter count comparable to Swin-T. Its performance is slightly lower than the Swin model (Table 8), mainly due to a few challenging datasets (GB-chesa., GB-cas.).

Size	Model	Optical	MS	Overall
Large models	DEO (contrast.)	81.98	57.45	69.72
	DEO (MIM)	78.18	56.39	67.28
	DEO (no CR)	82.02	54.47	68.25
Small models	DEO (Swin-T)	79.88	55.45	67.67
	DEO (ViT-S p8)	76.62	52.09	64.35
	DEO (shift)	79.86	55.56	67.71

Table 8. Additional experiments.

E. Compute and carbon footprint

We provide the compute report and estimated pretraining emissions using *carbontracker* in Table 9. DEO's inference performance is comparable to DINOv3-B, which has a similar parameter count. Patch size 4 in our Swin-based DEO yields only a small increase in compute due to windowed attention, whereas a ViT with a reduced patch size shows a larger increase.

Model	img/s	Mem. [MB]	TFLOPS	kgCO2eq
DEO (Swin-**B**)	5.917	5272	10475	34.5
DINOv3-**B**	9.80	5595	9823	18000
DEO (ViT-**S** p8)	2.43	18152	31302	

Table 9. Inference experiments.

F. Additional qualitative

We provide additional qualitative results in Figures 1 to 4.

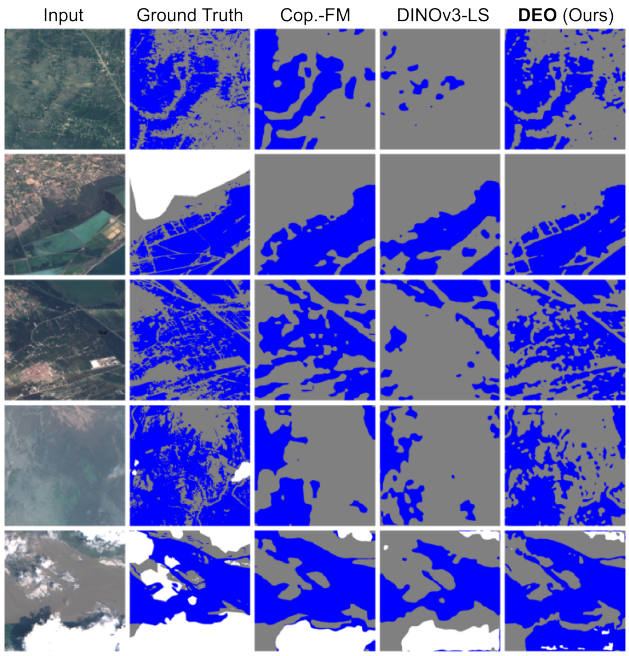

Figure 1. Extended qualitative results for Sen1Floods11.

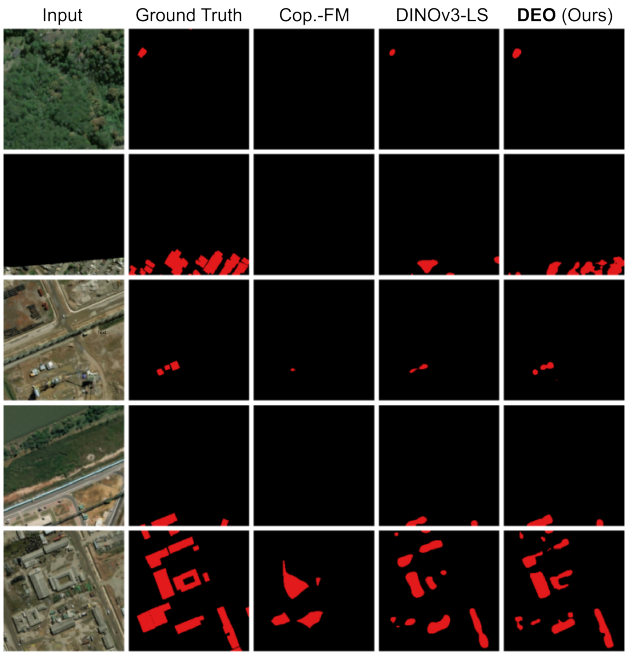

Figure 3. Extended qualitative results for SpaceNetv1.

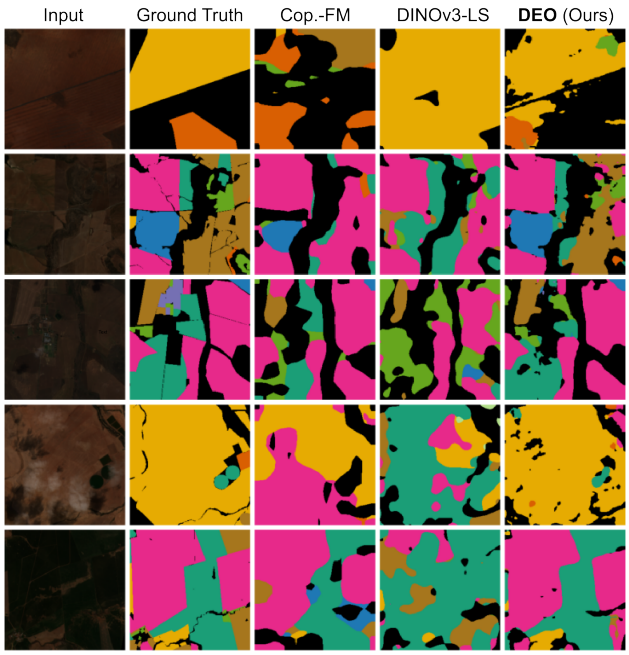

Figure 2. Extended qualitative results for m-SA-crop-type.

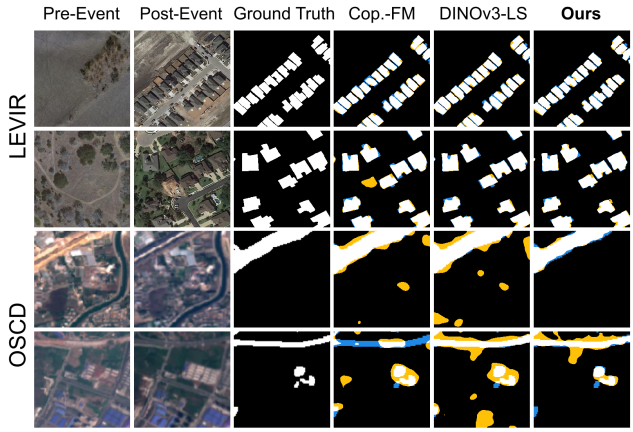

Figure 4. Extended qualitative results for LEVIR and OSCD.

References

[1] Muhammad Sohail anish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. Terrafm: A Scalable Foundation Model for Unified Multisensor Earth Observation. In *9th International Conference on Learning Representations, ICLR*, 2026. 3

[2] 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. 2

[3] Hao Chen and Zhenwei Shi. A Spatial-Temporal Attention-Based Method and A New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12:1662, 2020. 2

[4] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. In *IEEE International Geoscience and Remote Sensing Sympo-*

sium, pages 2115–2118. IEEE, 2018. 2

[5] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote Sensing Representations With Contrastive Radar-Optical Masked Autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023. 3

[6] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic Segmentation of Satellite Image Time Series With Convolutional Temporal Attention Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021. 2

[7] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved Baselines for Agglomerative Vision Foundation Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497, 2025. 1

[8] Yuru Jia, Valerio Marsocci, Ziyang Gong, Xue Yang, Maarten Vergauwen, and Andrea Nascetti. Can generative geospatial diffusion models excel as discriminative geospatial foundation models? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8429–8440, 2025. 3

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[10] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geobench: Toward Foundation Models for Earth Monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023. 2

[11] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 3

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 1, 3

[13] 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. 1

[14] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 3

[15] Blaž Rolih, Matic Fučka, Filip Wolf, and Luka Čehovin Zajc. Be the Change You Want to See: Revisiting Remote Sensing Change Detection Practices. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[17] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 3

[18] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A Remote Sensing Dataset and Challenge Series. *arXiv preprint arXiv:1807.01232*, 2018. 2

[19] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. MTP: Advancing Remote Sensing Foundation Model via Multi-Task Pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2

[20] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a Unified Copernicus Foundation Model for Earth Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[21] Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via Coding Rate Regularization. In *Forty-second International Conference on Machine Learning*, 2025. 1

[22] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 2