Generalised Distillation of Geospatial Foundation Models

Amina Said[1 [0009−0007−7752−7790]],
Julia Dietlmeier[2 [0000−0001−9980−0910]],
Margaret McCaul[2 [0000−0002−0996−8435]], and
Noel O'Connor[1,2 [0000−0002−4033−9135]]

[1] Research Ireland Centre for Research Training in Machine Learning,
Dublin City University, Ireland
amina.said3@mail.dcu.ie
[2] Insight Research Ireland Centre for Data Analytics,
Dublin City University, Ireland
{julia.dietlmeier, margaret.mccaul, Noel.OConnor}@dcu.ie

Abstract. Geospatial Foundation Models (GeoFMs) are large-scale models with millions of parameters, pre-trained on vast amounts of data to support diverse downstream tasks. However, they are typically computationally demanding, which restricts deployment in resource-limited environments. Furthermore, they are predominantly trained using optical data despite the inherent limitations of this modality due to natural phenomena such as cloud cover and atmospheric conditions. Our work aims to leverage multiple modalities by specifically transferring knowledge from a multimodal teacher using optical and radar data, to train a radar-only student model. To ensure computational efficiency, we apply offline feature-based knowledge distillation to a frozen TerraMind-B teacher to compress its multimodal knowledge into a lightweight unimodal TerraMind-S student model. Despite using a smaller student architecture, our distillation approach outperforms TerraMind's state-of-the-art (SOTA) performance on radar data. In particular, we obtain an improvement in mean intersection over union (mIoU) of 11.11% and 2.29% on the Crop Type Mapping South Sudan and Sen1Floods11 datasets, respectively. Lastly, we find that the later layers are more important in enhancing the performance of the student model and that distillation performance is thematically dependent. Code is accessible from: https://github.com/amisaid/GeoFM-GenDistill.

Keywords: Generalised Distillation · GeoFMs · Multimodality.

1 Introduction

Remote sensing, commonly referred to as Earth Observation (EO), entails capturing information about features or phenomena on the surface of the Earth at a distance without being in physical contact with them [13]. The process employs two sensing mechanisms: active and passive sensing. Active remote sensing

emits its own electromagnetic signals and records the reflected signal, it includes synthetic aperture radar (SAR) [35] and light detection and ranging (LiDAR). Conversely, passive remote sensing relies on solar illumination and the sensors only record the reflected energy by the targets. A common example is optical remote sensing, which includes multispectral and hyperspectral imagery.

In the context of computer vision, radar and optical data are referred to as modalities. Optical remote sensing is generally considered *"information rich"* as it provides multispectral information with channels ranging from visible to infrared bands. However, as a passive remote sensing technology which is dependent on solar energy, it has its limitations. In particular, it is constrained by atmospheric conditions and cloud cover, which are the primary hindrance to consistent availability of reliable optical imagery [8].

In contrast, radar remote sensing is weather-independent [15] because it uses the longer microwave wavelength. Hence, it is unaffected by weather and lighting conditions. While this ensures data continuity, radar is highly susceptible to geometric distortions, such as foreshortening and layover. The distortions are due to the side-looking orientation of the sensors. Furthermore, radar imagery often suffers from speckle noise, which affects detection, segmentation and extraction of features [23,35].

Despite the all-weather and cloud-penetrating capabilities of radar, the focus of EO has primarily been on optical imagery. The modality imbalance is evident from some Geo-Benchmark studies, such as the PANGAEA [21] which analyses 11 satellite datasets: seven are purely optical, four are multimodal and none are exclusively radar data. However, a more recent evaluation benchmark, GEO-Bench-2 [30] shows progress by including two exclusively radar datasets and five multimodal datasets that integrate both optical and radar. This advancement fosters the development of more robust models [25] by effectively leveraging the complementary strengths of both modalities.

Regardless of the imbalance in modalities, optical and radar imageries are complementary [24,18] and multimodality improves the performance of deep learning models [18,24,25,15]. Furthermore, combining these modalities can account for the limitations of each individual modality. For example, optical imagery can aid in geometric correction [23] and speckle noise filtering of the radar data. However, deep learning models require consistency between the train and test datasets [8]. In practice, having multimodal dataset during inference is not always possible due to the inconsistent availability of optical imageries. In particular, when natural disasters such as floods occur, they are typically accompanied by heavy rainfall and thick cloud cover, introducing atmospheric and weather interferences to the optical data. Consequently, these interferences obscures the optical imagery and renders multimodal feature extraction highly challenging.

Knowledge distillation (KD) is a model compression technique that facilitates training a large teacher model and transferring it learned representation to a smaller, computationally efficient student model [14]. A powerful extension of this framework is generalised distillation [8], whereby a student model learns from a teacher model trained with privileged information [19,33]. Privileged informa-

tion refers to data available exclusively during training and absent at inference. Generalised distillation has been applied in EO by training an information richer teacher model, such as a multimodal model trained on both optical and SAR imageries [8]. The knowledge is then transferred to a student model targeting a different modality such as SAR [8,10]. A closely related variant is cross-modal distillation, where the teacher and student operate entirely on different modalities. Collectively, these KD strategies can address the modality gap by distilling insights across heterogenous modalities [19,8]. This ensures that critical monitoring of features or events can continue seamlessly via radar-only inference, where leveraging this privileged information during training actively enhances the performance even when optical data is entirely obscured or unavailable at deployment

This study is tailored to overcome the challenges of missing modalities and computational constraints. To achieve this, we apply offline generalised distillation to leverage optical properties during training while *"breaking free"* from this dependence during inference by using only radar imagery. Moreover, this framework aims to enhance computational efficiency by transferring the knowledge within TerraMind geospatial foundation model (GeoFM) base variant (TerraMind-B) to a compact small model (TerraMind-S). Our contributions include:

- To the best of our knowledge, this is the first implementation of generalised distillation on a GeoFM, utilising a multispectral privileged knowledge to train a SAR-only student model.
- First study on distilling a base GeoFM teacher model into a small GeoFM student model.
- Comparison of generalised distillation performance on both uni-temporal and multi-temporal datasets.

2 Related Works

2.1 Multimodal Geospatial Foundation Models

Foundation models have recently emerged as a dominant paradigm in computer vision, a trend that is also influencing the field of EO. Most GeoFMs such as Clay-v1 [5], Prithvi-EO [16,31], among others, are mainly unimodal. Nevertheless, this is currently changing, as Hong et al. [15] highlight a significant transition towards multimodal GeoFMs. They note that in 2024, there were nearly twice as many multimodal models developed compared to unimodal (45 versus 23). This is also witnessed by the introduction of multimodal models such as TerraMind [18], OlmoEarth [12] and Galileo [32] in the year 2025. These models use two satellite modalities by incorporating both radar (Sentinel-1) and optical (Sentinel-2) imageries. Notably, TerraMind model incorporates additional heterogeneous datasets such as land cover maps, vegetation index, elevation models, texts and spatial coordinates [18]. Likewise, the OlmoEarth model extends its cross-modal capabilities to three distinct satellite sensors by also including Landsat imagery in addition to Sentinel-1 and Sentinel-2.

Marsocci et al. [21] find that using multimodal data on GeoFMs like S12-MAE, CROMA and DOFA underperforms and optical-only data achieves better performance. In particular, the study presents a mean intersection over union (mIoU) performance drop of at least 4% on the Crop Type Mapping South Sudan (CTM-SS) dataset relative to optical-only data. In contrast, Jakubik et al. [18] state that multimodality on TerraMind-B outperforms unimodal data, reporting a mIoU improvement of almost 5% on the same CTM-SS dataset over the optical-only data. Similarly, Simumba et al. [30] also report that TerraMind outperforms other models such as DOFA and Clay when using multimodal data. Given that TerraMind is one of few GeoFMs consistently shown to benefit from multimodality, our study focuses on TerraMind to leverage this advantage effectively.

2.2 Knowledge Distillation

Model compression is a useful technique that has been applied in deep learning to reduce model sizes and in some cases has shown to improve performance [29,27]. Few studies have focused on compressing GeoFMs [27,1,2]. Barreiro et al. [1] report pruning of Prithvi-EO-2 by up to 79% while retaining 94% mIoU performance on segmentation task. In another study, Said et al. [27] finds that pruning can improve segmentation performance by up to 2.38% while compressing the model by 1.4 on Clay-v1 model.

Dino et al. [8] present a radar-only student model developed from a teacher model trained on both optical and radar imageries on land-use land cover (LULC) dataset. The student is trained on SAR-only data and performance is compared to classic machine learning (ML) models such as multilayer perceptron, random forest and convolutional neural network (CNN). The study reports that the student model outperforms the other models with a mean accuracy of 65%, a difference ranging between almost 1-4% compared to the classic ML models. Similarly, Wang et al. [34] present a study on transferring knowledge from a complex vision foundation model DINOv2 to a compact student model for landslides segmentation . The study reports performance improvement of 1.75% mIoU over the state-of-the-art (SOTA). The study notes that performance is also dependent on consistency between the teacher and student architectures, i.e., CNNs vs transformer-based architectures.

Due to the aforementioned reasons, this study focuses exclusively on TerraMind models. We aim to transfer privileged optical information from the teacher model, TerraMind-B, to a smaller student model, TerraMind-S. This approach leverages the rich optical data and the multimodal TerraMind GeoFM to improve both accuracy and computational efficiency.

3 Datasets and Implementation

3.1 Datasets

The experiment is carried out on three multimodal segmentation datasets of optical and radar modalities, as shown in Table 1. The datasets are part of the PANGAEA [21] Geo-Benchmark.

Table 1. Key features of the multimodal datasets used.

Dataset	Classes	Patches	Patch Size	Bands	Feature	Year Released
CTM-SS [26,36]	4	837	64^2	12	multi-temporal	2019
PASTIS-R [9,28]	18	2433	128^2	10	multi-temporal	2021
Sen1Floods11 [3]	2	431	512^2	15	uni-temporal	2020

CTM-SS [26] is an agricultural multimodal dataset containing Sentinel-1, Sentinel-2 and high resolution imagery of Planet. For this study, we only use Sentinel-1 and Sentinel-2 imageries. Sentinel-1 is composed of two bands of polarizations vertical transmit-vertical receive (VV) and vertical transmit-horizontal receive (VH) [10]. While Sentinel-2 contains 10 bands, leaving out the coastal aerosol, water vapour and cirrus bands. CTM-SS is a satellite image time series (SITS) dataset in which each patch comprises six temporal acquisitions. We use the temporal mean across these acquisitions to mitigate the effects of atmospheric noise, cloud cover and cloud shadow, without increasing computational overhead, as explained in Section 4.1.

Sen1Floods11 [3] is a floods dataset, it features Sentinel-1 and Sentinel-2 imageries. The dataset contains two different labels: hand-labelled and weakly-supervised labels. Sentinel-1 contains three bands VV, VH and VV/VH. Sentinel-2 dataset contains full 13 bands of Sentinel-2. For our study, we leave out the band VV/VH and use full spectral bands of Sentinel-2 for high performance [21]. We use the hand-labelled and leave out the weakly-supervised labels. The hand label is split into 252 (60%), 89 (20%) and 90 (20%) images for training, validation and testing splits, respectively.

PASTIS-R [9,28] is an agricultural multimodal dataset integrating Sentinel-1 and Sentinel-2 imageries. The dataset contains 2433 patches split into 5 folds, first three folds used for training and the other two each for validation and testing. PASTIS-R is also a SITS dataset with a range of 38-61 temporal series. For this study, a mean of the minimum temporal period of 38 series is computed and used for the same reasons as CTM-SS dataset.

3.2 Models and Experimentation

The implementation is carried out using TerraTorch library [11], on NVIDIA GeForce RTX 5090 GPU with 32GB of memory. For this study, the generative,

multimodal TerraMind model is used as the teacher with its based-sized vision transformer (TerraMind-B), while the small variant (TerraMind-S) is used as a student model, as shown in Table 2. As Jakubik et al. [18] report that multi-modality in TerraMind-B outperforms uni-modality on the three datasets, this study aims to evaluate its performance under generalised distillation. Moreover, to enhance computational efficiency for deployment on resource-limited devices such as edge-devices and satellites, the multimodal representations from the base model are transferred to a lighter student model. This ensures that we bridge the modality gap by transferring privileged optical information to a unimodal model. It also bridges the capacity gap by compressing the base model into a smaller student model.

Table 2. Key features of the GeoFMs used in our experiments, parameters are expressed in millions (M).

Model	Params. (M)	Property	Year
TerraMind-S [22,18]	23	generative, multimodal	2025
TerraMind-B [18]	89	generative, multimodal	2025

Experimentation started by first fully fine-tuning (FT) each of the three multimodal datasets. The best performing FT model is used as the teacher for knowledge distillation. For consistency with the work of PANGAEA [21] and TerraMind [18], UperNet decoder is used for all models. A batch size of 8 is used on Sen1Floods11 and 16 for the SITS datasets, i.e., CTM-SS and PASTIS-R. For all models, AdamW optimiser is used with a learning rate of 1×10^{-4} and a random seed of 42. As the experimentation focuses on segmentation tasks, mean intersection over union (mIoU, Eq. (1)) is used to evaluate performance and the model is saved based on the best validation mIoU [30].

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{1}$$

While the teacher model (TerraMind-B) was fully optimised by training both its encoder and decoder, we follow the approach of Marsocci et al. [21] and Jakubik et al. [18] for the student model. For fair comparison, the TerraMind-S student encoder is also frozen and only its UperNet decoder is trained during distillation. We train the teacher model for 80 epochs, while for the student we apply an early stopping with a patience of 30 epochs on the validation loss.

Fig. 1 illustrates the feature-based knowledge distillation architecture used for this study. Since the teacher is already fully fine-tuned (FT), we freeze its weights during distillation and train only the student model. This ensures that

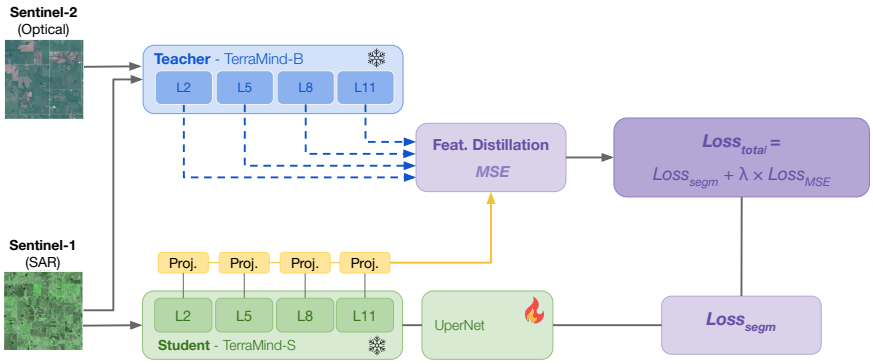


Fig. 1. Feature-based knowledge distillation implementation architecture, with a frozen teacher model and a trained frozen-encoder student model. Blue represents the teacher encoder, L is layer starting from zero and L=[2, 5, 8, 11] layers used to transfer the representations from the teacher to student (green).

the teacher's knowledge is transferred to the student without additional computational cost, as the teacher's weights remain unchanged throughout training. It reduces model complexity and improves efficiency [20] without introducing additional parameters. Freezing the teacher also ensures that knowledge transfer remains strictly unidirectional, from teacher to student [4] and prevents student's representations from distorting the teacher's learned representations [7].

Distillation is tested by transferring representations from three different layer setups: last layer only (1-layer) which is the foundational approach introduced by Hinton et al. [14]. The last layer contains the rich "dark knowledge" of the teacher model and distilling only the last layer is reported to perform better than distilling a single middle later [23]. Second setup is spreading the distillation layers to ensure student model learns from different ranges of representations [13]. The last setup is distilling the last 6 layers as they are richer in representations [4].

$$L_{\text{total}} = L_{\text{segm.}} + (\lambda) L_{\text{MSE}} \tag{2}$$

Eq. (2) presents the total distillation loss function (L_{total}) [7], which is a sum of two losses. L_{task} is the loss function from the ground truth, also known as hard labels. While L_{MSE} is the mean squared error, difference of the teacher and student features. Lambda (λ) is the weighting coefficient, balancing the trade off between the hard labels and the teacher feature representations (soft labels).

4 Results and Discussion

While prior research [22] suggests up to 10% performance decline on TerraMind-S, our study finds the opposite. Table 3 shows radar-only fully FT TerraMind-

S model slightly outperforms our TerraMind-B models on CTM-SS (+2.13%) dataset. Our performance on full FT multimodal models conforms with the findings of Jakubik et al. [18], as multimodal outperforms unimodal models of optical and radar. This suggests that the model can leverage on the strengths of both modalities to improve the performance.

Table 3. Performance comparison on frozen encoder (FE), full FT and distilled model, performance presented as mIoU (%) and trainable parameters in millions (M). Sen-1 is Sentinel-1 radar data, while Sen-2 is Sentinel-2 optical and MM is multimodality of radar and optical. Dashes (–) indicate unreported information. Bold presents best performance and underlined is second best performance.

Data	Model	Technique	CTM-SS		Sen1Floods11	
			Perf.↑	Param.	Perf.↑	Param.
Sen-1	Terra-S	Full FT	<u>25.73</u>	30.1	80.88	30.1
		FE	19.31	8.6	81.03	8.6
		Distilled	**35.56**	11.3	**82.68**	11.3
	Terra-B	Full FT	23.60	97.4	<u>81.31</u>	97.4
		FE [18]	24.45	–	80.39	–
Sen-2	Terra-B	Full FT	39.45	121.0	89.59	95.4
		FE [18]	50.90	–	89.57	–
MM	Terra-B	Full FT	51.38	99.4	90.53	103.0
		FE [18]	55.80	-	90.62	–

The distilled radar-only models outperform both the frozen encoder and fully FT models on same modality, as presented in Table 3 and Table 4. This aligns to the work of Dino et al. [8], of which the distilled model outperforms the classic machine learning models as mentioned in Section 2.2. The distilled Sen1Floods11 radar model has a performance improvement of 1.65% and 1.80% from its corresponding frozen encoder and full FT models, respectively. It also outperforms OlmoEarth-B model [12], which reports mIoU of 79.20%.

As the distilled CTM-SS dataset has a higher performance range (7.12%) as compared to the other two datasets, we report the mean value from three runs [6] with a standard deviation 4.90%. The higher performance range could be due to the high class imbalance, as out of the four classes of CTM-SS dataset, two thirds at 67% is sorghum class, while the three other classes have distribution of 10% maize, 9% rice and 7% groundnut [26]. Increasing batch size ensures the minority classes are represented [21], nonetheless, it did not improve the performance fluctuations. However, the distilled CTM-SS model has a higher performance improvement of 9.83% and 16.25% from its respective full FT and frozen encoder models.

All the three distilled radar models outperform the accuracy reported by Jakubik et al. [17] on TerraMind-B. There is a performance improvement of 11.11% on CTM-SS and 2.29% on Sen1Floods11. However, for PASTIS-R, we use the minimum available time series of 38, unlike the work of Jakubik et al. [18] which uses 6 time series. Our performance with 6 time series was low, as presented in Table 4. Moreover, the distilled CTM-SS radar-only model also outperforms CROMA and S12-MAE models from PANGAEA [21] by using linear temporal mapping strategy by 0.17% and 7.05%, respectively.

Table 4. PASTIS-R performance comparison on frozen-encoder (FE), full FT and distilled models, performance presented as mIoU (%) and trainable parameters in millions (M), temporal represents pooling strategy and series count of temporal series used. Sen-1 is Sentinel-1 radar data and MM is multimodality of radar and optical. Dashes (–) indicate unreported information. Bold presents best performance and underlined is second best performance.

Modality	Model	Techn.	Perf.↑	Params.	Temporal	Series
Sen-1	Terra-S	FE	13.25	8.1	mean	6
		FE	24.89	8.1	mean	38
		Distilled	**27.40**	9.9	mean	38
	Terra-B	FE [18]	20.04	–	–	6
		FE	24.61	9.6	mean	30
		FE	<u>26.84</u>	9.6	mean	38
MM	Terra-B	FE [18]	40.51	–	–	6
		Full FT	46.31	97	mean	30
		Full FT	46.73	188	concat	30
		Full FT	46.91	97	mean	38

Despite the performance variability of CTM-SS and PASTIS-R, the higher performance improvement on the agricultural datasets and lower improvement on Sen1Floods11 datasets is also consistent with the findings of Dino et al. [8]. The high performance improvement on the distilled agricultural datasets could be due to the model leveraging the optical information from the teacher model. Optical data is information richer and commonly used in the study of agriculture and vegetation [8], mainly due to its red and near-infrared bands.

4.1 Ablation Study

As CTM-SS and PASTIS-R are a multi-temporal datasets, while TerraMind is a uni-temporal model, we applied temporal wrapping from TerraTorch [11] to account for the extra temporal factor. We compare two temporal wrapping techniques provided in TerraTorch: concatenation and mean pooling. Concatenation technique outperforms mean pooling, however it is computationally costly as

number of parameters are much higher. For example, Table 5 shows mIoU on TerraMind-S model on 38 timestamps, concatenation has mIoU of 36.47% with 66.3M trainable parameters. In contrast, the mean model of 38 time series has mIoU of 24.89% and 8.1M trainable parameters, same number of parameters as using 6 timestamps but with better performance. Despite the 11.58% performance improvement, this results in more than eight fold trainable parameters for the 38 temporal series. While varying temporality is reported to present high performance fluctuation [30], we aim to leverage on it without additional computation cost and thus use mean temporal pooling.

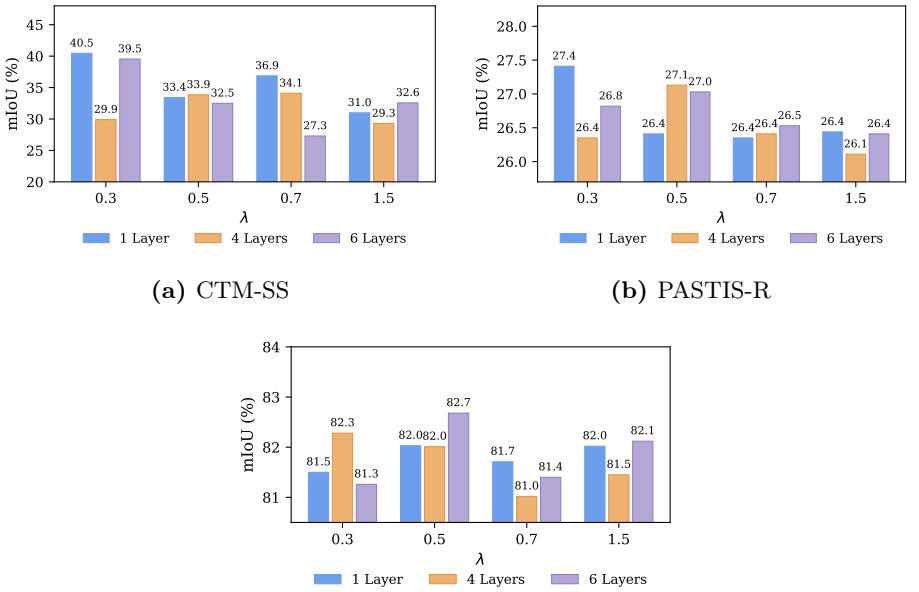
Table 5. PASTIS-R comparison of performance versus number of parameters (M), on using different temporal wrapping strategies. Series presents timestamps used in temporal wrapping with a frozen-encoder on Sentinel-1 dataset.

Model	Perf.↑	Params.	Temporal	Series
Terra-S	13.25	8.1	mean	6
	24.89	8.1	mean	38
	36.47	66.3	concat.	38
Terra-B	24.61	9.6	mean	30
	33.69	100	concat.	30

Fig. 2 shows how the model performs on distilling different layers and on different λ values on the datasets. We compared distilling the last layer only (1 layer, 12th layer), 4 layers (3rd, 6th, 9th and 12th) and the last 6 (7th - 12th) layers. For CTM-SS dataset, distilling the last layer only is best performing as shown in Fig. 2a, with an average of 33.74% across the 4 different λ values, followed by six layers at 31.33% and 4 layers at 29.86%. This may imply that the later layers are more meaningful for the student model, while the intermediate layer may be adding noise to the model. Best performing λ is 0.5 with an average of 32.00% and followed with 0.7 at 31.96% and least performing is λ 1.5 at 30.66%.

PASTIS-R distillation performance is presented in Fig. 2b. It has λ 0.5 as best performing with an average of 27.02%, followed by 0.3 at 26.51% and least performing 1.5 at 26.40%. The best performing distillation layer is the last 6 with an average of 26.70%, followed closely by 1 layer at 26.62% and lastly the 4 layers at 26.50. The distillation performance is consistent across the different layers and λ, as the difference with the best performing of each is less than one percent.

For Sen1Floods11 dataset, distilling 6 last layers is best performing as presented in Fig. 2c with a mean of 81.87%, compared to one layer (81.82%) and 4 layers (81.69%). λ value of 0.5 closely outperforms the other values compared by at least 0.38% to 0.86%, a margin of less than 1%. The distillation ablation study on Sen1Flood11 dataset is more robust and consistent as values ranged



(a) CTM-SS

(b) PASTIS-R

(c) Sen1Floods11

Fig. 2. Comparison of mIoU (%) performance on different λ and number of layers combinations. 1 layer is the last 12th layer, 4 layers are 3rd, 6th, 9th and 12th layers, while 6 layers are the last 6 layers (7th–12th).

from 81.02% to 82.68% and all outperforms almost all the models presented in Table 3.

Based on the average values from both datasets, it is clear that the later layers are more meaningful for the student model than intermediate and skipping some layers. Additionally, moderate weighting of λ=0.5 shows consistency on the three datasets. The better performance on high λ=1.5 is on the one layer for the agricultural datasets and last 6 for Sen1Floods11 dataset, which conforms with Hinton [14], as the high weights are more meaningful for the later layers.

5 Conclusions

In summary, our study finds that the TerraMind-S model is efficient on radar data, as it outperforms our CTM-SS full FT base model. Despite the high performance variation on CTM-SS dataset, the performance of all the three distilled models consistently outperform corresponding frozen-encoder models. Sen1Floods11 dataset has an improvement of mIoU 1.65%, 2.51% for PASTIS-R and up to 16.25% on the CTM-SS dataset, from respective frozen encoder models. Moreover, we find that generalised distillation is thematically dependent and highly likely to improve performance on vegetation by harnessing rich

optical information from the teacher. Additionally, the later layers have more influence on improving performance on the student model. Last but not least, we conclude that using less information and smaller models on the generalised distillation does not hinder enhancing performance and can help bridge the gap between multi-modalities. Our future work will focus on analysing how generalised distillation GeoFMs perform on other land covers such as built-up areas.

Acknowledgments. This work was funded by Taighde Éireann – Research Ireland through the Centre for Research Training in Machine Learning (18/CRT/6183). Sincere appreciation to Ms. Vicky Flanagan for her invaluable support and for facilitating acquisition of the necessary computational infrastructure.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Barreiro, V., Jakubik, J., Argüello, F., Heras, D.B.: SIMPLER: Efficient Foundation Model Adaptation via Similarity-Guided Layer Pruning for Earth Observation (Mar 2026). https://doi.org/10.48550/arXiv.2603.19873, http://arxiv.org/abs/2603.19873, arXiv:2603.19873 [cs.CV]
2. Basile, G., Jakubik, J., Blumenstiel, B., Brunschwiler, T., Moreno, J.B.: Quantizing Space and Time: Fusing Time Series and Images for Earth Observation (Jan 2026). https://doi.org/10.48550/arXiv.2510.23118, http://arxiv.org/abs/2510.23118, arXiv:2510.23118 [cs.CV]
3. Bonafilia, D., Tellman, B., Anderson, T., Issenberg, E.: Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 835–845. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPRW50498.2020.00113, https://ieeexplore.ieee.org/document/9150760/
4. Chiumento, F., Dietlmeier, J., Killeen, R.P., Curran, K.M., O'Connor, N.E., Liu, M.: Cross-Modal Knowledge Distillation for PET-Free Amyloid-Beta Detection from MRI (Apr 2026). https://doi.org/10.48550/arXiv.2604.12574, http://arxiv.org/abs/2604.12574, arXiv:2604.12574 [cs]
5. Clay Foundation: Clay Foundation Model: An Open Foundation Model for Earth Observation (2024), https://clay-foundation.github.io/model/index.html
6. Corley, I., Lehmann, N., Robinson, C., Tseng, G., Anthony Fuller, Alemohammad, H., Shelhamer, E., Marcus, J., Kerner, H.: No One Knows the State of the Art in Geospatial Foundation Models (2026). https://doi.org/10.48550/ARXIV.2605.12678, https://arxiv.org/abs/2605.12678, version Number: 1
7. Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., Zhang, H.: Decoupled Kullback-Leibler Divergence Loss (Oct 2024). https://doi.org/10.48550/arXiv.2305.13948, http://arxiv.org/abs/2305.13948, arXiv:2305.13948 [cs]
8. Dino Ienco, Gbodjo, Y.J.E., Raffaele Gaetano, Interdonato, R.: GENERALIZED KNOWLEDGE DISTILLATION FOR MULTI-SENSOR REMOTE SENSING CLASSIFICATION: AN APPLICATION TO LAND COVER MAPPING. ISPRS Annals of the Photogrammetry, Remote Sensing and

Spatial Information Sciences **V-2-2020**, 997–1003 (Aug 2020). https://doi.org/10.5194/isprs-annals-V-2-2020-997-2020, https://isprs-annals.copernicus.org/articles/V-2-2020/997/2020/

9. Fare Garnot, V.S., Landrieu, L.: Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4852–4861. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00483, https://ieeexplore.ieee.org/document/9711189/
10. Garg, S., Feinstein, B., Timnat, S., Batchu, V., Dror, G., Rosenthal, A.G., Gulshan, V.: Cross-modal distillation for flood extent mapping. Environmental Data Science **2**, e37 (2023). https://doi.org/10.1017/eds.2023.34, https://www.cambridge.org/core/product/identifier/S2634460223000341/type/journal_article
11. Gomes, C., Blumenstiel, B., Almeida, J.L.d.S., Oliveira, P.H.d., Fraccaro, P., Escofet, F.M., Szwarcman, D., Simumba, N., Kienzler, R., Zadrozny, B.: TerraTorch: The Geospatial Foundation Models Toolkit (Mar 2025). https://doi.org/10.48550/arXiv.2503.20563, http://arxiv.org/abs/2503.20563, arXiv:2503.20563 [cs]
12. Herzog, H., Bastani, F., Zhang, Y., Tseng, G., Redmon, J., Sablon, H., Park, R., Morrison, J., Buraczynski, A., Farley, K., Hansen, J., Howe, A., Johnson, P.A., Otterlee, M., Schmitt, T., Pitelka, H., Daspit, S., Ratner, R., Wilhelm, C., Wood, S., Jacobi, M., Kerner, H., Shelhamer, E., Farhadi, A., Krishna, R., Beukema, P.: OlmoEarth: Stable Latent Image Modeling for Multimodal Earth Observation (Nov 2025). https://doi.org/10.48550/arXiv.2511.13655, http://arxiv.org/abs/2511.13655, arXiv:2511.13655 [cs]
13. Himeur, Y., Aburaed, N., Elharrouss, O., Varlamis, I., Atalla, S., Mansoor, W., Ahmad, H.A.: Applications of Knowledge Distillation in Remote Sensing: A Survey (Sep 2024). https://doi.org/10.48550/arXiv.2409.12111, http://arxiv.org/abs/2409.12111, arXiv:2409.12111 [cs]
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. In: NIPS 2014 Deep Learning Workshop. arXiv (Mar 2015). https://doi.org/10.48550/arXiv.1503.02531, http://arxiv.org/abs/1503.02531, arXiv:1503.02531 [stat]
15. Hong, D., Li, C., Li, X., Camps-Valls, G., Chanussot, J.: Foundation Models in Remote Sensing: Evolving from unimodality to multimodality. IEEE Geoscience and Remote Sensing Magazine **14**(2), 10–35 (Apr 2026). https://doi.org/10.1109/MGRS.2026.3669086, https://ieeexplore.ieee.org/document/11435138/
16. Jakubik, J., Chu, L., Fraccaro, P., Gomes, C., Nyirjesy, G., Bangalore, R., Lambhate, D., Das, K., Oliveira Borges, D., Kimura, D., Simumba, N., Szwarcman, D., Muszynski, M., Weldemariam, K., Zadrozny, B., Ganti, R., Costa, C., Edwards, Blair & Watson, C., Mukkavilli, K., Schmude, Johannes & Hamann, H., Robert, P., Roy, S., Phillips, C., Ankur, K., Ramasubramanian, M., Gurung, I., Leong, W.J., Avery, R., Ramachandran, R., Maskey, M., Olofossen, P., Fancher, E., Lee, T., Murphy, K., Duffy, D., Little, M., Alemohammad, H., Cecil, M., Li, S., Khallaghi, S., Godwin, D., Ahmadi, M., Kordi, F., Saux, B., Pastick, N., Doucette, P., Fleckenstein, R., Luanga, D., Corvin, A., Granger, E.: Prithvi-100M (Aug 2023). https://doi.org/10.57967/hf/0952
17. Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S.K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R.: Foundation Models for Generalist Geospatial Ar-

tificial Intelligence (Nov 2023), http://arxiv.org/abs/2310.18660, arXiv:2310.18660 [cs]
18. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., Ramachandran, R., Fraccaro, P., Brunschwiler, T., Cavallaro, G., Bernabe-Moreno, J., Longépé, N.: TerraMind: Large-Scale Generative Multimodality for Earth Observation. In: International Conference on Computer Vision (2025), http://arxiv.org/abs/2504.11171, arXiv:2504.11171 [cs]
19. Lopez-Paz, D., Bottou, L., Scholkopf, B., Vapnik, V.: UNIFYING DISTILLATION AND PRIVILEGED INFORMATION. In: Proceedings of International Conference on Learning Representations. International Conference on Learning Representations (2016)
20. Mansourian, A.M., Ahmadi, R., Ghafouri, M., Babaei, A.M., Golezani, E.B., Ghamchi, Z.Y.: A Comprehensive Survey on Knowledge Distillation. Transactions on Machine Learning Research (Sep 2025)
21. Marsocci, V., Jia, Y., Bellier, G.L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., Nascetti, A.: PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models (Apr 2025). https://doi.org/10.48550/arXiv.2412.04204, http://arxiv.org/abs/2412.04204, arXiv:2412.04204 [cs]
22. Murphy, M.: AI-powered satellites will upend how we observe our changing planet (Oct 2025), https://research.ibm.com/blog/terramind-prithvi-tiny-small-models-geospatial, publication Title: IBM Research Blog
23. Nair, R.N., Hänsch, R.: Let me show you how it's done - Cross-modal knowledge distillation as pretext task for semantic segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 595–603. IEEE Xplore, Seattle, WA, USA (Jun 2024). https://doi.org/10.1109/CVPRW63382.2024.00064, https://ieeexplore.ieee.org/document/10678411/
24. Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., Datcu, M.: FLOOD DETECTION IN TIME SERIES OF OPTICAL AND SAR IMAGES. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLIII-B2-2020**, 1343–1346 (Aug 2020). https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1343-2020, https://isprs-archives.copernicus.org/articles/XLIII-B2-2020/1343/2020/
25. Rao, A., Rolf, E.: Using Multiple Input Modalities Can Improve Data-Efficiency and O.O.D. Generalization for ML with Satellite Imagery. In: Proceedings of The TerraBytes {ICML} Workshop. Proceedings of Machine Learning Research (PMLR) (Jul 2025)
26. Rustowicz, R.M., Cheong, R., Wang, L., Ermon, S., Burke, M., Lobell, D.: Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (Jun 2019), https://openaccess.thecvf.com/content_CVPRW_2019/papers/cv4gc/Rustowicz_Semantic_Segmentation_of_Crop_Type_in_Africa_A_Novel_Dataset_CVPRW_2019_paper.pdf
27. Said, A., Dietlmeier, J., McCaul, M., O'Connor, N.E.: When Less is More: Evaluating Structural Pruning in Geospatial Foundation Models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. IEEE Xplore (2026), https://openaccess.thecvf.com/content/

WACV2026W/CV4EO/papers/Said_When_Less_is_More_Evaluating_Structural_Pruning_in_Geospatial_Foundation_WACVW_2026_paper.pdf
28. Sainte Fare Garnot, V., Landrieu, L., Chehata, N.: Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS Journal of Photogrammetry and Remote Sensing **187**, 294–305 (May 2022). https://doi.org/10.1016/j.isprsjprs.2022.03.012, https://linkinghub.elsevier.com/retrieve/pii/S0924271622000855
29. Shinde, T.: Model Compression Meets Resolution Scaling for Efficient Remote Sensing Classification. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 1200–1209 (Feb 2025)
30. Simumba, N., Lehmann, N., Fraccaro, P., Alemohammad, H., Mel, G.D., Khan, S., Maskey, M., Longepe, N., Zhu, X.X., Kerner, H., Bernabe-Moreno, J., Lacoste, A.: GEO-Bench-2: From Performance to Capability, Rethinking Evaluation in Geospatial AI (Feb 2026). https://doi.org/10.48550/arXiv.2511.15658, http://arxiv.org/abs/2511.15658, arXiv:2511.15658 [cs.CV]
31. Szwarcman, D., Roy, S., Fraccaro, P., Gislason, T.E., Blumenstiel, B., Ghosal, R., Oliveira, P.H.d., Almeida, J.L.d.S., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Gomes, C., Kumar, A., Truong, M., Godwin, D., Lee, H., Hsu, C.Y., Asanjan, A.A., Mujeci, B., Shidham, D., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Hain, C., Kennedy, R., Zadrozny, B., Bell, D., Cavallaro, G., Watson, C., Maskey, M., Ramachandran, R., Moreno, J.B.: Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications (Feb 2025), http://arxiv.org/abs/2412.02732, arXiv:2412.02732 [cs]
32. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J.R., Shelhamer, E., Kerner, H., Rolnick, D.: Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. In: International Conference on Machine Learning Workshop. Proceedings of Machine Learning Research, Vancouver, BC, Canada (2025)
33. Vapnik, V., Izmailov, R.: Learning Using Privileged Information: Similarity Control and Knowledge Transfer. Journal of Machine Learning Research (2015)
34. Wang, S., Li, L., Dong, X., Shi, L., Tao, P.: KDA: Knowledge Distillation Adversarial Framework with Vision Foundation Models for Landslide Segmentation. IEEE Geoscience and Remote Sensing Letters pp. 1–1 (2025). https://doi.org/10.1109/LGRS.2025.3597685, https://ieeexplore.ieee.org/document/11122516/
35. Wei, T., Chen, H., Liu, W., Chen, L., Gu, P., Wang, J.: Optical and SAR Cross-Modal Hallucination Collaborative Learning for Remote Sensing Missing-Modality Building Footprint Extraction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **19**, 1183–1196 (2026). https://doi.org/10.1109/JSTARS.2025.3638382, https://ieeexplore.ieee.org/document/11271136/
36. Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D., Ermon, S.: SUSTAINBENCH: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In: 35th Conference on Neural Information Processing Systems (NeurIPS 2021) (2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/950a4152c2b4aa3ad78bdd6b366cc179-Paper-round2.pdf