Fused Foundation Model Embeddings for Earth Observation Compression: A Winning Solution to the Embed2Scale Challenge

Dávid Kerekes *
KTH Royal Institute of Technology
Division of Geoinformatics
dkerekes@kth.se

Isabelle Wittmann
IBM Research Europe
isabelle.wittmann1@ibm.com

Eric Brune
KTH Royal Institute of Technology
Division of Geoinformatics
ebrune@kth.se

Ritu Yadav
KTH Royal Institute of Technology
Division of Geoinformatics
rituy@kth.se

Valerio Marsocci
European Space Agency
Φ-lab
valerio.marsocci@esa.int

Yuru Jia
KTH Royal Institute of Technology
Division of Geoinformatics
yuruj@kth.se

Andrea Nascetti
KTH Royal Institute of Technology
Division of Geoinformatics
nascetti@kth.se

Abstract

Foundation models (FMs) are critical for compact Earth Observation (EO) data representations. We present a neural compression method that won the Embed2Scale challenge by achieving state-of-the-art performance on the NeuCo benchmark. Our method compresses multi-seasonal, multi-modal Sentinel-1/2 data into a 1024-dimensional vector by using a simple two-layer autoencoder over a diverse, concatenated ensemble of five pre-trained FMs: CROMA, DOFA, Scale-MAE, Copernicus-FM, and Prithvi 2.0. We demonstrate that merging diverse FM embeddings significantly improves performance on unknown downstream tasks, highlighting the power of FM fusion for generalizable, extreme data compression.

1 Introduction

Data compression is critical for retrieving, storing and processing an ever-increasing amount of Earth Observation (EO) data. EO-specific compression can lower costs, reduce environmental impact, accelerate research workflows, and democratize access to remote sensing data [8]. Recent advances in machine learning open new possibilities for representing EO data more compactly and meaningfully. Neural embeddings condense spectral, spatial, and temporal information into small, fixed-size vectors

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: REO: Advances in Representation Learning for Earth Observation.

that can be used directly for downstream tasks, bridging the goals of compression and representation learning.

In this paper, we present a foundation model (FM)-based neural compression method that achieved joint first place in the Embed2Scale [2] data compression challenge. We evaluate our method on the accompanying NeuCo-Bench [1] benchmark, which was used to rank solutions during the challenge. NeuCo-Bench provides an evaluation framework for general-purpose EO embeddings, measuring how much task-relevant information is retained after compression into compact representations. Instead of assessing reconstruction fidelity, it directly evaluates embeddings on diverse downstream EO tasks via linear probing, independent of the underlying encoder or backbone.

The Embed2Scale challenge extended NeuCo-Bench into a novel *embedding-only* data challenge. Participants were tasked with compressing four-season, multi-modal Sentinel-1/2 inputs into a single 1024-dimensional vector. Given the linear probing setup, these embeddings needed to retain essential input information in a task-ready form. The downstream tasks were undisclosed during the challenge, requiring representations to be both compact and broadly informative. To ensure fairness, newly curated datasets were provided for each task, all sharing a standardized Sentinel-1/2 input format.

The eight main tasks focus on regression. They include **biomass** prediction, **crop** fraction estimation for soybean and corn, **land cover** prediction for agriculture and forest, **cloud** cover fraction, and summer surface **temperature**. The three supplementary tasks are prediction of the fraction of missing data in the images and regression/classification of random noise as a sanity check.

Our method development was guided by the principles of the challenge, and focused on producing compact, reusable, and task-ready EO embeddings capable of generalizing to unknown downstream tasks.

Our submission was ranked first based on absolute performance, and a close second behind Xu et al. [16] in an alternative scoring scheme, where task weights were adjusted to reflect difficulty based on the competitors' results. Our solution achieves its result by leveraging the representational diversity of a curated ensemble of five FMs: CROMA, DOFA, Scale-MAE, Copernicus-FM, and Prithvi 2.0. We demonstrate that merging these diverse FM features significantly outperforms singular models.

2 Data

We base our experiments on the SSL4EO-S12-downstream [3] dataset, an openly available multimodal and multiseasonal EO dataset introduced alongside the NeuCo-Bench framework and used in the Embed2Scale Challenge.

The dataset follows the same structure as SSL4EO-S12 v1.1 [5] and provides harmonized Sentinel-1 and Sentinel-2 observations across four seasonal time steps per location. Each sample corresponds to one geolocation and contains a four-season data cube with dimensions ($4 \times 27 \times 264 \times 264$), where 4 denotes the seasonal timestamps (winter, spring, summer, autumn), 27 the total input channels, and 264×264 the spatial resolution. The channels comprise 2 Sentinel-1 SAR GRD (VV, VH), 13 Sentinel-2 Level-1C (TOA), and 12 Sentinel-2 Level-2A (surface reflectance) bands. All Sentinel-1 and Sentinel-2 scenes are co-registered and temporally aligned, with one observation per season.

3 Methodology

Our method (see Fig. 2) leverages embeddings from five openly available FMs, selected for their relative performance in the PANGAEA benchmark [10] to ensure comprehensive coverage of its tasks, and the available spectral bands:

- **CROMA** [7], which combines contrastive and reconstruction objectives for unimodal and multimodal representations on aligned Sentinel-1 (SAR) and Sentinel-2 (Optical) data.
- **DOFA** [15], a wavelength-conditioned, multimodal FM. Uses a hypernetwork to predict patch-embedding weights based on central wavelength, allowing a single transformer to process inputs with arbitrary band counts/sensors. Pretrained with Masking Image Modeling (MIM) and distillation.
- **Scale-MAE** [12] explicitly learns multi-scale representations by using a scale-dependent positional encoding and reconstructing low/high-frequency images at lower/higher scales.

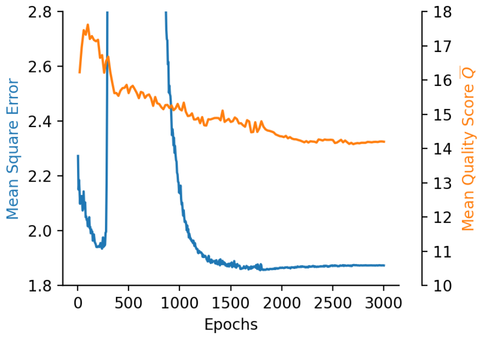

Figure 1: Plot of $\overline{Q}$ and validation MSE. The highest $\overline{Q}$ score is achieved at epoch 130.

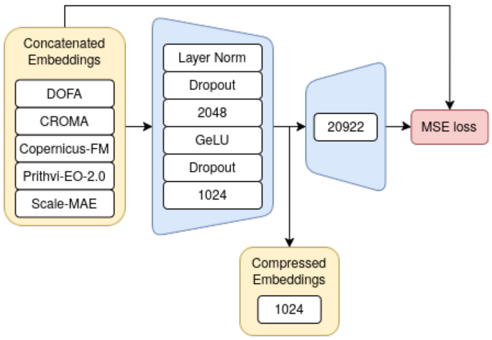

Figure 2: A flowchart of our compression method.

- **Copernicus-FM** [14] uses hypernetworks conditioned on band wavelength/bandwidth and non-spectral variable names, and injects geolocation/time metadata. Pretrained with MIM and continual distillation.
- **Prithvi 2.0** [13], a ViT-MAE architecture with 3D spatio-temporal embeddings, pretrained on 4.2M global HLS (Landsat/Sentinel-2) time-series samples, incorporating temporal and location encodings.

Where possible, we used spatially independent embedding outputs (e.g., tokens that were not directly related to one of the vision transformer input tokens), but Prithvi 2.0 was not trained to produce such embeddings. In this case, we calculate the spatial mean and standard deviation over the output tokens, and use these as our embedding. For networks that do not take multi-temporal inputs, an embedding was generated separately for each time step.

Since top-of-atmosphere (TOA) information is highly correlated with bottom-of-atmosphere (BOA) information, and no models were trained on TOA inputs, we decided to forego TOA data completely. We also do not employ any data augmentation on the input data of the foundation models, relying only on embeddings extracted from the unmodified data cubes.

Our method employs a two-layer fully connected autoencoder, with a GeLU [9] nonlinearity. Both layers utilize dropout with probability $p = 0.1$ for regularization. The stack of concatenated embeddings is normalized as one vector, by applying layer normalization [4]. We train our encoder jointly with a single-layer stand-in decoder that is intended to simulate the capability of the decoder used in the benchmark.

4 Results and Discussion

We evaluate our method using three metrics:

1. A Mean Squared Error (MSE) value based on reconstructing the embeddings from the "dev" split of the challenge dataset, which we do not use for training.
2. The mean quality score provided by the benchmark for every model, calculated as $\overline{Q} = \text{avg}_t\left(100\frac{\text{avg}_k(s_t^k)}{\text{std}_k(s_t^k)+\epsilon}\right)$ where s_t^k is the score on task t when the decoder is trained on a data fold $k \in [0, 39]$.
3. We introduce a task normalized score, $Z_m = \text{avg}_t\left(\frac{s_t^m - \text{avg}_m(s_t^m)}{\text{std}_m(s_t^m)}\right)$. This score is used to highlight the relative score differences between methods, with each task given the same importance, as models tend to produce Q values in a particular range for a given task.

With our challenge submissions, we wanted to specifically gain insight into using pre-trained foundation model encodings as the input to the compression encoder. Neural compression using foundation models is a topic that has not been well investigated in the literature, particularly when the target is not reconstruction.

Single Model	$\overline{Q}\uparrow$	$Z\uparrow$	MSE $\downarrow$
Copernicus-FM (Co)	10.82	-0.65	0.004
CROMA (CR)	11.51	-0.49	**0.001**
DOFA (D)	11.61	-0.43	0.008
Prithvi-EO-2.0 (Pr)	9.71	-0.14	0.843
Scale-MAE	9.48	-0.39	0.617
Fusion	$\overline{Q}\uparrow$	$Z\uparrow$	MSE $\downarrow$
D + CR	16.10	0.05	0.013
D + CR + Co	17.23	0.52	0.023
D + CR + Co + Pr	**18.15**	0.67	1.139
All	17.43	**0.87**	1.953

Table 1: Performance metrics for the ablation study, including FMs and their combinations. D, CR, Co and Pr stand for DOFA, CROMA, Copernicus-FM and Prithvi-2.0 respectively. We report mean quality ($\overline{Q}$), task normalized quality (Z), and mean squared error (MSE) on the validation set.

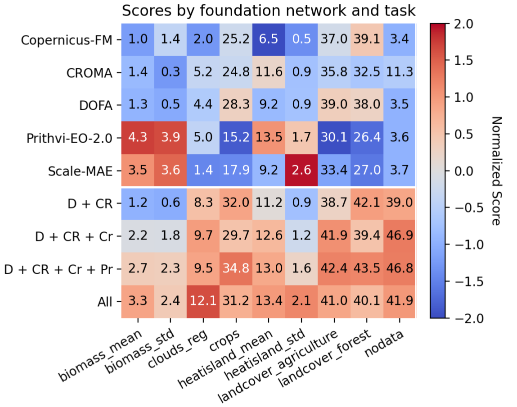

Figure 3: A per-task breakdown of the ablation study. Coloring is based on the task-normalized score Z, while the numbers show the task-wise Q quality scores from NeuCo Bench.

Due to the above, we wanted to find out how well a reconstruction metric such as MSE will correlate with the task performance, and we wanted to see whether using multiple foundation models improves the performance. The diversity in upstream data and tasks is probably an upside, but since our compression training is *not* guided by EO tasks, it might keep irrelevant or redundant details from the embeddings. When FMs are used for their classic tasks, such as image segmentation, the decoder architecture uses spatial features from multiple depth levels inside the encoder network. These together are usually larger than the input images themselves, so it might be that the (non-spatial) embeddings are not very efficient data compressions in the first place. Due to this, there was a chance that directly compressing the input images might yield similar results to compressing embeddings.

During the challenge, we went through multiple iterations of architectures. We found that trying to compress the image itself with a combination of engineered features and PCA performed worse than using engineered features on foundation model embeddings. Replacing the PCA with a convolutional autoencoder performed even better, but the best performance was achieved by the simple fully-connected two-layer autoencoder presented in Figure 2. The dropout layers also played a crucial role in avoiding overfitting: leaving them out decreases the $\overline{Q}$ score to 16.45.

In the challenge rankings, our method achieved a top $\overline{Q}$ value of 15.44, which was calculated using $K = 200$ folds and excluding the cloud and missing data prediction tasks. To make future comparisons easier, the results reported here use the default configuration settings from the since-published NeuCo-Bench repository, which sets $K = 40$ and includes the two extra tasks, increasing the Q value to 17.43 for our method.

The ablation results in Table 1 and Fig. 3 highlight the importance of including embeddings from multiple models. Compressing embeddings from the two best-performing FMs results in a 36% increase in $\overline{Q}$ score, and each additional network results in further improvements. Note that the reconstruction error on the Prithvi and Scale-MAE embeddings is quite large, indicating that the compression architecture struggles with reproduction. Despite this, the inclusion of the (otherwise relatively weakly performing) Scale-MAE embeddings worsens both $\overline{Q}$ and MSE, but the task-normalized score keeps increasing due to improvements in the weaker performing tasks. This supports our hypothesis that selecting a broad, diverse set of foundation models is more important than focusing on pure reconstruction capability, especially if the downstream tasks are unknown.

In the early stages of the challenge, we observed that prolonged training degraded our results, *even when* our validation set showed a lower reconstruction error. Despite increasing regularization, limiting the number of epochs was paramount for good performance. Interestingly, training for around 30 times longer than our best results (see Fig. 1) shows a really strong double descent phenomenon [6, 11] based on the reconstruction error, but this improvement does not translate to the benchmark results.

5 Conclusion

Our results open a possibility for using pre-trained foundational models as compression backbones for unknown downstream tasks in Earth Observation, hopefully reducing computational time and energy usage when developing and deploying such architectures.

We would like to highlight the gap in performance between the merged and singular embeddings, even if the FMs perform similarly before the merge. It seems that even when constrained by a compact representation and a limited encoder and decoder, feeding more diverse information in the form of multiple embeddings improves performance on downstream tasks. In our opinion, this indicates that there is space for improvement in the foundation models themselves. We wish to extend the current study to more and newer FMs, such as Terramind and Tessera.

Observing what might be a form of generalization beyond overfitting in the context of compression autoencoders was unexpected, but is not surprising in hindsight. If further work closes the gap between the reconstruction loss and the downstream task performance, this phenomenon could enable tuning compression methods using only a small amount of samples.

Acknowledgments and Disclosure of Funding

This research is part of the EO-AI4GlobalChange project funded by Digital Futures, Stockholm, Sweden.

References

[1] embed2scale/NeuCo-Bench - github.com. `https://github.com/embed2scale/NeuCo-Bench`. [Accessed 10-10-2025].

[2] EARTHVISION 2025 - Embed2Scale Challenge - grss-ieee.org. `https://www.grss-ieee.org/events/earthvision-2025/?tab=challenge`. [Accessed 10-10-2025].

[3] embed2scale/SSL4EO-S12-downstream · Datasets at Hugging Face — huggingface.co. `https://huggingface.co/datasets/embed2scale/SSL4EO-S12-downstream`. [Accessed 20-10-2025].

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

[5] B. Blumenstiel, N. A. A. Braham, C. M. Albrecht, S. Maurogiovanni, and P. Fraccaro. Ssl4eo-s12 v1.1: A multimodal, multiseasonal dataset for pretraining, updated, 2025. URL `https://arxiv.org/abs/2503.00168`.

[6] S. d'Ascoli, L. Sagun, and G. Biroli. Triple descent and the two kinds of overfitting: where and why do they appear?*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021 (12):124002, Dec. 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3909. URL `http://dx.doi.org/10.1088/1742-5468/ac3909`.

[7] A. Fuller, K. Millard, and J. R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders, 2023. URL `https://arxiv.org/abs/2311.00566`.

[8] C. Gomes, I. Wittmann, D. Robert, J. Jakubik, T. Reichelt, S. Maurogiovanni, R. Vinge, J. Hurst, E. Scheurer, R. Sedona, T. Brunschwiler, S. Kesselheim, M. Batič, P. Stier, J. D. Wegner, G. Cavallaro, E. Pebesma, M. Marszalek, M. A. Belenguer-Plomer, K. Adriko, P. Fraccaro, R. Kienzler, R. Briq, S. Benassou, M. Lazzarini, and C. M. Albrecht. Lossy neural compression for geospatial analytics: A review. *IEEE Geoscience and Remote Sensing Magazine*, 13 (3):97–135, Sept. 2025. ISSN 2473-2397. doi: 10.1109/mgrs.2025.3546527. URL `http://dx.doi.org/10.1109/MGRS.2025.3546527`.

[9] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

[10] V. Marsocci, Y. Jia, G. L. Bellier, D. Kerekes, L. Zeng, S. Hafner, S. Gerard, E. Brune, R. Yadav, A. Shibli, H. Fang, Y. Ban, M. Vergauwen, N. Audebert, and A. Nascetti. Pangaea: A global and inclusive benchmark for geospatial foundation models, 2025. URL https://arxiv.org/abs/2412.04204.

[11] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

[12] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023. URL https://arxiv.org/abs/2212.14532.

[13] D. Szwarcman, S. Roy, P. Fraccaro, Þorsteinn Elí Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, C. Gomes, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. A. Asanjan, B. Mujeci, D. Shidham, T. Keenan, P. Arevalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, D. Bell, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. B. Moreno. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications, 2025. URL https://arxiv.org/abs/2412.02732.

[14] Y. Wang, Z. Xiong, C. Liu, A. J. Stewart, T. Dujardin, N. I. Bountos, A. Zavras, F. Gerken, I. Papoutsis, L. Leal-Taixé, and X. X. Zhu. Towards a unified copernicus foundation model for earth vision, 2025. URL https://arxiv.org/abs/2503.11849.

[15] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu. Neural plasticity-inspired multimodal foundation model for earth observation, 2024. URL https://arxiv.org/abs/2403.15356.

[16] Z. Xu, R. Tang, M. Bianco, Q. Zhang, R. Madhok, N. Karianakis, and F. Yu. Geospatial foundational embedder: Top-1 winning solution on earthvision embed2scale challenge (cvpr 2025), 2025. URL https://arxiv.org/abs/2509.06993.

A Technical Appendices and Supplementary Material

A.1 Code

Code is available at `https://github.com/KerekesDavid/embed2scale-solution`

A.2 Asset licenses

- **SSL4EO-S12-downstream** [3]: CC-BY-4.0
- **CROMA** [7]: MIT license
- **DOFA** [15]: MIT license
- **Scale-MAE** [12]: CC Attribution-NonCommercial 4.0 International
- **Copernicus-FM** [14]: Apache License, Version 2.0.
- **Prithvi 2.0** [13]: MIT license