Fused Foundation Model Embeddings for Earth Observation Compression

A Winning Solution to the Embed2Scale Challenge

Dávid Kerekes[1], Isabelle Wittmann[2], Eric Brune[1], Ritu Yadav[1], Valerio Marsocci[3], Yuru Jia[1], Andrea Nascetti[1]

[1]KTH Royal Institute of Technology, [2]IBM Research Europe, [3]ESA Phi-lab

We merge diverse foundation model embeddings to achieve state-of-the-art performance on the NeuCo benchmark:

Team	$\overline{Q}\uparrow$
KTH and Friends (Ours)	**20.719**
AI4G Intern Squad	19.947
TeamGrelous	19.065
Sexy Scholars	17.530
Degas AI	16.441

This merging significantly improves performance on unknown downstream tasks.

The results highlight the flexibility of using an ensemble of Foundation Models as compression backbones for unknown downstream tasks.

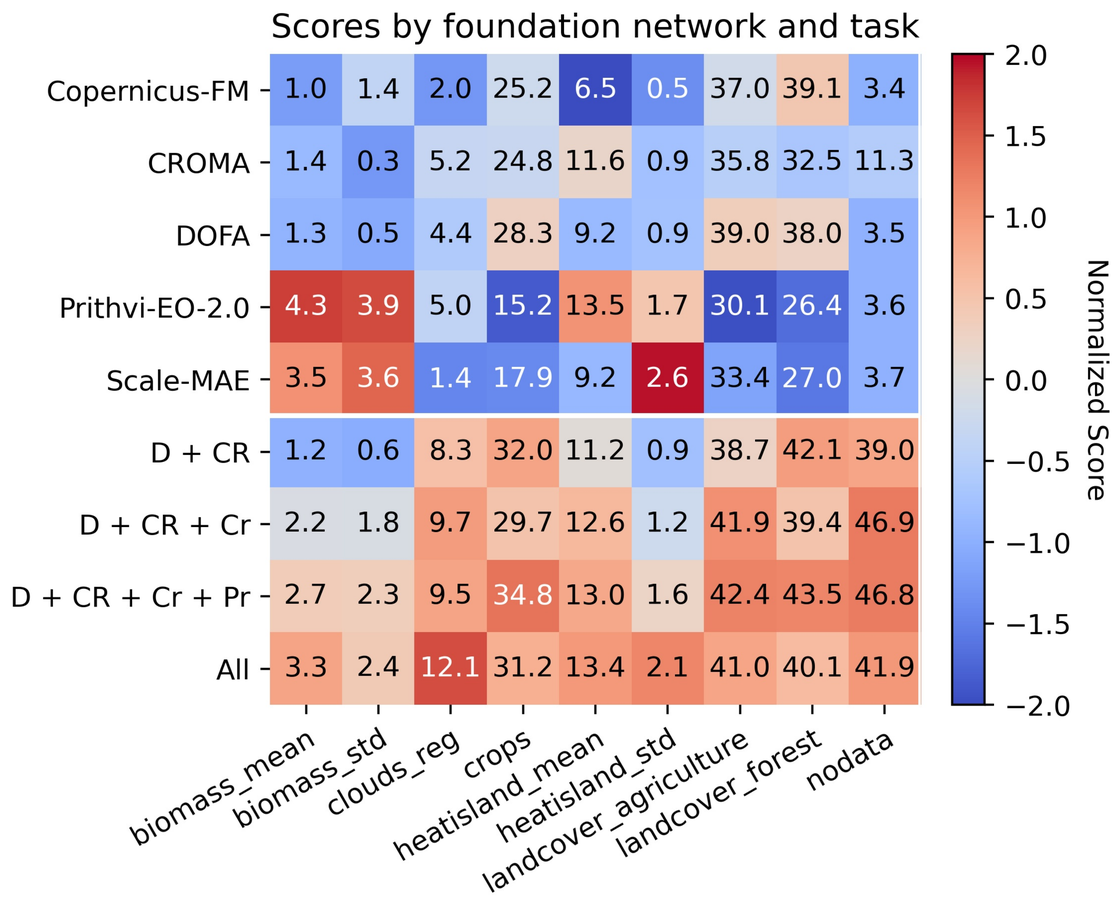

Figure 1: A per-task breakdown of the ablation study. Coloring is based on the task-normalized score Z, while the numbers show the task-wise Q quality scores from NeuCo-Bench.

Embed2Scale Challenge

The Embed2Scale challenge was an embedding-only data challenge. Participants were tasked with compressing four-season, multi-modal Sentinel-1/2 inputs into a single 1024-dimensional embedding vector.

These embeddings were evaluated using a simple linear probing, and as a consequence 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.

NeuCo Benchmark

NeuCo-Bench is an evaluation framework for general-purpose earth observation 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 tasks via linear probing, independent of the underlying encoder or backbone.

The eight main tasks focus on regression:

- Biomass
- Crop fraction of soybean and corn
- Land cover fraction for agriculture and forest
- Cloud cover fraction
- Summer surface temperature

Data

We base our evaluation on the SSL4EO-S12-downstream 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 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 27 input channels, and 264 × 264 spatial resolution.

Channels bands:

- 2 Sentinel-1 SAR GRD (VV, VH)
- 13 Sentinel-2 Level-1C (TOA)
- 12 Sentinel-2 Level-2A (surface reflectance)

Methodology

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.

We leverage the representational diversity of a curated ensemble of five foundation models:

- CROMA
- DOFA
- Scale-MAE
- Copernicus-FM
- Prithvi 2.0

Merging these diverse FM features significantly outperforms singular models (see ablation study in Figure 1 and Table 2).

We generate embeddings using the SAR and surface reflectance bands, concatenate them, and feed them to a simple two layer autoencoder architecture.

The decoder used for training is a single linear layer, to approximate the NeuCo-Bench conditions as well as possible.

Results

Our submission was ranked first based on absolute performance, and a close second behind AI4G Intern Squad in an alternative scoring scheme, where task weights were adjusted to reflect difficulty based on the competitors' results.

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 (see ablation study in Figure 1 and Table 2).

During 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 Figure 3) shows a really strong double descent phenomenon based on the reconstruction error, but this improvement does not translate to the benchmark results.

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 2: Ablation results. Mean quality, task normalized quality, and mean squared error on the validation set.

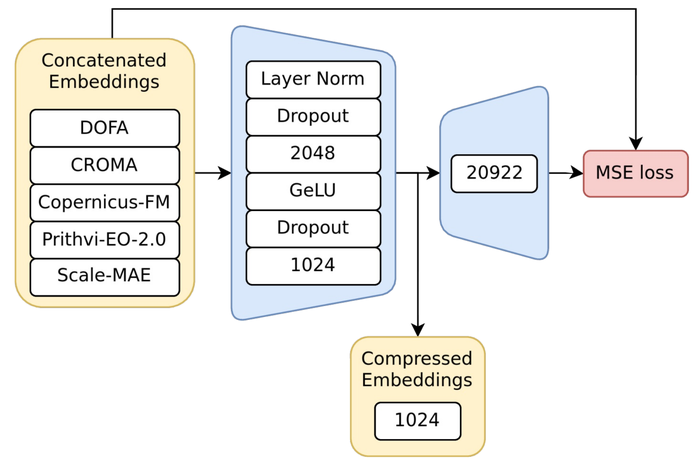

Figure 2: Encoder and training architecture

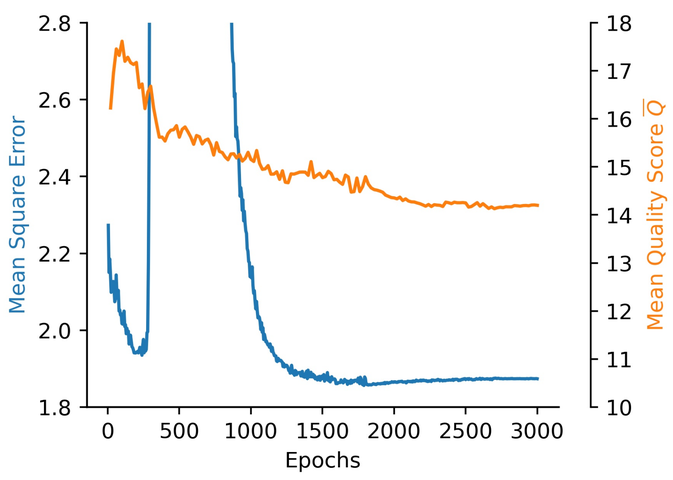

Figure 3: Plot of the Mean Quality Score and validation MSE over the training period. Highest Mean Q is at epoch 130.

Corresponding Author: Dávid Kerekes – dkerekes@kth.se

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