EMBED 2SCALE

Earth Observation & Weather Data Federation with AI Embeddings

D2.2:
Reference Code for AI-Compressors

Revision: v1.1

D2.2: Reference Code for AI-Compressors

Work package	WP 2
Task	T2.2
Due date	30/06/2025
Submission date	19/08/2025
Deliverable lead	German Aerospace Center (DLR)
Version	1.1
Authors	Conrad M Albrecht (DLR), Rikard Vinge (DLR), Isabelle Wittmann (IBM), Rocco Sedona (FZJ), Thomas Brunschwiler (IBM)
Reviewers	Jonas Hurst (WWU), Miguel Belenguer-Plomer (SatCen)
Abstract	We provide codes for state-of-the-art reference implementations of lossy neural compression of Sentinel-1 & -2 data based on our evaluations of the CVPR EARTHVISION data challenge.
Keywords	lossy neural compression, CVPR EARTHVISION data challenge, TerraMind, MOSAIKS, NeuCo-Bench, video compression, SSL4EO-S12

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	07/07/2025	Initial draft	Rikard Vinge (DLR), Isabelle Wittmann (IBM), Conrad M Albrecht (DLR), Thomas Brunschwiler (IBM), Rocco Sedona (FZJ)
v1.1	31/07/2025	Incorporate feedback of reviewer Nikos Fragkoulis	Conrad M Albrecht (DLR)

Funded by the European Union

Grant Agreement No: 101131841 | **Topic:** HORIZON-EUSPA-2022-SPACE-02-55

Call: HORIZON-EUSPA-2022-SPACE | **Type of action:** HORIZON-RIA

DISCLAIMER

Embed2Scale (Earth Observation & Weather Data Federation With AI Embeddings) project is funded by the EU's Horizon Europe program under Grant Agreement number 101131841. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI).

COPYRIGHT NOTICE

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	OTHER	
Dissemination Level	PU	
PU	Public	☐
SEN	Sensitive, limited under the conditions of the Grant Agreement	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report elaborates on the current status of E2S's efforts to advance the state-of-the-art in neural compression (NC) for Earth Observation (EO) data. The work is closely integrated with deliverable D2.3 providing the benchmarking framework NeuCo-Bench (NCB) as a spin-off from the 2025 IEEE Computer Vision and Pattern Recognition Conference's EARTHVISION workshop (CVPR-EVW). The CVPR-EVW challenged participants to compress multi-modal (radar and multi-spectral), multi-temporal (4 seasons) into 1024-dimensional embeddings equivalent to heavy lossy compression of a factor of about 7000. In particular, we evaluated and tested the source codes of participating teams to identify and jointly open source their solution. The corresponding GitHub repositories allow the community to further develop the winning image-level NC solutions for classification and regression tasks as provided by deliverable D4.5.

The methods developed to date compress the spatio-temporal-spectral SSL4EO-S12 data cube of Sentinel-1 and Sentinel-2 data to embeddings by factors of thousands. As identified by the CVPR-EVW E2S data challenge, it proves beneficial to utilize state-of-the-art geospatial Foundation Models (EO FMs) and blend them in a scheme including dimensional reduction techniques. Surprisingly, even the no-training, fixed feature extractor MOSAIKS provides promising results as a substitute for EO FMs which may serve as an energy-efficient solution when compared to compute-intensive pre-training of such models.

As a perspective based on our current research, we discuss the potential of video compression for time series of Earth observation and Weather Data (WD). Moreover, we provide performance evaluations on the any-to-any FM *TerraMind*[1]. As we are further exploring methodologies for efficient embedding generation with the *Earth2Vec* community (cf. deliverable D2.3), we plan to assemble a corresponding conference paper including our performance findings for future directions of the field.

[1] *ESA FAST-EO* project involving E2S partners DLR, FZJ, and IBM

ABBREVIATIONS

AI	Artificial Intelligence
EO	Earth Observation
NC	Neural Compression
WD	Weather Data
CV	Computer Vision
FM	Foundation Model
NLP	Natural Language Processing
CV	Computer Vision
CNN	Convolutional Neural Network
ViT	Vision Transformer
VLM	Vision Language Model
VCT	Video Compression Transformer
VQ-VAE	Vector-Quantised Variational Auto-Encoders
GAN	Generative Adversarial Network
GSQ	Grouped Spherical Quantisation
SSIM	Structural Similarity Index Measure
RGB	Red-Green-Blue (channels of an image)
CVPR-EVW	Computer Vision and Pattern Recognition Conference's EARTHVISION workshop, https://www.grss-ieee.org/events/earthvision-2025
NCB	NeuCo-Bench, https://github.com/embed2scale/NeuCo-Bench
SSL4EO	Self-Supervised Learning for EO, dataset: https://huggingface.co/datasets/embed2scale/SSL4EO-S12-v1.1
PCA, SVD	Principal Component Analysis, Singular Value Decomposition
MSE	Mean Square Error

INTRODUCTION: LOSSY NEURAL COMPRESSION FOR EARTH OBSERVATION

Earth Observation (EO) data repositories rank among the largest global data stores, with satellites generating terabytes of imagery every day [CopernDash] and spanning a broad spectrum of EO data types with varying spectral bands, spatial resolutions, and temporal frequencies which presents a unique challenge for data processing. Handling these large volumes of high-dimensional EO data requires substantial resources for storage, transmission, and computation. Compressed EO embeddings provide a promising approach to address these issues. By transforming raw imagery into compact, rich representations, embeddings can significantly lower storage and transmission costs while retaining the information needed for downstream tasks. This is underpinned by advances in neural methods from the fields of data compression and representation learning.

Today's EO workflows typically rely on hand-crafted codecs (e.g., [JPEG2000]) originally designed for generic imagery, cf. lossy NC review deliverable D2.1. In contrast, fully data-driven neural codecs optimize end-to-end rate–distortion objectives and can be tailored to the spectral, spatial and temporal structures present in EO data [E2SReview]. These learned methods produce latent representations that outperform classical codecs in reconstruction quality (PSNR, SSIM) and bitrate efficiency (bpp). Typically, NC utilizes a learnt encoder to map data into a compact latent space, and then applies entropy coding to generate a variable-length binary bitstream. Research generally prioritizes minimization of the bitrate of the final variable-length binary code and pixel-level fidelity over task-relevant information. The latter is particularly important in EO, where the value of data primarily lies in its effectiveness for downstream analytical models. Recent advances in EO FM research include any-to-any multimodal models such as *TerraMind*. It generates semantically rich embeddings capable of zero-shot inference across a variety of tasks. However, their latent tensors still incur significant storage and transfer burdens that do not fit E2S's scope of strong lossy compression. TerraMind's compression factor ranges from 0.75 to 2.

To advance NC of EO data into semantically rich embeddings, we evaluate and integrate methods from neural image codecs, representation learning, and the feature-compression literature. We target metrics aligned with the objectives of E2S:

- For **Embed2Transfer** – assess compression ratio and embedding size by evaluating the fixed-length latent representation (usable for retrieval and direct inference) alongside the additional gains from entropy-coding into a binary bitstream (average bitrate, bpp).
- For **Embed2Train & Embed2Infer** – measure downstream task performance and reductions in training and inference time when models are trained directly on compact embeddings rather than raw data.
- For **Embed2Find** – evaluate retrieval performance via mean average precision (mAP) and search latency of the generated embeddings.

By integrating these perspectives, E2S aims to comprehensively evaluate the efficiency and utility of compressed EO embeddings.

As an initial step along the agenda outlined above, we organized the E2S data challenge at the 2025 CVPR-EVW and subsequently released the NeuCo-Bench evaluation framework [NEUCO]. NCB benchmarks and compares compact EO embeddings across a common set of EO downstream tasks (cf. deliverable D4.5). This has sparked an active research community focused on compressed EO embeddings. Below, we review the current state-of-the-art approaches alongside our research efforts.

METHODS: REFERENCE IMPLEMENTATIONS

We introduce state-of-the-art NC for EO as

- developed within the E2S consortium
- developed in collaboration by third parties as part of the 2025 CVPR-EVW [NEUCO]
- evaluated and benchmarked by the NCB framework [NEUCO]

in order to gauge their performance.

IMAGE TOKENIZATION AS COMPRESSORS

- **Reference Code** (Apache 2.0 license)**:**
  https://github.com/erikscheurer/GSQ4EO

Petabyte-scale EO archives demand leaner storage and faster I/O. Inspired by artificial intelligence

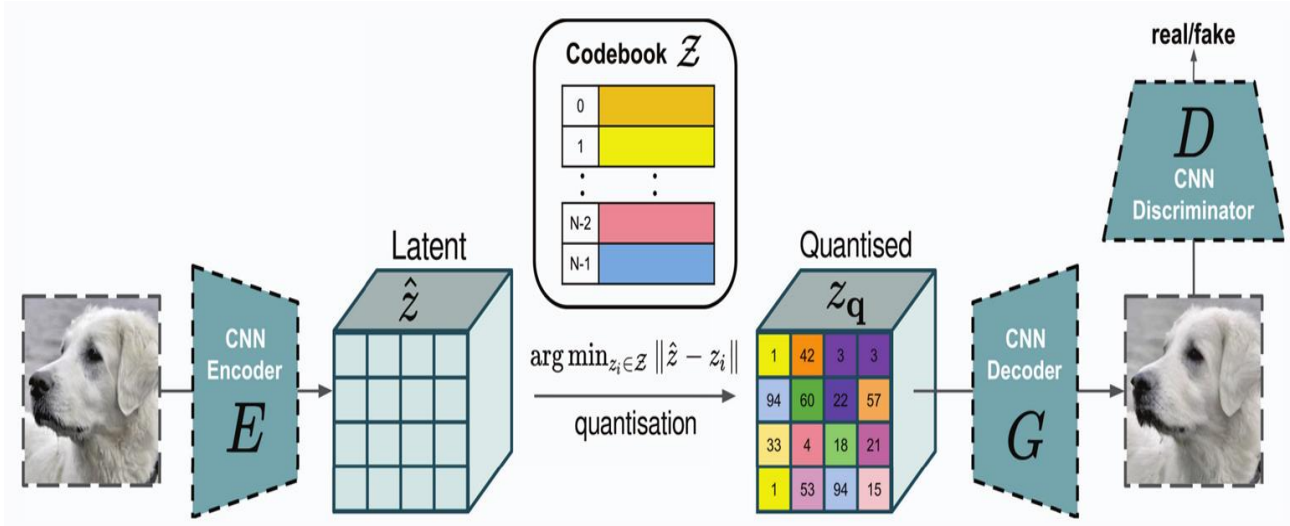

Figure 1: Cartoon of a VQ-VAE as depicted in [TrafoIS].

for natural language processing (NLP), embeddings can be interpreted as feature vectors in a metric space where their (Euclidean) distance resembles aspects of semantic relations. The concept of a transformer has been transferred from NLP to computer vision (CV) by means of vision transformers (ViT). Lossy NC generates embeddings of quality that need to be semantically enriched to serve a wide range of downstream tasks.

In an initial step towards a multi-modal FM, each EO modality requires a “tokenizer”. As depicted by Fig. 1, vector-quantised (VQ) variational auto-encoders (VQ-VAE) [VQVAE] store only code-book indices, trimming image size by a compression factor of more than 450 while satisfactorily reconstructing scenes. As alluded in deliverable D2.1 (Fig. 11), the statistics of EO data requires adaptation of standard NC methodologies. Our work *Scalable Efficient Compression in Large-Scale Earth Observation* [ScaleNC4EO] explores and addresses this aspect for an EO-aware VQ-VAE.

Decomposed VQ splits the latent space, easing optimization and squeezing extra fidelity from larger embeddings. As Fig. 2 illustrates, Grouped Spherical Quantisation (GSQ) further clusters codes on a unit sphere, tightening spectral error across the multispectral bands. When reconstructing RGB Sentinel-2 data, panels show near-identical imagery; only micro-textures fade after the down-sampling step. This observation confirms visually loss-tolerant compression. And indeed, the pixel statistics of Sentinel-2 images and their reconstructed counterparts well match. Moreover, example

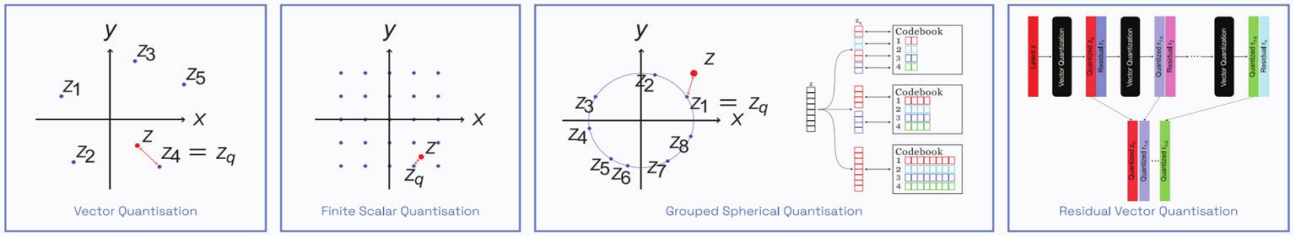

Figure 2: Visualization of Quantization Techniques.

imagery underline that our methodology keeps tile boundaries crisp indicating that spatial discontinuities survive the latent bottleneck.

Mixing unrelated embeddings we observe a global colour bleed: pixels previously depicting a desert turn green as the network extrapolates from tree-covered parts of the image. This information bleeding is caused by the network collecting global information in specific parts of the 2D representation. Image reconstruction quality plots (cf. Fig. 3) such as the Structural Similarity Index Measure (SSIM) vs. the compression factor (here purely defined by the number of data points in the input compared to the embedding) depict the trade-off between compression and reconstruction quality. Compared to RGB data, additional correlations in multichannel data can be exploited to preserve image quality while reducing the space required to store or transfer Sentinel 2 images. A discriminator in VQ-VAE (cf. Fig. 1) bumps up semantic meaning in the image but offers little gain in quantitative reconstruction quality of EO spectra. Given additional computational demand, we recommend a discriminator optional, but not necessary.

In a nutshell our VQ-VAE-based Sentinel-2 tokenizer slashes storage and I/O costs without derailing pixel-wise accuracy. Thus, it paves the way for a GPU-friendly, planet-scale tokenizer that serves even lossless compression scenarios.

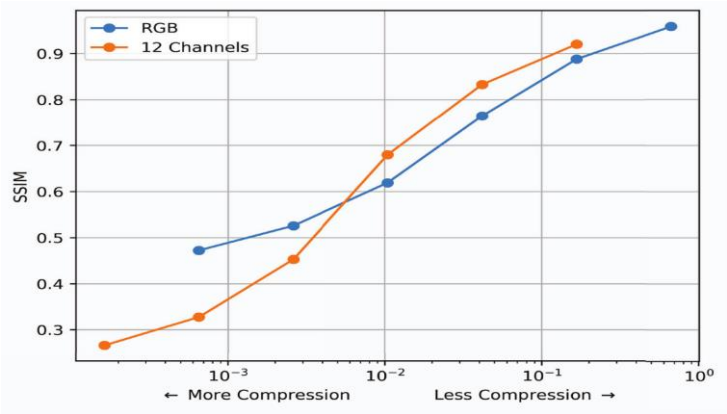

Figure 3: Rate-Distortion plot for VQ-VAE tuned to Sentinel-2 L2A imagery.

TerraMind: A Lossy Neural Compression Baseline

- **Reference Code** (Apache-2.0 license)[2]**:**
  https://github.com/IBM/terramind

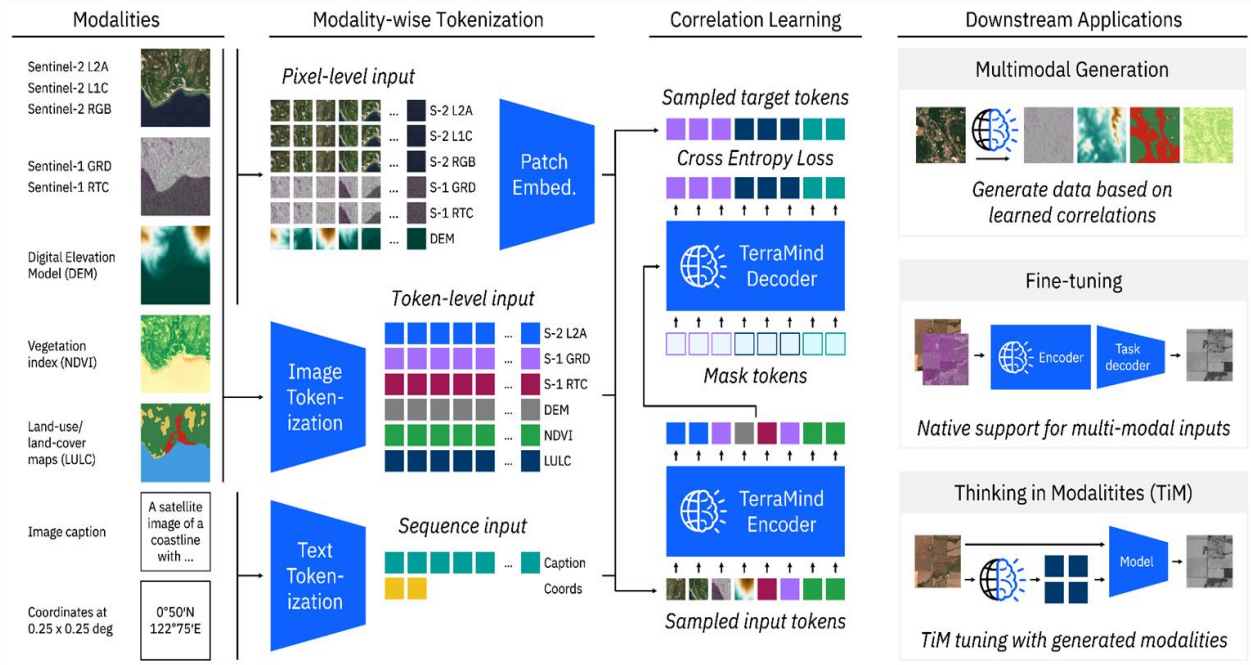

Figure 4: Overview of high-level workflow in TerraMind any-to-any FM as depicted in [TERRAMIND].

TerraMind is based on the 4M [4M] model and is trained in two steps as illustrated by Fig. 4:

1. A tokenizer is separately trained per modality to compress and reconstruct the respective modality.
2. The modality-specific tokens are passed through a transformer-based encoder-decoder network that fuses the modality-specific tokens into a single, combined, set of tokens, before reconstructing the original sets in the decoder.

The single set of combined tokens is what is of primary interest to Embed2Transfer, Embed2Train & Embed2Infer, and Embed2Find. As detailed in Fig. 4 (left to right), TerraMind operates as follows: Each modality is individually tokenized by separate *Tokenization* where each is separately trained to compress and reconstruct its individual modality. The modality-specific tokens are provided to a *TerraMind Encoder* bottleneck for *Correlation Learning*. The latter step exploits multi-modal correlations for fusing latent tokens (embeddings). It will become relevant in our future research related to task T2.4 and deliverable D2.4. To reconstruct all the modalities, a diffusion-based *TerraMind Decoder* reconstructs the modality-specific tokens. During training, those are compared with the modality-specific tokens generated by the *Tokenization* process.

TerraMind is natively not multi-temporal. To include multiple timesteps for each location of the SSL4EO-S12 dataset in the NCB framework, two methods of merging the temporal dimension are utilized:

[2] not part of E2S developments, but published within ESA's project FAST-EO: https://www.fast-eo.eu

1. the temporal dimension of the images for each location get averaged prior to the modality-specific tokenization
2. each timestep gets tokenized individually for subsequently averaging the model-specific tokens over the timesteps

To generate compact embeddings that can be evaluated alongside compression methods in NeuCo-Bench, we apply patch-averaging which yields small 768-dim embeddings per multi-modal input image.

The corresponding Python code for the application of patch-averaging to TerraMind embedding representations reads:

```
from terratorch.registry import BACKBONE_REGISTRY

model = BACKBONE_REGISTRY.build(
  'terramind_v1_base',
  pretrained=True,
  modalities=['S2L1C', 'S2L2A', 'S1GRD']
)

model.eval()
output = model(input_image)
output_embedding = output[-1].mean(dim=1)
```

TMOSAIKS: EMBEDDINGS FROM FIXED FEATURES WITHOUT TRAINING

- **Reference Code** (MIT license)[3]**:**
https://github.com/isaaccorley/temporal-mosaiks

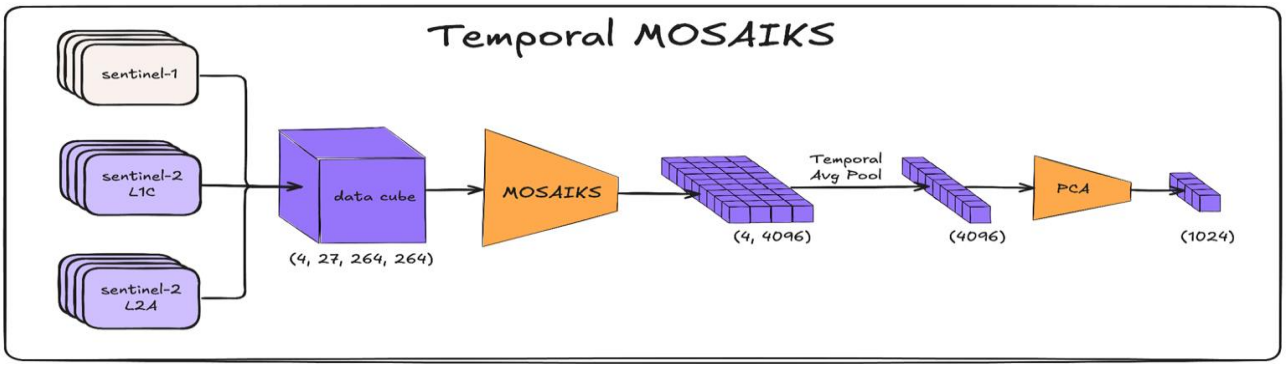

Figure 5: The different EO modalities are provided as is to the MOSAIKS feature creator that generates an embedding per season. Subsequently, the seasons are averaged and PCA is used to compress the MOSAIKS output further.

While AI for EO research tends to focus on the training of large neural networks, there are recently developed methods for training-free models. Such are particularly attractive w.r.t. energy-efficient solutions for NC. One such methodology is the feature extraction method MOSAIKS [MOSAIKS]. Temporal MOSAIKS, implemented in [AI4G] for the 2025 CVPR-EVW extends the MOSAIKS to handle the temporal aspect as present in the seasonal SSL4EO-S12 data (cf. D4.5). In the following, we denote the Temporal MOSAIKS methods simply as TMOSAIKS.

MOSAIKS avoids training by generating random features in a CNN-style approach. It creates K features from a set of N satellite images I_l, l=1,...,N, of dimensions (W, H, S), where W is the width of the image in pixels, H the height in pixels and S the number of spectral bands. The algorithm to create the features summarizes as follows:

1. Randomly generate K kernels of size (w, h, S), where w and h are the width and height of the kernels, respectively.
2. For each image I_l, convolve each of the K kernels with the image. This produces a new set of transformed images I'_l.
3. For each transformed image, apply a ReLU-activation function to create K "activation maps".
4. For each activation map, average all pixels into a single number, resulting in K features for each image.

TMOSAIKS extends MOSAIKS by allowing for temporal analysis of multiple timestamps at the same location. Instead of a single image I_l per location, we now assume T_l images $I_{l,t}$, l=1,…,N, t=1,...,T_l per location l. Starting from step 4 in MOSAIKS, TMOSAIKS extends the feature selection by:

5. Create the K features of each image at each location and timestamp repeating steps 2 through 4 which results in T_l sets of K features for each location.
6. Aggregate the T_l sets of K features into a single set of K features by averaging along the temporal dimension. This results in K features for each of the N locations.

[3] developed under the umbrella of the E2S CVPR-EVW data challenge (cf. D2.3) by a team from University of Bundeswehr and Microsoft, verification and evaluation by E2S team through NCB

7. If K is greater than the required number of features, perform a Principal Component Analysis (PCA) of the K features on the N locations and pick the number of principal components in order of largest to smallest eigenvalue as required to achieve a given compression ratio.

For illustration, the following code generates 4096 features utilizing TMOSAIKS to subsequently perform PCA reducing down to an embedding size of 1024 features (as per the E2S data challenge):

```
import torch
from einops import rearrange
from torchgeo.models import RCF
from sklearn.decomposition import PCA

# Each sample in a batch is a time-series of images
# batch = torch.randn(8, 4, 27, 64, 64)  # (b, t, c, h, w)

dataset = ...  # a torch dataset that returns dict(image=...)
Dataloader = torch.utils.data.DataLoader(
      dataset, batch_size=2, shuffle=False
)

# Initialize the model
model = RCF(in_channels=27, features=4096, kernel_size=3, mode="gaussian")
model.eval()

# Loop over your images and collect all embeddings
embeddings = []
for batch in dataloader:
    b, t = batch.shape[0], batch.shape[1]
    with torch.inference_mode():
        # Run MOSAIKS across all imagery independently
        x = rearrange(batch, "b t c h w -> (b t) c h w")
        emb = model(x)

        # Average pool over the time dimension
        emb = rearrange(emb, "(b t) c -> b t c", b=b, t=t)
        emb = emb.mean(dim=1)

        embeddings.append(emb)

# Perform PCA on the embeddings
embeddings = torch.cat(embeddings, dim=0).numpy()
embeddings = PCA(n_components=1024).fit_transform(embeddings)
```

MULTIFMF: FUSION OF FOUNDATION MODEL EMBEDDINGS

- **Reference Code** (Apache-2.0 license)[4]**:**
  https://github.com/KerekesDavid/embed2scale-solution

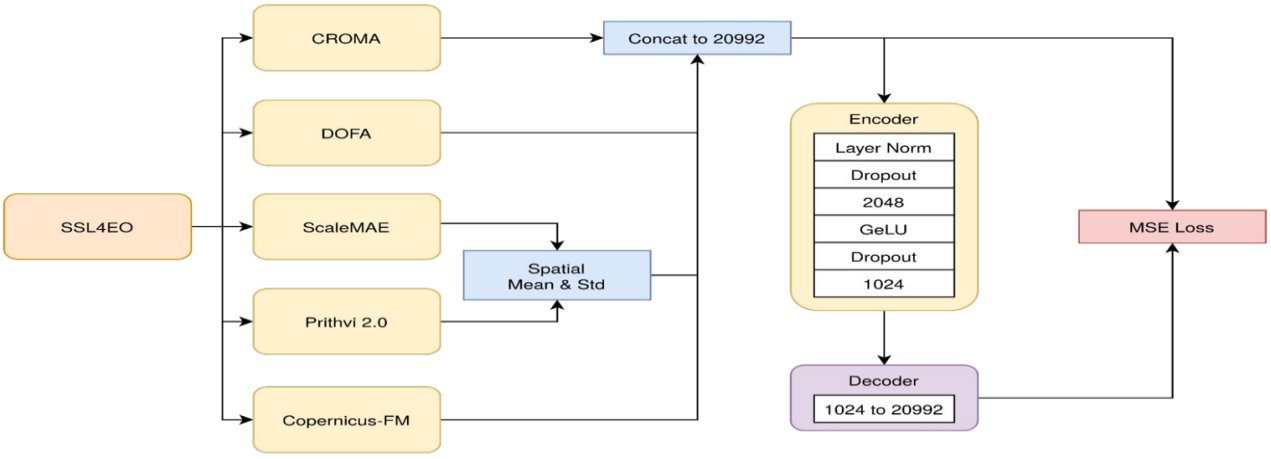

Figure 6: Workflow of the method: "Fusion of FM Embeddings" depicting five EO FMs who consume SSL4EO(-S12) data cubes to output embeddings which get merged by a two-layer MLP: A lightweight Encoder is jointly trained with a Decoder to regenerate the Concat(enated) output of the five FMs.

With the number of publicly available EO FMs constantly increasing, the combination of embeddings from multiple EO FMs becomes an attractive alternative compared to energy-expensive training from scratch. However, the way to combine the embeddings generated by various EO FMs is underexplored as of today. One option is to simply concatenate the embeddings of all the FMs into a single, long vector. To achieve competitive compression rates, a small encoder-decoder network can reduce the dimensionality of the embedding space to the dimension implied by a given compression ratio. Such a method has been implemented and open-sourced by [KTH] for the E2S data challenge in the 2025 CVPR-EVW. In what follows, we denote this approach MultiFMF as in *Multi-Foundation Model Fusion*.

As depicted in Fig. 6, the Encoder consists of two fully connected layers of dimensions 2048 and 1024, with dropout and Gaussian Error Linear Unit (GELU) activation in-between. Such are preceded by layer normalization and another dropout layer. The encoder reduces the dimensionality from an initial 20992 number of features down to the required 1024 floating point numbers. The decoder is an even more lightweight, single fully connected layer increasing the dimensionality back from 1024 to the original 20992 dimensions, cf. the "Decoder" block.

For the sake of the E2S data challenge, five EO FMs were utilized: CROMA [CROMA], DOFA [DOFA], Scale-MAE [SCALEMAE], Prithvi-2.0 [PRITHVI2] and Copernicus-FM [COPERNICUSFM]. The output from the five FMs are concatenated into a 20992-dimensional vector and fed into the encoder. If needed spatial or temporal means and standard deviations of embeddings are computed if the FM model outputs multiple given the SSL4EO-S12 input data.

[4] developed under the umbrella of the E2S CVPR-EVW data challenge (cf. D2.3) by winning team from KTH University, verification and evaluation by E2S team through NCB

During training, the encoder reduces the dimensionality of the input down to 1024 dimensions, before the decoder aims to recreate the original 20992 vector. The recreated vector is compared to the original by virtue of a Mean Squared Error (MSE) loss per FM. The five MSE losses are weighted based on initial experiments given a subset of downstream tasks available. This strategy reduces the impact of FMs with lower performance. The output from the encoder acts as the embeddings used to compare performance in the NCB framework (cf. D2.3).

MULTIFMS: EMBEDDINGS SPACE ENGINEERING UTILIZING VISION-LANGUAGE MODELS

- **Reference Code** (MIT license)[5]**:** https://github.com/xboboji/torchgeo/tree/zx/2025_earthvision_winning_solution/experiments/ssl4eo/

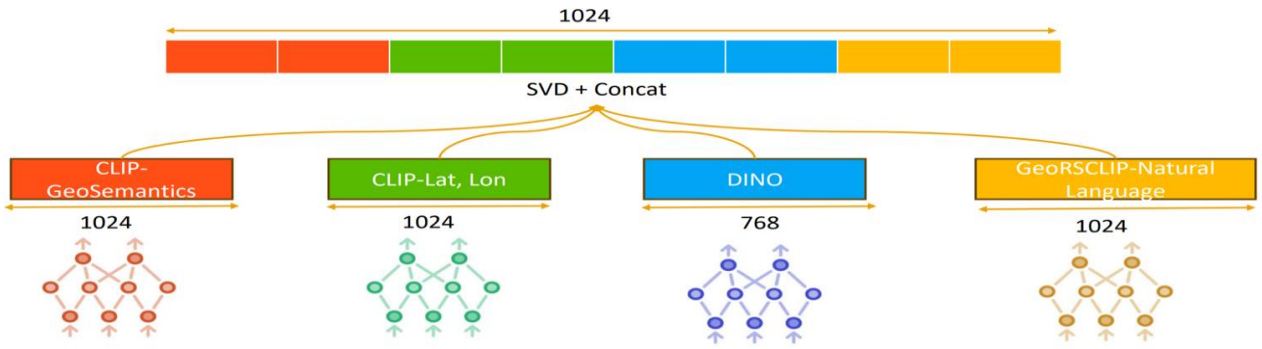

Figure 7: Architecture of the embedding space engineering utilizing VLMs models. Four FMs provide the base for the embedding space, each assuming one part of the embedding vector. SVD is employed to reduce the size of the output of the GeoRSClip-Natural Language model.

The previous methodology fused the embeddings of multiple FMs through the use of a small neural network. To achieve the same goal, i.e., to combine the strengths of multiple FMs, there are alternative routes. One such approach is to simply let each FM determine a fraction of the final (N=1024-dimensional) embedding space. This may get implemented by reducing the dimensionality of the M=4 FMs embeddings to N/M=256 features before concatenating them into the final embedding vector of total size N=1024.

The E2S data challenge winning solution presented at the 2025 CVPR-EVW we refer to as *MultiFMS* as in *Multi-Foundation Model Stacking*. MultiFMS combines four EO FMs, two of which are based on ViT-Huge [VIT] and ConvNeXt-XXL [CONVNEXT]. Those are trained using Contrastive Language-Image Pre-training (CLIP) [CLIP] by incorporating image and semantic information. Semantic information is provided in the form of text strings pointing to geo-location and details on the fraction of the image covered by forest, elevation information, the presence of night-lights, and population density figures. Semantic information is provided during pre-training, only. The third foundation model leverages ViT-Base trained with DINO [DINO] – a contrastive learning technique. The fourth FM utilizes GeoRSCLIP [GEORSCLIP], a pre-trained Vision-Language Model (VLM) fine-tuned on the RS5M dataset – a set of EO image and image caption pairs. Fig. 7 summarizes the rationale of the MultiFMS model architecture.

The CLIP- and DINO-based EO FMs of MultiFMS have been fine-tuned on SSL4EO-S12 v1.1 [SSL4EO]. In contrast, the GeoRSCLIP is kept frozen in its pre-trained state. To generate embeddings, GeoRSCLIP is fed the RGB channels of the Sentinel-2 L2A product as available in the

[5] developed under the umbrella of the E2S CVPR-EVW data challenge (cf. D2.3) by winning team from Microsoft, verification and evaluation by E2S team through NCB - the solution is currently merged into https://github.com/microsoft/torchgeo as per pull request https://github.com/microsoft/torchgeo/pull/2868

evaluation data of NCB given through SSL4EO-S12-downstream (without labels). The four seasons of SSL4EO-S12 generate four embeddings. The CLIP- and DINO-based EO FMs simultaneously receive all four seasons to output a total of three embeddings.

To engineer the final, 1024-dimensional output embedding from the four models the following dimensions need to get stacked:

- the two CLIP ViT-Huge models have an output dimension of 256
- CLIP ConvNeXt-XXL and DINO ViT-Base have 128 output dimensions
- GeoRSCLIP outputs 768 dimensions per season

The latter demands additional processing for dimensional reduction. In order to compress the four 768-dimensional embeddings, a Singular Value Decomposition (SVD) gets applied which results in 128 dimensions per season.

The final 1024-dimensional embedding is composed of:

1. 128 values from CLIP ConvNext-XXL
2. 256 values from CLIP ViT-Huge
3. 128 from DINO ViT-Base
4. 4 times 128 from GeoRSCLIP

PERFORMANCE EVALUATION OF AI-COMPRESSORS WITH NEUCO-BENCH

This section is dedicated to the performance of four neural compression methodologies and two non-machine learning baselines utilizing the NCB framework developed in D2.3. The benchmarking framework expects a fixed-size embedding space that was set to N=1024 in our E2S data challenge at the 2025 CVPR-EWV. As it stands, NCB does not evaluate methods which output embeddings of variable size, e.g., methods including an Arithmetic Encoding bottleneck. The methods we evaluate are the ones introduced above: *TerraMind, TMOSAIKS, MultiFMF,* and *MultiFMS*.

The methodologies are evaluated on five downstream tasks aligned with SSL4EO-S12-downstream as delivered in D4.5. Specifically, the regression tasks estimate …

- … the mean per-pixel biomass of the image (`biomass` use case).
- … the standard deviation of the per-pixel biomass (`biomass` use case).
- … the fraction of the image covered in clouds (`cloud` use case).
- … the fraction of the image covered in corn or soybean farms (`crops` use case).
- … the fraction of the image covered by agriculture (`landcover` use case).
- … the fraction of the image covered by forest (`landcover` use case).
- … the mean per-pixel surface temperature during summer months (`heatisland` task).
- … the standard deviation of the per-pixel surface temperature during summer months (`heatisland` task).
- … the fraction of missing values in the satellite imagery (`nodata` task).

A sixth task relating to `ships` is under development. In addition to the six tasks related to the Embed2Scale use cases, we include three additional tasks related to urban surface temperature (`heatisland`) and data quality (`nodata`), with the goal to estimate the fraction of missing data in Sentinel-2 satellite imagery. Missing data may originate from tile boundaries, data transmission errors, etc.

The two baselines we kept represent naive solutions to those tasks:

1. The *Random Baseline* creates random embeddings from a normal distribution – it serves a case where methodologies completely lack to distill any information from the EO data.
2. The *Mean Baseline* represents a simple, no-training method relying on image resolution reduction:
    a. First, we reduce the spatial resolution of each of the 27 channels from 264x264 pixels to 8x8 by bi-linear interpolation.
    b. Next, we exploit correlation by reducing the number of channels from 27 down to four. Specifically, we average the channels B1 through B9 of both S2L1C and S2L2A and we do similar for channels B11 and B12, respectively. Channel B10 of S2L1C is kept separate since no corresponding band exists in S2L2A.
    c. The four seasons have so far been treated separately, reducing each of them to data cubes of size 8x8x4.
    d. Finally, we flatten the data cubes into an N = 1024 = 8·8·4·4-elements array.

The Mean Baseline will serve as reference to estimate the magnitude of how much task-relevant information survives such an aggressive spatial and spectral aggregation with compression factor of about 7000 as set by the E2S data challenge at the 2025 CVPR-EVW.

The performance results of the methodologies listed above presents the following table, where rows with gray background correspond to Embed2Scale use cases relevant to WP4:

Task	MultiFMS	MultiFMF	TMOSAIKS	TerraMind	Mean Baseline	Random Baseline
biomass (mean)	0.33	0.28	0.42	0.42	-0.88	-5.06
biomass (std)	0.20	0.19	0.33	0.34	-0.55	-4.0
clouds	0.49	0.50	0.48	0.71	-0.69	-8.36
crops	0.73	0.77	0.79	0.86	0.14	-3.16
landcover (agriculture)	0.84	0.83	0.80	0.91	0.04	-1.16
landcover (forest)	0.82	0.82	0.81	0.90	0.19	-0.89
heatisland (mean)	0.57	0.61	0.49	0.70	-0.08	-12.1
heatisland (std)	0.14	0.13	0.11	0.24	-0.46	-3.1
nodata	0.85	0.87	0.97	0.96	-0.15	-0.31

The performance of the six methods on the nine downstream regression tasks is measured in terms of R-squared, averaged over 50 training re-runs of the linear probing procedure. Higher values are better with 1.0 the perfect score and 0.0 indicating that the model performs equally well as a model that always predicts the mean value of all the labels – negative R-squared is even worse than this simple prediction.

The two baselines perform according to expectations:

1. The Random Baseline is worse than always predicting the mean across all tasks, while
2. The Mean Baseline hovers around zero, indicating a performance similar to always predicting the mean of the target labels.

The performance of the neural compression methods are far exceeding the baselines.

On the two agriculture-related tasks, `crops` and `landcover agriculture`, our novel NC methods yield R-squared values that stay above 0.7, sometimes passing 0.8 or even ranging beyond 0.9 for the TerraMind model on the `landcover` downstream task. Such figures indicate that a high-level of semantic information is retained in the embeddings. Direct utilization of these embeddings for real-world applications is an option. Our methodologies perform exceedingly well on the `nodata` task. Correspondingly, we have options to test an image's integrity from highly compressed embeddings prior to downloading the entire scene.

The `clouds` and `biomass` tasks result in lower performance, although still significantly above the zero-threshold. In the case of the `biomass` task, we observed significant uncertainty in the target labels caused by the sparsity of the GEDI data ground truth data (cf. D4.5). Similarly, cloud segmentation is also known to be difficult due to the ambiguity to define cloud boundaries.

Funded by the European Union

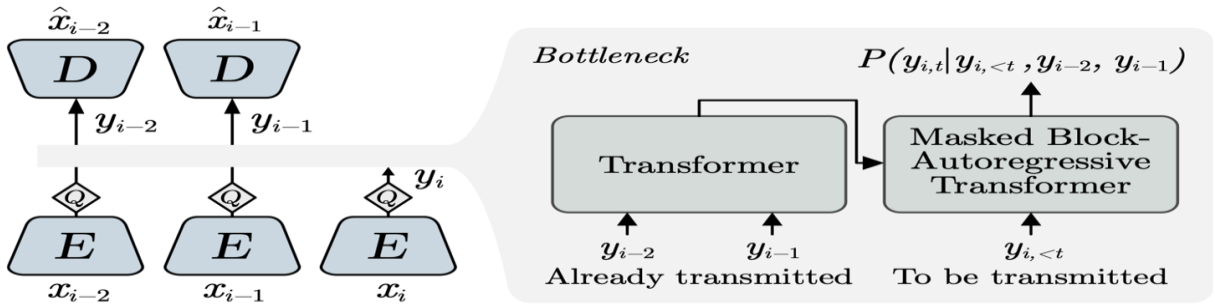

Figure 8: The Video Compression Transformer (VCT) architecture implements lossy compression on individual input images down to a quantized latent representation. Temporal relations are modelled through a latent transformer model to establish Entropy Coding. The illustration is borrowed from VCT.

PERSPECTIVE: TIME-SERIES COMPRESSION FOR EO FROM VIDEO TRANSFORMERS

INTRODUCTION

Alongside image-based methods, we currently explore video compression approaches for EO data. It is motivated by the extensive temporal structure and inter-frame correlations inherent in satellite imagery time series. Recent works started to experiment with *reference-based coding*, i.e., exploitation of historical images of the same geolocation to compress only temporal variations instead of full frames [Earth+] similar to event cameras [EventCAM]. This is motivated by the rather static nature of EO scenes such as in Sentinel-2 imagery. Yet, video compression remains an active field of research with significant potential for EO data. Modern video codecs typically rely on explicit motion estimation and compensation ([VCScale-Space], [DVC])—allocating substantial capacity to optical-flow and warping modules that are less relevant for largely static or slowly varying EO scenes. We therefore consider motion-free temporal modeling a more promising direction in EO data compression with temporal components.

Accordingly, we adopt and adapt the Video Compression Transformer (VCT) [VCT] architecture for EO data as summarized in Fig. 8. VCT first applies a learned image codec to compress each frame independently, then uses a Transformer over the frame latent representation to model temporal dependencies directly in latent space. VCT predicts the distribution of each new frame's latent embedding conditioned on its predecessors to capture long-range correlations.

Funded by the European Union

PERFORMANCE OF VCT FOR EO

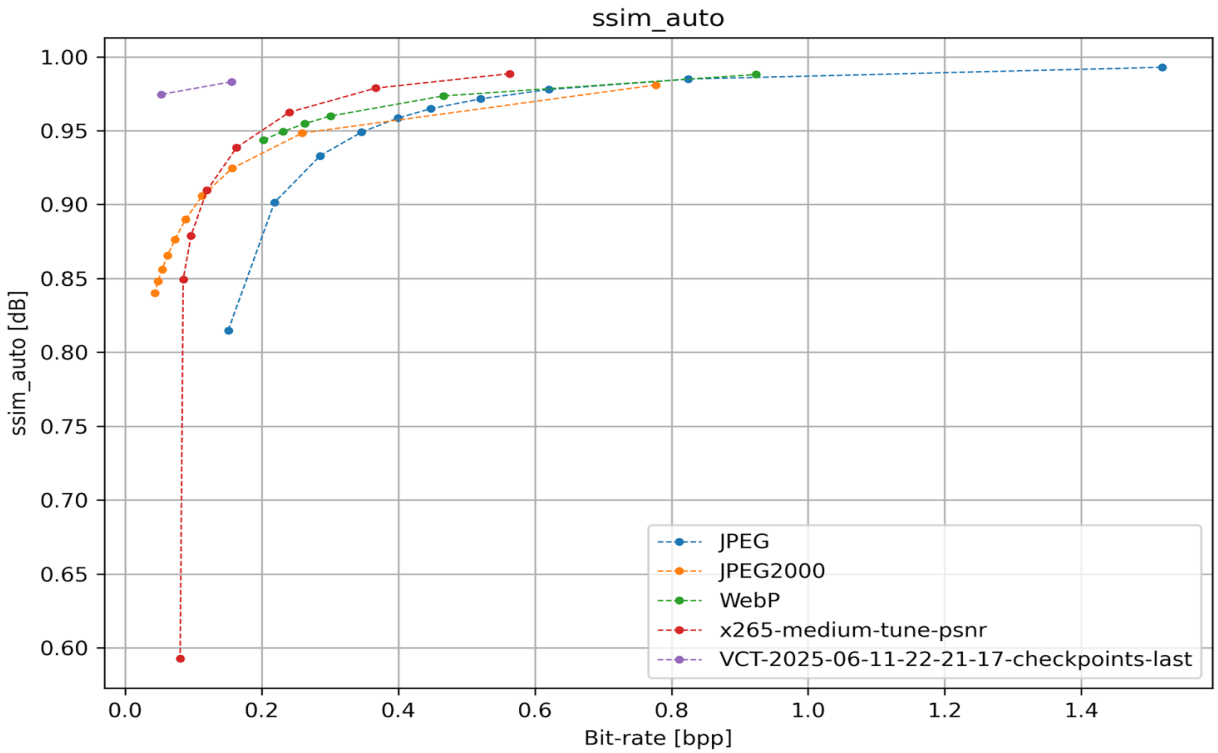

Figure 9: Rate-Distortion plot for VIT to highlight its potential to maintain structural quality across compression levels compared to traditional, image- and video-based approaches.

With our first architectures pre-trained on rate–distortion, we confirm that VCT is able to both outperform neural image–compression baselines as well as classical video codecs, demonstrating the advances this architecture provides, cf. Fig. 9. By conditioning timesteps and image tokens on early encoded frames, the average bits-per-pixel needed for successive images significantly drops, since knowledge about static shapes can be leveraged and is captured in the learned prior distribution for the entropy-coding step. We currently train VCT on the four timesteps of SSL4EO-S12 data and configure it to use a context length of two previous frames. We expect further improvements by extending the context to three reference frames and by adding B-frame prediction—using both past and future reference frames during training—so that we can condition more flexibly during inference.

Given VCT learns temporal priors with a transformer architecture that ingests both past and current frame–patched latent representations, these priors can extract highly informed, data-specific knowledge about the EO time series at hand. We probe this by decoding the model's prior belief—i.e., what it predicts the next frame and patch will look like—rather than the actual encoded latent embedding. Our analysis in Fig. 10 demonstrates how accurately the model anticipates future frames: even with little or no information about the current timestep, it can generate visually similar reconstructions. As we transmit more tokens from the current frame (increasing k in Fig. 10), the

reconstruction becomes incrementally more precise—offering the flexibility to adjust how much information is sent, while still generating plausible satellite images from past context alone.

Moving forward, we aim to extend our findings by a more semantic loss to produce embeddings that are informative for both reconstruction and downstream tasks. To that end, we're adding new temporal evaluation tasks (e.g., agriculture monitoring) to our benchmark suite. We will test the integration of masked autoencoding into

- either the transformer that jointly processes patches from past and current frames
- or the per-image transformers.

We aim to reinforce the model's understanding of the general image content and the relationships between spatial and temporal tokens. Our approach will pave the way to novel neural video compression methods that we hope to eventually apply to the vast stream of weather model data.

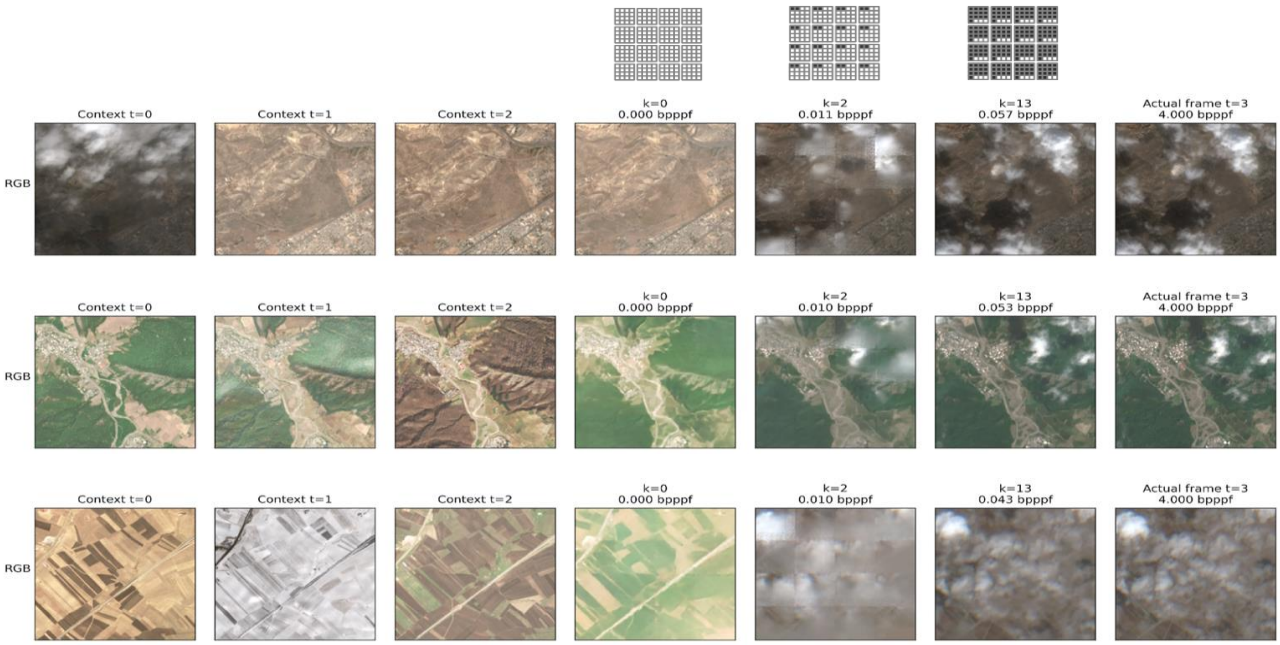

Figure 10: Ability of VCT to predict future frames given a fraction of tokens (k=0 corresponds to no token) of the frame to predict at t=3 – Model Belief Sharpens with More Decoded Tokens.

INTEGRATION WITH E2S USE CASES & ROADMAP

E2S USE CASES

Given the methods evaluated and developed here based on the SSL4EO-S12-downstream benchmark dataset (D4.5) utilizing NeuCo-Bench (D2.3), we draw the following conclusions regarding use the use cases of WP4:

1. As our NC methodology development and performance testing stands today, **image-level agriculture monitoring** with highly compressed embeddings is an option in practical reach. A lightweight, no-training NC model such as TMOSAIKS could easily get deployed on the edge of satellites to downlink monitoring embeddings for close-to real-time applications.

2. For **global biomass applications** a less sparse, hard-to-come-by ground truth dataset is required to increase confidence in performance evaluation through linear probing of image-level regression tasks. However, this use case may shift to dense, pixel-level predictions which TerraMind is able to provide. Nevertheless, it remains to increase TerraMind to compression factors beyond single digits. An alternative route for the biomass use case is feature upscaling of highly compressed embeddings as elaborated in, e.g., [FeatUP].

3. As it stands, **applications involving clouds** are more challenging from the perspective of the NCB framework. However, single timestamp compression may improve the situation over the current 4 seasons SSL4EO-S12 datacubes. Our developments of [ClimateBenchPress] will pave the way towards a more systematic study of that subject. In addition, our work on time series compression of EO data may help to identify cloud artifacts. Also, video compression is a natural subject to distill embeddings from temporally dense weather forecast data, in general.

4. The **maritime awareness use case** may require most attention for future developments. We see the following way forward: utilize TerraMind's any-to-any multi-modal capabilities and thinking-in-modalities feature to predict vessel localization from a diverse set of modalities such as Sentinel-1 and Sentinel-2 data aided by auxiliary data such as land cover information. Moreover, a multi-SAR sensor approach is in the discussion with advisory board member ICEYE given their large fleet of X-band radar satellites of multiple generations. As it stands, we aim to also test the *TMOSAIKS, MultiFMF,* and *MultiFMS* models on the `ships` SSL4EO-S12-downstream task for vessel localization.

ROADMAP FOR EXPLOITATION AND DISSEMINATION

Recently, IBM held workshops for partners involved in Embed2Scale use cases (WP4) to transfer knowledge of how to utilize the TerraMind foundation model. The E2S tech leads (DLR & IBM) are available to all consortium partners for consultation on how to integrate the methods presented here into their work flows.

As DLR leads the developments on data fusion of embeddings (D2.4), the team currently performs research on how to exploit D2.2. In fact, the application of the NeuCo-Bench framework to SSL4EO-S12 data cubes demonstrates the fusion of two modalities, namely: radar (Sentinel-1) and multi-spectral (Sentinel-2). A paper titled “NeuCo-Bench: A Novel Benchmark Framework for Neural Compression” has been crafted and is currently under peer review.

At the same time we are currently growing the *Earth2Vec* community (D5.3, D2.3) which gathers more than 30 partners from academia, the corporate sector, and government organizations as of Aug 1, 2025. Earth2Vec members communicate through a dedicated slack channel including monthly meetings to discuss topics and synergies around embeddings for Earth observation. Specifically, the winning teams (KTH University, Bundeswehr University, and Microsoft Research) and their professional network (among them the main developers of MOSAIKS and Google Research [AlphaEarth]) plan to draft a paper on lessons learned from the CVPR EARTHVISON data challenge in order to guide future developments of lossy neural compression.

Furthermore, we envision the following benefits for Embed2Scale’s pillars to leverage embeddings:

1. **Embed2Transfer**: For our recent general assembly at DLR in July 2025 [E2S-DLR-GA] we invited tech experts of our advisory board partners OroraTech and ICEYE. Their satellite solutions offer on-board processing capabilities including GPU units or FPGAs. For use cases such as near real-time wildfire monitoring by infrared sensors (OroraTech) or large-scale observation of forest inventories and building collections through X-band SAR (ICEYE), we currently sense opportunities to more closely collaborate.

2. **Embed2Infer**: Extending the NeuCo-Bench framework in conjunction with the Earth2Vec community will be vital to deploy our developments related to expressive embeddings that apply to a wide range of downstream tasks through simple linear probing. As quoted above, in WP2 we crafted a corresponding paper under peer review, with another paper planned in collaboration with the CVPR EARTHVISION data challenge winners and their professional network.

3. **Embed2Find**: as per an informal vote of the Earth2Vec community at its kick-off in Frascati, Italy [ESA-NASA-GEOFM], Embed2Find seems the theme that resonates most with stakeholders. IBM investigates integration of scalable vector databases for efficient storage and retrieval of neurally compressed embeddings (WP3). In the light of developing data fusion methodologies (D2.4), joint embedding spaces generated by multi-modal geo-foundation models will play a key role to implement Embed2Find solutions. Earth2Vec members such as CloudFerro will play a vital role to push the boundaries of this front.

REFERENCES

[TERRAMIND] 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. (2025). Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*.

[4M] Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., & Zamir, A. (2023). 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, *36*, 58363-58408.

[MOSAIKS] Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., & Hsiang, S. (2021). A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery. Nature Communications, 12(1), 4392.

[AI4G] Corley, I., Ekim, B., Robinson, C., Rolf, E. Temporal MOSAIKS, (2025), GitHub repository, https://github.com/isaaccorley/temporal-mosaiks

[NEUCO] NeuCo-Bench: A Novel Benchmark Framework for Neural Compression (attachment to deliverable D2.3)

[KTH] Kerekes, D., Brune, E., Yadav, R., Nascetti, A. Foundation model embedding compression for the EMbed2Scale challenge, (2025), Github repository, https://github.com/KerekesDavid/embed2scale-solution

[CROMA] Fuller, A., Millard, K., & Green, J. (2023). CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. Advances in Neural Information Processing Systems, 36, 5506-5538.

[DOFA] Xiong, Z., Wang, Y., Zhang, F., Stewart, A., Hanna, J., Borth, D., Papoutsis, I., Le Saux, B., Camps-Valls, G., & Zhu, X. (2024). Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. arXiv e-prints, arXiv:2403.15356.

[SCALEMAE] Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., & Darrell, T. (2023). Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. AGU23.

[PRITHVI2] Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, Þ. E., Blumenstiel, B., Ghosal, R., de Oliveira, P .H., de Sousa Almeida, J. L., 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.. (2025). Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications.

[COPERNICUSFM] Wang, Y., Xiong, Z., Liu, C., Stewart, A. J., Dujardin, T., Bountos, N. I., ... & Zhu, X. X. (2025). Towards a unified copernicus foundation model for earth vision. *arXiv preprint arXiv:2503.11849*.

[VIT] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[CONVNEXT] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).

[CLIP] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[DINO] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).

[GEORSCLIP] Zhang, Z., Zhao, T., Guo, Y., & Yin, J. (2024). RS5M and GeoRSCLIP: A large scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing.

[SSL4EO] Blumenstiel, B., Braham, N. A. A., Albrecht, C. M., Maurogiovanni, S., & Fraccaro, P. (2025). Ssl4eo-s12 v1. 1: A multimodal, multiseasonal dataset for pretraining, updated. arXiv preprint arXiv:2503.00168.

[VCT] Fabian Mentzer et al. VCT: A Video Compression Transformer. arXiv:2206.07307

[VCScale-Space] Eirikur Agustsson et al. "Scale-Space Flow for End-to-End Optimized Video Compression".
In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[MAE] Kaiming He et al. Masked Autoencoders Are Scalable Vision Learners. 2021. arXiv: 2111.06377

[Earth+] K. Du, Y. Cheng, P. Olsen, S. Noghabi, R. Chandra, and J. Jiang, “Earth+: On-board satellite imagery compression leveraging historical earth observations,” 2024, arXiv:2403.11434.

[DVC] Guo Lu et al. DVC: An End-to-end Deep Video Compression Framework. arXiv:1812.00101

[CopernDash] https://dashboard.dataspace.copernicus.eu

[JPEG2000] Pen-Shu Yeh *et al.*, "The new CCSDS image compression recommendation," *2005 IEEE Aerospace Conference*, Big Sky, MT, USA, 2005, pp. 4138-4145, doi: 10.1109/AERO.2005.1559719.

[E2SReview] C. Gomes *et al.*, "Lossy Neural Compression for Geospatial Analytics: A review," in *IEEE Geoscience and Remote Sensing Magazine*, doi: 10.1109/MGRS.2025.3546527.

[VQVAE] Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." *Advances in neural information processing systems* 30 (2017).

[TravoIS] Esser, Patrick, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[ScaleNC4EO] https://meetingorganizer.copernicus.org/EGU25/EGU25-19016.html

[EventCAM] Li, Hongmin, et al. "Cifar10-dvs: an event-stream dataset for object classification." *Frontiers in neuroscience* 11 (2017): 309.

[FeatUP] Fu, Stephanie, et al. "Featup: A model-agnostic framework for features at any resolution." *arXiv preprint arXiv:2403.10516* (2024).

[ClimateBenchPress] https://doi.org/10.5194/egusphere-egu25-15912

[AlphaEarth] https://deepmind.google/discover/blog/alphaearth-foundations-helps-map-our-planet-in-unprecedented-detail , note: Conrad M Albrecht (DLR) got invited to LPS session (organized by ESA) where Google DeepMind presented AlphaEarth: https://lps25.esa.int/programme/programme-session/?id=2CED0B00-B9E1-41B7-8288-3FE11524FADB

[ESA-NASA-GEOFM] https://nikal.eventsair.com/nasa-esa-international-workshop-on-geospatial-ai-foundation-model-for-earth-observation-and-earth-sciences

[E2S-DLR-GA] https://www.dlr.de/en/eoc/latest/news/2025/ai-in-space-embed2scale