LEPA: Learning Geometric Equivariance in Earth Observation with a Predictive Architecture

Erik Scheurer[1,2] **Rocco Sedona**[1] **Stefan Kesselheim**[1] **Gabriele Cavallaro**[1,3]
[1]Forschungszentrum Jülich, Germany [2]University of Stuttgart, Germany
[3]University of Iceland, Iceland
{e.scheurer, r.sedona, s.kesselheim, g.cavallaro}@fz-juelich.de

Abstract

Recent work has introduced embedding datasets for Earth observation (EO), where pretrained foundation models provide precomputed feature vectors to reduce downstream computational cost. Yet, geometric interpolation of embeddings remains underexplored and does not necessarily yield meaningful representations. We address this by training Joint-Embedding Predictive Architecture (JEPA) models to learn geometric operations in embedding space, using both NASA/USGS HLS data and ImageNet-1k. Evaluation on the PANGAEA benchmark shows that, without architectural changes, JEPA achieves competitive performance to other foundation models on EO tasks. Conditioning the predictor on geometric augmentations to get a learned equivariance-predicting architecture (LEPA) improves mean reciprocal rank (MRR) from 0.2 to 0.7, with further gains from fine-tuning. An embedding analysis reveals classification-specific noise in ImageNet-1k, which we mitigate by introducing a CLS-token, improving MRR.

1 Introduction

To increase accessibility, speed up prototyping, and alleviate computational and memory bottlenecks, embedding datasets are becoming common in Earth observation (EO) [5, 8, 10]. Instead of downloading and preprocessing terabytes of satellite data, embeddings, vectorial proxies of the data, are precomputed with foundation models, yielding representations more expressive and compact than the raw inputs [29, 16, 5]. However, users may have proprietary data at positions or resolutions mismatched with the embeddings. To avoid repeated expensive encoder passes, it is desirable to transform embeddings directly rather than transform images first. While image interpolation is well-studied [24], geometric operations on embeddings, especially patch embeddings, may not be meaningful: latent-space structure depends on the foundation model and can be highly non-convex [11]. If a transformation t in embedding space of encoder E satisfies $t(E(x)) = E(T(x))$ for an image-space transformation T, then E is *equivariant* to T. Approximating such equivariance is an active research area [7, 20, 19, 15, 3, 14], though most approaches have not been scaled, are not steerable, or use non-geometric transformations. To address these limitations, we train an I-JEPA [1] model on ImageNet-1k [26] and HLS data [29] as a baseline and evaluate it on PANGAEA [23]. We show that an unmodified architecture already yields competitive results. We then improve equivariance through a pretraining strategy that enforces predictor sensitivity to geometric transformations, inspired by image world models [14]. This substantially increases mean reciprocal rank (MRR), a measure of equivariance, with further gains from equivariance fine-tuning. We call this a learned equivariance-predicting architecture (LEPA). Finally, we analyze the embeddings to compare datasets and adjust the model architecture to further enhance equivariance in I-JEPA models.

2 Method

I-JEPA The pretraining task of the image-based joint-embedding predictive architecture (I-JEPA) [1] is latent inpainting. A teacher produces representations from the full image, and blocks of patches are selected as targets. A ViT encoder processes a non-overlapping region to provide context to the key novelty of I-JEPA: a predictor that inpaints missing blocks conditioned on the context via cross-attention. The teacher is an exponential moving average of the student. I-JEPA has been applied in EO [6, 2, 30, 21], but these works focus on representation quality rather than equivariance. A follow-up, image world models [14], introduces an additional pretraining task to enforce equivariance: given an augmented context and the augmentation parameters, predict target embeddings generated from the unmodified image. The predictor thus learns both latent inpainting and approximate reversal of augmentations such as color jitter, grayscale, and contrast changes, producing embeddings equivariant to the augmentations used [14]. We adapt this framework to geometric transformations.

LEPA Geometric transformations such as rescaling, rotation and translation are not trivially defined for patch-wise embeddings, since parts of an image patch move to neighboring patches under these operations. A linear combination of embeddings, as in interpolation, may not yield a valid vector in a highly non-convex embedding manifold. Thus the predictor must capture both spatial information within each patch-embedding and the desired geometric transformation. Training pairs are obtained by transforming the unmodified context into a target image using translation in x/y, rotation, and scaling. These parameters are appended to the predictor's mask-token and passed through a 3-layer MLP projecting back to the embedding dimension, enabling the predictor to inpaint missing patches while performing the transformation in embedding space. For better performance, the full target image is predicted rather than blocks as in I-JEPA. We also test an inductive bias via novel conditioned positional encodings whose sinusoidal indices are centered on the image rather than a corner, allowing them to transform according to the target parameters. A formal definition and illustration are provided in Appendix A.

3 Experiments

3.1 Model Training

Both I-JEPA and LEPA train for 50 epochs either on ImageNet-1k [26] or the HLS dataset used for pretraining Prithvi-EO-2.0 [29] based on a ViT-base architecture [9]. We found that 50 epochs are sufficient for performance convergence even though the representations continue to change. The number of epochs is chosen to decrease training time and to avoid overfitting to the data. With longer training the noise discussed in Sec. 3.3 increases and therefore the equivariance decreases.

3.2 Representation Quality

The PANGAEA [23] benchmark is used to measure embedding quality. For each dataset in PANGAEA a UPerNet decoder [31] is trained to map from frozen encoder representation to semantic segmentation map. Despite UPerNets's design, no intermediate outputs are passed to the decoder to imitate a fixed embedding dataset. As a baseline, we compare Prithvi-EO-2.0 [29], TerraMind [16], RemoteCLIP [22], DOFA [32] and CROMA [12] with our JEPA variants in Tab. 1. JEPA models perform competitively without architectural changes. While TerraMind-L outperforms all variations in most datasets, their model and dataset is much larger and can even use multimodal information in cases where Sentinel-1 information exists, i.e. Sen1Floods11, CropTypeMapping and PASTIS. We visualize these results and compute an overall normalized score in appendix B.

Impact of Dataset When training with a JEPA target, we obtain high-quality embeddings. Pretraining data strongly influences downstream performance: ImageNet-1k-pretrained I-JEPA achieves competitive scores despite being out-of-distribution. Unlike ImageNet's fixed, class-balanced sampling, HLS is stratified across land-cover classes and ecoregions, with entropy-based tiles for heterogeneous areas. Temporal sequences (four seasonal timestamps) were extracted per tile, filtered for clouds, and split into patches for spatial and temporal diversity [29]. Thus, ImageNet emphasizes single-object classification, while HLS captures diverse landscapes. The ImageNet variant of I-JEPA outperforms others on the Marine Debris and Oil Spill (MADOS) dataset, which is challenging due to 15 classes

Table 1: Mean IoU comparison of pretrained foundation models and JEPA variations of the PANGAEA benchmark [23]. The **best** and second best models are marked for each dataset.

Model	Dataset	CLS	PosEnc	Params	HLSBurnscars	Sen1Floods11	MADOS	AI4SmallFarms	PASTIS	DynamicEarthNet	SpaceNet7
I-JEPA	IN-1k	No	default	85.8M	77.15	71.75	59.28	26.58	21.31	35.13	60.02
I-JEPA	IN-1k	Yes	default	85.8M	77.58	74.74	**65.59**	25.98	24.07	37.68	61.12
I-JEPA	HLS	No	default	86.4M	82.82	86.69	45.74	24.33	33.00	29.95	56.93
I-JEPA	HLS	Yes	default	86.4M	83.08	85.87	46.40	25.48	35.03	31.46	56.67
I-JEPA	HLS	Yes	CondPos	86.4M	81.73	85.88	47.59	24.58	34.43	32.64	56.64
LEPA	HLS	No	CondPos	86.4M	83.03	87.29	51.40	23.82	33.84	33.84	56.17
LEPA	HLS	Yes	CondPos	86.4M	82.71	85.44	45.14	24.01	33.77	32.13	55.61
LEPA	HLS	No	default	86.4M	82.72	87.37	43.36	23.26	33.61	28.06	56.31
NoMask	HLS	No	CondPos	86.4M	80.81	83.89	33.51	21.12	29.12	26.65	51.96
Prithvi-EO-2.0-100M [29]				86.4M	80.28	87.12	41.46	26.03	31.12	26.27	57.12
RemoteCLIP [22]				87.5M	70.98	70.28	53.11	23.09	15.01	38.28	54.33
TerraMind-L [16]				305.7M	**83.16**	**89.58**	59.74	**26.80**	**43.08**	**39.39**	**61.35**
DOFA [32]				111.3M	76.63	85.51	50.32	25.70	28.81	38.00	59.04
CROMA [12]				201.5M	79.56	87.26	54.97	23.90	33.72	37.97	56.66

and skewed distribution [17]. The focus of ImageNet on classes likely aids detection. ImageNet variants also exceed HLS when no multispectral data are available.

Architecture Modifications Prepending a classification (CLS) token to the patch embeddings to aggregate global information [9] improves semantic segmentation for ImageNet models across datasets (except AI4SmallFarms), even though only patch embeddings feed the decoder. HLS models show no consistent effect when adding a CLS-token, suggesting differences lie within decoder variance. Modified positional encodings and the auxiliary transformation prediction objective had no measurable impact, confirming earlier findings [9] that positional encoding choice matters less than its presence.

3.3 Representation Equivariance

We evaluate the equivariance of representations using the Mean Reciprocal Rank (MRR) [14, 18]. For each image, a series of 256 different augmentations is computed. Then an unmodified version of the image is passed through the encoder and predictor trying to predict the embeddings of one of the previous transformations. The different augmentations are sorted by cosine similarity to the predicted embedding. The mean reciprocal rank is computed by averaging these ranks as

$$\text{MRR} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\text{rank}_n}, \tag{1}$$

where rank_n is the rank of the n-th sample and N the number of images in the evaluation set. This metric is a useful proxy to the capacities of the model, as it can differentiate between invariance and equivariance. An invariant model produces similar embeddings for each augmentation, yielding an unstable rank and thus a lower MRR. This makes the metric more informative than a simple L^2 distance or cosine similarity between target and predicted embeddings, which are also small for invariant models. For encoders without a predictor, the augmented embedding is computed by bilinearly interpolating the patch embeddings.

Embedding analysis When analyzing the embeddings of the ImageNet model, a class-specific noise is noticeable, as seen in the background of Fig. 1. This noise resembles that observed in DINOv3 [27], since regions with different content show high similarity to parts containing the main subject, and because the effect increases with longer training. Adding a CLS-token reduces or removes these artifacts. The HLS model, not trained on a fixed set of concepts and lacking a central subject, does not exhibit the same artifact pattern. However, artifacts still appear (Fig. 1) and do not fully disappear even with a CLS-token, which is reflected in the MRR score in Tab. 2. A CLS-token improves interpolation equivariance for ImageNet models but not for HLS I-JEPA models, and for LEPA it even decreases MRR. We hypothesise that this occurs because the CLS-token is never masked during training: if the model stores class-related information in this token, the predictor may use it for inpainting. The open-ended nature of HLS data prevents useful global image information from being concentrated in a single CLS-token.

Predicting augmentations Conditioning the predictor on augmentation parameters markedly improves equivariance. With LEPA, the MRR increases from around 0.2 to nearly 0.7. Further finetuning of the predictor, while keeping the encoder frozen and training it solely to predict augmentations

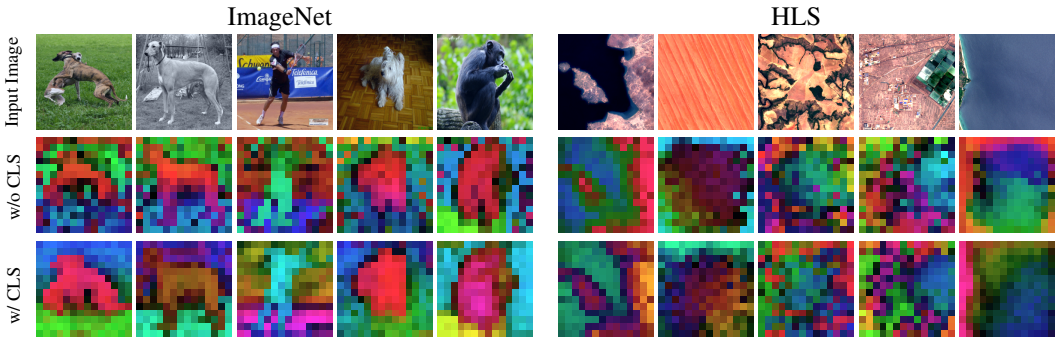

Figure 1: Example images with class-specific noise. Bottom rows show the first two PCA components (per image) mapped to a color wheel. In the ImageNet model without a CLS-token, background embeddings resemble the subject, a pattern not seen consistently in HLS data.

Table 2: MRR of models given different hyperparameters with and without finetuning of the predictor.

Model	Dataset	CLS	Pos. Enc.	MRR	After Finetune
I-JEPA	IN-1k	No	default	0.1732	N/A
I-JEPA	IN-1k	Yes	default	0.2019	N/A
I-JEPA	HLS	No	default	0.1743	N/A
I-JEPA	HLS	Yes	default	0.1755	N/A
LEPA	HLS	No	CondPos	**0.6975**	0.8062
LEPA	HLS	Yes	CondPos	0.6630	0.7994
LEPA	HLS	No	default	0.6183	**0.8355**
NoMask	HLS	No	CondPos	0.5906	N/A
Prithvi-EO-2.0-100M				0.1973	N/A

without inpainting, pushes the score to 0.8. This setup illustrates how the predictor can be deployed to adjust a grid of embeddings for sampling specific points. However, we cannot use this as the exclusive objective, as shown by the no-masking training in Tab. 2: it underperforms in semantic segmentation (Tab. 1) and does not necessarily yield a high MRR score. With this single objective, training becomes unstable, and the low equivariance is due to collapse into simpler invariant embeddings. Conditioning through positional encodings produces more equivariant embeddings than merely appending the parameters to the mask tokens and passing them through an MLP; after finetuning, however, the positional encodings reduce embedding equivariance.

4 Future work

We identify several directions to strengthen evaluation, equivariance and performance. First, we will test whether a fixed number of classes induces the observed artifacts by training one model on a classification-oriented EO dataset and another on a general-purpose RGB dataset. Second, additional foundation models can be probed for equivariance; since prior work [13, 3, 33] shows that linear predictors often suffice, a smaller predictor for LEPA and other models may reduce inference cost. Third, LEPA's inductive bias could be improved with relative positional encodings such as ALiBi [25] or RoPE [28], which also yield size invariance. Finally, downstream performance and embedding consistency may benefit from an auxiliary loss, analogous to the CLIP objective in AlphaEarth [5], aligning embeddings of the same land type across images using the multimodal TerraMesh dataset [4] for pretraining.

5 Acknowledgements

This research is carried out as part of the Embed2Scale project and is co-funded by the EU Horizon Europe program under Grant Agreement No. 101131841. Additional funding for this project has been provided by the Swiss State Secretariat for Education, Research and Innovation (SERI) and UK Research and Innovation (UKRI).

References

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.01499.

[2] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19530–19540, June 2025.

[3] S. Bhardwaj, W. McClinton, T. Wang, G. Lajoie, C. Sun, P. Isola, and D. Krishnan. Steerable equivariant representation learning. *arXiv preprint arXiv: 2302.11349*, 2023.

[4] B. Blumenstiel, P. Fraccaro, V. Marsocci, J. Jakubik, S. Maurogiovanni, M. Czerkawski, R. Sedona, G. Cavallaro, T. Brunschwiler, J. Bernabe-Moreno, et al. TerraMesh: A planetary mosaic of multimodal earth observation data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

[5] C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, W. J. Rucklidge, M. Samsikova, C. Zhang, E. Shelhamer, E. Lahera, O. Wiles, S. Ilyushchenko, N. Gorelick, L. L. Zhang, S. Alj, E. Schechter, S. Askay, O. Guinan, R. Moore, A. Boukouvalas, and P. Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv: 2507.22291*, 2025.

[6] S. Choudhury, Y. Salunkhe, S. Mehrotra, and B. Banerjee. Rejepa: A novel joint-embedding predictive architecture for efficient remote sensing image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2373–2382, June 2025.

[7] T. Cohen and M. Welling. Group equivariant convolutional networks. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 20–22 Jun 2016. URL `https://proceedings.mlr.press/v48/cohenc16.html`.

[8] M. Czerkawski, M. Kluczek, and J. S. Bojanowski. Global and dense embeddings of earth: Major tom floating in the latent space, 2024. URL `https://arxiv.org/abs/2412.05600`.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[10] Z. Feng, C. Atzberger, S. Jaffer, J. Knezevic, S. Sormunen, R. Young, M. C. Lisaius, M. Immitzer, T. Jackson, J. Ball, D. A. Coomes, A. Madhavapeddy, A. Blake, and S. Keshav. Tessera: Precomputed fair global pixel embeddings for earth representation and analysis, 2025. URL `https://arxiv.org/abs/2506.20380`.

[11] M. F. Frenzel, B. Teleaga, and A. Ushio. Latent space cartography: Generalised metric-inspired measures and measure-based transformations for generative models. *arXiv preprint arXiv: 1902.02113*, 2019.

[12] A. Fuller, K. Millard, and J. R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[13] Q. Garrido, L. Najman, and Y. Lecun. Self-supervised learning of split invariant equivariant representations. In *International Conference on Machine Learning*. PMLR, 2023.

[14] Q. Garrido, M. Assran, N. Ballas, A. Bardes, L. Najman, and Y. LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv: 2403.00504*, 2024.

[15] S. Gupta, J. Robinson, D. Lim, S. Villar, and S. Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lgaFMvZHSJ.

[16] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, et al. TerraMind: Large-scale generative multimodality for earth observation. *ICCV'25*, 2025.

[17] K. Kikaki, I. Kakogeorgiou, I. Hoteit, and K. Karantzalos. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024.

[18] T. Kipf, E. van der Pol, and M. Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gax6VtDB.

[19] T. Kouzelis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UWhW5YYLo6.

[20] S. Kundu and R. Kondor. Steerable transformers for volumetric data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Ax550Vokon.

[21] W. Li, W. Yang, T. Liu, Y. Hou, Y. Li, Z. Liu, Y. Liu, and L. Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2024.09.013. URL https://www.sciencedirect.com/science/article/pii/S0924271624003514.

[22] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

[23] 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. *arXiv*, 2024. URL https://arxiv.org/abs/2412.04204.

[24] E. Meijering. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342, 2002. doi: 10.1109/5.993400.

[25] O. Press, N. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Apr. 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y.

[27] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. DINOv3. *arXiv preprint arXiv: 2508.10104*, 2025.

[28] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

[29] 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. *arXiv preprint arXiv: 2412.02732*, 2025. doi: 10.48550/arXiv.2412.02732.

[30] G. Tseng, A. Fuller, M. Reil, H. Herzog, P. Beukema, F. Bastani, J. R. Green, E. Shelhamer, H. Kerner, and D. Rolnick. Galileo: Learning global local features of many remote sensing modalities, 2025. URL `https://arxiv.org/abs/2502.09356`.

[31] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[32] 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 foundation model for observing the Earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024.

[33] J. Yang, N. Dehmamy, R. Walters, and R. Yu. Latent space symmetry discovery. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56047–56070. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/yang24g.html`.

A Positional Encodings

2D positional encodings of a standard ViT [9] are sinusoidal encodings for height and width position of each patch. The full embedding dimension is split into two parts as follows:

$$\begin{aligned} PE_{(pos_h,2i)} &= \sin(pos_h/10000^{2i/d_{model}}), \\ PE_{(pos_h,2i+1)} &= \cos(pos_h/10000^{2i/d_{model}}), \\ PE_{(pos_w,2i)} &= \sin(pos_w/10000^{2i/d_{model}}), \\ PE_{(pos_w,2i+1)} &= \cos(pos_w/10000^{2i/d_{model}}), \end{aligned} \tag{2}$$

with pos_w and pos_h denoting the x- and y-coordinates in a grid and $i = 0...D/2$ for embedding dimension D. These positional encodings are then concatenated into a single embedding vector. For our conditioned embeddings we only change the underlying grid that is used to obtain the positions. We show the different grids used to index the patches in Fig. 2. Our grid is centered at the origin of (0,0) instead of the edge of an image. With this modification, changing the angle, translation and scaling can easily be realized. These conditioned positional encodings are added to the mask-tokens in the same way as other absolute positional encodings. The generated positional encodings for each embedding dimension are shown in Fig. 3.

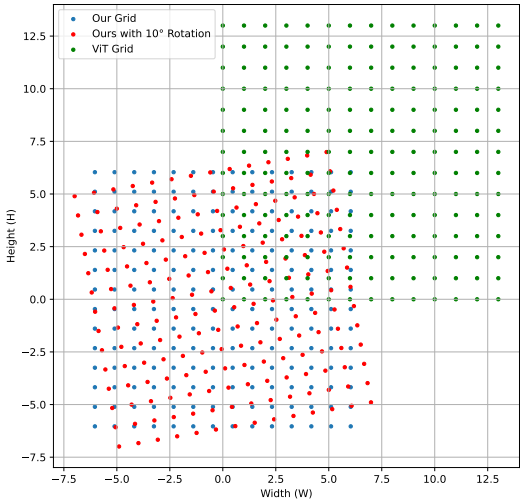

Figure 2: Grid that is used for sinusoidal positional encodings by default, ours without a transformation and ours with a rotation angle of 10°.

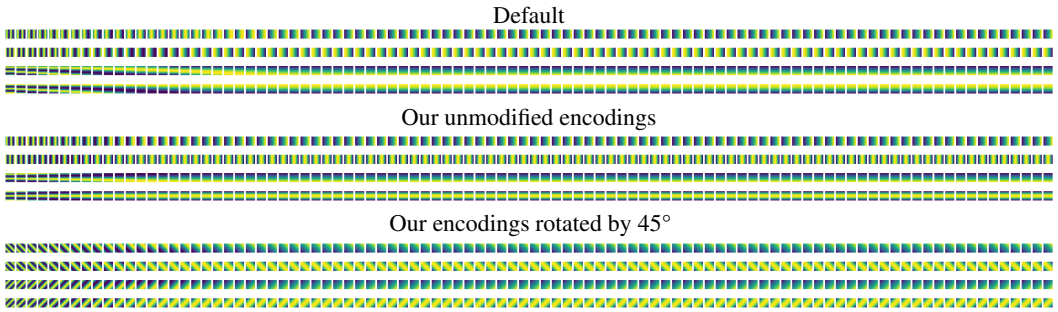

Figure 3: Comparison of positional encodings for each embedding dimension out of 768. Each row consists of one line of equation 2.

B Normalized Performance

To compute a global score for each model that takes into account all datasets, we first need to normalize the results to later compute the average for a model. We obtain mean and variance of each dataset using all PANGAEA models and two I-JEPA models. We decide on only two, one for ImageNet-1k and one for HLS pretraining data, to prevent bias in the distribution and because the different hyperparameter variations perform similarly. With a normalized score for each dataset, the model performance can be analyzed using a box plot in Fig. 4. TerraMind-L clearly outperforms all other models also with lower variation across datasets and our JEPA variants perform competitively with other foundation models. The HLS models show an overall lower variance across datasets compared to the ImageNet models which underperform on multispectral tasks. The plots indicate trends but are not perfect representations yet. Multiple runs are necessary to get an average score for each dataset before averaging as decoder initialization has a large impact on some PANGAEA datasets.

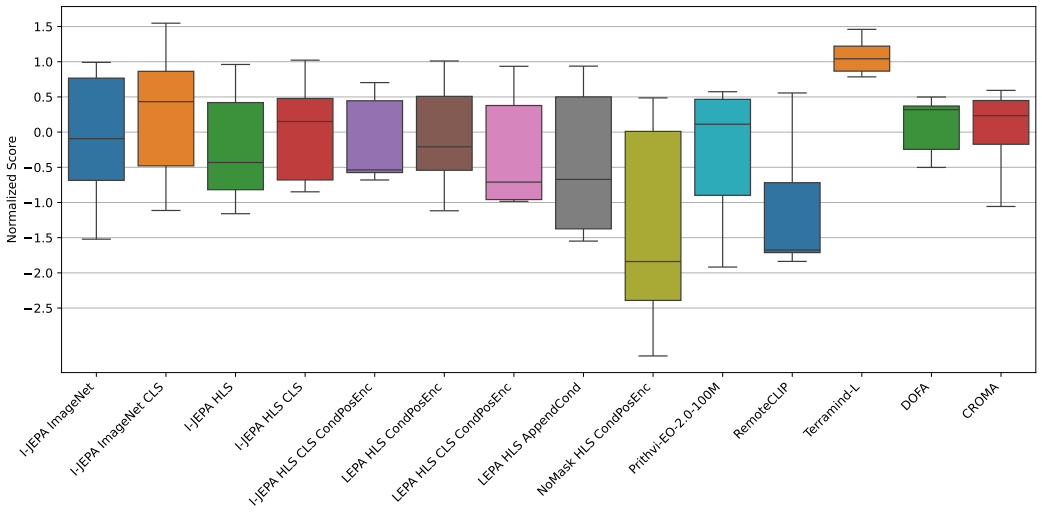

Figure 4: Normalized Performance with mean and standard deviation computed from ImageNet CLS, HLS CLS, Prithvi-EO, RemoteCLIP, TerraMind, DOFA and CROMA.