Multi-modal, multi-scale representation learning for satellite imagery analysis just needs a good ALiBi

4th Space Imaging Workshop

Patrick Kage

October 8, 2024

Problem statement

We want a representation learning system that fuses:

- Multi-modal (optical and SAR) geospatial data
- Multi-scale (high and low GSD) geospatial data

Problem statement

We want a representation learning system that fuses:

- Multi-modal (optical and SAR) geospatial data
- Multi-scale (high and low GSD) geospatial data

We're going to use a novel attention **(Scale-ALiBi)** and a combination of contrastive and reconstructive objectives to build this system.

Problem statement

We want a representation learning system that fuses:

- Multi-modal (optical and SAR) geospatial data
- Multi-scale (high and low GSD) geospatial data

We're going to use a novel attention **(Scale-ALiBi)** and a combination of contrastive and reconstructive objectives to build this system.

Right now there are systems which tackle both multi-scale (ScaleMAE[1], etc.) and multi-modal (CROMA[2], etc.) representations, but not both.

Scale-ALiBi

We present a transformer linear bias attention mechanism which incorporates cross-GSD-scale attention.

We use this to tie together three encoders for multi-scale multi-modal data (low-res optical, low-res SAR, and high-res optical) with a contrastive and reconstruction objective to form a representation learning system invariant to data modality and scale.

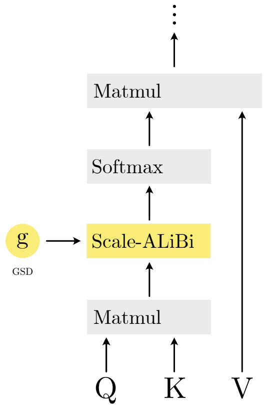


Figure: Scale-ALiBi transformer attention

Linear biases (ALiBi)

Linear bias positional encodings lets transformers learn sequence lengths longer than those presented at training time.[3]

Instead of adding sinusoidal positional encodings, this is added directly to the query-key product before softmax'ing the product.

$$a_{hij} = \sqrt{d} \cdot q_{hi} \cdot k_{hj} - distance(i, j) \cdot m(h) \quad (1)$$

$$m(h) = \left[\frac{1}{2^1}, \frac{1}{2^2}, \cdots, \frac{1}{2^8}\right] \quad (2)$$

Linear bias attention for each a_{hij} in attention matrix $A \in \mathbb{R}^{h \times L \times L}$ for h heads, sequence length L and head depth d.[2, 3]

Scale-ALiBi

CROMA[2] extended this to 2D representations by adding a Euclidean distance factor to the image patches.

We additionally scale this distance factor by the GSD of the sample, inspired by Scale-MAE[1].

We're calling the resulting attention "Scale-ALiBi."

$$a_{hij} = \underbrace{\sqrt{d} \cdot q_{hi} \cdot k_{hj}}_{\text{normal attention}} - \underbrace{g(i,j) \cdot m(h)}_{\text{Scale-ALiBi}} \tag{3}$$

$$g(i,j) = \text{distance(i,j)} \cdot \boxed{\text{GSD}} \tag{4}$$

Scale-ALiBi attention. Similar to before, but now with a GSD scaling variable. Attention matrix $A \in \mathbb{R}^{h \times L \times L}$ for h heads, sequence length L and head depth d.

ViT Tokenization Recap

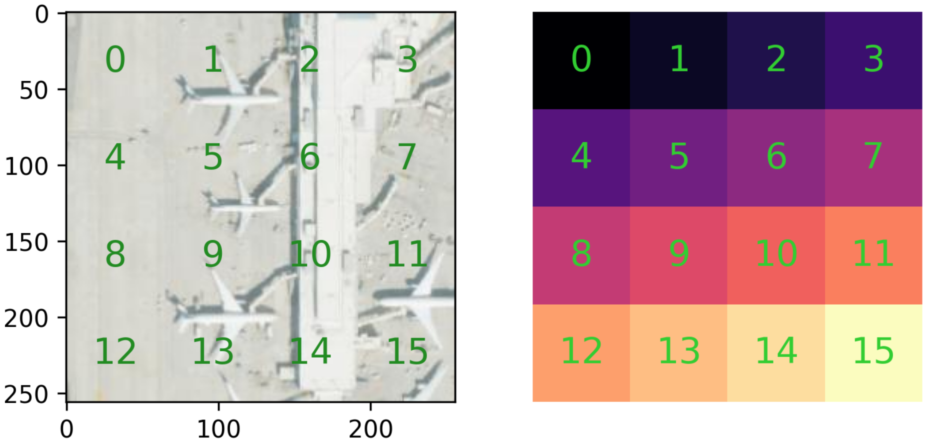


Figure: ViT tokenization of a 256×256 pixel image into 16 patches of size 64×64.

Scale-ALiBi attention: same GSD

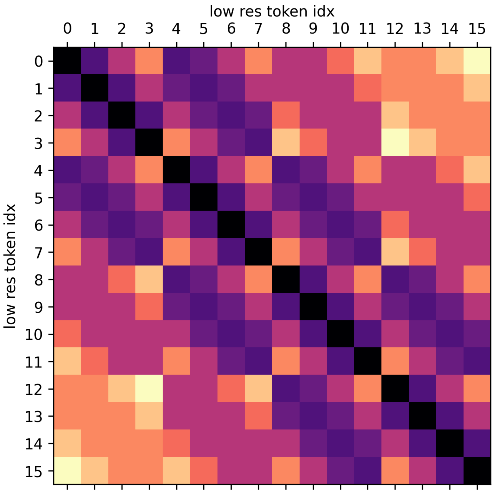


Figure: Scale-ALiBi cross-attention for images with the same GSDs and the same areas.

ViT tokenization recap (2× resolution)

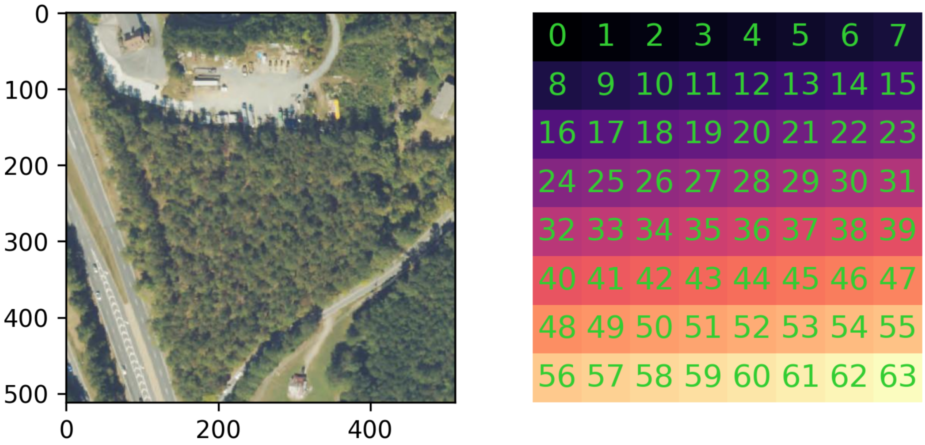


Figure: Double-resolution (512×512) ViT tokenization with 64 patches of size 64×64.

ViT tokenization recap (2× resolution)

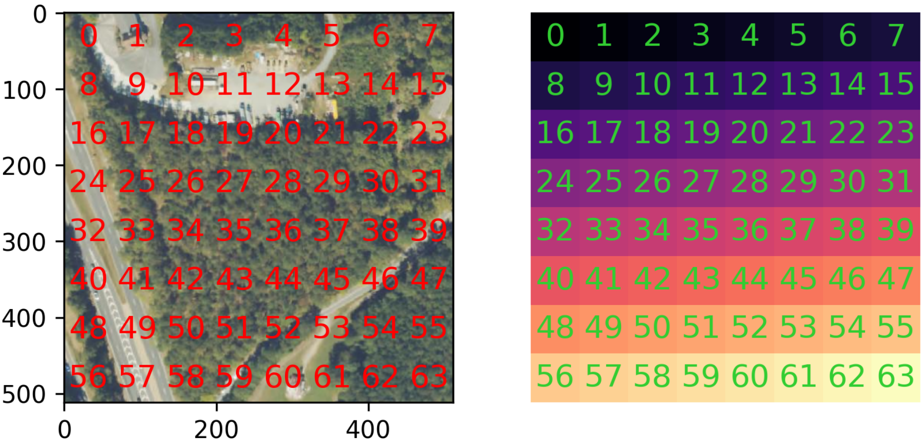


Figure: Double-resolution (512 × 512 pixel) ViT tokenization with 64 patches of size 64 × 64.

Scale-ALiBi attention: differing GSDs

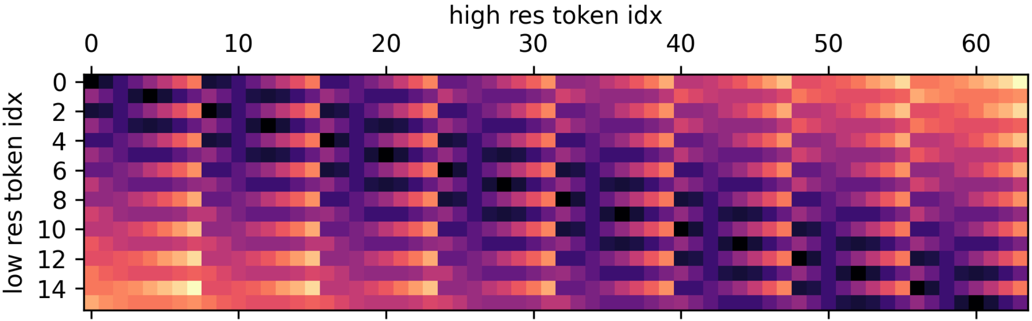


Figure: Scale-ALiBi cross-attention showing images with different sizes. Note that since the two images cover the same area but with double the resolution, the scale factor is 0.5.

Contrastive vs reconstructive representation learning

Contrastive learning is better at combining separate views, but performs poorly on high-frequency information.

Conversely, reconstruction objectives are much better at reconstructing fine-grained details[4].

We combine both to form the Scale-ALiBi architecture to ensure that we learn high-quality representations.

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Cont}} + \mathcal{L}_{\text{Recon}} \tag{5}$$

Our loss is a simple addition of these two components, CROMA showed that there was no benefit to weighting[2].

Full model architecture

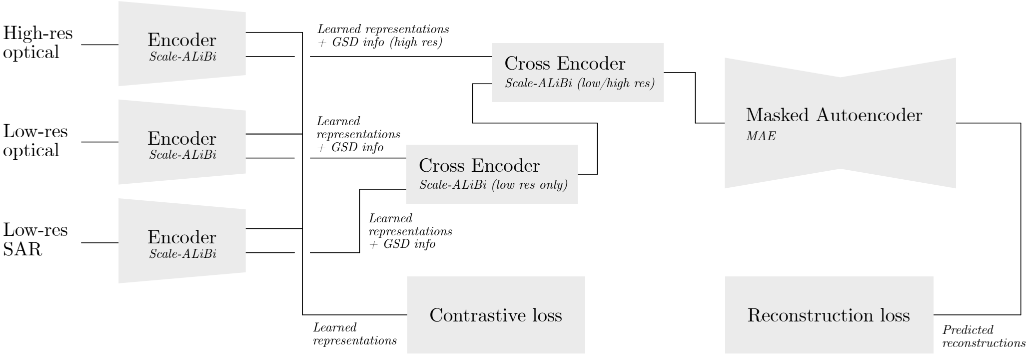


Figure: Scale-ALiBi training architecture.

Full model architecture

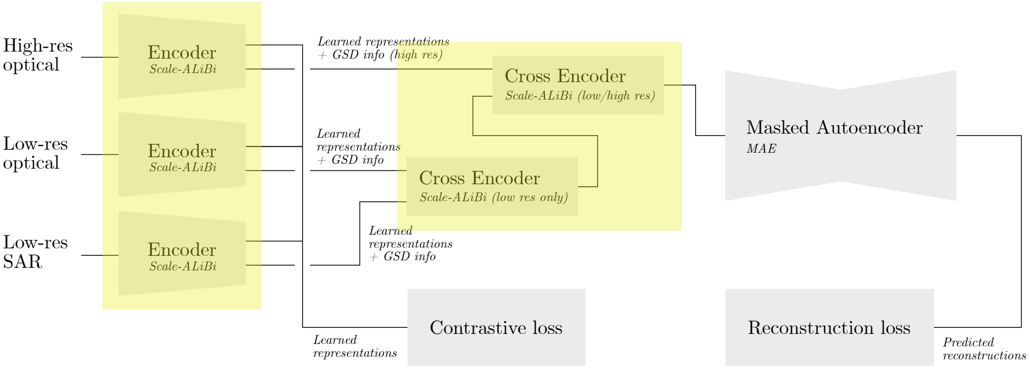
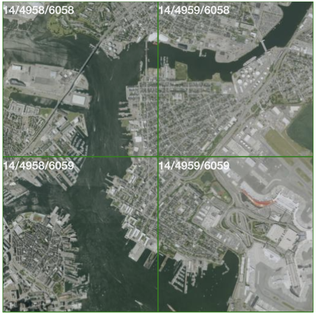


Figure: Scale-ALiBi training architecture, with Scale-ALiBi attention uses highlighted.

We collected a combination of Sentinel-1 (SAR)[5], Sentinel-2 (10m optical)[5], and NAIP (60cm optical)[6] imagery.

All samples aligned by *XYZ* tiles, using the *Z* difference as the GSD scale parameter.

Three datasets released: `small` (21,497 samples), `full` (188,060 samples), and `micro` (146,502 samples).



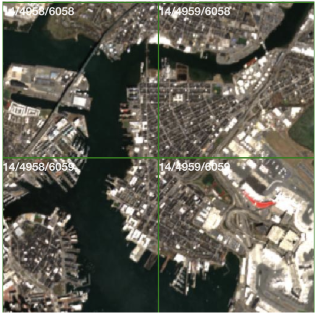


Figure: Comparison of XYZ tiles from NAIP & Sentinel-2 tiles[5, 6].

Dataset samples

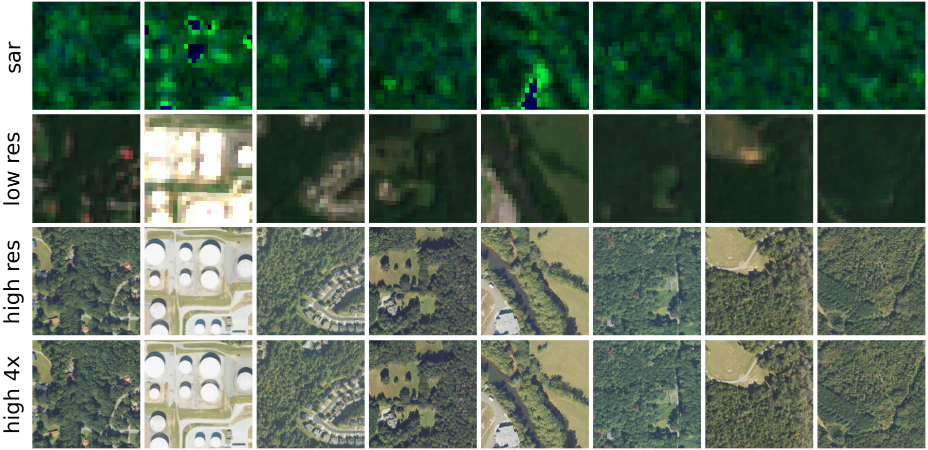


Figure: Samples from the Scale-ALiBi dataset `micro`. These tiles were generated from $Y = 17$.

Dataset samples

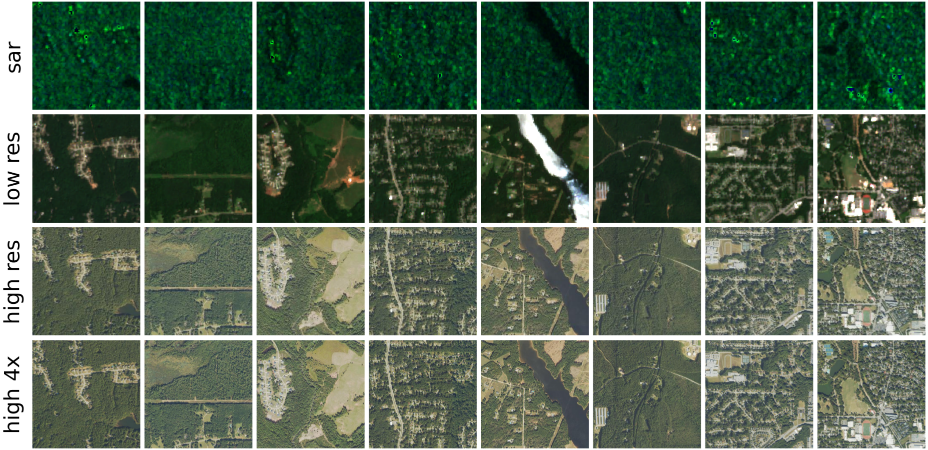


Figure: Samples from the Scale-ALiBi dataset `small`. These tiles (and `full`) were generated from $Y = 15$.

Benchmarking

We tested against GEO-Bench[7], consists of six classification and six segmentation tasks across data modes and GSDs.

We saw an improvement on GEO-Bench scores with our foundation model as compared to an identically-trained SOTA model (CROMA[2]).

Figure: pv4ger classification and segmentation benchmarks from GEO-Bench. Reproduced from [7].

We tested against GEO-Bench[7], consists of six classification and six segmentation tasks across data modes and GSDs.

We saw an improvement on GEO-Bench scores with our foundation model as compared to an identically-trained SOTA model, CROMA[2].

Name	SA-high	SA-low	CROMA
	k-NN		
`m-pv4ger`	**92.39**%	91.89%	92.29%
`m-forestnet`	**38.26**%	37.26%	35.44%
`m-euronet`	58.70%	64.40%	**66.30**%
`m-brick-kiln`	75.37%	74.97%	**76.47**%

Figure: Selected benchmarks comparing non-parametric embedding performance over classification tasks in GEO-Bench.

Conclusion and Future work

Conclusion

Overall, we showed that we are able to use the Scale-ALiBi attention to fuse low-resolution/high-resolution optical and low-resolution SAR images into a unified representation. We also released our dataset publicly for further representation learning work.

Future work

- Longer training run across larger cluster.
- Add additional modality encoders into the contrastive step.
- Retain the reconstruction autoencoder after training.

Q&A

Patrick Kage
p.kage@ed.ac.uk

Artificial Intelligence and its Applications Institute
The University of Edinburgh
Edinburgh, Scotland

ml.ka.ge/scale-alibi

[1] Colorado J. Reed et al. *Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning*. Sept. 2023. DOI: 10.48550/arXiv.2212.14532. arXiv: 2212.14532 [cs]. (Visited on 03/10/2024).

[2] Anthony Fuller, Koreen Millard, and James R. Green. "CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh et al. 2023.

[3] Ofir Press, Noah A. Smith, and Mike Lewis. *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*. Apr. 2022. arXiv: 2108.12409 [cs]. (Visited on 05/27/2024).

[4] Namuk Park et al. *What Do Self-Supervised Vision Transformers Learn?* May 2023. DOI: 10.48550/arXiv.2305.00729. arXiv: 2305.00729 [cs]. (Visited on 05/27/2024).

[5] European Space Agency. *Copernicus Sentinel Data, Processed by ESA*. 2024.

[6] U.S. Geological Survey. *National Agriculture Imagery Program (NAIP)*. 2024. DOI: 10.5066/F7QN651G. (Visited on 05/31/2024).

[7] Alexandre Lacoste et al. *GEO-Bench: Toward Foundation Models for Earth Monitoring*. Dec. 2023. DOI: 10.48550/arXiv.2306.03831. arXiv: 2306.03831 [cs]. (Visited on 05/28/2024).