UiT The Arctic University of Norway

NORA summer school on multi-modal learning

Multi-Modal Learning: Beyond Vision and Language

Challenges in Remote Sensing

Rwiddhi Chakraborty
UiT Machine Learning Group and Visual Intelligence

Schedule Today

- 09 - 12: Beyond Vision-Language – I and II
- 12 – 13: Lunch
- 13 – 16: Group Project

In this talk

Domain Challenges

Remote Sensing

Pre-training

Zero-Shot

In this talk

Domain Challenges

Remote Sensing

Pre-training

Zero-Shot

Beyond Vision and Language

Pre-trained vision-language models lack fine-grained understanding

Unsuitable for domains where specificity is key – e.g. remote sensing, biology, etc.

Text as a data mode has its own issues

Beyond Vision and Language

Real world datasets exist in multiple views and modalities

Only vision-language as a pre-training objective is too restrictive

We want to leverage similar ideas for diverse tasks

In this talk

Domain Challenges

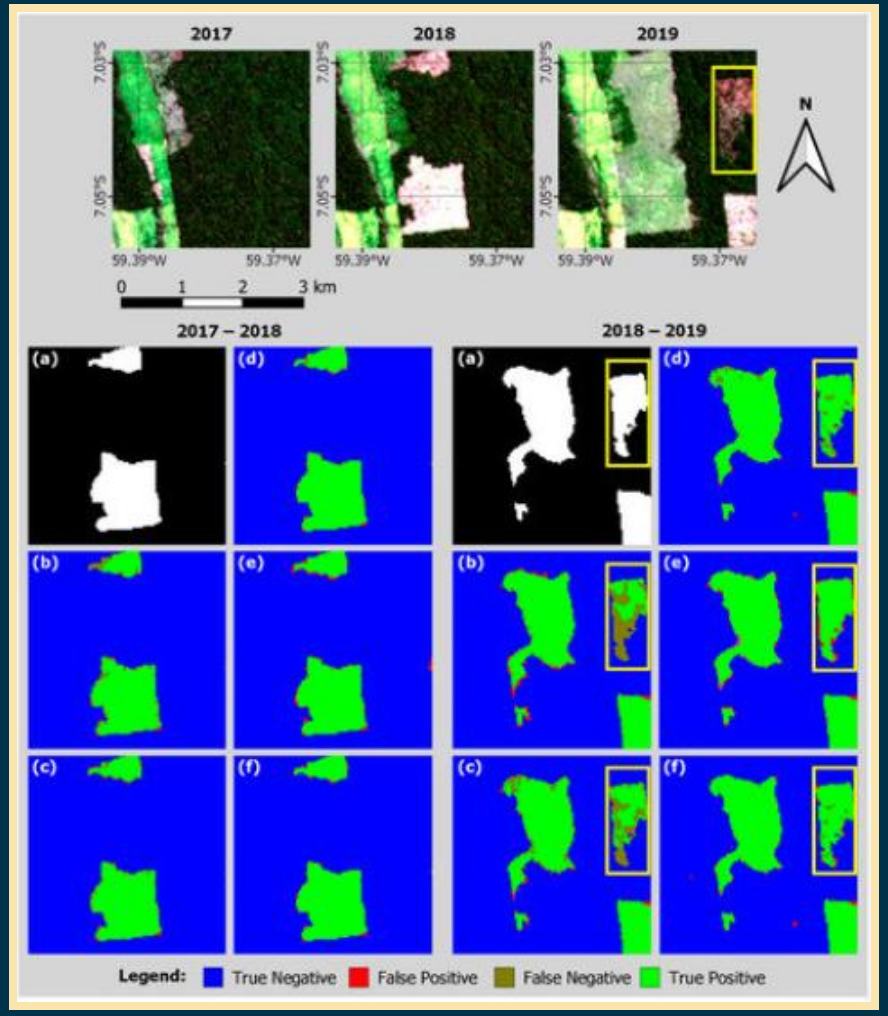
Remote Sensing

Pre-training

Zero-Shot

Monitoring Deforestation in the Amazon

Monitoring Deforestation in the Amazon

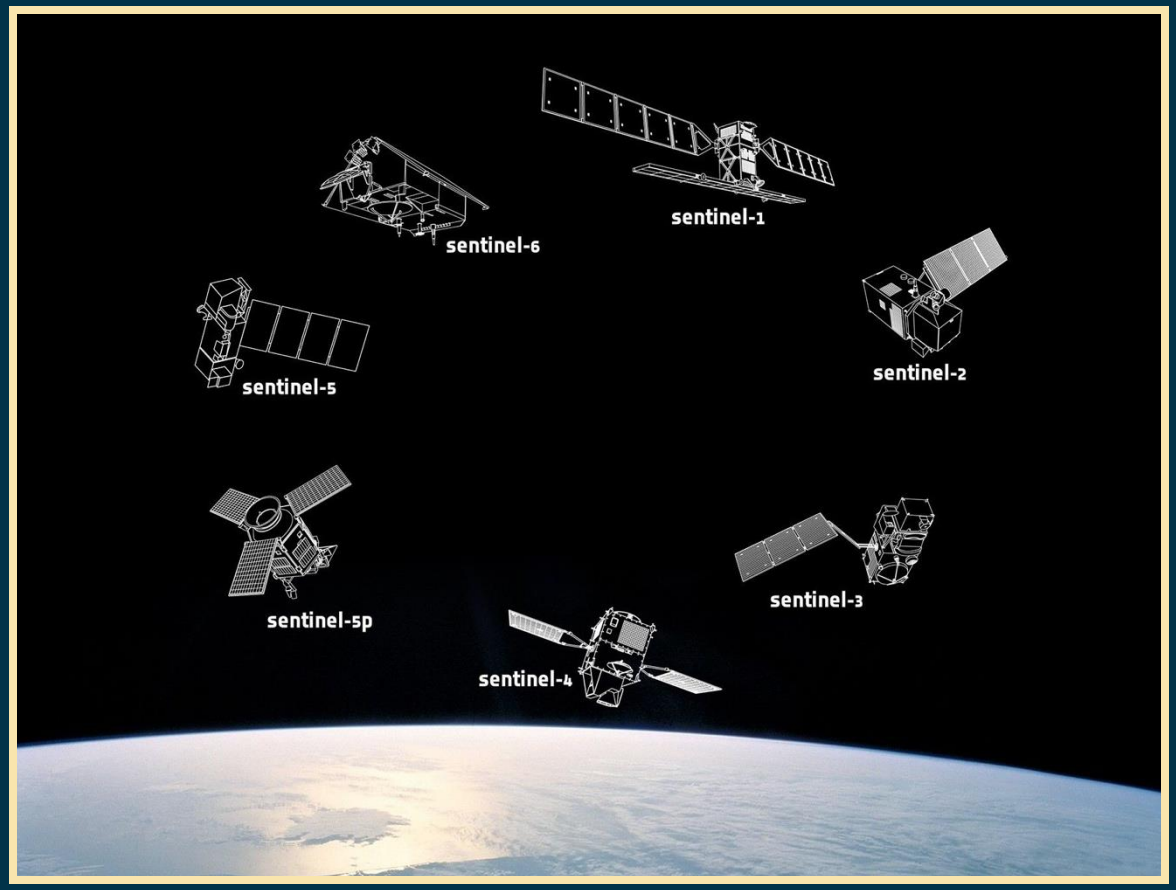


de Bem, P.P.; de Carvalho Junior, O.A.; Fontes Guimarães, R.; Trancoso Gomes, R.A. Change Detection of Deforestation in the Brazilian Amazon Using Landsat Data and Convolutional Neural Networks. *Remote Sens.* **2020**

Remote Sensing: The Sentinels

Part of Europe's *Copernicus program*

Monitor changes on the Earth's surface



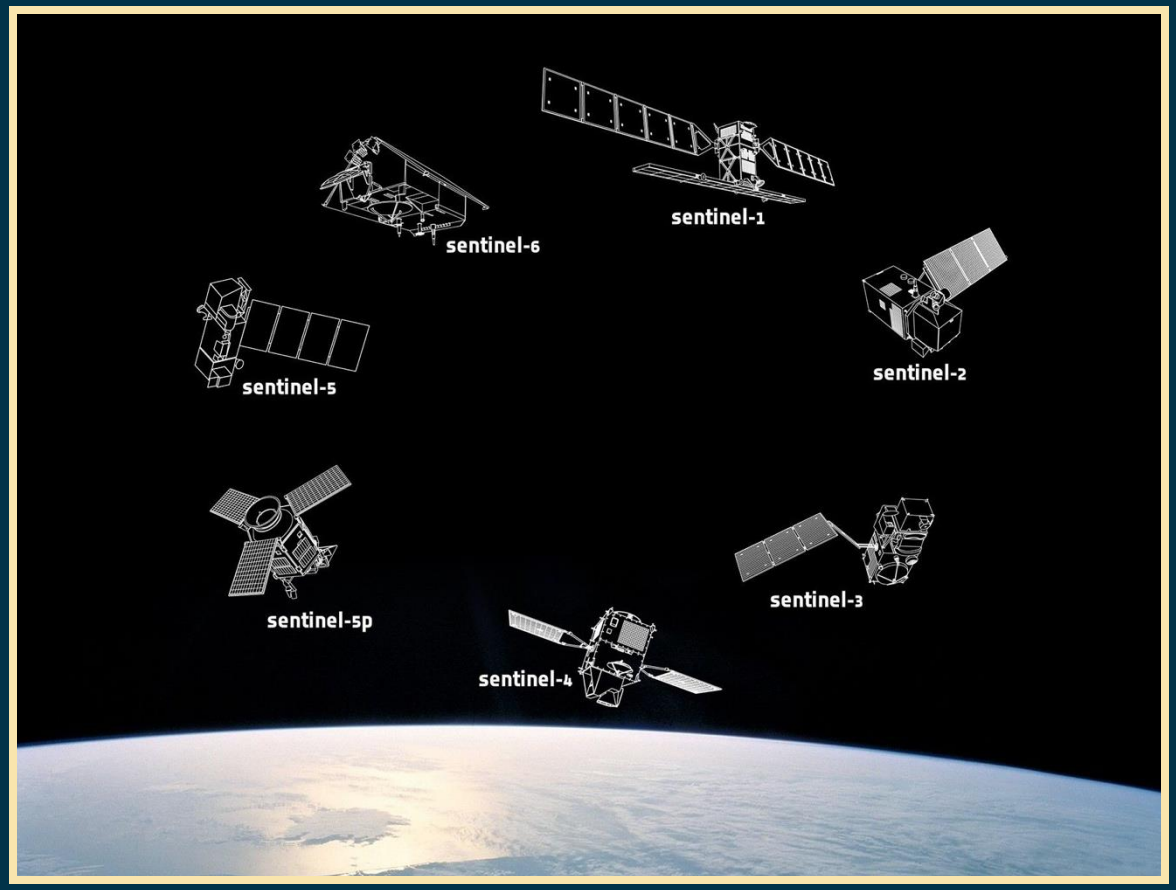
Multispectral optical (MO) imagery

Synthetic Aperture Radar (SAR) data

Remote Sensing: The Sentinels

MO imagery reveals material composition of objects (440-2200nm)

SAR reveals geometry, roughness, and electrical properties of objects (5.5cm)



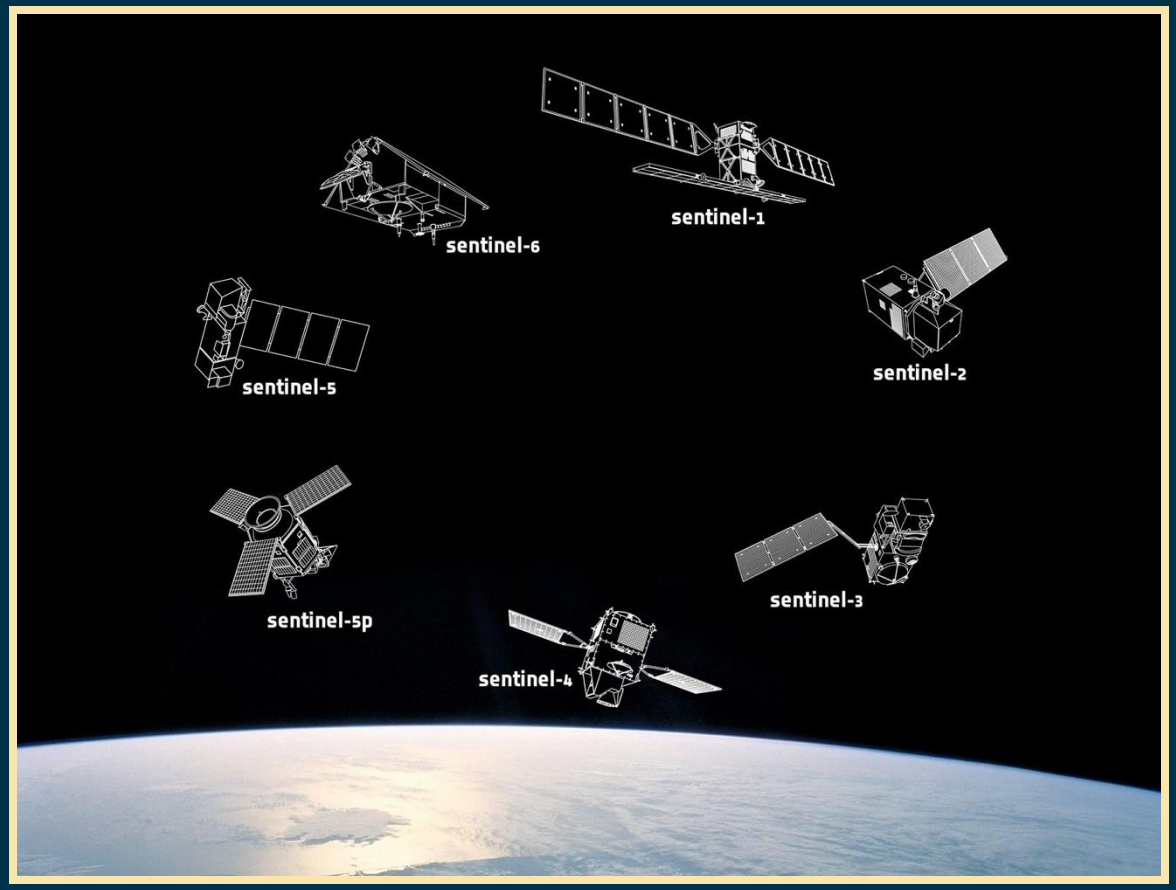
Agriculture, Climate Change, Forestry….

Cornerstone of Earth Observation

Remote Sensing: The Sentinels

Multiple views of the same features on ground

Spatially aligned

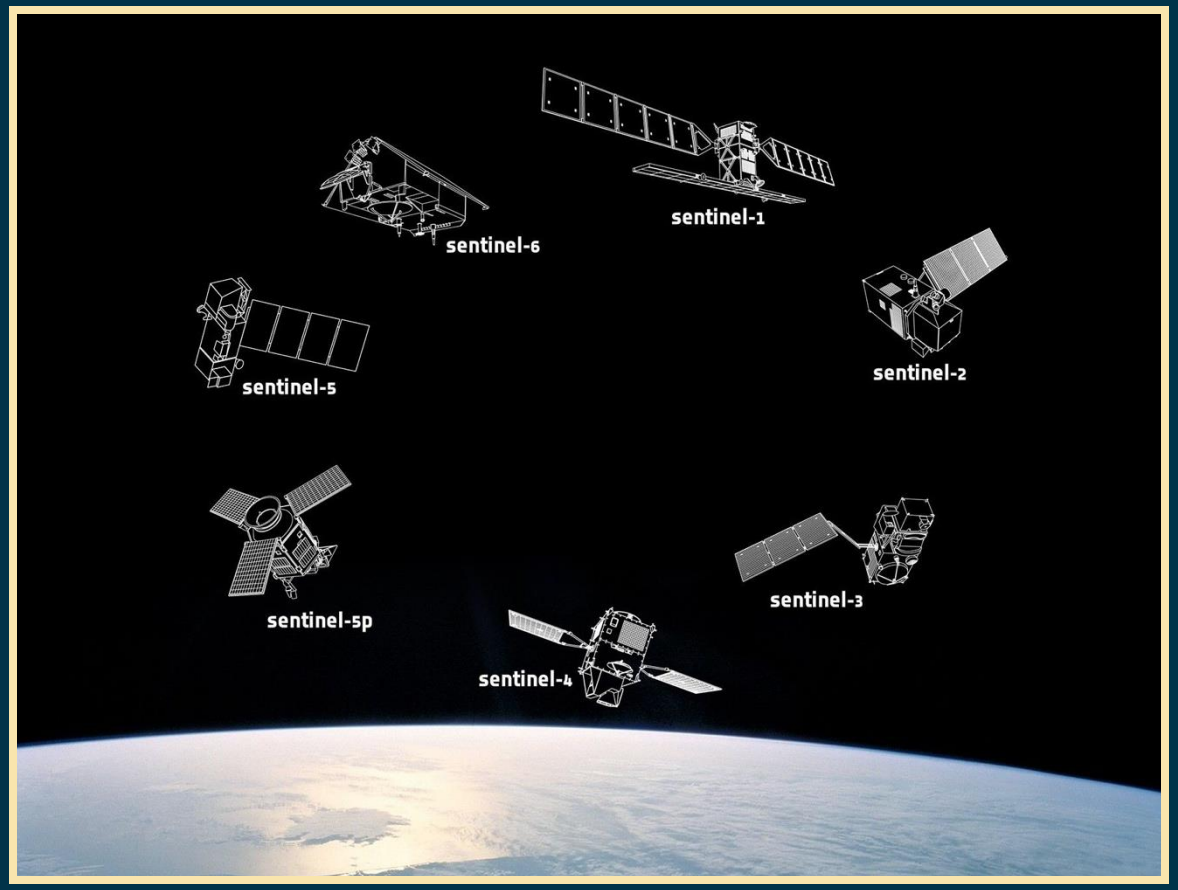


Can self-supervised learning help?

Remote Sensing: The Sentinels

Multiple views of the same features on ground

Spatially aligned



Can self-supervised learning help?

Yes

In this talk

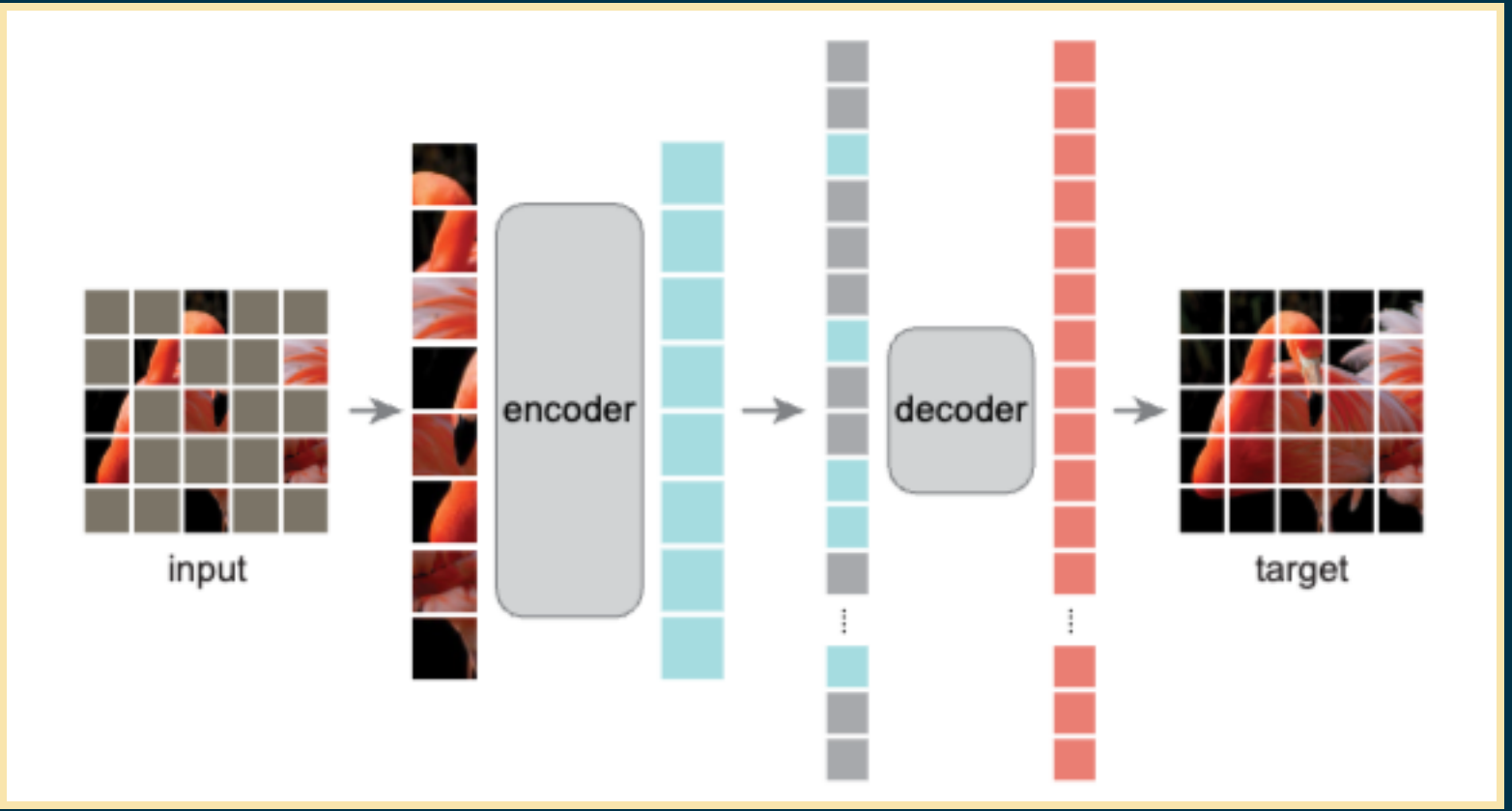
Domain Challenges

Remote Sensing

Pre-training

Zero-Shot

The Masked Autoencoder



He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

MAE in Remote Sensing?

Good starting point

MAE in Remote Sensing?

Good starting point

BUT

MAE in Remote Sensing?

Good starting point

BUT

Does *not* leverage scale

Absolute/Relative positional encoding

MAE in Remote Sensing?

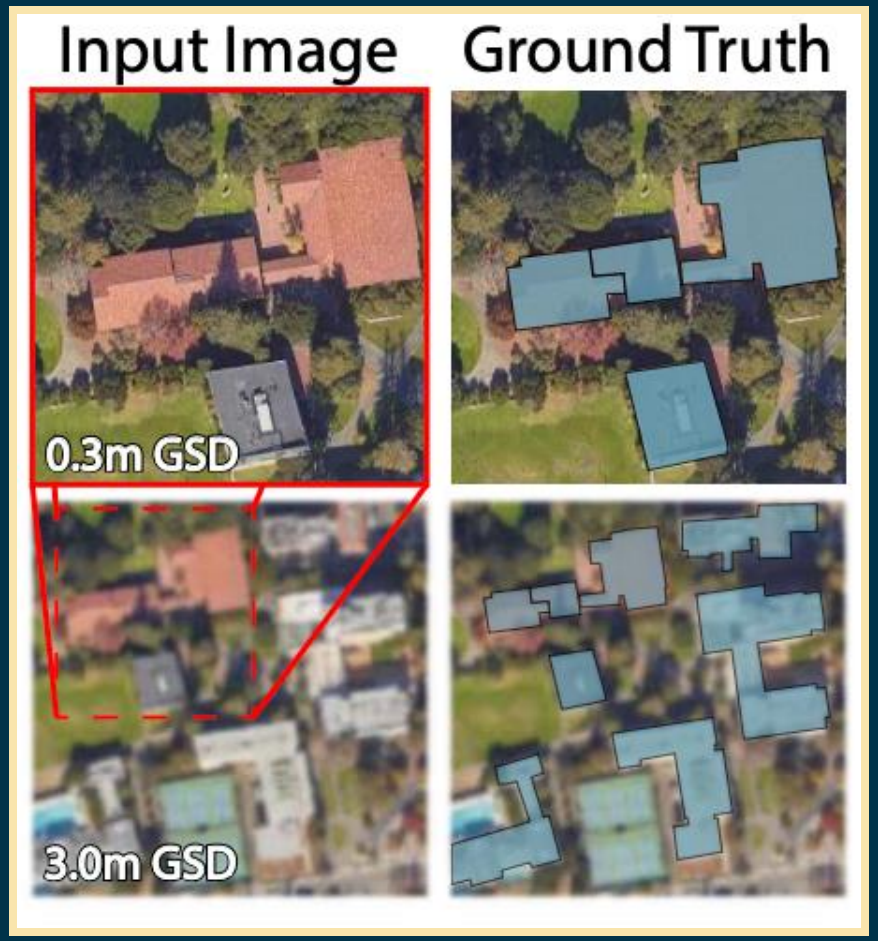
Good starting point

BUT

Does *not* leverage scale

Absolute/Relative positional encoding

MAE in Remote Sensing?



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

MAE in Remote Sensing?

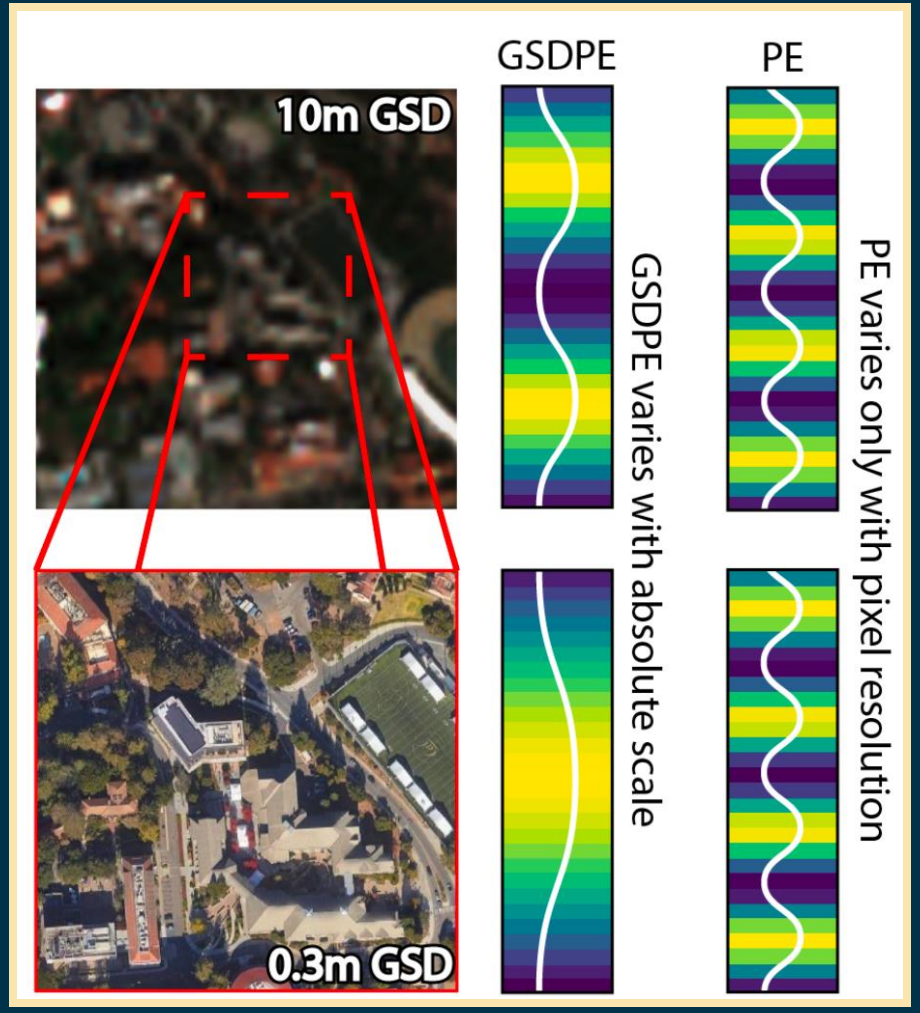
Good starting point

BUT

Does *not* leverage scale

Absolute/Relative positional encoding

MAE in Remote Sensing?



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

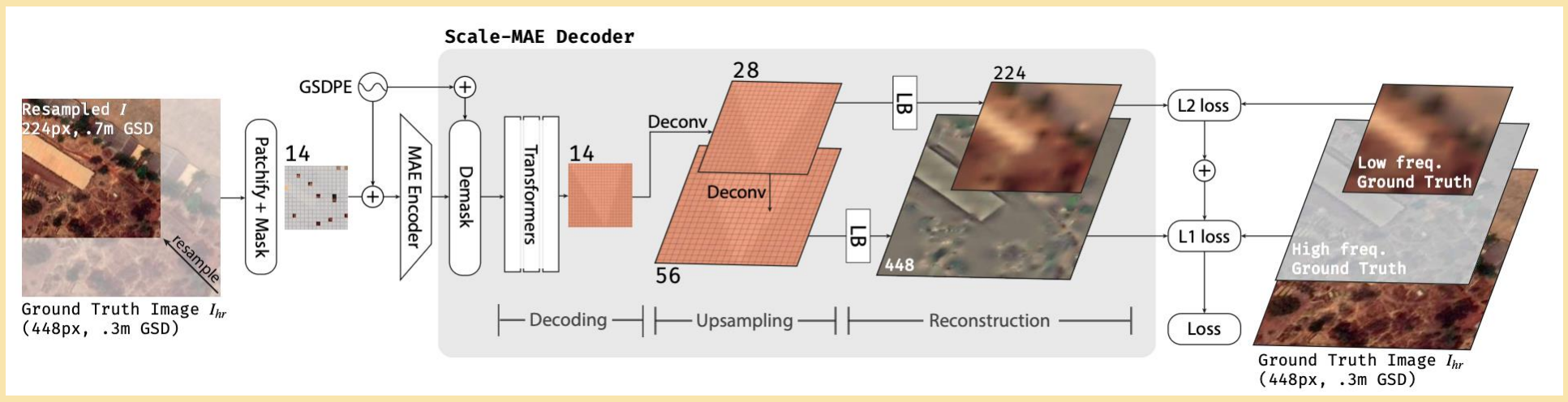
MAE in Remote Sensing?

In essence, we need a decoder that has:

(i) Scale-aware positional encoding
(ii) Multi-resolution sampling

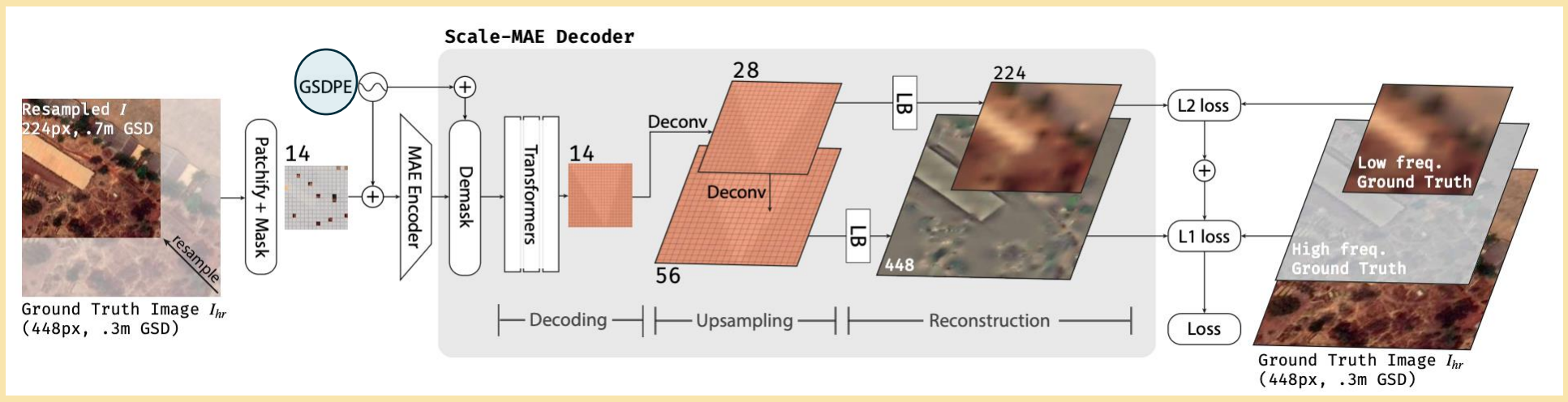
Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing

$$v_x(pos, 2i) = \sin \frac{pos}{10000^{\frac{2i}{D}}}$$
$$v_y(pos, 2i+1) = \cos \frac{pos}{10000^{\frac{2i}{D}}}$$

Before

Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing

$$v_x(pos, 2i) = \sin \frac{pos}{10000^{\frac{2i}{D}}}$$
$$v_y(pos, 2i+1) = \cos \frac{pos}{10000^{\frac{2i}{D}}}$$

Before

$$v_{gsd,x}(pos, 2i) = \sin \frac{g}{G} \frac{pos}{10000^{\frac{2i}{D}}}$$
$$v_{gsd,y}(pos, 2i+1) = \cos \frac{g}{G} \frac{pos}{10000^{\frac{2i}{D}}}$$

Now

Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing

$$v_x(pos, 2i) = \sin \frac{pos}{10000^{\frac{2i}{D}}}$$
$$v_y(pos, 2i+1) = \cos \frac{pos}{10000^{\frac{2i}{D}}}$$

Before

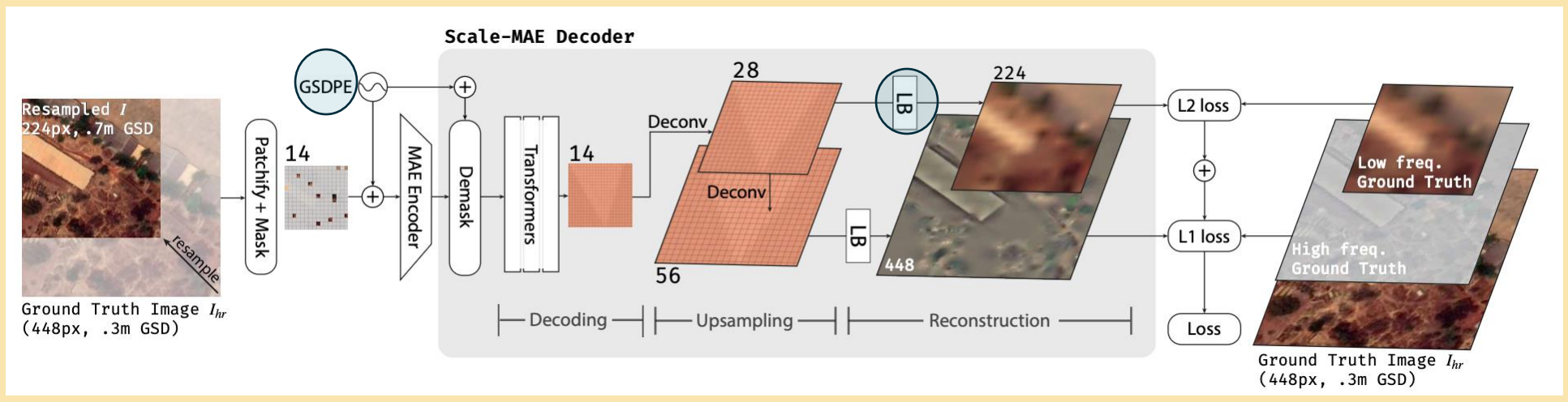
$$v_{gsd,x}(pos, 2i) = \sin \frac{g}{G} \frac{pos}{10000^{\frac{2i}{D}}}$$
$$v_{gsd,y}(pos, 2i+1) = \cos \frac{g}{G} \frac{pos}{10000^{\frac{2i}{D}}}$$

Now

An object at a finer resolution has more pixels representing it
The same object at a coarser resolution must then map to fewer pixels

Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing

The Laplacian Block (LB) helps sample at a specified frequency (resolution)

Reconstruct low res images for low frequencies, high res images for high frequencies

Final image reconstruction combines these two resolutions

Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing

The Laplacian Block (LB) helps sample at a specified frequency (resolution)

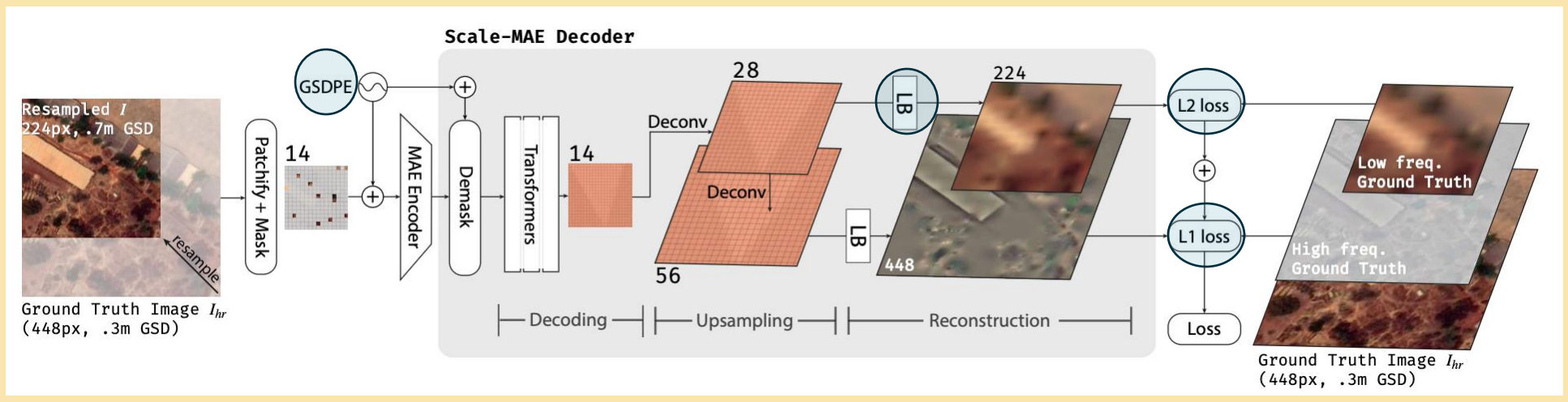
Reconstruct low res images for low frequencies, high res images for high frequencies

Final image reconstruction combines these two resolutions

Multi-scale reconstruction like this leads to the term Laplacian *pyramid*

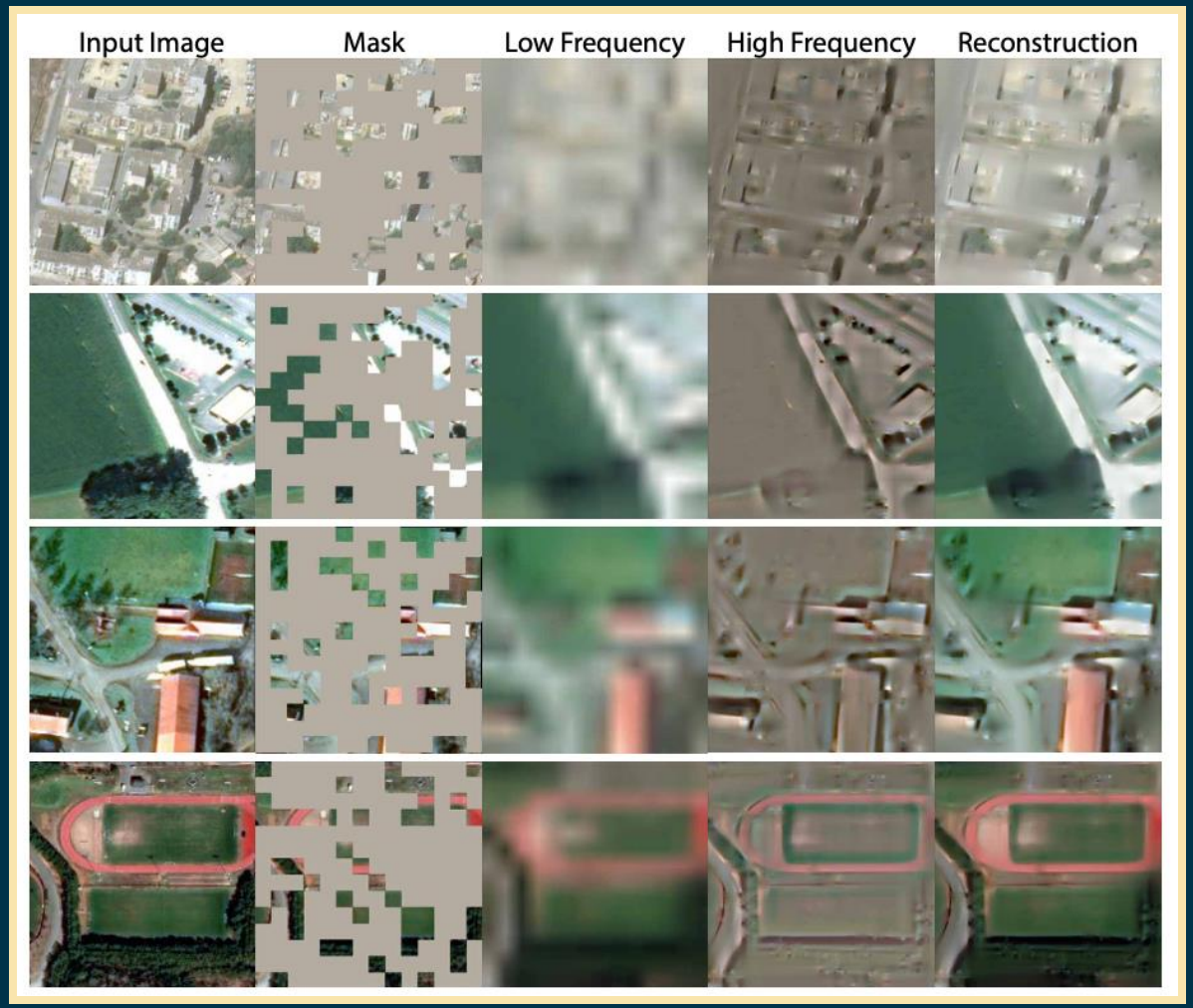
Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ScaleMAE for Remote Sensing



Reed, Colorado J., et al. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

In this talk

Domain Challenges

Remote Sensing

Pre-training

Zero-Shot

Contrastive ideas as well?

Masked reconstruction algorithms are great:

(i) Scalable
(ii) No need for paired data

BUT

Contrastive ideas as well?

Masked reconstruction algorithms are great:

(i) Scalable
(ii) No need for paired data

BUT

Significant fine-tuning required

Contrastive ideas as well?

Masked reconstruction algorithms are great:

(i) Scalable
(ii) No need for paired data

BUT

Significant fine-tuning required

Contrastive algorithms are great:

(i) Rich view information
(ii) Great for downstream tasks

BUT

Contrastive ideas as well?

Masked reconstruction algorithms are great:

(i) Scalable
(ii) No need for paired data

BUT

Significant fine-tuning required

Contrastive algorithms are great:

(i) Rich view information
(ii) Great for downstream tasks

BUT

Data and compute hungry
Sensitive to paired info

Contrastive ideas as well?

Masked reconstruction algorithms are great:

(i) Scalable
(ii) No need for paired data

BUT

Significant fine-tuning required

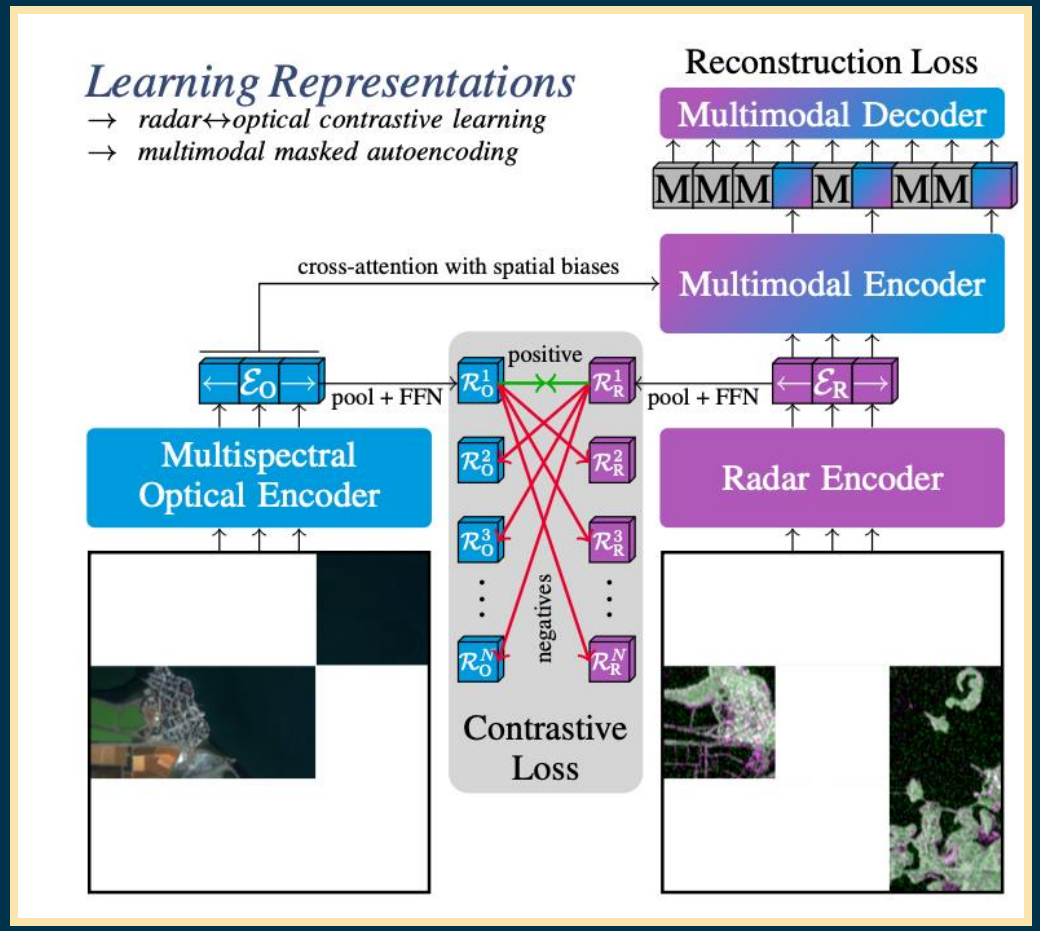
Contrastive algorithms are great:

(i) Rich view information
(ii) Great for downstream tasks

BUT

Data and compute hungry
Sensitive to paired info

Contrastive + Masked Reconstruction!

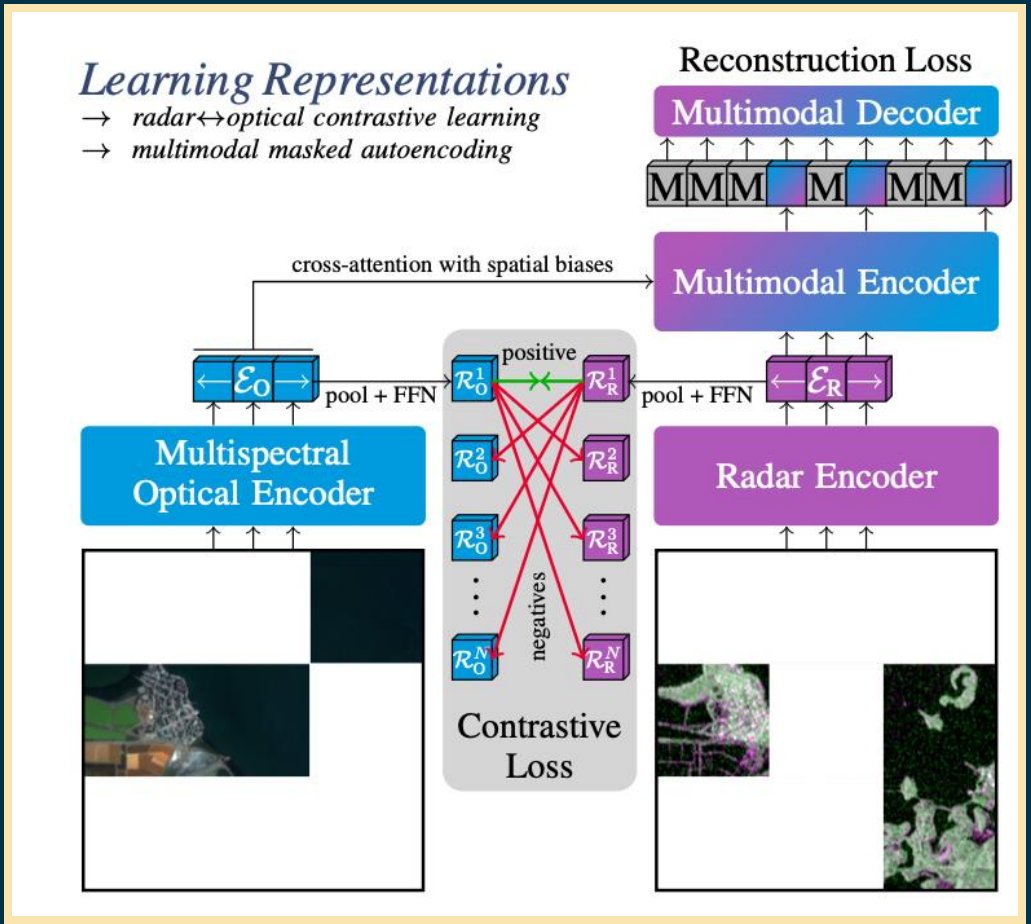


Fuller, A., Millard, K., & Green, J. (2023). CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, *36*, 5506-5538.

Contrastive + Masked Reconstruction!

Optical Data:
12-channel MO from Sentinel-2

Radar data:
2-channel backscatter from Sentinel-1

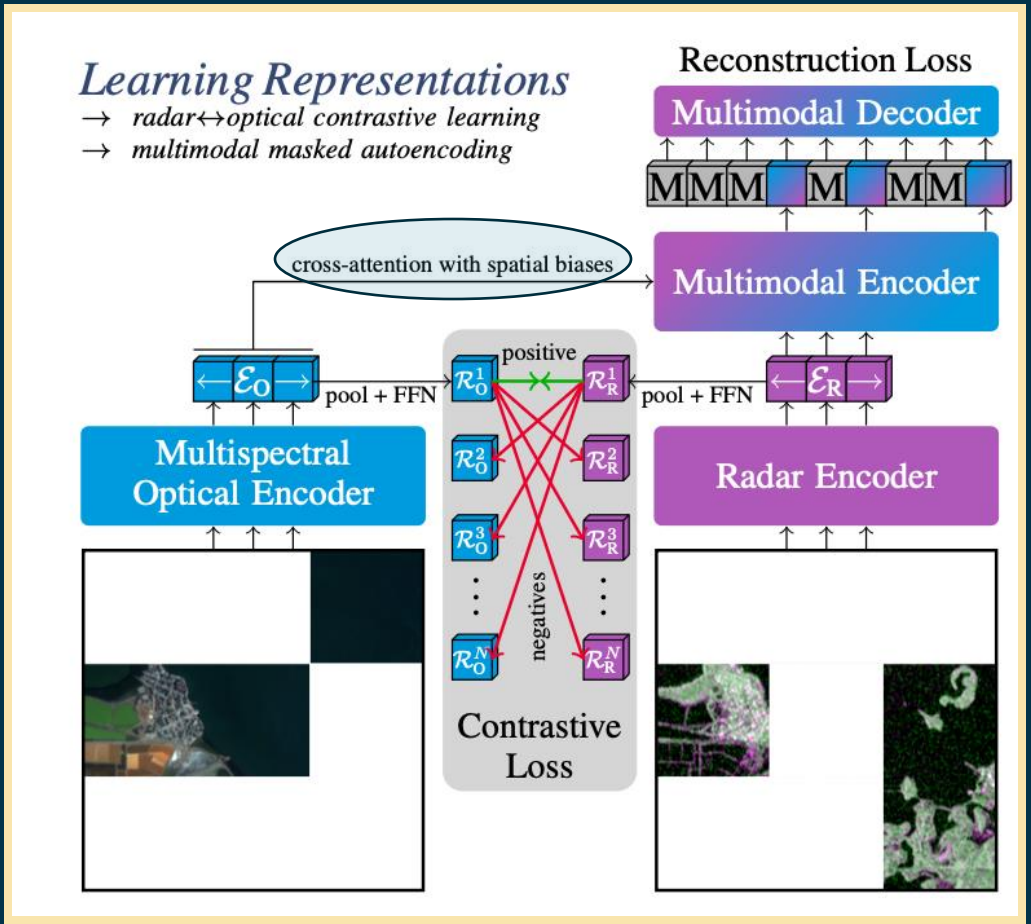


Fuller, A., Millard, K., & Green, J. (2023). CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, *36*, 5506-5538.

Contrastive + Masked Reconstruction!

Optical Data:
12-channel MO from Sentinel-2

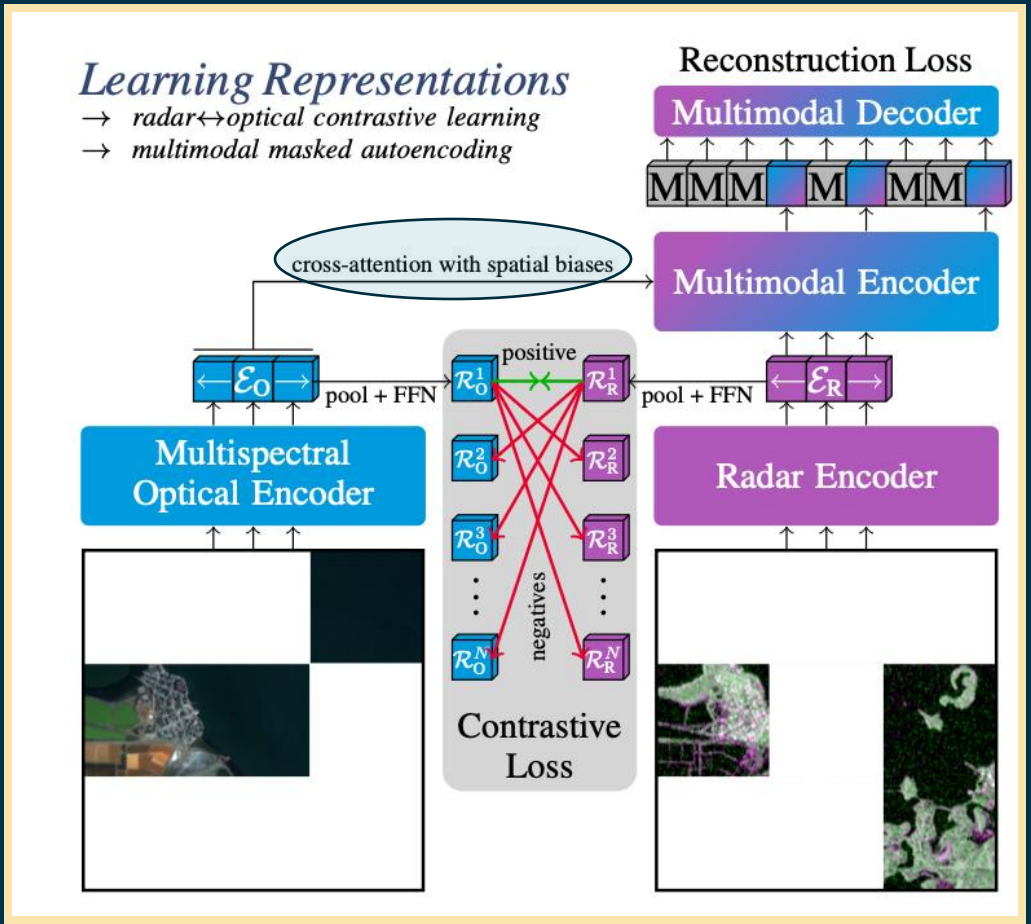
Radar data:
2-channel backscatter from Sentinel-1



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.

Contrastive + Masked Reconstruction!

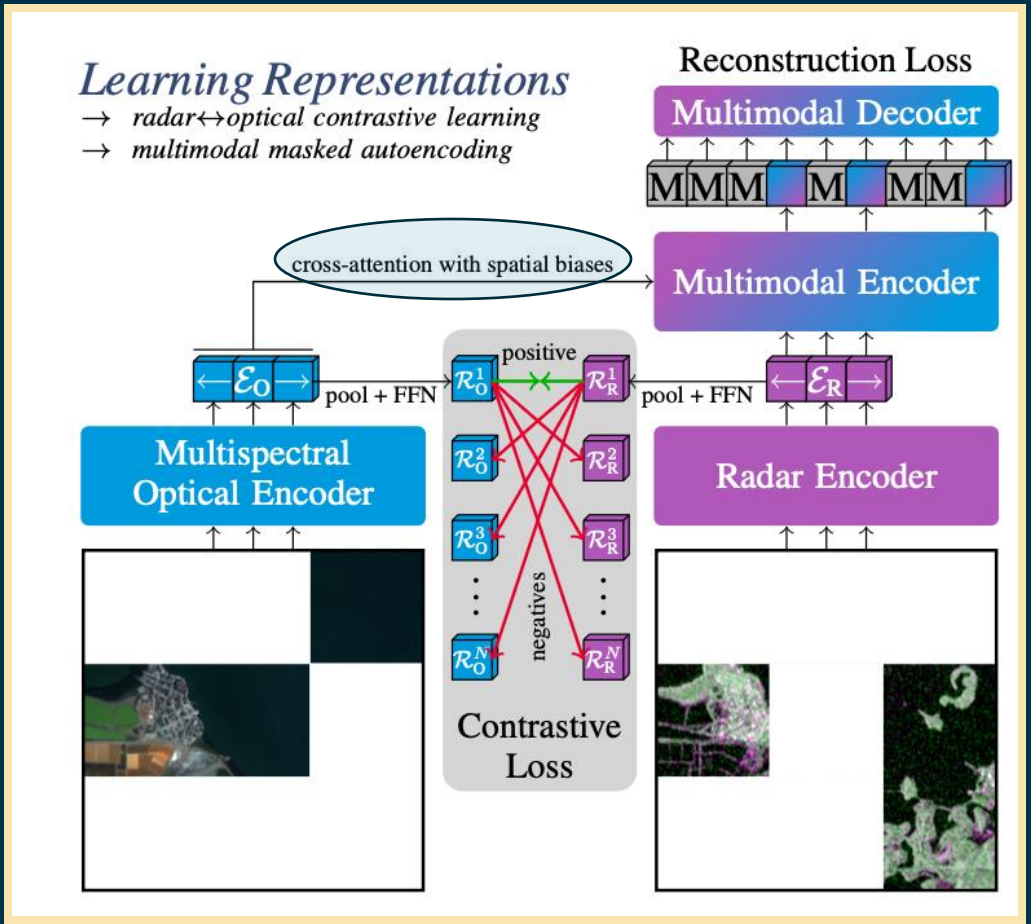
Each ViT backbone has its self-attention heads biased based on the Euclidean distance between query-key pairs



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.

Contrastive + Masked Reconstruction!

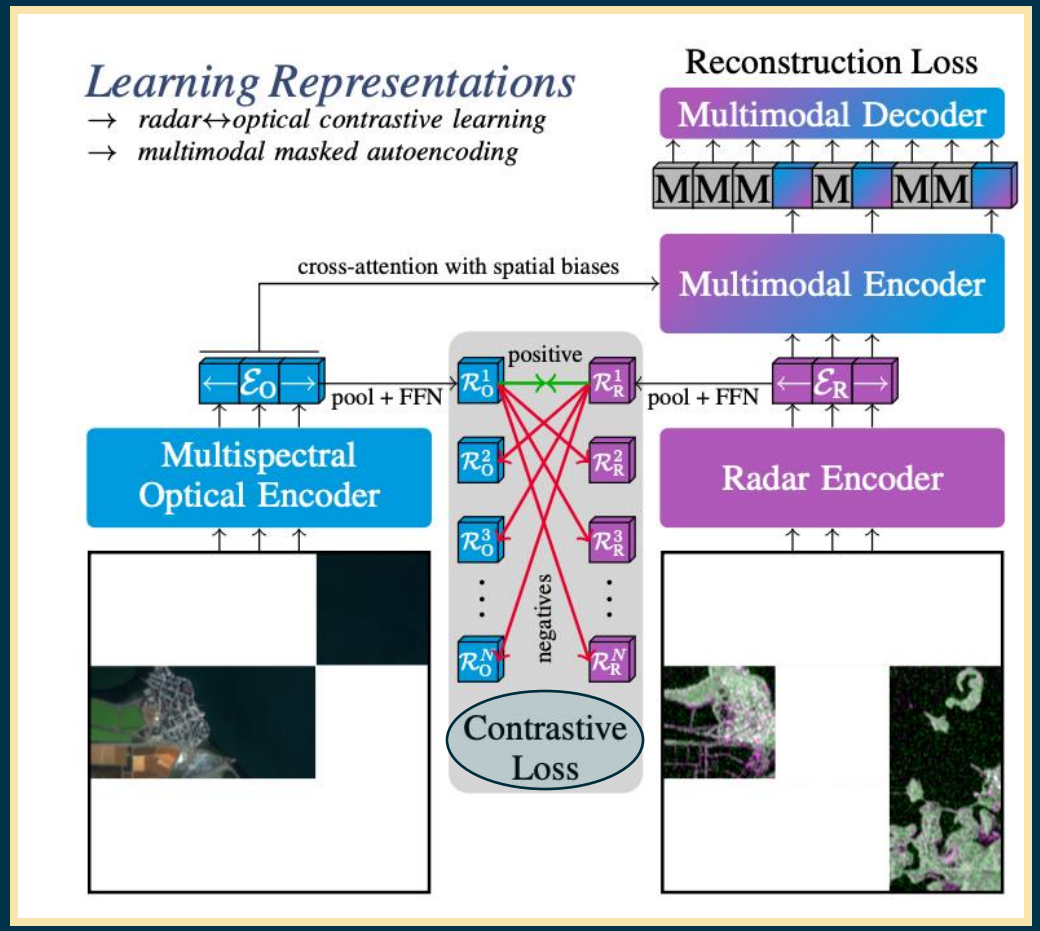
Each ViT backbone has its self-attention heads biased based on the Euclidean distance between query-key pairs



The cross-attention matrix is modified by the Euclidean distance between *cross-modal* key-query pairs

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.

Contrastive + Masked Reconstruction!

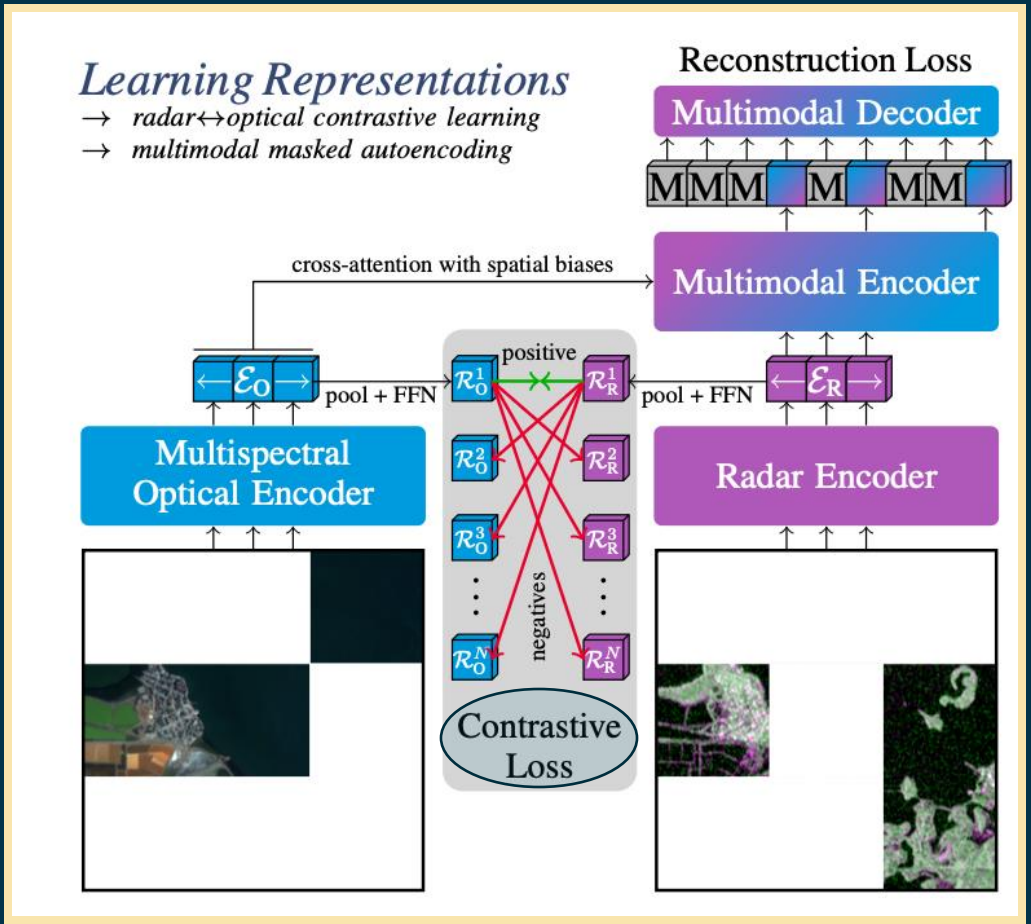


Fuller, A., Millard, K., & Green, J. (2023). CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, *36*, 5506-5538.

Contrastive + Masked Reconstruction!

For optical anchor image, positive sample is the geographic, temporal match in the radar sample

Negative sample? *All other radar samples in the batch!*

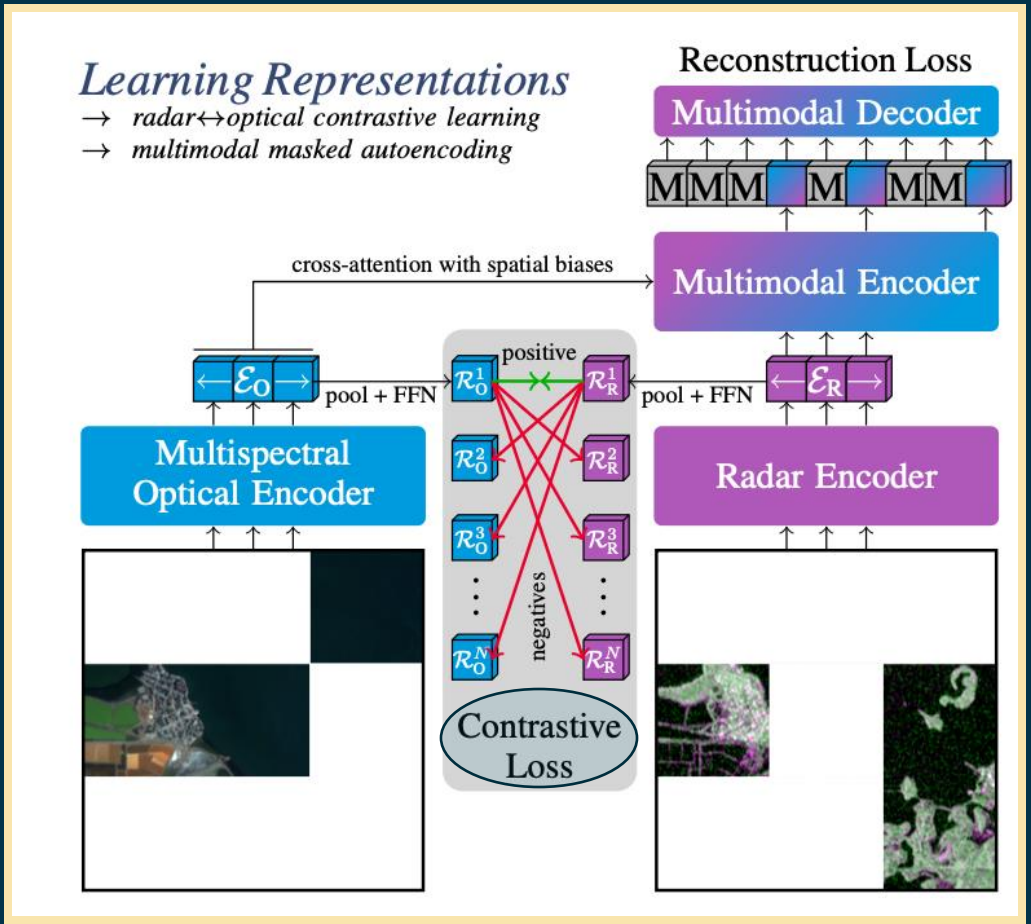


Fuller, A., Millard, K., & Green, J. (2023). CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, *36*, 5506-5538.

Contrastive + Masked Reconstruction!

For optical anchor image, positive sample is the geographic, temporal match in the radar sample

Negative sample? *All other radar samples in the batch!*

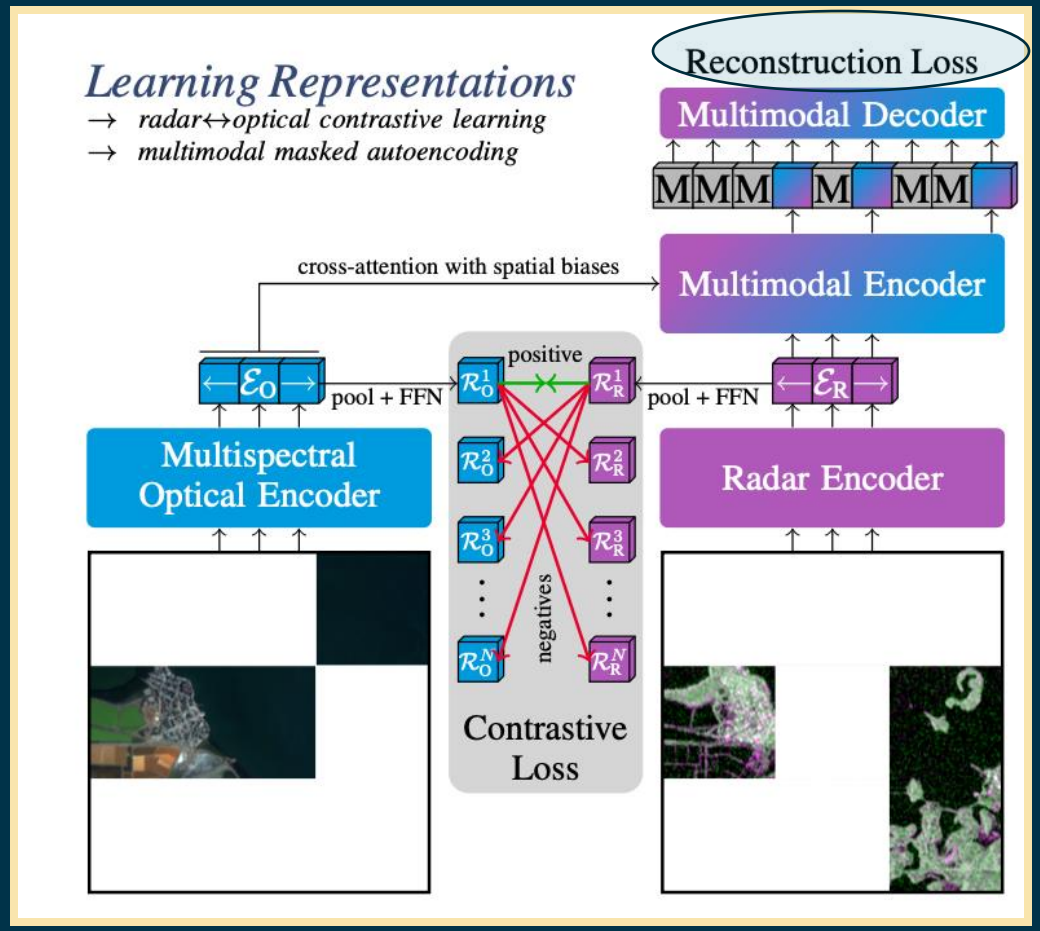


And vice versa!

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.

Contrastive + Masked Reconstruction!

Independently mask 75% of radar and optical patches



Loss term incorporates both radar and optical reconstruction distances

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.

In this talk

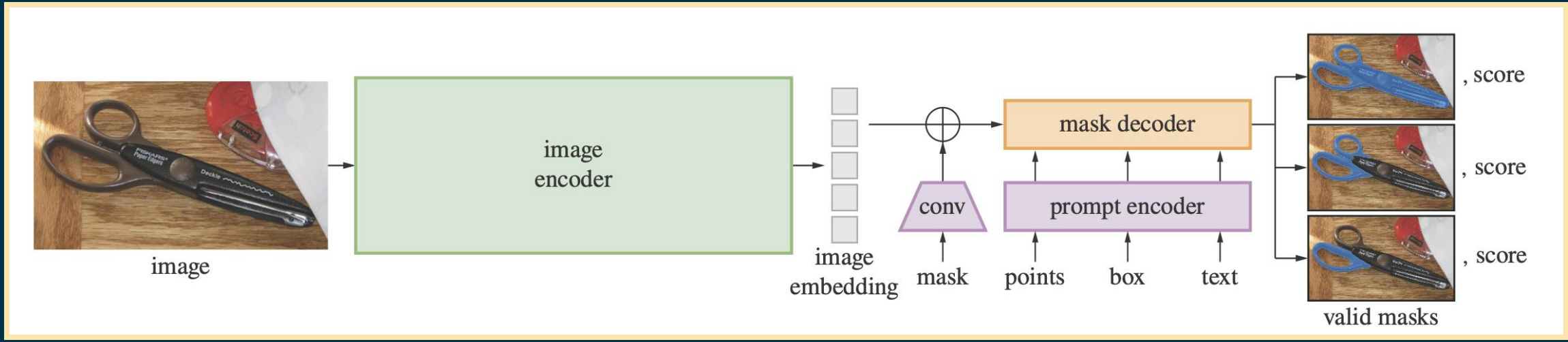
Domain Challenges

Remote Sensing

Pre-training

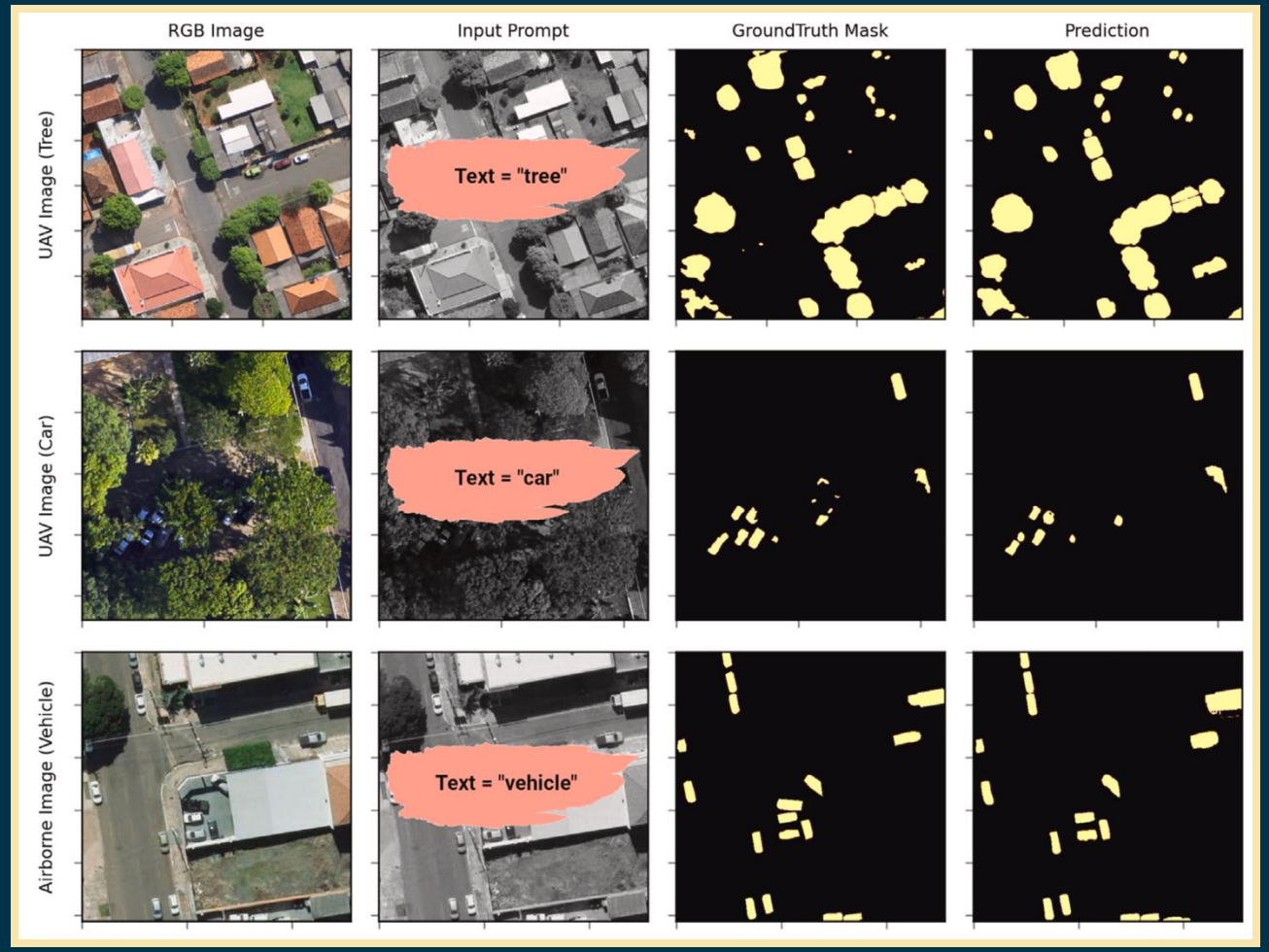
Zero-Shot

Segment Anything



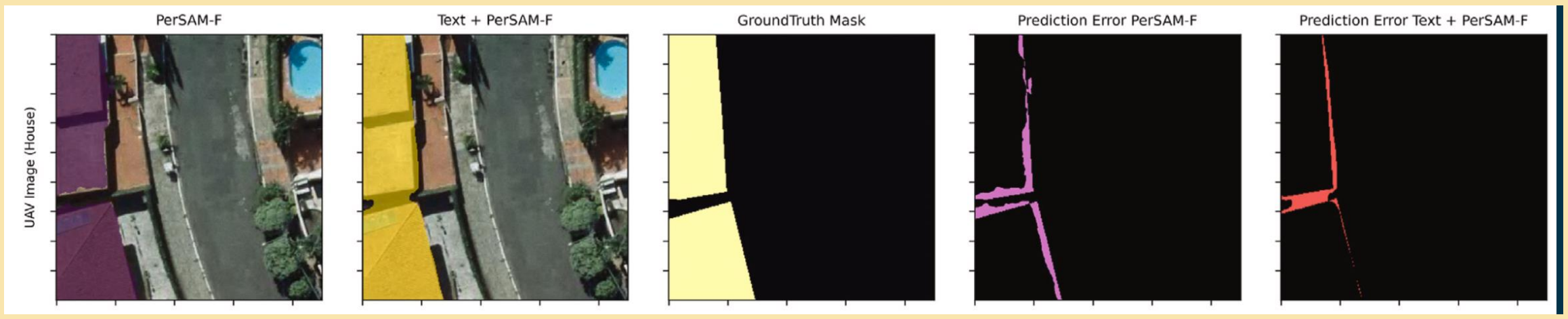
Kirillov, Alexander et al. "Segment Anything." *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023): 3992-4003.

Zero-Shot with SAM



Osco, Lucas Prado, et al. "The segment anything model (sam) for remote sensing applications: From zero to one shot." *International Journal of Applied Earth Observation and Geoinformation* 124 (2023): 103540.

One-Shot with SAM



Osco, Lucas Prado, et al. "The segment anything model (sam) for remote sensing applications: From zero to one shot." *International Journal of Applied Earth Observation and Geoinformation* 124 (2023): 103540.

Caution!

Concepts are still broad – special cases may be challenging

Tends to overestimate object boundaries – may be a prompt issue

Summary

Domain Challenges

Remote Sensing

Pre-training

Zero-Shot