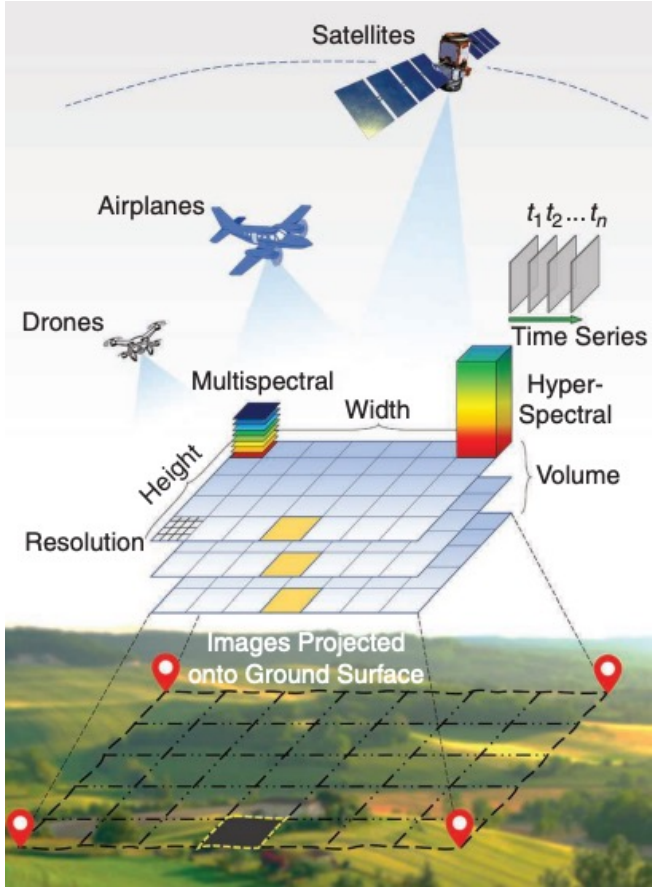
Foundation Models for Earth Observation

Philipe Dias, Ph.D.
GeoAI Group
Geospatial Science and Human Security Division

Aristeidis Tsaris (CCSD), Dalton Lunga (GeoAI), Abhishek Potnis, Jacob Arndt, Jordan Bowman

Earth Observation

- Gathering of information about the physical, chemical, and biological systems of the planet Earth
- **Remote-sensing technologies**, direct-contact sensors in ground-based, airborne platforms
- Applications: human dynamics, precision agriculture, disaster management, humanitarian assistance, national security



Schmitt, Michael, et al. *"There are no data like more data: Datasets for deep learning in earth observation."* IEEE Geoscience and Remote Sensing Magazine (2023).

GeoAI @ ORNL

GeoAI

- *Spatial explicit AI models*
- *Infusing spatial temporal reasoning into AI models*

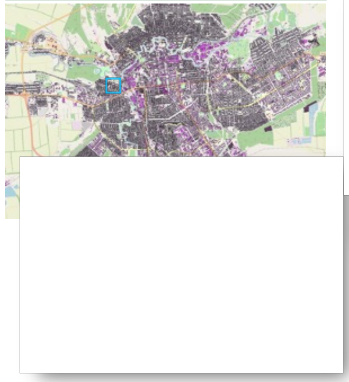
Capabilities developed over the years

- Mapping physical and built environments
- Disaster impacts analysis
- Help population distribution mapping
- Assess urban growth
- HPC-enabled mapping at large-scale

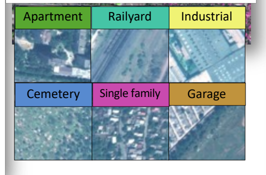
Examples of capabilities & applications

Characterization of built environment: model population, human dynamics

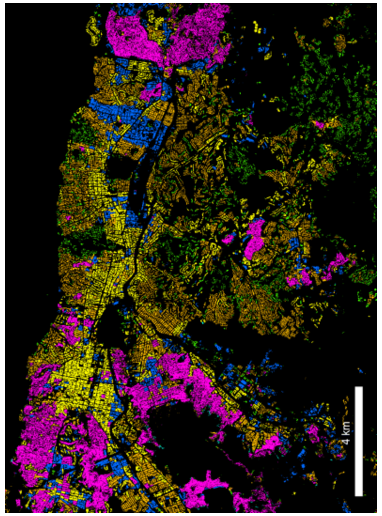
Building footprint extraction



Land use attribution

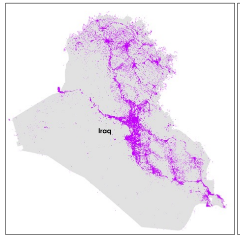


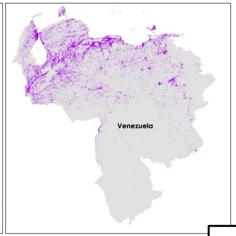
Road segmentation



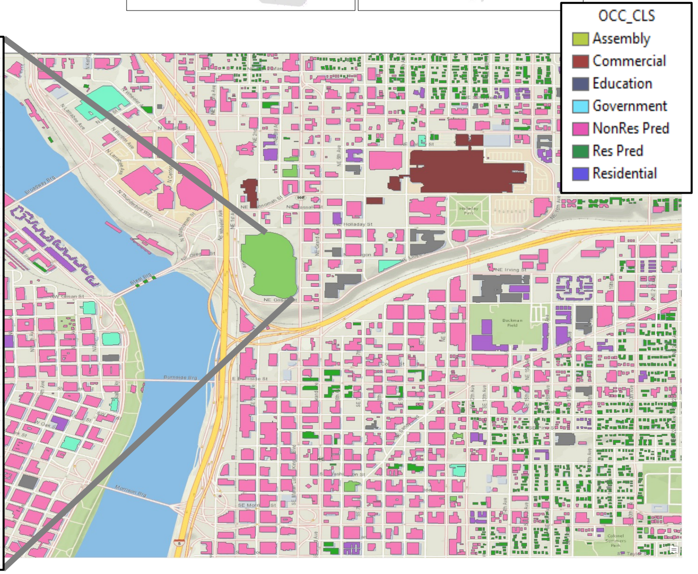
GeoAI example: Detected building footprints with socio-economic neighborhood delineations Image Credit: ORNL

From local to country-scale





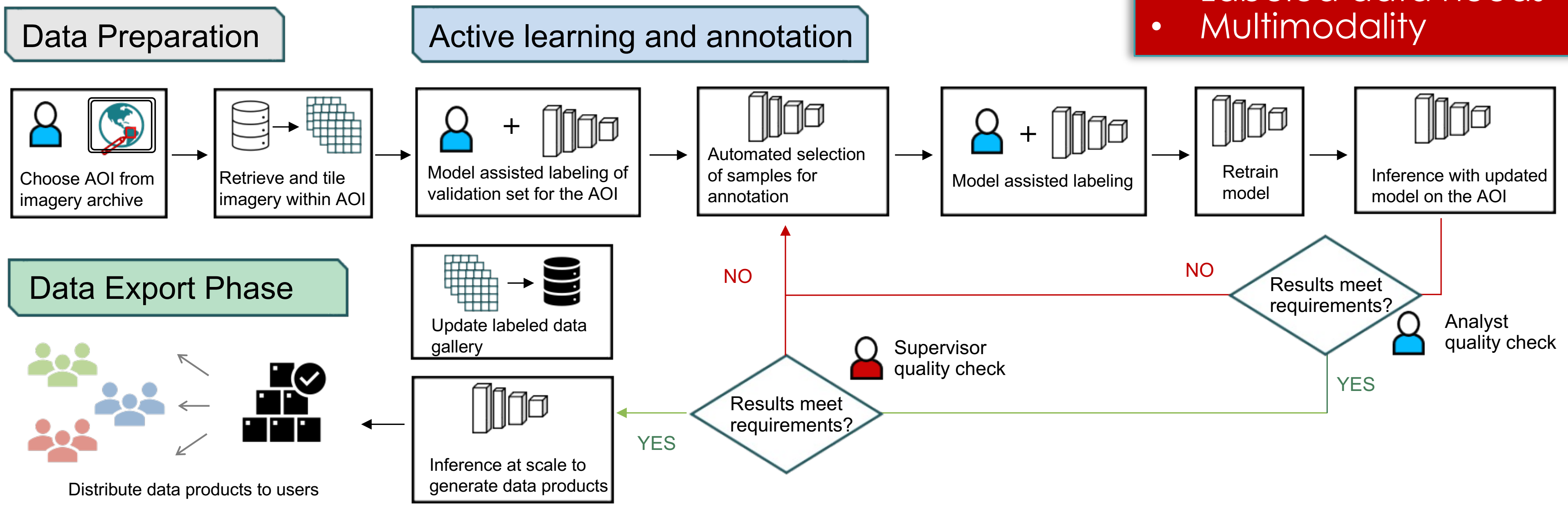
SHAPE	Polygon
BUILD_ID	142816
OCC_CLS	Assembly
PRIM_OCC	Convention Centers
SEC_OCC	<null>
PROP_ADDR	777 NE Martin Luther King Jr Blvd
PROP_CITY	Portland
PROP_ST	Oregon
PROP_ZIP	97232
OUTBLDG	<null>
HEIGHT	14.15
SQMETERS	53684.48
SQFEET	577852.6
H_ADJ_ELEV	<null>
L_ADJ_ELEV	<null>
FIPS	41051
CENSUSCODE	41051002303
PROD_DATE	9/21/2010
SOURCE	NGA
USNG	10T ER 26305 41699
LONGITUDE	-122.663135
LATITUDE	45.528332
IMAGE_NAME	7644016274442
IMAGE_DATE	9/21/2010
VAL_METHOD	Unverified
REMARKS	<null>
OCC_SOURCE	HIFLD Convention Centers
SHAPE_Length	0.010606
SHAPE_Area	0.000006



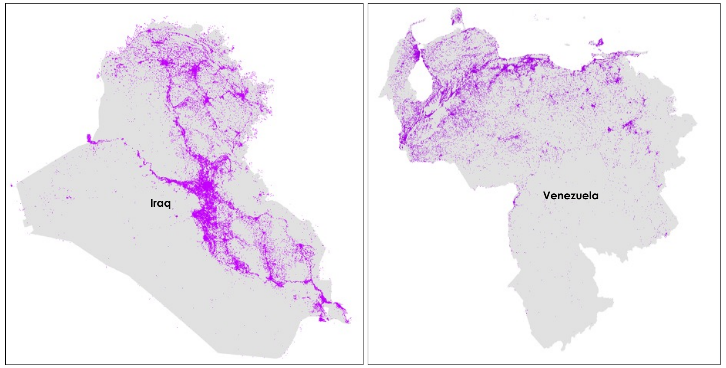
Applications at large-scale

Challenges

- Task-specific models
- Labeled data needs
- Multimodality

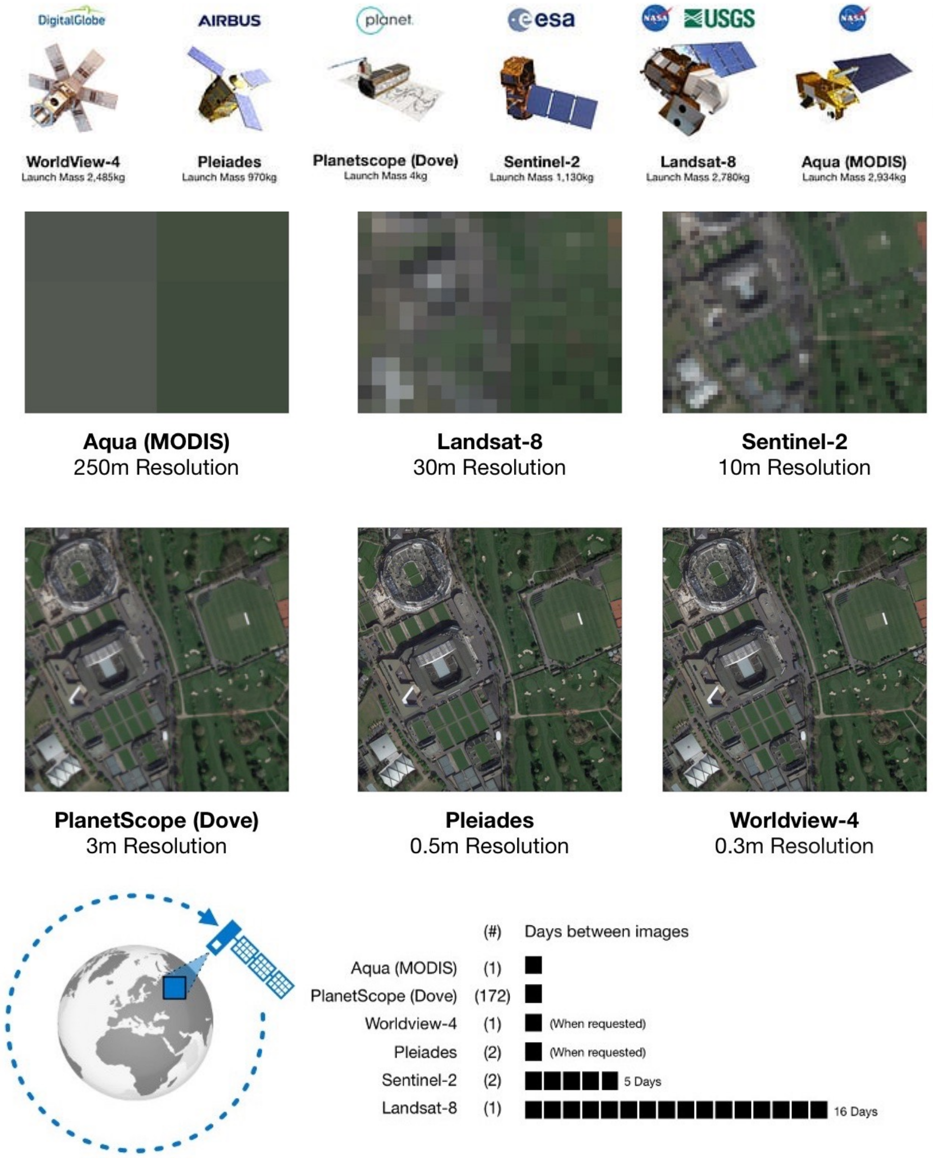


From local to country-scale

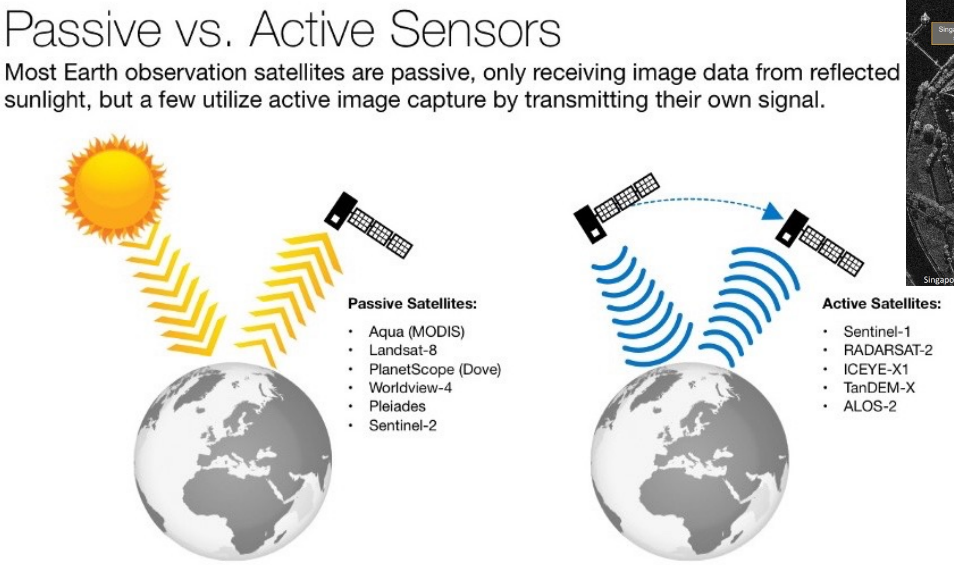


Dias, P., Arndt, J., Bowman, J., Myers, A., Yang, L., and Lunga, D. "*Human-Machine Collaboration for Reusable and Scalable Models in Remote Sensing Imagery Analysis*". Presented at the ICML 2022 Workshop on Human-Machine Collaboration and Teaming

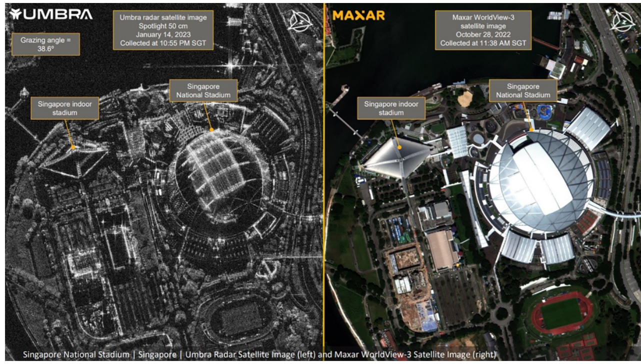
The multiple modalities in EO

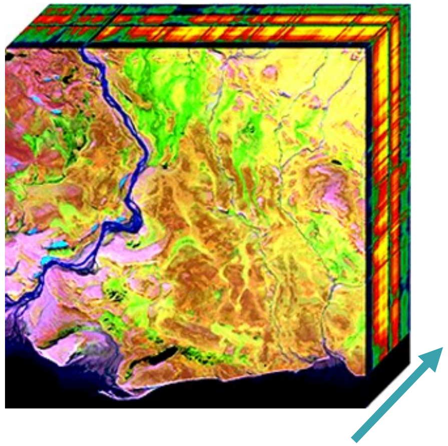


Radiant.Earth
Earth Imagery for Impact



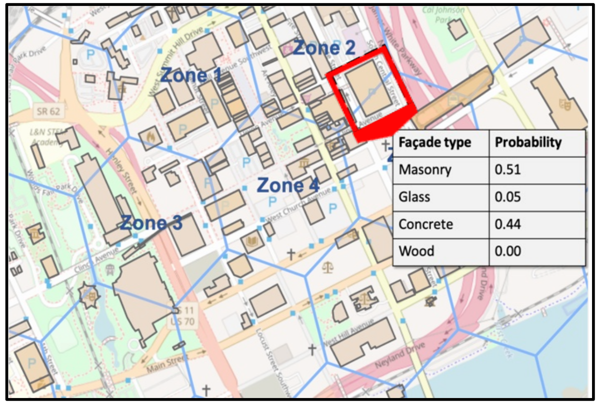
https://breakingdefense.com/2023/02/maxar-contracts-startup-umbra-to-supply-sar-satellite-data/





Spectral bands

SpaceNet 4
https://www.cosmiqworks.org/archived-projects/spacenet-4/



Façade type	Probability
Masonry	0.51
Glass	0.05
Concrete	0.44
Wood	0.00

Bayesian Modeling: Estimation of *probable* material types for each building in Knoxville, TN. Image Credit: ORNL

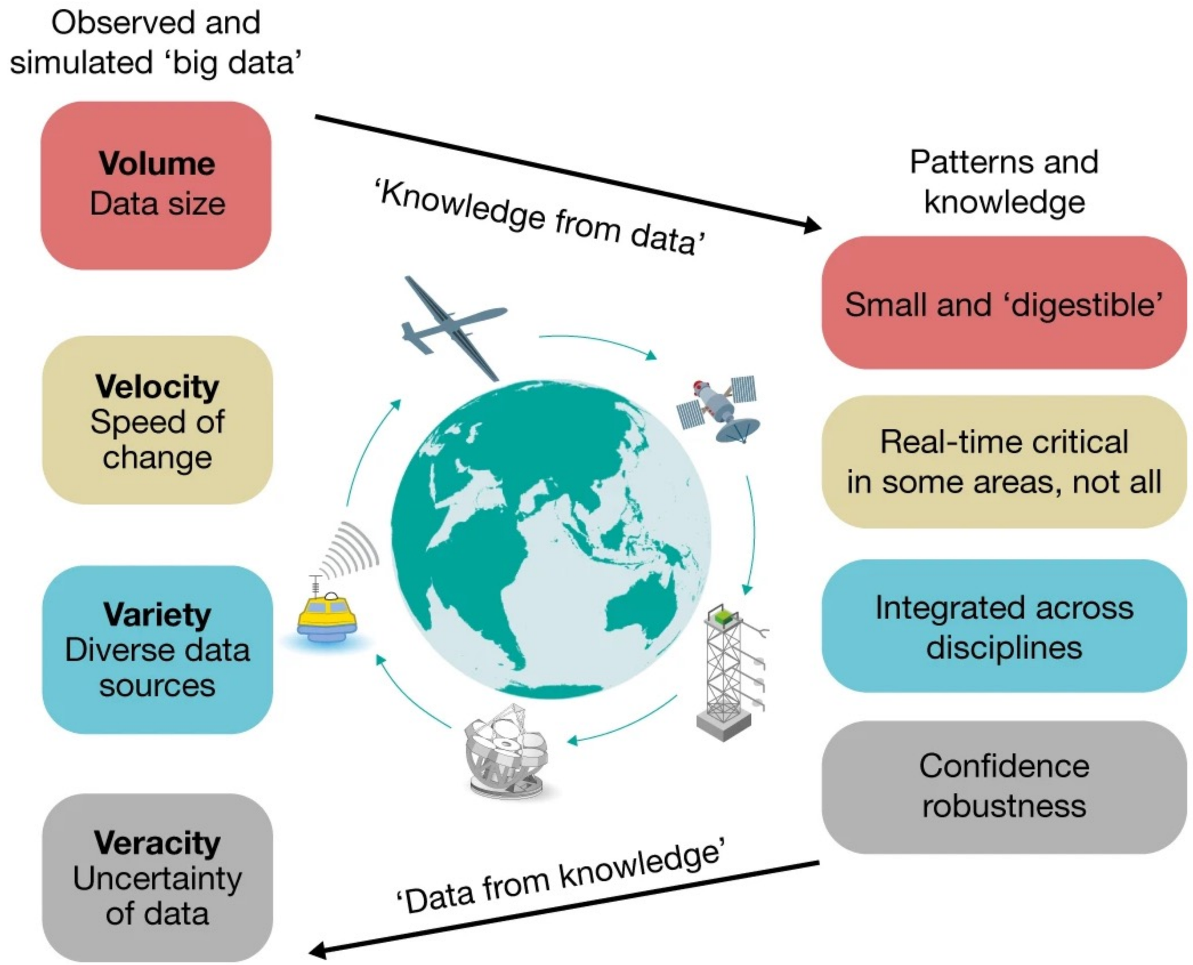
Characteristics/challenges of Earth Observation data

Data volumes

- Current EO satellite constellations: 100+TBs of data/day
- Images can be billions of pixels large (e.g., 30,000 x 30,000 x 4)
  - At a modest 5m resolution:
    Earth's surface = 100 trillion pixels
  - Nigeria at ~0.5 m resolution
    - 20,000 Individual scenes, 90TB
- Data management, training/inference challenges
- But great potential for applications & large models!

Fig. 1: Big data challenges in the geoscientific context.

From: Deep learning and process understanding for data-driven Earth system science



Data size now exceeds 100 petabytes, and is growing quasi-exponentially (tapering of the figure to the right indicates decreasing data size.) The speed of change exceeds 5 petabytes a year; data are taken at frequencies of up to 10 Hz or more; reprocessing and versioning are common challenges. Data sources can be one- to four-dimensional, spatially integrated, from the organ level (such as leaves) to the global level. Earth has diverse observational systems, from remote sensing to in situ observation. The uncertainty of data can stem from observational errors or conceptual inconsistencies.

Foundation models for Earth Observation

A shared backbone pretrained using self-supervised learning (SSL) that can be efficiently tuned for multiple tasks

Model scaling & Data scaling → **Emergent Abilities**

Key aspects / building blocks

Data
- Vast / rich volumes

Pretraining objectives
- Self-Supervised Learning (SSL)

Architectures
- Scalable / parallelizable

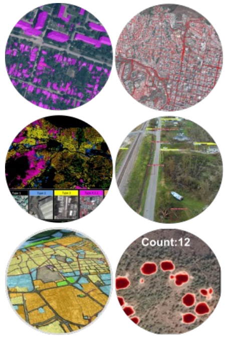
HPC resources
- Billions - Trillion of parameters → lots of FLOPS

Downstream adaptation & evaluation
- Efficient adaptation to multiple tasks
- Emergent properties

Applications

FM4EO
Foundation Model for Earth Observation

Adaptation →
- Fast
- Low need for task-specific labeled data



- Land-Use/Cover
- Built infrastructure
- Change detection
- Object detection
- Search & retrieval
- …

Rolling database of **~400k+ high-resolution satellite images**, **~3 PB of data**

HPC resources

Quetzal Foundation Model(s)

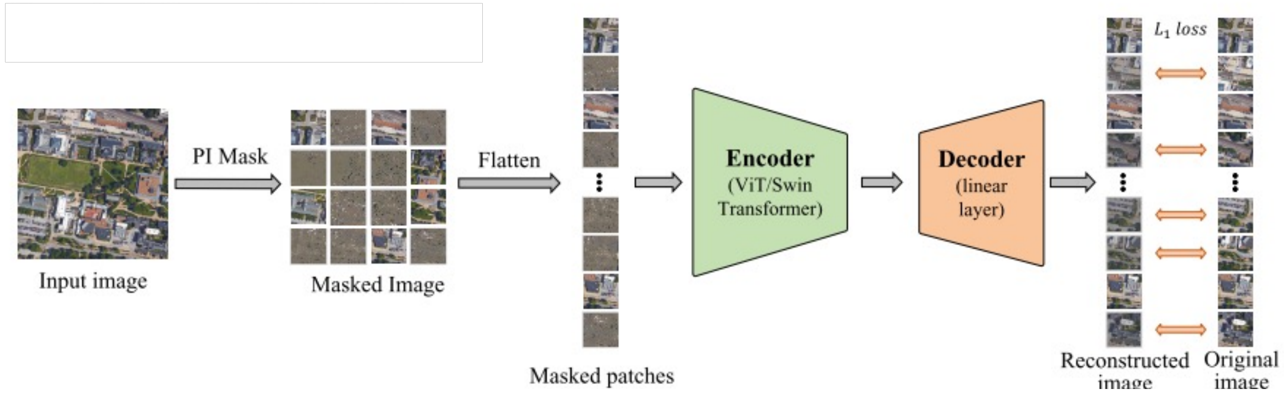
Why Quetzal?

Resplendent Quetzal by Phoo Chan, Shutterstock

- *Mesoamerican cultures: messenger between Earth and heavens/sky*
- *Multiple colors: metaphor for multiple modalities*

- Quetzal-HR: High-resolution (HR)
  - Optical imagery (RGB+NIR)

1. Pretraining using SSL



2. Finetune task-specific decoders

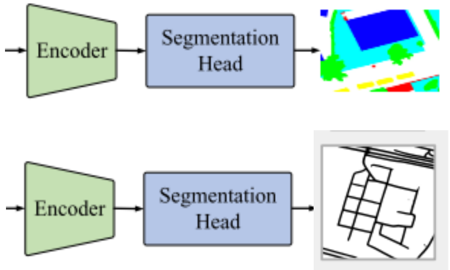
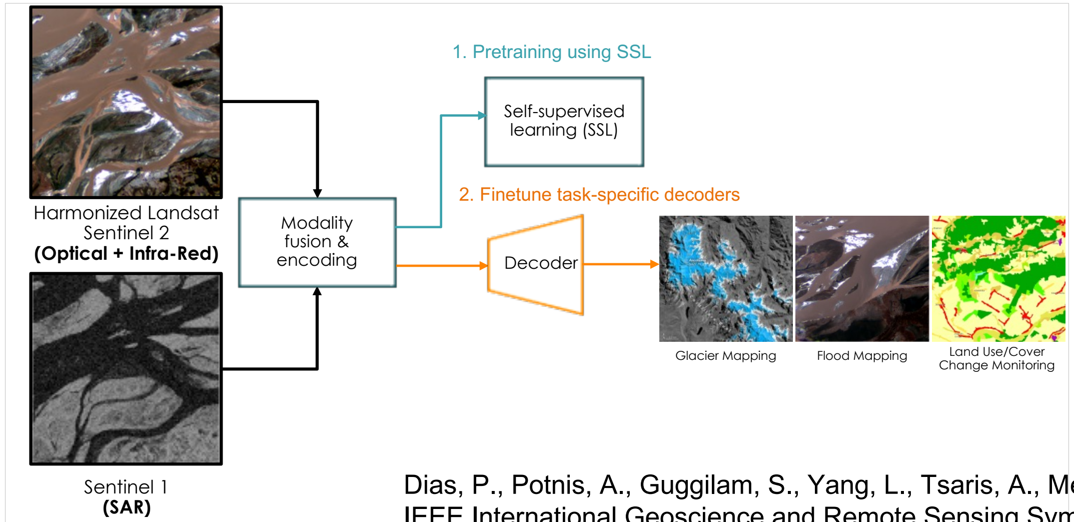
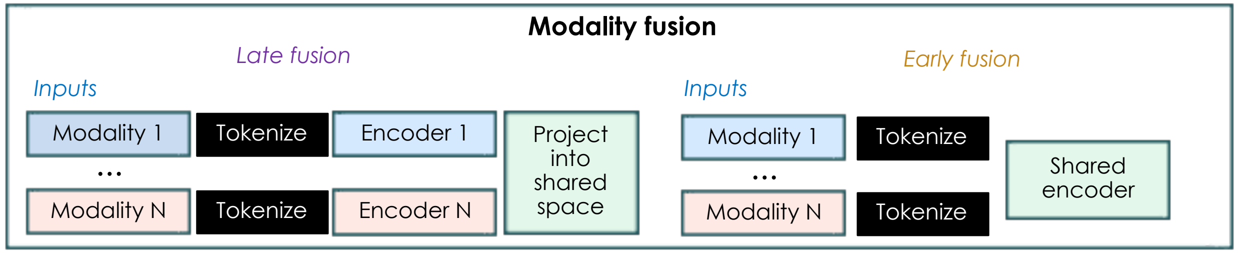


Diagram adapted from: Sun, X. et al. *"RingMo: A remote sensing foundation model with masked image modeling"*. IEEE TGRS (2022)

- Quetzal-LR: Low-resolution (LR) + multimodality (SAR)





Dias, P., Potnis, A., Guggilam, S., Yang, L., Tsaris, A., Medeiros, H. and Lunga, D. "*An Agenda for Multimodal Foundation Models for Earth Observation*". IEEE International Geoscience and Remote Sensing Symposium 2023

“Mise en place” toward such models

Data

- Maxar WorldView 3 imagery
  - RGB+NIR, ~0.5meter/pixel

Labeled

- ORBITaL-Net (ORNL BFE) [1]
  - North America, South America, Africa, Asia
  - variety of viewing angles, vernacular architecture styles, land-use contexts, atmospheric conditions
  - 130k tiles, 512 x 512 pixels each

Unlabeled

Access to rolling database of **~400k+ high-resolution satellite images**, **PBs of data**

Key aspects / building blocks

Data
- Vast / rich volumes

Pretraining objectives
- Self-Supervised Learning (SSL)

Architectures
- Scalable / parallelizable

HPC resources
- Billions - Trillion of parameters → lots of FLOPS

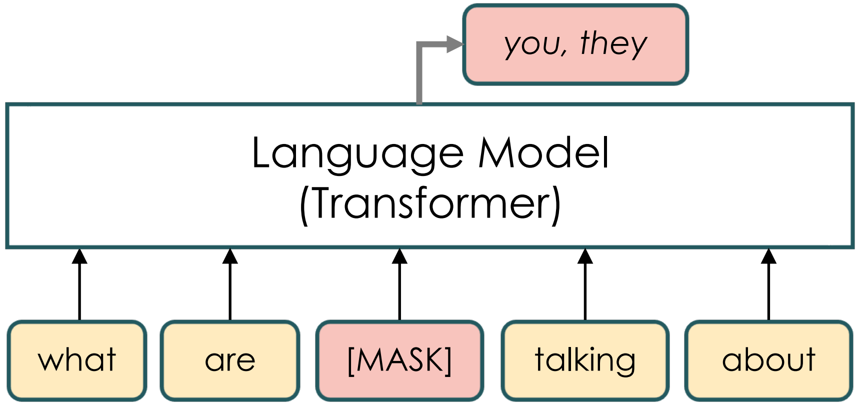
Downstream adaptation & evaluation
- Efficient adaptation to multiple tasks
- Emergent properties

[1] Swan, B.; Pyle, J.; Roddy, D.; Rose, A.; Yang, H. L.; Laverdiere, M. (2024). “*ORBITaL-Net Training Library for Building Extraction. Figshare+. Dataset*”. https://doi.org/10.25452/figshare.plus.25282225.v1

Architecture | **Pretraining objectives**

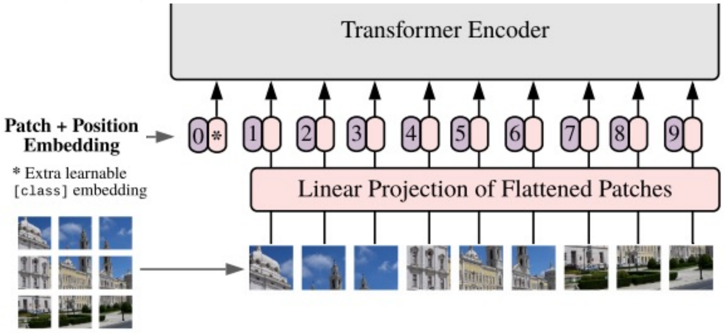
Large Language Models (LLMs)

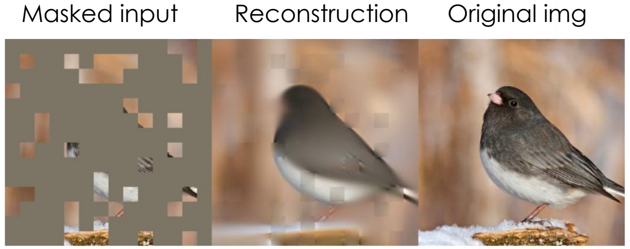
- Transformer-based architectures
  - Data tokenization: “words”
- Masked Language Modeling
  - Randomly mask a portion of the input tokens in a sentence
  - Task model to predict masked tokens
  - e.g., BERT, GPT



Large Vision Models

- Vision Transformers (ViT)
  - Data tokenization: ~~pixels?~~ Patches!
- Masked Image Modeling
  - Randomly mask a portion of the patches in an image
  - Task model to reconstruct masked patches
  - e.g., Masked Autoencoders (MAE)





From [1]

[1] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR* 2020.
[2] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." IEEE/CVF CVPR 2022.

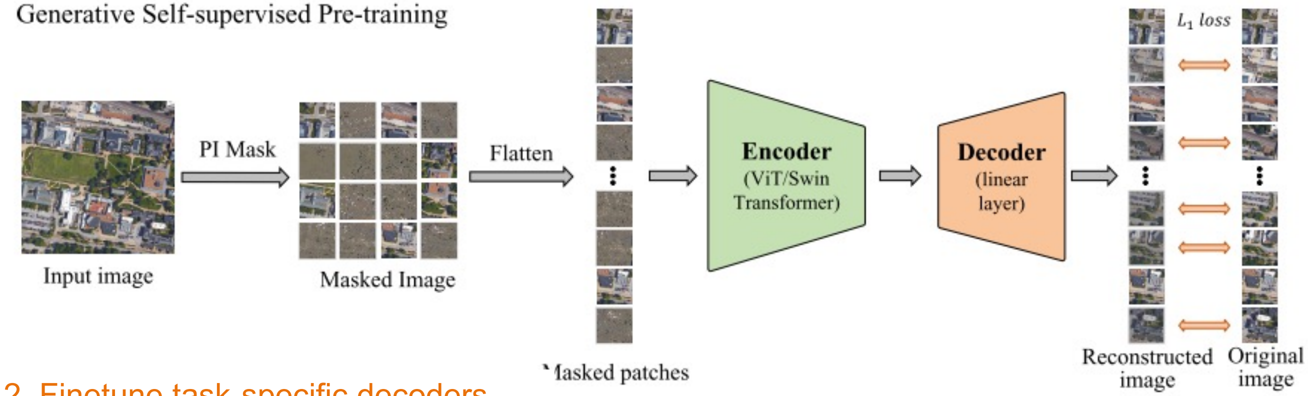
Pretraining objectives

Downstream adaptation

Quetzal-HR: High-resolution (HR)

- **Pretraining**: Masked Autoencoder (MAE)
- **Downstream:** finetune task-specific decoders

1. Masked Autoencoder (MAE) for SSL



2. Finetune task-specific decoders

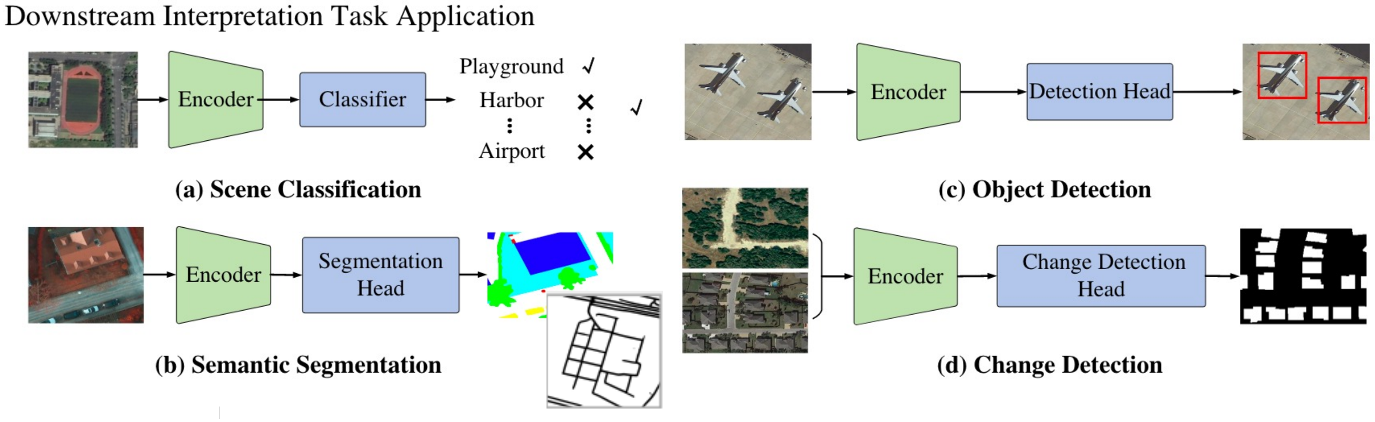


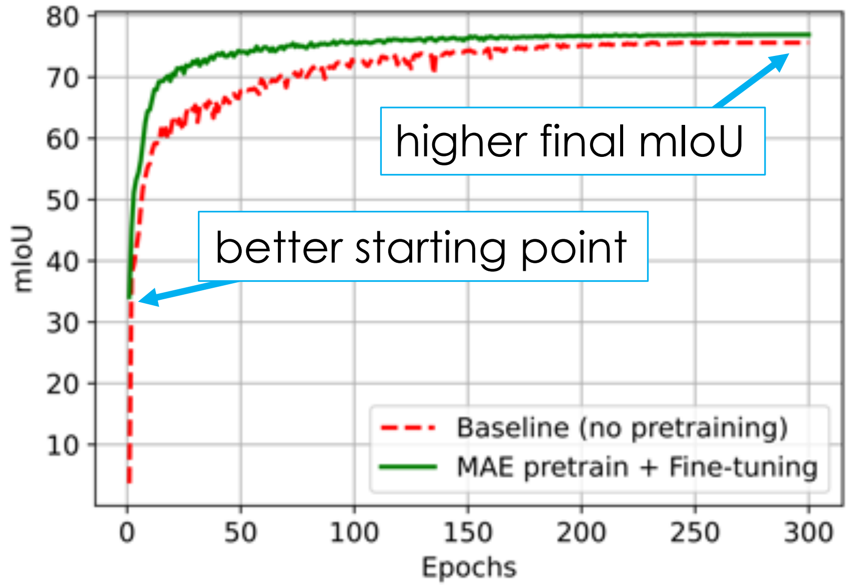
Diagram adapted from: Sun, X. et al. *"RingMo: A remote sensing foundation model with masked image modeling"*. IEEE TGRS (2022)

Quetzal-HR: 4-band

Building Footprint Extraction as application

- Pretraining & finetuning using same image tiles
- ViT-B (86M parameters) + UperNet

- Computing setup
  - PyTorch with Distributed Data Parallel (DDP)
  - Summit and now Frontier
  - Pretraining: 8 nodes (64 GPUs) – BS=2046



4,000+ validation tiles

	F1	Recall	Precision
Baseline (no pretrain)	90.78	89.31	92.30
MAE pretrain + FT	91.79	90.78	92.83

180 out-of-geography (test) tiles

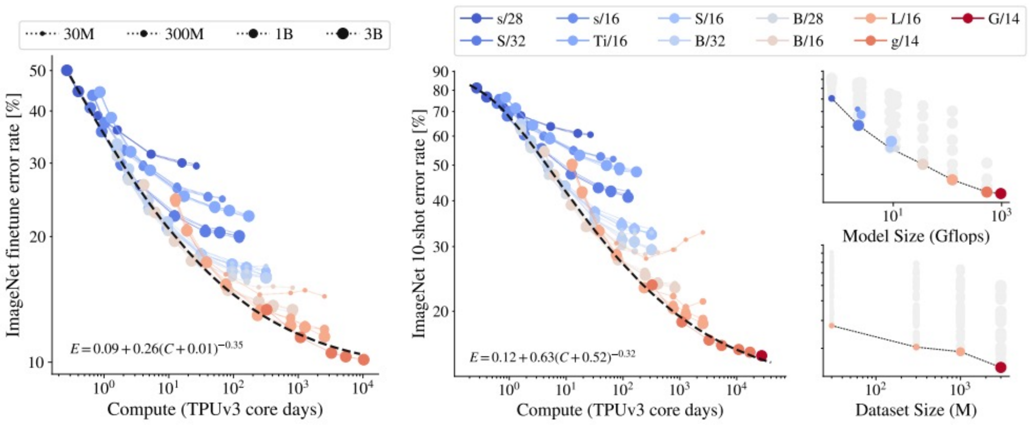
	F1	Recall	Precision
Baseline (no pretrain)	86.58	81.23	92.69
MAE pretrain + FT	90.51	89.65	91.38

Model scaling

- Multiples works taking place in Remote Sensing
  - Contrastive learning, Masked Autoencoders
  - **But restricted to small scale (model sizes)**
    - Mostly conducted by academia

An incomplete summary of FMs developed for EO

Reference	Model size	GPUs
GASSL	ResNet (~25M)	N/A
Sat-MAE, Scale-MAE	ViT-Large (300M)	8 V100 GPUs N/A
RVSA	ViT-Base	8 A100 GPUs
RingMo	Swin/ViT-Base	N/A V100 GPUs
Prithvi	ViT-Large	64 A100 GPUs
SeCo	ResNet (~25M)	N/A
Satlas	Swin-Base	N/A
GFM	Swin-Base	8 V100 GPUs
*SkySense	ViT-L/Swin-H (654M)	80 A100 GPUs

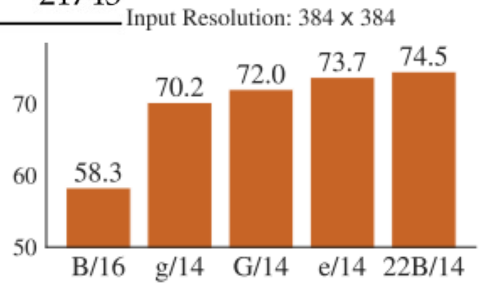


Zhai, X., et al. "Scaling vision transformers." IEEE/CVF CVPR 2022.

Table 1: ViT-22B model architecture details.

Name	Width	Depth	MLP	Heads	Params [M]
ViT-G	1664	48	8192	16	1843
ViT-e	1792	56	15360	16	3926
ViT-22B	6144	48	24576	48	21743

Dehghani, M., et al. "Scaling vision transformers to 22 billion parameters." *ICML* 2023.



Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

Quetzal-HR – Open data

- MAE pretraining with 1M samples (MillionAID)
- ViT configurations up to **3B** parameters
  - Frontier, Pytorch DDP, 2048 global batch size, 100k iterations
- Semantic Segmentation (fine-tuning)
  - more complex decoder (3B requires sharding)
    - 64 nodes (512 GPUs), BS=1 for ViT-1B model
  - limited gains with limited data

Model	Parameters [M]
ViT-Base	87
ViT-Huge	635
ViT-1B	914
ViT-3B	3067

Image segmentation (fine-tuning)

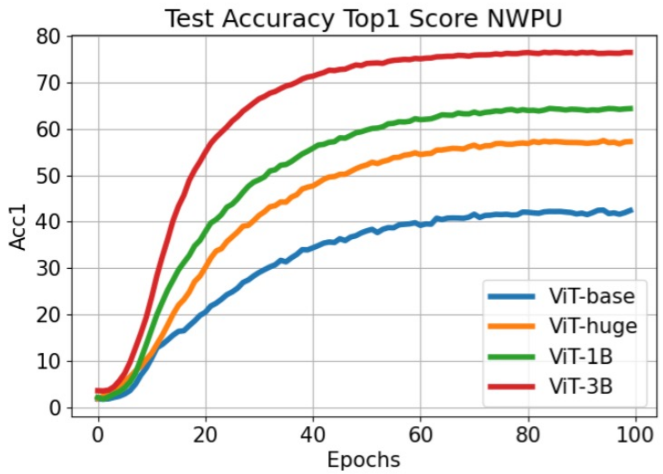
	LoveDA [mIoU % – test]	Potsdam [mF1 % - val]
ViT-Base	50.92	90.83
ViT-Huge	51.94	91.36
ViT-1B	52.58	91.49

Image classification (linear probing)

Image Classification

Datasets	Training Samples	Testing Samples	Classes
MillionAID	1000	9000	51
UCM	1050	1050	21
AID	2000	8000	30
NWPU	3150	28350	45

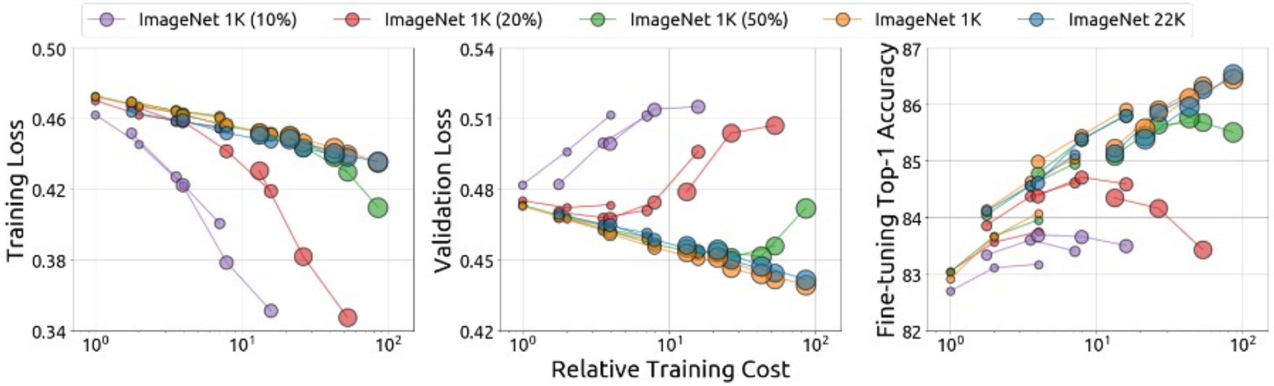
Model	Pretrain epochs	Top1 Acc (%) *UCM* (*TR*=50%)	*AID* (*TR*=20%)	*NWPU* (*TR*=10%)	*MillionAID*
ViT-Base	400	45.17	52.11	54.28	47.20
ViT-Base	100	40.62	41.72	42.40	41.31
ViT-Huge	100	50.00	60.78	57.24	53.28
ViT-1B	100	57.10	68.89	64.35	59.14
ViT-3B	100	74.05	79.96	76.43	72.98



Tsaris, A.; Dias, P.; Potnis, A.; Yin, J.; Wang, F.; Lunga, D. *"Pretraining Billion-scale Geospatial Foundational Models on Frontier"* To be published at IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC 2024)

Data scaling

Larger models require more data to avoid MIM overfitting



Xie, Z. et al. "On data scaling in masked image modeling". *IEEE/CVF CVPR 2023.*

Model	Iter	IN1K (10%)	IN1K (20%)	IN1K (50%)	IN1K (100%)	IN22K (100%)
SwinV2-S	125K	43.4	44.9	45.3	44.2	-
	250K	43.5	46.7	46.6	45.8	-
	500K	43.5	47.2	47.2	48.3	-
SwinV2-B	125K	44.2	45.4	46.1	46.0	46.8
	250K	43.3	46.0	48.5	47.7	47.3
	500K	42.1	46.9	49.0	49.3	48.2
SwinV2-L	125K	43.4	46.4	48.0	48.0	47.4
	250K	43.1	47.3	49.6	50.2	50.0
	500K	41.9	45.6	50.3	51.1	51.2

Table 5: Results (mIoU) on validation set of ADE20K semantic segmentation.

Ineffective to just "dump" a bunch of data

- ORBITaL-Net (ORNL BFE) vs Ukraine only:
  - larger volume, but worse results → diversification issues
  - ORBITaL-Net (ORNL BFE) [1]
    - *North America, South America, Africa, Asia*
    - *variety of viewing angles, vernacular architecture styles, LU/LC contexts, and atmospheric conditions*

	Volume	F1 – Ukraine data	F1 – Global data
Global tiles	0.7 TB	90.51 %	91.79 %
Ukraine images	18 TB	90.55 %	91.40 %

[1] Swan, B.; Pyle, J.; Roddy, D.; Rose, A.; Yang, H. L.; Laverdiere, M. (2024). "*ORBITaL-Net Training Library for Building Extraction. Figshare+. Dataset*". https://doi.org/10.25452/figshare.plus.25282225.v1

Data biases

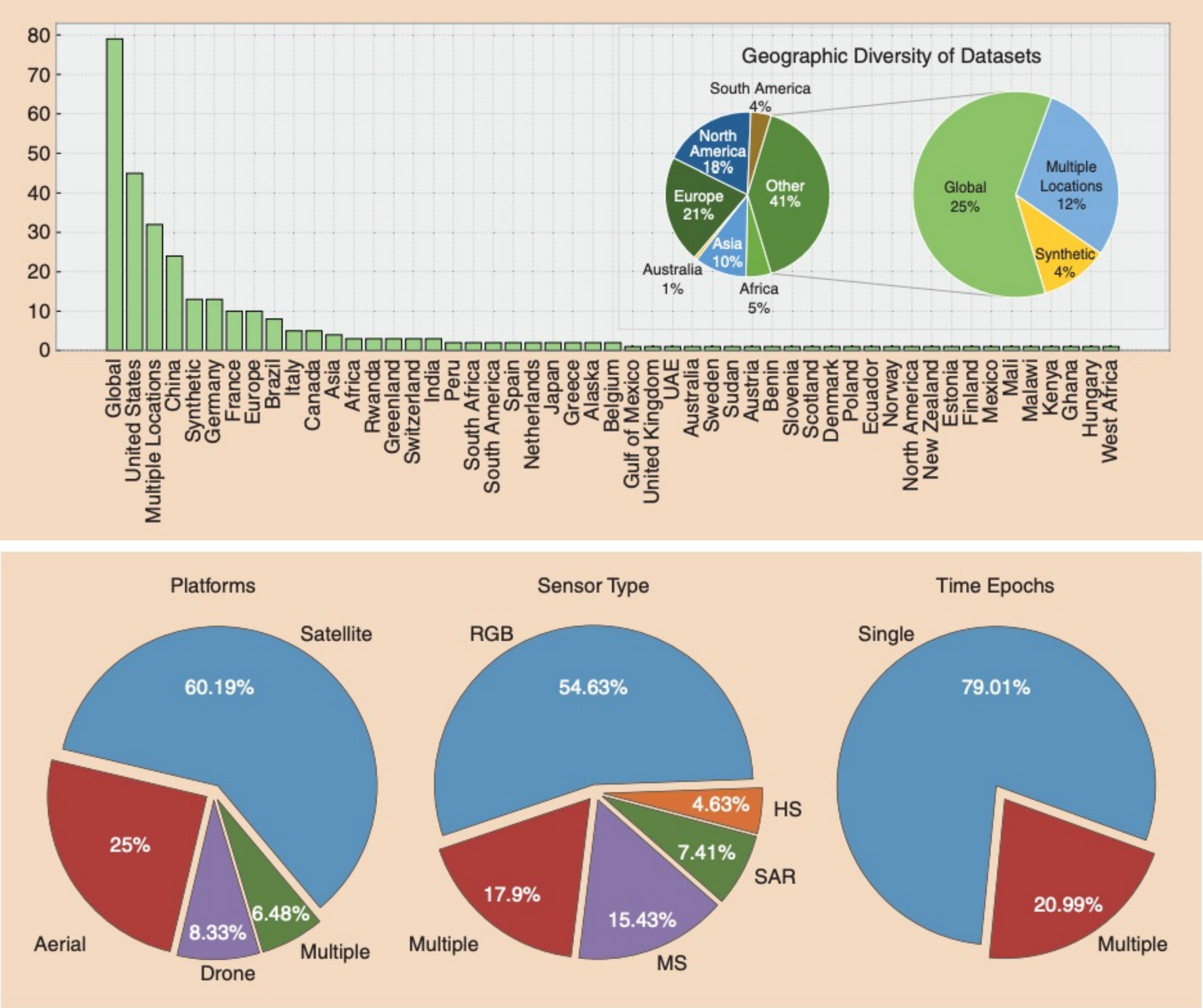
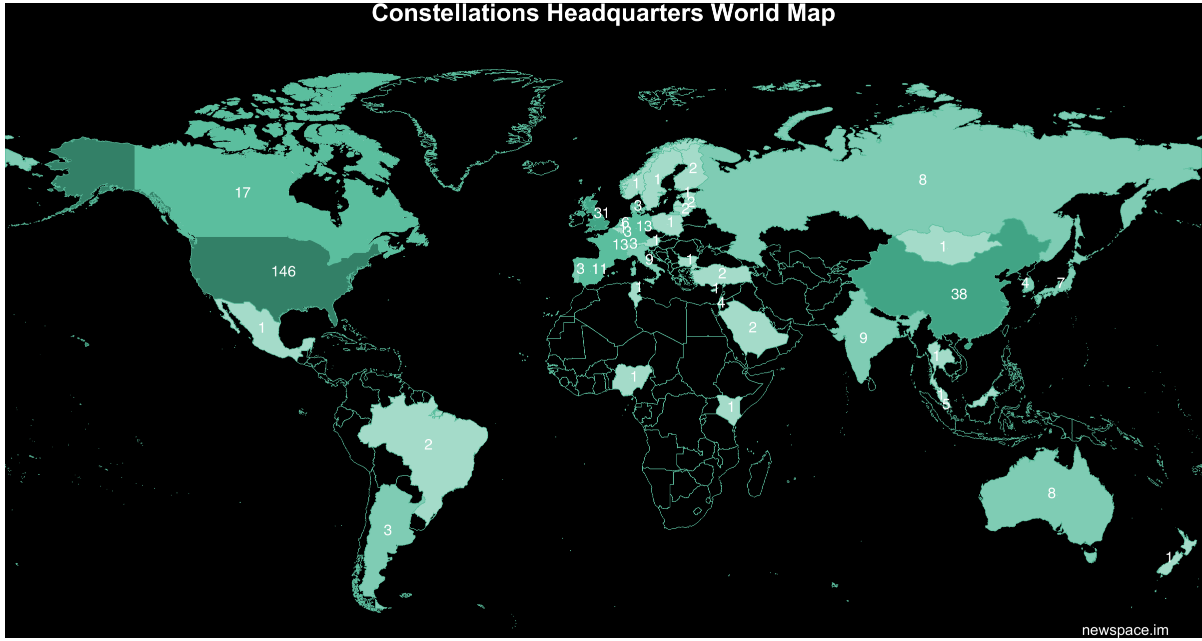


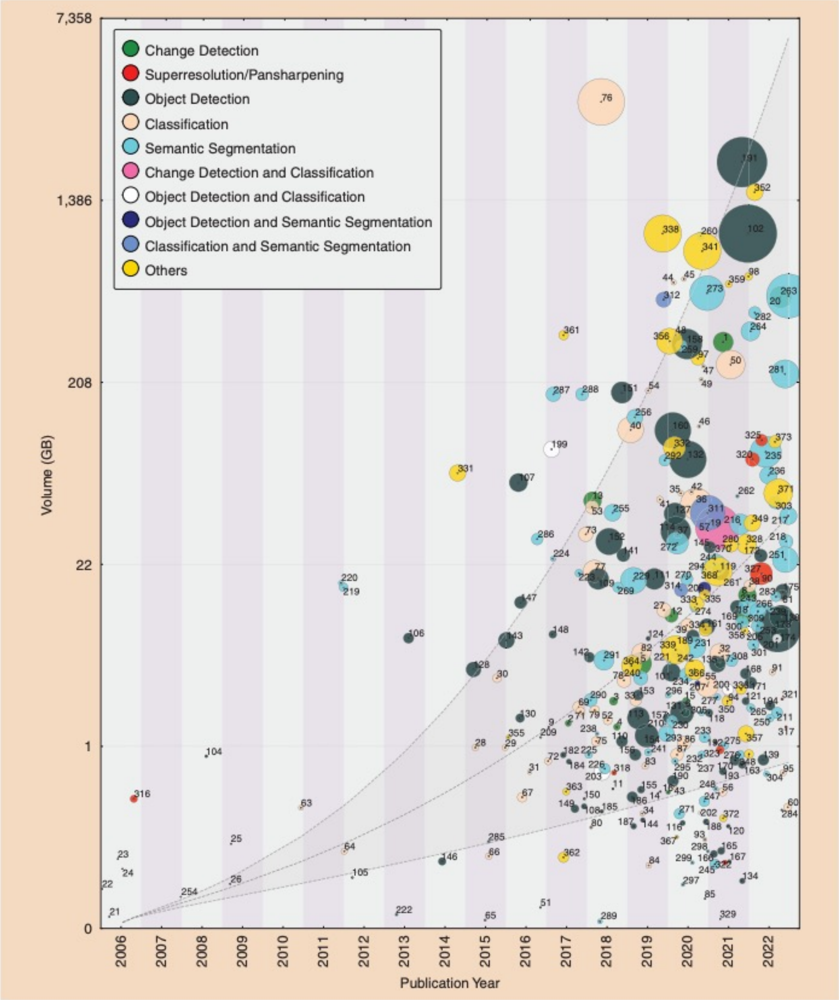
FIGURE 3. A distribution of available EO datasets over different platforms, sensor types, and number of acquisition times. Single-image red, green, blue (RGB) images acquired by satellites are clearly the dominating modality. MS: multispectral; HS: hyperspectral.



Schmitt, Michael, et al. *"There are no data like more data: Datasets for deep learning in earth observation."* IEEE Geoscience and Remote Sensing Magazine (2023).

Dataset needs for pretraining and benchmarking

Currently 😵‍💫



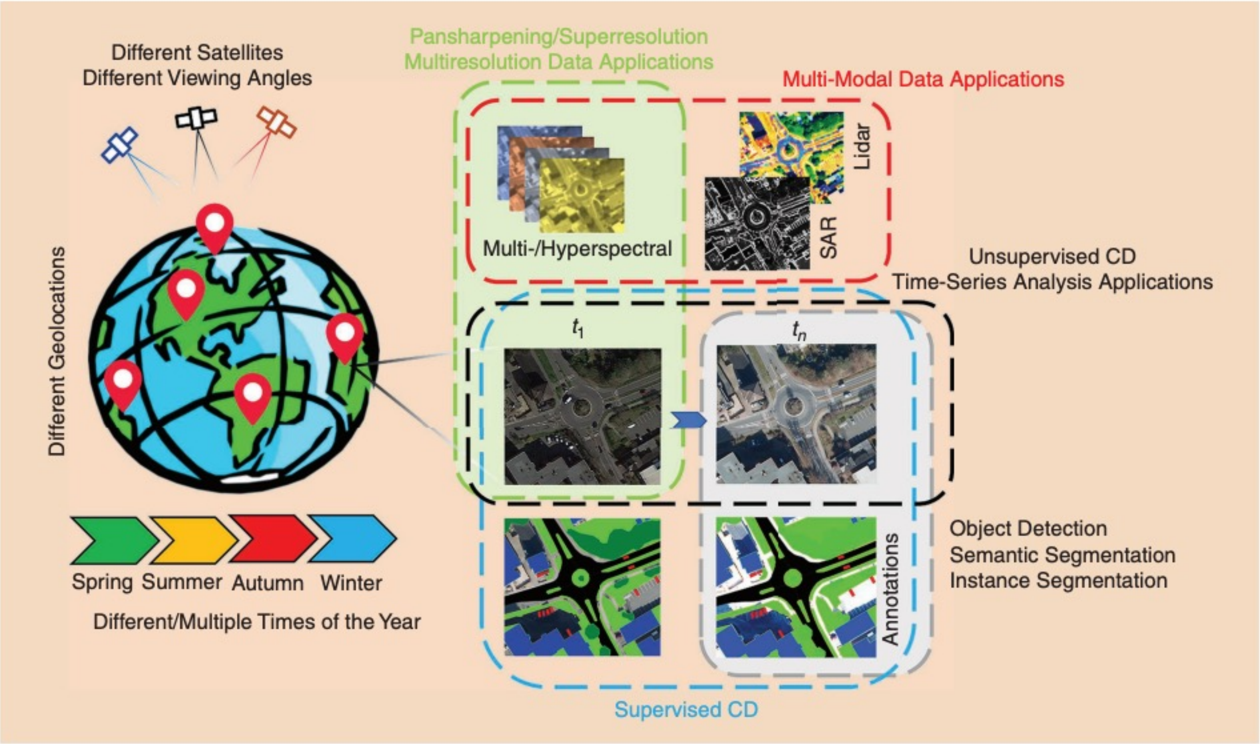
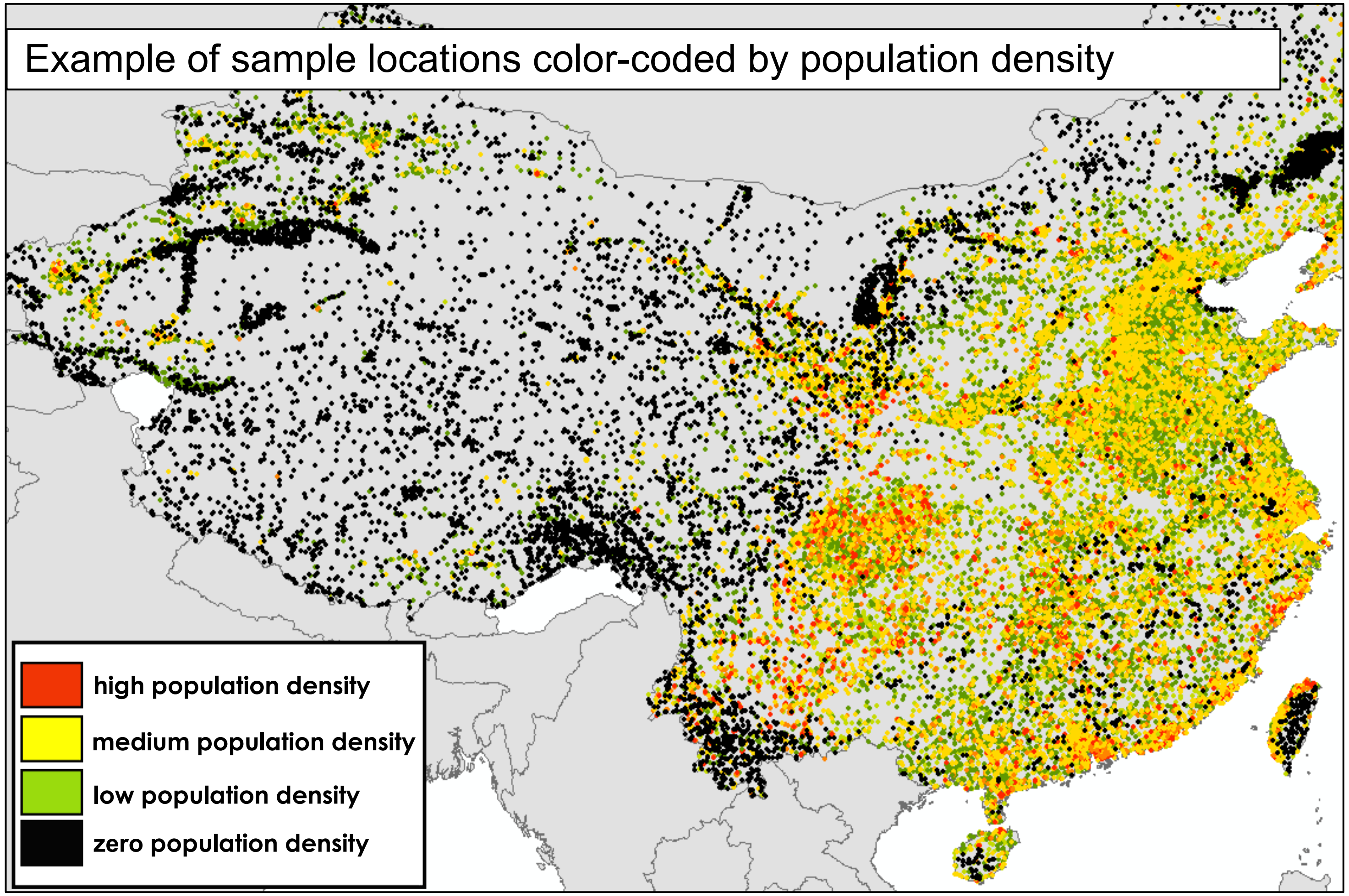


FIGURE 28. An illustration that shows the authors' view of the paramount properties that an ideal benchmark dataset needs to satisfy, including the type of tasks, sensors, temporal constraints, and geolocalization.

Schmitt, Michael, et al. *"There are no data like more data: Datasets for deep learning in earth observation."* IEEE Geoscience and Remote Sensing Magazine (2023).

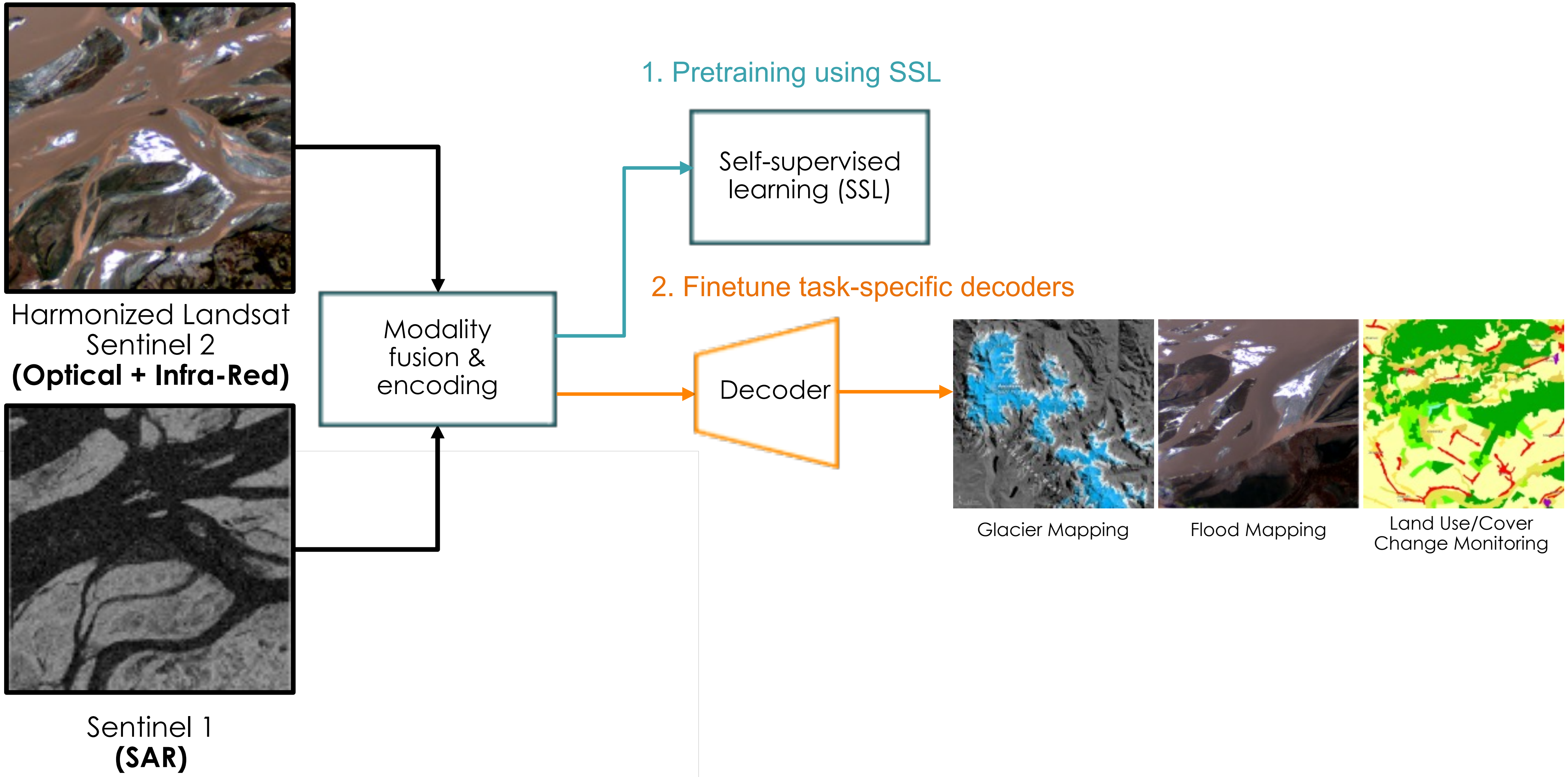
How we are curating a Pretraining Dataset

- Key requirements
  - Geographic Diversity
  - Temporal Diversity
  - Acquisition Parameter Diversity
  - Support for varied Pretext Tasks and Dataset Sizes
- Sampling
  - Geo-clusters based on biome, realm, and climate zone information
    - Koppen-Geiger Climate Zones
    - 2017 Ecoregions Layer
  - Guided sampling based on landcover, population density, and geo-cluster
    - Land Cover: ESA WorldCover v200
    - Population: ORNL LandScan Global



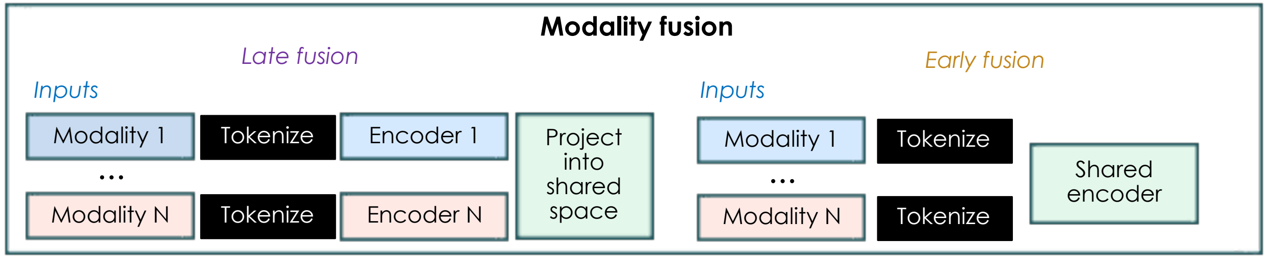
Arndt, J., Dias, P., Potnis, A., and Lunga, D. "*Towards Diverse and Representative Global Pretraining Datasets for Remote Sensing Foundation Models*". Accepted for publication at IEEE International Geoscience and Remote Sensing Symposium 2024

Quetzal-LR: Low-resolution (LR) + multimodality (SAR)



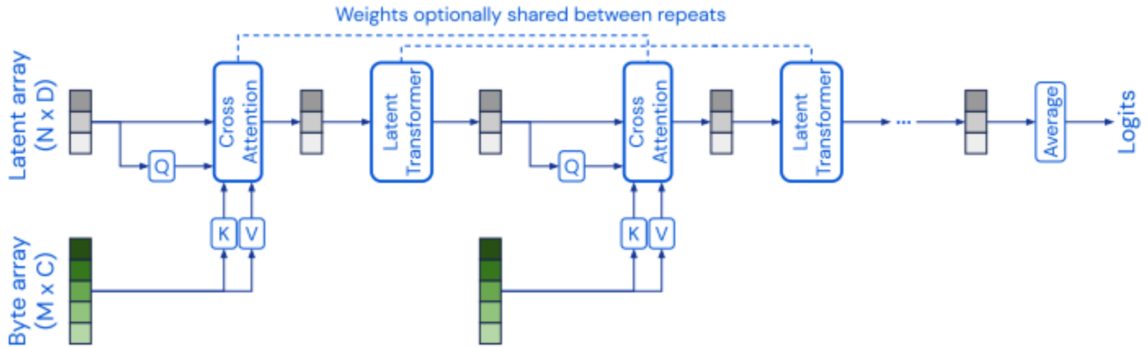
Multimodal reasoning

- Modality alignment
  - different input types, dimensionalities, resolutions
- Early fusion vs late fusion vs mixed
  - e.g., cross-attention mechanisms
- Challenges
  - Risk of model relying mostly in certain modalities over others
  - Different data availability for different modalities



Perceiver

- concept of cross-attention
- inputs of different dimensionalities projected into fixed-dimensional space



Jaegle, A., et al. "Perceiver: General perception with iterative attention." *ICML* 2021.

ClimaX:
A foundation model for weather and climate

- modality-specific tokenization + *aggregation*
- *aggregation:* cross-attention outputs single vector per spatial position

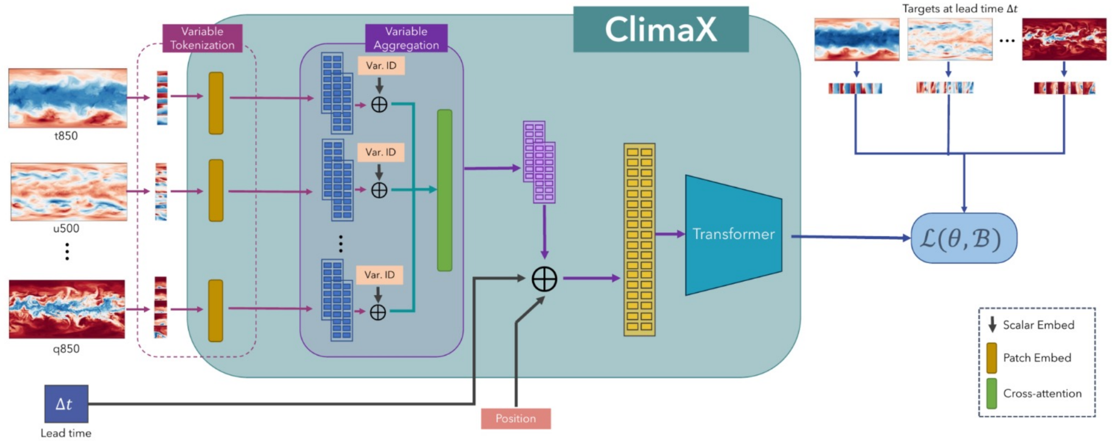


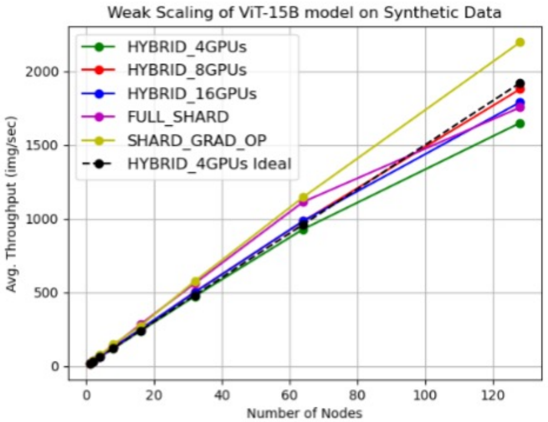
Figure 2: The ClimaX architecture as used during pretraining. Variables are encoded using variable-separate tokenization, and subsequently aggregated using variable aggregation. Together with position and lead time embedding those are fed to the ViT backbone.

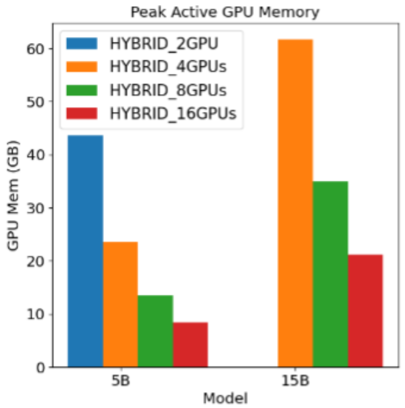
Nguyen, T., et al. "ClimaX: A foundation model for weather and climate." *ICML* 2023.

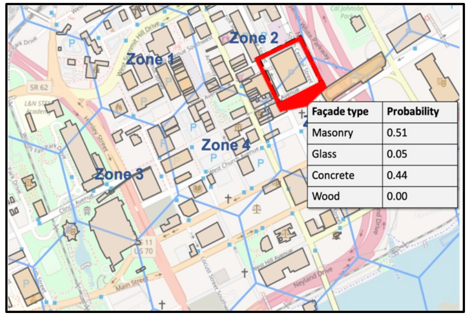
Closing thoughts

Tsaris, A.; Dias, P.; Potnis, A.; Yin, J.; Wang, F.; Lunga, D. *“Pretraining Billion-scale Geospatial Foundational Models on Frontier”* To be published at IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC 2024)

Model	Width	Depth	MLP	Heads	Parameters [M]
ViT-Base	768	12	3072	12	87
ViT-Huge	1280	32	5120	16	635
ViT-1B	1536	32	6144	16	914
ViT-3B	2816	32	11264	32	3067
ViT-5B	1792	56	15360	16	5349
ViT-15B	5040	48	20160	48	14720

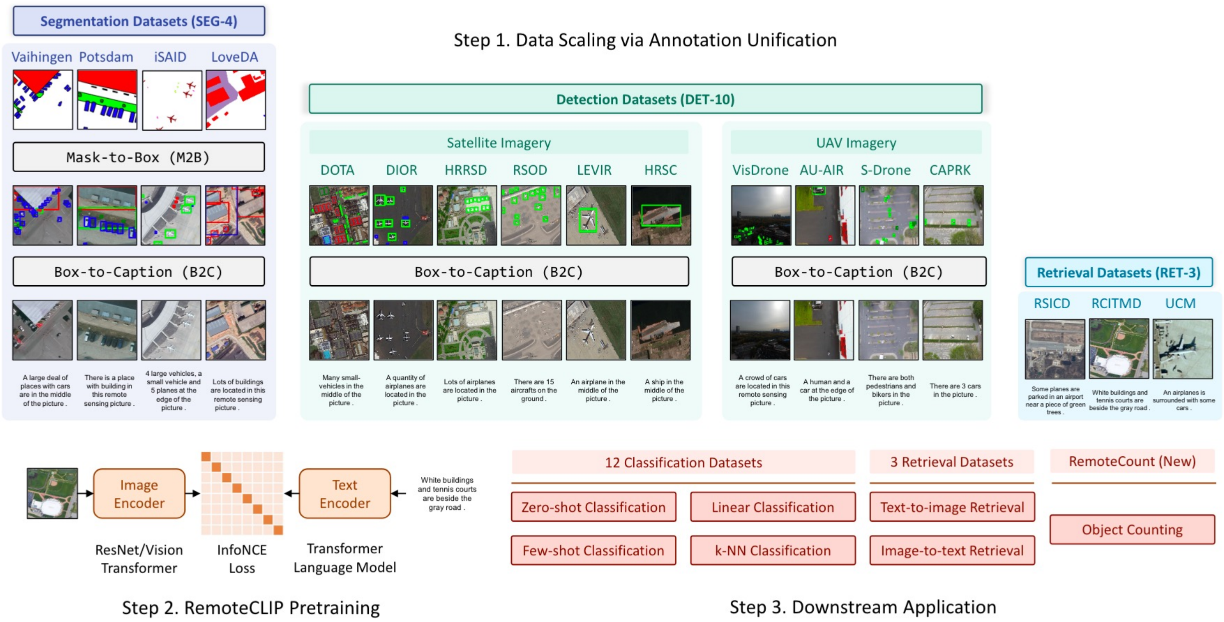




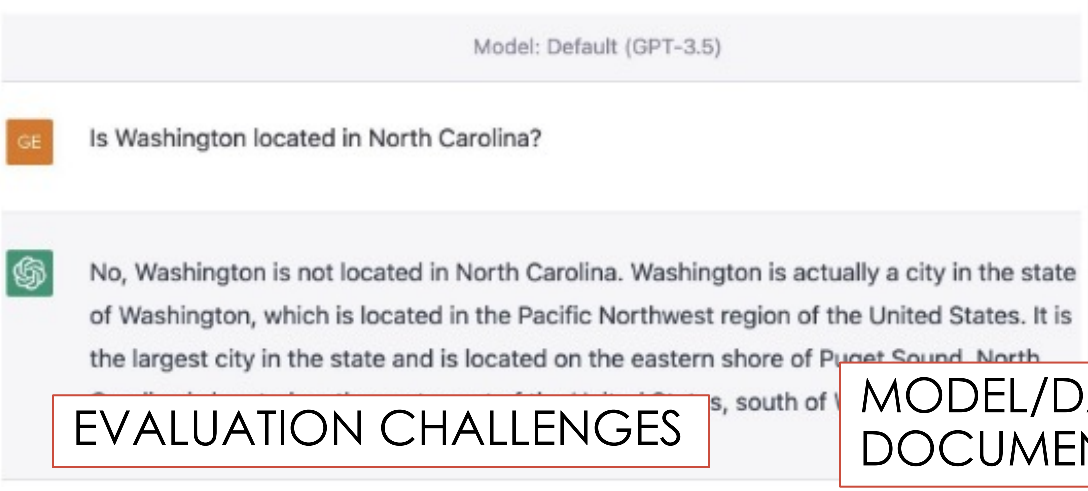


Bayesian Modeling: Estimation of *probable* material types for each building in Knoxville, TN. Image Credit: ORNL

Liu, F., et al. "RemoteCLIP: A vision language foundation model for remote sensing." *arXiv preprint arXiv:2306.11029* (2023).



Mai, G., et al. "On the opportunities and challenges of foundation models for geospatial artificial intelligence." arXiv preprint (2023).



EVALUATION CHALLENGES

MODEL/DATA DOCUMENTATION

OPTIMIZATION CHALLENGES W/ LARGE BATCH SIZES

Thank you!

Contacts:

ambroziodiap@ornl.gov

geoai.ornl.gov

- Acknowledgements
  - Aristeidis Tsaris (CCSD)

  GeoAI colleagues:
  - Dalton Lunga
  - Abhishek Potnis
  - Jacob Arndt
  - Jordan Bowman
  - Lexie Yang
  - OLCF/Frontier
  - CADES