FROM PIXELS TO LANGUAGE: THE EVOLUTION OF AI FOR EARTH OBSERVATION

2026 Summer School on AI Advances in Earth Observation

Luca Chiarabini
German Aerospace Center (DLR)
Wessling-Oberpfaffenhofen, Germany

DLR

1. About me

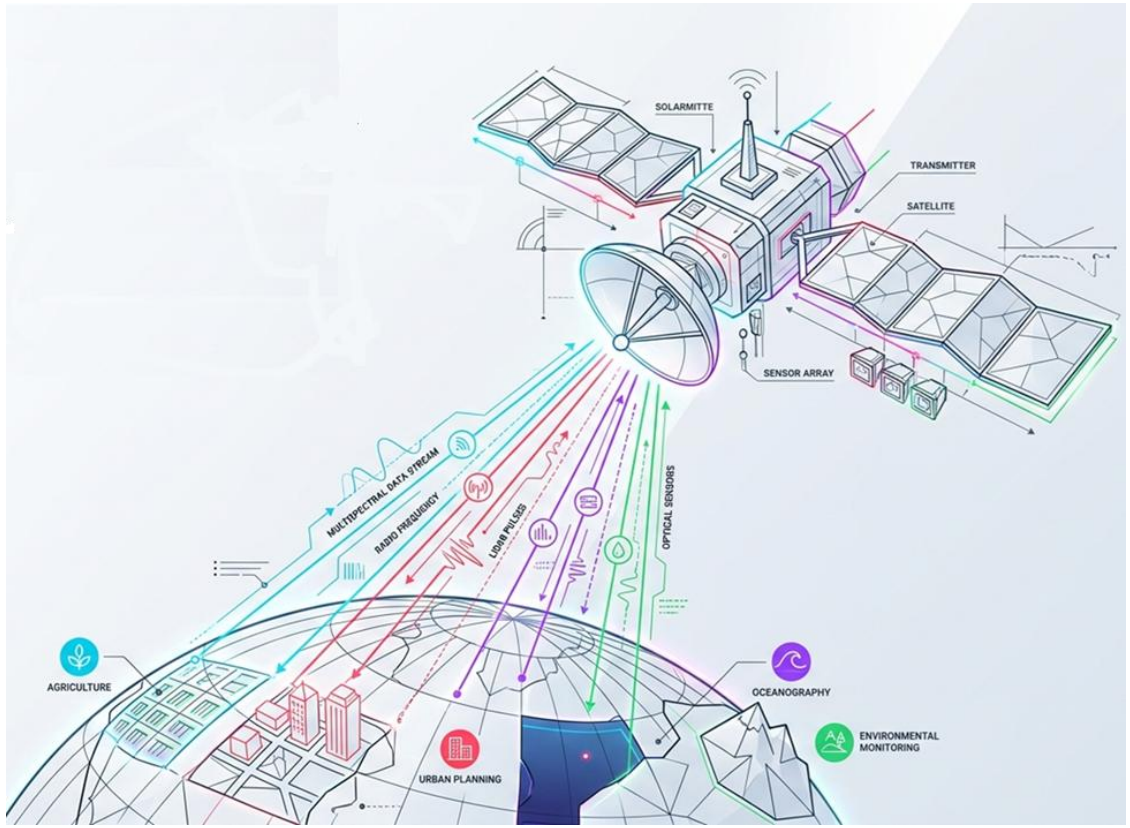
About me

Luca Chiarabini, Ph.D.

- AI Engineer at DLR
- Working on Artificial Intelligence and Vision Language Models, with a background in research, industry and university teaching
- Ph.D. in Computer Science

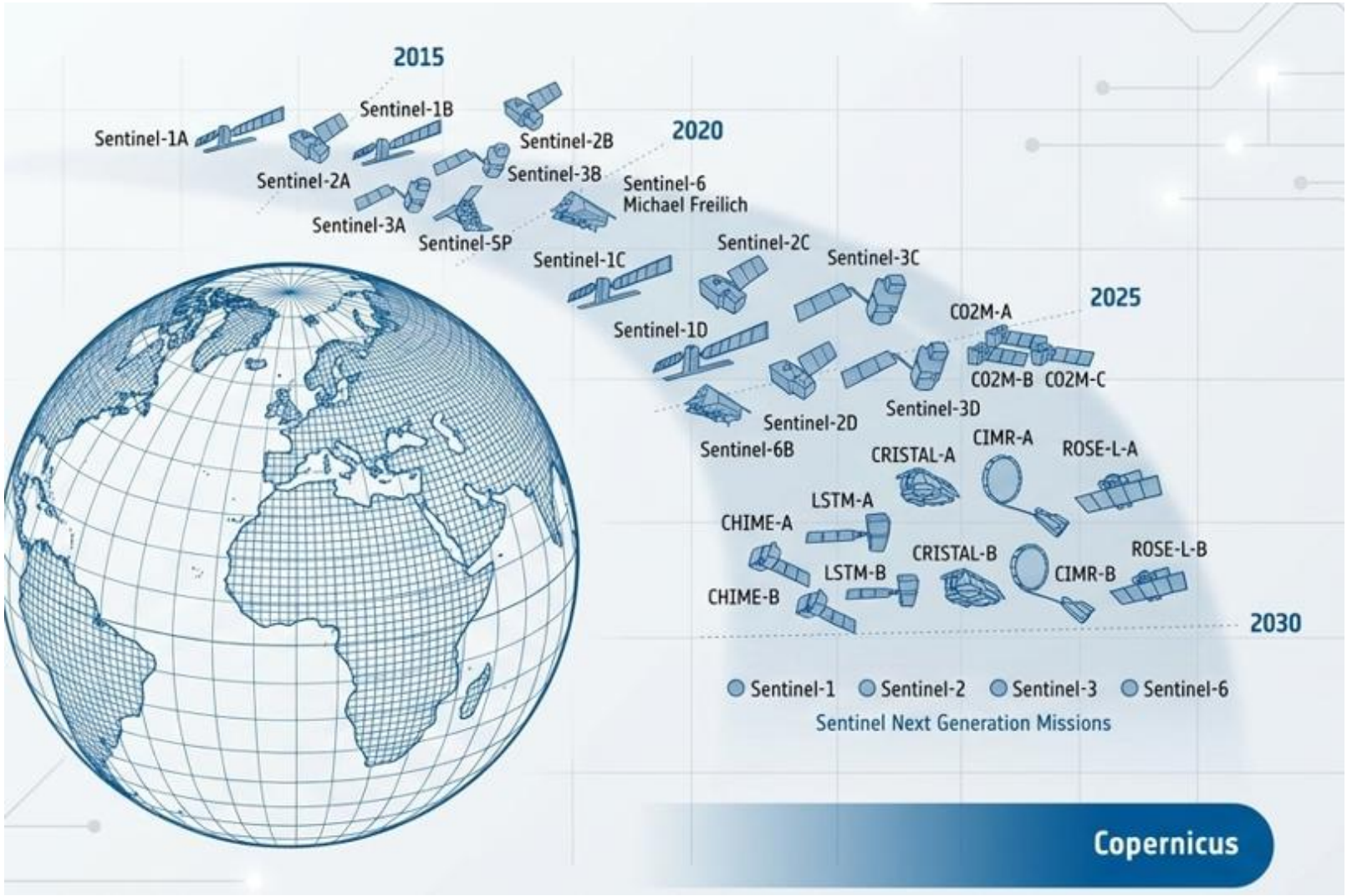
2. Earth Observation the Essentials

What is Earth Observation?

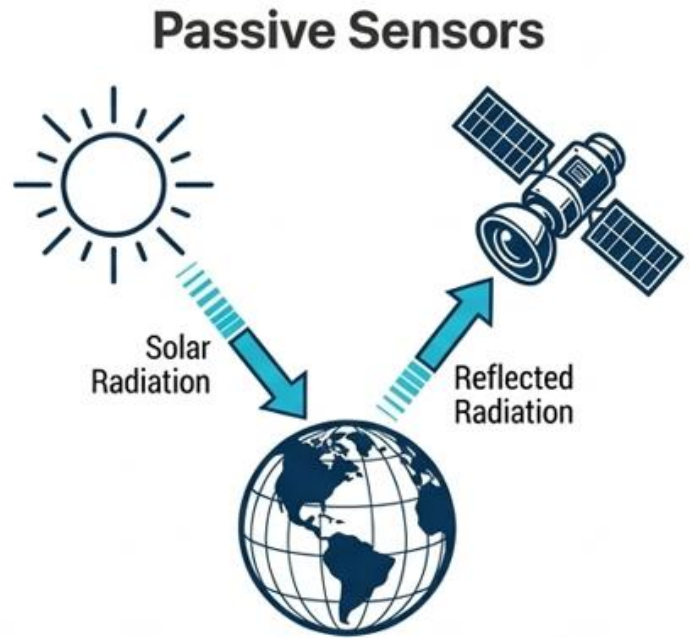


- **Earth Observation (EO)** is the acquisition of information about the Earth‘s surface using remote sensing instruments deployed on:
  - Satellites
  - Aircraft
  - Drones
  - Ground-based platforms
- **Goal** Transform observations into useful information for decision-making.
- **Applications** Environmental monitoring, Agriculture, Urban planning, Disaster management.

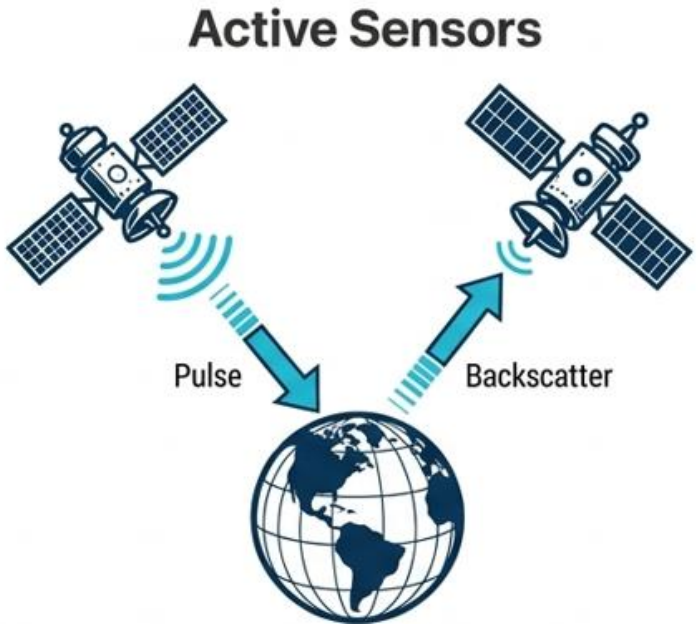
The Copernicus Programme & Sentinel Missions



Active vs. Passive Sensors

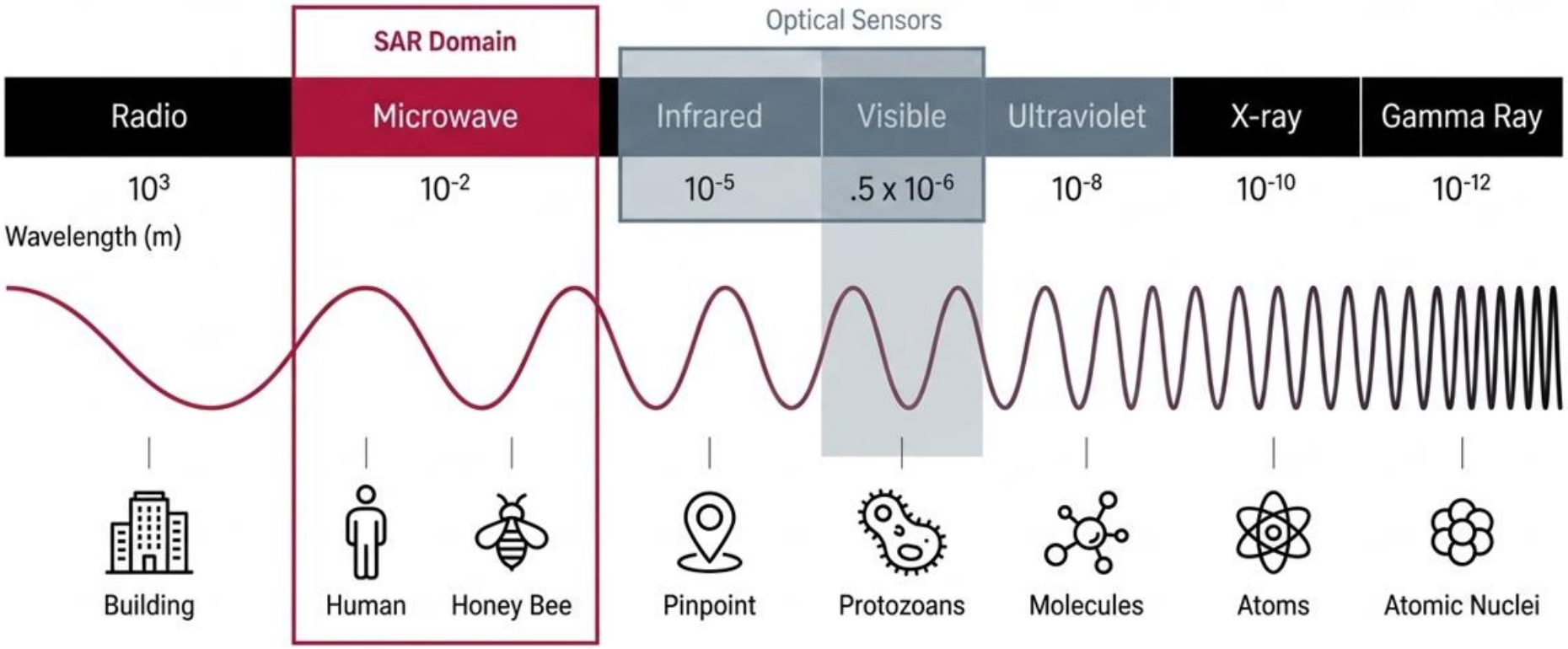


- Measure naturally available radiation
- Optical sensors are the key example
- Depend on illumination and atmospheric conditions



- Emit their own signal and measure the returned energy
- Radar/SAR
- Can operate day/night

The Microwave Advantage



Optical vs SAR: Complementary Information

Sentinel-2 (Passive Sensor)

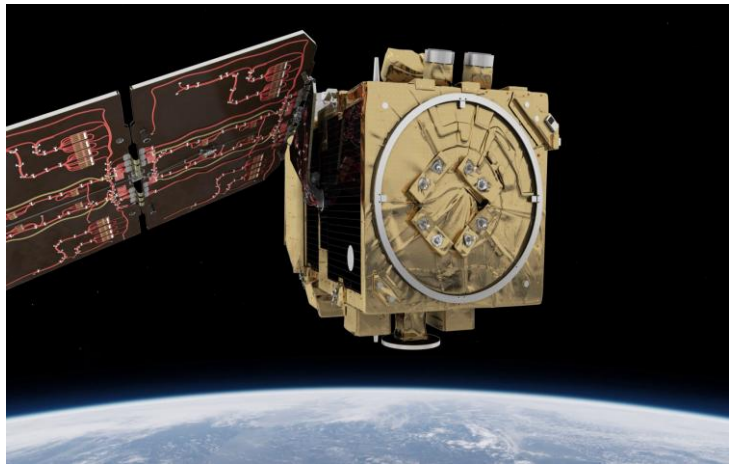
- Multispecral/Optical
- Reflected sunlight
- 13 spectral bands at 10, 20 and 60 m
- Vegetation, land, cover, water …

Sentinel-1 (Active Sensor)

- SAR
- C-band microwave backscatter
- All-weather, day-and-night observations
- Sensitive to surface structure, roughness, moisture, flooding

Sentinel-2: Multispectral Optical Imaging

@Wikipedia. Artistic representation of Landsat-8

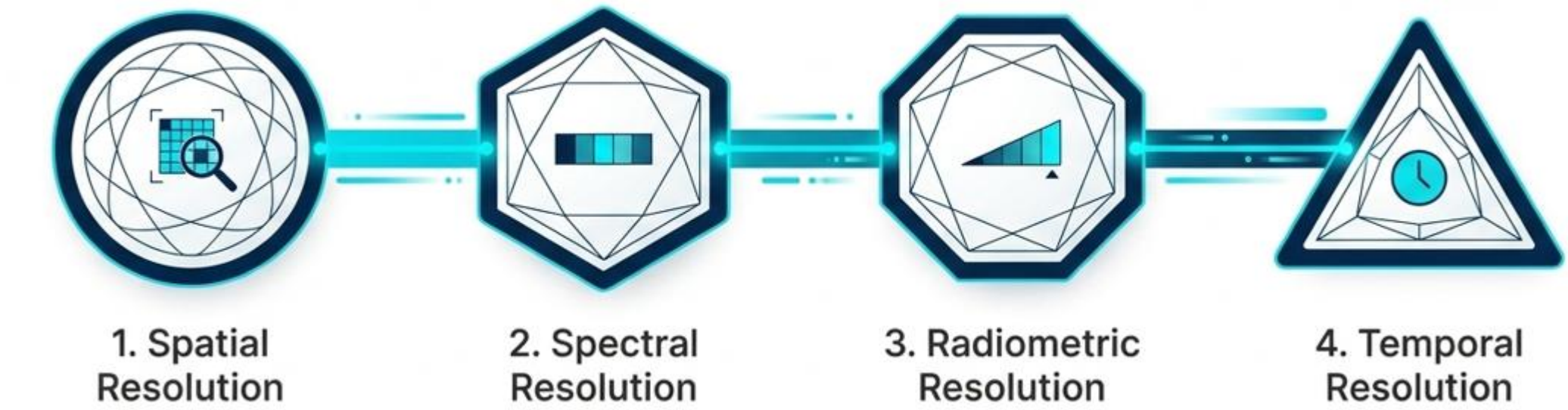


@ESA-Sentinel-2

Copyright © AIRBUS Defence & Space

- 13 spectral bands
- 10 / 20 / 60 m spatial resolution
- ~5 dayd revisit time

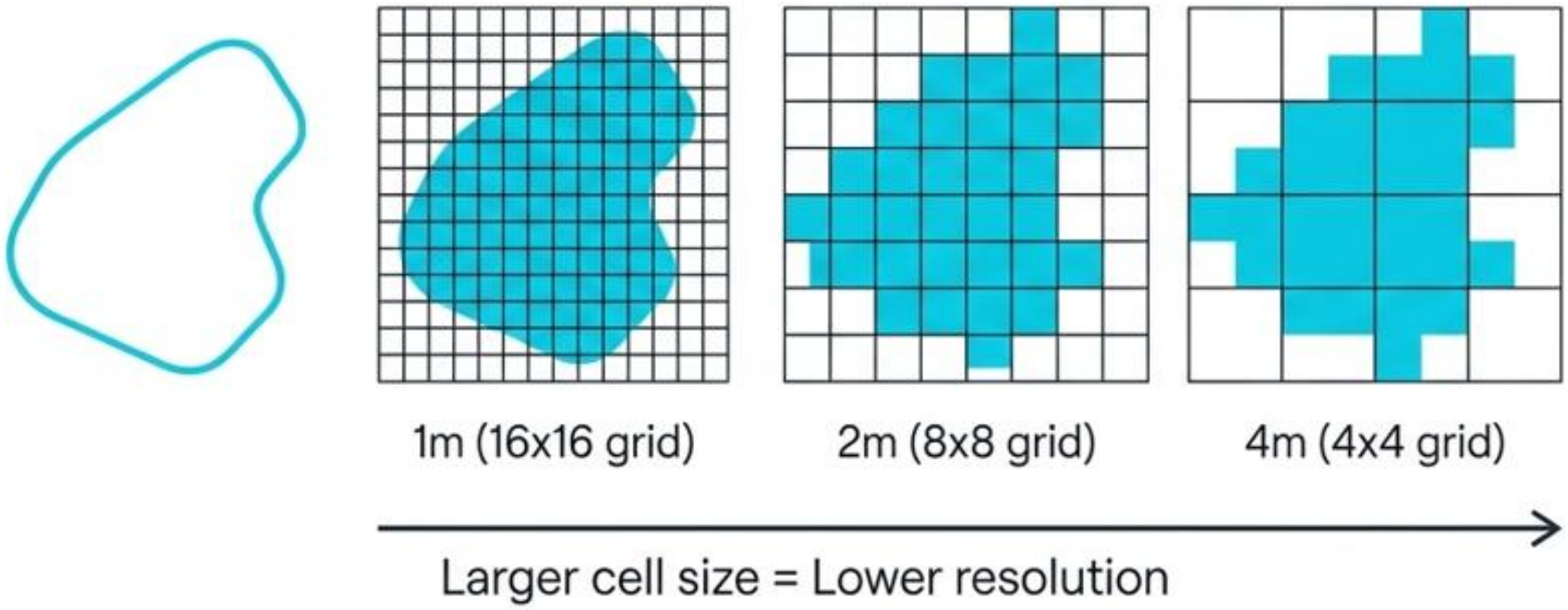
Understanding Resolution in Earth Observation



- **Spatial Resolution** How much spatial details can we observe?
- **Spectral Resolution** How finely do we sample the electromagnetic spectrum?
- **Radiometric Resolution** How precisely can we distinguish differences in measured intensity?
- **Temporal Resolution** How frequiently do we observe the same location?

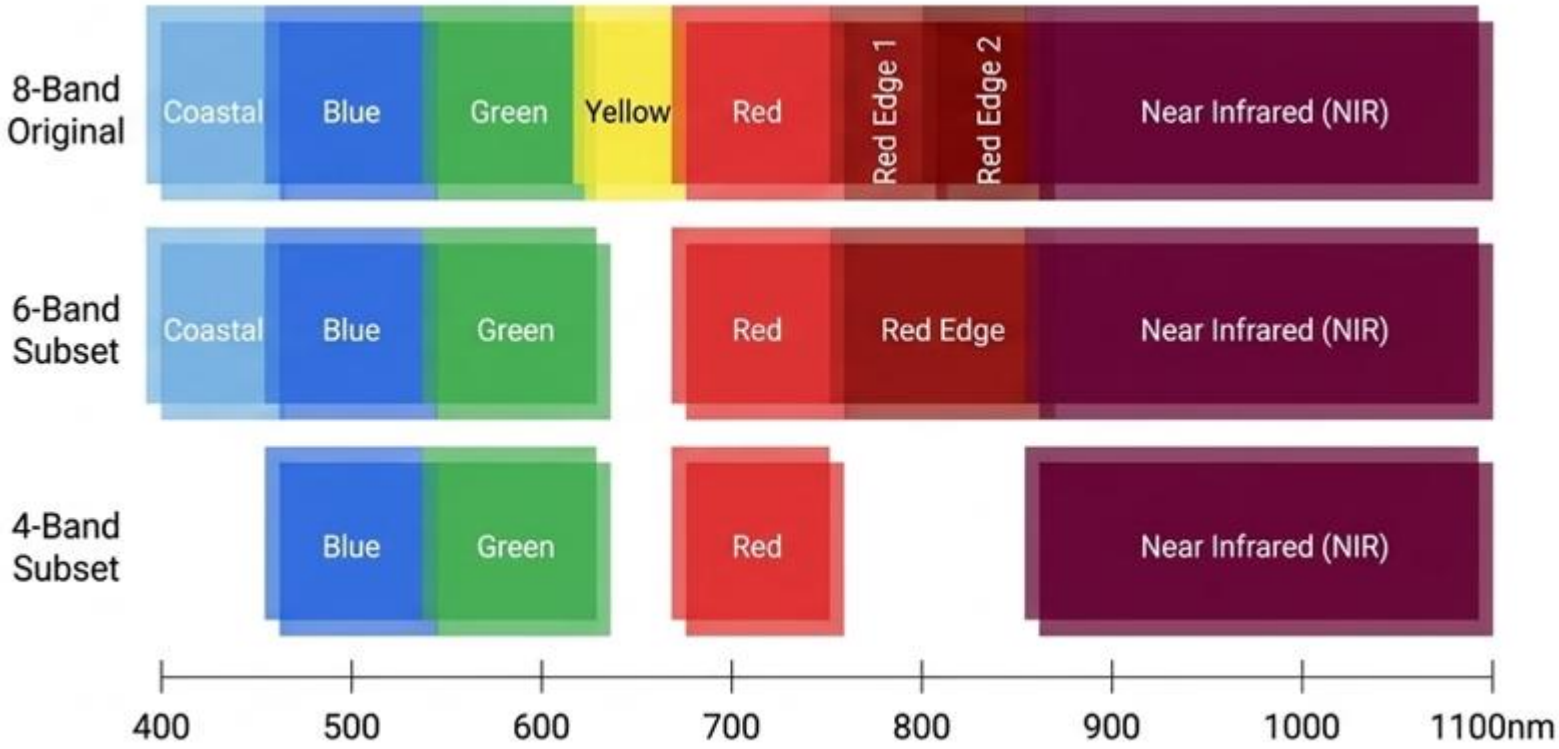
Spatial Resolution

How much spatial details can we observe?



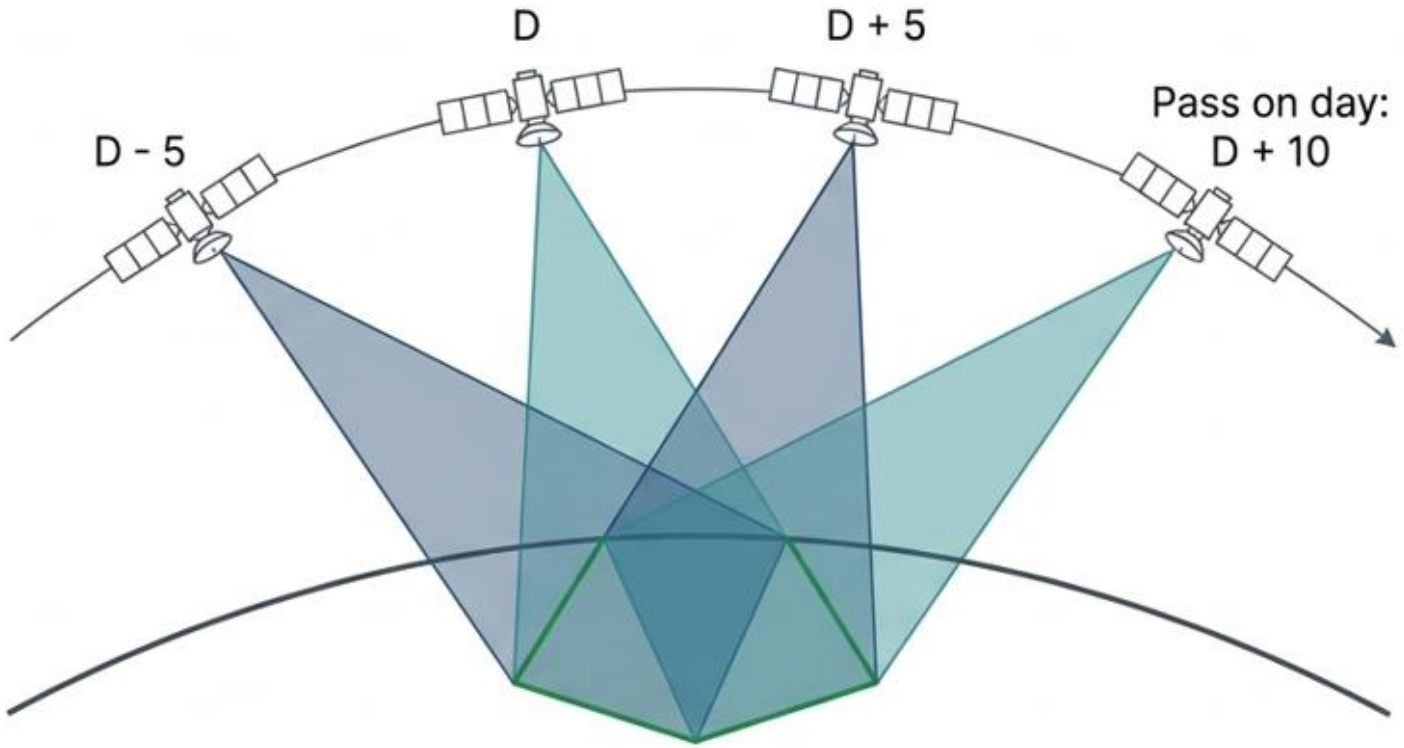
Spectral Resolution

How finely do we sample the electromagnetic spectrum?



Temporal Resolution

How frequently do we observe the same location?



3. AI for Earth Observation

From Earth Observation Data to Information

The Wealth of Satellite Data has Little Value Without Interpretation

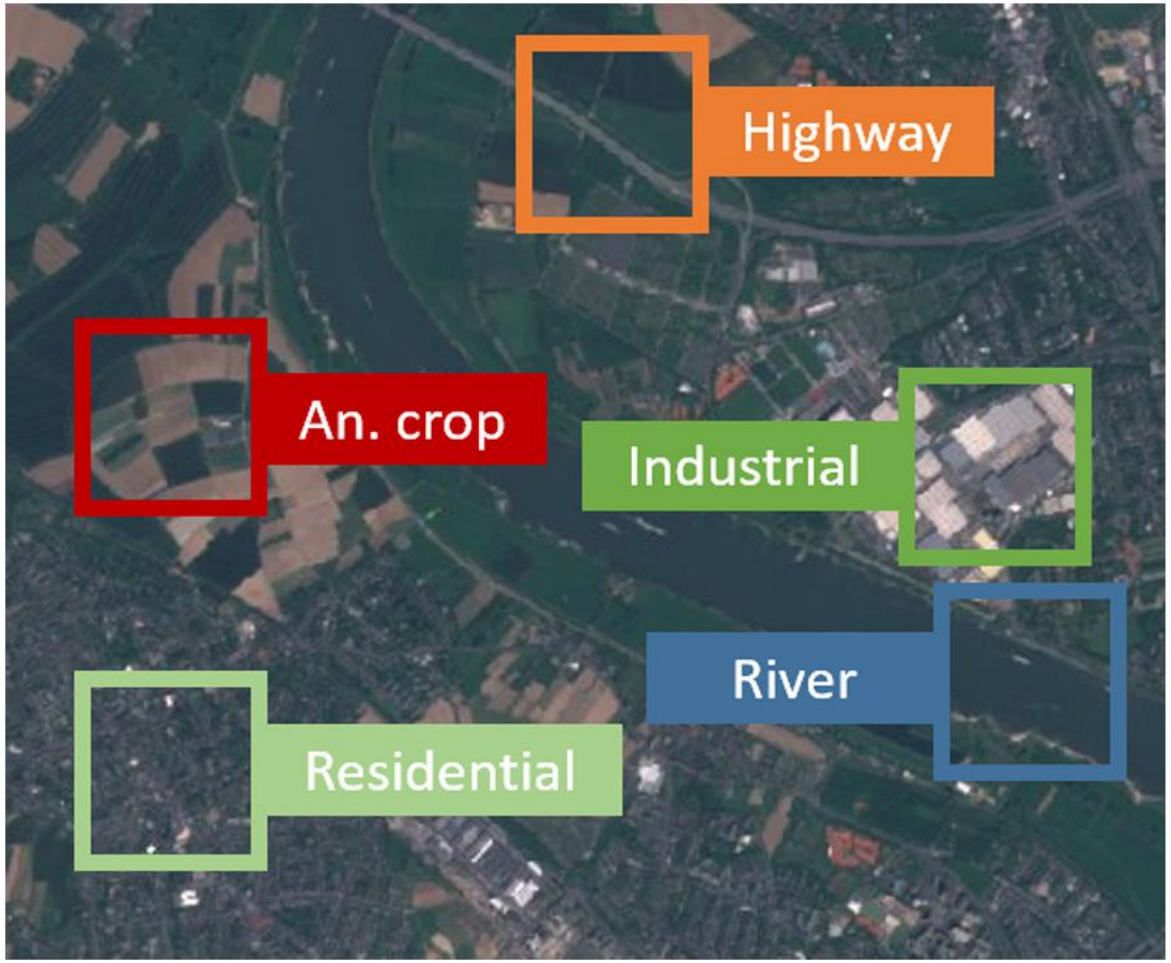
Machine Learning for Earth Observation

- Machine Learning (ML) enables us to learn patterns from Earth Observation data and automatically extract useful information.
- Computer Vision (CV) plays a central role in image-based Earth Observation
- EO-specific challenges:
  - **Multi-sensor data:** EO combines different sensors and modalities, such as optical and SAR.
  - **Label scarcity:** High-quality ground-truth data are often limited or expensive to obtain.
  - **Data imbalance:** Rare but important phenomena may be poorly represented in training data.

Different Questions, Different Tasks

Question	AI Task
What is in the image?	Image Classification
Where is each class?	Semantic Segmentation
How much?	Regression
What objects are there, and where?	Object Detection

Image Classification



- Image Classification: the model assigns a semantic class to an entire image or image patch.

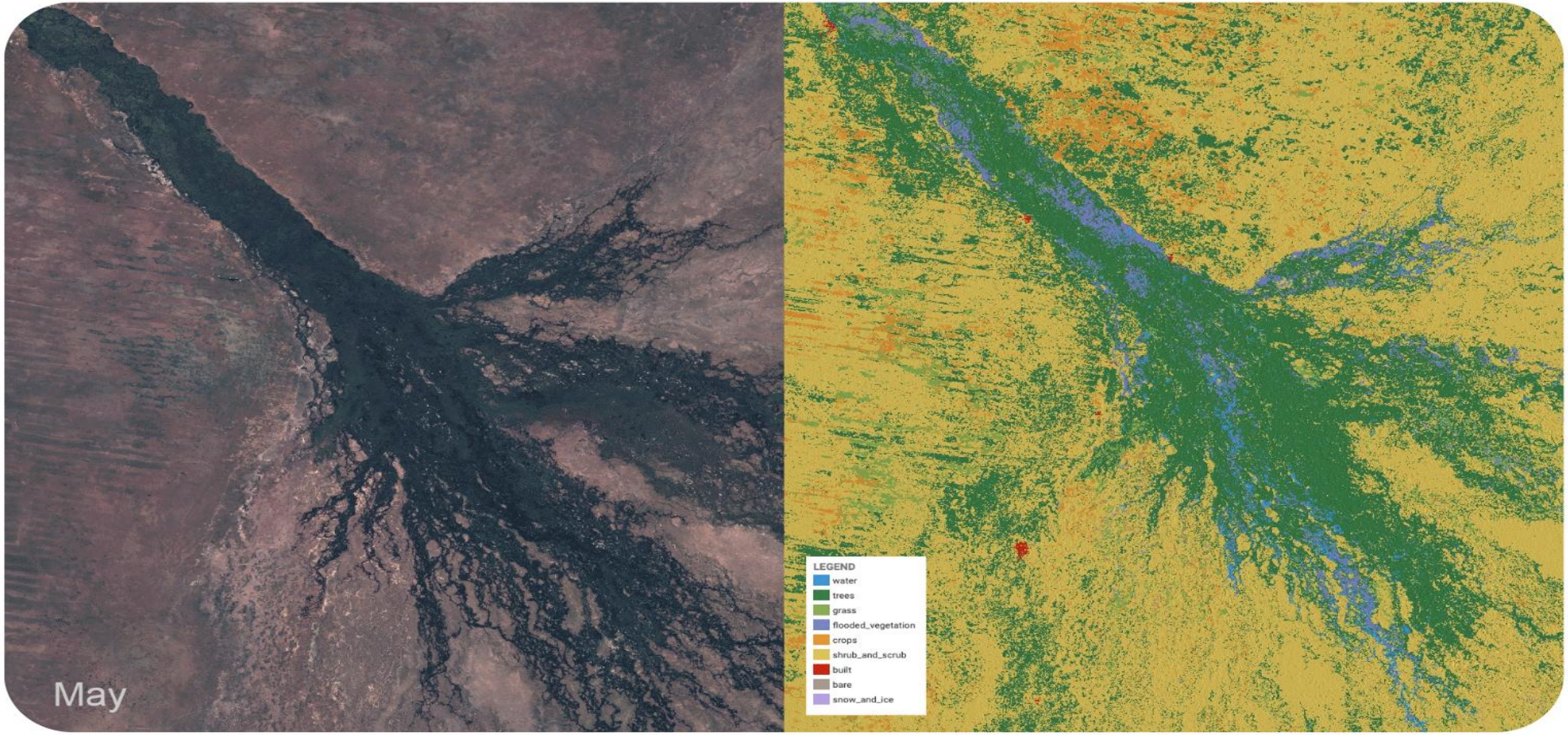
https://github.com/phelber/eurosat

Image Classification – Remote Sensing Scene Classification

- **Architecture:** *RemoteCLIP* (2023), vision-language foundation model for remote sensing.
- **Downstream Task:** Zero-shot remote sensing scene classification.
- **Input Data:** Remote sensing image (+ candidate class labels)
- **Model Prediction:** „*A busy airport with many airplanes*“

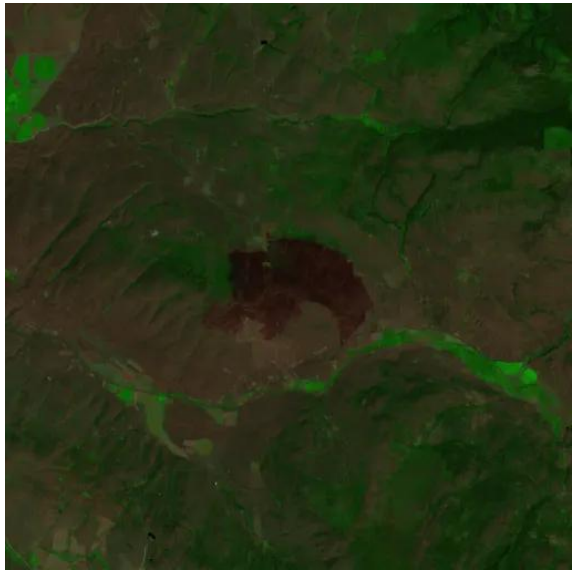
https://github.com/ChenDelong1999/RemoteCLIP

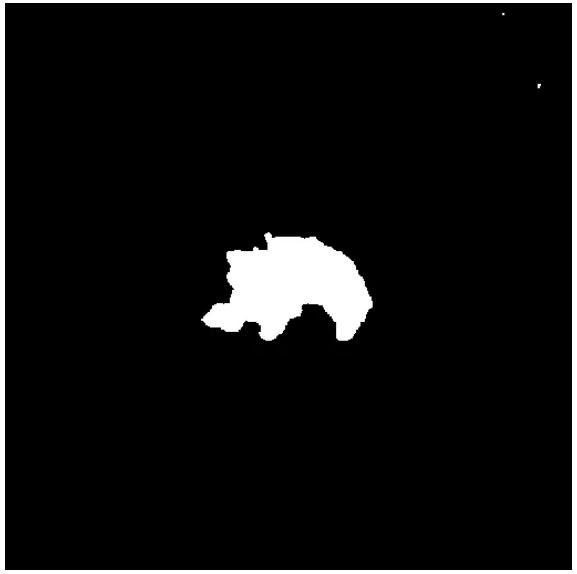
Semantic Segmentation



https://blog.google/products-and-platforms/products/earth/dynamic-world-land-cover-data/

Sematic Segmentation – Burn Scars Detection

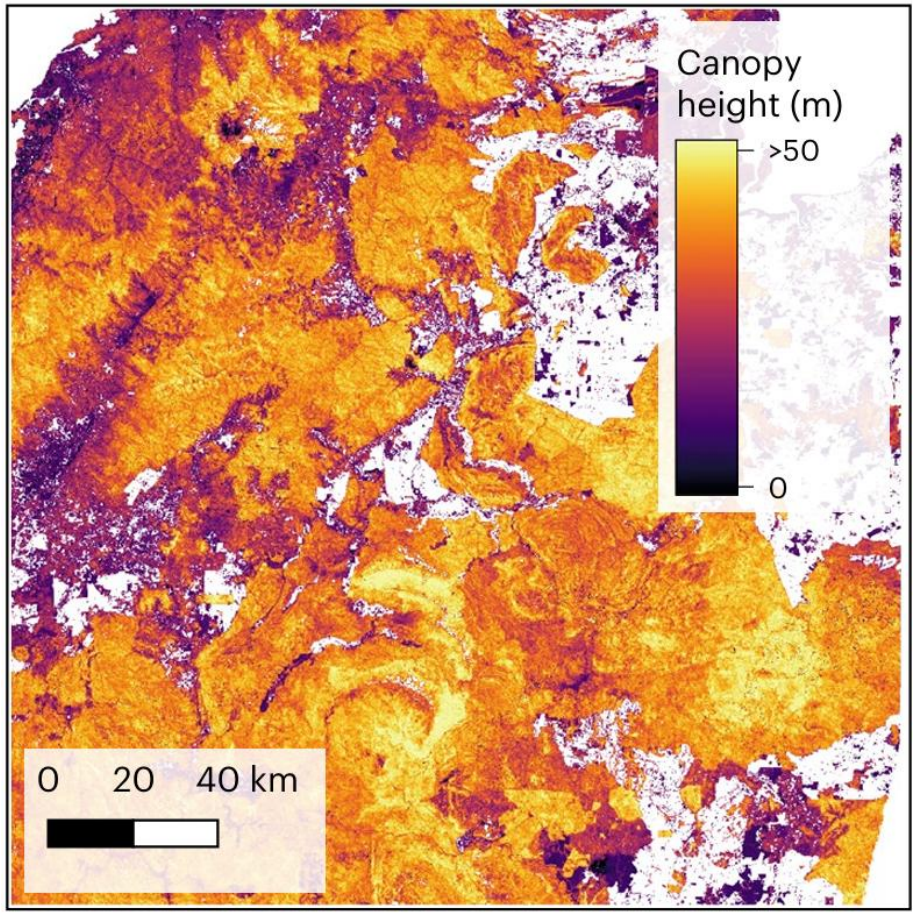




- **Architecture:** *Prithvi* (2023), GFM developed by IBM-NASA
- **Downstream Task:** Post-wildfire burn scar delineation.
- **Input Data:** False color composite (SWIR, Narrow NIR, Red) to isolate charred vegetation and penetrate smoke.
- **Model Prediction:** Binary segmentation mask (Black: No burn scar; White: Burn scar).

https://huggingface.co/spaces/ibm-nasa-geospatial/Prithvi-100M-Burn-scars-demo

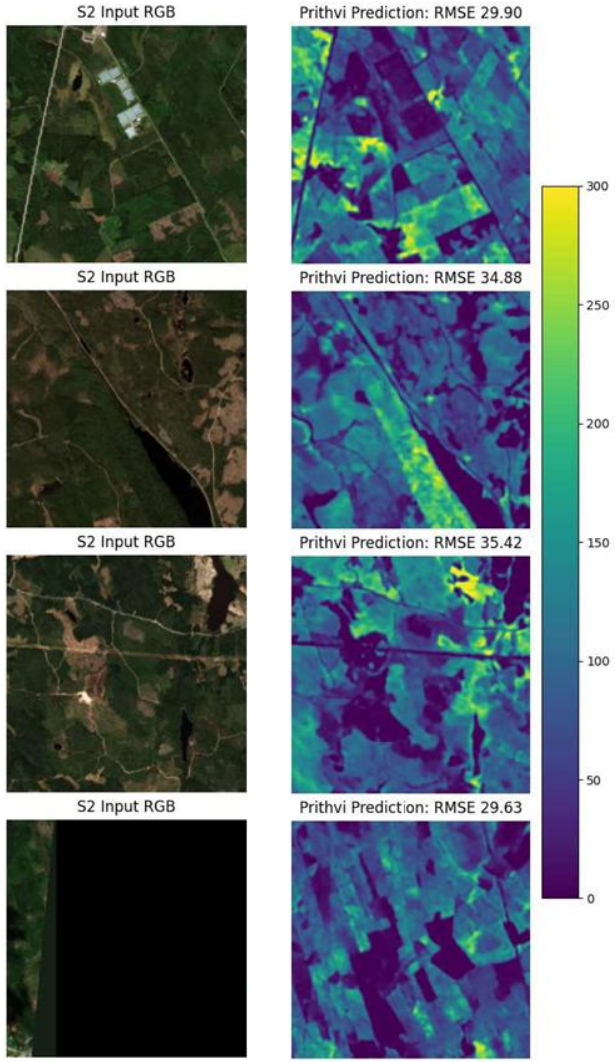
Regression



- Image Regression predicts continuous numerical values from EO data.
- Example: estimating vegetation or canopy height from satellite imagery.

https://github.com/langnico/global-canopy-height-model

Regression – Above-Ground Biomass Estimation



- **Architecture:** *Prithvi-EO-2.0-300M* (2026), GFM developed by IBM-NASA
- **Downstream Task:** Pixel-wise Above-Ground Biomass (AGB) estimation.
- **Input Data:** 12 monthly Sentinel-2 observations.
- **Model Prediction:** Continuous AGB map (metric tonnes per pixel).

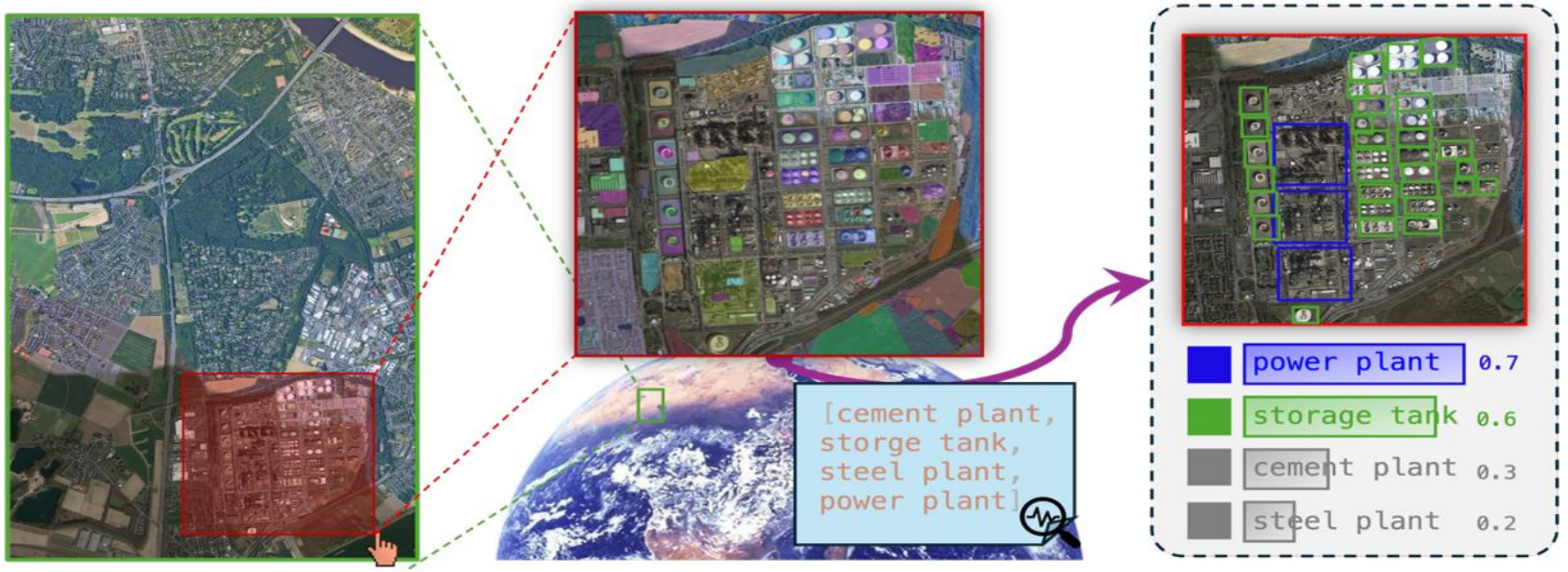
https://arxiv.org/abs/2412.02732

Object Detection

- Object Detection identifies and localizes individual objects in EO imagery.
- Example: detection and localization of individual buildings in satellite imagery.

https://xviewdataset.org

Object Detection – Open-Vocabulary Object Detection



- **Architecture:** *LAE-DINO* (2025), foundation object detector for remote sensing
- **Downstream Task:** Open-vocabulary object detection
- **Input Data:** High-resolution remote sensing imagery + object categories
- **Model Prediction:** Bounding boxes, object classes and confidence scores

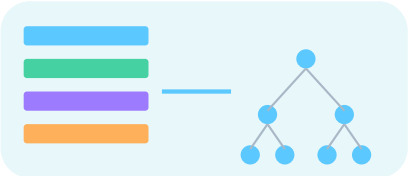
https://ojs.aaai.org/index.php/AAAI/article/view/32672

From traditional ML to modern AI4EO

ENGINEERED FEATURES → LEARNED REPRESENTATIONS → CONTEXT → REUSABLE PRETRAINED REPRESENTATIONS

01 TRADITIONAL ML

Analyst defines features

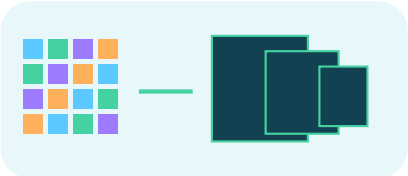


- Bands • indices • texture
- SAR backscatter • time-series stats
- Strong with smaller tabular datasets

Random Forest • SVM • XGBoost

02 DEEP LEARNING

Model learns spatial features

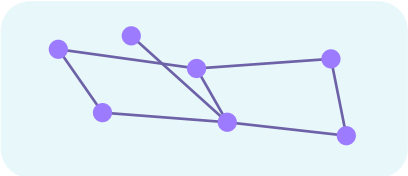


- Pixels / patches as input
- Hierarchical spatial representations
- Strong for mapping & segmentation

CNN • U-Net

03 TRANSFORMERS

Model learns wider context

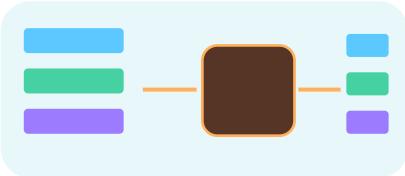


- Long-range spatial relationships
- Temporal / sequence context
- Useful for large scenes & time series

Vision / temporal transformers

04 FOUNDATION MODELS

Pretrain once, adapt many times



- Multimodal pretraining
- Reusable representations across tasks
- Less task-specific training from scratch

Fine-tune • probe • adapt

4. The Foundation Model Paradigm

From Task-Specific Models to Foundation Models

Traditional Paradigm **Dataset → Train Model → Specific Task**	**Foundation Model Paradigm** **Large Scale EO data → Pretraining → Foundation Model**
Classification → Model A	Fine tuning → Classification
Segmentation → Model B	Fine tuning → Segmentation
Regression → Model C	Fine tuning → Regression
Object-Detection → Model D	Fine tuning → Object Detection

Prithvi-EO-2.0 – A Foundation Model for EO

Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications

Daniela Szwarcman[1,†], Sujit Roy[2,3,†,‡] (Senior Member, IEEE), Paolo Fraccaro[1,†,‡], Þorsteinn Elí Gíslason[4], Benedikt Blumenstiel[1], Rinki Ghosal[3], Pedro Henrique de Oliveira[1], Joao Lucas de Sousa Almeida[1], Rocco Sedona[5], Yanghui Kang[6], Srija Chakraborty[12], Sizhe Wang[7], Carlos Gomes[1], Ankur Kumar[3], Vishal Gaur[3], Myscon Truong[8], Denys Godwin[9], Sam Khallaghi[9], Hyunho Lee[7], Chia-Yu Hsu[7], Rohit Lal[3], Ata Akbari Asanjan[12], Besart Mujeci[12], Disha Shidham[12], Rufai Omowunmi Balogun[9], Venkatesh Kolluru[3], Trevor Keenan[11], Paulo Arevalo[10], Wenwen Li[7], Hamed Alemohammad[9], Pontus Olofsson[2], Timothy Mayer[3], Christopher Hain[2], Robert Kennedy[8], Bianca Zadrozny[1], David Bell[12], Gabriele Cavallaro[4,5] (Senior Member, IEEE), Campbell Watson[1], Manil Maskey [2] (Senior Member, IEEE), Rahul Ramachandran[2], and Juan Bernabe Moreno[1]

†Equal Contribution;

***Abstract*—This paper presents Prithvi-EO-2.0, a new geospatial foundation model that offers significant improvements over its predecessor, Prithvi-EO-1.0. Trained on 4.2 million global time series samples from NASA's Harmonized Landsat and Sentinel-2 data archive at 30-m resolution, the new model incorporates temporal and location embeddings for enhanced performance across various geospatial tasks. Through extensive benchmarking with GEO-Bench, the model outperforms the previous Prithvi-EO model by 8% across a range of tasks. It also outperforms six other geospatial foundation models when benchmarked on remote sensing tasks from different domains and resolutions (i.e. from 0.1 m to 15 m). The results demonstrate the versatility of the model in both classical Earth observation and high-resolution applications. Early involvement of end-users and subject matter experts (SMEs) allowed constant feedback on model and dataset design, enabling customization across diverse SME-led applications in disaster response, land cover and crop mapping, and ecosystem dynamics monitoring. Prithvi-EO-2.0 is available as an open-source model on Hugging Face and IBM TerraTorch, with additional resources on GitHub. The project exemplifies the Trusted Open Science approach embraced by all involved organizations.**

[1]IBM Research (UK, Ireland, Brazil, Zurich, and USA); [2]NASA Marshall Space Flight Center, Huntsville, AL, USA; [3]Earth System Science Center, University of Alabama in Huntsville, AL, USA; [4]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland; [5]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany; [6]Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA, USA; [7]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA; [8]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA; [9]Center for Geospatial Analytics, Clark University, Worcester, MA, USA; [10]Department of Earth and Environment, Boston University, Boston, MA, USA; [11]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA; [12]Earth from Space Institute, Universities Space Research Association, USA

‡ Corresponding author: sujit.roy@nasa.gov, paolo.fraccaro@ibm.com

Model available at: https://huggingface.co/ibm-nasa-geospatial/Prithvi-EO-2.0. Code available at: https://github.com/NASA-IMPACT/Prithvi-EO-2.0

I. Introduction

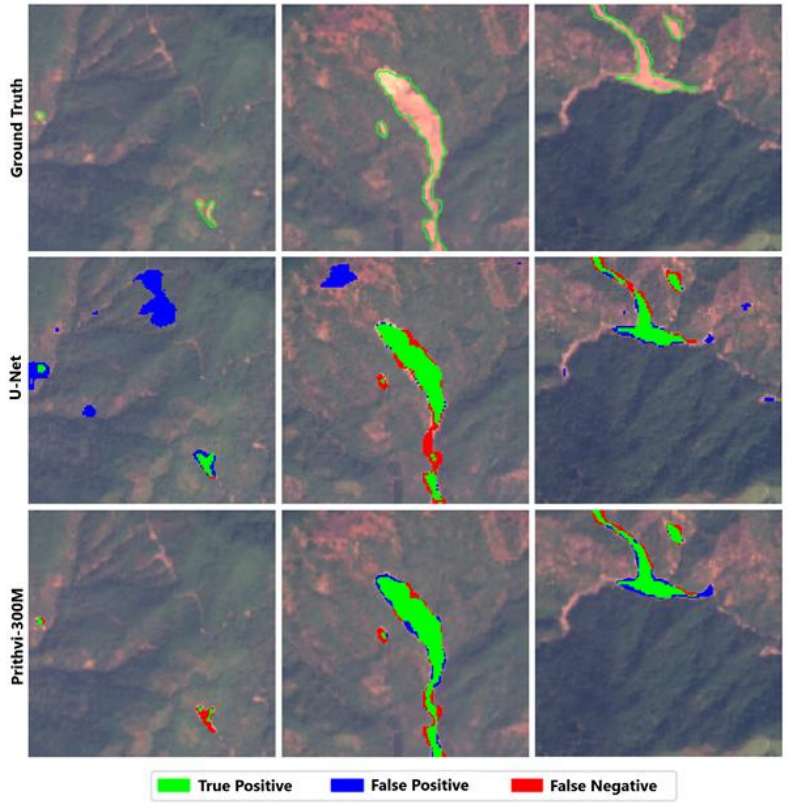
In recent years, Earth Observation (EO) has entered a new era with the rise of Geospatial Foundation Models (GFMs)—large-scale artificial intelligence (AI) systems trained on vast amounts of unlabeled satellite imagery using self-supervised learning [1]. Foundation models are general-purpose neural networks (often based on transformer architectures) that learn broad representations from massive datasets and can be fine-tuned for many downstream tasks with minimal labeled data [2]. These models can help address long-standing challenges in EO by reducing the need for manually labeled samples, which are usually hard to obtain at scale [1]. Remote sensing is especially well-suited for foundation model development due to its global coverage, frequent revisits, and the sheer scale of unstructured imagery. Once pretrained, GFMs require less data to achieve similar or even superior performance across various domains [3], [4]. Despite their potential, real-world adoption remains limited.

We identify three main limitations with available GFMs. First, although EO data is inherently multi-temporal, most GFMs do not account for this characteristic. Those that do either process only point data or focus on long time series restricted to small patches or individual pixels [4], [5]. Second, in-depth validation considering diverse types of tasks and clear comparison protocols remains limited. This hinders users' ability to assess whether the models are suitable for their use case. Third, adapting state-of-the-art GFMs to different applications may require AI expertise in the absence of the proper tools and guidance. While several GFMs have released the weights and model architecture [3], [4], [6]–[8], which is an important step toward community adoption, we believe that the lack of clear instructions or a streamlined code base for fine-tuning remains a significant barrier to broader use and further evaluation of GFMs.

To address these issues and increase the impact of GFMs

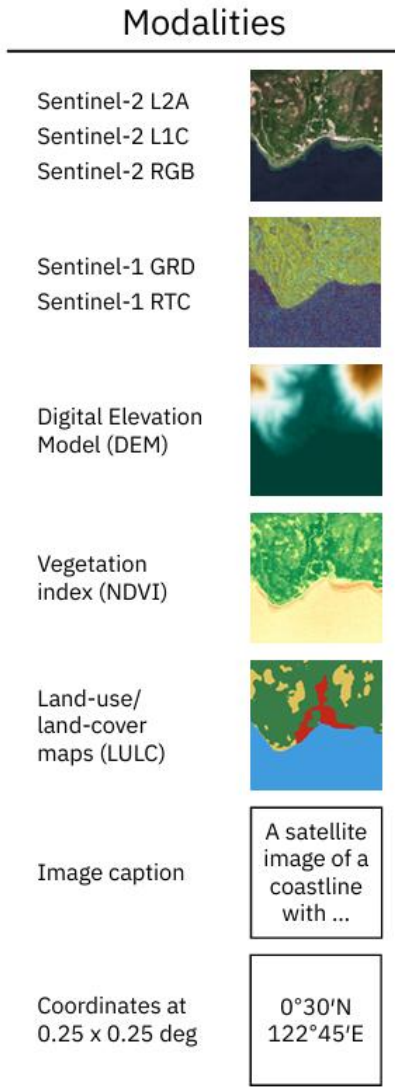
- Multi-Temporal EO Foundation Model
- Pretraining data: Harmonized Landsat Sentinel-2 (HLS)
- Pretraining dataset: 4.2M global time-series samples
- Model sizes: 300M & 600M parameters
- Designed for adaptation across diverse EO tasks

Prithvi-EO-2.0 – Adapting to Downstream Tasks



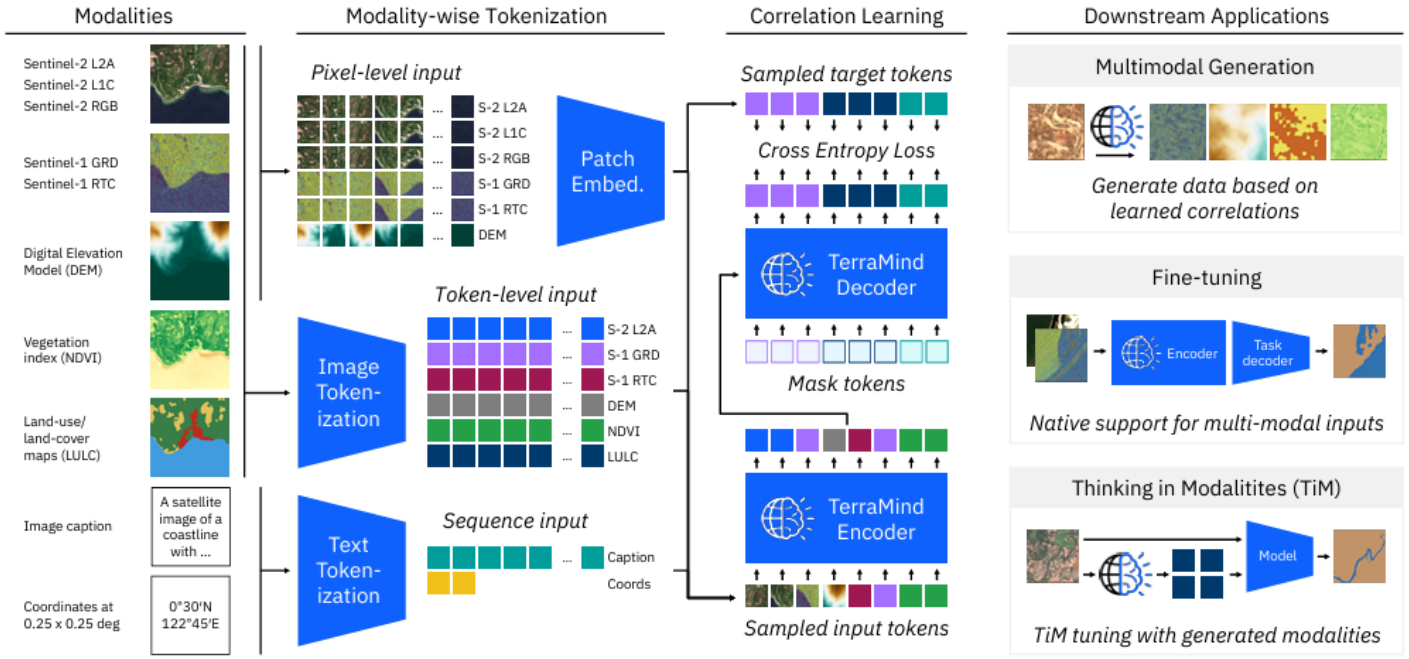
- Downstream task: Landslide segmentation
- Dataset: Landslide4Sense (L4S)
- Input: 6-band optical imagery
- Output: pixel-level landslide mask
- Adaptation: Fine Tuning

Earth Observation is Multimodal



- Modalities:
  - Optical imagery
  - SAR imagery
  - Elevation
  - Land cover & vegetation
  - Language
  - Geolocation
- Different modalities provide complementary information about the same Earth System.

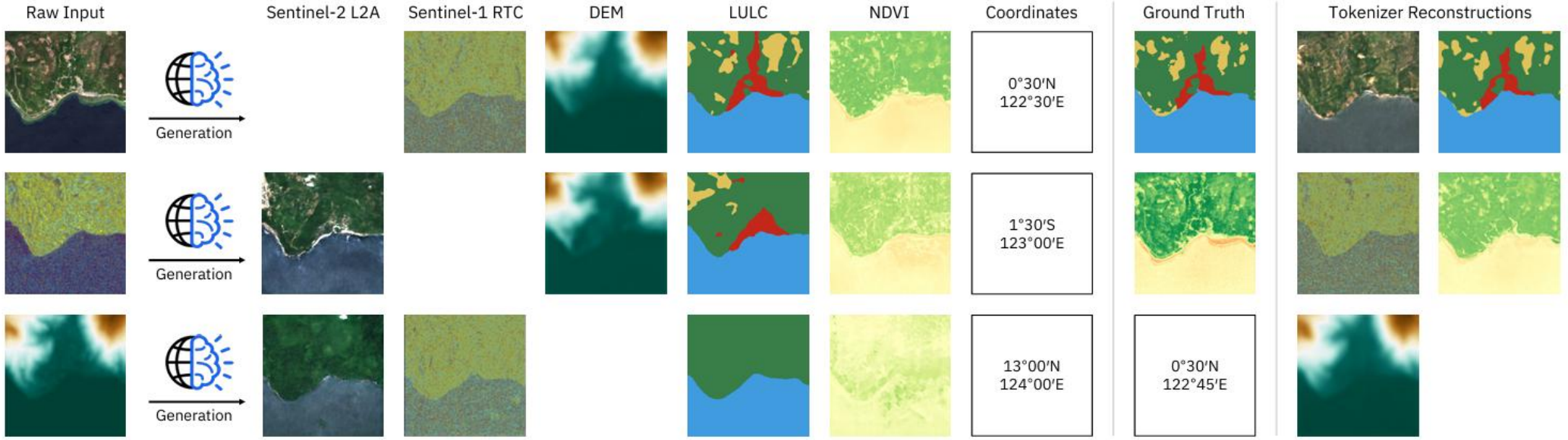
TerraMind: Generative Multimodality for Earth Observation



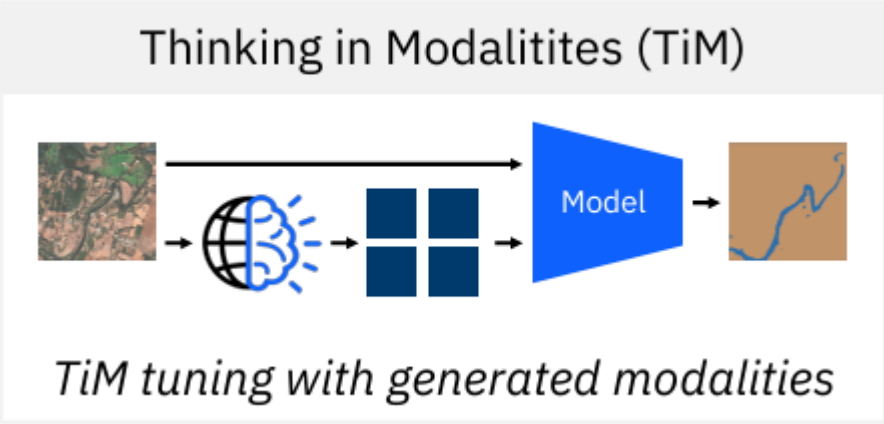
- Any-to-any generative multimodal EO model
- Pretrained on 500B tokens
- Pixel-level + token-level representations
- Supports multimodal generation and downstream adaptation

TerraMind: Any-to-Any Generation

- Given one modality, TerraMind can generate other EO modalities using the cross-modal relationships learned during pretraining.



TerraMind: Thinking in Modalities (TiM)



Observed Modality → Generate an additional modality
→ Observed + generate modalities → Downstream task

- TerraMind can generate missing modalities and use them as additional information for downstream tasks.

TerraMind: Thinking in Modalities (TiM)

- Generated modalities can provide useful contextual information for downstream tasks.

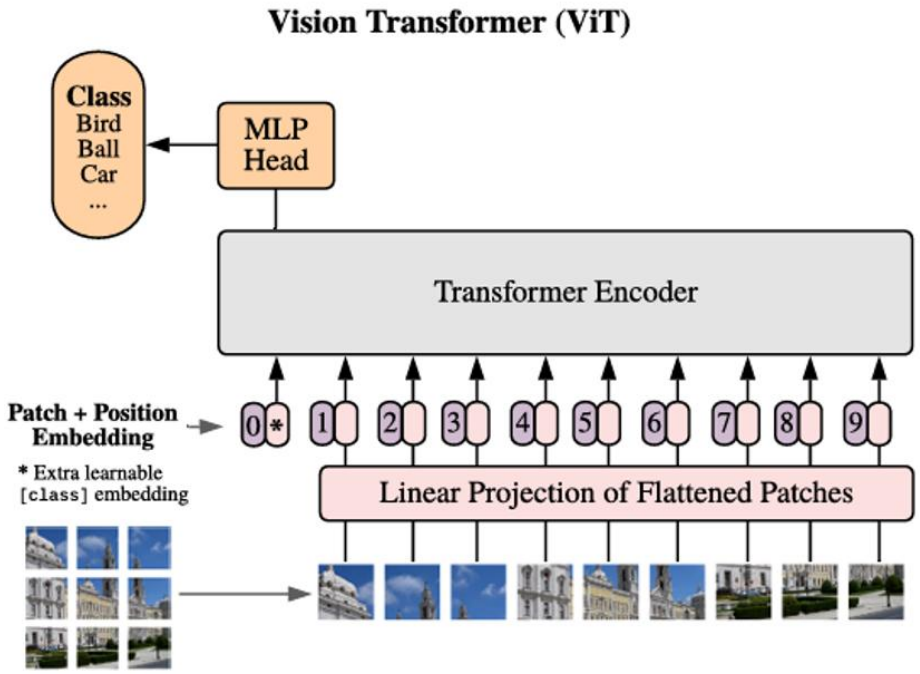
	Input	mIoU
TerraMindv1-B	S-2	41.87
TerraMindv1-B TiM	S-2 + *gen. LULC*	**42.74**

Table 13. Thinking-in-modalities (TiM) tuning compared with standard full fine-tuning approaches on the SA Crop dataset.

- Downstream task: crop type segmentation
- Input: sentinel-2 imagery
- Generated modality: LULC
- TiM: Sentinel-2 + generated LULC
- Result: +0.87 mIoU

5. When the Vision Meets Language

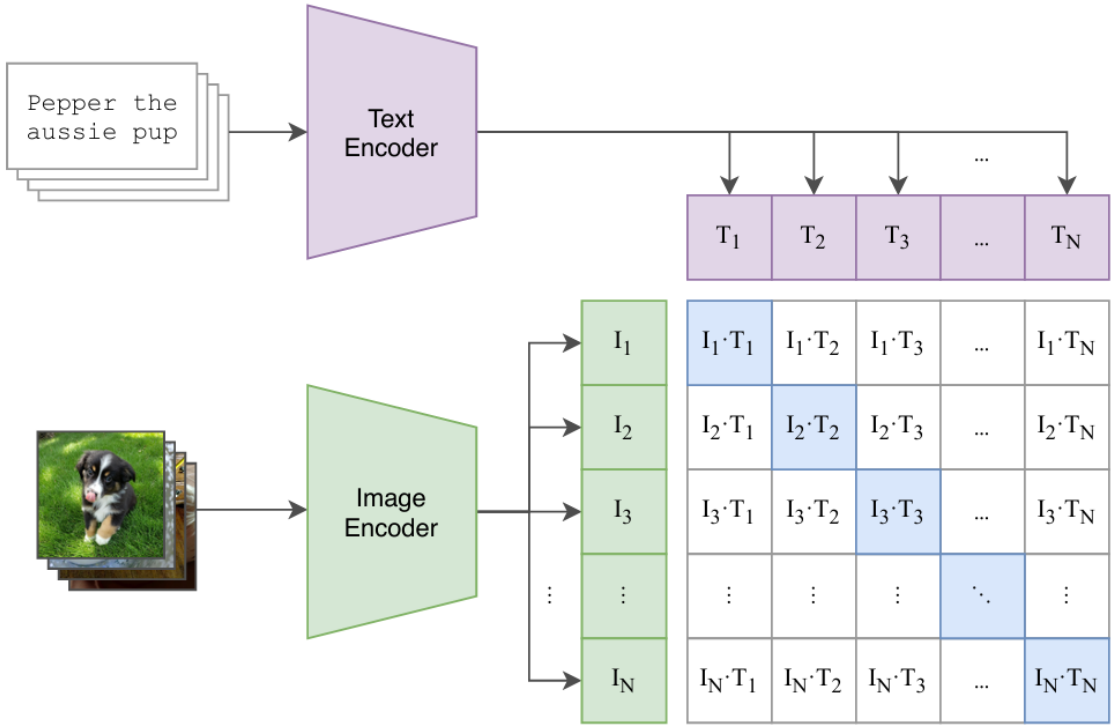
Images as Tokens



- Split the image into fixed-size patches
- Map each patch into a vector embedding
- Process the resulting sequence with a Transformer
- *Image patches can be treated as visual tokens!*

https://arxiv.org/pdf/2010.11929

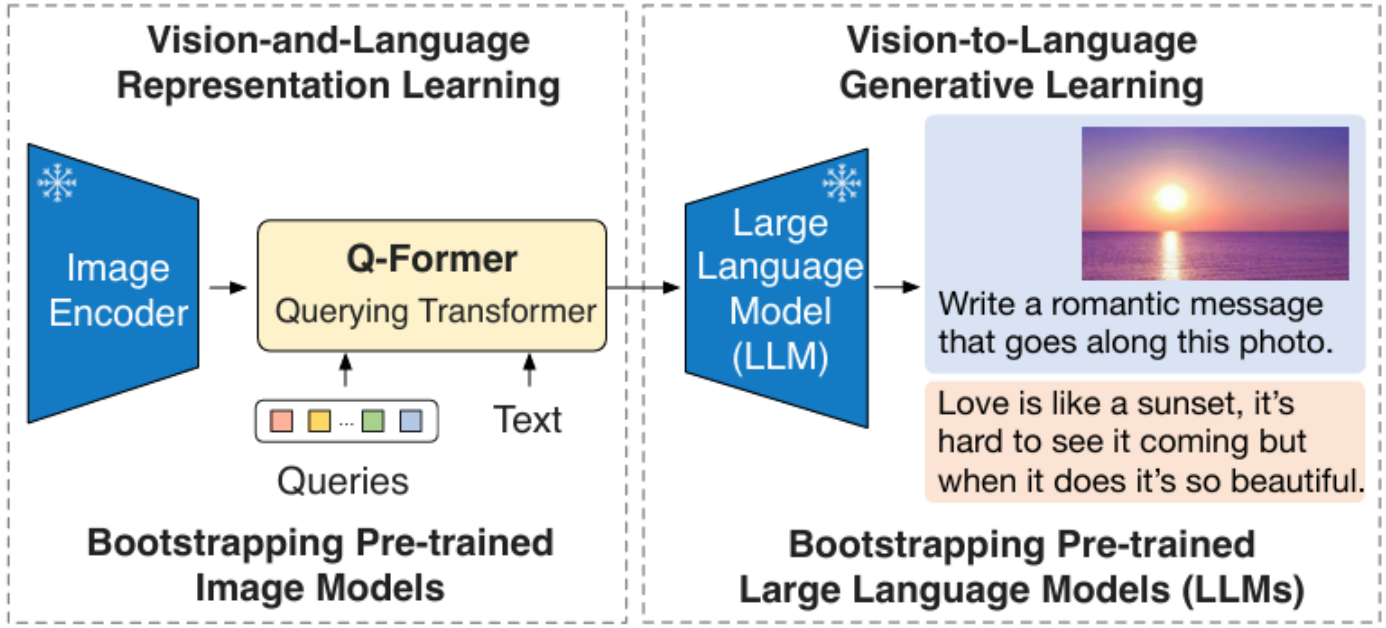
Aligning Vision and Language



- Encode images and text separately
- Contrastive learning
- Bring matching image-text (I_i,T_i) pairs closer
- Push non-matching pairs (I_i,T_j), with i≠j, apart
- *Vision and language are aligned in a shared embedding space*!

https://arxiv.org/pdf/2103.00020

Bridging Vision and Language Models



- Vision encoders produce visual representations.
- LLMs operate in a language embedding space.
- A learnable connector bridges the two modalities.
- Two-stage training
- *The connector translates visual information into representations the language model can use!*

https://arxiv.org/pdf/2301.12597

Visual Instruction Tuning

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

- Train on diverse image–instruction–response examples.
- Teach the model to follow different visual instructions.
- Move toward general-purpose visual assistants.
- *Natural-language instructions define the task! No task-specific output head is required.*

https://arxiv.org/pdf/2304.08485

6. Understanding LLaVA

LLaVA: Large Language and Vision Assistant

Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has been shown to improve zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. We present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: **L**arge **L**anguage **a**nd **V**ision **A**ssistant, an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general-purpose visual and language understanding. To facilitate future research on visual instruction following, we construct two evaluation benchmarks with diverse and challenging application-oriented tasks. Our experiments show that LLaVA demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model, and code publicly available.

1 Introduction

Humans interact with the world through many channels such as vision and language, as each individual channel has a unique advantage in representing and communicating certain concepts, and thus facilitates a better understanding of the world. One of the core aspirations in artificial intelligence is to develop a general-purpose assistant that can effectively follow multi-modal vision-and-language instructions, aligned with human intent to complete various real-world tasks in the wild [4, 27, 26].

To this end, the community has witnessed an emergent interest in developing language-augmented foundation vision models [27, 16], with strong capabilities in open-world visual understanding such as classification [40, 21, 57, 54, 39], detection [29, 62, 33], segmentation [25, 63, 58] and captioning [50, 28], as well as visual generation and editing [42, 43, 56, 15, 44, 30]. We refer readers to the *Computer Vision in the Wild* reading list for a more up-to-date literature compilation [12]. In this line of work, each task is solved independently by one single large vision model, with the task instruction implicitly considered in the model design. Further, language is only utilized to describe the image content. While this allows language to play an important role in mapping visual signals to language semantics—a common channel for human communication, it leads to models that usually have a fixed interface with limited interactivity and adaptability to the user's instructions.

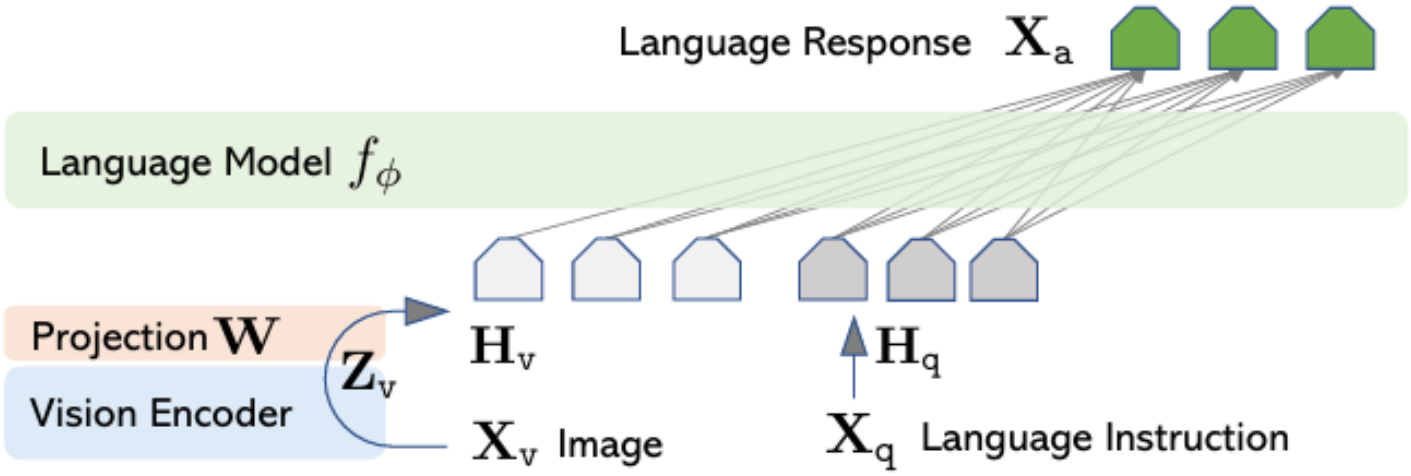
Large language models (LLM), on the other hand, have shown that language can play a wider role: a universal interface for a general-purpose assistant, where various task instructions can be explicitly represented in language and guide the end-to-end trained neural assistant to switch to the task of interest to solve it. For example, the recent success of ChatGPT [35] and GPT-4 [36] have demonstrated the power of aligned LLMs in following human instructions, and have stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [49] is an open-source LLM that matches the performance of GPT-3. Alpaca [48], Vicuna [9], GPT-4-LLM [38]

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

- Connects a pretrained vision encoder with a Large Language Model.
- Uses a simple learnable projection between vision and language.
- Trained with visual instruction tuning.
- Goal: *build a general-purpose visual assistant!*

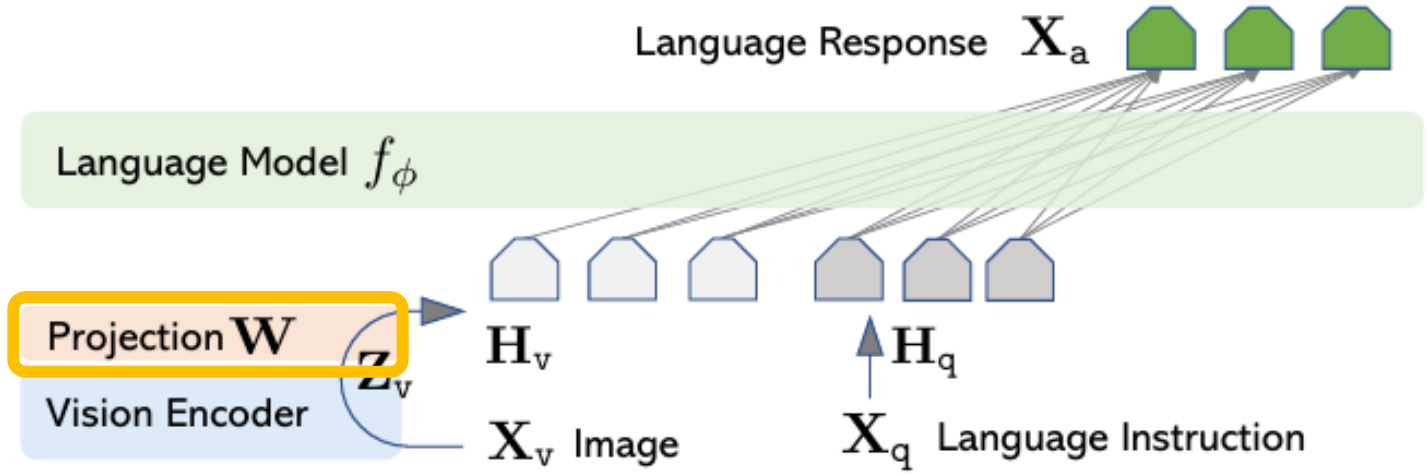
https://arxiv.org/pdf/2304.08485

Inside LLaVA: Connecting Vision and Language



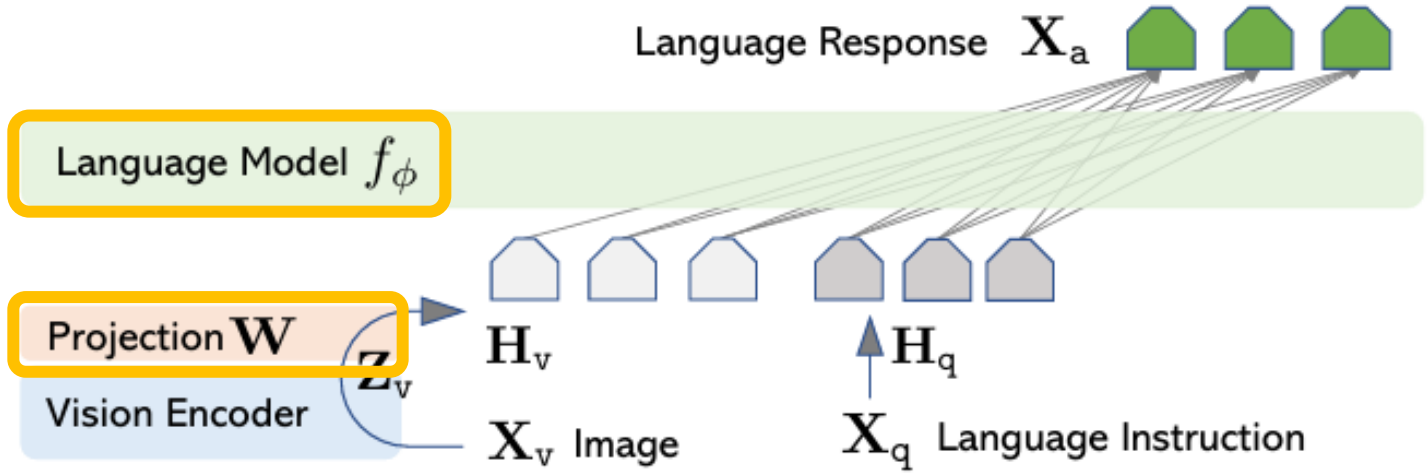
- Vision encoder: CLIP ViT-L/14
- Connector: Linear projection ***W***
- Language: Vicunia
- Output: Autoregressive text response

Training LLaVA: Stage 1, Feature Alignment



- Goal: align visual features with the LLM embedding space
- Vision encoder: Frozen
- Language model: Frozen
- Projection ***W***: Trainable.
- *Stage 1 learns the vision-to-language connector!*

Training LLaVA: Stage 2, Visual Instruction Tuning



- Goal: learn to follow multimodal instructions
- Training data: Image–instruction–response examples.
- Vision encoder: Frozen
- Projection ***W***: Trainable
- Language model: Trainable

LLaVA in Action: Visual Reasoning

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

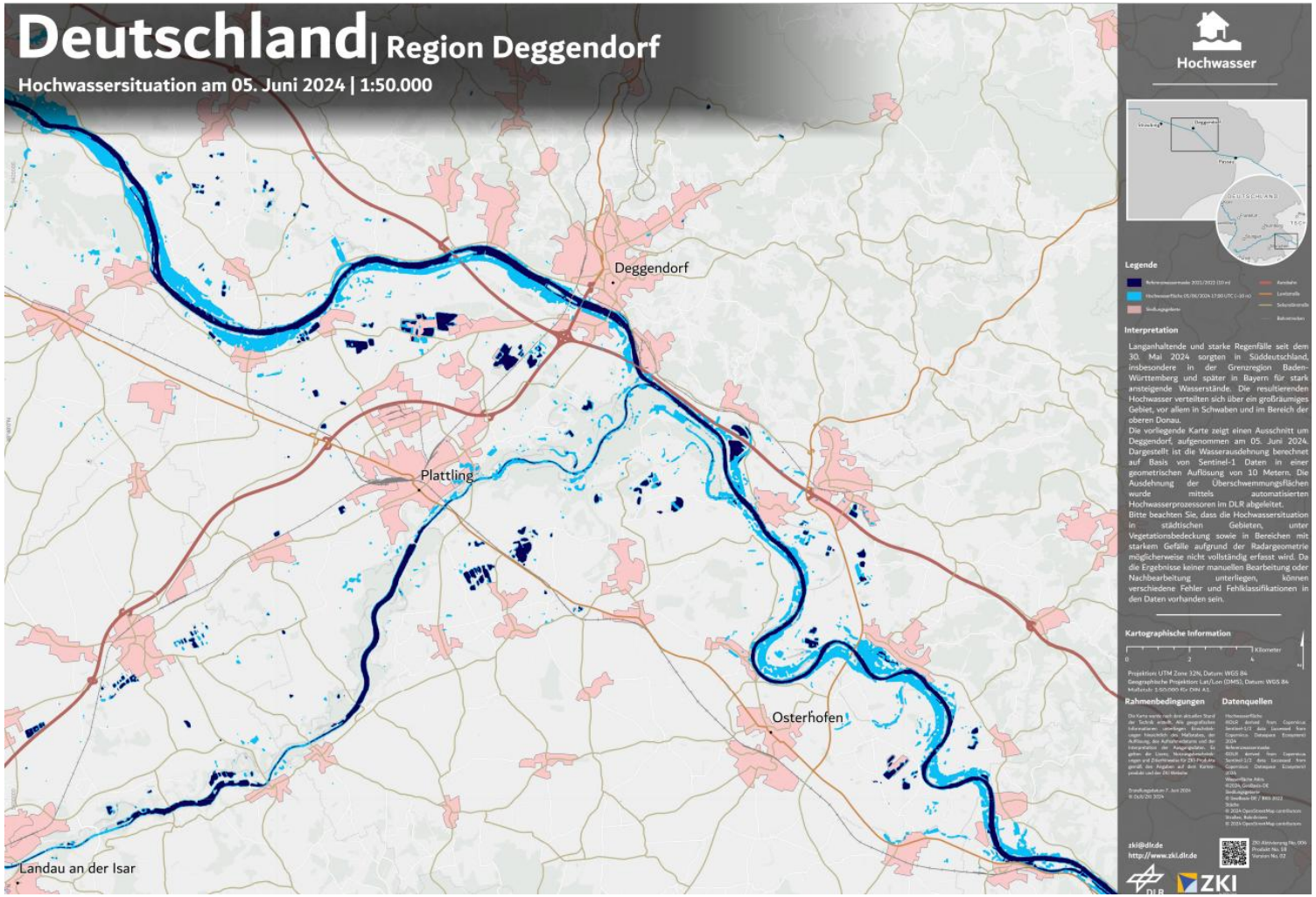
From LLaVA to Earth Observation

- Different viewpoint and spatial scale
- Different spectral and sensor characteristics
- Domain-specific objects and spatial patterns
- Specialized terminology and interpretation

7. Vision-Language Models for Disaster Response

DLR

From Crisis Mapping to Rapid Interpretation



- Detailed & validated
- Expert-driven
- Actionable products
- *Goal: not replacing expert crisis mapping, but complementing it with rapid, interactive information!*

Interactive Disaster Understanding

DLR aerial imagery — Deggendorf, 4 June 2024 | Source: DLR/ZKI

Can General-Purpose VLMs Understanding Floods from Space?

Copernicus Sentinel-2 image of severe flooding around Porto Alegre, Rio Grande do Sul, Brazil, 6 May 2024, Credit: European Union, Copernicus Sentinel-2 imagery

8. Case Study: LLaVA for Flood Understanding

Can a General-Purpose VLM Recognize a Flood?

LLaVA
General-purpose VLM

"The image you've provided appears to be a satellite view of a landscape, possibly a region with agricultural fields and a body of water. The majority of the image shows agricultural fields, with a grid-like pattern of plots that are likely used for growing crops. The fields are divided into sections, which is typical for organized farming practices. There is a large body of water that meanders through the landscape. It appears to be a river or a canal, given its winding path and the presence of what looks like a bank or a shoreline."

https://huggingface.co/llava-hf/llama3-llava-next-8b-hf

Can a General-Purpose VLM Recognize a Flood?

LLaVA
General-purpose VLM

"The image you've provided appears to be a satellite view of a landscape, possibly a region with agricultural fields and a body of water. The majority of the image shows agricultural fields, with a grid-like pattern of plots that are likely used for growing crops. The fields are divided into sections, which is typical for organized farming practices. There is a large body of water that meanders through the landscape. It appears to be a river or a canal, given its winding path and the presence of what looks like a bank or a shoreline."

No flood is mentioned

https://huggingface.co/llava-hf/llama3-llava-next-8b-hf

Introducing Flood-LLaVA

Flood-LLaVA
Flood-specialized VLM

"The satellite image reveals a coastal landscape characterized by a series of islands surrounded by expansive turquoise waters. The islands exhibit varied vegetation patterns, with some areas appearing densely vegetated and others more sparsely covered. The water has inundated a portion of the land, particularly along the edges of the islands. This flooding appears to be concentrated in areas where the land meets the water, suggesting a recent high-water event. The presence of the flooded areas indicates a potential impact on the local ecosystem and any human infrastructure that may be present"

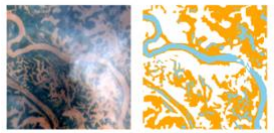
- Flood-LLaVA is a domain-adapted Vision-Language Model for flood detection and analysis from Sentinel-2 RGB imagery.
- Base model: LLaVA-NeXT-8B.
- Fine Tuning on WorldFlood data.
- Goal: Generate flood-aware descriptions and support visual dialogue.

Building a Flood-Aware Vision-Language Dataset

WorldFloods Dataset
Sentinel-2 imagery of flooded and non-flooded scenarios

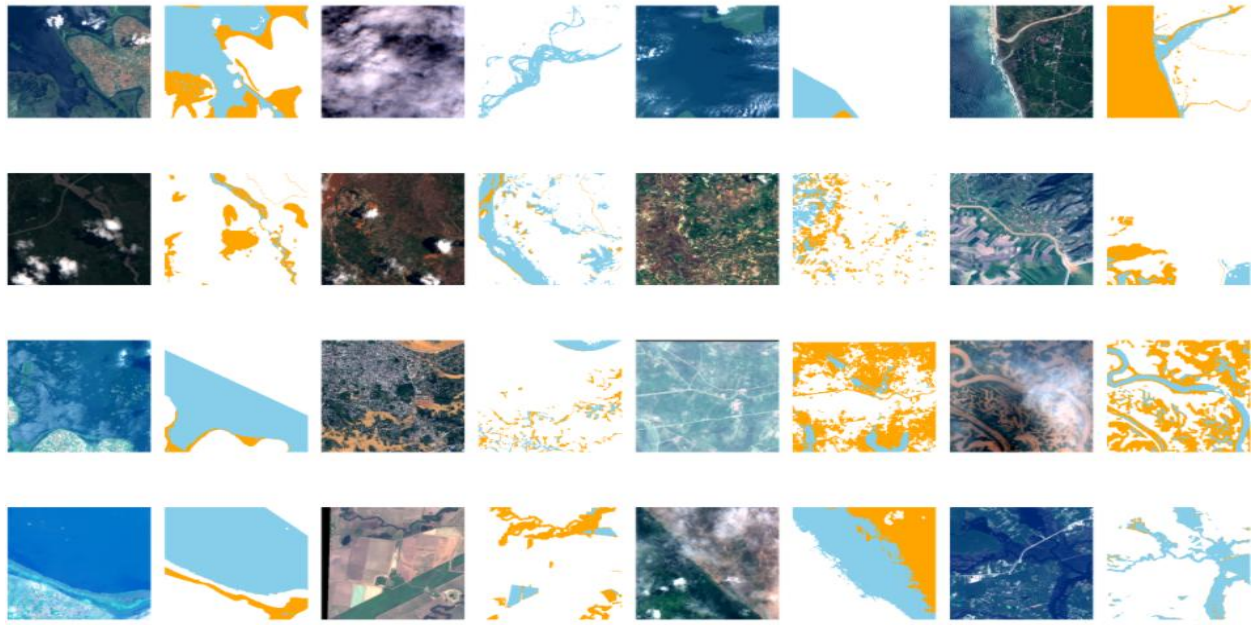
→

Sentinel-2 RGB Tiles + Flooded Masks



→

2,300 Selected Tiles
Flooded + Non-Flooded Scenarios, Automatic + Manual Quality Filtering



https://spaceml-org.github.io/ml4floods/content/worldfloods_dataset.html

From Flood Masks to Language

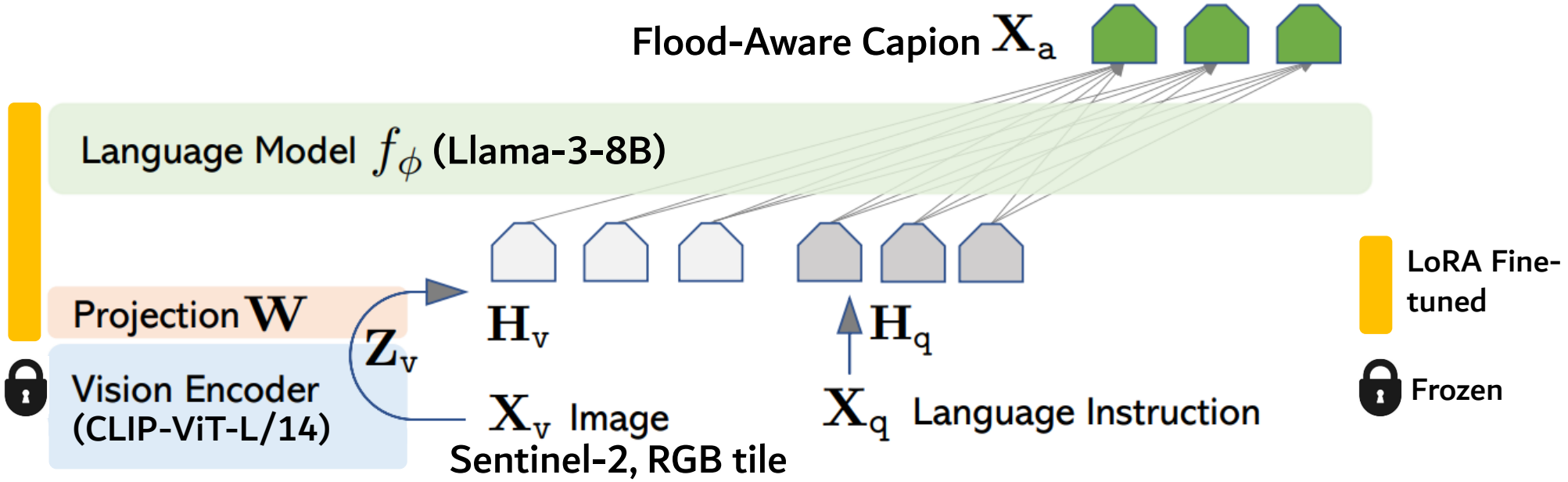
Sentinel-2 RGB + 3-Class Flood Mask → Gemini 2.0 *Expert Annotator* → Flood-Aware Caption

- RGB imagery and semantic masks are jointly provided to the annotator
- Semantic alignment propmpting generates technical description of the scene and flood impact

The satellite image reveals a coastal region characterized by dense vegetation and a network of waterways. A cluster of buildings sits along the coastline, connected by roads. A river flows from the interior toward the sea, and has noticeably expanded beyond its normal channel, inundating a narrow band of land adjacent to the coast. The floodwaters appear to be encroaching on the settlement and surrounding vegetated areas, indicating a potential risk to the local community. The waters extend along the coastline where the river meets the sea.

Flood-LLaVA: Domain Adaptation

- LLaVA-NeXT-8B adapted to the flood domain
- CLIP ViT-L/14 visual encoder + Llama-3-8B language backbone
- Parameter-efficient fine-tuning with LoRA adapters.



Mask-Guided Instruction Tuning

- **Training Phase** Prompts include *technical notes* derived from ground-truth masks (e.g. „*flood detected in 15% of area*“).
- **Inference Pahse** Technical notes are removed; the model receive RGB imagery only.
- **Result** The model learns to associate RGB patterns with flood events, enabling autononous flood recognition at inference time.

Does Fine-Tuning Works?

Models/Metrics	BLEU-4	ROUGE-L	METEOR	Semantic Similarity
LLaVa-Next-8B	0.0229	0.1898	0.2679	0.6848
Flood-LLaVA-8B (Fine Tuned)	**0.1612**	**0.3412**	**0.4044**	**0.8654**
Improvement (Δ%)	+603.9%	+79.8%	+50,9%	+26.4%

- **Takeaway** Domain adaptation substantially improves flood-specific terminology and semantic alignment.

Talking to Flood-LLaVA

llava-flood-model

The image you've provided appears to be a satellite or aerial view of a landscape that has been affected by flooding. The river has overflowed its banks, causing significant inundation of the surrounding agricultural fields. The water has turned the fields into a large, muddy expanse, which is typical of floodwaters. The area looks like it has been heavily impacted, with the water reaching far beyond the river's usual course.

There are several structures visible, including buildings that could be homes or farm buildings, and roads that are likely used for transportation. The flooding has likely disrupted the normal activities of the people living and working in this area, and it may have caused damage to crops and infrastructure. The image is a stark reminder of the power of natural

are there any flooded building?

llava-flood-model

Yes, there are buildings that appear to be flooded in the image. The water has risen to the level of the roofs, indicating that the floodwaters have entered the structures. This can cause significant damage to the buildings and their contents, as well as pose a risk to the safety of the people who live or work there.

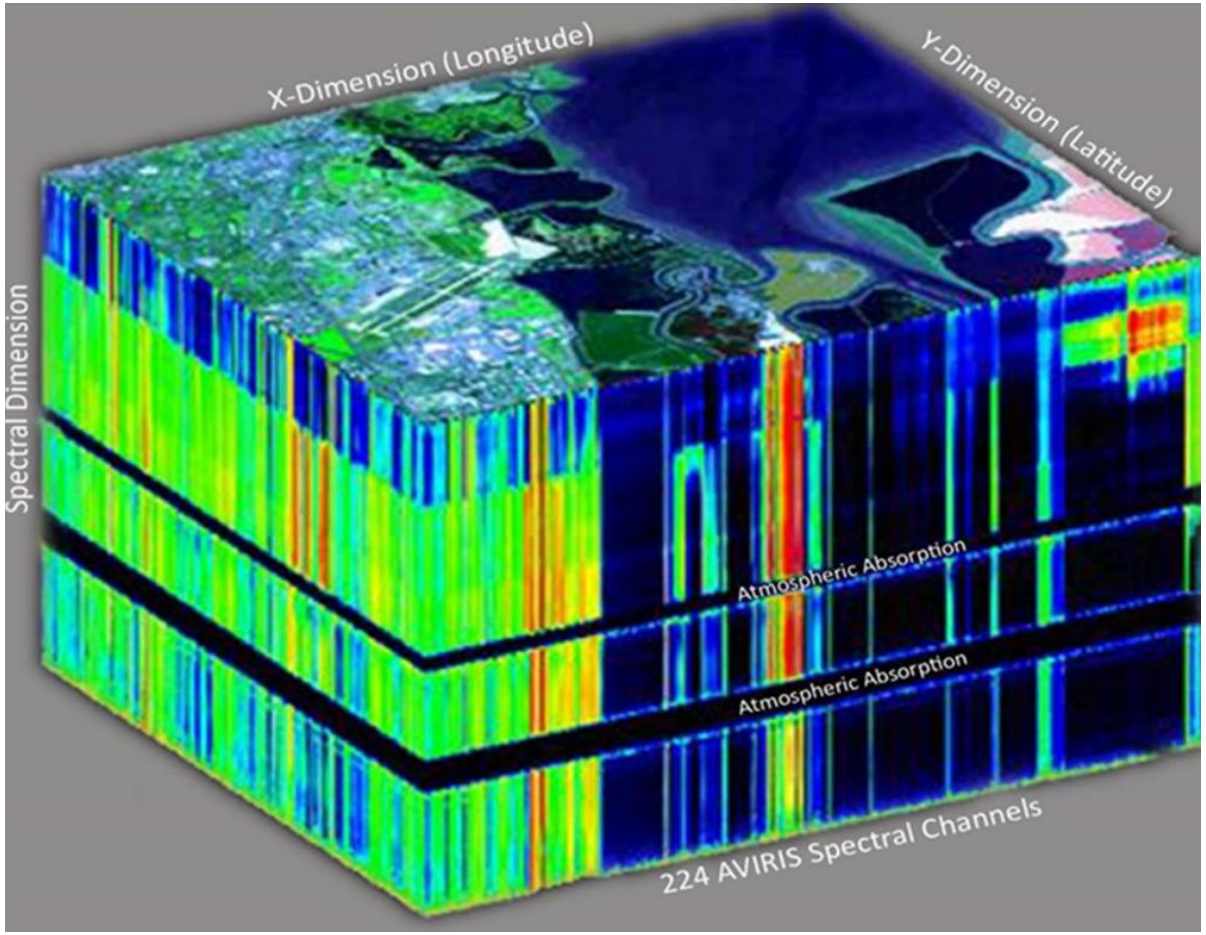
in which part of the image?

llava-flood-model

The flooded buildings are located in the lower right corner of the image. They are surrounded by the flooded fields and are part of the larger area that has been inundated by the river's overflow.

9. Challenges & Future Directions

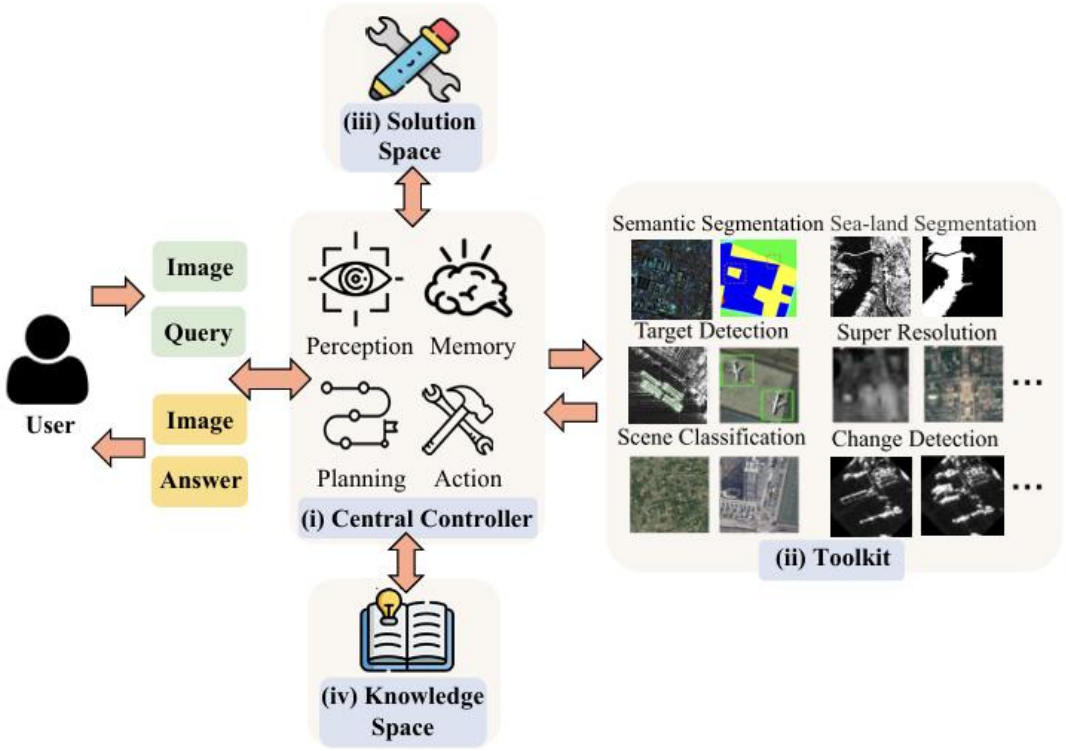
Beyond RGB: Hyperspectral Vision-Language Models



EnMAP hyperspectral imagery, 224 spectral bands, Source: DLR / EnMAP

- RGB 3 spectral bands →Hyperspectral, hundreds of narrow spectral bands
- Can vision-Language Models learn to understand spectral information beyond RGB?
- Ongoing work: adapting VLMs to hyperspectral Earth Observation imagery

Beyond a Single Model: Agentic EO



- Planning & reasoning
- Memory
- Specialized EO tools
- Multi-step execution.
- *From a single model to a system that can orchestrate specialized capabilities!*

https://arxiv.org/pdf/2406.07089

10. Conclusions & Take-Home Messages

Take-Home Messages

Earth Observation
→ Task-Specific AI
→ Foundation Models
→ Vision-Language Models
→ Domain-Specifc VLMs
→ Agentic EO

- AI for EO is moving from task-specific models to reusable foundation models.
- Vision-language models bring natural-language interaction to EO, but domain adaptation matters.
- Future EO systems may combine richer observations, specialized models and intelligent agents.
- *From pixels to language, and from individual models to intelligent EO systems!*

Thank You!

Questions?

From Pixels to Language:
The Evolution of AI for Earth Observation

Luca Chiarabini
German Aerospace Center (DLR)
luca.chiarabini@dlr.de