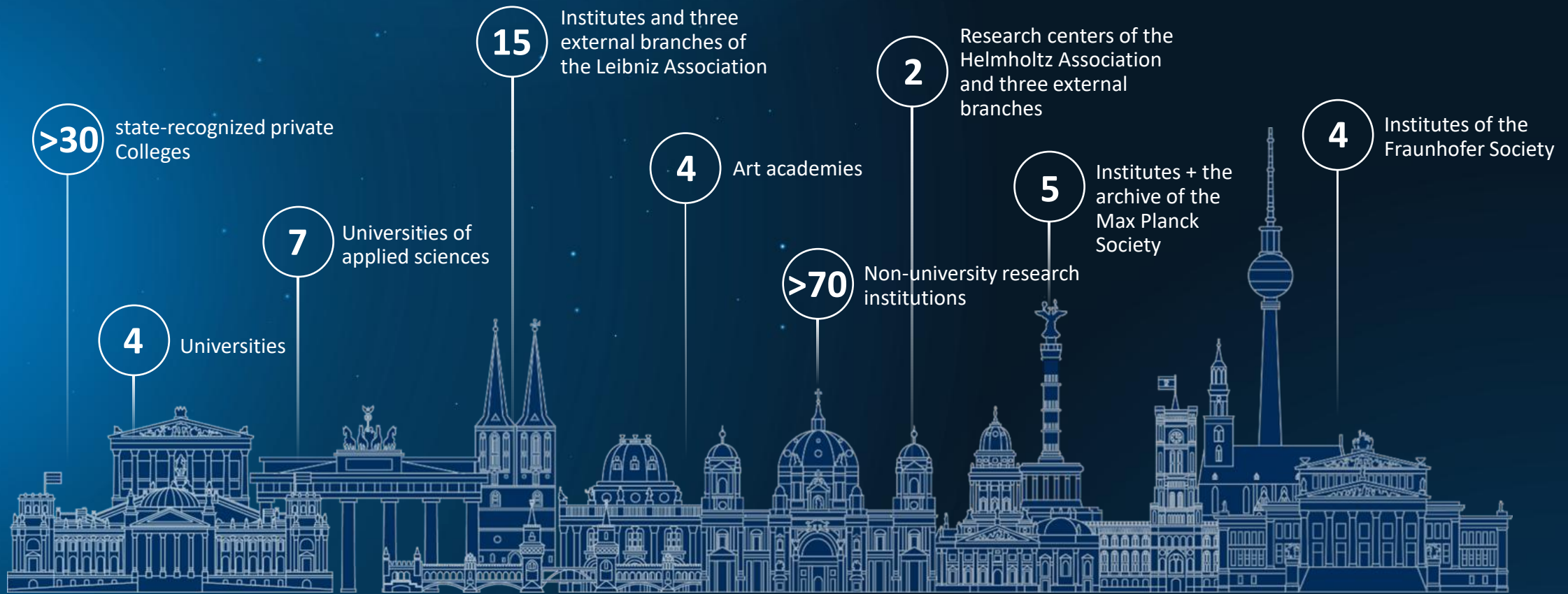


Recent Trends in AI for Earth Observation and Pathways Forward

Prof. Dr. Begüm Demir, BIFOLD and TU Berlin



Berlin – Science Metropolis



BIFOLD Facts & Figures

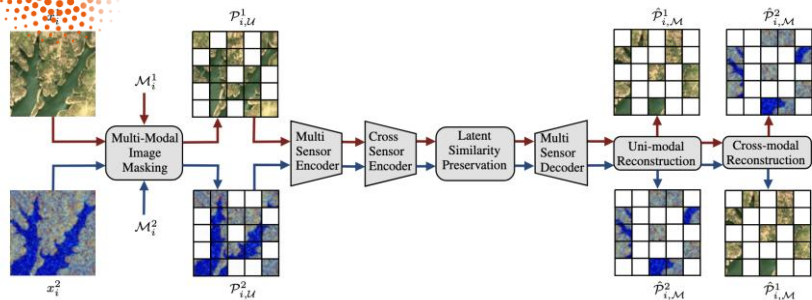
Mission Statement



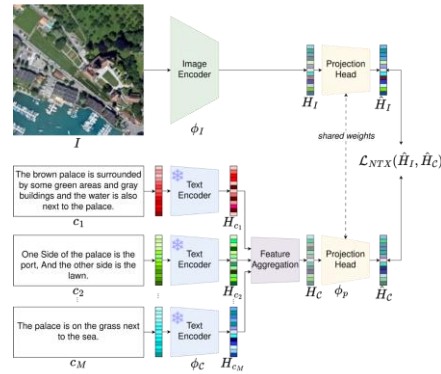
Research @ BigEarth and RSiM

Our groups perform research at the intersection of machine learning and big data management for Earth observation.

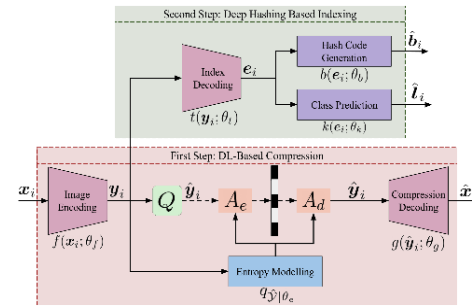
erc Satellite Image Search and Retrieval



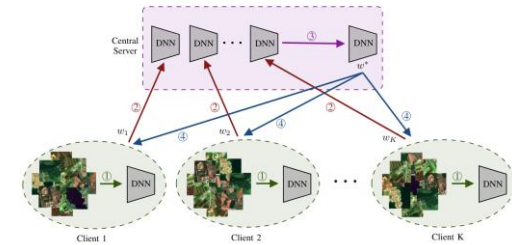
Vision-Language Models for EO



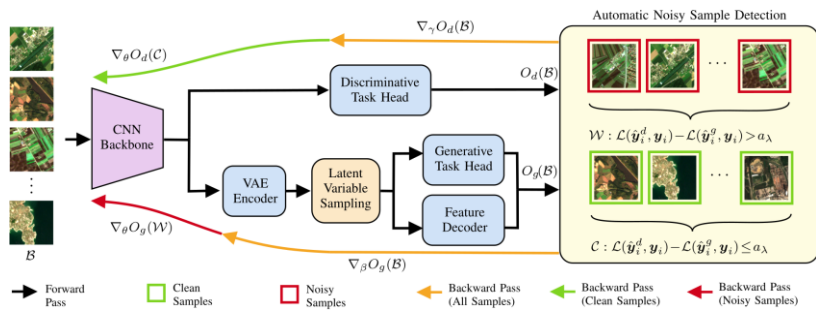
Compress Domain Data Analysis



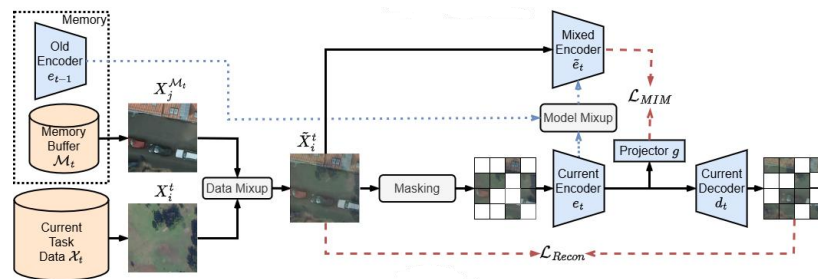
Decentralized Data Analytics



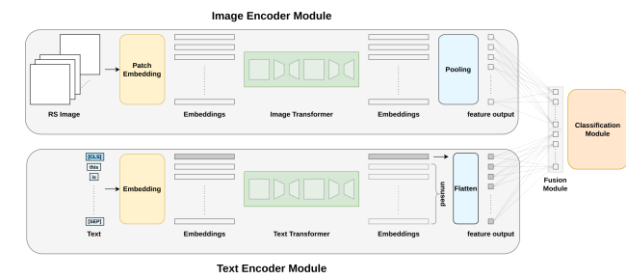
Scene Understanding



Continual and Explanation-Guided Learning



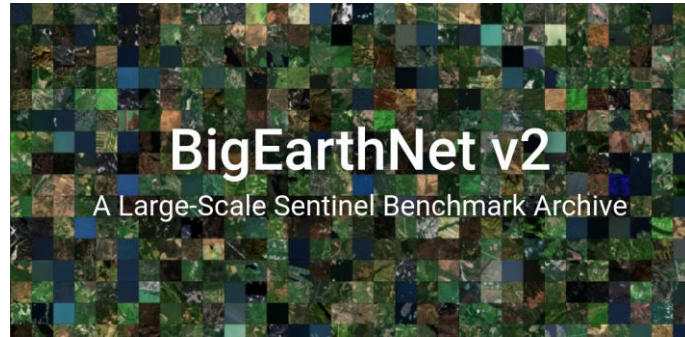
Efficient Model Development



Research @ BigEarth and RSiM

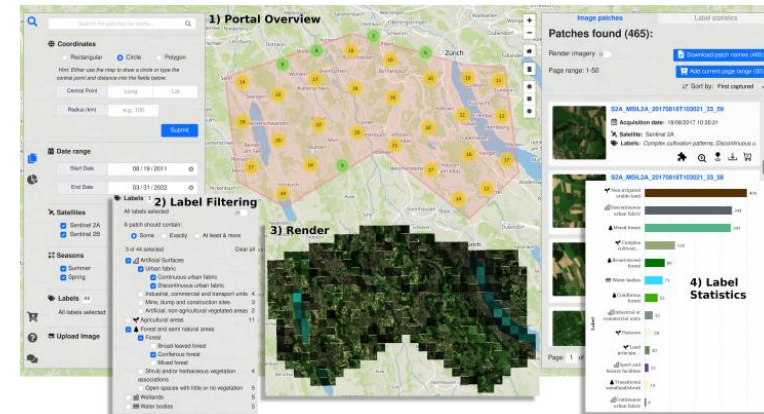
We have significant experience on the design and development of systems, software libraries and benchmark datasets for Earth observation.

Benchmark Datasets for Earth Observation

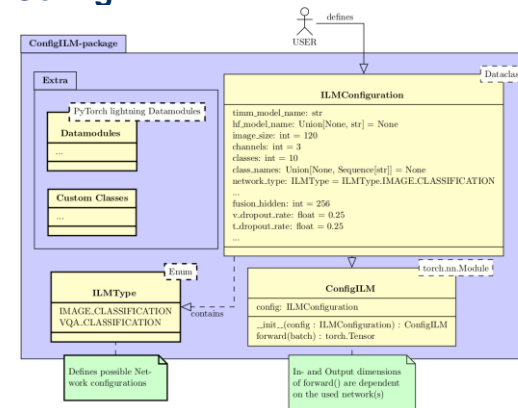


Systems and Software Libraries for Earth Observation

EarthCube



ConfigILM



DA4DTE

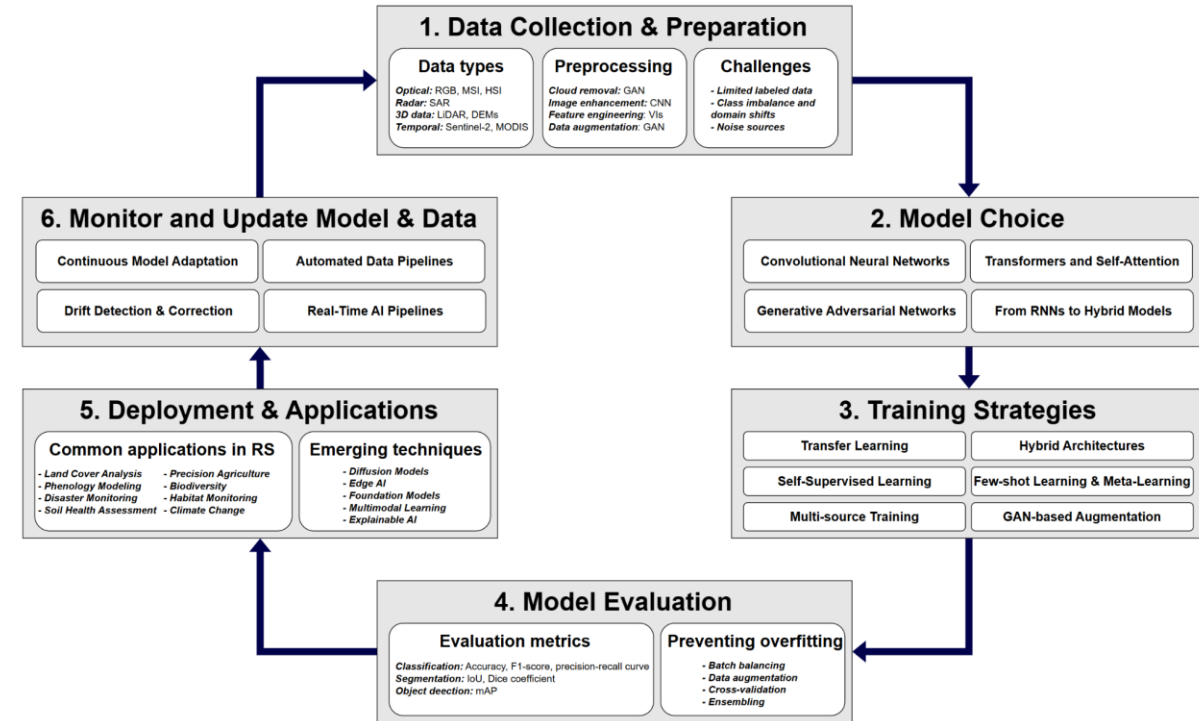
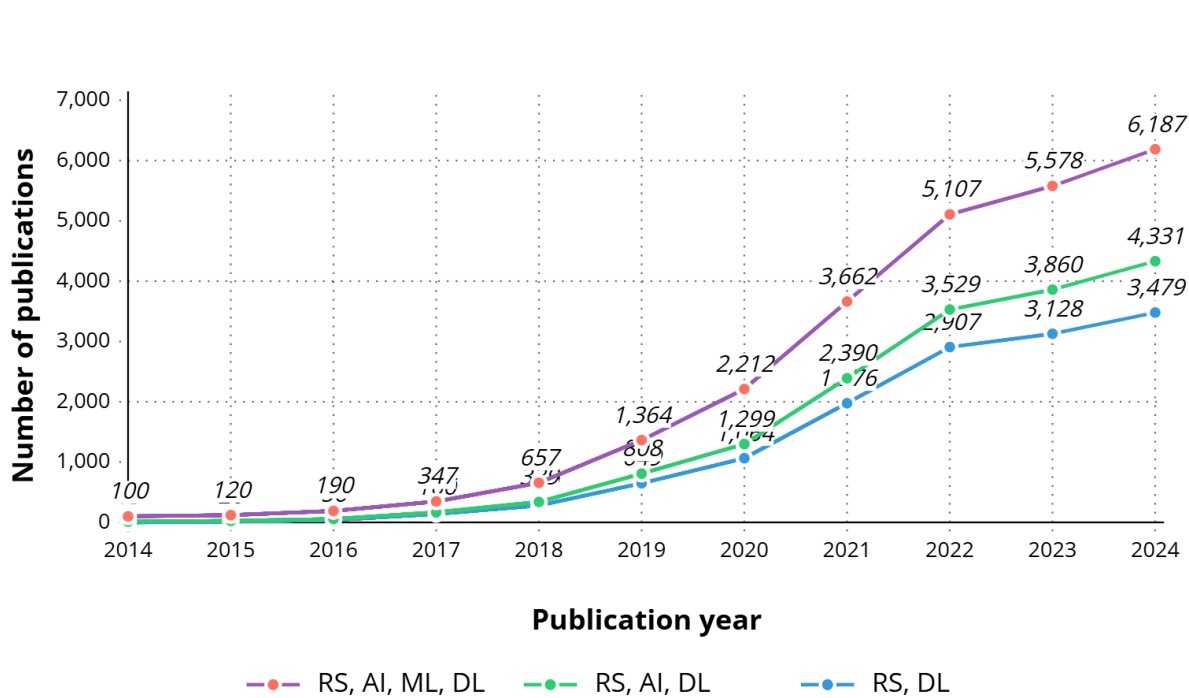


<https://rsim.berlin>

Recent Trends in AI for Earth Observation

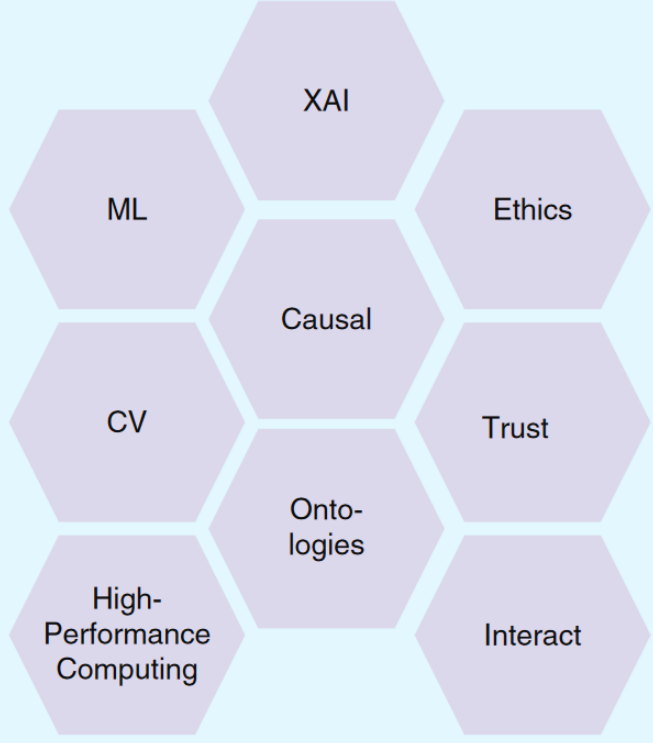
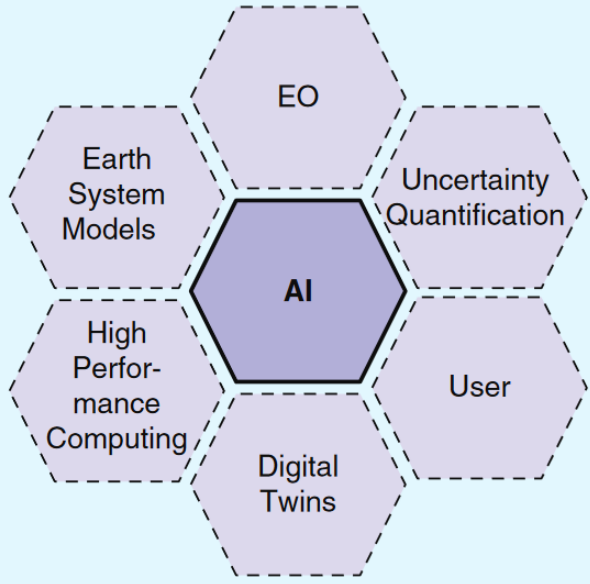
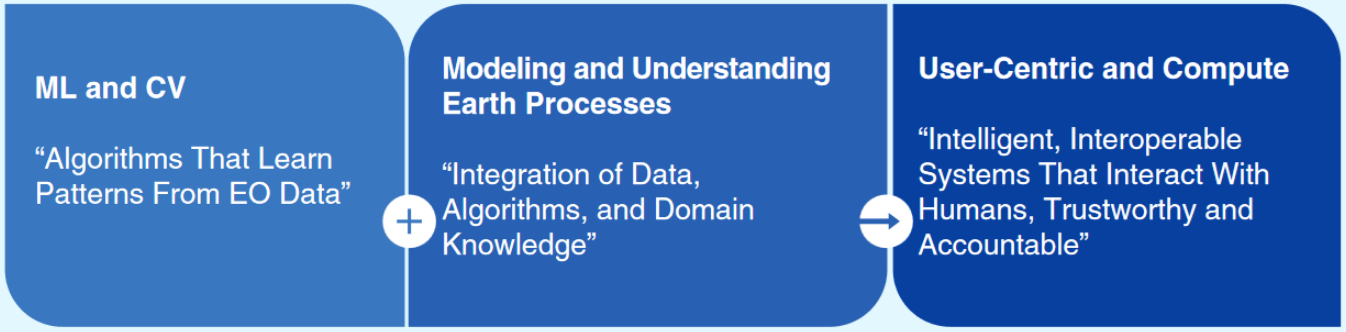


First Decade of Deep Learning Applications in Remote Sensing



A. Safonova, et al. “First Decade of Deep Learning Applications in Remote Sensing,” ISPRS Journal of Photogrammetry and Remote Sensing, under review, 2025.

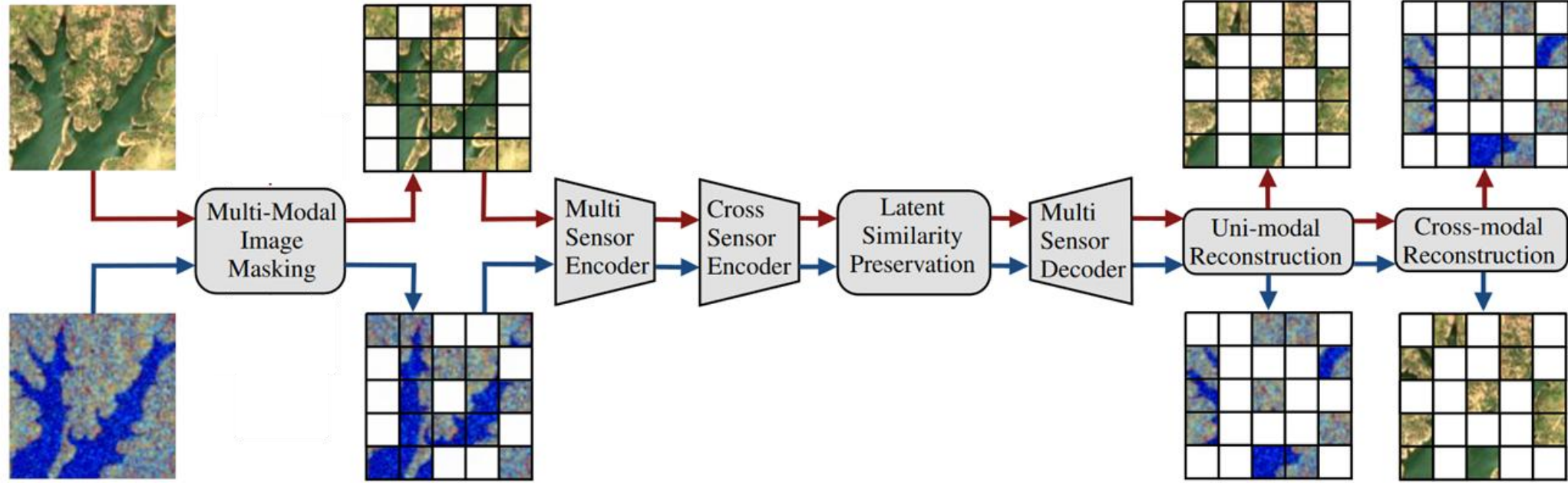
AI to Advance EO



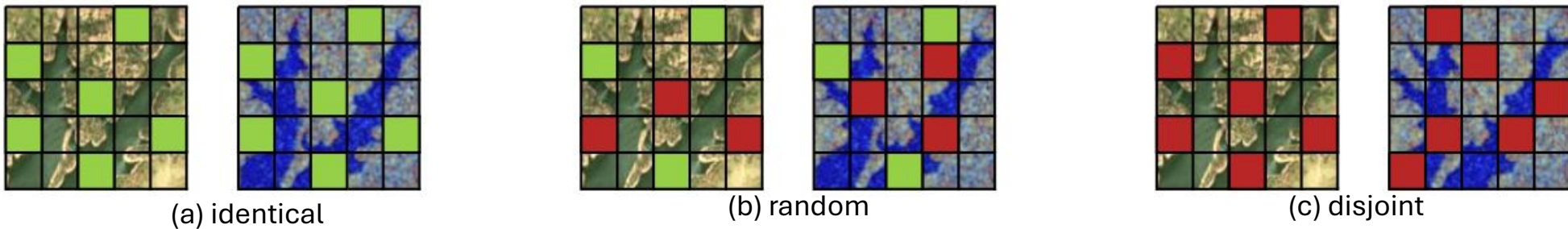
D. Tuia, et al. “Artificial Intelligence to Advance Earth Observation: A Review of Models, Recent Trends, and Pathways Forward”, IEEE GRSM, 2024.



Sensor-Agnostic Learning



J. Hackstein, et al. “Exploring Masked Autoencoders for Sensor-Agnostic Image Retrieval in Remote Sensing”, IEEE TGRS, 2024.



An illustration of three different multi-modal masking correspondences. For each one, if the same local areas are masked out on images from different sensors, they are shown in green. Otherwise, they are shown in red.

Search by Image Engine: Results

Cross-Modal Query-By-Image Retrieval: S1 → S2

1st

2nd

3rd

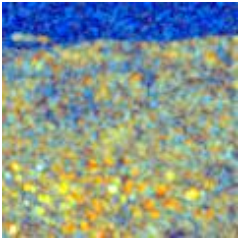
4th

5th

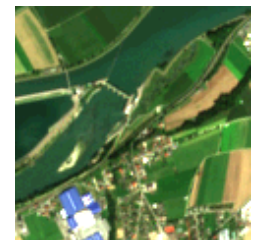
6th

7th

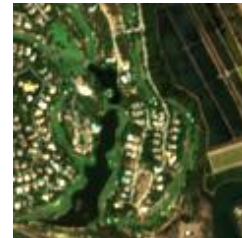
CM-MAE: most similar Sentinel-2 images



Query by
Sentinel-1



DUCH: most similar Sentinel-2 images



CMMAE: Cross-Modal Masked Autoencoders; DUCH: Deep Unsupervised Cross-Modal Contrastive Hashing

Search by Image Engine: Results

The comparison with state of the art without deep hashing-module and using a subset of BigEarthNet (~14k samples)

Method	Requirements of Labels	S1→S1	S2→S2	S1→S2	S2→S1
DUCH [1] (ours)	no	68.9 %	69.8 %	69.4 %	69.1 %
CMMAE [2] (ours)	no	69.7 %	71.5 %	70.1 %	70.6 %
S2MC [3]	yes	-	-	41.7 %	45.6 %
Deep-SM [4]	yes	-	-	65.0 %	68.7 %
DCCA [5]	no	-	-	49.1 %	47.3 %
SimCLR [6]	no	-	-	47.3 %	53.3 %

[1] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Deep Unsupervised Contrastive Hashing for Large-Scale Cross-Modal Text-Image Retrieval in Remote Sensing", arXiv:1611.08408, 2022.

[2] J. Hackstein, G. Sumbul, K. N. Clasen, B. Demir, "Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing", IEEE TGRS, 2024

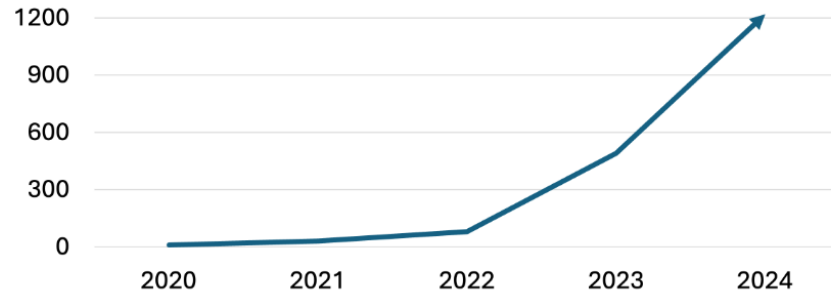
[3] M. Li, Y. Li, S. Huang, and L. Zhang, "Semantically supervised maximal correlation for cross-modal retrieval," in IEEE International Conference on Image Processing, 2020, pp. 2291-2295.

[4] Y. Wei et al., "Cross-Modal Retrieval With CNN Visual Features: A New Baseline," in IEEE Transactions on Cybernetics, vol. 47, no. 2, pp. 449-460, Feb. 2017

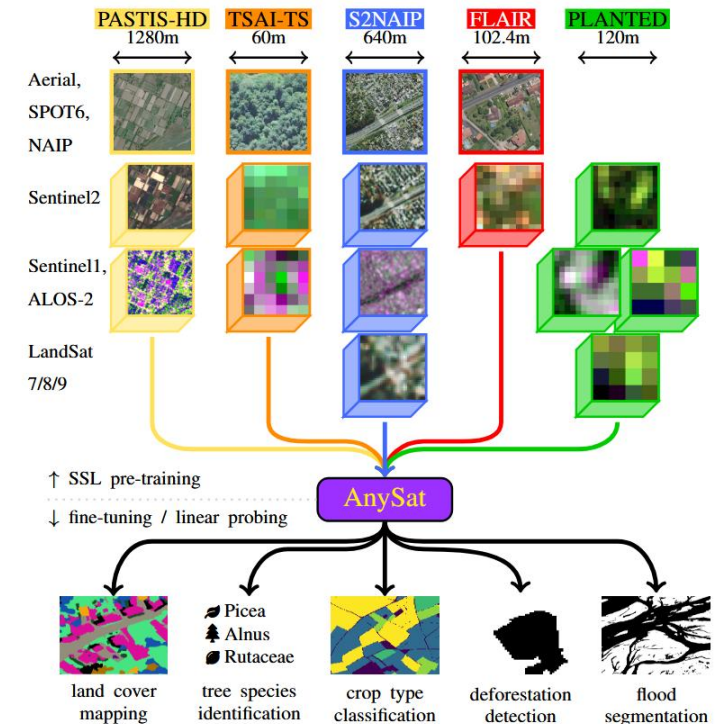
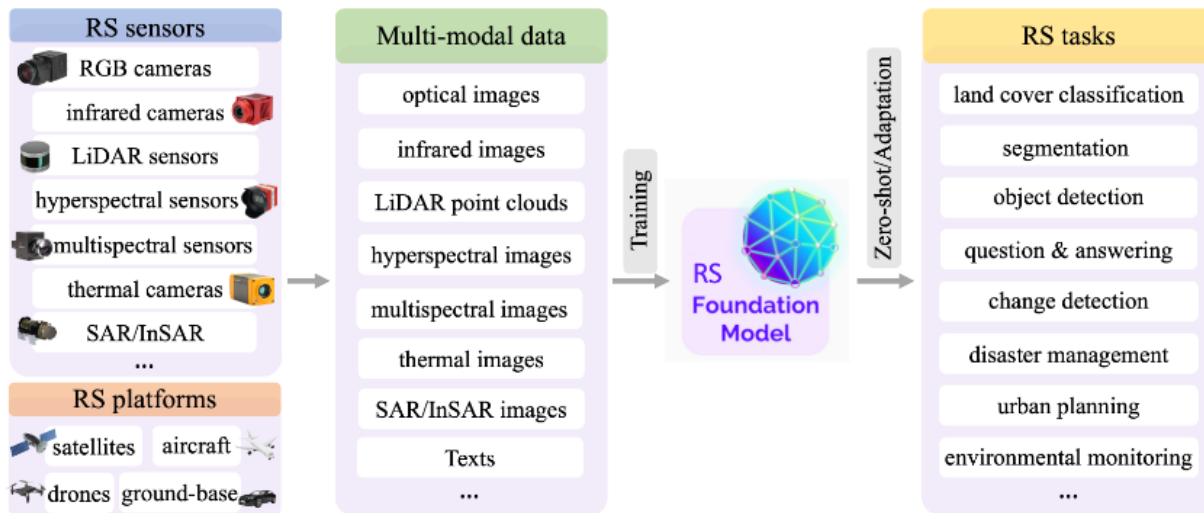
[5] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," International Conference on Machine Learning, vol. 28, no. 3, pp. 1247-1255, 2013

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," International Conference on Machine Learning, pp. 1597-1607, 2020.

Foundation Models for EO



Cumulative number of Google Scholar papers containing the keyphrases 'foundation model' and 'remote sensing' (2020 onward).

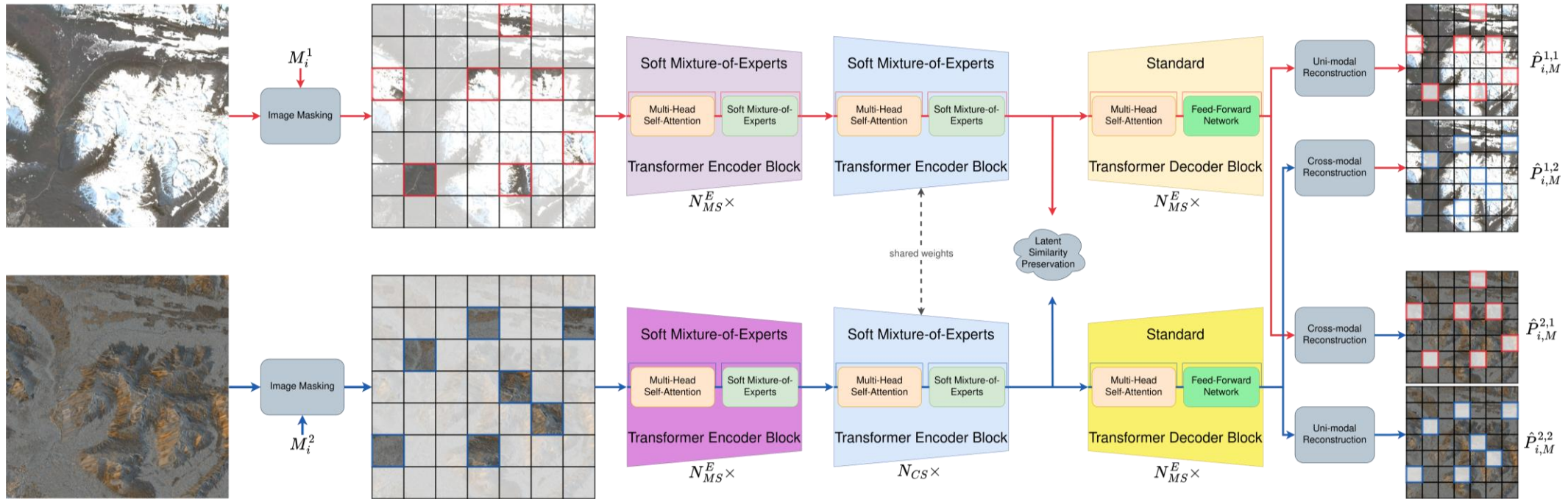


Using a scale-adaptive joint embedding predictive architecture (JEPA), AnySat can train in a self-supervised manner on highly heterogeneous datasets.

A. Xia, et al. "Foundation Models for Remote Sensing and Earth Observation: A Survey," Arxiv 2410.16602, 2024.

G. Astruc, et al. "AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities," CVPR, 2025.

Efficient Remote Sensing Foundation Models with Mixture-of-Experts



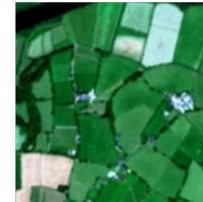
L. Hackel, et al. "Efficient Remote Sensing Foundation Models with Mixture-of-Experts," IEEE TGRS, in review 2025.

J. Puigcerver et al., "From sparse to soft mixtures of experts," ICLR, 2024.

Soft MoE performs an implicit soft assignment by passing different weighted combinations of all input tokens to each expert. Experts in Soft MoE only process a subset of the (combined) tokens, enabling larger model capacity (and performance) at lower inference cost.

Vision-Language Foundation Models in Remote Sensing

- ✓ Inspired by the success of LLM based FMs in NLP such as LLama and GPT-3, VLMs in RS have recently exhibited substantial advancements.
- ✓ The VLMs learn image-language alignments from a large number of image-text (i.e., image-caption) pairs and are then fine-tuned with a small amount of labeled data.
- ✓ Existing VLMs in RS demonstrate remarkable capability in several image-language tasks.



Q: Does *Arable Land* cover more than 75% of the area?
A: Yes.

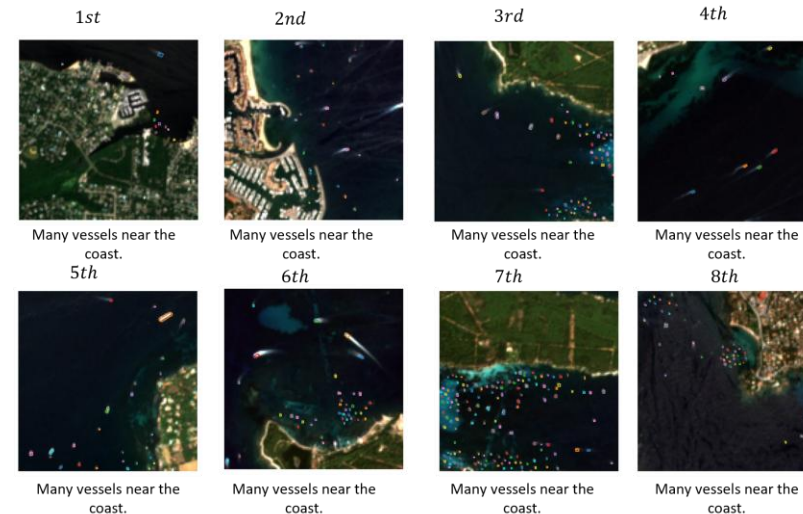
Q: Does the *Inland Water* cover between 50-75%?
A: No.



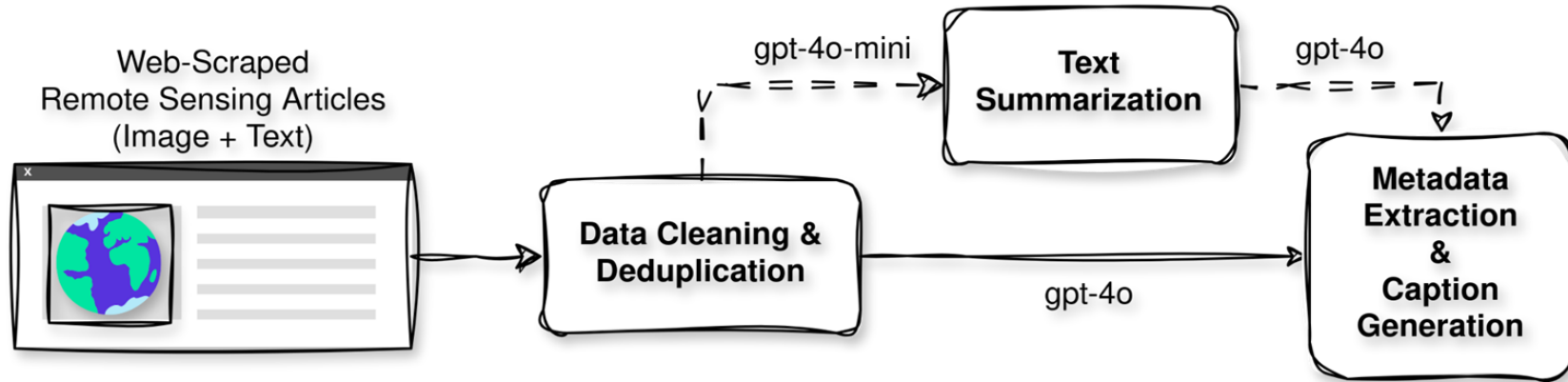
Q: Do the *Pastures* present next to *Inland Water*?
A: No.

Q: Does the *Urban Fabric* present next to the *Inland Water*?
A: Yes.

Query Text
Many vessels near the coast.



GAIA: Vision-Language Dataset for EO



Scale: ~41,030 unique RS images featuring 5 diverse, synthetic captions per image (205,150 image-text pairs) in addition to the original alt-text caption and metadata.



- **Oil slicks** prominently visible in **San Francisco Bay**, captured using **radar imagery**. The **oil is seen as dark areas**, contrasting with **brighter land and rougher water surfaces** due to reduced surface tension caused by the spill.
- **Spread of oil north and west in San Francisco Bay, encircling Angel Island**. Smooth, **dark patches** on the water indicate **oil presence**, visible in **radar imagery** due to its ability to detect **textural differences**.
- A significant **oil spill extending from San Francisco Bay into the Pacific Ocean**. Long **ribbons of oil** south of the **Bay Bridge** are visible, illustrating the impact of **tidal movements on oil dispersion**.
- Environmental impact on **marine habitats**, with **oil spreading over 30 kilometers of shoreline**. **Radar imagery** reveals smoother **water surfaces** where oil is present, affecting nearby **wildlife and ecosystems**.
- Radar detection of **oil slicks** near **urban areas and natural reserves**. The contrast between **oil-darkened water** and brighter **urban land** highlights the **spill's extent** and the **ongoing clean-up efforts visible through ships in the bay**.



A. Zavras, et al. "A Global, Multi-modal, Multi-scale Vision-Language Dataset for Remote Sensing Image Analysis," IEEE GRSM, in review 2025.

Redundancy-aware Pretraining of Vision-Language Foundation Models in Remote Sensing

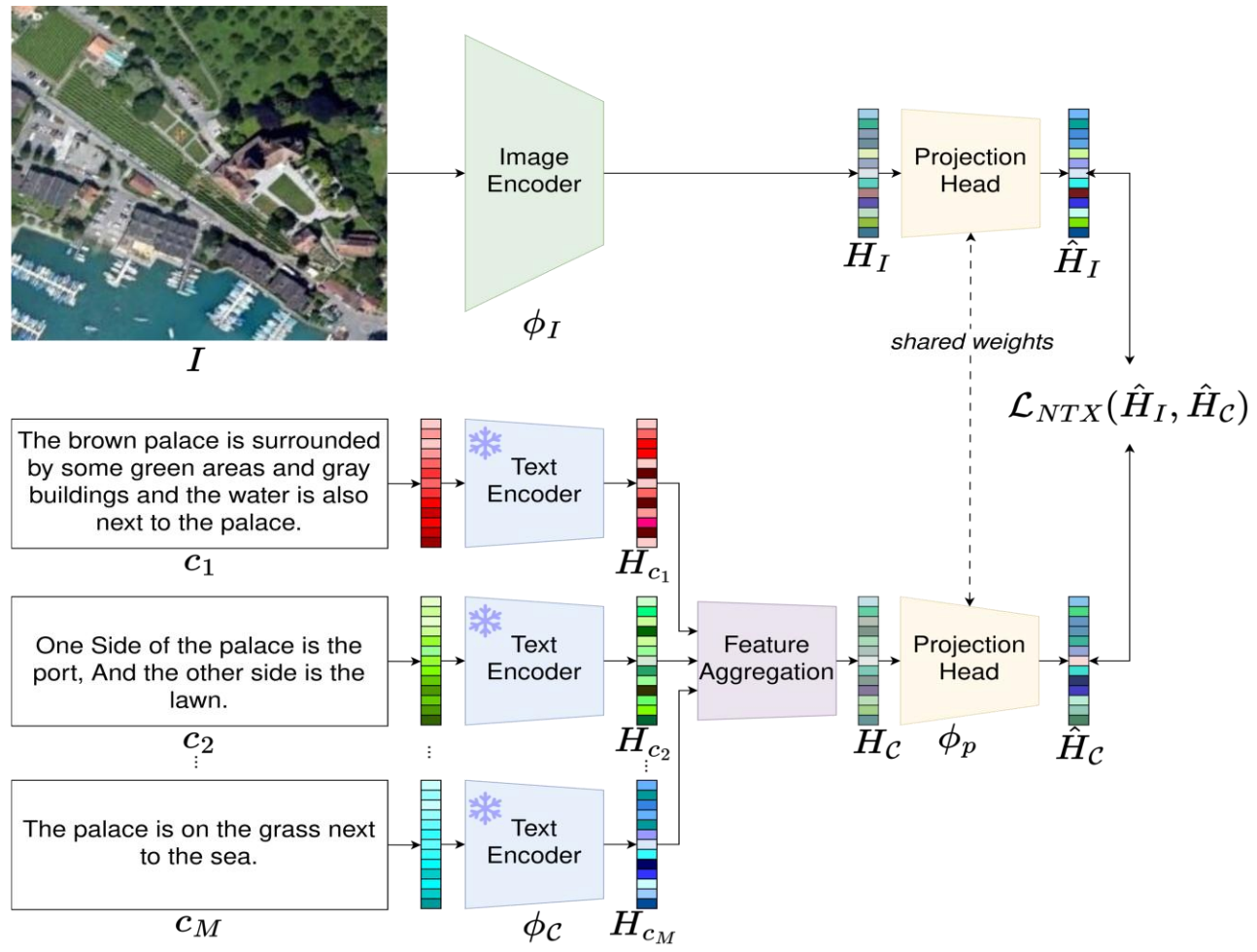


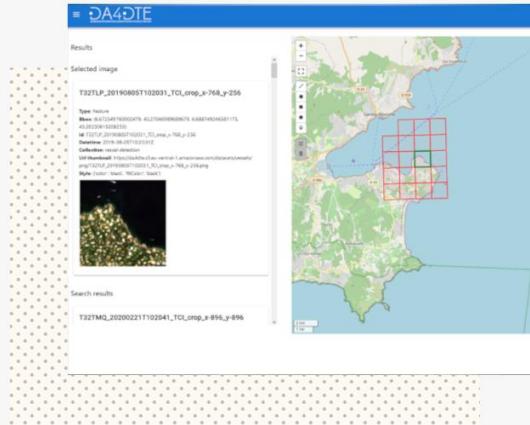
- 1) Four airplanes are parked at the airport.
- 2) There are some planes and cars in the airport.
- 3) Four different kinds of airplanes are in the airport.
- 4) Four different sizes of airplanes are in the airport.
- 5) Here are some airplanes and cars in the airport.

- ✓ We introduce a weighted feature aggregation (WFA) strategy that:
 - extracts and exploits complementary information from multiple captions per image;
 - reduces redundancies through feature aggregation with importance weighting.
- ✓ To calculate adaptive importance weights for different captions of each image, we propose two different techniques: i) non-parametric uniqueness; and ii) learning-based attention.



M. Adler, et al. "Redundancy-aware Pretraining of Vision-Language Foundation Models in Remote Sensing," IEEE IGARSS, 2025.



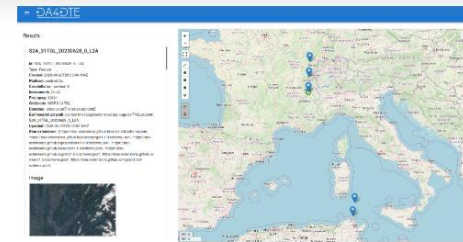
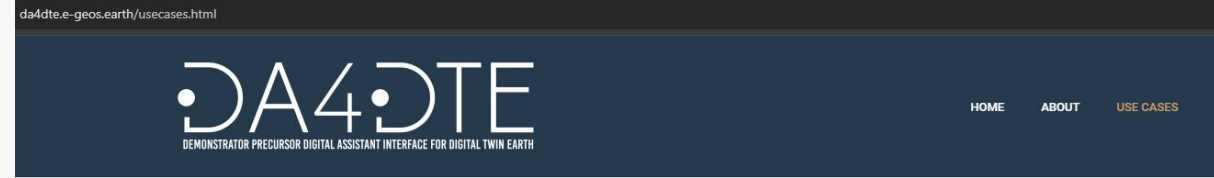


About the project

Digital Assistant capable of understanding complex requests related to geospatial data searches

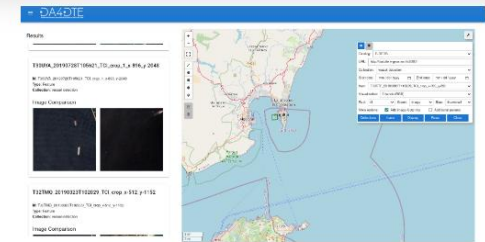
We present an **AI-powered** digital assistant that extracts and utilizes the content of satellite images, leveraging cutting-edge technologies and advances in Natural Language Processing (NLP), Machine Learning (ML), and Computer Vision (CV) to address Earth Observation challenges.

[Go To The Platform](#)



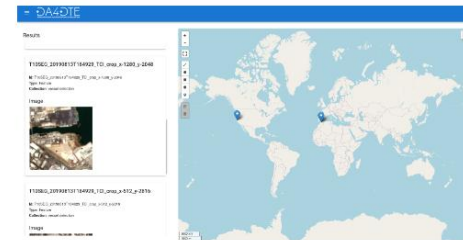
Knowledge Graph Question Answering

The Knowledge Graph Question Answering (KGQA) engine accepts questions in natural language (English) that request satellite images meeting specific criteria and returns links to such datasets. The questions can refer to image metadata as well as geographic entities, both of which are included in the target knowledge graph.



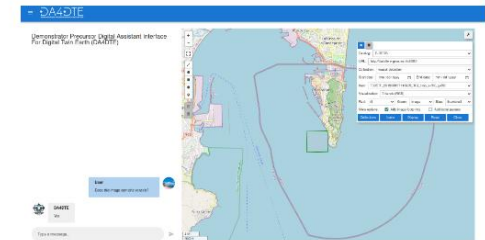
Search By Image

The Search by Image engine takes a query image and computes the similarity function between the query image and all archive images to find the most similar images in a scalable way. We developed this engine based on the self-supervised DUCH and CM-MAE methods.



Search By Caption

The Search by Caption engine takes a text sentence to search for images, achieving cross-modal text-image retrieval. We developed this engine by adapting the self-supervised DUCH and CM-MAE methods to be operational



Visual Question Answering

The Visual Question Answering (VQA) engine allows users to ask questions about the content of remote sensing (RS) images in a free-form manner, extracting valuable information. In this context, an efficient and accurate

Funded by

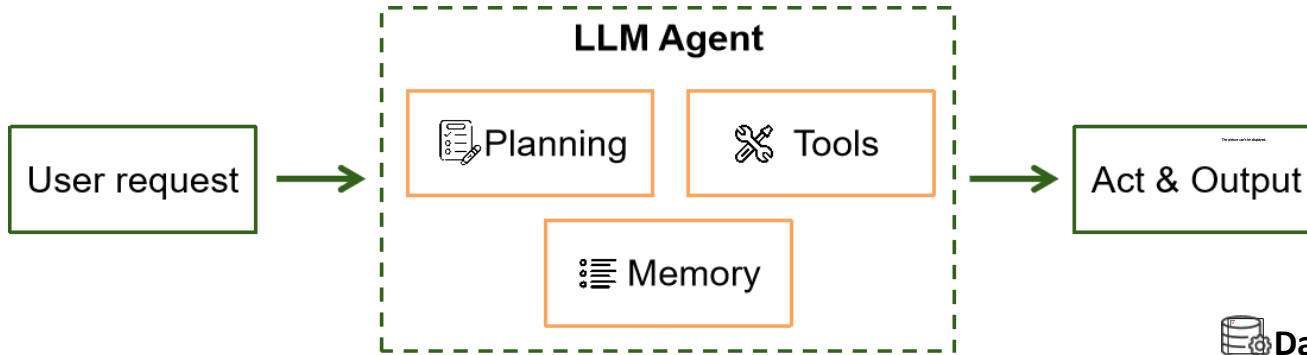


Pathways Forward

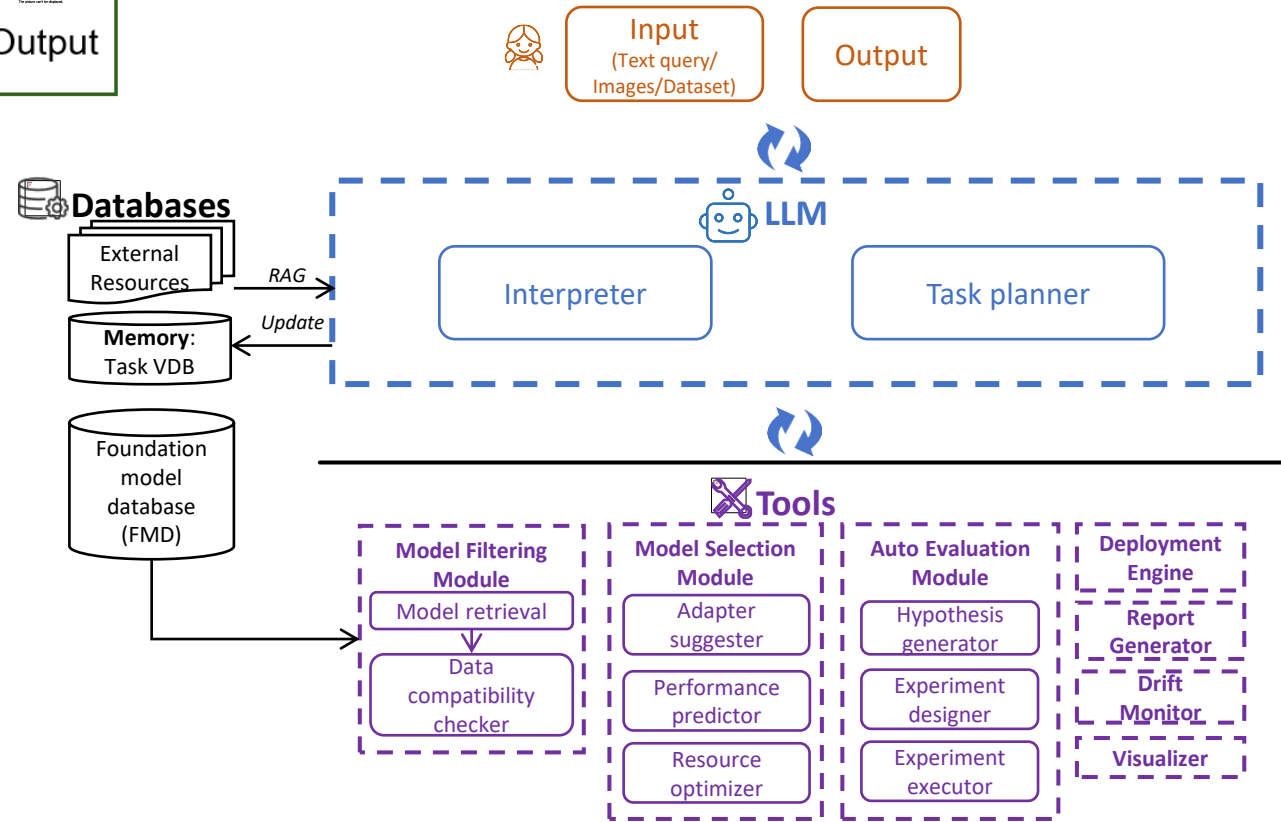


LLM Agents

Our Vision – Agents for Foundation Model Selection



- **Planning:** decide what tasks to do and in what order
- **Tools:** execute tasks from the plan
- **Memory:** remember what has been done



Example Chat Demonstration



I want to map different crop types in Africa. I have some Sentinel-2 data and a few labels from last year.



Can you check if this setup actually works?

What's your hardware setup?
Are the labels sparse or well-distributed?



✓ Ran test on Ethiopia + Morocco
Baseline F1: 0.68 → Adapted F1: 0.78
+0.10 gain → Accept adaptation



Running on cloud, no GPU worries. But label data is sparse.



Please deploy the model for my task.

🔍 **Selected base model:** Presto
⚠️ **Issues detected:**
label sparsity, task shift, efficiency need
💡 **Suggested adapters:**
LoRA, Spatial resizing, Pseudo-labeling
⚙️ **Optimized Config:**
{ precision: "fp16", batch_size: 8, engine: "ONNXRuntime" }
📊 **Performance Estimation:**
Expected F1: 0.78, Confidence: Medium

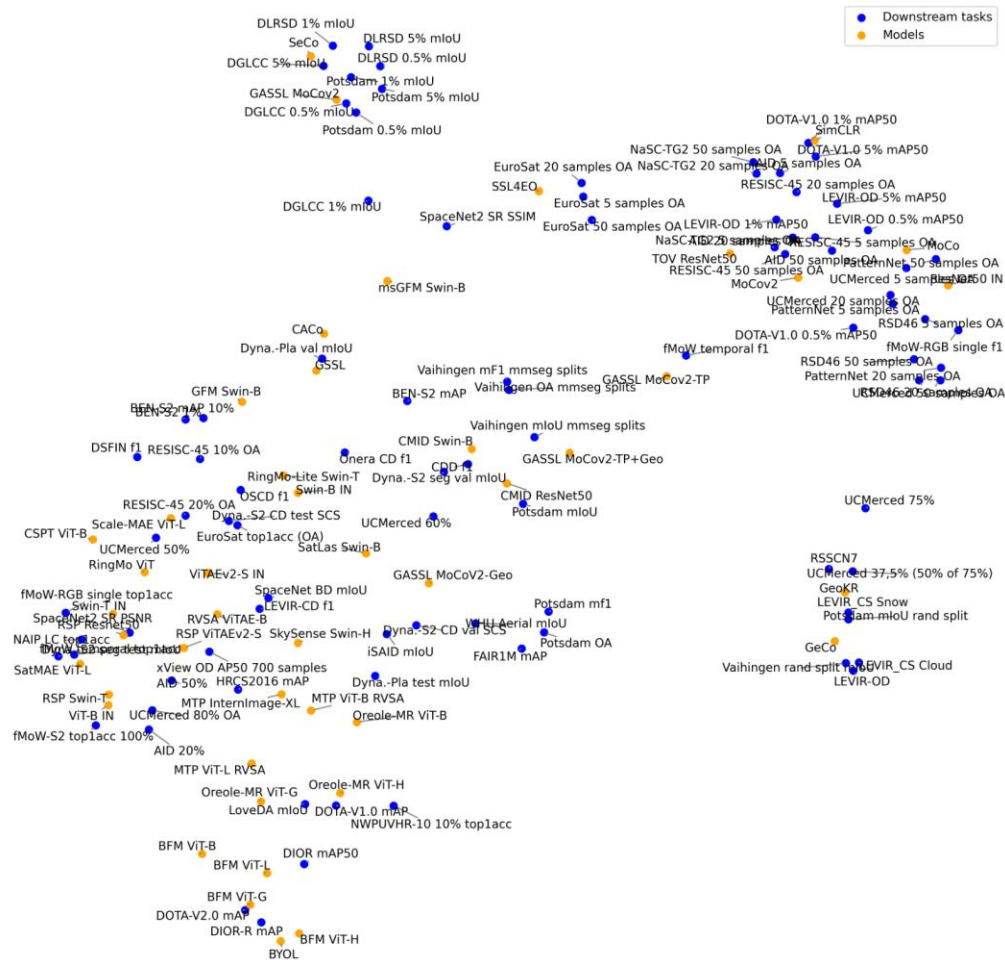


✓ Model deployment.

Final Report:



Evaluation



P. Adorni, et al. “Towards Efficient Benchmarking of Foundation Models in Remote Sensing: A Capabilities Encoding Approach”, CVPR MORSE Workshop, 2025.



L. Maier-Hein, Reinke, A., Godau, P. et al. “Metrics reloaded: recommendations for image analysis validation,” Nat Methods 21, 195–212 (2024).

Metrics Reloaded

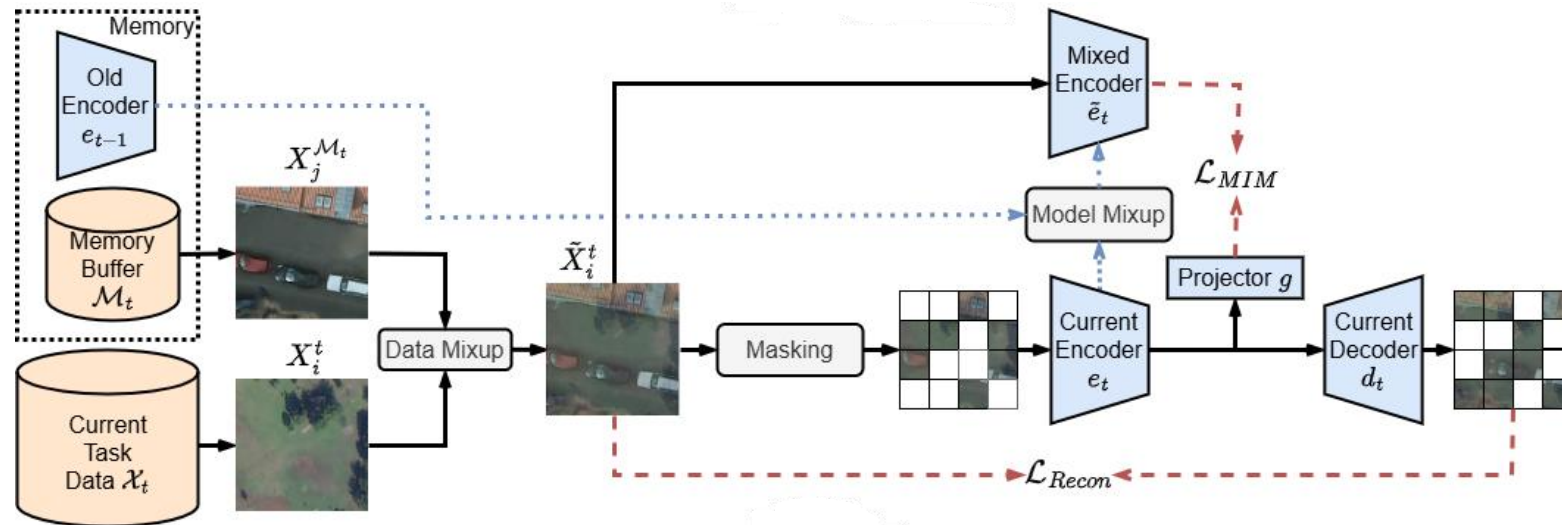
A framework for trustworthy image analysis validation

(Quick access to Metrics Reloaded main paper [↗](#))

Metrics Reloaded Mission

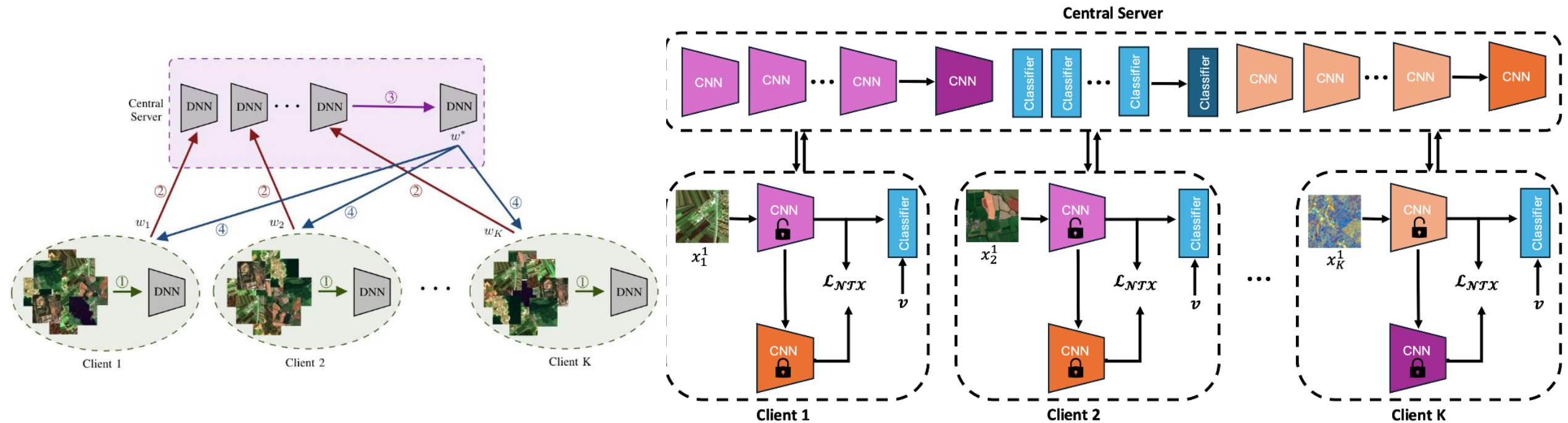
The mission of Metrics Reloaded is to guide researchers in the selection of appropriate performance metrics for biomedical image analysis problems, as well as provide a comprehensive online resource for metric-related information and pitfalls.

Towards Continual Learning



L. Mollenbrok, et al. "Continual Self-Supervised Learning with Masked Autoencoders in Remote Sensing," IEEE GRSL, in review, 2025.

Open Collaboration: Federated Learning for FM Training



B. Buyuktas et al., Federated Learning Across Decentralized and Unshared Archives for Remote Sensing Image Classification, IEEE GRSM, 2024.

B. Buyuktas et al., A Multi-Modal Federated Learning Framework for Remote Sensing Image Classification, IEEE TGRS, under review, 2025.



The First Workshop on Foundation and Large Vision Models in Remote Sensing (MORSE)

at the IEEE/CVF Computer Vision and Pattern Recognition Conference 2025

Nashville, TN, June 12, 2025

GAIA 2025 Symposium

Geospatial AI and Applications with Foundation Models

24 - 26 September 2025 | Sofia, Bulgaria

Visit Our Group Webpage
<https://rsim.berlin>

