Institute
Jožef Stefan
Ljubljana, Slovenia

From Benchmarks to Knowledge Bases

FAIR and Trustworthy Resources for AI4EO

Dragi Kocev
Bias Variance Labs & Jožef Stefan Institute

AutoML, Benchmarking and Open Science Infrastructure for Earth Observation
Workshop at AutoML, Ljubljana, 01. 10. 2026

FAIR-EO is funded under the OSCARS 1st Open Call (Horizon Europe, GA No. 101129751)

EO has no shortage of models or datasets. It has a shortage of comparable results.

Re-run under one controlled protocol, much of the reported progress does not survive. AutoML — meta-learning, performance modelling, pipeline design — inherits whatever empirical record we give it.

1 **Outdated baselines**
New architectures are compared against numbers copied from older papers, not re-trained under the same setup.

2 **Leaky splits**
A widely used random split of WHU-CD leaked roughly 85% of test patches into training, inflating scores.

3 **Misattributed gains**
Improvements credited to new modules often come from the training recipe: loss, schedule, augmentation.

Reality check for change detection: Corley et al. (2024), as discussed in Tomanič et al. (2026)

Two halves of the same problem

01

INFRASTRUCTURE

FAIR-EO & the FAIR-EO Hub

A semantic layer that describes datasets, methods, runs and results as linked data — and a hub that serves it.

02

EVIDENCE

Benchmarks & lessons learned

What survives a single controlled protocol: AiTLAS Benchmark Arena, the change detection benchmark, and five tasks in the hub.

Benchmarks are only as useful as their record is trustworthy and queryable — and infrastructure is only as useful as the evidence it holds.

01

FAIR aspects & the FAIR-EO Hub

A shared semantic layer for EO datasets, methods, experiments and results

EO–AI collaboration is slowed by three recurring gaps

1

Fragmented data

EO datasets are scattered across sources in inconsistent formats — hard to discover and prepare for AI.

2

Weak FAIR adherence

Resources are poorly documented and rarely interoperable, so they are difficult to find and reuse.

3

Not reproducible

Models are evaluated in different settings and conditions — results are hard to compare or repeat.

FAIR-EO addresses all three with a shared, standards-based semantic layer.

The FAIR-EO Hub

An integrated, AI-ready catalogue — every resource semantically annotated

Web explorer · SPARQL endpoint · JSON-LD / Turtle exports

DatasetHub

278

AI-ready EO datasets

Preprocessed, standardised, annotated: bands, classes, splits, licences, literature metrics

ModelHub

149

methods & models

Traditional and foundation models implemented in AiTLAS toolbox

ExperimentsHub

880

benchmark runs

Every run with parameters, resource estimates, artefacts and all evaluation metrics

AnalysisHub

5

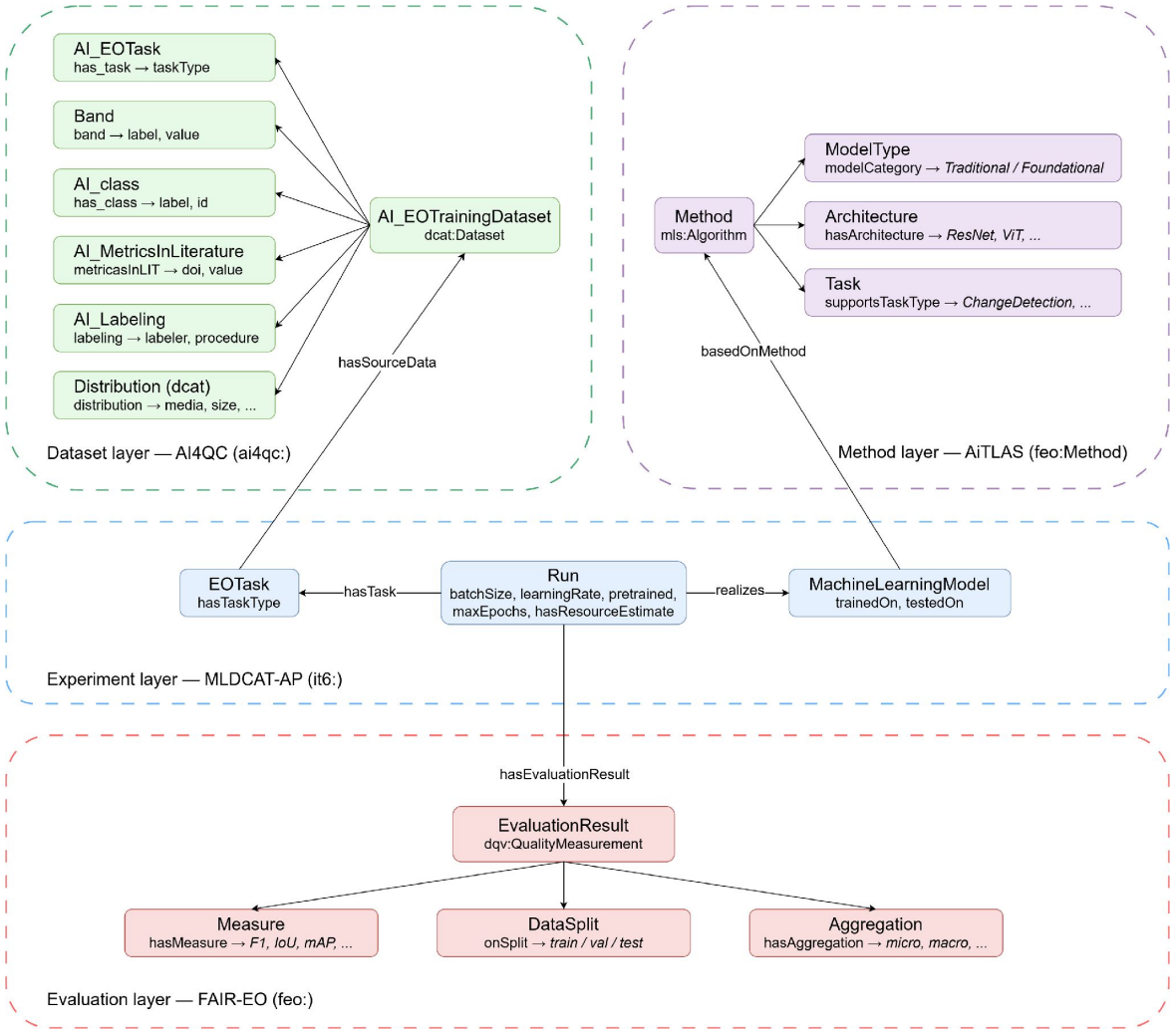
EO task types

Ready-to-use comparisons: performance, run time, GFLOPs, scratch vs pre-trained

The FAIR-EO schema is the semantic layer that ties these together into one queryable knowledge base.

Counts as served by eodata.bvlabs.ai/ai4eo, October 2026

The FAIR-EO schema: four layers, one connected model



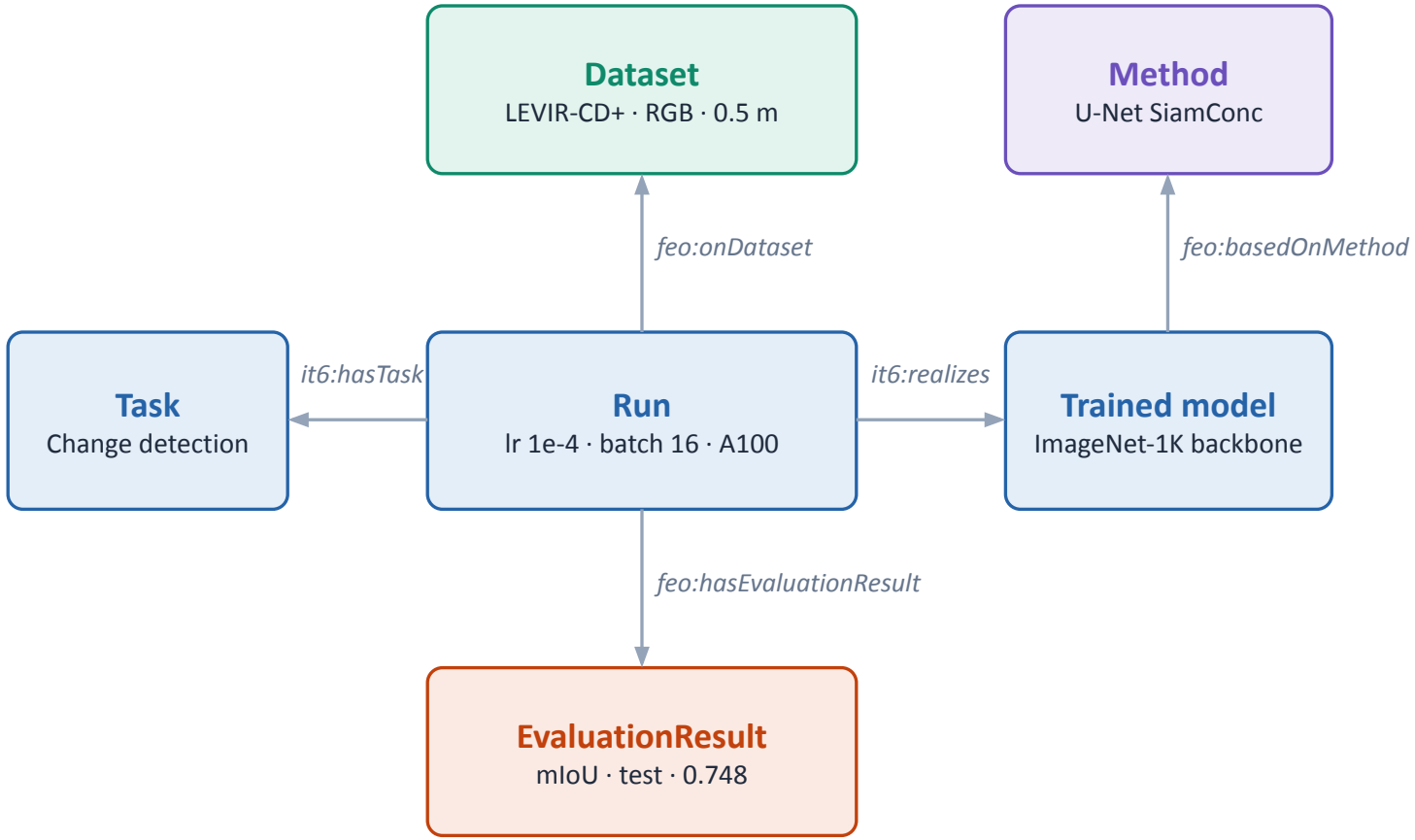
- **Dataset** AI4QC schema reuse
  Task, spectral bands, label classes, distributions, labelling, metrics reported in the literature
- **Method** simple taxonomy of methods
  Architecture, model family, traditional vs foundation, supported tasks, input modality
- **Experiment** MLDCAT-AP 2.0 schema
  Run, task and trained model, with parameters and resource estimates
- **Evaluation** FAIR-EO · DQV
  One metric = one node: measure, aggregation, split and value

Three bridges make it one graph: hasSourceData · basedOnMethod · hasEvaluationResult

Figure: Tomanič et al. (2026), arXiv:2608.28247, Fig. 2

One experiment = one connected graph

The Run is the hub that ties the four layers together



Metrics become meaning, not strings

`F1_score Macro/test`

measure	aggregation
F1	**Macro**
split	value
Test	**0.838**

- Full provenance: one run reaches its data, method, task and results.
- Shared SKOS vocabularies: “Classification” matches “Multiclass” via skos:broader.

Example: U-Net SiamConc, pre-trained, on LEVIR-CD+ (change detection benchmark)

Built on standards, not reinvented

The schema reuses established vocabularies and aligns to them with OWL/SKOS axioms

MLDCAT-AP 2.0 schema

EU profile for ML in DCAT — Task, Run, Flow, Model. The experiment layer is used largely as-is.

AI4QC schema

EO training-dataset ontology — bands, classes, anomalies, labelling. feo:EODataset ≡ ai4qc:AI_EOTrainingDataset.

AiTLAS taxonomy

Architectures, model families, learning types, input modalities — as SKOS controlled vocabularies.

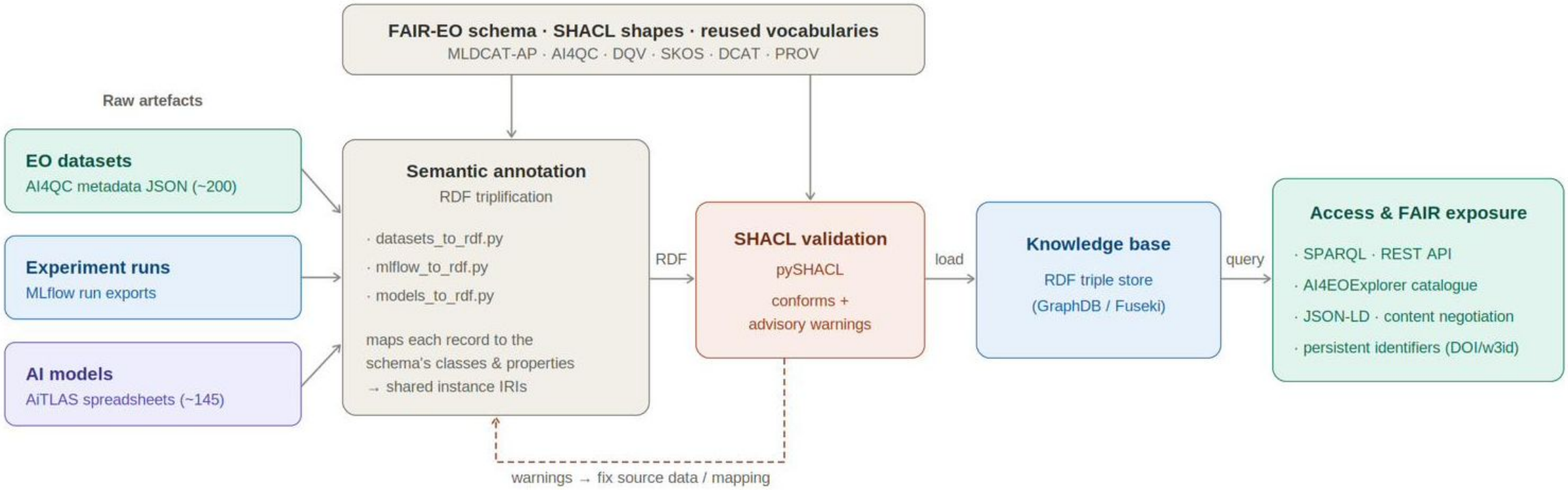
DQV · SKOS · DCAT · PROV

Quality measurements, controlled vocabularies, catalogue metadata and provenance.

Reuse over invention → interoperable and FAIR by construction. The feo: namespace is used only where no standard term fits.

Annotating datasets, models and runs

From raw EO artefacts to a validated, queryable, FAIR knowledge base



Schema-driven

Three converters triplify dataset records, MLflow runs and AiTLAS spreadsheets onto the schema.

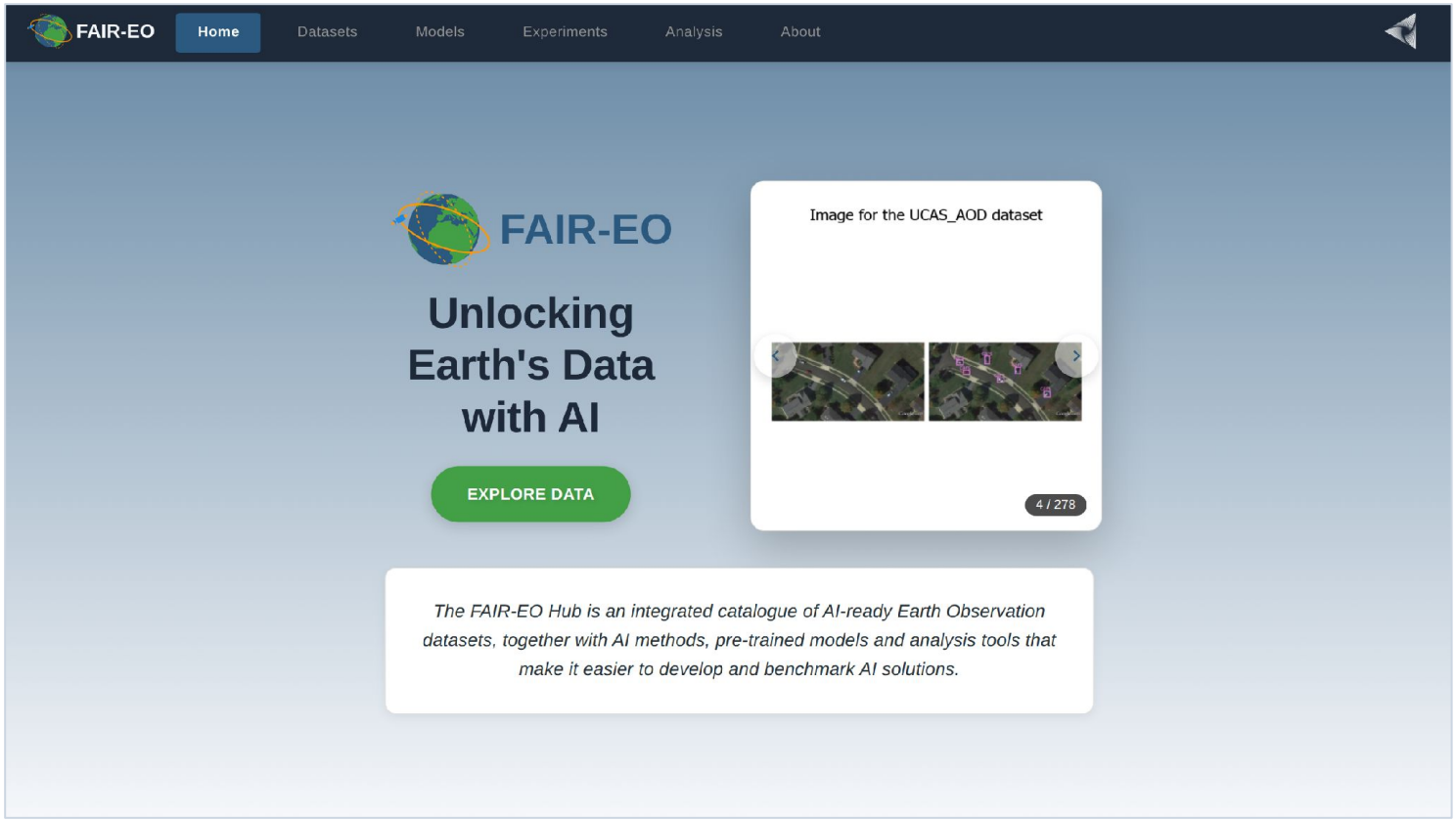
SHACL is the contract

Violations block loading; data-quality issues surface as advisory warnings that loop back to the source.

Shared instance IRIs

Independently generated datasets, methods and runs join on the same nodes in the triple store.

The FAIR-EO Hub is live



278

AI-ready EO datasets

149

methods · 76 traditional, 73 foundation

880

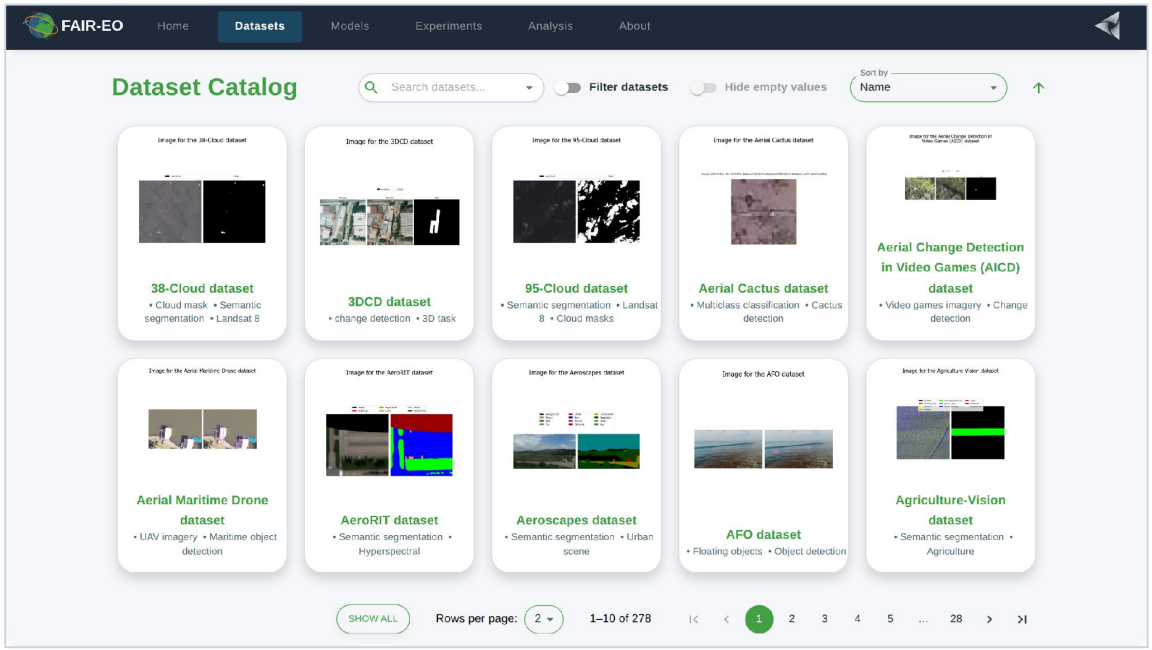
experiment runs on 45 datasets

181k

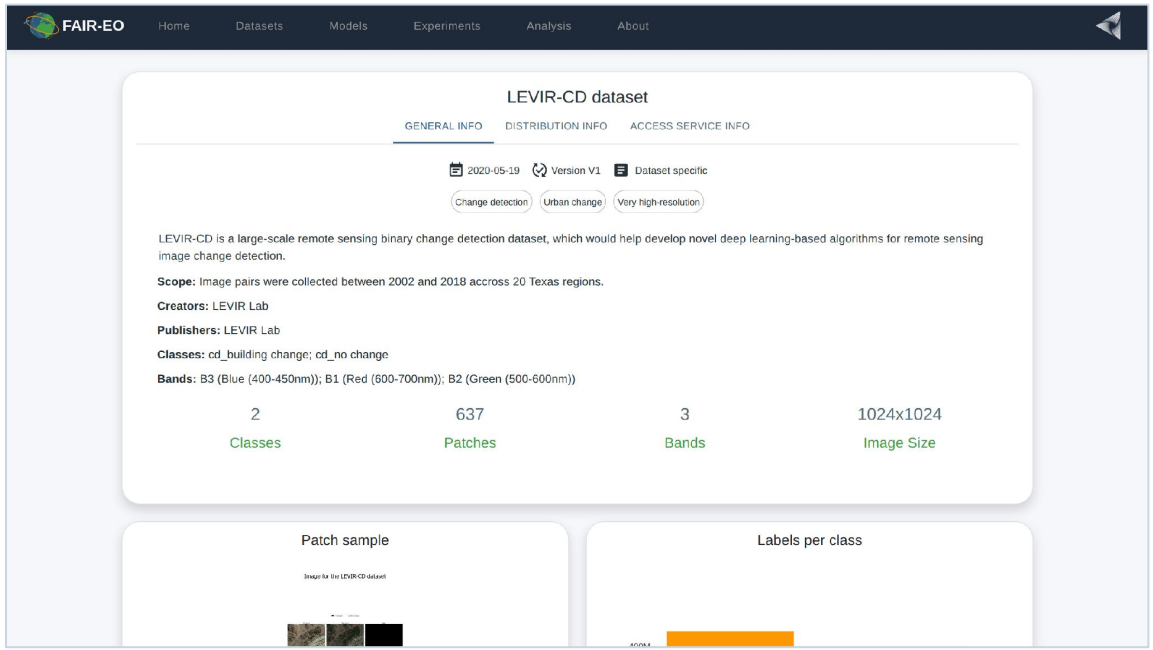
RDF triples in the knowledge base

eodata.bvlabs.ai/ai4eo · open access, CC BY 4.0 metadata

Datasets: discoverable, comparable, AI-ready



Catalogue — search and filter by task, sensor and keyword



Record — scope, classes, bands, patches, distribution and access

124
semantic segmentation

64
multi-class classification

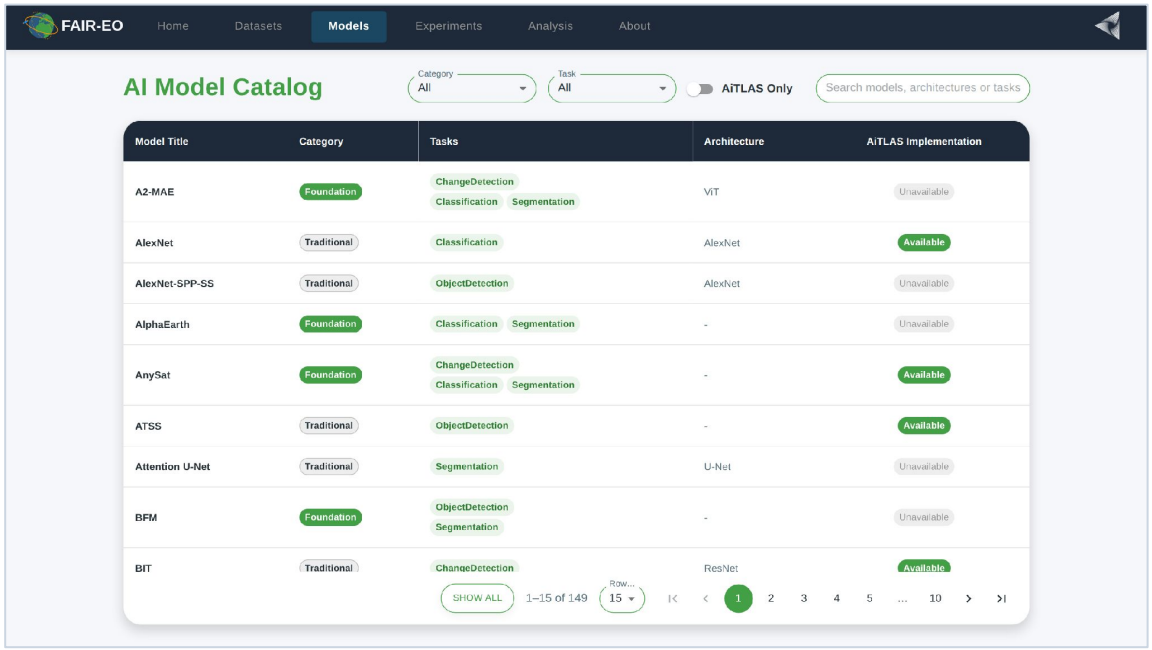
47
object detection

29
change detection

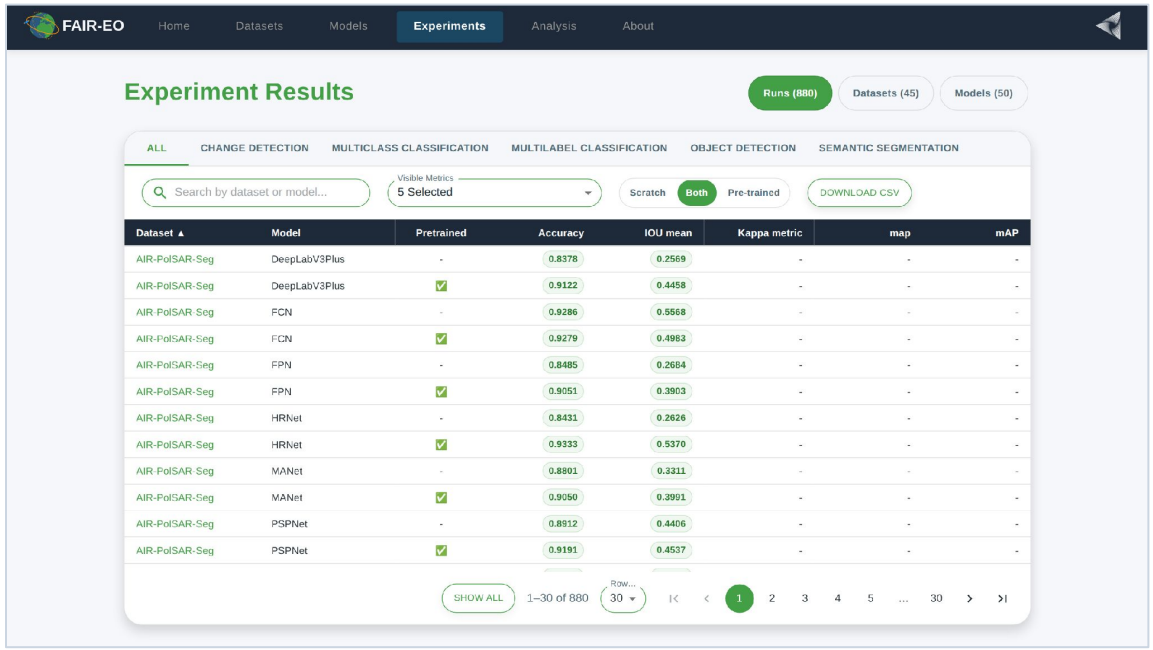
12
multi-label classification

Datasets per task type, from datasets.jsonld

Models and experiment results, side by side



Model catalogue — category, architecture, tasks, AiTLAS availability



Experiment results — scratch vs pre-trained, all metrics, CSV export

76
traditional models

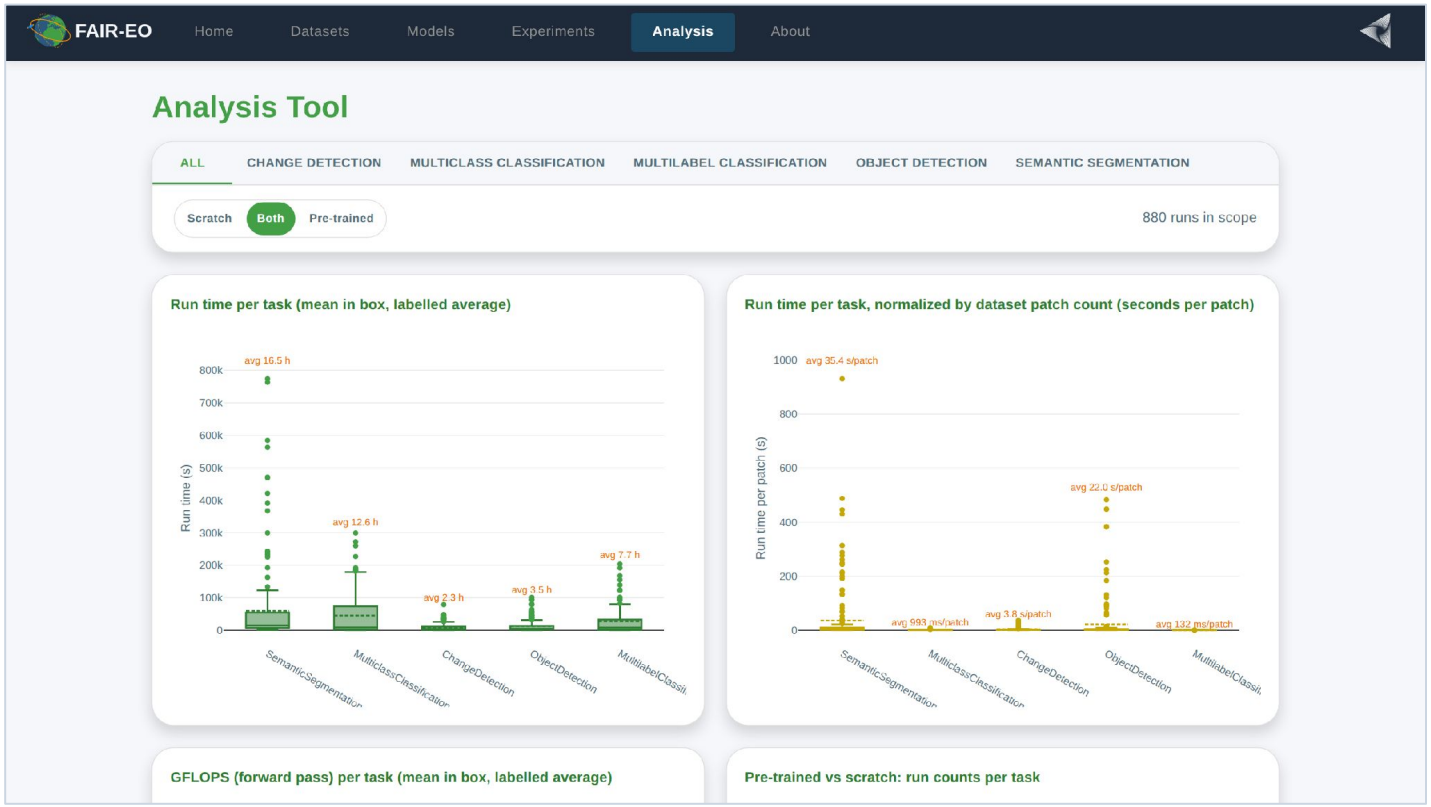
73
foundation models

45
benchmark datasets

50
benchmarked models

5
EO task types

Every page is also data



Analysis — run time, cost per patch and GFLOPs across all 880 runs

Machine-actionable access

- **SPARQL endpoint**
  /ai4eo/query
- **JSON-LD & Turtle**
  catalog · datasets · models · experiments
- **schema.org markup**
  Data Catalog / Dataset for web search
- **Persistent identifier**
  purl.org/faireo
- **Open licence**
  CC BY 4.0

From a leaderboard in a PDF to one SPARQL query

```
PREFIX feo: <http://w3id.org/fair-eo#>
PREFIX it6: <http://data.europa.eu/it6/>

SELECT ?method ?dataset ?value WHERE {
  ?run it6:realizes ?model ;
       feo:onDataset ?dataset ;
       feo:hasEvaluationResult ?er .
  ?model feo:basedOnMethod ?method .
  ?er feo:hasMeasure     feo:F1Score ;
      feo:hasAggregation feo:Macro ;
      feo:onSplit        feo:TestSplit ;
      feo:value          ?value .
} ORDER BY DESC(?value)
```

Leaderboards & cost-vs-quality

Rank methods on any measure, split and aggregation; join with GFLOPs and run time.

Cross-layer joins

Group results by architecture, foundation-model family, sensor or dataset property.

Reproducibility & green AI

Each value traces back to its run, split and checkpoint — reuse results, don't re-run.

Query pattern from the FAIR-EO schema documentation

FAIRness assessment using the F-UJI tool

Assessment Results:

Evaluated Resource:

FAIR-EO Hub
https://eodata.bvlabs.ai/ai4eo/

Save | {JSON} | New

FAIR level:	advanced
Resource PID/URL:	https://eodata.bvlabs.ai/ai4eo/
DataCite support:	enabled
Metric Version:	0.8
Metric Specification:	https://doi.org/10.5281/zenodo.15045911
Software version:	4.0.1
Download assessment results:	{JSON}
Save and share assessment results:0	
Harvested metadata:	show metadata

Summary:

	Score earned:	Fair level:
Findable:	7 of 7	advanced
Accessible:	7 of 7	advanced
Interoperable:	6 of 6	advanced
Reusable:	6 of 6	moderate

02

Benchmarks & lessons learned

What survives a single controlled protocol — in accuracy and in cost

Protocol: Trustworthy benchmarking rests on four pillars

1

Methodological fairness

Identical pipelines — splits, preprocessing, loss, hardware — isolate the architecture as the variable.

2

Transparency & reproducibility

Fixed, released splits; complete MLflow logging; open code, checkpoints and training logs.

3

Reliability

Evaluate across heterogeneous datasets, not a single leaderboard; report variance.

4

FAIR release

DOIs, open licences, model cards — and every run described in RDF with the FAIR-EO schema.

AiTLAS Benchmark Arena

500+ models · 10 architectures · 22 classification datasets

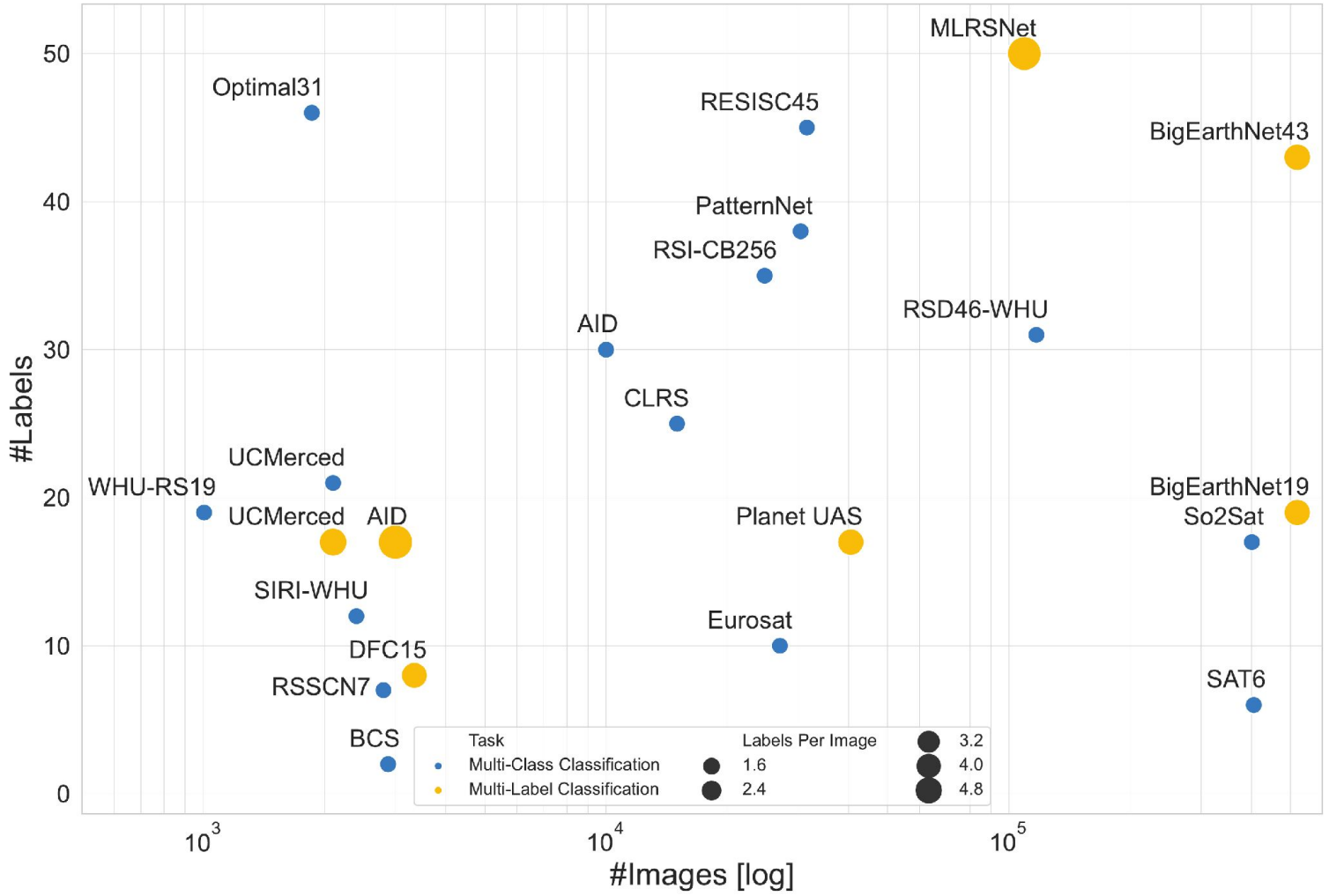
Dimitrovski et al., 2023, Current trends in deep learning for Earth Observation: An open-source benchmark arena for image classification, ISPRS Journal of Photogrammetry and Remote Sensing

Change detection benchmark

10 models · 10 datasets

Tomanič et al., 2026, A comprehensive and trustworthy benchmark of AI methods for change detection in Earth observation, ISPRS Journal of Photogrammetry and Remote Sensing (under review)

Image classification: 500+ models on 22 datasets



500+

trained models

10

architectures — AlexNet to ViT, Swin, ConvNeXt, MLP-Mixer

22

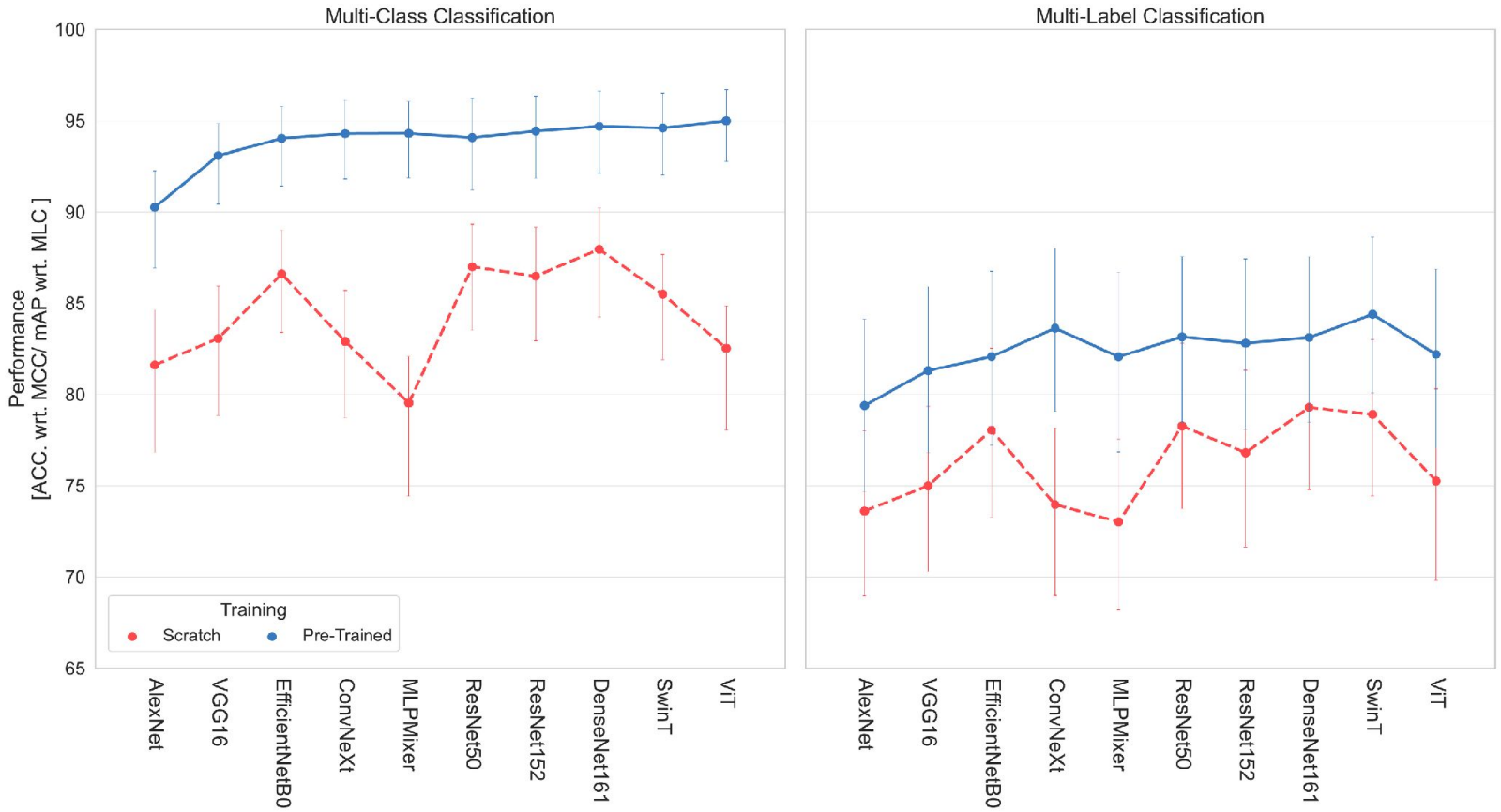
datasets: 15 multi-class, 7 multi-label

2

regimes: from scratch vs ImageNet-1K

Figure: Dimitrovski et al., arXiv:2207.07189, Fig. 1(a)

Pre-training pays off — and reshuffles the ranking



Scratch (red) vs ImageNet-1K pre-trained (blue); left multi-class, right multi-label

All 10 architectures improve

Largest gains on small datasets such as WHU-RS19, Optimal31 and UC Merced.

Transformers gain most

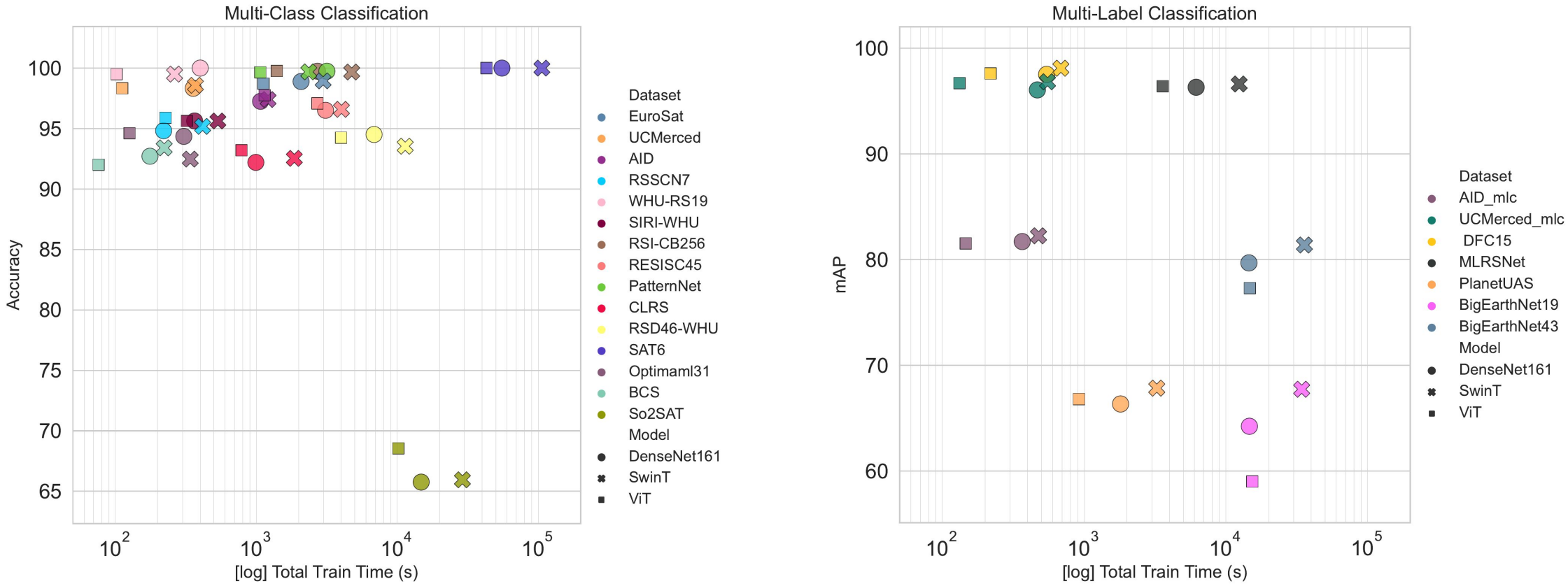
ViT and Swin rank near the bottom from scratch, at the top when pre-trained.

ImageNet is a hard baseline

In-domain EO pre-training alone did worse; ImageNet → EO helped on some tasks (up to 5%) but rarely justified the cost.

Figure: Dimitrovski et al., arXiv:2207.07189, Fig. 2

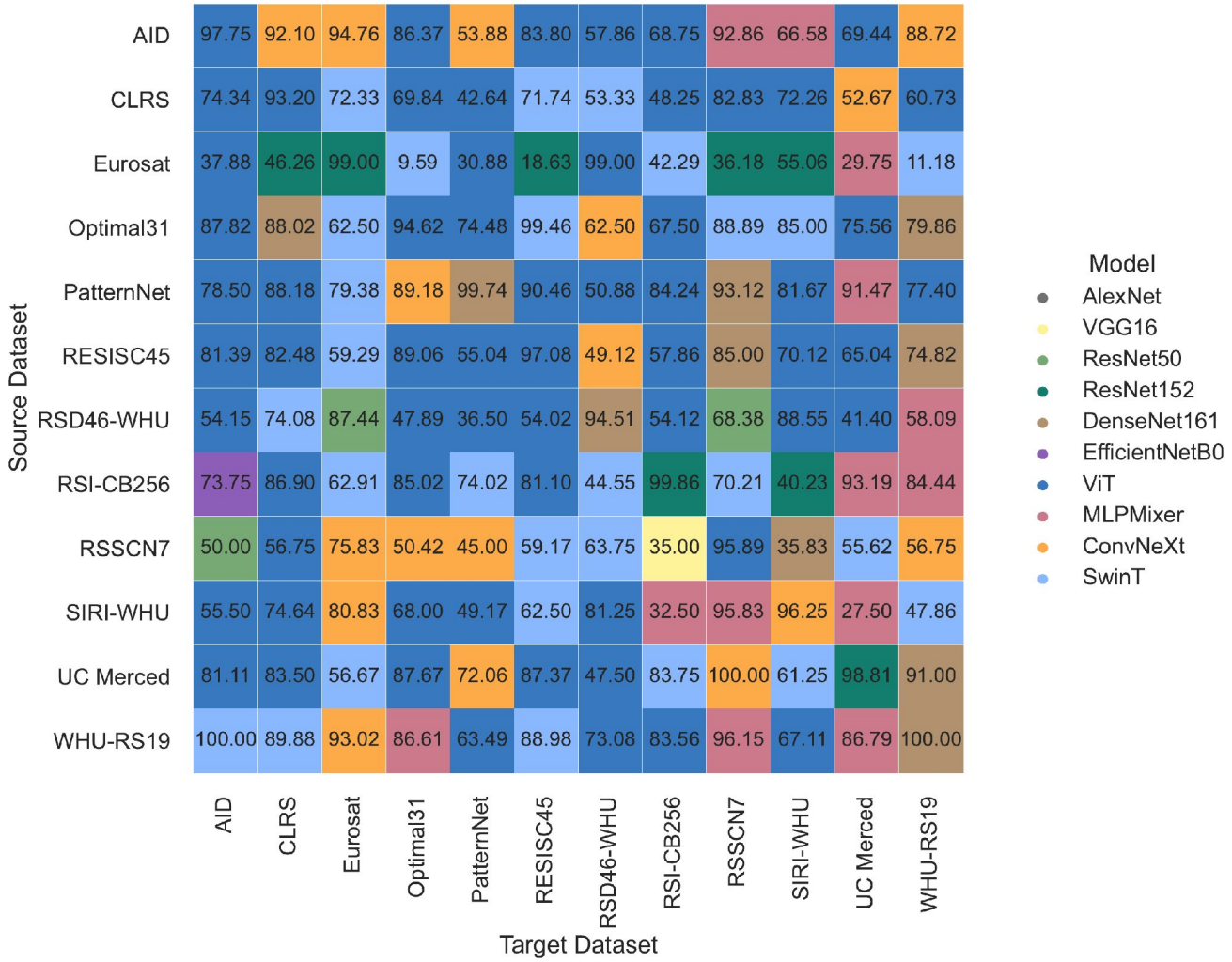
Similar accuracy, very different training cost



Top-3 pre-trained (ViT, DenseNet161, Swin): ViT matches DenseNet at about half the training time; Swin is best on multi-label but 2–3× slower. Fine-tuning roughly halves training time vs. scratch.

Figure: Dimitrovski et al., arXiv:2207.07189, Fig. 5 (log-scale training time)

In-domain accuracy is not generalisation



Easy datasets mislead

PatternNet and RSI-CB256 reach ~99% in-domain, but drop to 65–67% on a shared-label holdout set.

Diverse data transfers

Multi-resolution, label-rich datasets (CLRS, RESISC45) give the best transfer — ViT reaches 92.6% from CLRS.

Transformers generalise best

Best in about 2/3 of cross-dataset pairs; ViT alone in about 40%.

Figure: Dimitrovski et al., arXiv:2207.07189, Fig. 3 (source → target accuracy, best model per pair)

Change detection: 10 architectures × 10 datasets

Datasets — 0.03 to 10 m GSD, optical, SAR, off-nadir

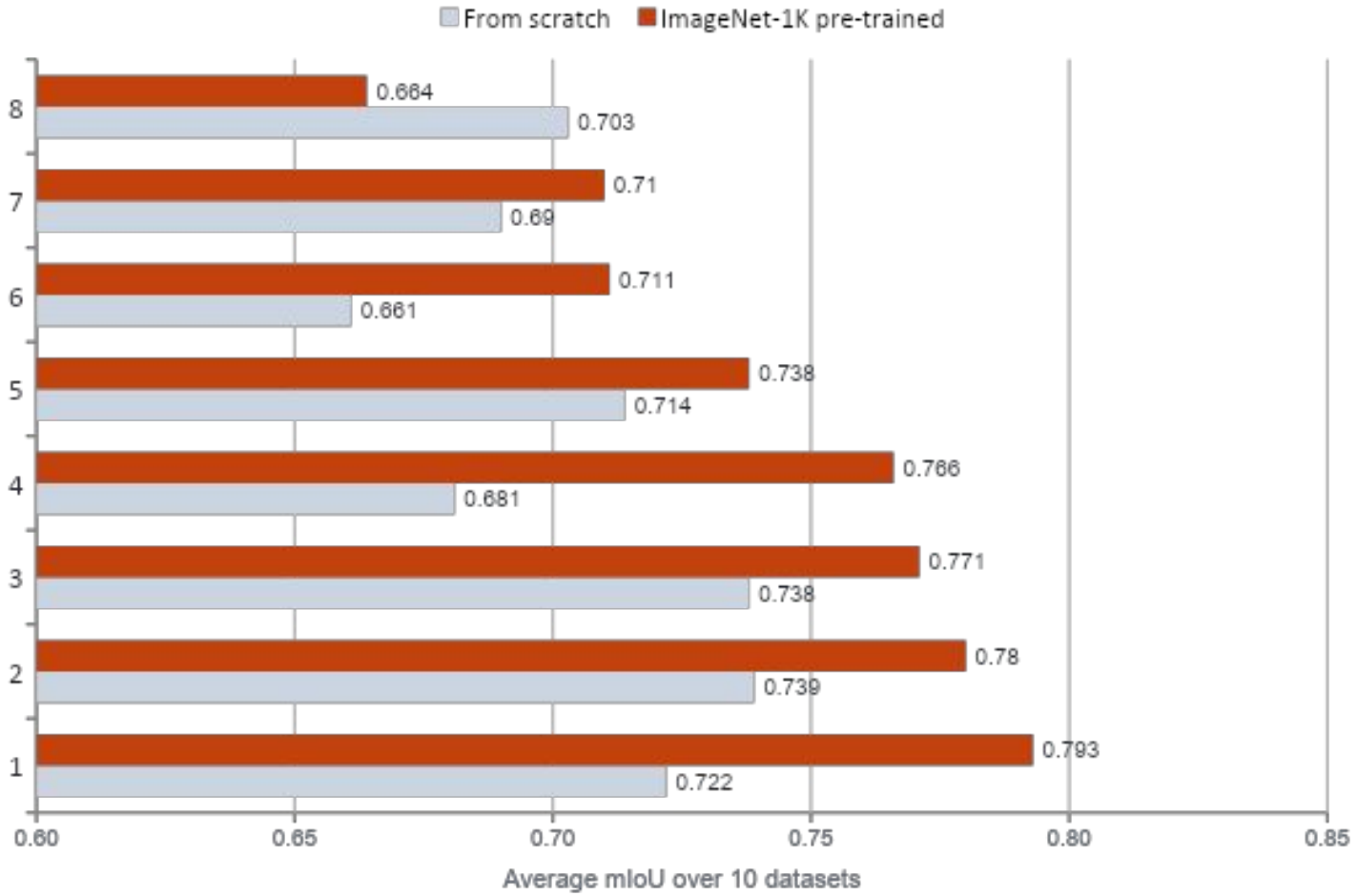
- **Urban & infrastructure**
  LEVIR-CD+, EGY-BCD, BANDON (off-nadir)
- **Disaster mapping**
  OMBRIA (Sentinel-1/2 floods)
- **Agriculture & land cover**
  CLCD
- **General & multi-sensor**
  Season-varying CDD, DSIFN, SYSU-CD, MSBC, MSOSCD

One protocol for all: 60/20/20 splits · 256×256 · focal + dice loss · 100 epochs, early stop 10 · lr 1e-4 · batch 16 · A100 40 GB · scratch vs ImageNet-1K

Model	Family	Year	Params (M)	GFLOPs
U-Net SiamConc	CNN	2018	40.35	19.37
HRNet SiamConc	CNN	2019	21.48	9.55
SiamCRNN	CNN + RNN	2020	28.51	65.49
TinyCD	CNN, lightweight	2023	0.29	1.46
STANet	Attention	2020	12.21	19.15
CGNet	Prior-guided	2023	33.68	87.55
BIT	Transformer	2021	3.50	10.88
ChangeFormerV6	Transformer	2022	41.03	138.77
ChangeViT	ViT	2026	20.66	26.36
CSSM	State-space (Mamba)	2025	2.59	2.78

Tomanič et al., arXiv:2608.28247, Tables 1–2 (GFLOPs for a 2×3×256×256 input)

A 2018 Siamese U-Net tops the benchmark



0.793

U-Net SiamConc, pre-trained: best average mIoU, top on 7 of 10 datasets

7 of 8

models gain from pre-training (+0.02 to +0.085 mIoU); CGNet degrades — unstable training

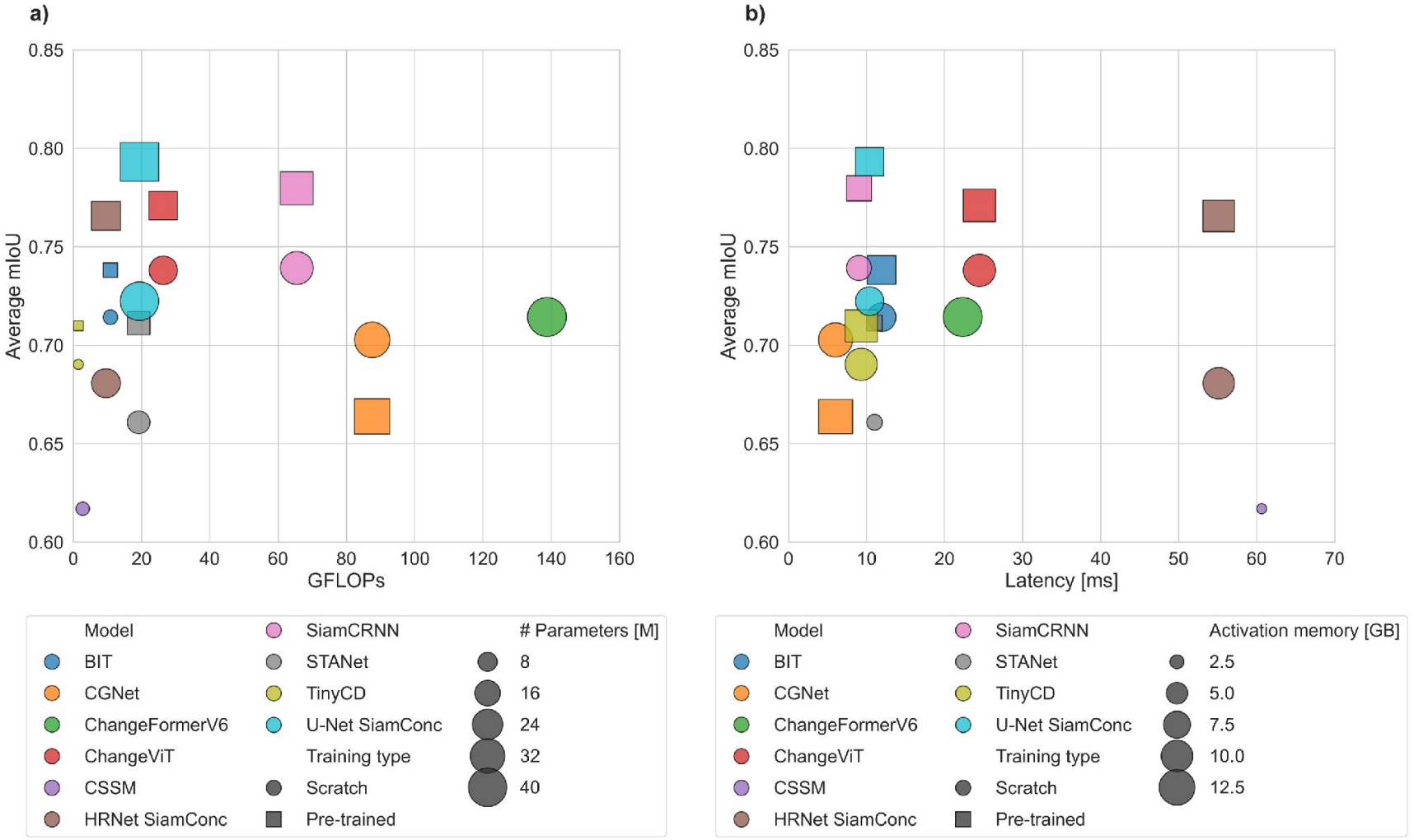
0.54 → 0.89

mean mIoU from BANDON (off-nadir) to Season-varying CDD: dataset spread exceeds model spread

ChangeFormerV6 (0.714) and CSSM (0.617): scratch only — no ImageNet weights available.

Data: Tomanič et al., arXiv:2608.28247, Table 3 — verified against the 180 hub runs

Once cost is counted, simple architectures win



U-Net SiamConc

0.793 mIoU at 10.4 ms and 19 GFLOPs

SiamCRNN

0.780 at 9.0 ms and 6.6 GB — a leaner alternative

ChangeFormerV6

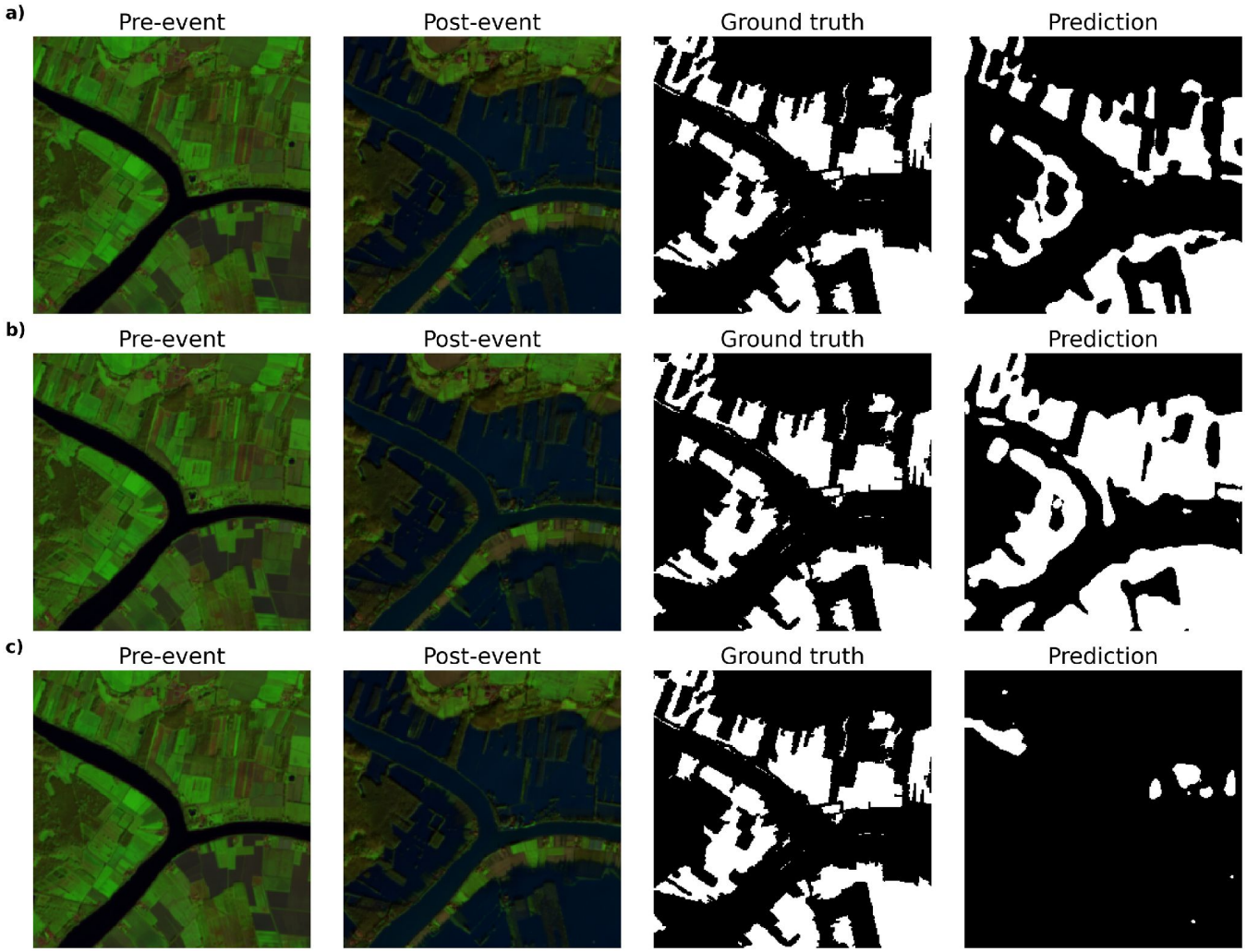
139 GFLOPs, 14.8 GB, mid-pack accuracy (0.714)

Pre-training

An accuracy gain at zero extra inference cost

Figure: Tomanič et al., arXiv:2608.28247, Fig. 4 (top-left and small = best)

Averages hide failure modes



OMBRIA flood mapping (Sentinel-1/2), models trained from scratch

a ChangeFormerV6 — best on OMBRIA (0.693): smooth, faithful boundaries

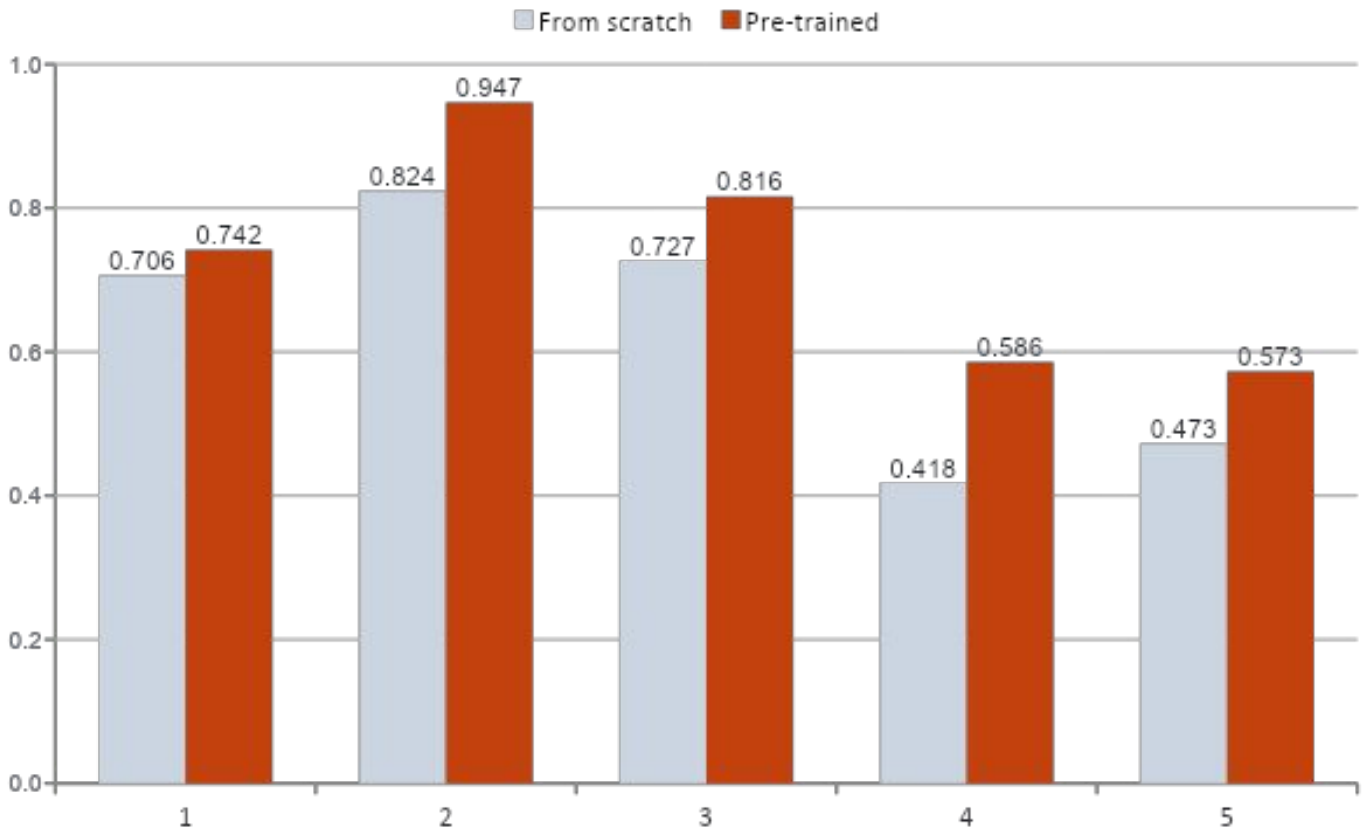
b U-Net SiamConc — best overall, 0.630 here

c STANet — worst here (0.580): misses most of the change

The overall winner is not the per-dataset winner — select models per task and sensor.

Figure: Tomanič et al., arXiv:2608.28247, Fig. 6

Five tasks, 880 runs, one protocol



Mean primary metric over paired runs (same dataset and model, both regimes)

385 / 430

paired runs where pre-training wins

Best pre-trained model (avg. rank)

Change detection	**U-Net SiamConc (1.4)**
Multi-class	**Swin Transformer (2.3)**
Multi-label	**ConvNeXt-T (1.4)**
Object detection	**FCOS (1.8)**
Segmentation	**HRNet (2.0)**

DETR and Sparse R-CNN score mAP@50 ≈ 0 from scratch on all 10 detection datasets — usable only when pre-trained.

Preliminary: queried from the FAIR-EO Hub (45 datasets, 50 models), October 2026

Lessons on results

1

Controlled re-runs deflate claimed progress

Under one protocol a 2018 Siamese U-Net beats transformers and state-space models for change detection.

2

Pre-training is the most reliable gain

All 10 Arena architectures, 7 of 8 change detection models, 385 of 430 hub pairs — and transformers need it most.

3

There is no universal winner

Winners change with task, sensor and dataset; dataset difficulty often varies more than model quality.

4

In-domain scores are not generalisation

Saturated datasets (~99%) neither separate models nor produce models that transfer.

Lessons on performance, cost and practice

139 vs 19

GFLOPs

Cost does not buy accuracy: the most expensive change detection model is mid-pack; the best one is cheap and fast.

0 ms

extra inference

Pre-training improves accuracy at no inference cost, and roughly halves fine-tuning time.

33 runs

accuracy > 0.85, mIoU < 0.2

Accuracy hides failure on imbalanced segmentation — report mIoU / F1 and the full metric set.

1 protocol

exposes fragility

CGNet and Swin (multi-label) are unstable under fixed hyperparameters: report seeds, variance and tuning budget.

~7,700

GPU-hours in the hub

Released runs and checkpoints let others reuse results instead of re-running them.

Run counts and GPU-hours from the FAIR-EO Hub (preliminary)

From benchmarks to knowledge bases

What AutoML needs from the empirical record — and where the graph provides it

AutoML needs	The FAIR-EO graph provides
Task meta-features	Dataset layer: bands, GSD, classes, size, modality, task type
A configuration space	Method layer: architecture, family, foundation vs traditional, pre-training
Performance and cost	Evaluation & resource estimates: measure, split, aggregation, GFLOPs, run time
Trustworthy provenance	Run: protocol, splits, parameters, checkpoints and logs

Benchmarks become training data for meta-learning, performance modelling and cost-aware selection — queried, not scraped from PDFs.

Thank you

Open, standards-based and queryable — a foundation for reproducible AI4EO.

FAIR-EO Hub
eodata.bvlabs.ai/ai4eo

Change detection benchmark
arXiv:2608.28247 · github.com/biasvariancelabs/FAIR-EO-CD-benchmark

AiTLAS Benchmark Arena
arXiv:2207.07189 · github.com/biasvariancelabs/aitlas-arena

AiTLAS toolbox
github.com/biasvariancelabs/aitlas

Dragi Kocev, Panče Panov, Alice Baudhuin, Tadej Tomanič, Jan Sotošek, Jure Brence, Nikola Simidjievski

Bias Variance Labs & Jožef Stefan Institute · FAIR-EO funded under OSCARS (Horizon Europe, GA No. 101129751)

OSCARS
Open Science Clusters' Action for Research & Society

Funded by the European Union

GeoAI

SLAIF
Slovenian AI Factory

Explore the hub