



Design Choices Sensitivity

in Remote Sensing Multi-Label Classification

Maryam Gholami Shiri^{1 2} · Eva Tuba^{3 4 5} · Sašo Džeroski¹ · Tome Eftimov^{2 3} · Ana Nikolij^{2 3}

¹ Jožef Stefan Institute ² Jožef Stefan International Postgraduate School ³ Computer Systems Dept., JSI ⁴ Trinity University ⁵ Singidunum University

Keywords: multi-label classification · remote sensing · functional ANOVA · benchmarking · deep learning design choices



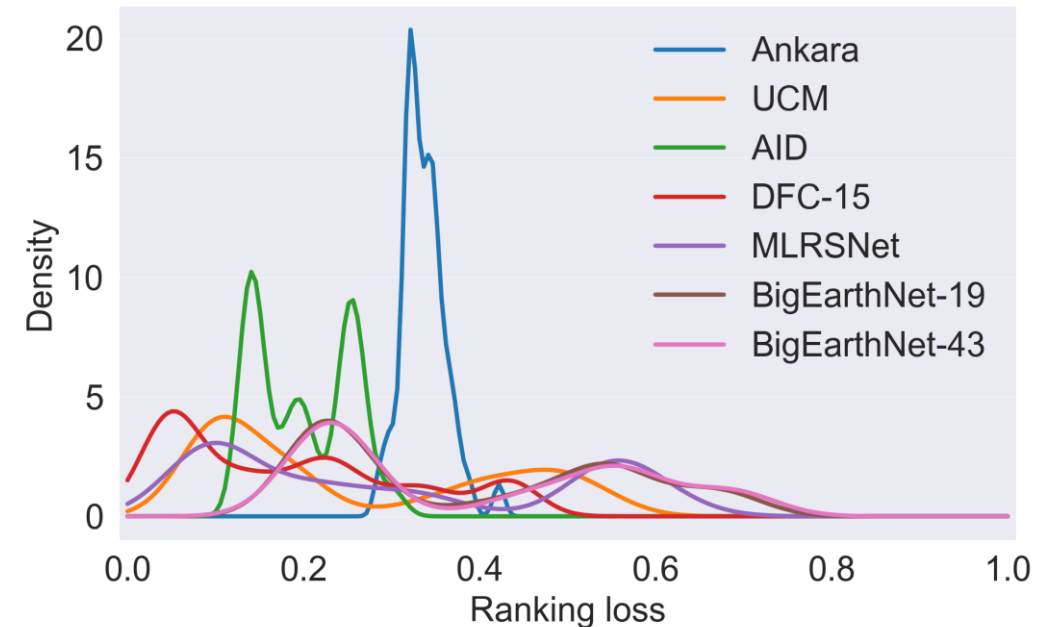
Problem Definition: The Limits of Aggregate Benchmark Rankings

The MLC benchmarking challenge

Which design choices govern deep learning (DL) performance on remote sensing (RS) images?

- Prior benchmark studies [1] compare DL approaches — architectures, fine-tuning and learning strategies — for RS MLC.
- Reporting a single “winner” conflates two distinct effects:
 - it hides the importance of individual design choices and their interactions;
 - it does not relate performance to the properties of the dataset it was evaluated on.

Do the design choices that matter for one dataset generalize to others, or is their importance dataset-dependent?

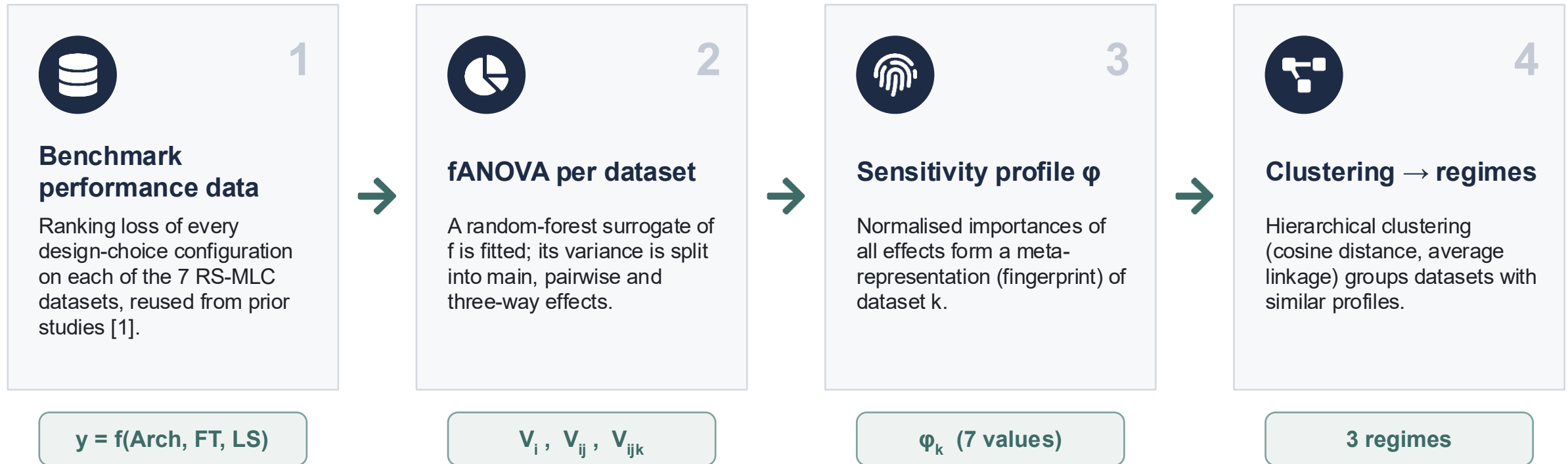


Raw performance is highly variable across datasets, even before asking why.

Fig. 1 — Kernel density estimates of raw performance (ranking loss) across the seven datasets.

[1] Stoimchev, M., Koccev, D., Džeroski, S. Deep network architectures as feature extractors for multi-label classification of remote sensing images. Remote Sensing 15(2), 538, 2023.

Methodology: From Benchmark Results to Performance Regimes



Contributions



[4] Dimitrovski, I., Kitanovski, I., Kocev, D., Simidjievski, N. Current trends in deep learning for Earth Observation: an open-source benchmark arena for image classification. ISPRS J. Photogramm. Remote Sens. 197, 2023.

Methodology: Functional ANOVA Variance Decomposition

The idea: decompose performance variance by design choice

For a model described by design-choice modules $x = (x_1, \dots, x_d)$ with performance $y = f(x)$, the total variance is decomposed into main effects and interactions between modules.

$$V(f) = \sum_i V_i + \sum_{i < j} V_{ij} + \sum_{i < j < k} V_{ijk} + \dots$$

$$V_i = \text{Var}(E[f(x) \mid x_i])$$

Main effect of module i — variance of the marginal (expected performance holding x_i fixed, averaging over the other modules).

$$V_{ij} = \text{Var}(f_{ij}) - V_i - V_j$$

Pairwise interaction — joint effect of modules i and j beyond their individual contributions.

$$R^2 \text{ (surrogate fit)}$$

Coefficient of determination of the random-forest surrogate that approximates f — reported per dataset as a reliability check [2].

Exact computation is intractable: fANOVA approximates it via a random-forest surrogate [2], extended to multi-dataset benchmarking following Nikolikj & Eftimov [3].

[2] Hutter, F., Hoos, H., Leyton-Brown, K. An efficient approach for assessing hyperparameter importance. ICML, 2014. [3] Nikolikj, A., Eftimov, T. Exploring module interactions in modular CMA-ES across problem classes. Swarm Evol. Comput. 98, 2025.

Experimental Setup: Design Choices

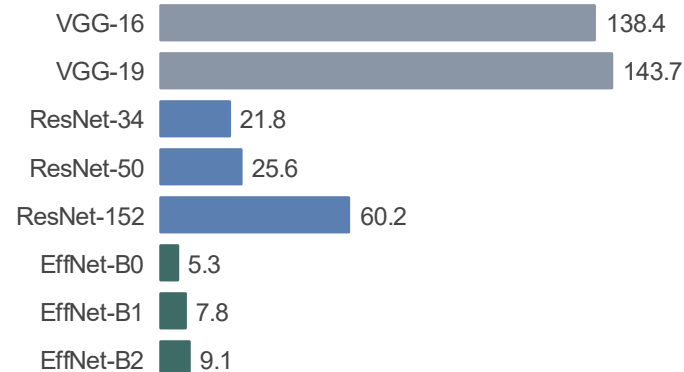


Goal of the reference study [1]: find which CNN backbone and transfer-learning strategy work best for RS-MLC — used end-to-end, or as a feature extractor for tree ensembles. Its 48 configurations $\times$ 7 datasets are the performance data we decompose with fANOVA.

Architecture (Arch)

8 options

VGG · ResNet · EfficientNet families



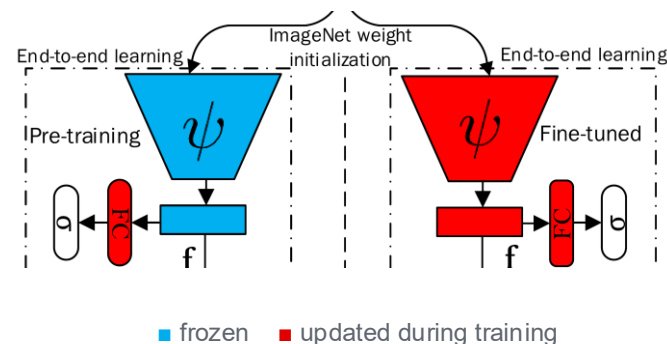
Parameters (millions) — VGG · ResNet · EfficientNet families

$8 \times 2 \times 3 = 48$ models per dataset

Fine-tuning (FT)

2 options

Pre-trained (frozen backbone) vs. fine-tuned (all layers)



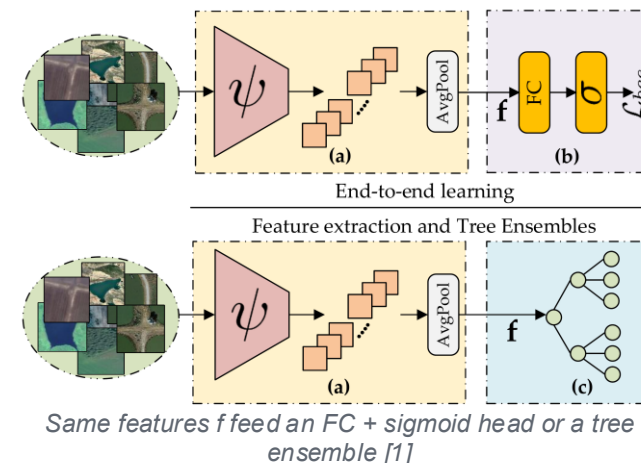
ImageNet weights: freeze the backbone ψ (left) or fine-tune it (right) [1]

$\dim(\phi) = 3 + C(3,2) + C(3,3) = 7$
3 main + 3 pairwise + 1 three-way effect

Learning strategy (LS)

3 options

End-to-end · Random Forest · Extra Trees



Metric: ranking loss
threshold-independent · lower is better

Experimental Setup: Datasets and Meta-Features



Ankara



DFC-15



UCM



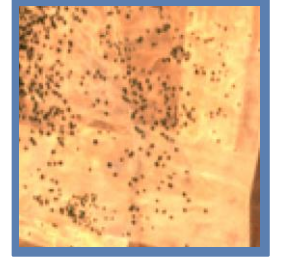
AID



MLRSNet



BigEarthNet-19



BigEarthNet-43

Dataset	#Labels	Cardinality	Density	#Samples	Image size
Ankara	29	9.12	0.536	216	64 ²
DFC-15	8	2.80	0.465	3,341	600 ²
UCM	17	3.33	0.476	2,100	256 ²
AID	17	5.15	0.468	3,000	600 ²
MLRSNet	60	5.77	0.144	109,151	256 ²
BigEarthNet-19	19	2.90	0.263	590,326	256 ²
BigEarthNet-43	43	2.97	0.247	590,326	256 ²

Groups = clusters found by the methodology (Slide 8), not chosen a priori

 Data-limited (Cluster 3)

 Medium-scale (Cluster 2)

 Large-scale (Cluster 1)

Cardinality = average number of labels per image.
Density = cardinality / #labels (average proportion of labels present).

Scale spans three orders of magnitude (216 → 590,326 samples); label space ranges from 8 to 60 classes.

Results: Main Effects Dominate the Variance Decomposition

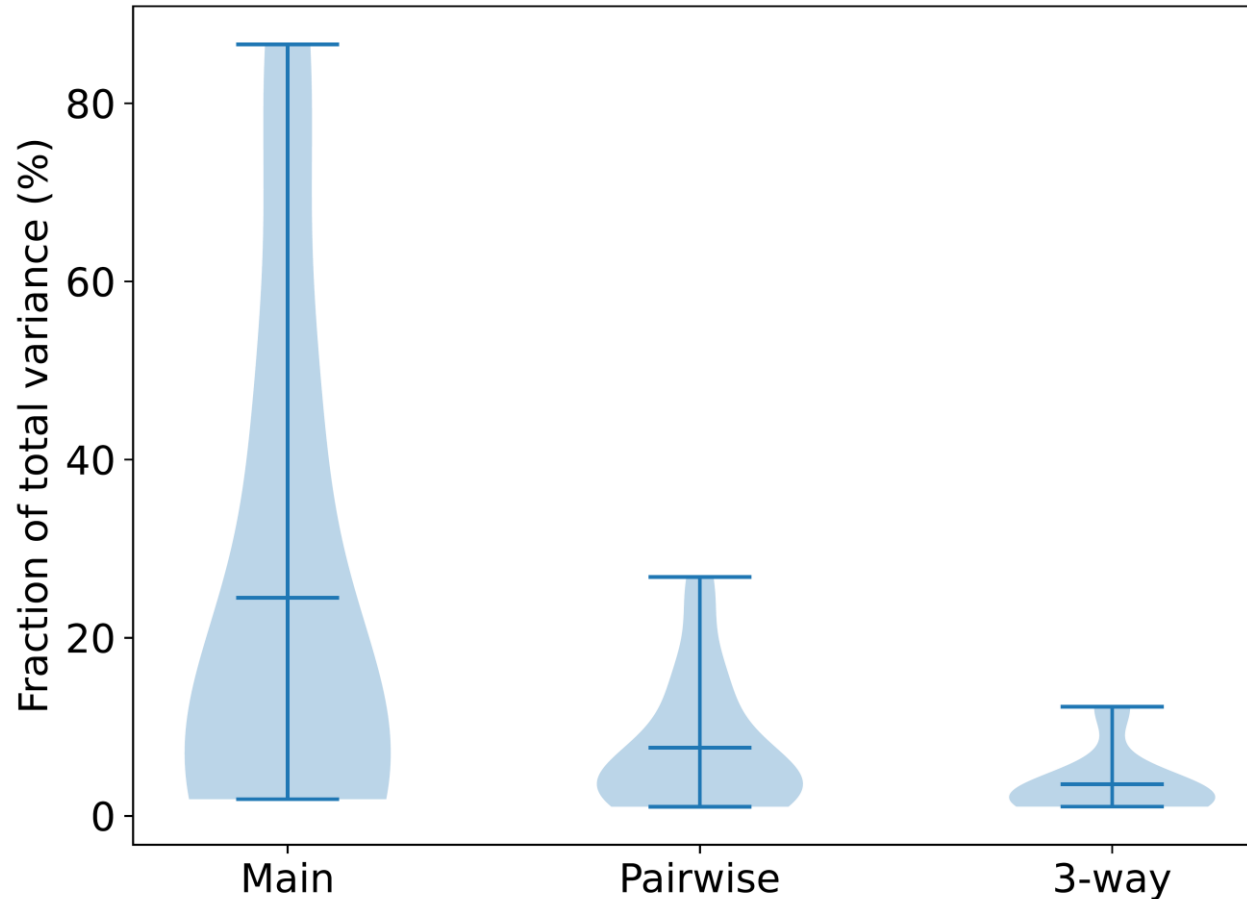


Fig. 2 — Share of total variance explained by each effect, grouped by effect order (all datasets).

0.966

Median R^2 of the surrogate

The random-forest surrogate reproduces the benchmark almost perfectly, so the decomposition is reliable.

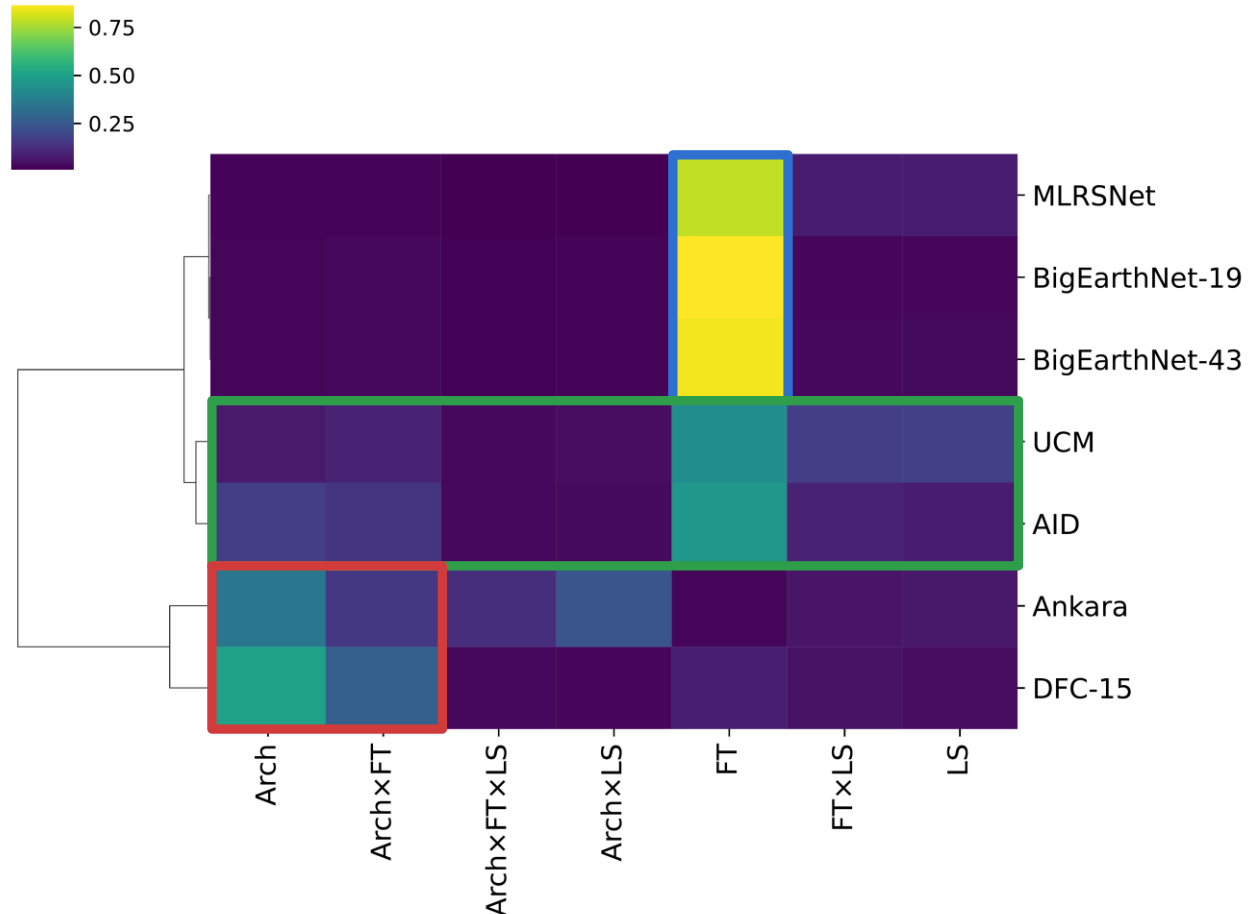
>80%

Main-effect share, best case

A single design choice can explain most of the variation in ranking loss on a dataset.

Interactions matter less on average — but, as the next slides show, not on every dataset.

Results: Dataset Clustering by Design-Choice Importance



Cluster 1 — Large-scale

BigEarthNet-19/43, MLRSNet

Fine-tuning dominates; the other design choices are negligible.

Cluster 2 — Medium-scale

UCM, AID

Fine-tuning importance drops; performance depends on combinations of choices.

Cluster 3 — Data-limited

Ankara, DFC-15

Architecture becomes critical, and its interaction with fine-tuning grows.

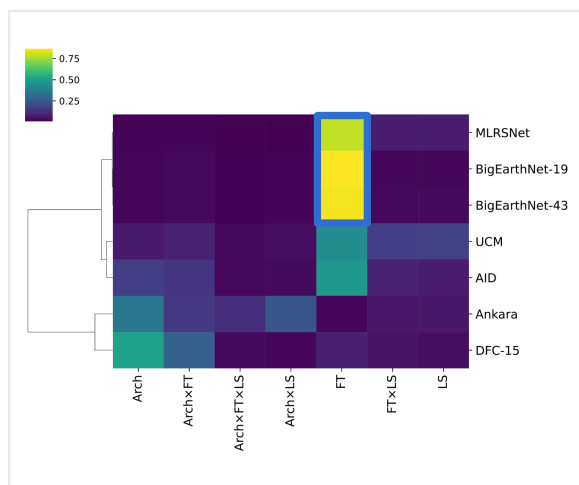
Fig. 3 — Clustermap of module-effect contributions per dataset (yellow = higher). Silhouette = 0.757, cophenetic corr. = 0.967.

Results: What Drives Each Cluster?

Cluster 1 · Large-scale · Fine-tuning dominant

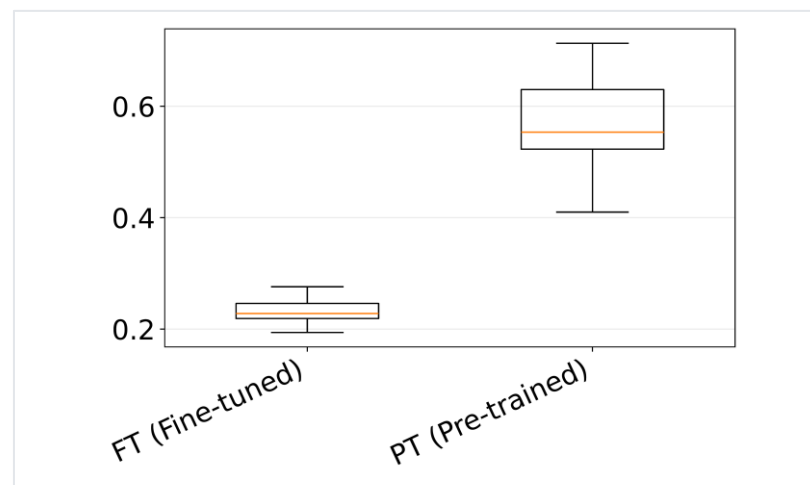
Representative dataset: BigEarthNet-43

Reminder: Fig. 3 clustermap



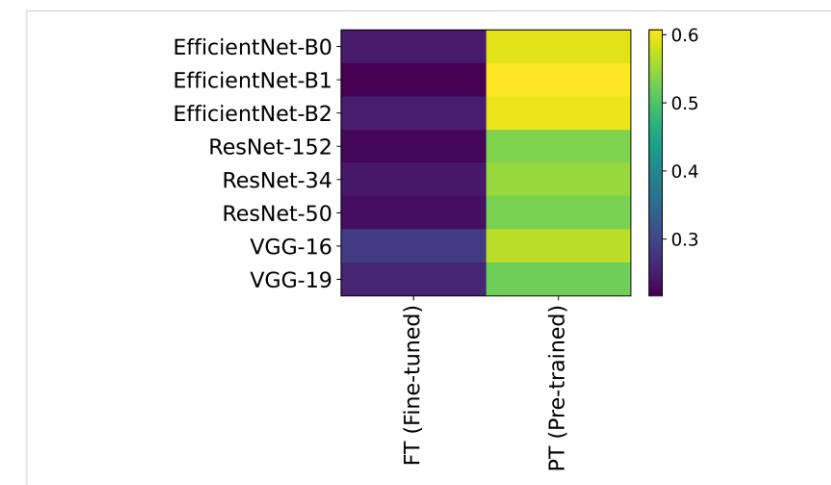
Cluster 1 · Large-scale
BigEarthNet-19/43, MLRSNet

(a) Main effect: fine-tuning strategy



Why it matters: Fine-tuned models are clearly and consistently better. Satellite images differ strongly from ImageNet photos, and with this much data the whole network can adapt safely.

(b) Interaction: architecture × fine-tuning



Why it matters: Once fine-tuned, the backbones become interchangeable; architecture matters only for frozen features. Adapting the features outweighs which network produces them.

Large data → adapting the representation is the lever; which backbone you pick is secondary.

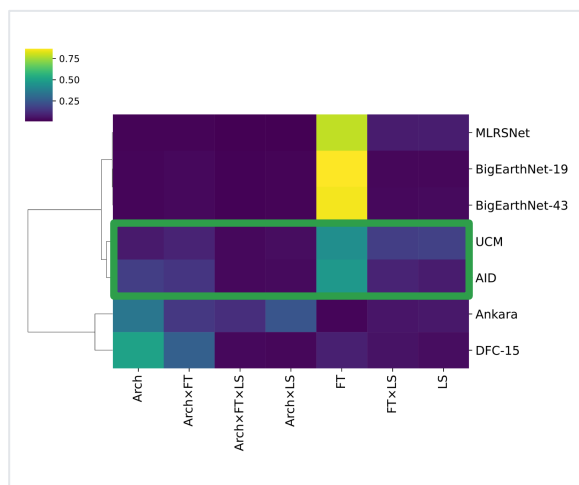
Fig. 4 — BigEarthNet-43: (a) main-effect box plot, (b) interaction heatmap of mean ranking loss (lower is better).

Results: What Drives Each Cluster?

Cluster 2 · Medium-scale · Interactions dominant

Representative dataset: UCM

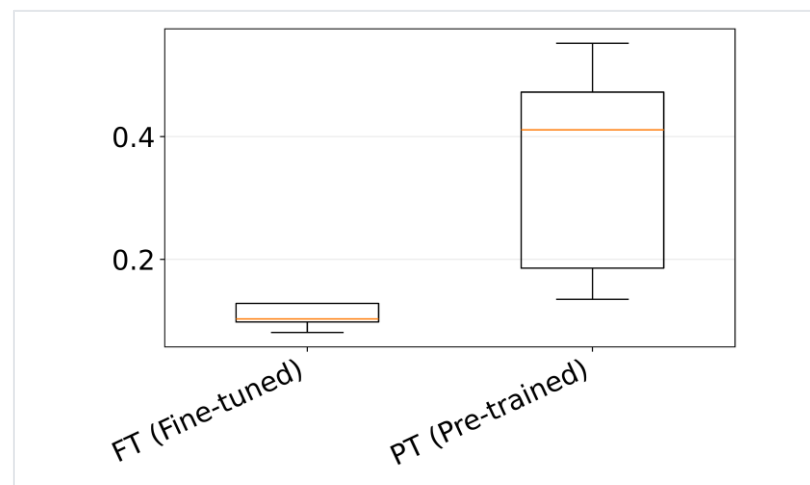
Reminder: Fig. 3 clustermap



Cluster 2 · Medium-scale

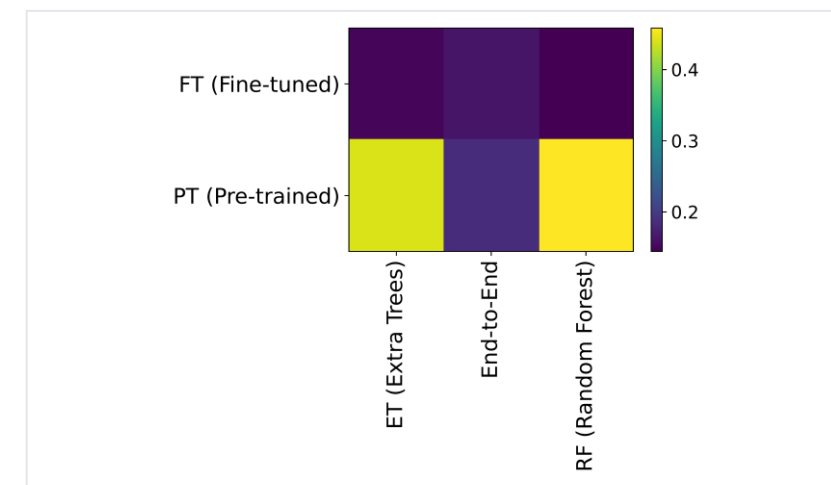
UCM, AID

(a) Main effect: fine-tuning strategy



Why it matters: Fine-tuning still helps, but frozen models range from nearly as good to much worse. Fine-tuning alone no longer predicts performance.

(b) Interaction: fine-tuning × learning strategy



Why it matters: Frozen features still work with an end-to-end head, but tree ensembles struggle with them. Which learning strategy works best depends on the fine-tuning choice.

Medium data → no single choice rules; the combination of fine-tuning and learning strategy decides.

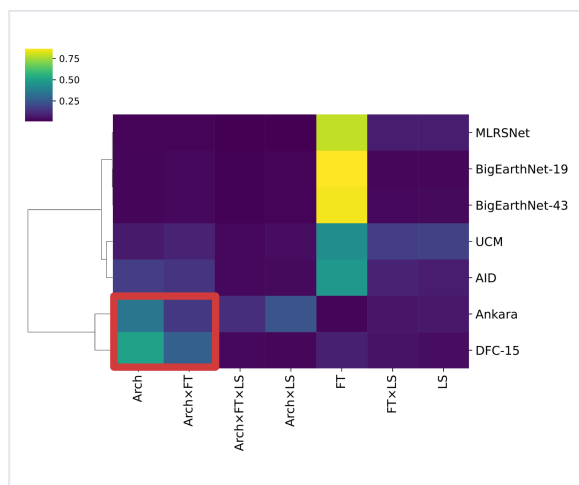
Fig. 5 — UCM: (a) main-effect box plot, (b) interaction heatmap of mean ranking loss (lower is better).

Results: What Drives Each Cluster?

Cluster 3 · Data-limited · Architecture dominant

Representative dataset: DFC-15

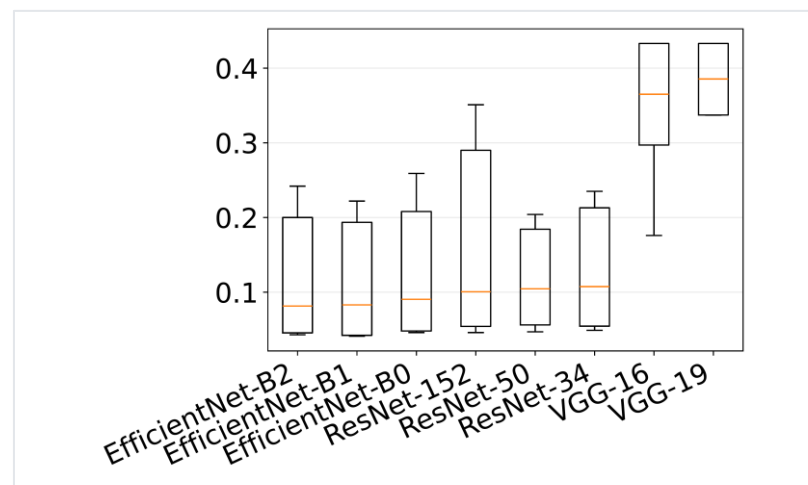
Reminder: Fig. 3 clustermap



Cluster 3 · Data-limited

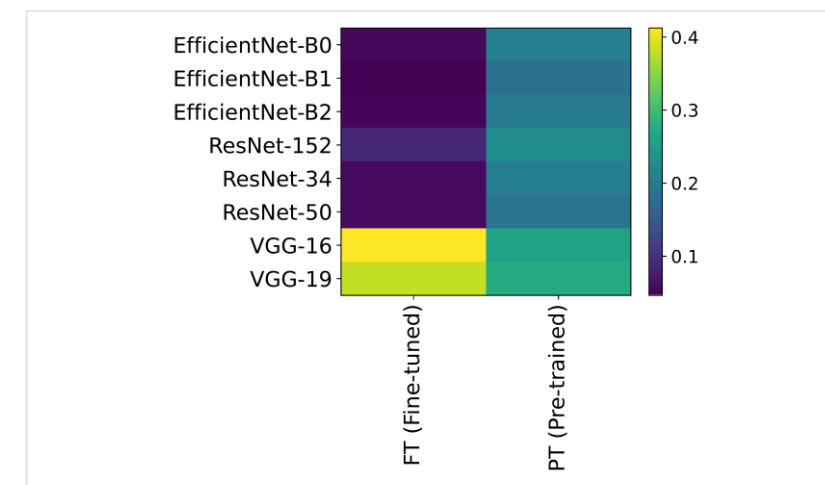
Ankara, DFC-15

(a) Main effect: architecture



Why it matters: Modern, compact backbones (EfficientNet, ResNet) perform well, while the much larger VGG models lag behind. With few images, the choice of network carries most weight.

(b) Interaction: architecture × fine-tuning



Why it matters: Fine-tuning improves modern backbones but makes VGG worse, plausibly because its ~140M parameters overfit scarce data. Whether it helps depends on the backbone.

Limited data → choose a lightweight, well-matched backbone first; fine-tune only if it is robust.

Fig. 6 — DFC-15: (a) main-effect box plot, (b) interaction heatmap of mean ranking loss (lower is better).

Regime Summary

Regime	Datasets	#Samples	Dominant factor	Mechanism
Large-scale	BigEarthNet-19 / -43, MLRSNet	109K – 590K	Fine-tuning	Adapting pretrained representations to the target domain is the critical lever; backbones perform alike once fine-tuned.
Medium-scale	UCM, AID	2.1K – 3K	Interaction of modules	No single factor rules; the joint choice of fine-tuning and learning strategy matters most.
Data-limited	Ankara, DFC-15	216 – 3.3K	Architecture (× fine-tuning)	Simple, well-matched backbones compensate for scarce data; unstable architectures (VGG) fail when fine-tuned.

Regimes align with intrinsic dataset meta-features (scale, resolution, label complexity) — dataset properties, not just model choice, shape which design decision is worth optimizing.

Universal Conclusions Become Dataset-Conditional

Original claim	This work's dataset-conditional finding
“Fine-tuning generally preferred” [1]	Holds strongly for large-scale data (BigEarthNet, MLRSNet); weakens for medium-scale (UCM, AID); architecture dominates instead for data-limited Ankara / DFC-15.
“EfficientNet-B2 among strongest architectures” [1]	Architecture importance is regime-dependent — decisive only where fine-tuning is not, i.e., in data-limited settings.

Large-scale
matters most: Fine-tuning

Medium-scale
matters most: Interactions

Data-limited
matters most: Architecture

Benchmarking conclusions reported as universal are, on closer inspection, contingent on where a dataset sits along the scale / resolution / label-complexity spectrum.

Limitations and Future Work

Scope constraints

- Restricted to CNN and early transformer architectures; at most three modules considered.
- Performance data reused from prior studies — clustering based on only seven datasets per scenario.

Future work

- Predict a dataset's sensitivity profile directly from its meta-features (regression on ϕ).
- Extend the analysis to modern visual foundation models and RS-specific pretrained backbones.
- Broaden the dataset pool to strengthen clustering and meta-feature regression.
- Treat training-level choices (augmentation, optimizer) as additional fANOVA modules.

Does the regime-dependent pattern hold once modern foundation models and RS-specific pretraining are included?

Code & data: zenodo.org/records/19630021



Thank you for your attention

Questions?

Code & data: zenodo.org/records/19630021

