QUANTIFYING DATA CURATION EFFECTS IN VISION TRANSFORMERS FOR RADIO ASTRONOMY IMAGING

L-Università ta' Malta
Institute of Space Sciences & Astronomy

MACHINE LEARNING FOR ASTROPHYSICS
3RD EDITION

Hayley Camilleri
Ian Fenech Conti
Alessio Magro
Andrea DeMarco

The Model vs the Data

Model-centric

Invest in the architecture; refine the ViT backbone, tune self-supervised objectives, optimise training procedures. Ask: can we build a better model?

- ViT backbone
- SSL objectives
- Hyperparameter tuning

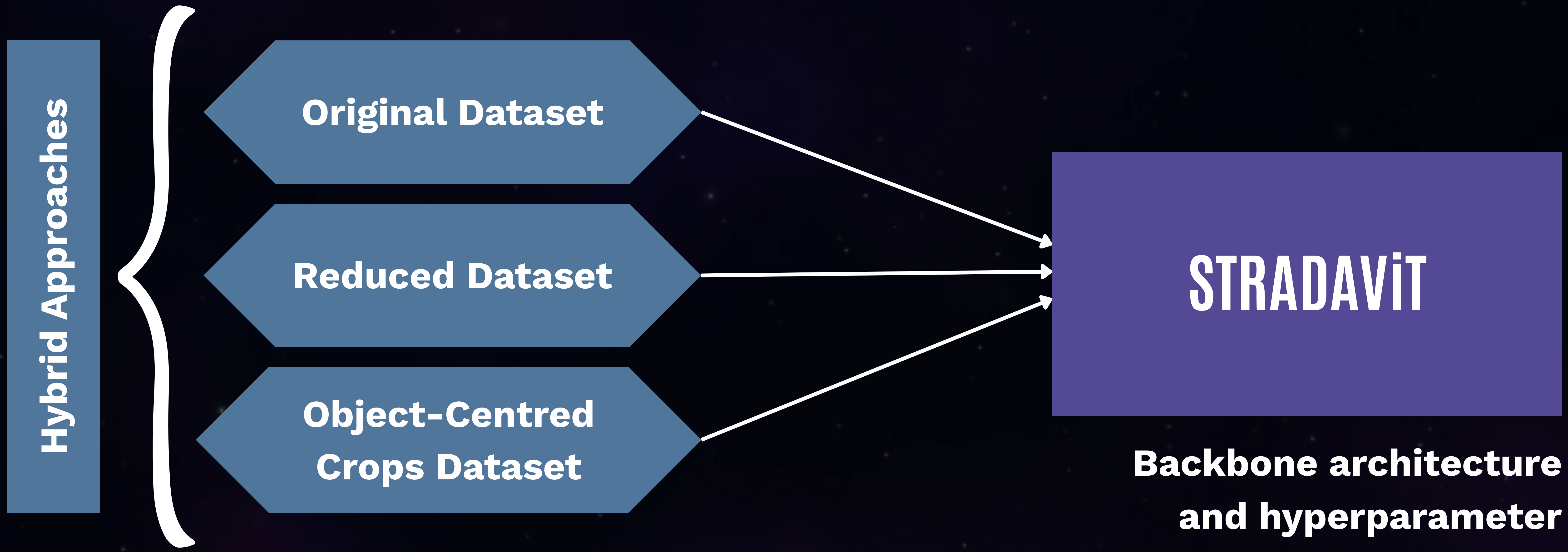
Data-centric

Hold the model fixed. Invest in what goes in; curation, augmentation, diversity. Ask: can better data unlock better performance?

- Dataset curation
- Cross-telescope diversity
- Augmentation strategies

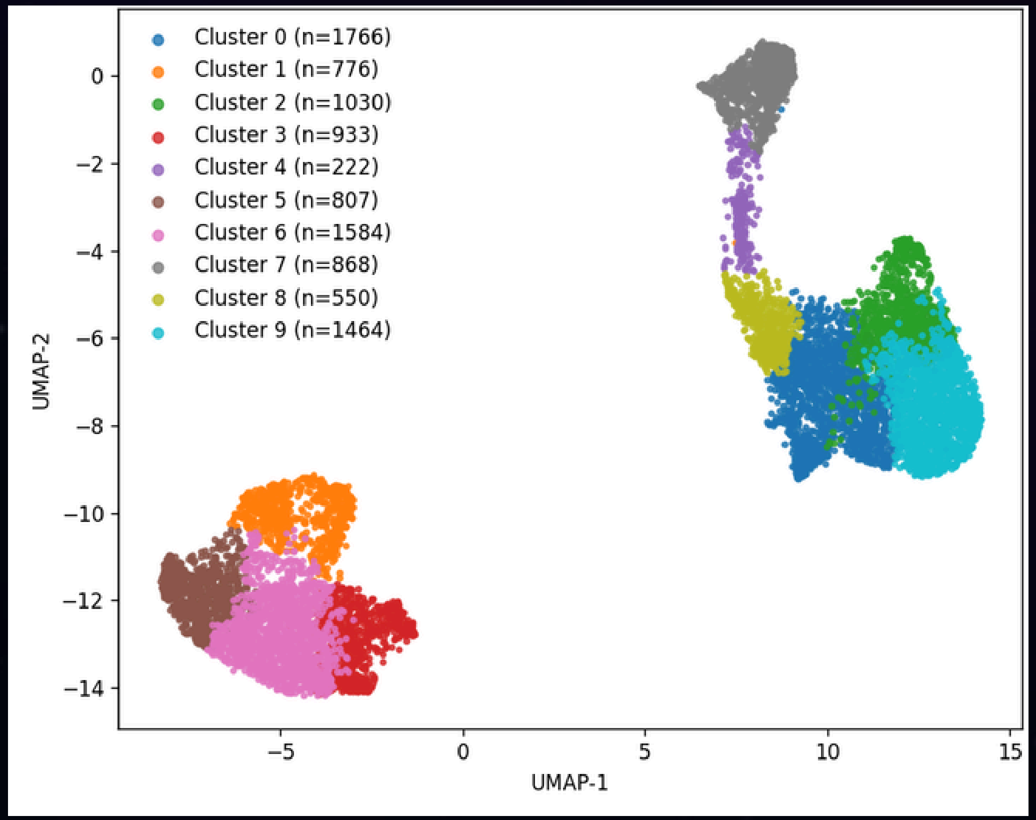
Can dataset quality and curation compensate for data volume or online selection in Vision Transformers?

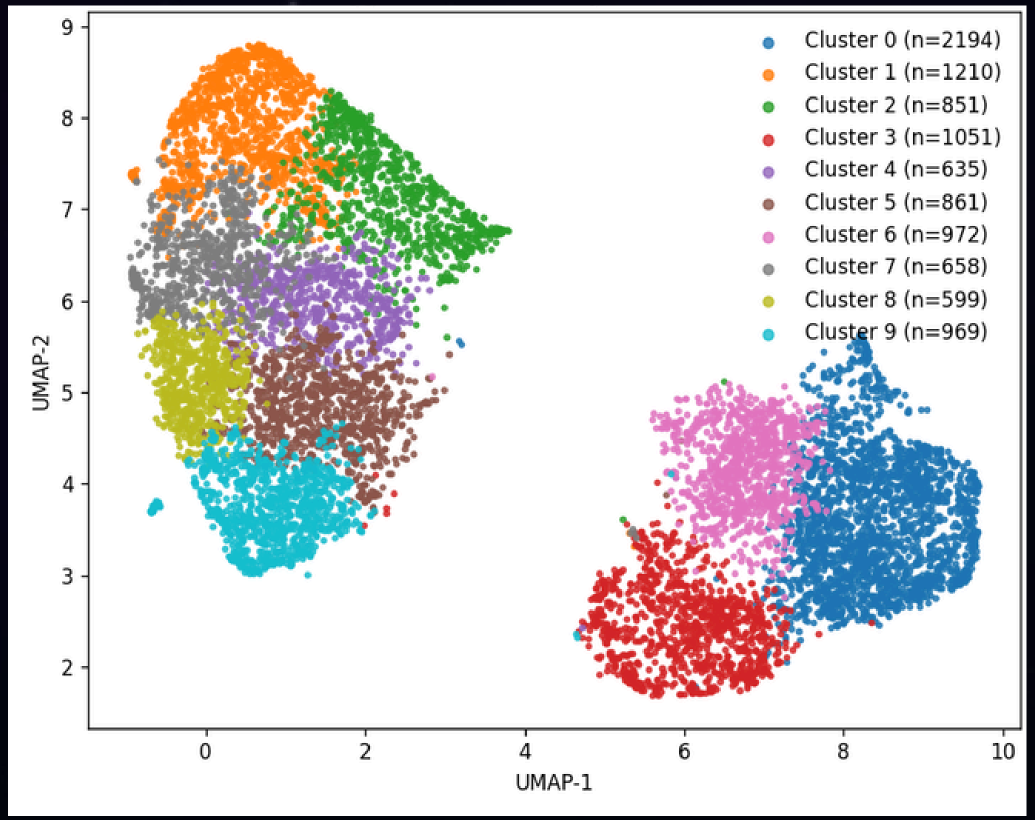
Controlled Curation Interventions

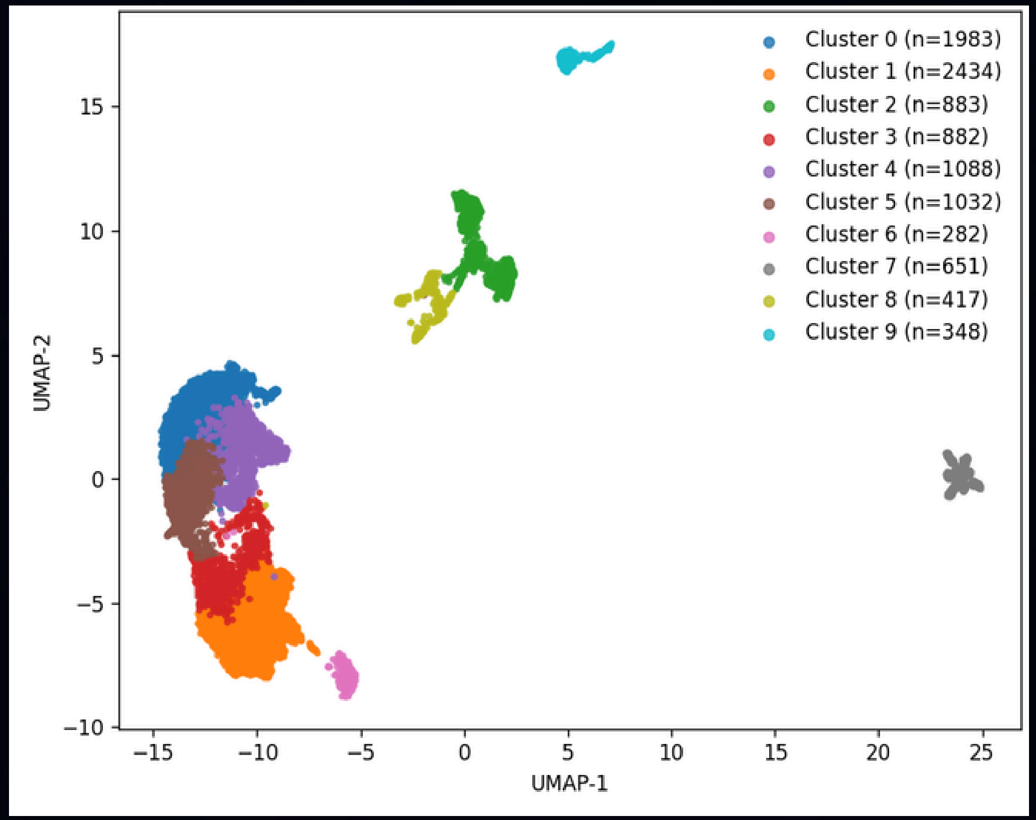


Backbone architecture and hyperparameter settings are kept identical

Reduced Dataset

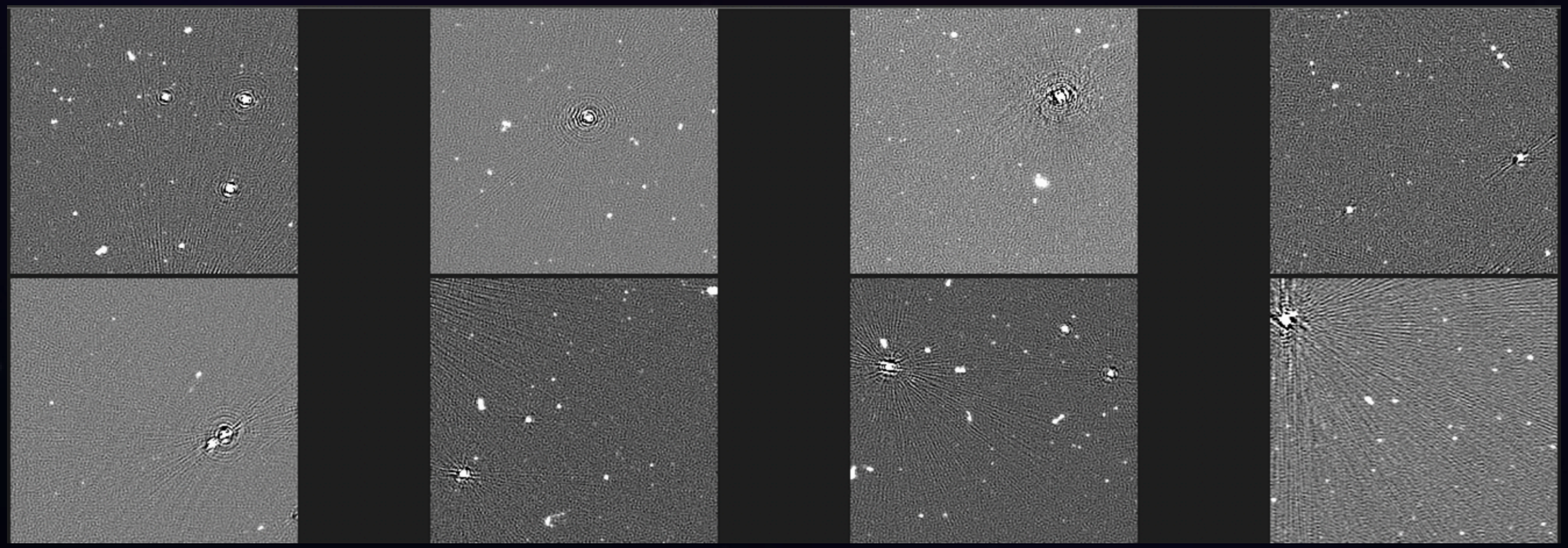
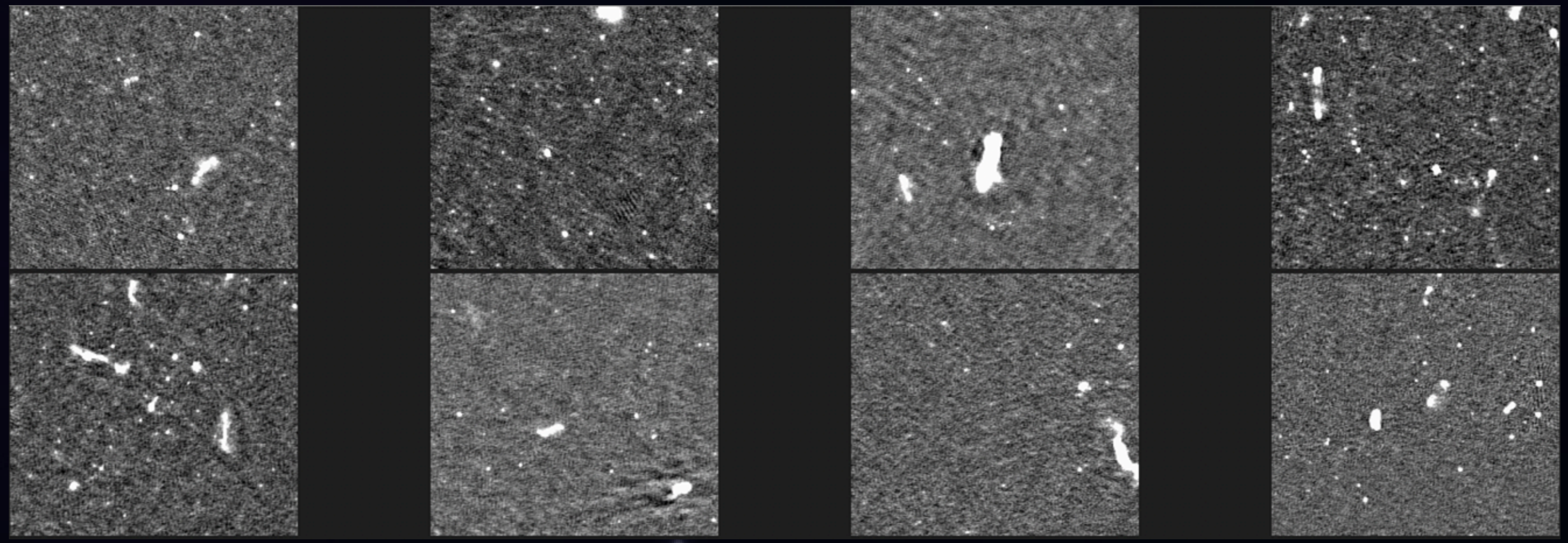
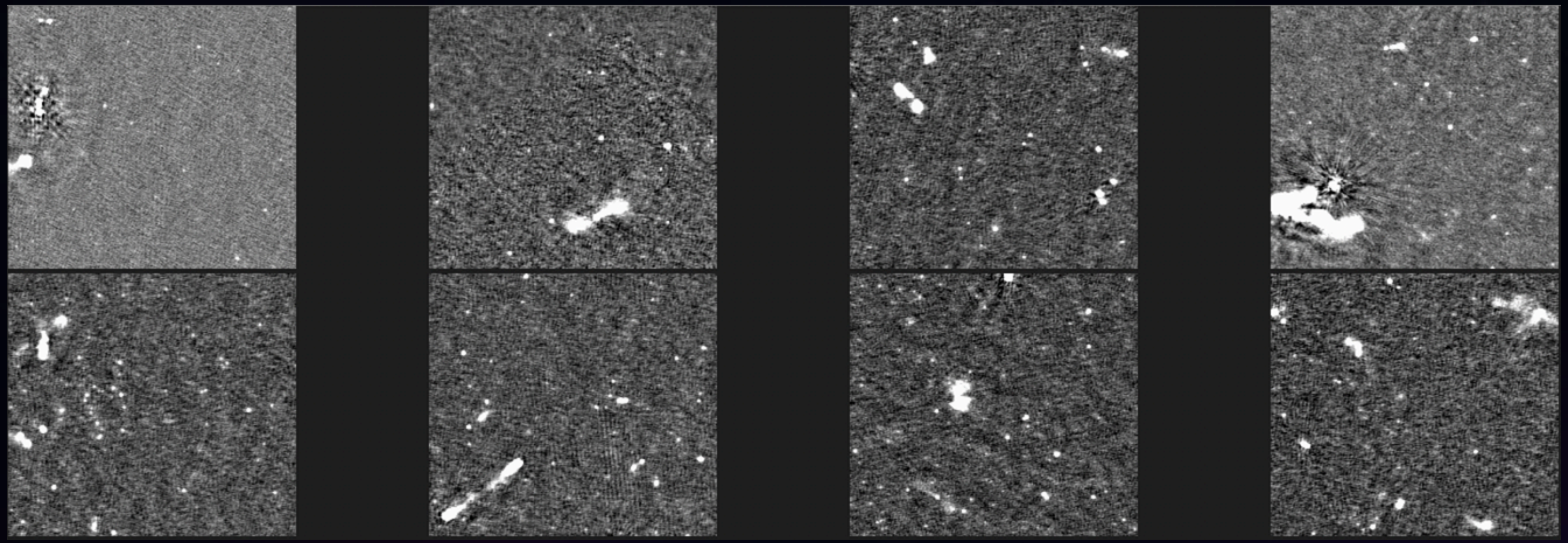
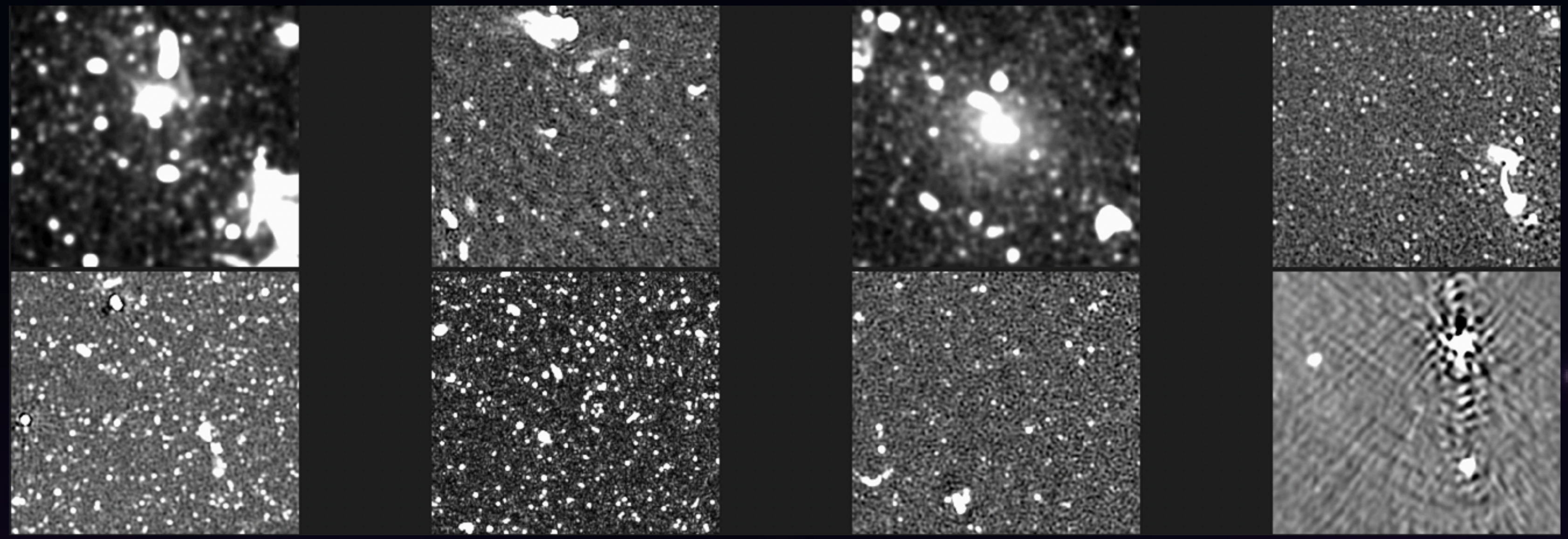




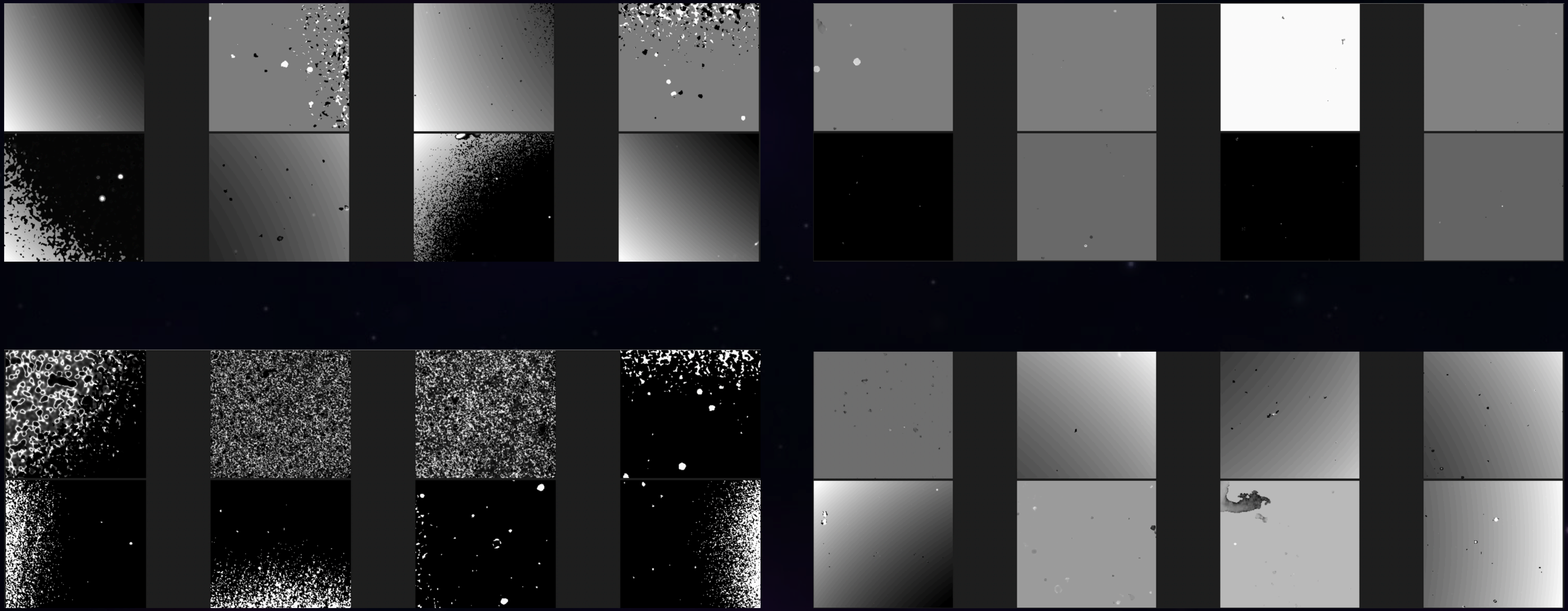


- Image representations projected via UMAP projection → unsupervised clustering using HDBSCAN
- Clusters containing low-SNR empty frames, boundary artifacts, and repetitive point sources are pruned
- Clusters containing extended morphologies and high-fidelity sources are retained

Clusters Kept in Dataset

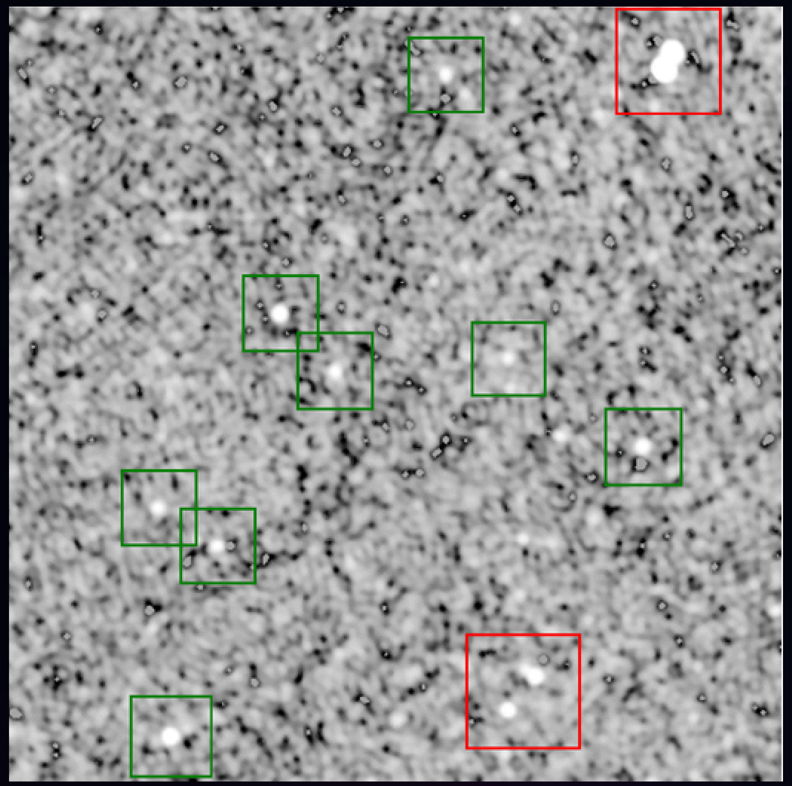





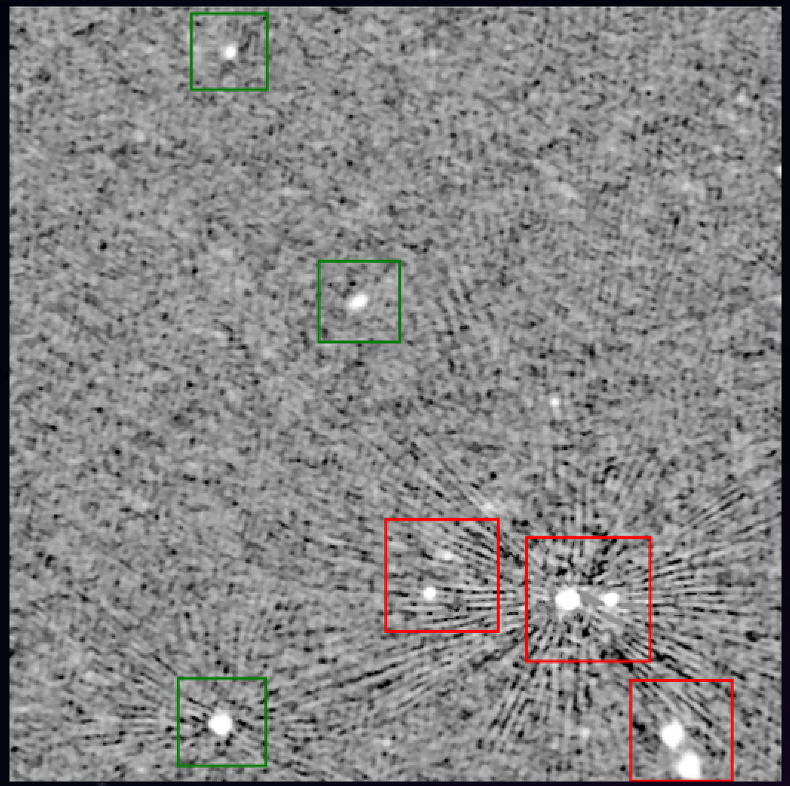
Clusters Removed from Dataset

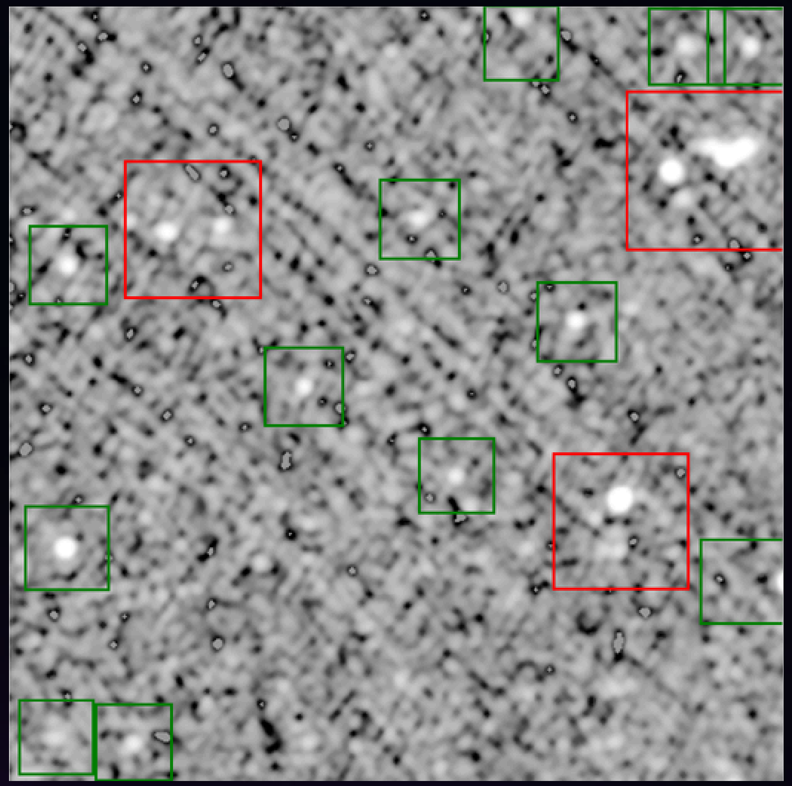


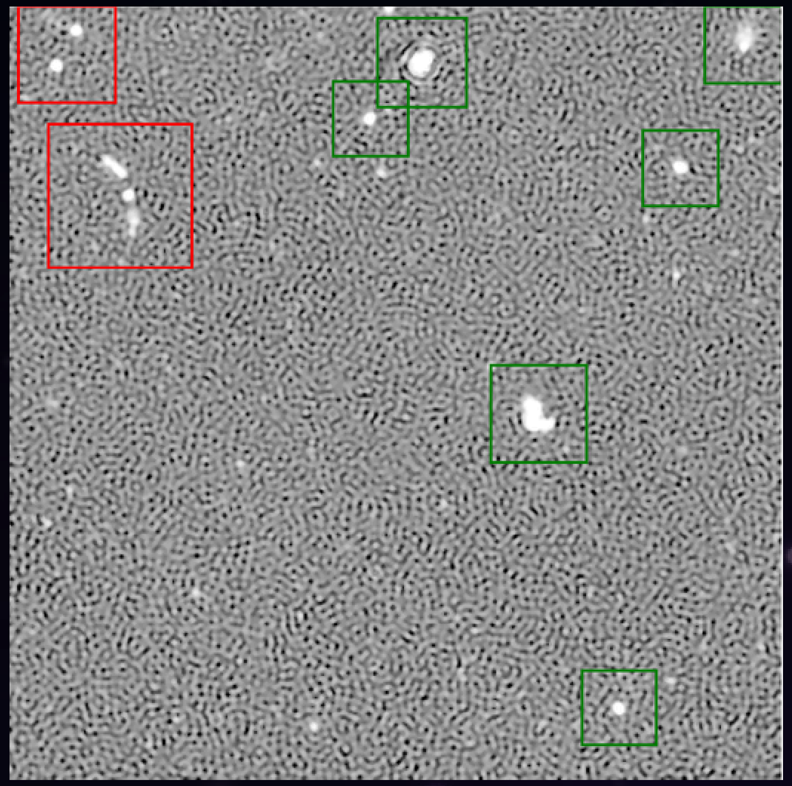
Object-Centred Crops

Full Survey Field Map (High Noise) → SNR Thresholding → Object Detection & Bounding Box Extraction → If Overlapping, combine Crops





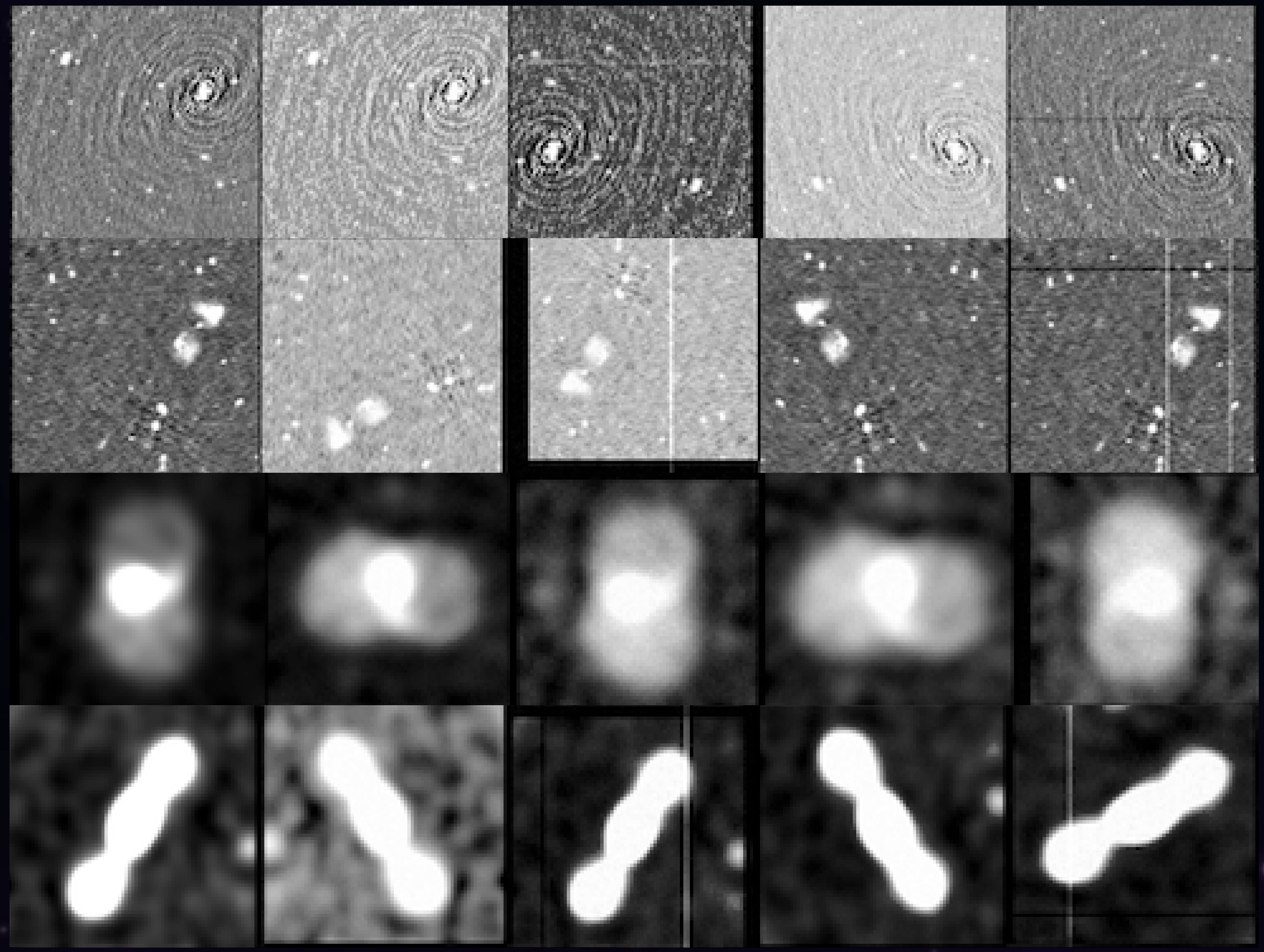




Augmentations

Morphology-preserving operations performed:

- Dihedral transforms
- Asinh stretch
- Affine jitter
- Gaussian noise
- Banding
- Flux/background perturbation



Preliminary Results

Table 1: Top five model configurations under linear probing, ranked by mean Macro-F1 across MiraBest, LoTSS DR2, and RGZ DR1. Best Macro-F1 among the displayed models for each validation dataset is highlighted in bold.

Rank	Model configuration	MiraBest		LoTSS DR2		RGZ DR1	
		Macro-F1	Weighted-F1	Macro-F1	Weighted-F1	Macro-F1	Weighted-F1
1	Phase 2 – Reduced (ROI, Aug.)	0.670 ± 0.013	0.671 ± 0.013	**0.583 ± 0.006**	0.753 ± 0.003	0.733 ± 0.004	0.896 ± 0.002
2	Phase 2 – Original (ROI, Aug.)	**0.679 ± 0.011**	0.680 ± 0.011	0.567 ± 0.005	0.738 ± 0.005	0.728 ± 0.007	0.894 ± 0.002
3	Phase 2 – Original (ROI, No Aug.)	0.661 ± 0.005	0.662 ± 0.006	0.556 ± 0.009	0.734 ± 0.006	0.725 ± 0.002	0.892 ± 0.001
4	Phase 2 – Reduced (ROI, No Aug.)	0.636 ± 0.016	0.635 ± 0.018	0.568 ± 0.001	0.746 ± 0.004	**0.734 ± 0.001**	0.896 ± 0.001
5	Phase 2 – Reduced (No ROI, No Aug.)	0.662 ± 0.026	0.662 ± 0.025	0.543 ± 0.012	0.726 ± 0.011	0.654 ± 0.006	0.864 ± 0.003

Reduced ROI models match or exceed raw dataset performance using significantly fewer training samples

Preliminary Results

Table 2: Top five model configurations under full fine-tuning, ranked by mean Macro-F1 across MiraBest, LoTSS DR2, and RGZ DR1. Best Macro-F1 among the displayed models for each validation dataset is highlighted in bold.

Rank	Model configuration	MiraBest		LoTSS DR2		RGZ DR1	
		Macro-F1	Weighted-F1	Macro-F1	Weighted-F1	Macro-F1	Weighted-F1
1	Phase 2 – Reduced (ROI, No Aug.)	0.719 ± 0.020	0.720 ± 0.020	0.669 ± 0.007	0.803 ± 0.005	0.819 ± 0.003	0.922 ± 0.001
2	Phase 2 – Reduced (ROI, Aug.)	0.718 ± 0.007	0.719 ± 0.006	0.663 ± 0.006	0.803 ± 0.001	**0.825 ± 0.001**	0.924 ± 0.001
3	Phase 2 – Original (ROI, Aug.)	**0.721 ± 0.008**	0.721 ± 0.009	0.663 ± 0.017	0.798 ± 0.010	0.821 ± 0.003	0.923 ± 0.000
4	Phase 2 – Original (ROI, No Aug.)	0.708 ± 0.004	0.709 ± 0.004	0.659 ± 0.014	0.798 ± 0.009	0.813 ± 0.002	0.921 ± 0.000
5	Phase 1 – Reduced (ROI, Aug.)	0.695 ± 0.009	0.696 ± 0.010	**0.670 ± 0.015**	0.805 ± 0.006	0.798 ± 0.005	0.914 ± 0.001

Fine-tuning on reduced curated sets achieves top Macro-F1 across LoTSS DR2 (0.669) and RGZ DR1 (0.819)

Morphological Class Rebalancing

Goal: Prevent self-attention collapse on simple, high-frequency core points

Work in Progress

1. Analysis to quantify how rebalancing changes sampling across telescopes and source types
2. Perform Class Rebalancing on competitive Datasets
3. Evaluate Results
4. Automate process of reducing dataset

Thank You for your Attention!

This work is being funded by the
MDIA Pathfinder Scholarship Fund