STRADAViT

Self-Supervised Transfer for Radio Astronomy Vision Transformers

Andrea DeMarco[1] Ian Fenech Conti[1] Hayley Camilleri[1]
Ardiana Bushi[3] Simone Riggi[2]

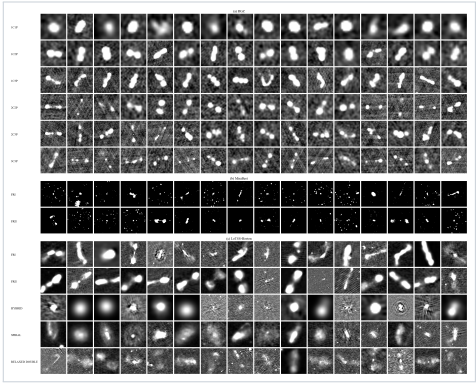
[1]Institute of Space Sciences and Astronomy, University of Malta
[2]INAF–Osservatorio Astrofisico di Catania
[3]School of Physics and Astronomy, University of Edinburgh

Machine Learning for Astrophysics 2026
Valletta, Malta

Radio morphology is a transfer problem

- Labels are scarce, expensive, and dataset-specific.
- Resolution, noise, calibration, and imaging artefacts change across surveys.
- A useful backbone must preserve morphology *before* task-specific fine-tuning.

Classification is task-specific.
The backbone question is: *what transfers?*



Shared input pipeline: MiraBest, LoTSS DR2, and RGZ DR1 after ZScale normalization.

Three questions structure the ViT-only study

Data and views

Can object-aware views prevent sparse backgrounds from dominating self-supervision across a mixed-survey dataset?

Training objective

How do masked reconstruction, contrastive learning, and their ordering affect transfer?

Transfer behaviour

Do gains survive both frozen-backbone probing and end-to-end fine-tuning across datasets?

Scope: ViT and ViT-derived backbones. This is not a universal state-of-the-art claim over task-specialized CNNs or equivariant models.

Mixed-survey pretraining, independent benchmarks

512 × 512 cutouts; per-image ZScale normalization

590,654 unlabeled cutouts

SKA SDC1 2%

Design rationale

The pretraining dataset exposes the encoder to heterogeneous beams, noise statistics, dynamic ranges, and imaging pipelines.

Independent transfer tasks

MiraBest: binary FR I/FR II
LoTSS DR2: five morphology labels
RGZ DR1: six component/peak classes

Continued pretraining adapts an existing ViT

Architecture: 224×224 input, 196 image-patch tokens, 768-dimensional pooled representation. **Registers** are learned auxiliary tokens that participate in attention but are excluded from patch pooling.

Controlled variables

Views, losses, stage ordering, and register tokens.

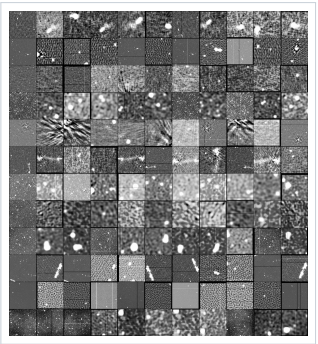
Held fixed

Backbone scale and downstream evaluation protocol.

Object-aware crops keep radio sources in view

Naive random crops can contain only background—or place different sources in a positive pair. ROI anchoring preserves informative, shared source structure.

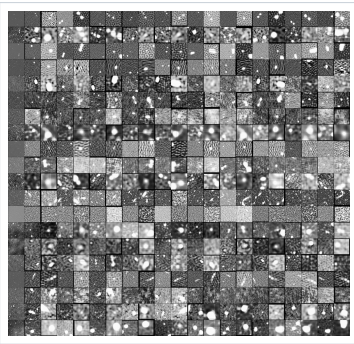
Reconstruction branch



Sampling

ROI-aligned: 20% wide-field samples

Contrastive branch

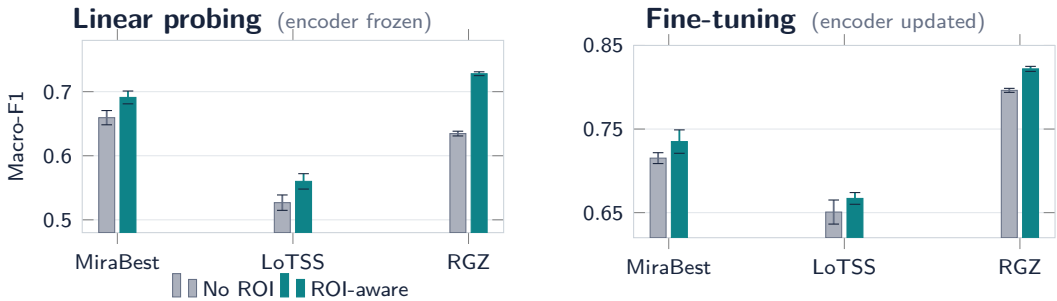


Pairing

Shared ROI: one wide + one overlapping local view

ROI-aware views improve every matched comparison

Matched contrastive-only Soft-HCL; same preprocessing and three stratified folds; class-balanced Macro-F1 mean ± standard deviation.



ROI-aware sampling is higher in all six comparisons; the largest difference is RGZ linear probing (+9.34 pp).

Reconstruction losses ask where fidelity matters

L2

Masked-patch error

Penalizes squared reconstruction error only at masked patch locations.

Baseline MAE objective; large errors receive stronger penalties.

L2 + L1

Adds global absolute error

Extends the signal across the reconstructed image rather than only masked patches.

Encourages overall pixel fidelity with less emphasis on extreme residuals.

L2 + BL1

Adds brightness weighting

Upweights absolute reconstruction errors in brighter image regions.

Tests whether source-dominated pixels deserve greater influence than background.

Observed outcome: reconstruction-loss choice had limited influence, and reconstruction-only training gave the weakest transfer results.

How contrastive losses weight negatives

Cosine similarity compares feature vectors. A different image close to anchor **A** in feature space is a **hard negative**. Every negative contributes; thicker line means greater influence on the loss.
HCL: Hard Contrastive Learning.

InfoNCE

Base similarity score

Every negative contributes. A negative closer to **A** receives a larger penalty.

HCL

Full similarity boost

Start from InfoNCE. Give a larger extra boost to negatives closer to **A**. The closest can dominate.

Soft-HCL

Average two scores

For each InfoNCE negative, average:
Same-for-all score: identical for every negative.
Resemblance score: larger when closer to **A**.

The average becomes the extra boost; resemblance controls only half.

Contrastive learning provides the largest step

Exploratory envelope: the highest mean is selected separately within each branch and dataset; these cells do not represent one checkpoint.

Linear probe: highest observed mean

	MiraBest	LoTSS	RGZ
Reconstruction	.612	.485	.565
Contrastive	.692	.560	.728
Two-stage	**.699**	**.575**	**.732**

Fine-tuning: highest observed mean

	MiraBest	LoTSS	RGZ
Reconstruction	.717	.664	.795
Contrastive	.735	.672	**.826**
Two-stage	**.736**	**.677**	.825

- Reconstruction-only training gives the weakest transfer results here.
- Contrastive refinement produces the main representation gain.
- Two-stage training is the most consistent branch, but its margin over contrastive-only is small and dataset-dependent.

The adaptation procedure also transfers to DINOv2

DINOv2-Base(R): before and after HCL adaptation

Protocol	Dataset	Off-the-shelf	Adapted	Δ pp
LP	MiraBest	.685	.686	**+0.1**
LP	LoTSS	.529	.561	**+3.2**
LP	RGZ	.629	.739	**+11.0**
FT	MiraBest	.704	.738	**+3.4**
FT	LoTSS	.708	.678	**−3.0**
FT	RGZ	.812	.828	**+1.6**

Same radio-domain HCL recipe and downstream fold assignments; Macro-F1 means shown.

Observed pattern
Base(R) improves in five of six comparisons; the largest shift is RGZ linear probing (**+11.0 pp**).

Register result: adapted Base(R) exceeds adapted Base in all six cells.

Interpretation: the same recipe changes both DINOv2 initializations, supporting procedural portability beyond ViT-MAE.

DINOv2 uses native 518 × 518 inputs and a smaller micro-batch; point estimates use the same downstream fold assignments.

A balanced release checkpoint

Fixed selection rule: choose the two-stage checkpoint whose weakest dataset-level average across linear probing and fine-tuning is highest.

Δ class-balanced Macro-F1 versus ViT-MAE (pp)

Dataset	Linear probe	Fine-tuning
MiraBest	+3.0	+0.1
LoTSS DR2	+7.5	−1.9
RGZ DR1	+14.6	+2.0

Released configuration

ViT-MAE Base
four register tokens
L2+L1 reconstruction
Soft-HCL refinement

196

patch tokens per view
versus **1369** for native DINOv2-Base

Read the deltas carefully: absolute fine-tuning scores are generally higher than linear-probe scores; the smaller fine-tuning values here are gains relative to the initialization baseline.

A stronger radio-domain ViT starting point

STRADAViT[1] improves frozen radio-morphology representations while retaining competitive fine-tuning and substantially lower token-processing requirements than the DINOv2-based alternative.

What the experiments support

- ROI-aware views improve all six matched linear-probe and fine-tuning comparisons.
- Contrastive refinement drives the main gain; the released checkpoint improves frozen transfer on all three datasets.

What remains unresolved

- Fine-tuning gains are modest; LoTSS remains below the off-the-shelf references.
- ROI-centred crops and fixed tiling can underrepresent extended emission.

Companion talk: The released checkpoint recipe is held fixed while alternative data-curation regimes are evaluated.

Questions and discussion

[1] Hugging Face: `ISSA-ML/stradavit-base`

References

Vision Transformers and self-supervision

Dosovitskiy, A. et al. (2021). *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. ICLR.

He, K. et al. (2022). *Masked Autoencoders Are Scalable Vision Learners*. CVPR.

Chen, T. et al. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. ICML.

Robinson, J. D. et al. (2021). *Contrastive Learning with Hard Negative Samples*. ICLR.

Oquab, M. et al. (2023). *DINOv2: Learning Robust Visual Features without Supervision*. arXiv:2304.07193.

Radio astronomy datasets and SSL

Porter, F. A. M. and Scaife, A. M. M. (2023). *MiraBest: a data set of morphologically classified radio galaxies for machine learning*. RASTI.

Horton, M. A. et al. (2025). *Complex morphology and precession indicators of active galactic nuclei jets in LoTSS DR2*. A&A.

Wong, O. I. et al. (2024). *Radio Galaxy Zoo data release 1: 100185 radio source classifications from the FIRST and ATLAS surveys*. MNRAS.

Slijepcevic, I. V. et al. (2023). *Radio Galaxy Zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning*. RASTI.

Backup: training and compute

	Reconstruction	Contrastive
Epochs	35	35
Views per cutout	1	2
Per-device batch	256	64
Effective image batch	2048 via gradient accumulation	
Optimizer	AdamW, cosine schedule, base LR 5×10^{-4}	
Masking	75% patches	unmasked encoder
Readout	mean pooling over patch tokens; $D = 768$	
Hardware	4× NVIDIA RTX 6000 Ada (48 GB)	

Per forward pass	Each GPU processes 64 images, with two views per image. Across four GPUs, the contrastive loss therefore uses 512 view embeddings.
Per optimizer update	Gradients from eight forward passes are accumulated before updating the model: $64 \times 4 \times 8 = 2048$ images.
Negative pool	Each loss uses only the 512 embeddings from its current forward pass. Earlier passes contribute gradients to the update, but their embeddings are not reused as additional negatives; each query therefore has 510 negatives.

Backup: checkpoint versus references

Protocol	Model/comparison	MiraBest	LoTSS	RGZ
LP	Release checkpoint	.674	.564	.732
	Δ vs ViT-MAE (pp)	$+3.0$	$+7.5$	$+14.6$
	Δ vs stable DINOv2 (pp)	-4.3	-0.5	$+7.1$
FT	Release checkpoint	.725	.677	.825
	Δ vs ViT-MAE (pp)	$+0.1$	-1.9	$+2.0$
	Δ vs stable DINOv2 (pp)	$+1.8$	-3.1	$+1.3$

Point estimates from the common three-fold protocol; fine-grained rankings are interpreted descriptively.