Learning the Faint Radio Sky:
A Deep Ensemble Framework for Diffuse Surface-Brightness Inference

Nicoletta Sanvitale, INAF - Institute for Radio astronomy

Claudio Gheller - INAF - Institute for Radio astronomy
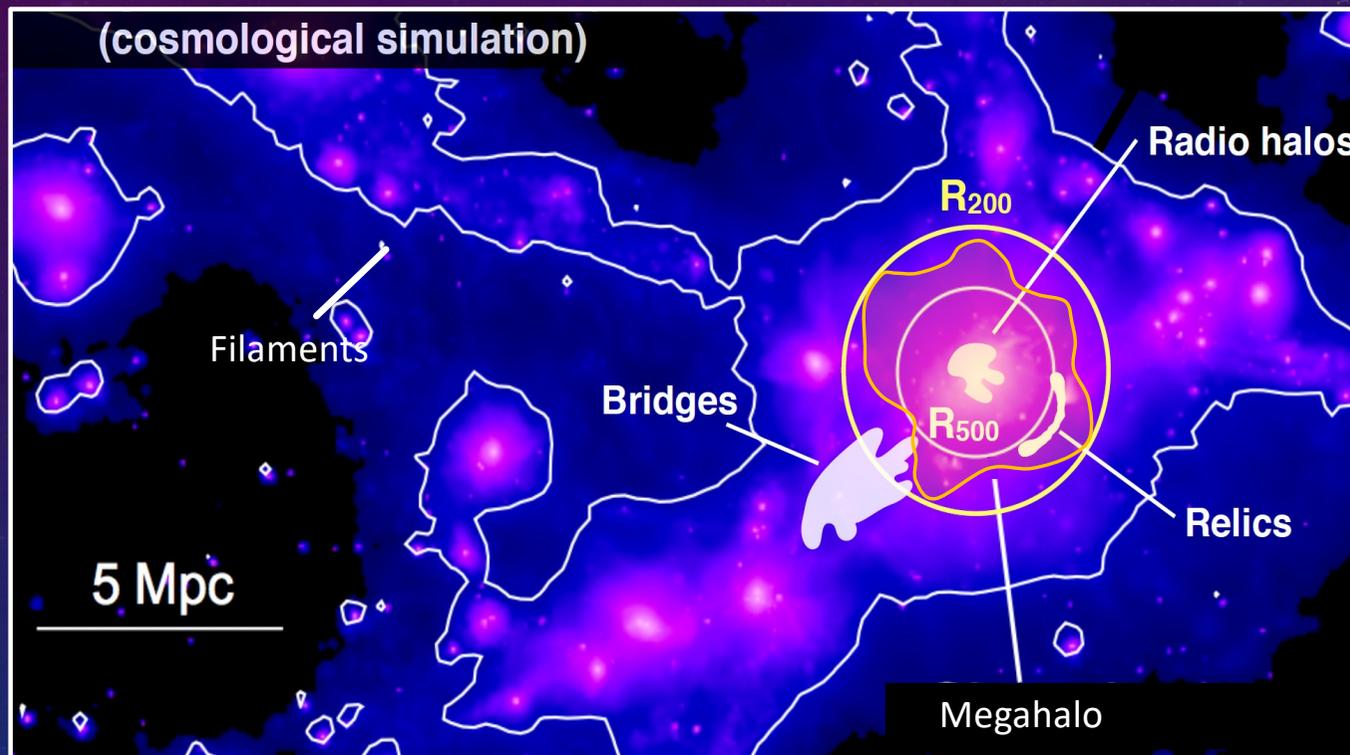
Franco Vazza- University of Bologna
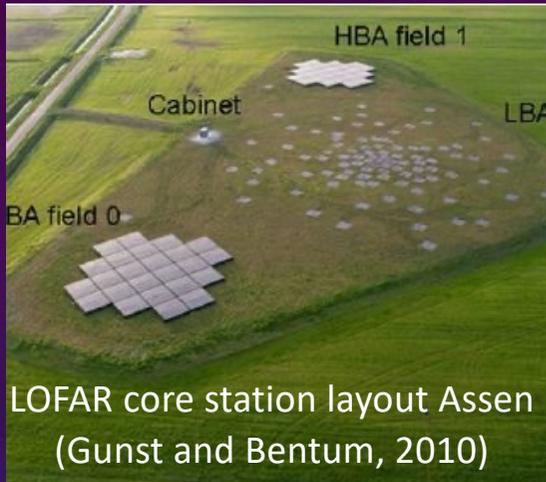
GOAL

Detecting and characterizing (flux estimate) low surface brightness diffuse radio sources through machine learning based approaches

credits Cuciti (2020)

LOFAR core station layout Assen
(Gunst and Bentum, 2010)

A composite of the SKA telescopes (SKAO)

WHY IS IT A PROBLEM?

- Low surface brightness, low S/N

- Telescope sensitivity

- Imaging complications for interferometry observations

- Compact source contamination

- Big data issue for current and upcoming surveys:

  - ❑ Computational cost

  - ❑ human supervision
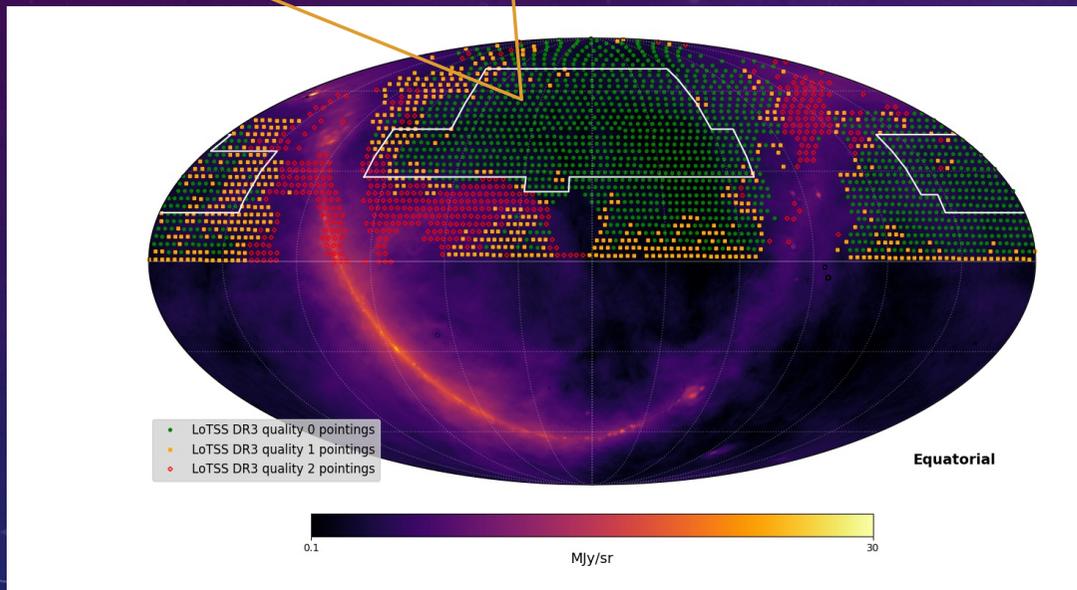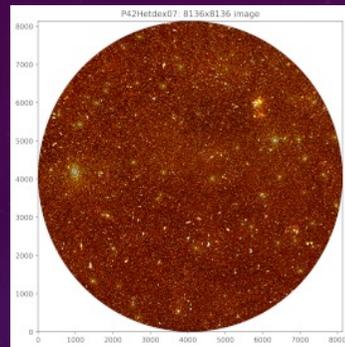
RADIO OBSERVATIONS OF THE SKY

LOFAR Data Releases (DR):

DR2 (2022) https://lofar-surveys.org/dr2_release.html
- 120-168 MHz, 6'' resolution
- 841 pointings
- ~4º x 4º each (moon diameter: 0.5º)
- ~8000^2 64 bits pixels
- ~ 500 MB/img
- TOT ~ 420 GB for high res image

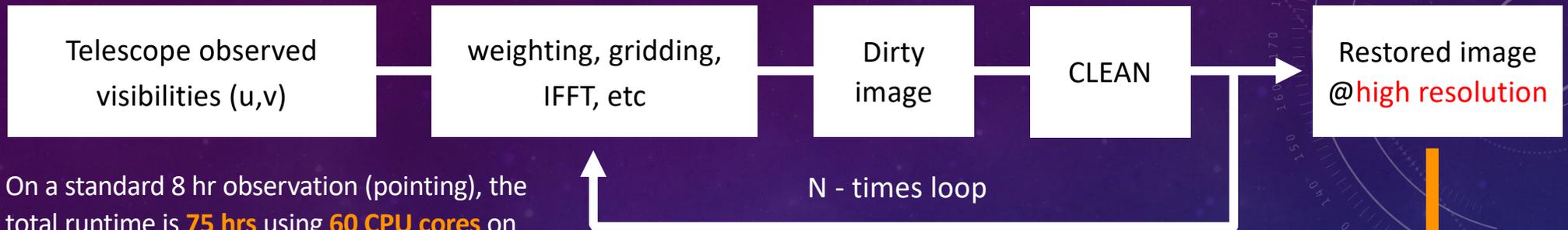**DR3** (February 2026):
- 2552 pointings
- TOT ~ 1.3 TB for high res image
- 18600 TB of data (visibilities)

WHY IS IT A PROBLEM? MANY POINTINGS, MANY SOURCES

RADIO IMAGING PIPELINE FOR HIGH RESOLUTION

Telescope observed visibilities (u,v) → weighting, gridding, IFFT, etc → Dirty image → CLEAN → Restored image @high resolution

N - times loop

On a standard 8 hr observation (pointing), the total runtime is **75 hrs** using **60 CPU cores** on a node with AMD EPYC Bergamo processors
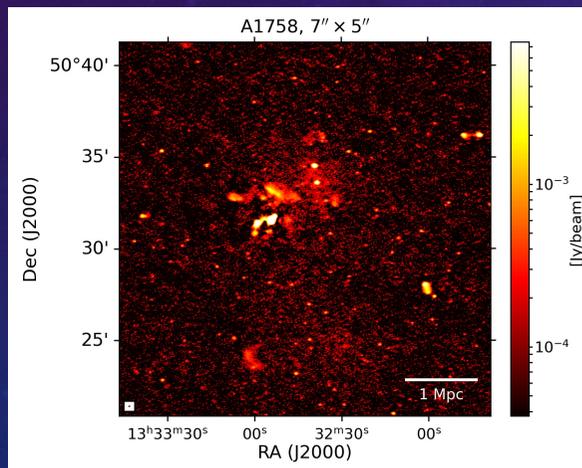
LOFAR DR2
P42Hetdex07 pointing 6''
(RA 204.4° DEC 49.7°)

A1758, 7" × 5"

A1758 LOFAR image at 144 MHz @ 7 arcsec x 5 arcsec

4

RADIO RE-IMAGING PIPELINE FOR LOW RESOLUTION

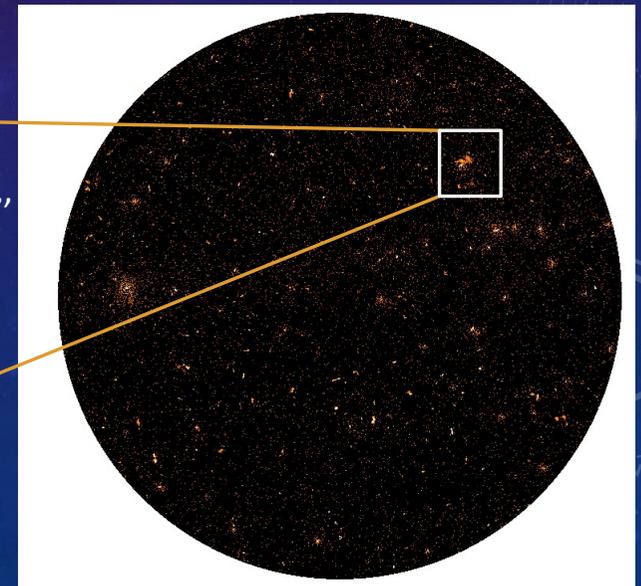| Telescope observed visibilities (u,v) | → | weighting, tapering, gridding, FFT, etc | → | Dirty image | → | CLEAN | → | Restored image @low resolution |

N - times loop

On a standard 8 hr observation (pointing), the total runtime is **75 hrs** using **60 CPU cores** on a node with AMD EPYC Bergamo processors

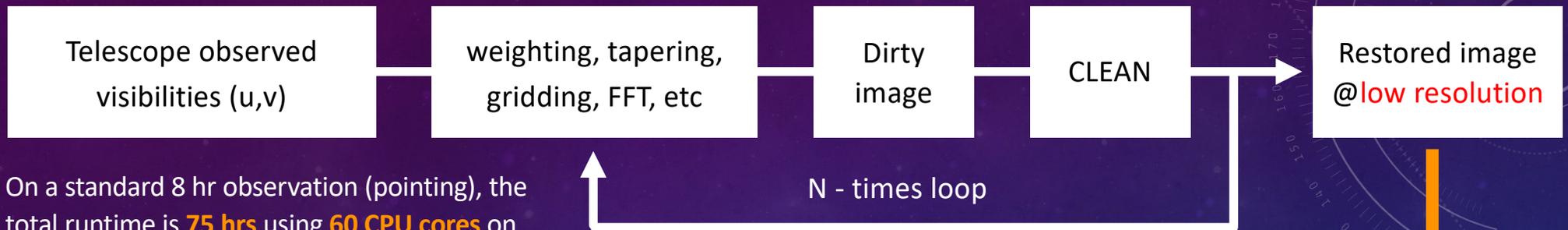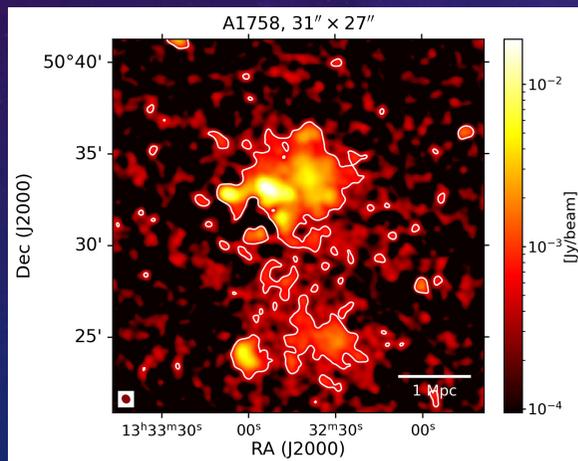**LOFAR DR2**
P42Hetdex07 pointing 20''
(RA 204.4° DEC 49.7°)

A1758, 31″ × 27″

A1758 LOFAR image at 144 MHz @ 7 arcsec x 5 arcsec
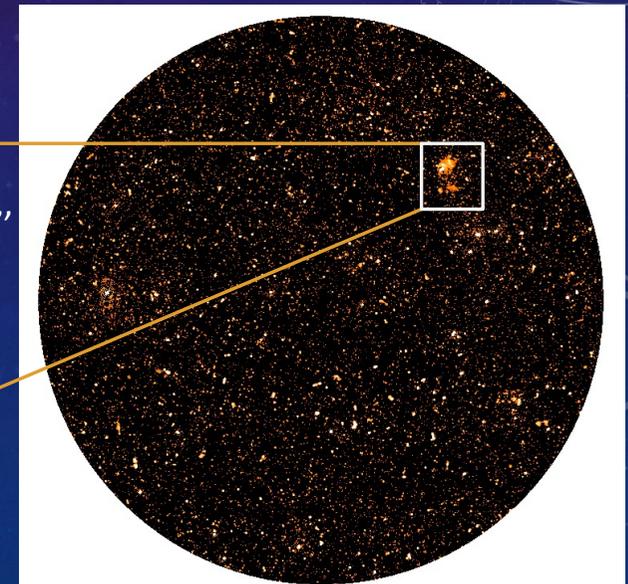
INTRODUCING TUNA (TransUNet for Astrophysics)

TransUNet is a hybrid network combining:

- A Vision Transformer (ViT) with Multi-head Self-Attention (MSA) mechanism encoder for capturing global context

- A U-Net decoder for precise, pixel-level prediction extracting local features

$$Loss = \lambda_1 \cdot L_{Cross\ entropy} + \lambda_2 \cdot L_{Dice}$$

$$\lambda_1 = \lambda_2 = {}^1\!/_2$$

Conv 3x3, ReLu

Upsample

Feature concatenation

Segmentation head (Softmax + argmax)
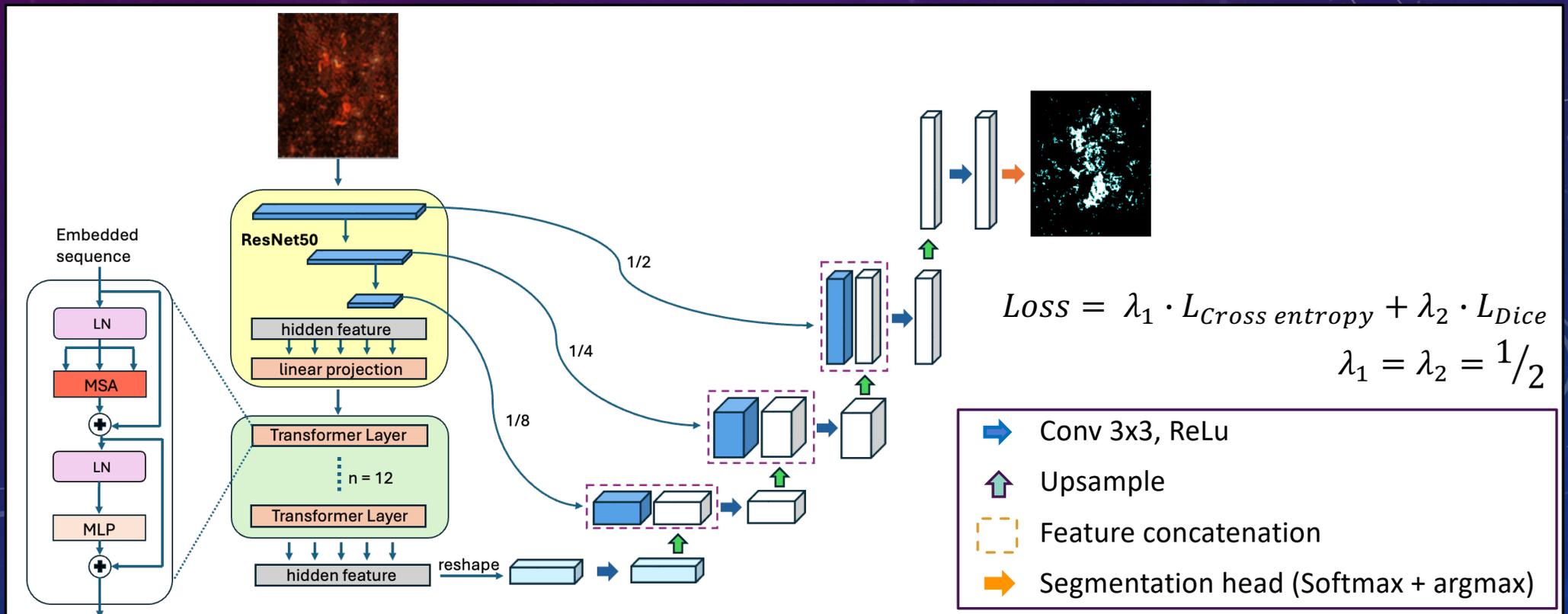
RESULTS OF SEGMENTATION ON REAL OBSERVATIONS

(Sanvitale et al., 2025)

- Evaluation data: LoTSS-DR2 observational data at 120-168MHz (nominal central frequency 144 MHz)

- Benchmark sample: 309 galaxy clusters from the PSZ2 catalog within LoTSS-DR2.

- Results: detects and reconstructs sources requiring 4–6× lower resolution, preserving morphology and completeness. (Precision 0.61± 0.02 ; Recall 0.50 ± 0.01)
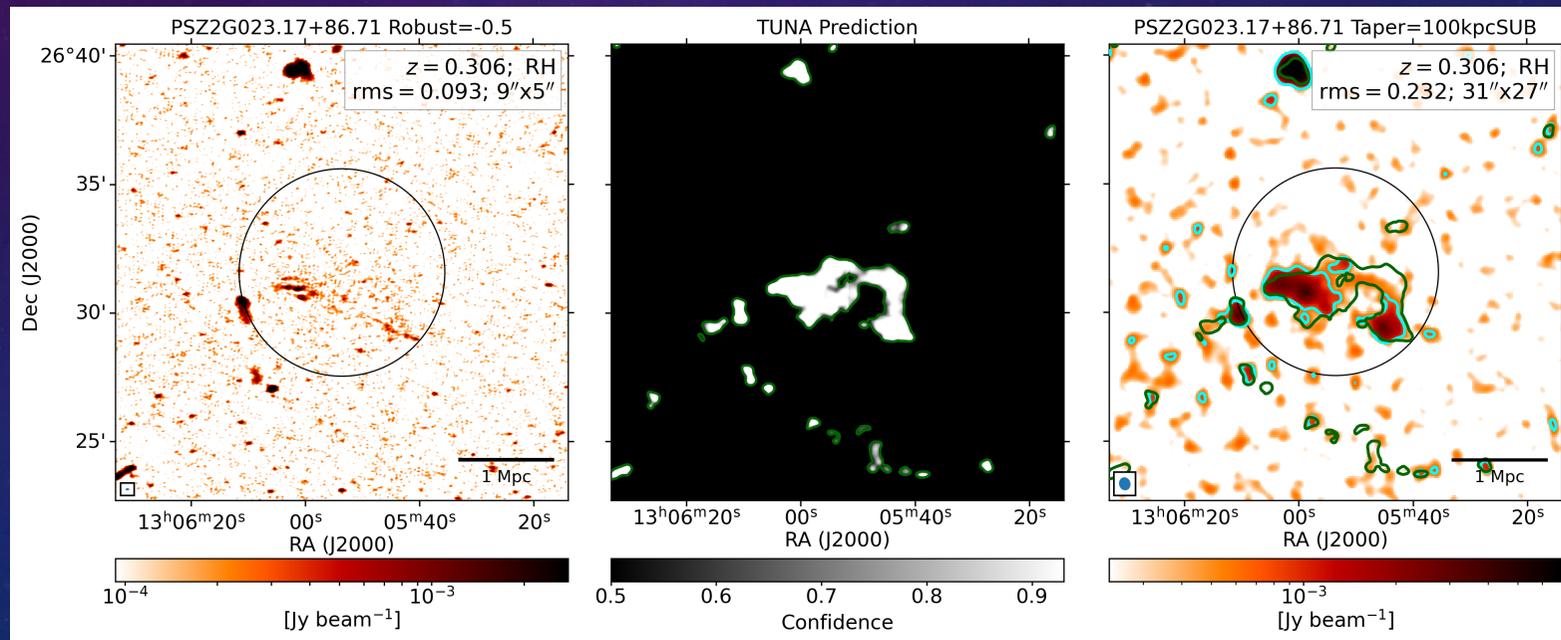
TUNA FOR SURFACE BRIGHTNESS REGRESSION

From segmentation to regression

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(pred_i - GT_i)^2 + \epsilon^2}$$
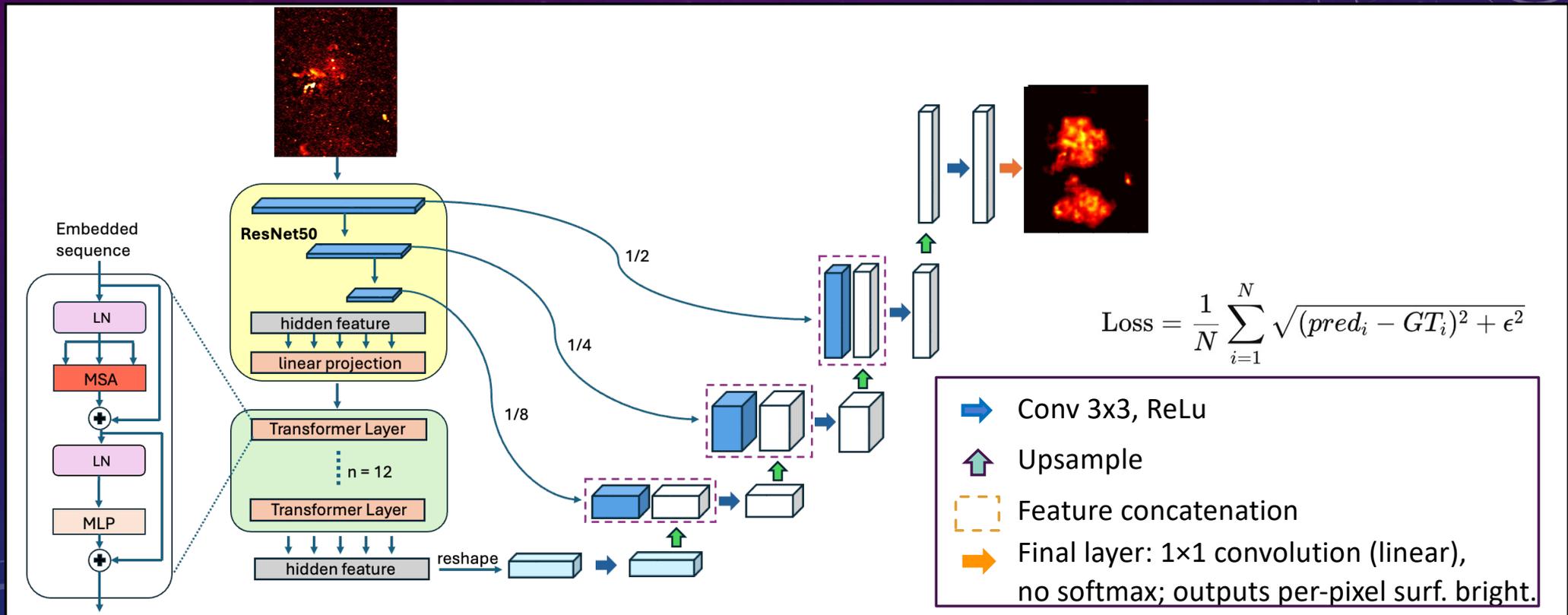
MOCK OBSERVATION GENERATION FOR TRAINING DATASET

Sky Model:

- Light cones from cosmological simulations using ENZO code

- Synchrotron emission calculated on single frequency @144MHz (LOFAR HBA central frequency)

- Real radio galaxies (extracted from LoTTs Dr2 pointings)

Mock Observation mimicking LOFAR HBA – 8 hrs observation:
Obtained using WSClean in combination with custom tools to mimic realistic radio interferometric data, incorporating:

- Random instrumental noise

- Dirty beam artifacts

- average noise $\simeq 10^{-6}$ Jy/arcsec^2

Dataset:

- 300 mock images
  (2000 × 2000 px, i.e. 1.1° × 1.1°)

- training 250 images, testing 50 images

- nominal spatial resolution 2 arcsec/px

Mock image — Ground Truth

9
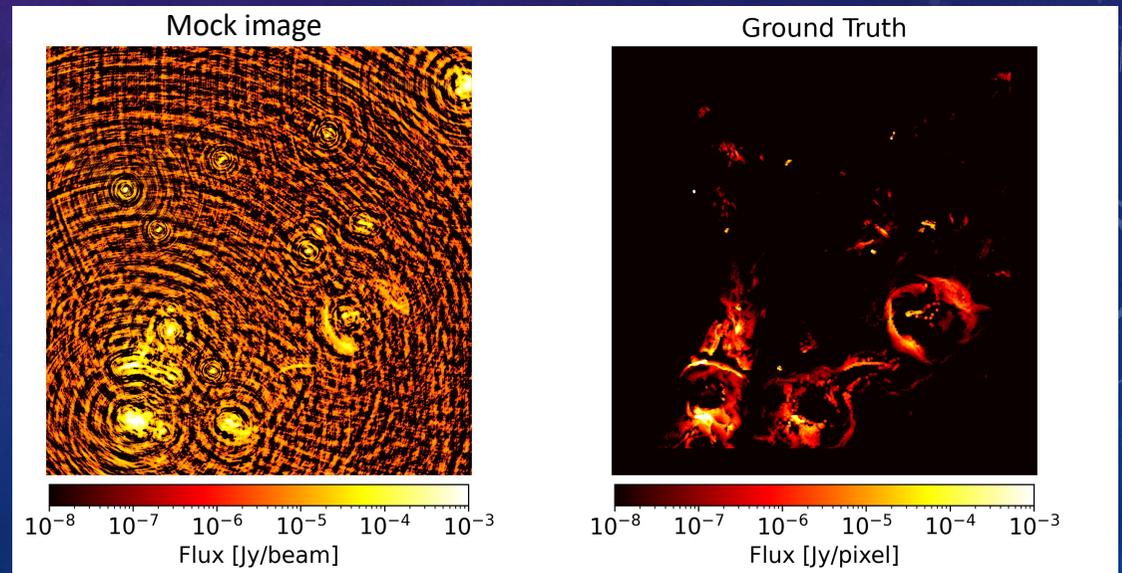
DEEP ENSEMBLE APPROACH

Deep ensembles are a widely used state-of-the-art approach for improving prediction robustness and estimating uncertainty in deep learning models.

Non-convex neural network loss

- Training can converge to **different local minima**
- Different runs produce **different predictors**

Deep ensemble

- Train multiple models with different initializations
- Average predictions → reduced variance and more robust regression

Each network produces a prediction: $f_1(x), \ f_2(x), \ ..., \ f_N(x)$

Final prediction is the average: $\hat{y} = \dfrac{1}{N} \sum_{i=1}^{N} f_i(x)$

The spread of predictions reflects model uncertainty: $\sigma^2(x) = \text{Var}(f_1(x), ..., f_N(x))$

DEEP ENSEMBLE PERFORMANCE

- 9 Models with different seedings

- 350 epochs, lr 0.005, batch size 24, tile size 512 px

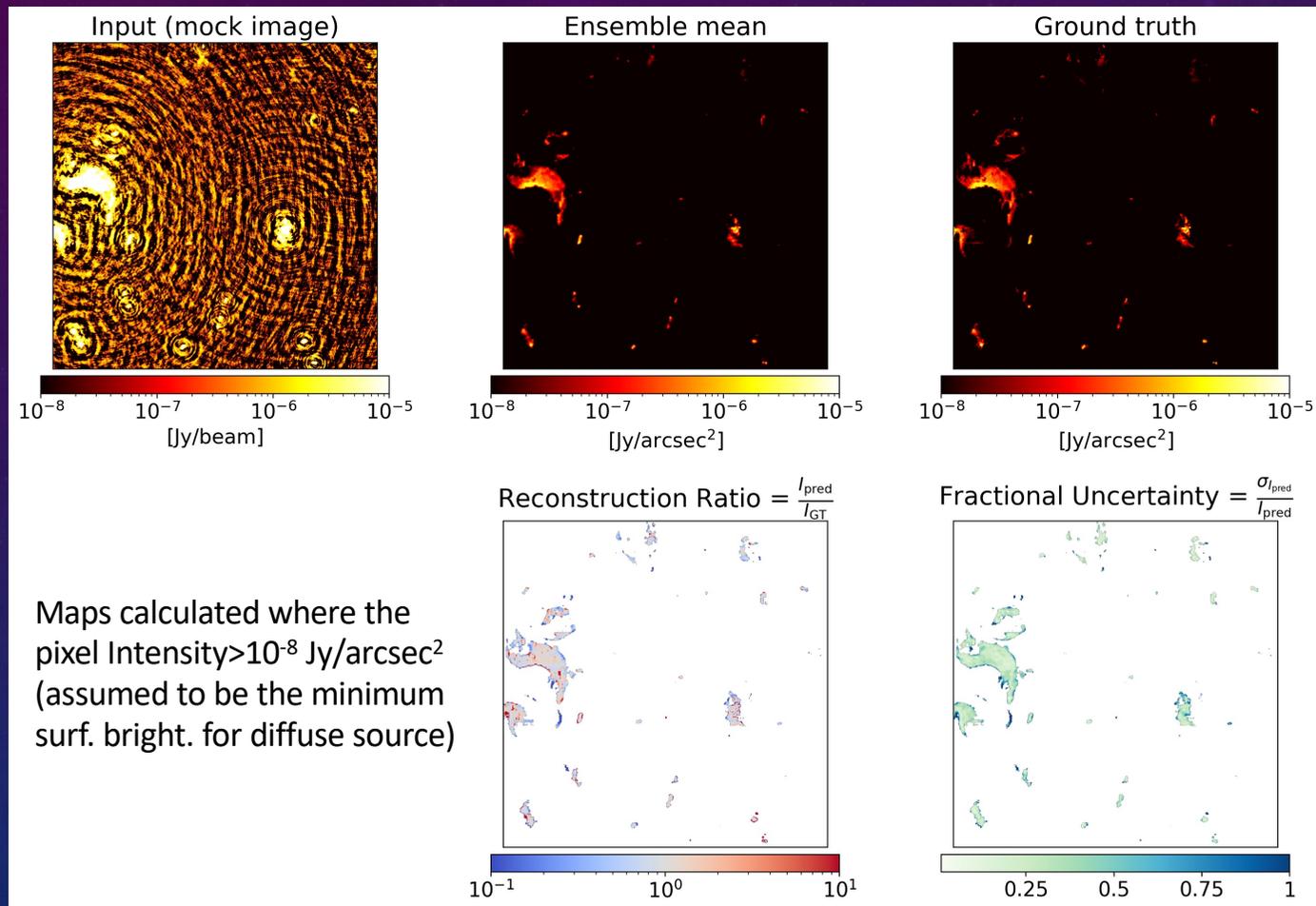- inference time 18 sec/image on NVIDIA Ampere 100 GPU of Leonardo @ CINECA

Pixel Metrics			
Metric		Mean	Std
MAE (Jy/arcsec^2)	Mean absolute pixel error	2.02E-08	7.98E-08
RMSE (Jy/arcsec^2)	Square root of MSE	3.54E-07	1.21E-06
R^2	Coefficient of Determination	7.92E-01	1.99E-01
SSIM	Structural similarity index	9.99E-01	2.88E-03

- Near-perfect structural reconstruction (SSIM ≈ 0.999)

- small pixel-level errors

- The scatter in R^2, MAE, and RMSE show that a small number of test images are much harder than the rest.

Flux-based metrics			
Metric		Value	Std
Signed flux error mean (%)	Average photometric bias	-13.99	25.11
Absolute flux error mean (%)	Typical photometric error	22.48	17.91
P25 absolute flux error (%)	25th percentile of absolute flux errors	11.40	
P75 absolute flux error (%)	75th percentile of absolute flux errors	25.21	

- Typical integrated-flux error of about 20%

- The ensemble shows a moderate underestimation of total flux

11

Input (mock image)

Ensemble mean

Ground truth

Reconstruction Ratio $= \frac{I_{\text{pred}}}{I_{\text{GT}}}$

Fractional Uncertainty $= \frac{\sigma_{I_{\text{pred}}}}{I_{\text{pred}}}$

Maps calculated where the pixel Intensity$>10^{-8}$ Jy/arcsec^2 (assumed to be the minimum surf. bright. for diffuse source)

12

Input (mock image) Ensemble mean Ground truth

Maps calculated where the pixel Intensity>10^{-8} Jy/arcsec^2 (assumed to be the minimum surf. bright. for diffuse source)

Reconstruction Ratio $= \frac{I_{\text{pred}}}{I_{\text{GT}}}$

Fractional Uncertainty $= \frac{\sigma_{I_{\text{pred}}}}{I_{\text{pred}}}$

13

Input (mock image)

Ensemble mean

Ground truth

10^{-8} 10^{-7} 10^{-6} 10^{-5}
[Jy/beam]

10^{-8} 10^{-7} 10^{-6} 10^{-5}
[Jy/arcsec^2]

10^{-8} 10^{-7} 10^{-6} 10^{-5}
[Jy/arcsec^2]

Reconstruction Ratio $= \frac{I_{\text{pred}}}{I_{\text{GT}}}$

Fractional Uncertainty $= \frac{\sigma_{I_{\text{pred}}}}{I_{\text{pred}}}$

Maps calculated where the pixel Intensity>10^{-8} Jy/arcsec^2 (assumed to be the minimum surf. bright. for diffuse source)

10^{-1} 10^{0} 10^{1}
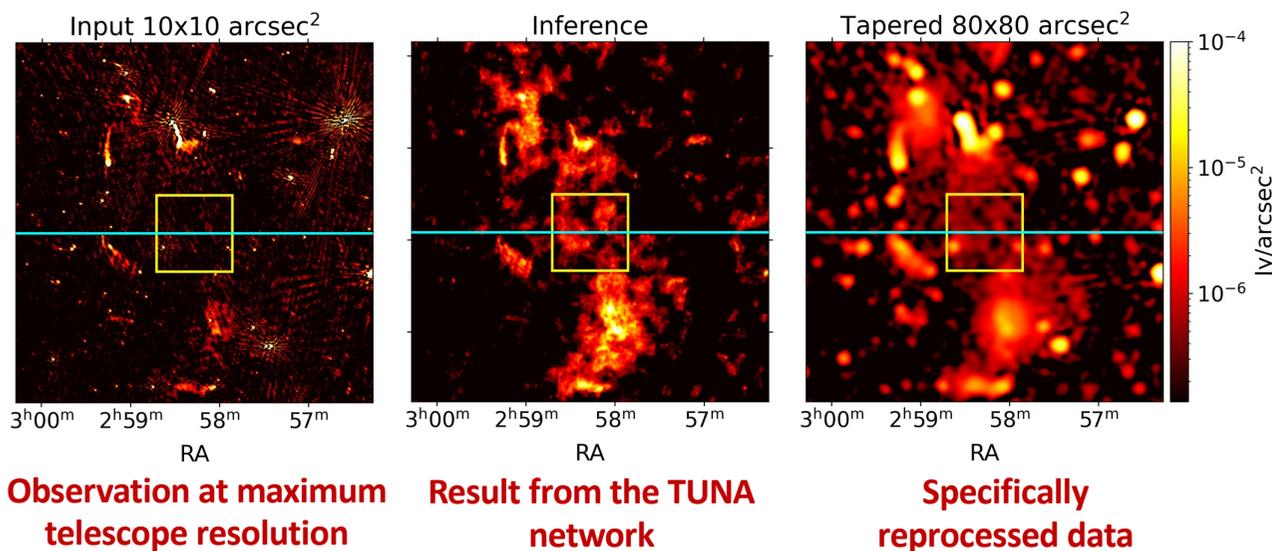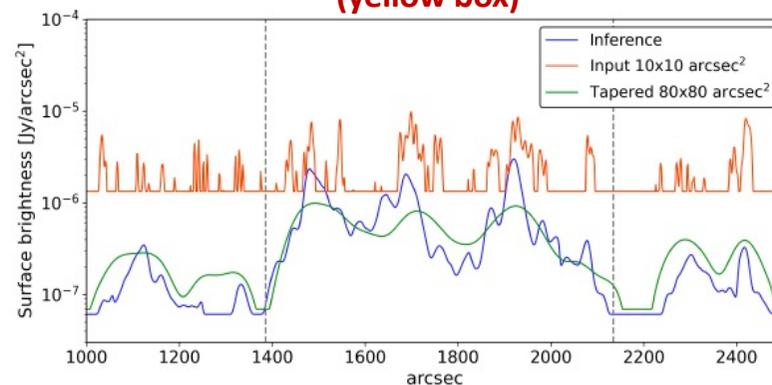
0.25 0.5 0.75 1

PRELIMINARY RESULTS ON REAL OBSERVATIONS

(Sanvitale et al., 2025)

- LOFAR HBA observations of the radio ridge connecting Abell 399 and Abell 401 (Govoni et al. 2019)

- Input image: 140 MHz, 10" resolution, $\sigma = 0.3$ mJy/beam

- TUNA prediction compared with the original image and an 80" tapered map ($\sigma = 1$ mJy/beam)

Zoom-in on the 1D surface brightness profile across the ridge region (yellow box)

Observation at maximum telescope resolution

Result from the TUNA network

Specifically reprocessed data

Thank you for the attention!