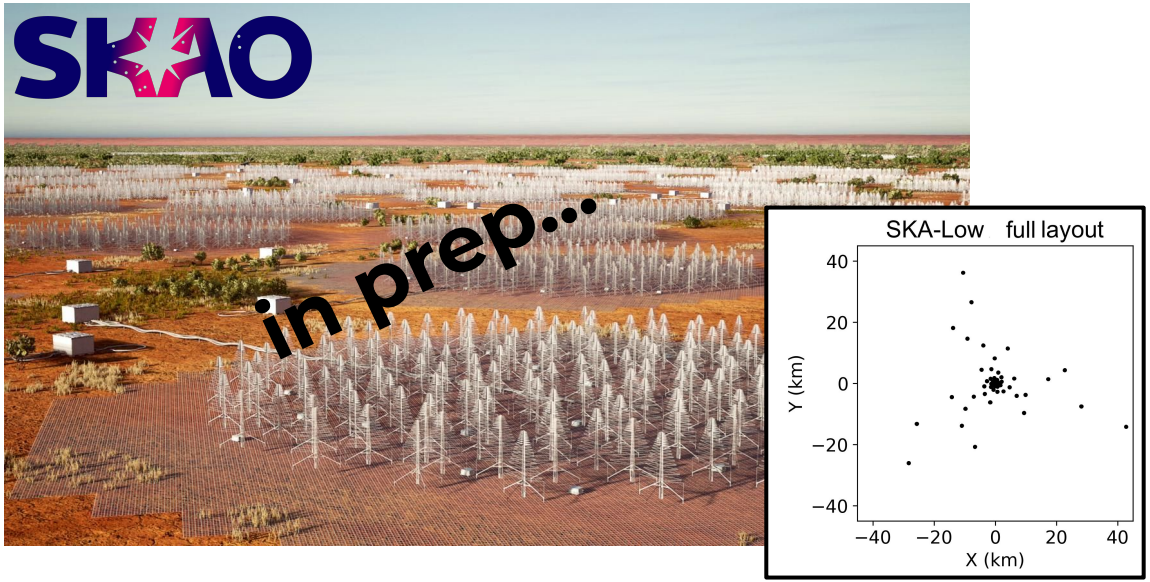
Detection and characterization of galaxies with deep-learning in radio continuum surveys, preparation to SKAO

Adam Zarka - D. Cornu, B. Semelin, G. Sainton
Observatoire de Paris, LUX
SF2A - 22/06/2026

Context - *SKA Observatory*

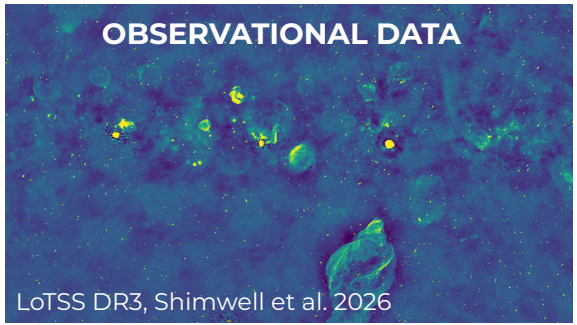
SKA-low, mapping the structure of the early Universe

- 512 stations across Western Australia (419,000 m2)
- Frequency range of 50 MHz - 350 MHz
- Science ready data flow of 700 PB/year
- First science ready data ~ 2029 (2027 for science verification)



Precursor instruments

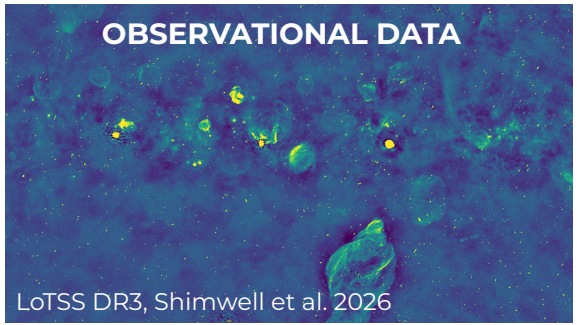
Aim - *Detection pipeline for radio continuum, a computer vision challenge*

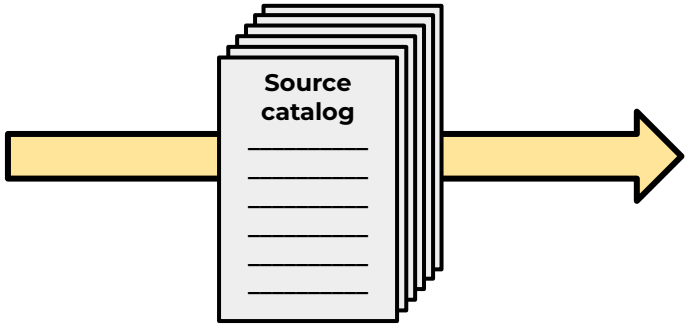


SCIENCE !

- AGN physics
- Star Formation
- Cosmology
- Large-scale structures
- ...

Aim - *Detection pipeline for radio continuum, a computer vision challenge*

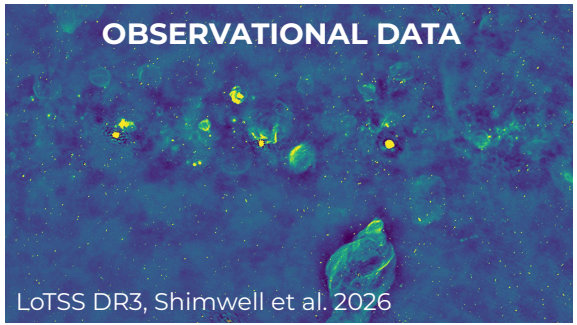


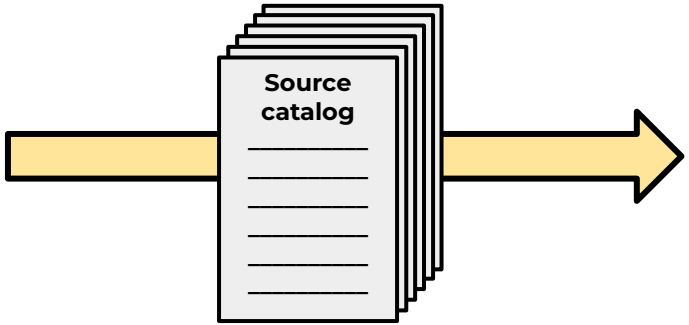


SCIENCE !

- AGN physics
- Star Formation
- Cosmology
- Large-scale structures
- ...

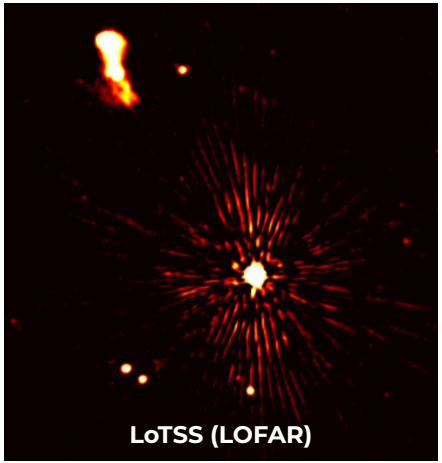
Aim - *Detection pipeline for radio continuum, a computer vision challenge*



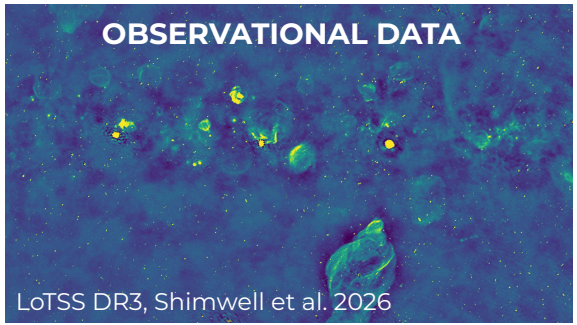


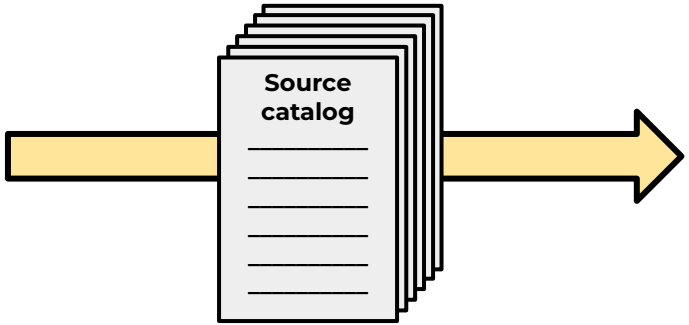
SCIENCE !

- AGN physics
- Star Formation
- Cosmology
- Large-scale structures
- ...



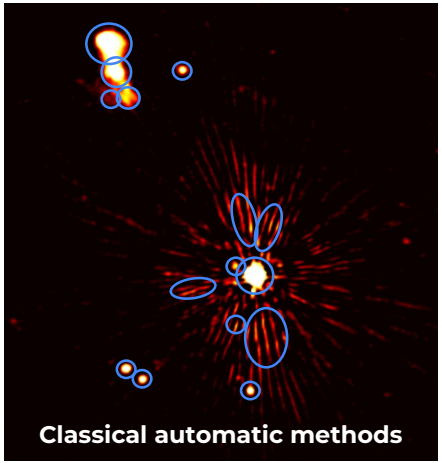
Aim - *Detection pipeline for radio continuum, a computer vision challenge*



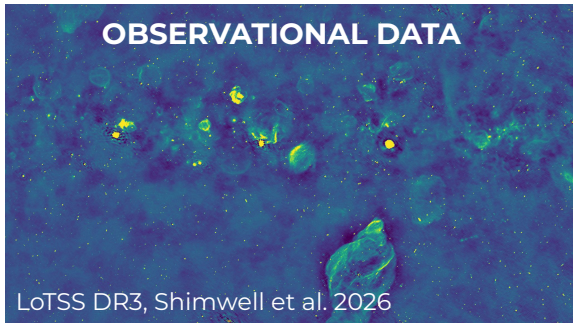


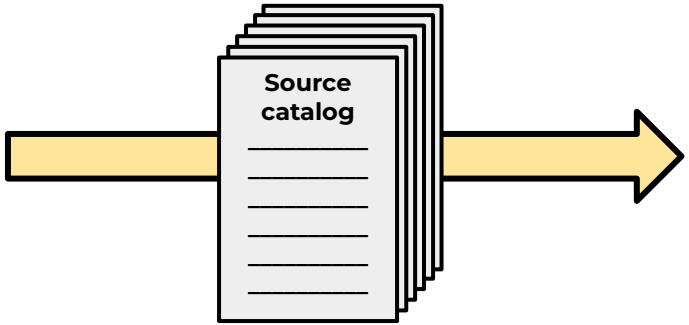
SCIENCE !

- AGN physics
- Star Formation
- Cosmology
- Large-scale structures
- ...



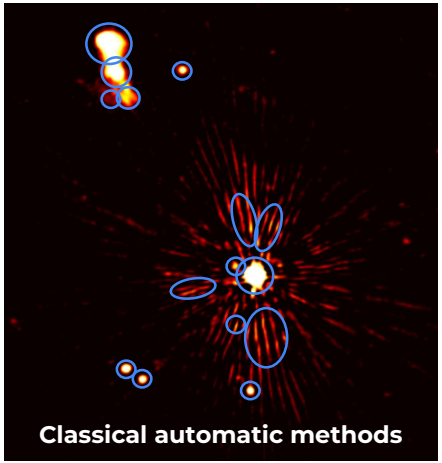
Aim - *Detection pipeline for radio continuum, a computer vision challenge*

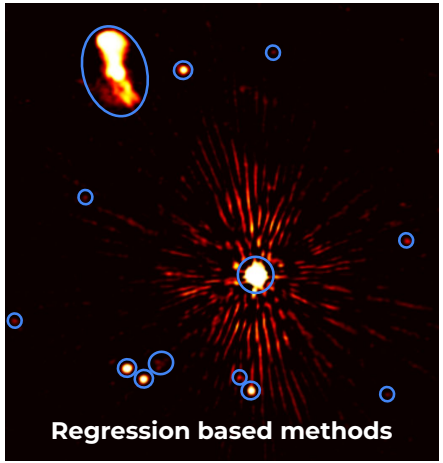




SCIENCE !

- AGN physics
- Star Formation
- Cosmology
- Large-scale structures
- ...



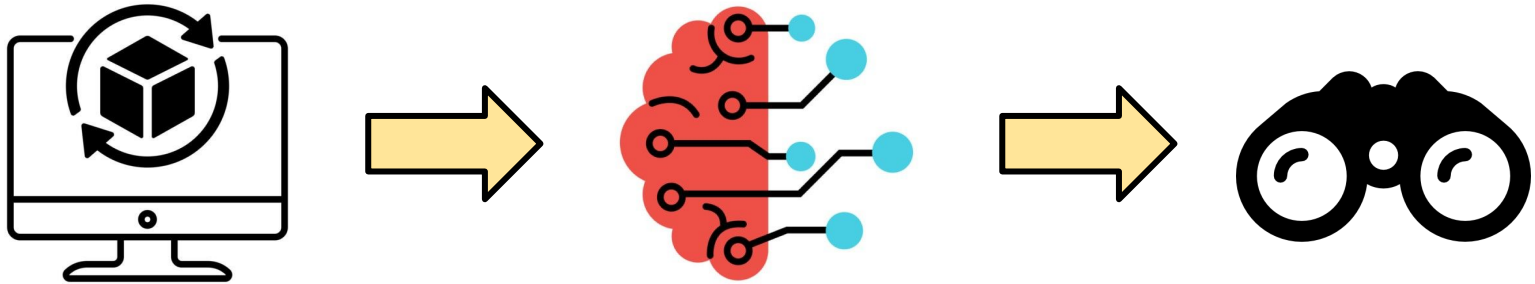


Challenges :

- Complex morphologies
- Artifacts
- Low SNR sources
- Expensive computing

➔ Developing a **Machine Learning-based fast-processing detection pipeline** for 2D **radio continuum** data on precursor instruments to SKA (here LOFAR) in order to build robust catalogs for non-biased science.

Approach - *Convolutional Neural Networks (CNN) trained on simulations*



Data **simulation**
products : mock data + targets

YOLO-CIANNA
Regression based detector
supervised **training**...

Inference on real data
(direct/indirect)

Mock dataset must :

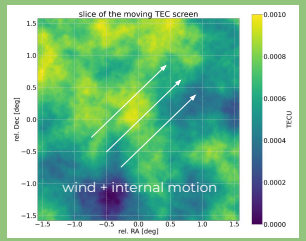
- be **complete** : contains all observational features (objects, instrumental effects, distributions...)
- have **pure** target catalog : training catalog should be as close as possible to detection expectancy

Simulations - *Overview of the full pipeline*

Sky Model
Taken from the SKA Data Challenge 3a (Bonaldi et al. 2025), generated from T-RECS (Bonaldi et al. 2018)

Interferometer layout
LOFAR High Band Antenna layout

Ionospheric screen
Generated from a Kolmogorov turbulent cascade power spectrum

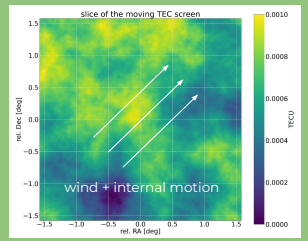


Simulations - *Overview of the full pipeline*

Sky Model
Taken from the SKA Data Challenge 3a (Bonaldi et al. 2025), generated from T-RECS (Bonaldi et al. 2018)

Interferometer layout
LOFAR High Band Antenna layout

Ionospheric screen
Generated from a Kolmogorov turbulent cascade power spectrum



OSKAR
(Oxford SKA simulator, Dulwich et al. 2009)

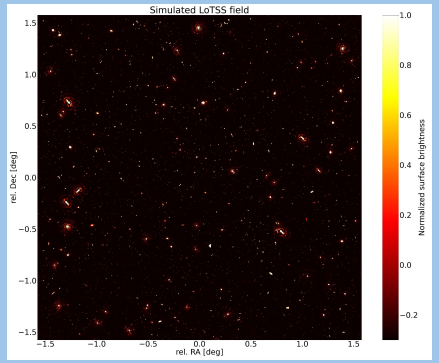
Astronomical radio interferometer simulator that produces the instrumental response with a given interferometer layout, ionospheric screen & sky model

Dirty image

WSClean
(W-Stacking Clean, Offringa et al. 2014)

Produce dirty image from visibilities
Removes artifacts/PSF with multi-scale model cleaning & occasional dirty residual re-computation (Coton-Schwab alg.)

Partial product

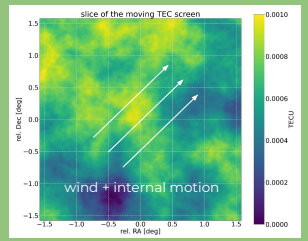


Simulations - *Overview of the full pipeline*

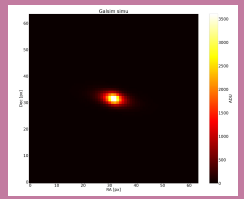
Sky Model
Taken from the SKA Data Challenge 3a (Bonaldi et al. 2025), generated from T-RECS (Bonaldi et al. 2018)

Interferometer layout
LOFAR High Band Antenna layout

Ionospheric screen
Generated from a Kolmogorov turbulent cascade power spectrum



Resolved galaxies
Simulation of the resolved galaxies using the Galsim package with Sersic profile and random inclination, orientation...



OSKAR
(Oxford SKA simulator, Dulwich et al. 2009)

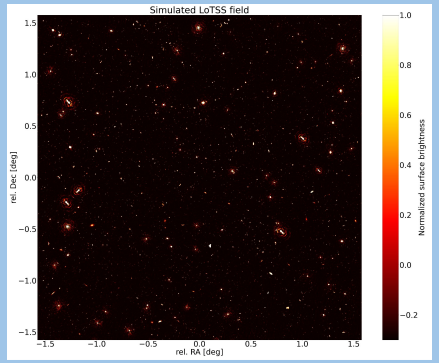
Astronomical radio interferometer simulator that produces the instrumental response with a given interferometer layout, ionospheric screen & sky model

Dirty image

WSClean
(W-Stacking Clean, Offringa et al. 2014)

Produce dirty image from visibilities
Removes artifacts/PSF with multi-scale model cleaning & occasional dirty residual re-computation (Coton-Schwab alg.)

Partial product

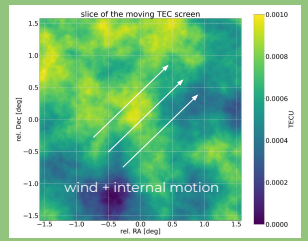


Simulations - *Overview of the full pipeline*

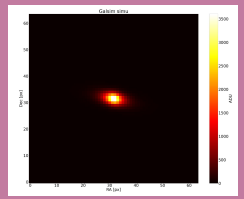
Sky Model
Taken from the SKA Data Challenge 3a (Bonaldi et al. 2025), generated from T-RECS (Bonaldi et al. 2018)

Interferometer layout
LOFAR High Band Antenna layout

Ionospheric screen
Generated from a Kolmogorov turbulent cascade power spectrum



Resolved galaxies
Simulation of the resolved galaxies using the Galsim package with Sersic profile and random inclination, orientation...

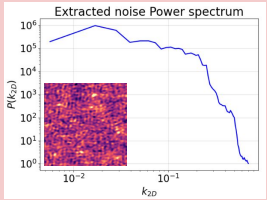


OSKAR
(Oxford SKA simulator, Dulwich et al. 2009)

Astronomical radio interferometer simulator that produces the instrumental response with a given interferometer layout, ionospheric screen & sky model

Dirty image

Thermal noise
Extracted from the data through power spectrum computation and regenerated with random re-sampling of the Fourier space

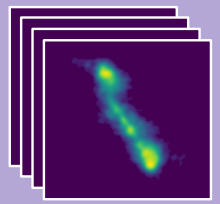


DATA
Here LOFAR data

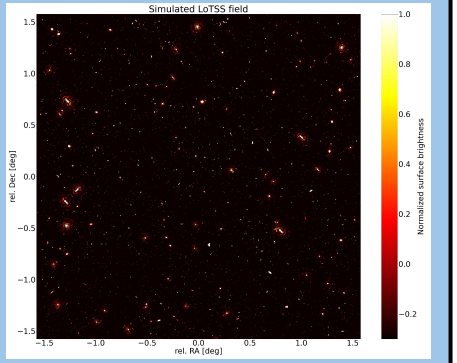
WSClean
(W-Stacking Clean, Offringa et al. 2014)

Produce dirty image from visibilities
Removes artifacts/PSF with multi-scale model cleaning & occasional dirty residual re-computation (Coton-Schwab alg.)

Resolved AGNs
Mask extraction from the data based on local maxima analysis

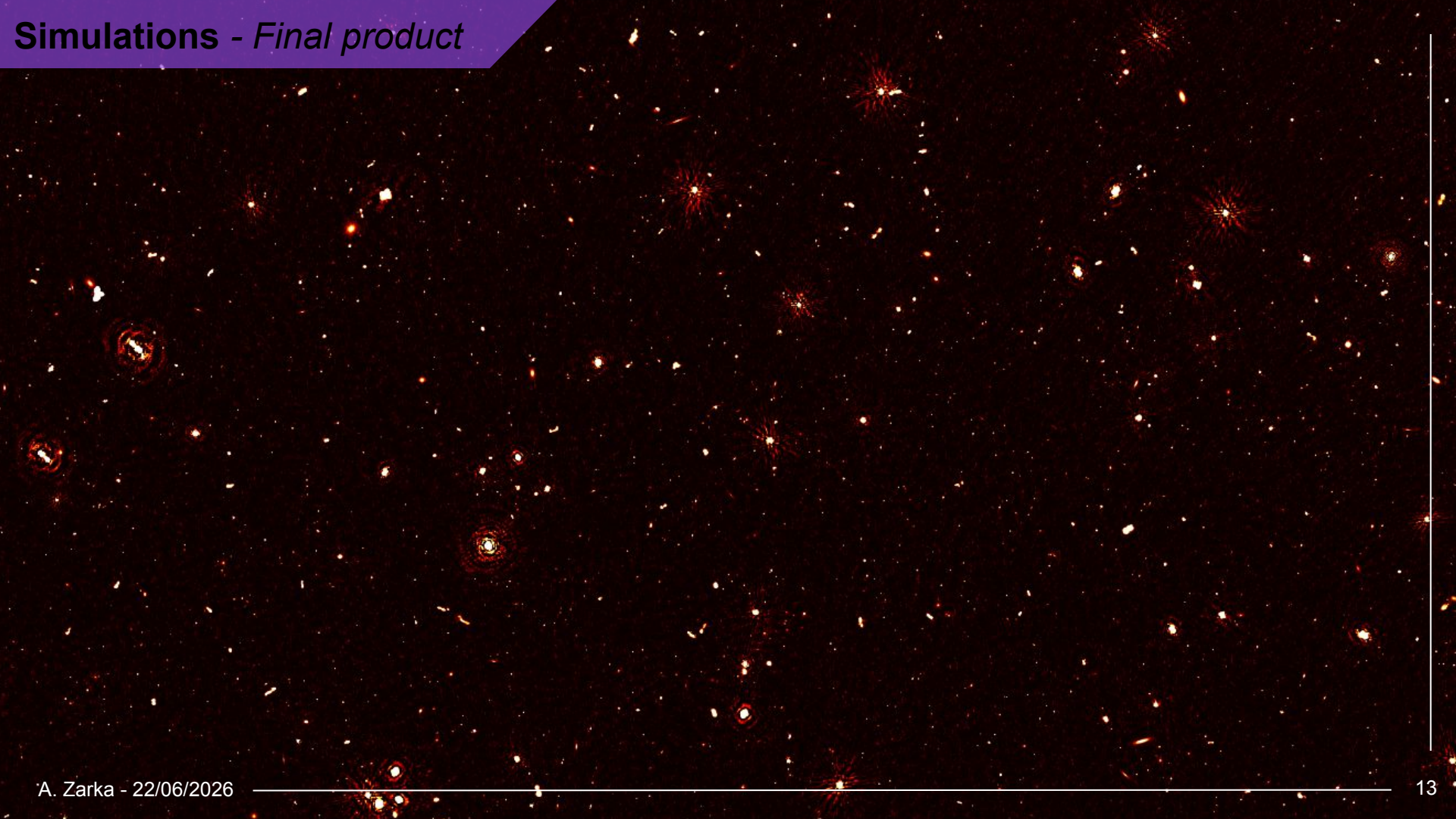


Final product

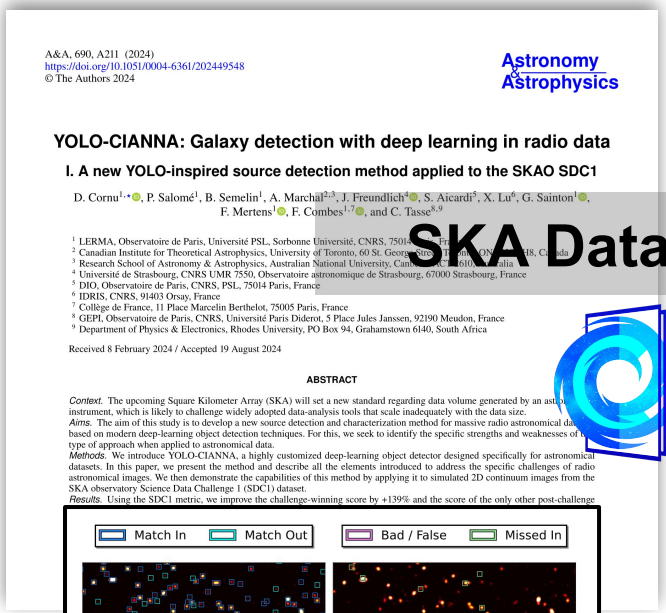


Used along with its associated source catalog (sky model) for network training

Simulations - *Final product*



Deep-learning detector - *YOLO/CIANNA framework*



A&A, 690, A211 (2024)
https://doi.org/10.1051/0004-6361/202449548
© The Authors 2024

Astronomy & Astrophysics

YOLO-CIANNA: Galaxy detection with deep learning in radio data

I. A new YOLO-inspired source detection method applied to the SKAO SDC1

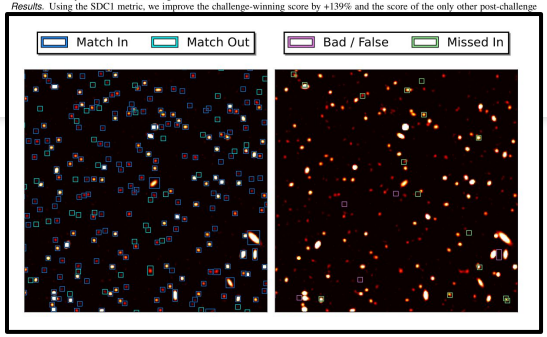
D. Cornu[1,*], P. Salomé[1], B. Semelin[1], A. Marchal[2,3], J. Freundlich[4], S. Aicardi[5], X. Lu[6], G. Sainton[1], F. Mertens[1], F. Combes[1,7], and C. Tasse[8,9]

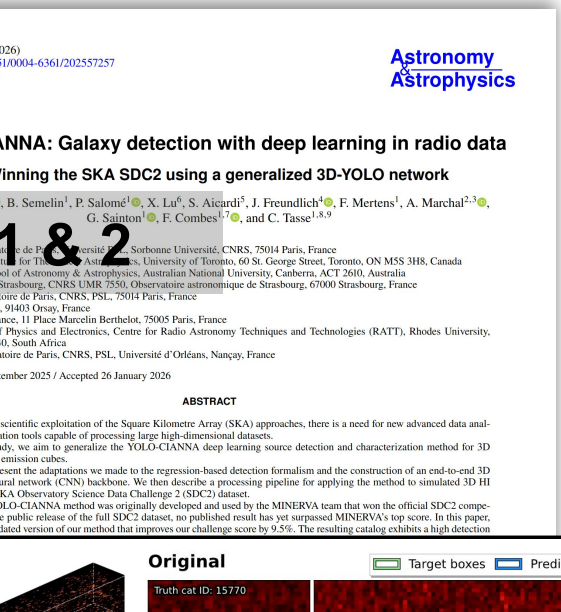
[1] LERMA, Observatoire de Paris, Université PSL, Sorbonne Université, CNRS, 75014 Paris, France
[2] Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada
[3] Research School of Astronomy & Astrophysics, Australian National University, Canberra, ACT 2610, Australia
[4] Université de Strasbourg, CNRS UMR 7550, Observatoire astronomique de Strasbourg, 67000 Strasbourg, France
[5] DIO, Observatoire de Paris, CNRS, PSL, 75014 Paris, France
[6] IDRIS, CNRS, 91403 Orsay, France
[7] Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France
[8] GEPI, Observatoire de Paris, CNRS, Université Paris Diderot, 5 Place Jules Janssen, 92190 Meudon, France
[9] Department of Physics & Electronics, Rhodes University, PO Box 94, Grahamstown 6140, South Africa

Received 8 February 2024 / Accepted 19 August 2024

ABSTRACT

Context. The upcoming Square Kilometer Array (SKA) will set a new standard regarding data volume generated by an ast... instrument, which is likely to challenge widely adopted data-analysis tools that scale inadequately with the data size.
Aims. The aim of this study is to develop a new source detection and characterization method for massive radio astronomical da... based on modern deep-learning object detection techniques. For this, we seek to identify the specific strengths and weaknesses of t... type of approach when applied to astronomical data.
Methods. We introduce YOLO-CIANNA, a highly customized deep-learning object detector designed specifically for astronomical datasets. In this paper, we present the method and describe all the elements introduced to address the specific challenges of radio astronomical images. We then demonstrate the capabilities of this method by applying it to simulated 2D continuum images from the SKA observatory Science Data Challenge 1 (SDC1) dataset.
Results. Using the SDC1 metric, we improve the challenge-winning score by +139% and the score of the only other post-challenge





A&A, 707, A203 (2026)
https://doi.org/10.1051/0004-6361/202557257
© The Authors 2026

Astronomy & Astrophysics

YOLO-CIANNA: Galaxy detection with deep learning in radio data

II. Winning the SKA SDC2 using a generalized 3D-YOLO network

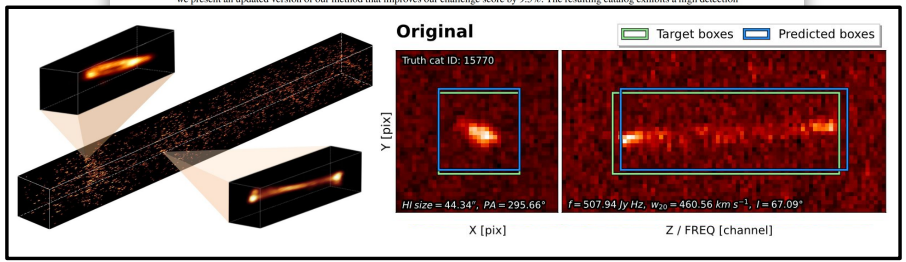
D. Cornu[1,*], B. Semelin[1], P. Salomé[1], X. Lu[6], S. Aicardi[5], J. Freundlich[4], F. Mertens[1], A. Marchal[2,3], G. Sainton[1], F. Combes[1,7], and C. Tasse[1,8,9]

[1] LERMA, Observatoire de Paris, Université PSL, Sorbonne Université, CNRS, 75014 Paris, France
[2] Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada
[3] Research School of Astronomy & Astrophysics, Australian National University, Canberra, ACT 2610, Australia
[4] Université de Strasbourg, CNRS UMR 7550, Observatoire astronomique de Strasbourg, 67000 Strasbourg, France
[5] DIO, Observatoire de Paris, CNRS, PSL, 75014 Paris, France
[6] IDRIS, CNRS, 91403 Orsay, France
[7] Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France
[8] Department of Physics and Electronics, Centre for Radio Astronomy Techniques and Technologies (RATT), Rhodes University, Makhanda 6140, South Africa
[9] ORN, Observatoire de Paris, CNRS, PSL, Université d'Orléans, Nançay, France

Received 15 September 2025 / Accepted 26 January 2026

ABSTRACT

Context. As the scientific exploitation of the Square Kilometre Array (SKA) approaches, there is a need for new advanced data analysis and visualization tools capable of processing large high-dimensional datasets.
Aims. In this study, we aim to generalize the YOLO-CIANNA deep learning source detection and characterization method for 3D hyperspectral HI emission cubes.
Methods. We present the adaptations we made to the regression-based detection formalism and the construction of an end-to-end 3D convolutional neural network (CNN) backbone. We then describe a processing pipeline for applying the method to simulated 3D HI cubes from the SKA Observatory Science Data Challenge 2 (SDC2) dataset.
Results. The YOLO-CIANNA method was originally developed and used by the MINERVA team that won the official SDC2 competition. Despite the public release of the full SDC2 dataset, no published result has yet surpassed MINERVA's top score. In this paper, we present an updated version of our method that improves our challenge score by 9.5%. The resulting catalog exhibits a high detection

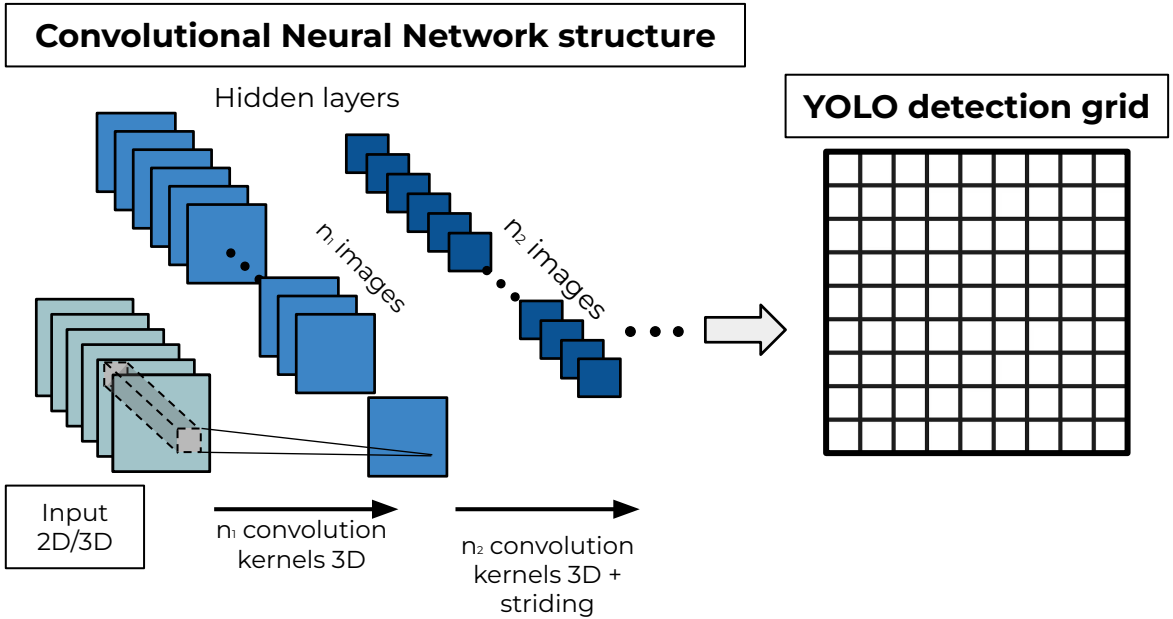


SKA Data Challenges 1 & 2

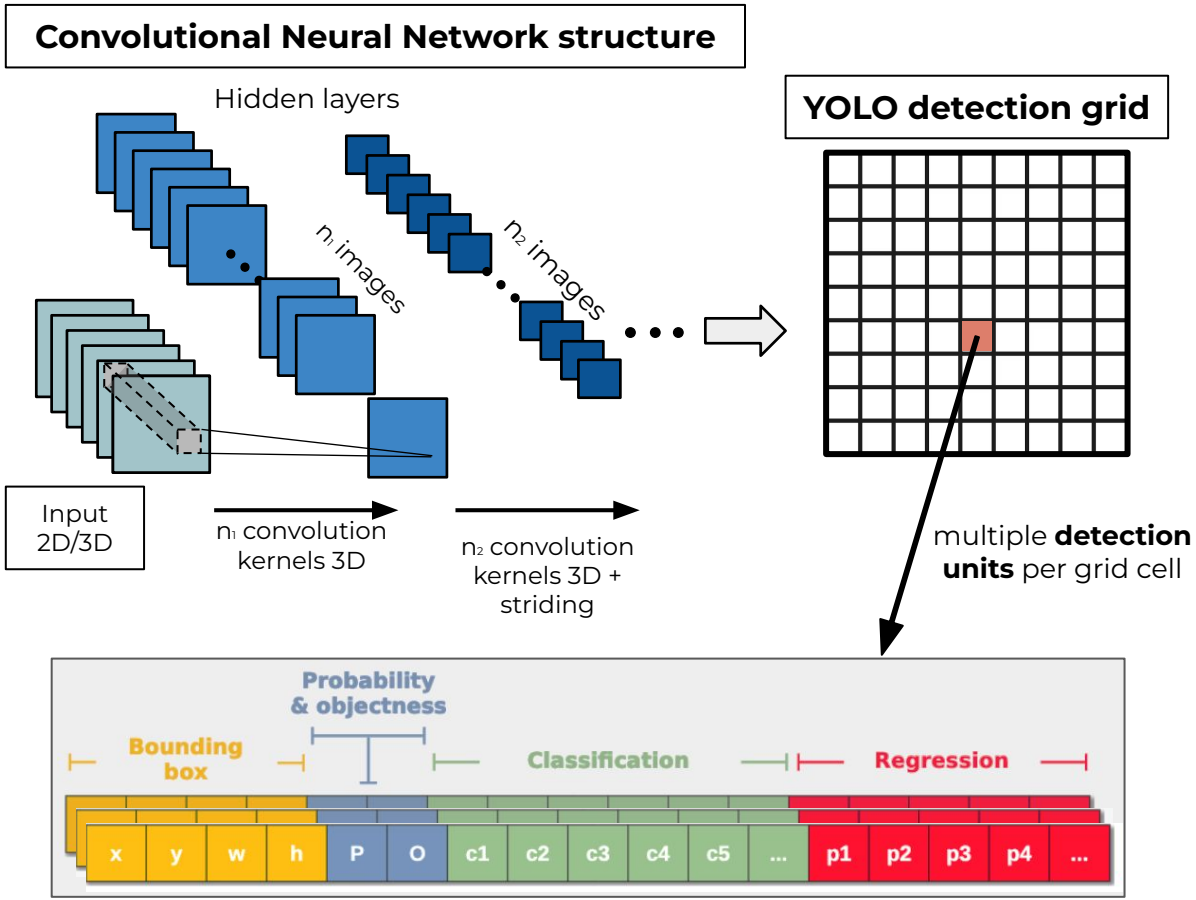
CIANNA

Method reaching the highest score on both SDC 1 & 2 data !

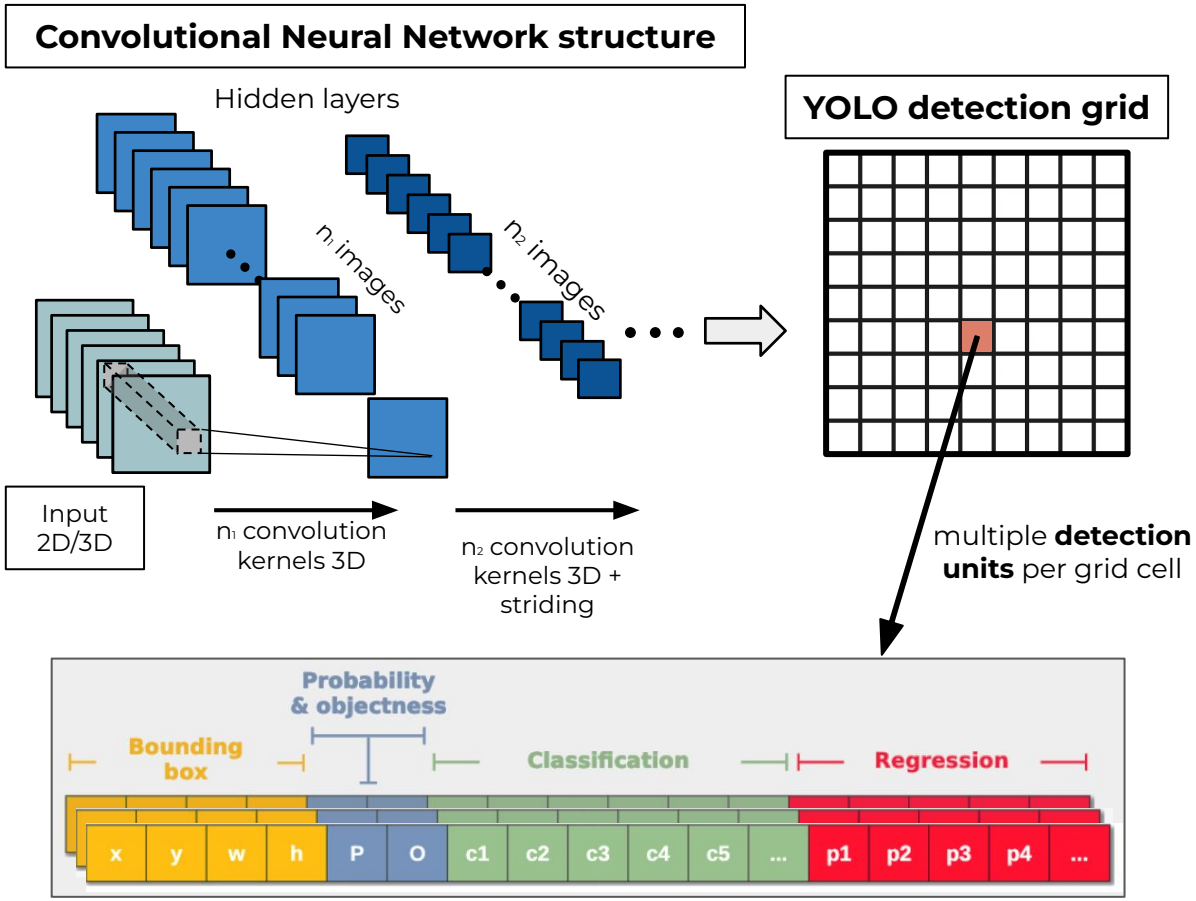
Deep-learning detector - *YOLO/CIANNA framework*

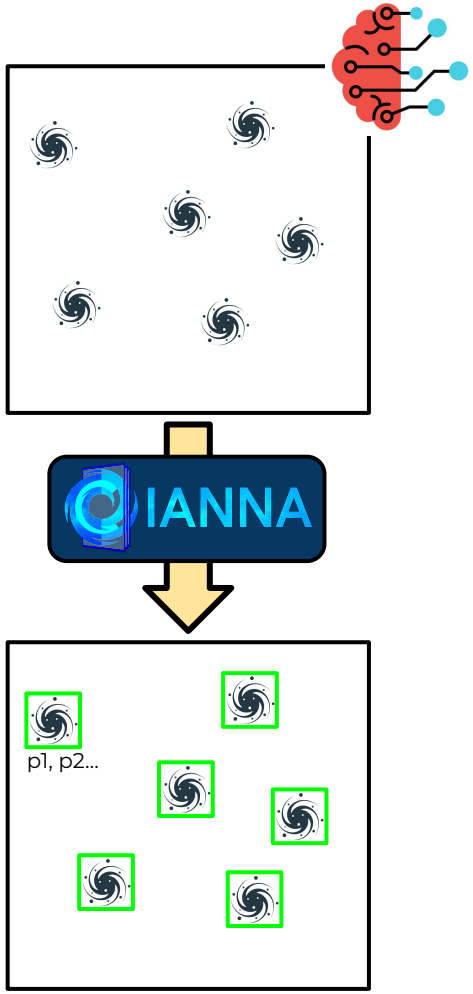


Deep-learning detector - *YOLO/CIANNA framework*



Deep-learning detector - *YOLO/CIANNA framework*

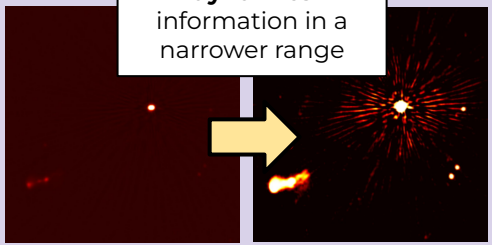




Deep-learning detector - *Training*

Pre-process

Normalization to compress the **dynamics** information in a narrower range

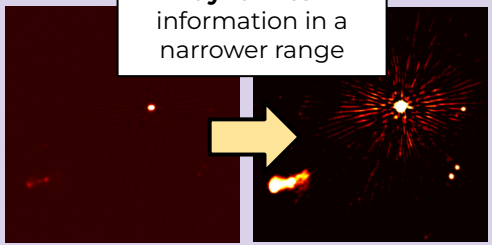


Selection/filtering function of training **targets** based on integrated flux and filtering (NMS, artifacts cleaning)

Deep-learning detector - *Training*

Pre-process

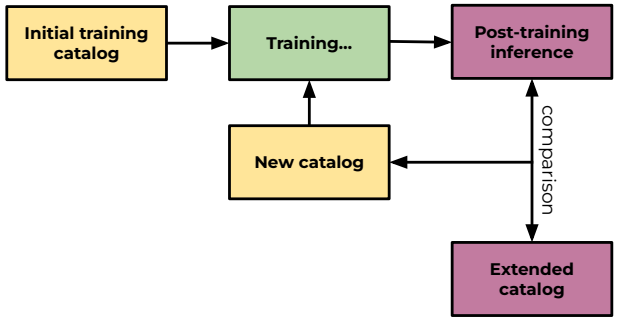
Normalization to compress the **dynamics** information in a narrower range



Selection/filtering function of training **targets** based on integrated flux and filtering (NMS, artifacts cleaning)

Training *of the YOLO-CIANNA model...*

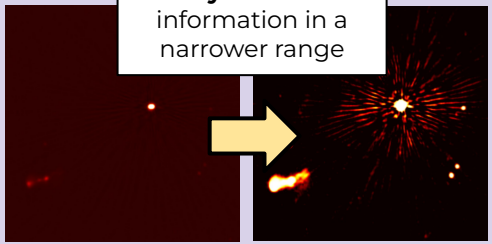
Multi-step **bootstrap training** to adapt the catalog purity



Deep-learning detector - *Training*

Pre-process

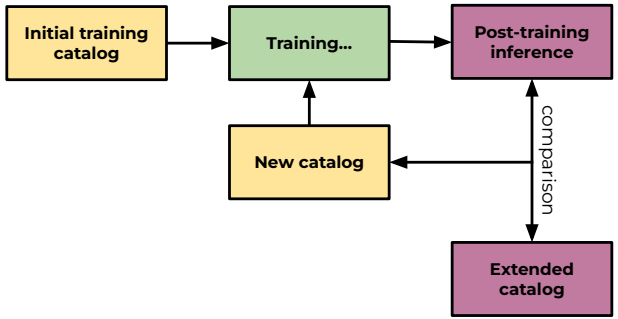
Normalization to compress the **dynamics** information in a narrower range



Selection/filtering function of training **targets** based on integrated flux and filtering (NMS, artifacts cleaning)

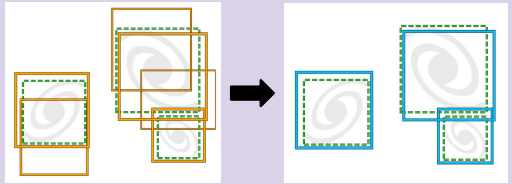
***Training** of the YOLO-CIANNA model...*

Multi-step **bootstrap training** to adapt the catalog purity

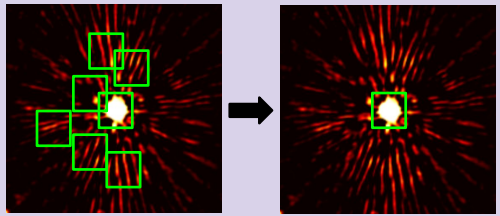


Post-process

1. **Threshold on detection score**
2. **Non-max suppression**



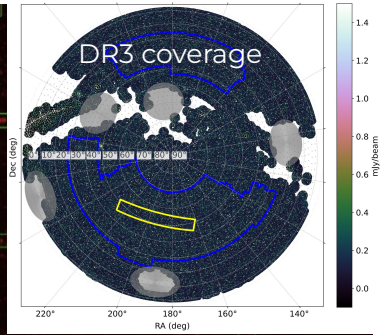
3. **Artifacts filtering**



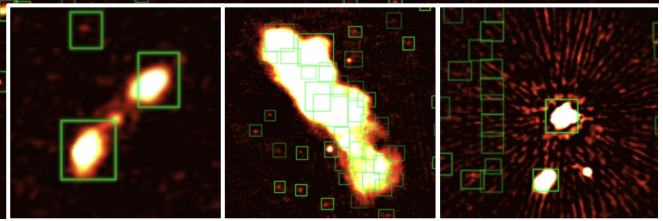
Inference - *Test inference on a LoTSS field*

Selected field : 8h pointing from the region of the **HETDEX Spring field - P18Hetdex03**
~3.2°x3.2° wide field at a resolution of 6”
Noise RMS at 83 µJy/beam

Inference time per mosaic : ~ 5s

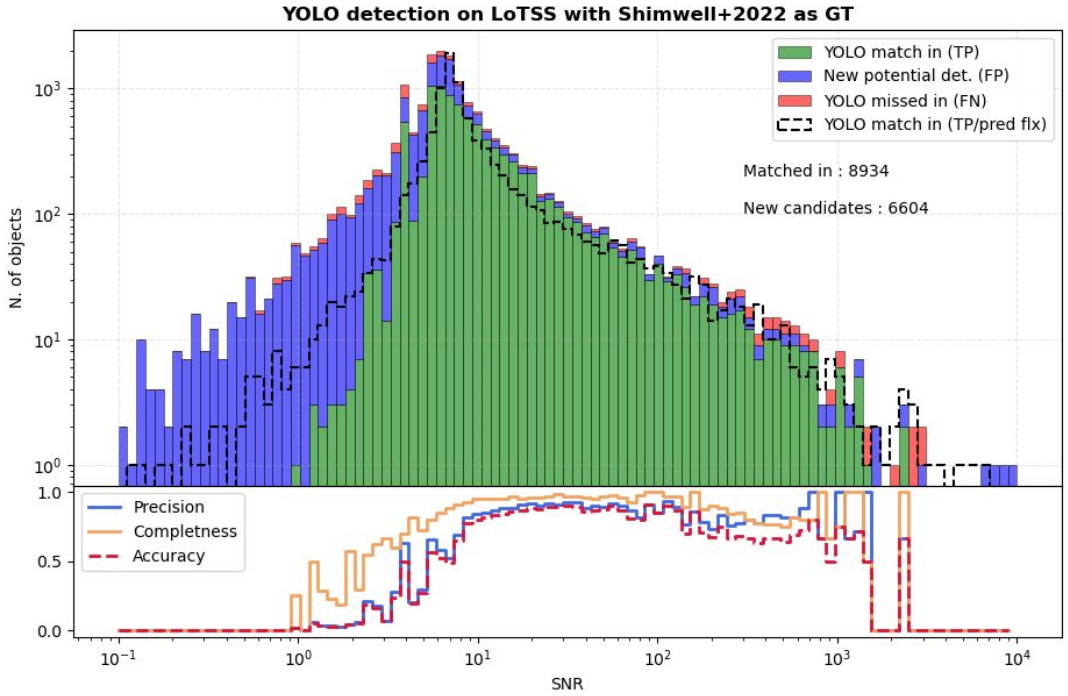


Difficult cases



1’

Inference - *Test inference on a LoTSS field*



Comparison with the **Shimwell+2022 raw radio component catalog of LoTSS-DR2**

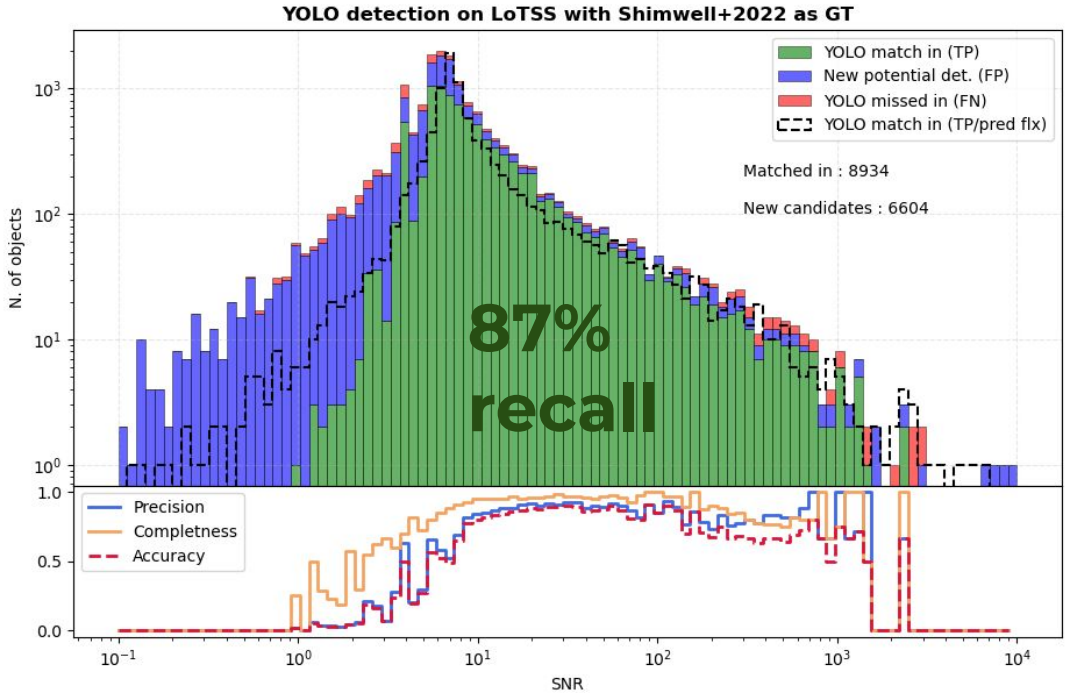
- Catalog computed using **PyBDSF** (Mohan & Rafferty 2015)
- Refined through source association & deblending, and **cross-identification** with optical/infrared

➔ YOLO shows ~60% precision at a recall of 87%

BUT ~1.5 times more sources detected

➔ **63%** of **IR counterpart** among **new candidates** (allWISE)

Inference - *Test inference on a LoTSS field*



Comparison with the **Shimwell+2022 raw radio component catalog of LoTSS-DR2**

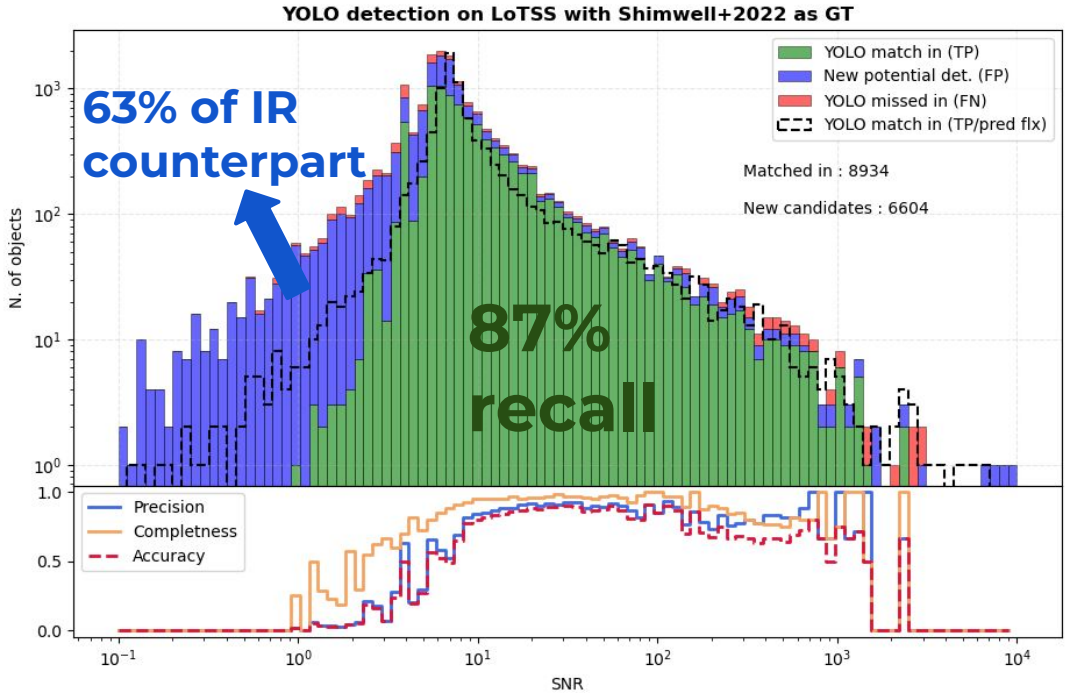
- Catalog computed using **PyBDSF** (Mohan & Rafferty 2015)
- Refined through source association & deblending, and **cross-identification** with optical/infrared

➔ YOLO shows ~60% precision at a recall of 87%

BUT ~1.5 times more sources detected

➔ **63%** of **IR counterpart** among **new candidates** (allWISE)

Inference - *Test inference on a LoTSS field*



Comparison with the **Shimwell+2022 raw radio component catalog of LoTSS-DR2**

- Catalog computed using **PyBDSF** (Mohan & Rafferty 2015)
- Refined through source association & deblending, and **cross-identification** with optical/infrared

➔ YOLO shows ~60% precision at a recall of 87%

BUT ~1.5 times more sources detected

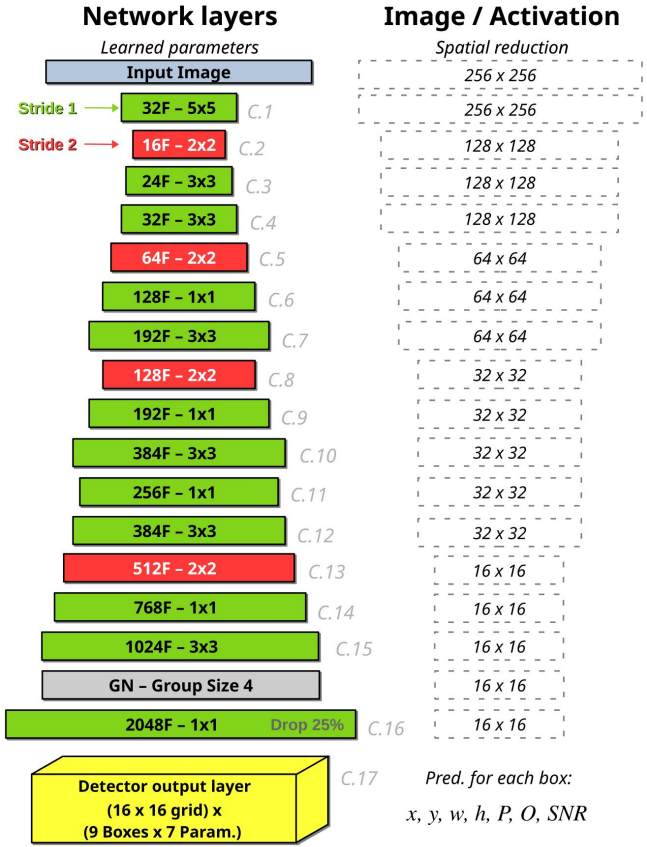
➔ **63%** of **IR counterpart** among **new candidates** (allWISE)

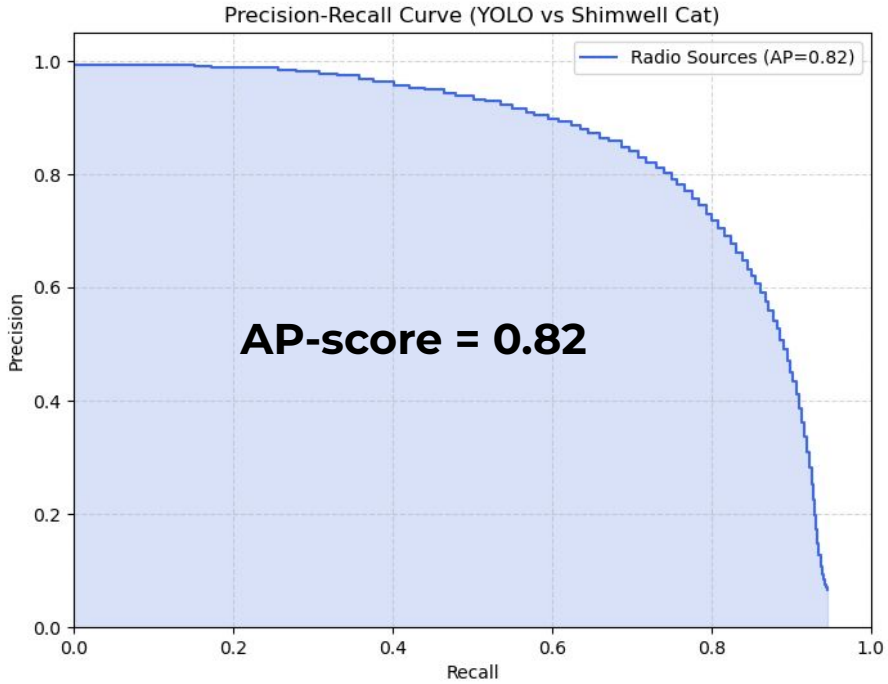
Summary

- **Simulation pipeline** development for supervised training
- Applying the **YOLO-CIANNA** method optimized for source detection
- Robust pipeline for source detection that **matches state of the art performances** on real-world data
- Higher performances in the management of “hard detections” (**low SNR sources**, artifacts management, various source morphologies…)
- **New candidates** + cross-identification with IR survey.
- Alternative **self-supervised** methods could be used to refine model during training

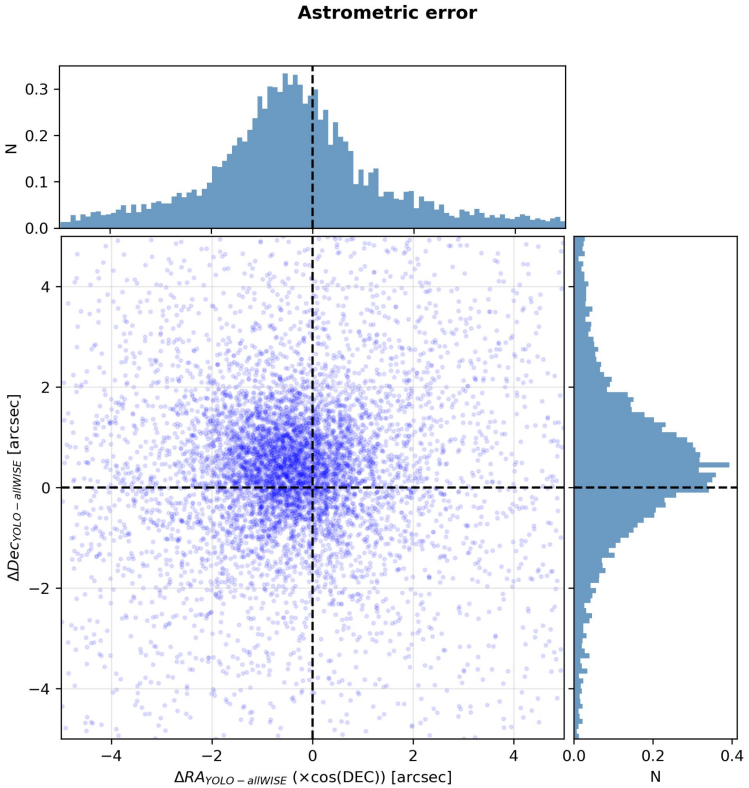
Appendix

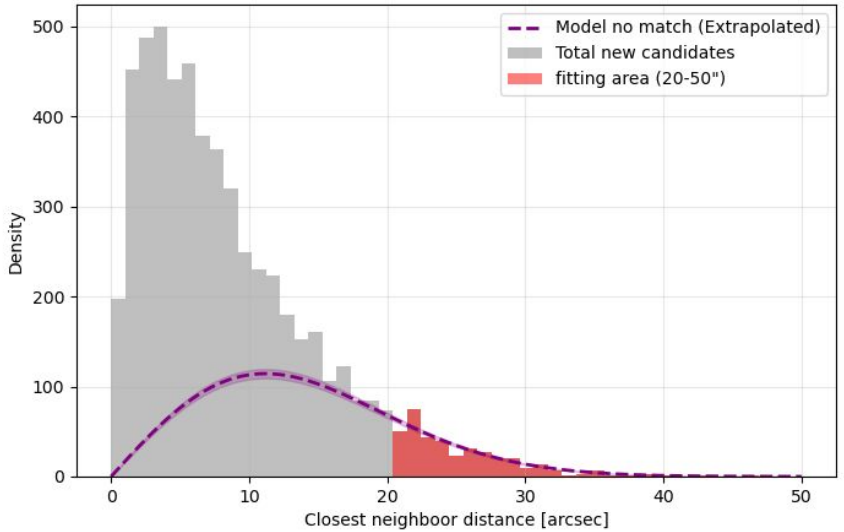
Inference - *Test inference on a LoTSS field, mAP-curve*



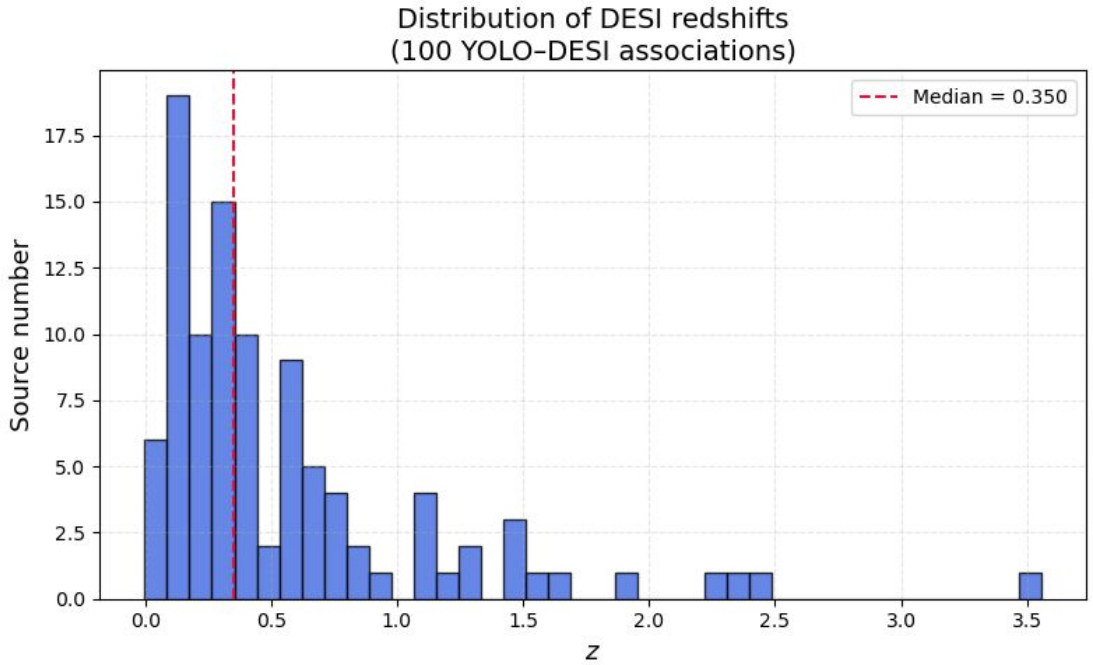


Inference - *Astrometric precision & new candidates purity*



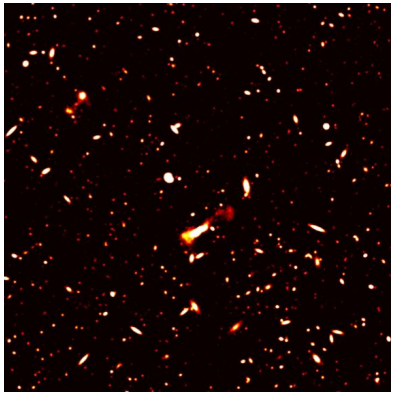


Inference - *Astrometric precision & new candidates purity*



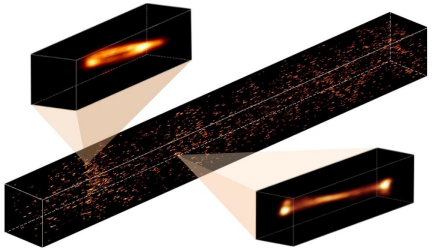
SKA Data Challenges

Precursor SKA data challenges :



SDC1 : 2D continuum data

=> Detect and characterize radio sources
Best score a posteriori
Cornu+24



SDC2 : 3D hyper-spectral HI data

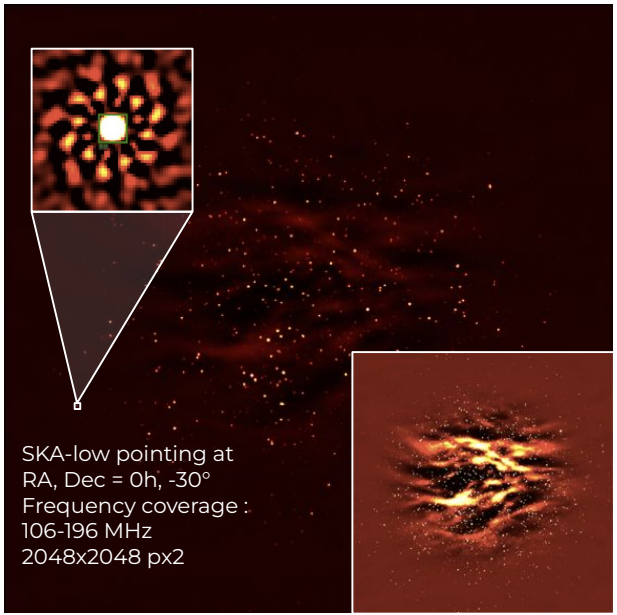
=> Detect and characterize radio sources
1st place with best score
Cornu+25

SDC3a : 3D hyper-spectral data for EoR experiment

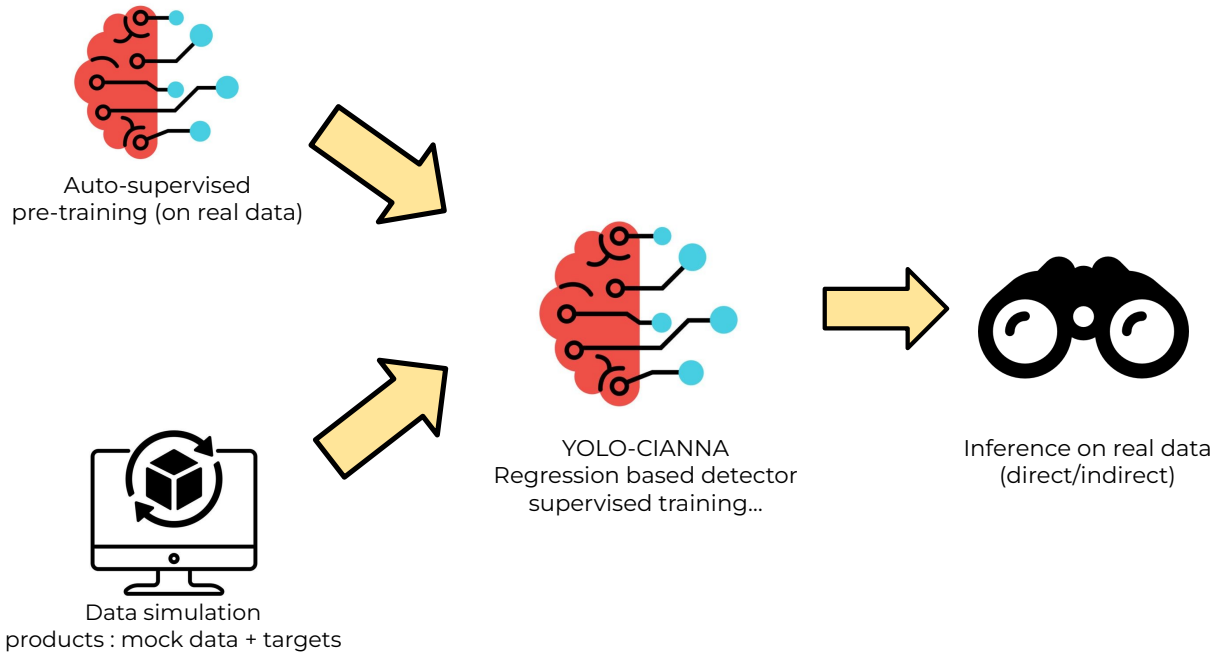
=> Remove obscuring sources for an underlying H-21 cm EoR signal analysis

=> Integrated to 2D continuum data for our detection purpose

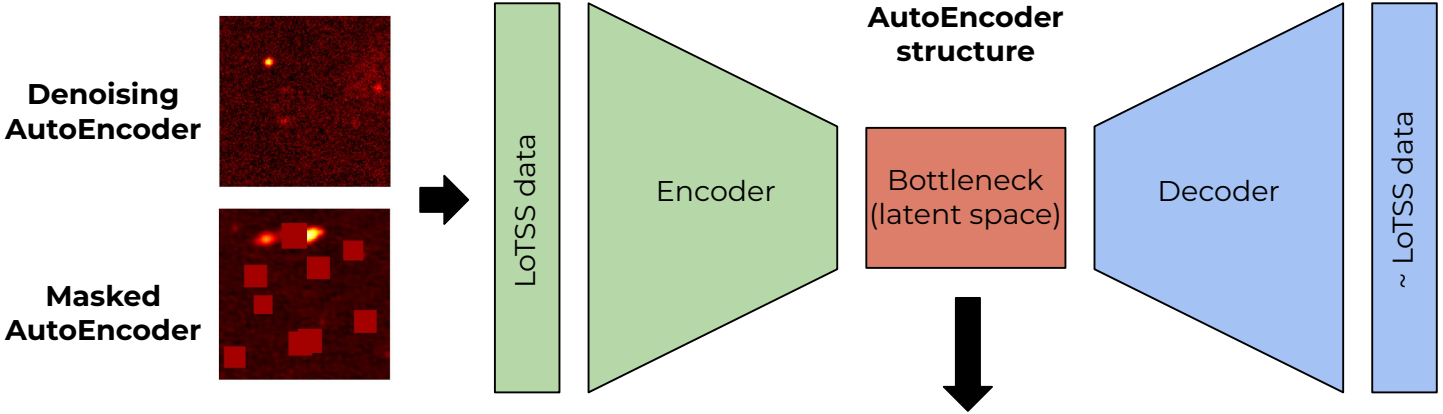
SDC3 full field - 10°x10°



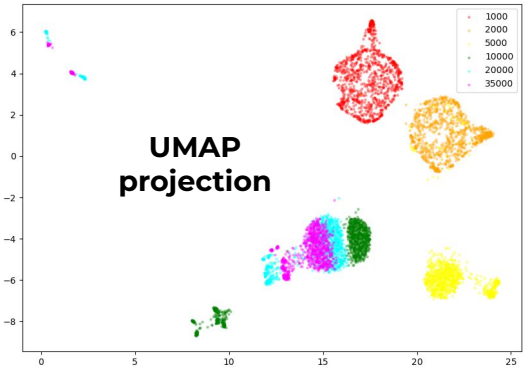
Second approach - *Pre-training through auto-supervised task on real-world data*



Auto-supervised pre-training



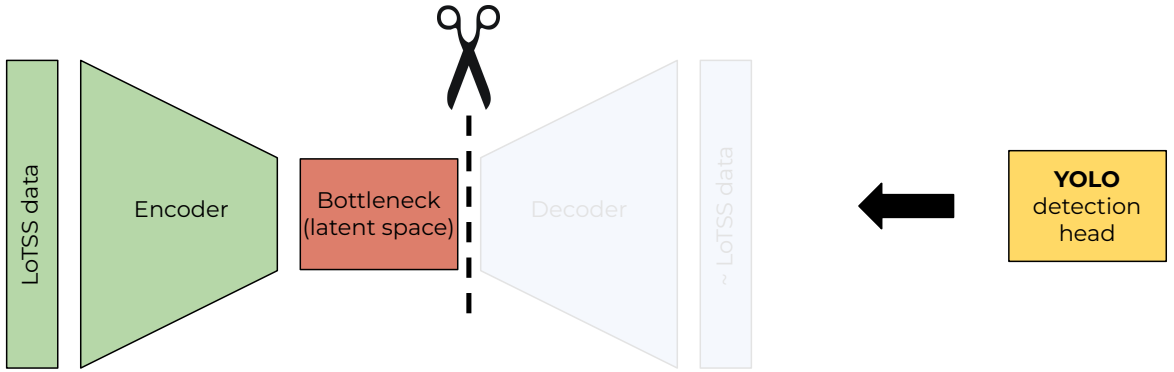
These tasks allow the network to really “learn” the data structure, not just an identity function

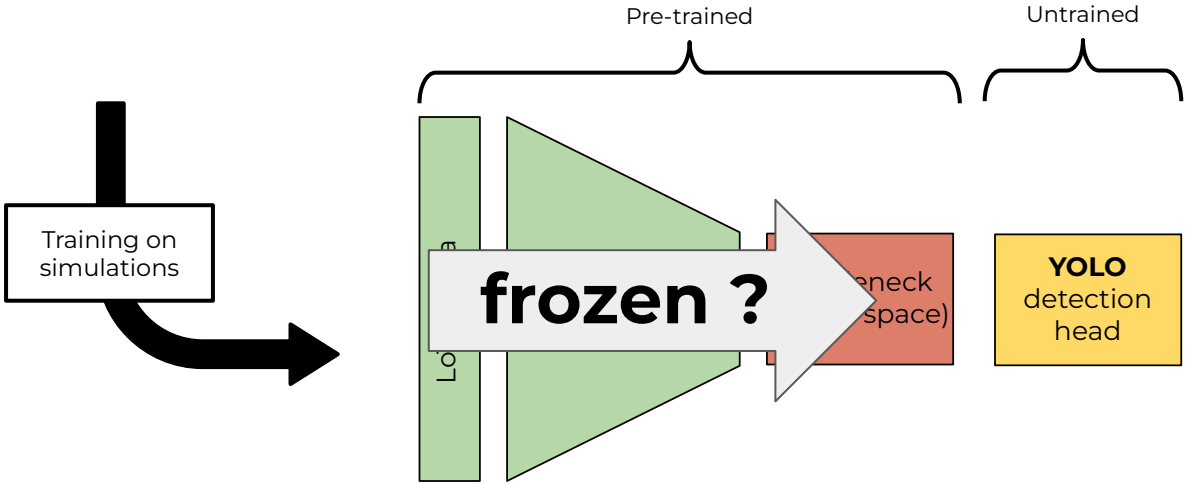


Architecture ConvNext inspired
-> Few-Many-Few residual blocks

-> Few-Few-Many

Auto-supervised pre-training





After pre-training and construction of the full detection architecture, the rest of the training is exactly the same as the first approach with the encoder part frozen during training (simulated data pre-process, neural training, post-process during inference...)