

Foundation Models & Deep Learning for Astrophysics

Collaborative Research Projects with CentraleDigitalLab

Anne-Laure MEALIER

AI Research Engineer · Head of CentraleDigitalLab

Centrale Méditerranée

anne-laure.mealier@centrale-med.fr



Séminaire IA & Astrophysique
Laboratoire d'Astrophysique
de Marseille (LAM)

May 2026

Centrale Méditerranée · CentraleDigitalLab

About Me

Anne-Laure MEALIER

AI Research Engineer & Head of CentraleDigitalLab

PhD — Cognitive Robotics

Thesis: *How does language impose the structure of meaning: construal and narration* - European project WYSIWYD "What You Say Is What You Did"

- Automatic grounding of natural language instructions onto robot action sequences.
- Combines speech, gesture recognition and human-robot interaction.

Research & Teaching

Specialisation in deep learning, NLP, computer vision and scientific AI.
Teaches AI methods to engineering students at Centrale Méditerranée.

Industrial & Scientific Projects

Supervises student projects with companies, labs and institutions across many sectors.



PhD

Cognitive Robotics

AI Focus

DL · NLP · CV · GenAI

Workplace

Centrale Méditerranée
Affiliated with the Institut Fresnel



Collab

Laboratory · Industry · Institut

My Science Background: WYSIWYD

What You Say Is What You Did — Cognitive Robotics, PhD Thesis

The Problem

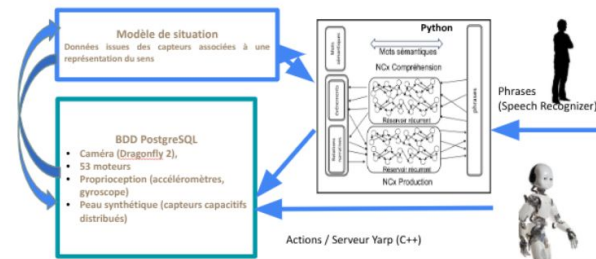
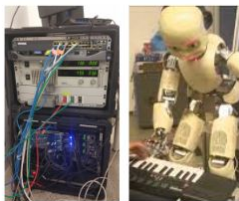
Robots need to understand natural language instructions and link them to the physical actions they have already performed.

Key Contribution

A multimodal grounding framework combining speech recognition, gesture interpretation and temporal reasoning to close the loop between language and robot motor sequences.

Methods

- Automatic speech recognition (ASR)
- Gesture & motion segmentation
- Sequence alignment between verbal descriptions and action traces
- Tested on physical robot platforms (manipulation tasks)



Speech Input

"Pick up the red cube"
→ ASR + NLU parsing

Robot Action

Motor sequence recorded
from sensors + vision

Grounding

Align language tokens
to action subsequences

Learning

Iterative self-supervised
update of action vocabulary

Research: Eye Tracking, Electrophysiology & Polyhandicap

Eyes Tracking · Electrophysiology · Non-verbal Language · AAC Communication

Goal

Develop AI-assisted communication tools for non-verbal individuals with profound multiple disabilities (polyhandicap) and aphasia, using EEG signals to decode emotional and affective states in real time and support caregivers in understanding patient needs.

Partners

ARI · CPPM · Centrale Méditerranée · Ampiric · Aix-Marseille Université

Hardware

Bitbrain Air wireless EEG headset with 8 dry electrodes, enabling rapid setup, real-time acquisition, and deployment in ecological environments. Optional integration of eye tracking for future multimodal interaction and calibration improvements.

AI

Patient-specific Self-Organizing Maps (SOM), EEG spectral and connectivity analysis, frontal alpha asymmetry biomarkers, and interpretable machine learning for emotional-state decoding.

Methods

Numéro de dossier, vague N°, année (à remplir par le CNRS)



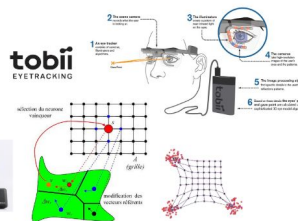
A-Identité du projet			
Titre du projet		Associer des algorithmes à l'interprétation des Emotion	
Acronyme		ALIE	
Porteur	M. ☐ Mme ☐	Nom : RIVIERE	Prénom :
	☐ Chercheur		☐ Ingénieur
	☐ Enseignant chercheur		☐ Autre :
e courriel : riviere@cpm.in2p3.fr		Téléphone :	
Institut principal du CNRS		Nucléaire & Particules	Délégation r4
Unité (nom, numéro, adresse)		UMR7346	
Centre financier de l'unité (demandeur au gestionnaire du laboratoire)		0933	
Contact SPV :		Clara DIANCOURT - clara.diancourt@cpm.in2p3.fr	
Objet de l'invention/innovation			
Un nouveau logiciel ?	☐ OUI ☐ NON	Un nouveau procédé de fabrication	☐ OUI ☐ NON
Un nouveau logiciel libre	☐ OUI ☐ NON	Un nouvel usage	☐ OUI ☐ NON
Autre (précisez) ? /commentaires ?			

Résumé du projet en 10 lignes (à destination des non-spécialistes du dom
Nous proposons de développer une solution technologique innovante po
de personnes aphasiques à partir de signaux physiologiques :



Eyes Tracking - Electrophysiology
Langage non verbal - Polyhandicap

- 16 illuminateurs
- quatre caméras oculaires intégrées
- Caméra de scène avec un champ de vision de 106°
- Microphone
- Gyroscope
- Accéléromètre (sampled at 100 Hz);
- Magnétomètre (sampled at 10 Hz)



https://fr.wikipedia.org/wiki/Carte_automatiquement_générée_à_descriptions

Patient-Specific EEG-Based Emotion Detection Using Self-Organizing Maps: A Technical Validation and Roadmap for Clinical Application in Polyhandicapped Patients with Aphasia

Anne-Laure MEALIER^{1,2,3,4}, Françoise RIVIERE^{5,6}, Victoria CIARLET THAON⁵

¹ Institut FRESNEL (Faculté des Sciences et Techniques, Marseille, France)

² Aix-Marseille Université : UMR7249 / UMR 6133

³ École Centrale de Marseille

⁴ CNRS : UMR7249 / UMR 6133

⁵ Centre de Physique des Particules de Marseille (CPPM)

⁶ CNRS/IN2P3 : UMR7346

March 4, 2026

Abstract

Brain injury, polyhandicap, and aphasia often lead to profound impairments in communication, limiting caregivers' ability to detect internal emotional states such as pain, fear, discomfort, or positive affect. Electroencephalography-based brain-computer interfaces (EEG-BCIs) offer a promising avenue for restoring minimal communication in individuals unable to communicate verbally or through gestures. However, inter-patient variability in neural organization makes population-level models insufficient for this clinical context.

We propose a patient-specific framework for emotion decoding using an 8-electrode EEG system combined with Self-Organizing Maps (SOMs) trained independently for each patient. In this technical validation study, we present results from a 4.7-minute session on a healthy female volunteer (age 42), demonstrating the feasibility of the system from acquisition to classification. Despite the short duration, the system successfully extracted 279 analysis windows, identified 6 distinct emotional states from 9

CentraleDigitalLab

A 12-month intensive AI training program - combining courses, industrial projects & internships

Two student cohort intakes per year - One training program intake in September and another in February.

Intensive Courses

7 weeks

Refresher courses (79h)

Mathematics for AI (22h) · Programming (25h) · Data Analysis & Visualisation (22h) · Web (10h)

AI & Data Science (94h)

Machine Learning (22h) · Deep Learning (31) · Big Data (21) ML Quantum Computing (20h)

AI Ethics, Data Governance, Soft Skills (15h) · Agile / Scrum (7h)

Company Projects

2 × 2 months

Two real-world AI projects with companies, start-ups, research labs or public institutions.

Students deliver production-ready software and present results.

3 to 4 students per project group

Internship

Students apply acquired skills in a company or research lab context.

Assessed on: Innovation · Development · Project Management · Communication.

Duration of 4 to 6 months

Student Profiles: engineers & scientists (Bac+5) from diverse backgrounds — CS, physics, mathematics, engineering sciences

Example CDL Projects — Computer Vision

Drone Pilot Landing Performance · LVCim

Evaluation of drone pilot landing performance on simulators



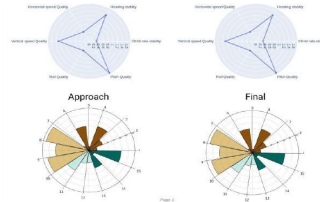
- Automatic identification of flight segments (approach, final, landing, taxiing).
- Trajectory prediction.
- Analysis of flight variations and stability.

Performance Evaluation Report

13/02/2023

1. The table illustrates the pilot performance:

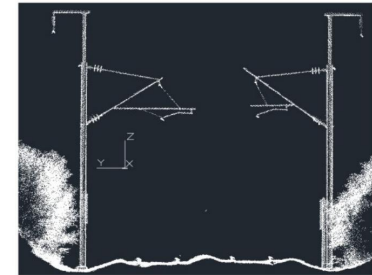
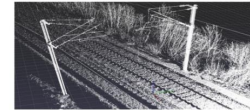
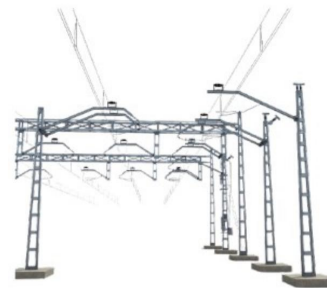
	horizontal_speed	climb_rate	heading_stable	pitch	roll
close mean	[strength: 0.75]	[weakness: 0.25]	[strength: 1.0]	[weakness: 1.0]	[weakness: 1.0]
little std	[weakness: 0.5]	[weakness: 0.5]	[weakness: 0.5]	[weakness: 1.0]	[weakness: 1.0]
gentle slope	[strength: 1.0]	[weakness: 0.5]	[strength: 0.5]	[strength: 1.0]	[weakness: 0.5]
no abrupt variations	[weakness: 0.75]	[weakness: 0.5]	[weakness: 0.5]	[weakness: 1.0]	[weakness: 1.0]



Automatic flight segment identification · Trajectory prediction
· Performance evaluation reports with radar charts

Catenary Detection on Railway Tracks · TSO Caténaires

Detection of catenary features on railway tracks



3D point cloud analysis from LiDAR scans · Detection of catenary support structures and overhead wire geometry to detect anomalies on catenaries

Example CDL Projects - Vision & NLP

Composite Rotor Defect Detection · Airbus Helicopters

Detection of defects in composite helicopter rotor parts

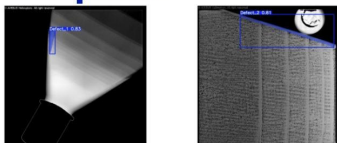
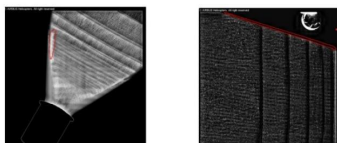


Figure 17: Results of YOLO



Defect 1

Defect 2

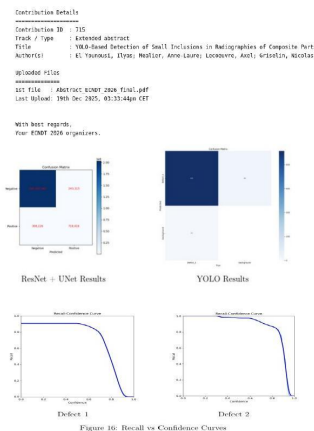


Figure 16: Recall vs Confidence Curves

Defect	mAP (Test Set)	mAP (Validation Set)
Defect 1	0.874	0.995
Defect 2	0.994	0.995

YOLO-based detection on X-ray images · mAP 0.874–0.994 · Accepted ECNDT 2026

[AI]rgos / AthenIA — ENSOSP



[AI]rgos

Extraction multimodale et formalisation des règles de sécurité incendie pour les préventionnistes : experts de l'analyse des risques et de l'application de la réglementation.

BUT : Faire une instruction automatisée des demandes d'autorisation de travaux en ERP : Établissements Recevant du Public.

Développement d'un POC qui en partant du pré-rapport technique génère une première ébauche de rapport réglementaire.



AthenIA

Accès intelligent et optimisé à la documentation de l'ENSOSP.

Valoriser et exploiter intelligemment sa base documentaire.

L'objectif est de permettre aux utilisateurs internes (élèves, formateurs, cadres) d'interroger efficacement cette base, d'identifier les savoirs existants et de faciliter leur réutilisation.

RAG system for fire safety regulation · LLM + ChromaDB · Automated inspection report generation

AI Projects Across Sectors

63% large corporations · 25% start-ups · 12% SMEs · Aix-Marseille & Nice ecosystem

Company projects examples

Computer Vision (Computer Vision / Deep Learning)

Defect detection in images

(quality control of helicopter rotor components, defect detection on banana bunches)

Biological and scientific image segmentation

(detection of phagocytes and lymphocytes, segmentation of planetary nebulae)

Automated segmentation of corneal structures for morphometric analysis (detection of the epithelium and endothelium to measure corneal thickness and the intermediate space)

Automatic anonymization in medical videos (cardiac surgery)

Automatic detection of delivery notes from images or scanned documents

Natural Language Processing (NLP / LLM)

Development of specialized chatbots (industrial maintenance, legal data, documentary repositories for educational foundations)

Semantic content extraction from scientific reports

Evaluation and Autonomous Systems

Flight evaluation for drone pilot qualification

Bioinformatics and Drug Discovery

Detection of mutations in DNA sequences

Large-scale molecular screening driven by Deep Learning for the identification of “viral factory” disruptors

Optimization and Operations Research

Optimization of delivery routes for construction-site vehicles

Anticipation of resource availability for emergency services at the Bataillon de marins-pompiers de Marseille

Scientific Modeling and Simulation

Wave propagation studies on oil risers

Applied Predictive Analytics

Financial forecasting (revenue prediction)

Marketing analysis and customer profiling studies

Time Series Analysis and Sensor Data

Forecasting time series data from connected sensors

Extraction and analysis of in-flight data from Dassault Falcon aircraft sensors

Computer Vision & Deep Learning

Defect detection · PNe detection · Image segmentation · Anonymous medical videos

NLP / LLM

Specialised chatbots · Regulatory reports · Knowledge base queries · Semantic extraction

Time Series & Signal Processing

Flight video analysis · Drone pilot evaluation · Eye tracking · IoT sensor prediction

Scientific AI & Simulation

Euclid spectral detection · Galaxy catalogue generation · Wave propagation · Drug discovery

CentraleDigitalLab — Partner Ecosystem

30+ industrial and academic partners · Aix-Marseille, Nice, national and international collaborations

PARTNER COMPANIES



**AIX-MARSEILLE
AND NICE
ECOSYSTEM,
INCLUDING:
START-UPS
SMES
FRENCH TECH
TOP 20 CLUB**

Today's Focus: Three Astrophysics Projects

Collaboration with CPPM (Centre de Physique des Particules de Marseille, UMR 7346) - UHA

01

Planetary Nebula Detection

Object detection · YOLO · IPHAS & VPHAS+ surveys
Automated detection of PNe candidates in galactic surveys.
Published: Astronomy & Astrophysics (Dec. 2025)

02

Euclid / DISPERS

Signal separation · FFT + YOLO · Slitless spectroscopy
Spectral order detection directly from Euclid/NISP simulation images.
ANR project · CPPM RENOIR team

03

MockCraft — Generative AI

Foundation models · AION · Galaxy spectra & images
Synthetic source generation for incomplete astronomical catalogues.
ArXiv: 2510.17960

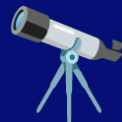
01

Project 1

Planetary Nebula Detection in Large Galactic Surveys

Identifying Planetary Nebulae Candidates from IPHAS and VPHAS+ Surveys via YOLO Deep Learning Models
AAS Astromical American Society

Team : Titouan SERANDOUR, Anne-Laure MEALIER, Bruno Colicchio, Dominique Daniel, Philippe Gilberton, Philippe Oros, Thomas Boch, Djaffar Ould Abdeslam
Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550 . Intitut de Recherche en Informatique, Mathématiques, Automatique et Signal (IRIMAS), Université de Haute-Alsace . Centrale Méditerranée



1 Identifying Planetary Nebulae Candidates from IPHAS and VPHAS+ Surveys
2 via YOLO Deep Learning Models
3 TITOUAN SERANDOUR ^{1,2} ANNE-LAURE MÉALIER,³ BRUNO COLICCHIO ¹ DOMINIQUE DANIEL,⁴ PHILIPPE GILBERTON,⁴ PHILIPPE OROS,⁴
4 THOMAS BOCH ² AND DJAFFAR OULD ABDESLAM ¹
5 AIP-PN
6
7 ¹Institut de Recherche en Informatique, Mathématiques, Automatique et Signal (IRIMAS), Université de Haute-Alsace, 61 Rue Albert Camus, 68093 Mulhouse,
8 France
9 ²Centrale Méditerranée, 38 Rue Frédéric Joliot-Curie, 13451 Marseille, France
10 ³Aix Marseille Univ, CNRS, Centrale Med, Institut Fresnel, Marseille, France
11 ⁴Astro Images Processing (AIP)
12 ⁵Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France
13
14 ABSTRACT
15 We present a deep learning detection method capable of effectively identifying planetary nebulae (PNe) on
16 IPHAS and VPHAS+ surveys. Our detection pipeline is based on the YOLO architecture, and is adaptable to
17 other datasets. Using the HiPS system, we pre-process and augment the r, i, and H α bands to produce two RGB
18 image datasets: (1) 640 $\times$ 640 px images at full resolution, and (2) 1 280 $\times$ 1 280 px images designed to capture
19 extended objects up to 0.2 deg. Images of $\sim$ 2 600 PNe are labelled using the HASH database. The loss function
20 and evaluation metrics are adapted to the diffuse and low-contrast nature of PNe; a very low IoU threshold (1%)
21 is used. A k-folds cross-validation shows that the YOLOv12m model achieves an average precision of 0.979
on the first dataset. At a confidence threshold of 0.15, it achieves a precision rate of 98.7% and a recall rate
of 97.4%. We deployed our system (on 1 280 $\times$ 1 280 images) **on the full surveys' coverage** ($\sim$ 3 755 deg²).

What are Planetary Nebulae?

End-of-life ionized gas shells produced by low- and intermediate-mass stars.

Short-lived (~20,000–30,000 years) and often extremely faint, detection in large Galactic surveys remains challenging.

Why they matter:

- Tracers of stellar evolution & Galactic chemical enrichment
- Useful kinematic and distance indicators
- Most of them are located in the Galactic plane, thousands of Galactic planetary nebulae likely remain undiscovered (a very crowded region containing huge numbers of stars, gas clouds, and instrumental artefacts)
- Strong science case for next-generation sky surveys and AI-based detection pipelines

IPHAS

INT : Isaac Newton Telescope (La Palma, Spain)
Photometric H α Survey of the Northern Galactic Plane
Northern Galactic plane · 3 bands (r, i, H α)
~1,800 deg²

VPHAS+

VST : VLT Survey Telescope (Chile) Photometric H α Survey of the Southern Galactic Plane and Bulge
Southern Galactic plane · 5 bands (u, g, r, i, H α)
~2,000 deg²

Challenge: Faint, diffuse, morphologically variable objects in crowded fields - at scale.

Scientific Context & Data Pipeline

How detect Planetary Nebulae?

PNe are shells of dying stars ($1-8 M_{\odot}$), lasting only $\sim 25,000$ years. Of $\sim 6,600-45,000$ estimated Galactic PNe, fewer than 4,000 are catalogued. They trace stellar evolution, chemical enrichment and serve as distance indicators. IPHAS + VPHAS+ cover $\sim 3,755 \text{ deg}^2$ of the Galactic plane.

NGC 2438 — r, i, H α bands (VPHAS+)

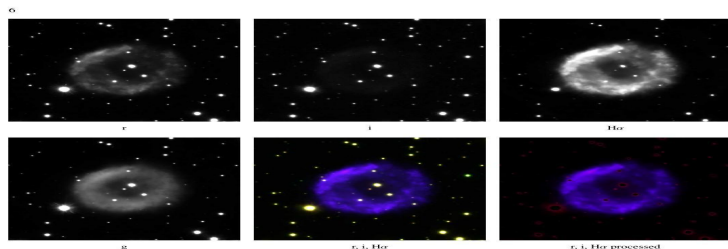


Figure 1. Small (640×640 px) images of NGC 2438 in various spectral bands (VPHAS+). The bottom right image is the r, i, H α image processed by applying the transformation $x \mapsto x \cdot (1 - z/255)$ to the r and H α bands, and $x \mapsto 0$ to the i band. This processing, which we do not use, is a mere illustration how the i band may be used to discriminate PNe signal versus background.

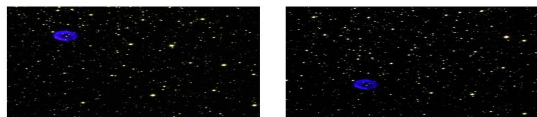


Figure 2. Large de-centered images of NGC 2438 ($1,280 \times 1,280$ px, VPHAS+).

1 HiPS Cut-outs

CDS HiPS2FITS · r, i, H α bands · 640×640 px ($211'' \times 211''$) and 1280×1280 px ($720'' \times 720''$)
Unified pipeline for both surveys

2 Pre-processing

Histogram stretching (200-quantiles) · Clipping (0.5%) · 8-bit RGB stacking
Random de-centering for data augmentation

3 Annotation (HASH)

$\sim 2,600$ known PNe labelled · Square bounding box from HASH (H-alpha Planetary Nabula database) diameter
Grouped k-fold split ($k=4$) stratified by PN status

4 YOLOv12m Training

20.2M parameters · COCO pretrained · AdamW · 160–200 epochs
A100 40 GB · IoU threshold = 0.01 (diffuse morphology)

YOLO - You Only Look Once

A deep learning architecture for object detection

The astronomer's analogy

Think of YOLO as a trainee who has studied thousands of images of planetary nebulae (or spectral orders). Given a new image, it scans it once and outputs: "there is something here, at this position, with this confidence." No iterative fitting, no template-matching : a single forward pass.

Input / Output

Input: an RGB astronomical image (r, i, H α for PNe, or original+FFT-0+FFT-1 for spectral orders)

Output: a list of bounding boxes, each with a class label (PN / 0th order / 1st order) and a confidence score [0–1]

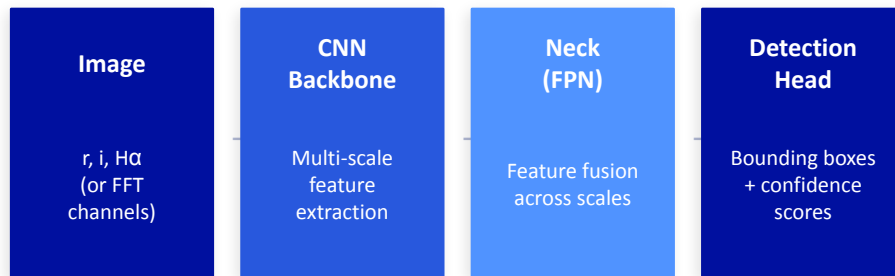
Nothing else: no spectra, no photometry, no prior source position needed

How it learns

1. Pre-training: starts from weights trained on COCO (millions of natural images : cats, cars, people). It already knows edges, textures, shapes.

2. Fine-tuning: we then train on our labelled astronomical images. The network adapts its filters to PN morphologies or spectral order shapes.

3. Inference: < 500 ms per image. No user interaction.



YOLOv12m: 20.2M parameters · pre-trained on COCO · fine-tuned on astronomical data

Why YOLO and not another architecture?

• vs Swin Transformer:

Sun et al. 2024 used Swin Transformer on VPHAS+. Our YOLOv12m achieves comparable or better P/R in a single unified pipeline covering both IPHAS and VPHAS+, with a much larger field of view. Our AP50=0.792 vs their AP50=0.944 — but our images are 12× larger in field of view.

• vs R-CNN / Mask R-CNN:

R-CNN is slower (two-stage: propose regions, then classify). YOLO is one-stage → 18–200 ms per image vs seconds for R-CNN. Critical when processing 164,858 images for the full galactic survey.

• vs classical SExtractor:

SExtractor detects sources by thresholding. It fails on diffuse, low-contrast PNe and heavily blended slitless spectra. YOLO learns morphological features statistically, without requiring a clean source profile.

Quantitative Results — YOLOv12m

0.985

AP01 (mean)
4-fold CV

98.7%

Precision
@conf 0.15

97.4%

Recall
@conf 0.15

98.0%

F1-score
mean k-fold

35,665

New candidates
full survey deploy

Precision : Among detected objects, how many are correct - **Recall** : Among real objects, how many were successfully detected

mAP (mean Average Precision) : Overall object detection performance metric combining precision and recall across confidence thresholds. - **F1-score** Global object detection performance over all confidence thresholds.

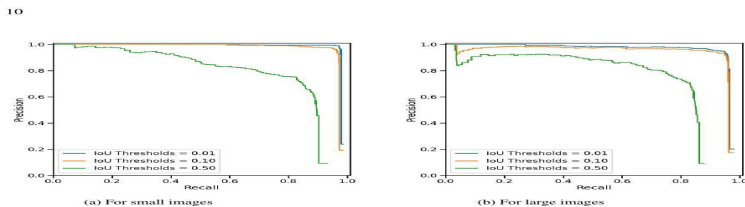


Figure 4. PR-curves for small and large image datasets. The confidence score threshold step is 0.01. For small images, AP01 = 0.979, AP10 = 0.970 and AP50 = 0.792. For large images, AP01 = 0.934 and AP50 = 0.745.

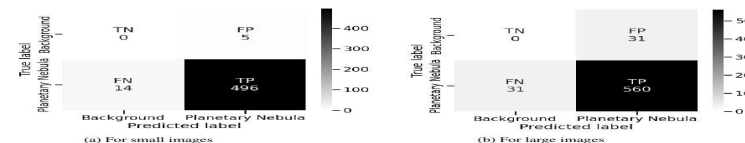


Figure 5. Confusion matrices for small and large image datasets.

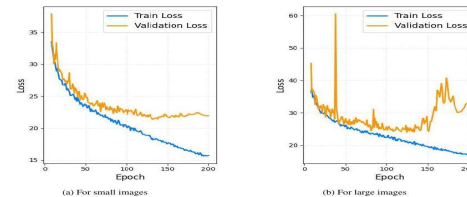


Figure 3. Loss functions graphs for small and large image datasets. These graphs were obtained during the training of YOLOv12 on the first cross-validation folds.

Table 3. YOLOv12m model k-fold cross validation results

($k = 4$) on the small image dataset. The confidence score threshold is set to 0.15.

Fold	AP01	Precision	Recall	F1-score
#1	0.979	99.0%	97.3%	98.1%
#2	0.996	99.6%	98.4%	99.0%
#3	0.990	98.0%	97.4%	97.7%
#4	0.974	98.0%	96.4%	97.2%
Mean	0.985	98.7%	97.4%	98.0%
Standard deviation	0.010	0.790	0.818	0.761

4.3. Discussion

Visual analysis of FP and FN.—A visual analysis of the 5 false positives found after training the model on the first cross-validation fold of the small image dataset (cf. Figure 5) shows that 2 of them are caused by CCD defects in IPHAS, having a morphology close to that of a compact PNe (compare Figure 6 and Figure 9(a)) and 1 is in fact targeting a true PN (called “Th 4-3”) with its apparent diameter value incorrectly set to 0 in the HASH database (thus leading to a zero IoU value). As a consequence, this object is also incorrectly considered as a false negative. Among the 14 false negatives, 2 are due to overlapping ground-truth bounding boxes, due to duplicates of PNe in the HASH database (indexed under the same “dPNMain” identifier).

Small objects detection performances.—Among the 70 ground truth boxes with a diagonal below 30 pixels in the first cross-validation fold of the small image dataset (cf. Figure 5), only 64 were detected by the model, yielding a recall rate of approximately 91.4%. In contrast, larger bounding boxes achieved a recall rate of approximately 98.2%, suggesting that the model exhibits reduced performance on smaller objects.

Label quality and low IoU thresholds choice.—Although the ground-truth bounding boxes used in this study were not produced by manual annotation, they are derived from the HASH PN catalogue, which represents the most comprehensive and widely used

Fig. 4 & 5: PR-curves (small & large images) and confusion matrices. Small images: AP01=0.979, AP10=0.970, AP50=0.792.

TP=496, FP=5, FN=14.

Table 3 & Fig. 3: k=4 cross-validation results + loss curves. Best fold:

AP01=0.996, P=99.6%, R=98.4%.

Detection Examples, Explainability & Deployment

Fig. 9 — Small image detections (test set, confidence thres. 0.15)

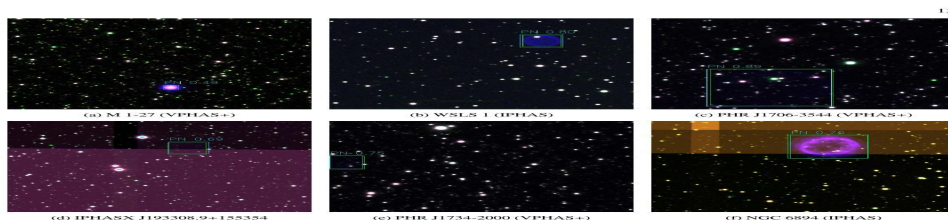


Figure 9. Small images selected from the test set

model’s predictions. Examining these activation maps shows that the model’s detection head focuses on meaningful PSN features across the input image, such as the presence of a person, a vehicle, or a building. This is evident in the visualization of the model’s internal state, which highlights the specific features that the model has learned to recognize. The model’s architecture includes a series of convolutional layers that extract features at various scales, allowing the model to detect objects of different sizes. The model’s output is a heatmap that shows the probability of each pixel belonging to a specific class. The model’s performance is evaluated using a variety of metrics, including precision, recall, and F1 score. The model’s performance is compared to other state-of-the-art models, and the results show that the proposed model achieves superior performance across all metrics. We used Shrivalkar’s implementation²², which is based on Ghidblat’s package (Ghidblat and Contributors 2023).

²²<https://github.com/ghidblat/psn-detection>

²³<https://github.com/ghidblat/psn-detection>

²⁴<https://github.com/ghidblat/psn-detection>

²⁵<https://github.com/ghidblat/psn-detection>

²⁶<https://github.com/ghidblat/psn-detection>

²⁷<https://github.com/ghidblat/psn-detection>

²⁸<https://github.com/ghidblat/psn-detection>

²⁹<https://github.com/ghidblat/psn-detection>

³⁰<https://github.com/ghidblat/psn-detection>

³¹<https://github.com/ghidblat/psn-detection>

³²<https://github.com/ghidblat/psn-detection>

³³<https://github.com/ghidblat/psn-detection>

³⁴<https://github.com/ghidblat/psn-detection>

³⁵<https://github.com/ghidblat/psn-detection>

³⁶<https://github.com/ghidblat/psn-detection>

³⁷<https://github.com/ghidblat/psn-detection>

³⁸<https://github.com/ghidblat/psn-detection>

³⁹<https://github.com/ghidblat/psn-detection>

⁴⁰<https://github.com/ghidblat/psn-detection>

⁴¹<https://github.com/ghidblat/psn-detection>

⁴²<https://github.com/ghidblat/psn-detection>

⁴³<https://github.com/ghidblat/psn-detection>

⁴⁴<https://github.com/ghidblat/psn-detection>

⁴⁵<https://github.com/ghidblat/psn-detection>

⁴⁶<https://github.com/ghidblat/psn-detection>

⁴⁷<https://github.com/ghidblat/psn-detection>

⁴⁸<https://github.com/ghidblat/psn-detection>

⁴⁹<https://github.com/ghidblat/psn-detection>

⁵⁰<https://github.com/ghidblat/psn-detection>

⁵¹<https://github.com/ghidblat/psn-detection>

⁵²<https://github.com/ghidblat/psn-detection>

⁵³<https://github.com/ghidblat/psn-detection>

⁵⁴<https://github.com/ghidblat/psn-detection>

⁵⁵<https://github.com/ghidblat/psn-detection>

⁵⁶<https://github.com/ghidblat/psn-detection>

⁵⁷<https://github.com/ghidblat/psn-detection>

⁵⁸<https://github.com/ghidblat/psn-detection>

⁵⁹<https://github.com/ghidblat/psn-detection>

⁶⁰<https://github.com/ghidblat/psn-detection>

⁶¹<https://github.com/ghidblat/psn-detection>

⁶²<https://github.com/ghidblat/psn-detection>

⁶³<https://github.com/ghidblat/psn-detection>

⁶⁴<https://github.com/ghidblat/psn-detection>

⁶⁵<https://github.com/ghidblat/psn-detection>

⁶⁶<https://github.com/ghidblat/psn-detection>

⁶⁷<https://github.com/ghidblat/psn-detection>

⁶⁸<https://github.com/ghidblat/psn-detection>

⁶⁹<https://github.com/ghidblat/psn-detection>

⁷⁰<https://github.com/ghidblat/psn-detection>

⁷¹<https://github.com/ghidblat/psn-detection>

⁷²<https://github.com/ghidblat/psn-detection>

⁷³<https://github.com/ghidblat/psn-detection>

⁷⁴<https://github.com/ghidblat/psn-detection>

⁷⁵<https://github.com/ghidblat/psn-detection>

⁷⁶<https://github.com/ghidblat/psn-detection>

⁷⁷<https://github.com/ghidblat/psn-detection>

⁷⁸<https://github.com/ghidblat/psn-detection>

⁷⁹<https://github.com/ghidblat/psn-detection>

⁸⁰<https://github.com/ghidblat/psn-detection>

⁸¹<https://github.com/ghidblat/psn-detection>

⁸²<https://github.com/ghidblat/psn-detection>

⁸³<https://github.com/ghidblat/psn-detection>

⁸⁴<https://github.com/ghidblat/psn-detection>

⁸⁵<https://github.com/ghidblat/psn-detection>

⁸⁶<https://github.com/ghidblat/psn-detection>

⁸⁷<https://github.com/ghidblat/psn-detection>

⁸⁸<https://github.com/ghidblat/psn-detection>

⁸⁹<https://github.com/ghidblat/psn-detection>

⁹⁰<https://github.com/ghidblat/psn-detection>

⁹¹<https://github.com/ghidblat/psn-detection>

⁹²<https://github.com/ghidblat/psn-detection>

⁹³<https://github.com/ghidblat/psn-detection>

⁹⁴<https://github.com/ghidblat/psn-detection>

⁹⁵<https://github.com/ghidblat/psn-detection>

⁹⁶<https://github.com/ghidblat/psn-detection>

⁹⁷<https://github.com/ghidblat/psn-detection>

⁹⁸<https://github.com/ghidblat/psn-detection>

⁹⁹<https://github.com/ghidblat/psn-detection>

¹⁰⁰<https://github.com/ghidblat/psn-detection>

¹⁰¹<https://github.com/ghidblat/psn-detection>

¹⁰²<https://github.com/ghidblat/psn-detection>

¹⁰³<https://github.com/ghidblat/psn-detection>

¹⁰⁴

¹² URL: <https://github.com/rigvedns/YOLO-V12-CAM>

M1-27 · WSLS1 · PHR J1706-3544 · IPHASX J193308.9 · PHR J1734-2000 · NGC6894

Large-scale deployment: $\sim 3,755 \text{ deg}^2$

164,858 images generated on CDS servers · 49,648 raw detections (conf ≥ 0.15)

After deduplication and cross-match with HASH (20" radius): **35,665 new PN candidates**

Follow-up spectroscopy campaign underway by AIP amateur astronomers

Fig. 11 — Eigen-CAM explainability (layers 4, 6, 19)

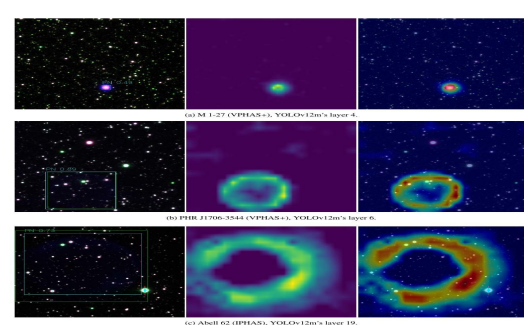


Figure 11. Original 640 × 640 px image (left), Eigen-CAM activation map (center) and their superposition (right)

trade-off. It delivers the highest F1-score (98.1%) while maintaining a reasonable parameter count (20.2M), avoiding the computational overhead of larger models without sacrificing detection quality. Given the marginal gains from scaling beyond medium size, we selected YOLOv12m for subsequent experiments. Avoiding the computational cost of larger

Activation maps confirm the model focuses on PN envelope, shell & boundary features.



Figure 2. On the left, Schenker 29 pictured by IPHAS (with H&R#9769 ellipses). Standard exposure times are 30 s (C), 10 s (B) and 120 s (H α). On the right, Schenker 29 pictured by HST. Data were acquired with a CFHT 300 mm F8.9 spectroscopic refractor equipped with a Polaris II camera and a QSI-CCD and PRISM slit-and-camera system. The dataset consists of three-filter frames (C, B, and H α), combined in H α -C-B and C-H α -B color images. Exposure times are 120 s (C), 120 s (B) and 110 s (H α) broad-band filters (H, G, B), for a total integration time of 11 h 32 min. Observations were conducted from a barycentric site.

Common PNe mimics derived by the model.—Predictions examinations are caused by a variety of PNe mimics and tangential artifacts including in particular some of the CCD images (cf. Figure 9) optical images (glow images) and most frequently the presence of extended F H regions leading to clusters of neighboring false positives, as shown in Figure 13.

Comparison with the literature.—In these work, Sun et al. (2004) deployed their Swin Transformer model on 43% of the VPHAS+ images. In total, they identified 10,000 PNe candidates. In this work, we identified 20,000 PNe candidates, among which 31 were confirmed by the literature. The difference between the two studies is due to the different analysis of their reduced spectra verified that 32 of these objects can be classified as True, Likely, or Possible PNe (Li et al. 2024).

[illegible]

Comparison with Literature & Perspectives

Method	AP	Precision	Recall	Coverage	IoU thresh.
Sun et al. 2024 (Swin Transformer)	AP50=0.944	96.5%	97.8%	979 deg ² (43% VPHAS+)	0.50
Ours — YOLOv12m IPHAS+VPHAS+	AP01=0.985 AP50=0.792	98.7%	97.4%	~3,755 deg ² (full surveys)	0.01 (deliberate)

Note: AP50 not directly comparable — our images have a much larger field of view. IoU=0.01 is a deliberate scientific choice: PNe are diffuse and their bounding boxes in HASH are approximate.

Current Limitations

- PNe mimics (H II regions) still generate false positives with high confidence
- Small objects (< 30 px diagonal): recall drops to 91.4% vs 98.2% for larger
- Precision drops on large-scale deployment due to rareness of PNe in the field
- Manual vetting required for spectroscopic follow-up selection

Future Work

- Multiclass detector: add H II regions, SNRs, CCD artefacts as additional classes
- Include g band for [OIII] line (VPHAS+ only)
- Cross-match candidates with Gaia white dwarf catalogue
- Develop automated bounding-box refinement pipeline
- Expand coverage + spectroscopic confirmation campaign (AIP + professionals)

02

Project 2

Spectral Order Detection in Slitless Spectroscopy for the Euclid Mission

YOLOv11 · FFT preprocessing · Computer Vision · HST WISP + Euclid SPV · DISPERS ANR · CPPM Marseille

Team: Prachi Kansal, Sheikh Obaid Ullah, Titouan Sérandour, Hadi Sqalli
Supervisor : Julien ZOUBIAN; Anne-Laure MEALIER



Euclid & Slitless Spectroscopy

The Euclid Mission (ESA, launched July 2023)

Maps large-scale structure of the Universe to constrain dark energy and dark matter across 10 billion years of cosmic history.

Slitless Spectroscopy (NISP grism)

Light from all sources is dispersed simultaneously, creating overlapping spectra that are hard to deblend.

DISPERS ANR Project — CPPM RENOIR team

Deep learning pipeline to detect spectral orders directly from slitless images, without relying on photometric input.

Datasets: HST WISP (development/robustness) + Euclid SPV simulation (full noise model: stray light, cosmic rays, persistence)



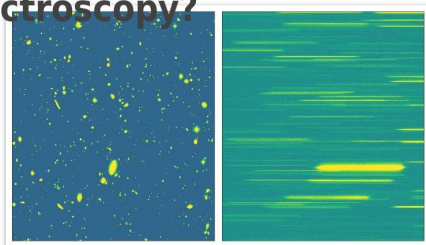
@Wikipedia artist's view of the Euclid telescope

Our aim:

Automate spectral order detection & classification directly from spectroscopic images — close to no reliance on photometric images. Code & data will be publicly released.

Detection using Slitless Spectroscopy?

But, what is slitless spectroscopy?

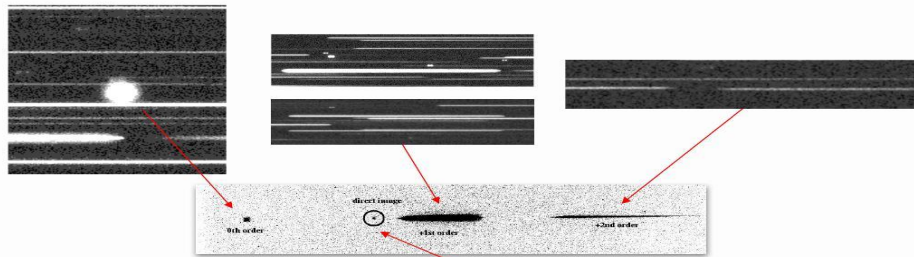


Photometric image

Spectroscopic image
(slitless)

Photometric image (left) vs slitless spectroscopic image (right).
Every source is dispersed into a spectrum, all overlapping.

More about spectroscopic images



Superimposition of the direct image of the star

The grism disperses each star into spectral orders: 0th (direct star image), 1st order (spectrum), 2nd order (fainter). They all overlap in the same image.

Why detection is hard

All spectra are dispersed simultaneously into the same image : there is no slit selecting a single source. Orders from different sources overlap, creating an extremely crowded image with no clean separation between signal and background.

What we want to detect

Given a spectroscopic image, automatically find and classify spectral orders (0th and 1st) using their box position and shape, with NO help from the photometric image. Current pipelines all require the photometric image as input.

The magnitude challenge

Brighter stars (low magnitude) produce bright, easy-to-detect spectral lines. Faint stars (high magnitude) produce lines that are barely above the noise floor. Setting the magnitude cut threshold determines how many sources can be detected.

Datasets & Why FFT Preprocessing?

Dataset 1 — HST WISP (Hubble)

387 images · 1845×2085 px FITS files

Very noisy: harder than expected Euclid data → good for building robustness

Used to develop and validate the full preprocessing + detection pipeline before Euclid data became available

Dataset 2 — Euclid SPV Simulation

388 simulation images · 2040×2040 px

Magnitude cut = 18 (bright enough to be detectable)

4 noise levels: stars only → + stray light → + persistence → + cosmic rays



Euclid SPV: stars only (left) vs full noise model including stray light, persistence and cosmic rays (right).

Why do we need FFT preprocessing?

✗ Problem: spectral orders overlap in spatial domain

In a raw slitless image, the 0th order (direct star image) sits right on top of the 1st order spectrum. A spatial detector like YOLO cannot separate them — they look like one big smear.

📡 Solution: frequency domain separation

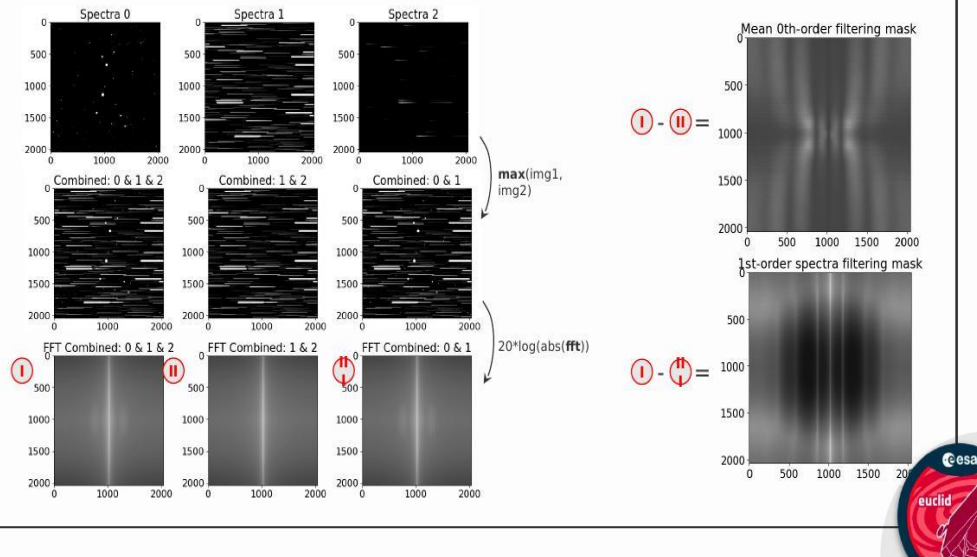
Spectral orders produce distinct horizontal periodic patterns in the image. FFT transforms the image into frequency space, where each order occupies a different frequency band. A targeted mask isolates the desired order.

✓ Result: two clean filtered images

After iFFT, we get one image highlighting only the 0th order and one highlighting only the 1st order. YOLO is then fed all 3 channels (original + FFT-0 + FFT-1) → richer input = better localisation.

The FFT Pipeline - Step by Step

Fast Fourier Transform Pipeline



Step-by-step FFT masking: raw images of Spectra 0/1/2 → combined via $\max()$ → FFT → masked → iFFT → separate 0th-order and 1st-order filtered images.

Combine images by $\max()$

①

Three spectral image variants (Spectra 0, 1, 2) are combined using a pixel-wise maximum operation. This keeps the strongest signal and ensures no order is lost.

FFT → magnitude spectrum

②

The 2D FFT transforms the combined image into frequency space. The magnitude spectrum ($20 \cdot \log |FFT|$) reveals horizontal frequency bands corresponding to each spectral order.

Apply selective frequency mask

③

A pre-computed mean mask targeting the 0th-order frequency band (or 1st-order) is multiplied element-wise with the FFT spectrum. This zeroes out all other frequency components.

Inverse FFT → filtered image

④

The iFFT converts the masked spectrum back to the spatial domain. The result is a clean image showing only the selected spectral order, ready for YOLO detection.

Key insight: running FFT with a 0th-order mask and a 1st-order mask gives two separate filtered images. YOLO takes a 3-channel input: original + FFT-0 + FFT-1 → significantly improves detection accuracy.

Results — HST WISP Dataset

YOLOv11 Medium · K-fold cross-validation ($k=10$) · Magnitude cut exploration

TPR = 0.80

Recall
Filtered, Mag 22

FNR = 0.20

Miss rate
Filtered, Mag 22

mAP@0.5
= 0.839
Best model
Filtered, Mag 22

F1 = 0.80

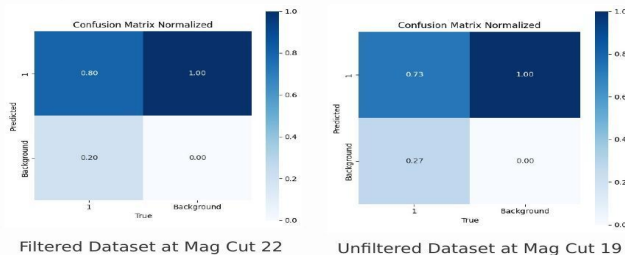
@conf 0.346
Filtered

mAP@0.5
= 0.758
Unfiltered
Mag cut 19

Confusion matrices — Filtered (Mag 22) vs Unfiltered (Mag 19)

Confusion Matrix

- Visualizes performance for true positives, false positives, true negatives, and false negatives.
- Higher true positives = better detection.



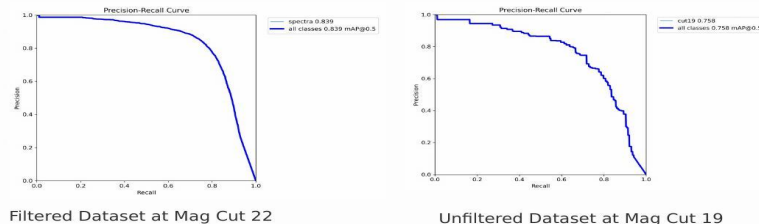
42

Filtered: TPR=0.80 (80% of spectra correctly detected). Unfiltered: TPR=0.73. FNR=false negative rate (missed detections).

Precision-Recall curves

Precision-Recall (PR) Curve

- Evaluates the trade-off between precision and recall across thresholds.
- Curve shape shows model's ability to balance true positives and false positives.



43

Filtered: mAP@0.5=0.839. Unfiltered: mAP@0.5=0.758. FFT filtering extends the magnitude cut ceiling from 19 to 22, making fainter spectral lines detectable.

Reading the matrix: the diagonal gives the True Positive Rate (Recall). Off-diagonal = missed detections (FN). The goal is a dark diagonal → high recall. Precision = of all detections made, how many are correct (not shown here but captured by the PR curve).

Results — Euclid SPV Simulation

YOLOv11 · Sim Stars (clean) and Sim Stars + Stray + Persistence + Cosmic Rays (full noise) · Mag cut = 18

Recall

94%

Class 0 (0th order)
Clean (Sim Stars)

Recall

88%

Class 1 (1st order)
Clean (Sim Stars)

Recall

92%

Class 0 (0th order)
Full noise

Recall

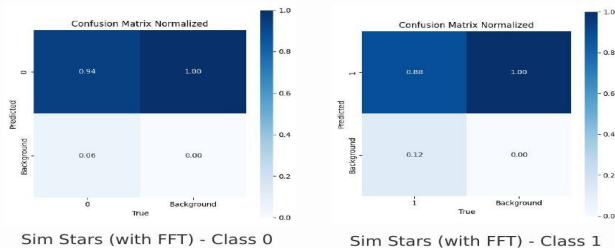
85%

Class 1 (1st order)
Full noise

Clean images (Sim Stars + FFT) — Class 0 & Class 1

Confusion Matrix

- Visualizes performance for true positives, false positives, true negatives, and false negatives.
- Higher true positives = better detection.

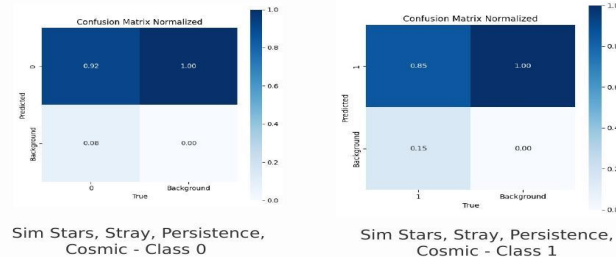


32

Noisy images (Stars + Stray + Cosmic + Persistence) — Class 0 & 1

Confusion Matrix

- Visualizes performance for true positives, false positives, true negatives, and false negatives.
- Higher true positives = better detection.



33

Class 0: TPR=0.94 (6% missed). Class 1: TPR=0.88 (12% missed). Clean images = stars dispersed only, no additional noise sources.

Class 0: TPR=0.92 (8% missed). Class 1: TPR=0.85 (15% missed). Adding stray light, persistence and cosmic rays only slightly reduces performance — showing robustness.

Key findings: (1) Clean → Noisy: performance drops only ~2–3% — the model is robust to Euclid noise. (2) Noisy-trained model generalises to clean images, but NOT the reverse → always train on the hardest data. (3) Individual class models (one per order) outperform a single combined model.

Model Comparison, Outcomes & Future Directions

Model	Order	Recall (TPR)	Precision	IoU@0.5	Inference
Computer Vision	Class 0	9.6%	—	0.5	190 ms
Computer Vision	Class 1	42.5%	—	0.5	300 ms
YOLOv11	Class 0	94% (clean) 92% (noisy)	high	0.5	402 ms
YOLOv11	Class 1	88% (clean) 85% (noisy)	high	0.5	455 ms

YOLO dramatically outperforms Computer Vision for Class 0 (94% vs 9.6% recall). CV is interpretable and fast but fails on faint spectra (thresholding treats them as background).

Scientific Outcomes

- First prototype to detect spectral orders directly from slitless images without photometric input
- Multi-channel FFT input (original + FFT-0 + FFT-1) substantially improves YOLO localisation
- Noisy training generalises to clean images, train on hardest data
- YOLOv11 recall 92–94% for 0th order and 85–88% for 1st order under full Euclid noise

Open Science

- Dataset, labels, code: public GitHub release
- Web app developed for interactive demonstration
- Full MLflow + DVC experiment tracking
- Directly integrated into DISPERS ANR project at CPPM

Future Directions

- Larger/synthetic dataset (only 387 HST + 388 Euclid images used)
- Advanced architectures: EfficientDet, Masked R-CNN, ViT
- Finer segmentation + hybrid DL+CV for faint orders
- Order 2 detection (first attempts promising with mag cut 16)

Scientific & Computational Challenges

Why applying deep learning to these two astrophysical problems is fundamentally hard — beyond standard computer vision

Shared challenges — the signal itself is the problem



No clean ground truth

PNe apparent diameters in HASH come from heterogeneous programs with different instruments, filters and definitions. Bounding boxes are approximate by construction. For slitless spectra, "ground truth" is a simulation — real data will always differ.



Extreme class imbalance

PNe are rare: fewer than 1 per 10 deg² on average. Most images are background. For slitless spectra, bright orders are easy but faint orders (high magnitude) are barely above noise. Standard loss functions fail — we had to re-weight them.



Non-standard IoU metrics

In natural images, IoU > 0.5 is the norm. Here we use IoU = 0.01 for PNe (diffuse, ill-defined edges) and IoU = 0.5 for spectral orders. These choices fundamentally change what the model optimises for — and make fair comparison with the literature harder.



Distribution shift

Models trained on survey data (IPHAS/VPHAS+) will face new noise patterns, instruments, and angular resolutions when applied to Rubin LSST or other future surveys. Robustness to distribution shift is an open problem.

Project-specific challenges

Project 1 — Planetary Nebulae

- **Morphological diversity:** round, elliptical, bipolar, point-source — a single model must cover all types without pre-defined criteria
- **Mimics:** compact H II regions, supernova remnants and CCD artefacts produce false positives indistinguishable photometrically from PNe — even for humans
- **Scale:** 164,858 images over 3,755 deg². The model must remain calibrated (not overconfident) across the full galactic density gradient

Project 2 — Euclid / DISPERS

- **Blending:** spectral orders from different sources overlap in the same image — there is no unique deblending solution without a prior source catalogue
- **Physics-informed preprocessing needed:** FFT is motivated by the known periodic structure of spectral orders — but choosing the right mask is non-trivial and survey-dependent
- **Simulation-to-reality gap:** only simulated data available so far. Performance on real Euclid data is unknown — and noise models may miss edge cases (saturation, transients, ghosts)

03

Project 3

MockCraft - Generative AI for Astronomical Catalogues



Foundation models · AION · Multimodal generation · Galaxy spectra & images · CPPM × Centrale Méditerranée

MockCraft

Generative AI for Astronomical Catalogues

Team

Bibiana STUMPF / João Victor MOTTA / Pedro LUNA

Supervision

Anne-Laure MEALIER / Julien ZOUBIAN



Centrale 
Méditerranée

What is MockCraft?

A generative python framework for producing realistic synthetic astronomical sources.

Catalogue limitations

- Astronomical catalogues contain rich information about galaxies and stars.
- Many sources still lack key modalities, such as spectra, images, photometry or physical properties.
- Existing generative models can help fill these gaps, but they are often complex and impractical to use directly.



MockCraft approach

MockCraft simplifies this process by using a generative AI model trained on large samples of **galaxies** and **stars**.

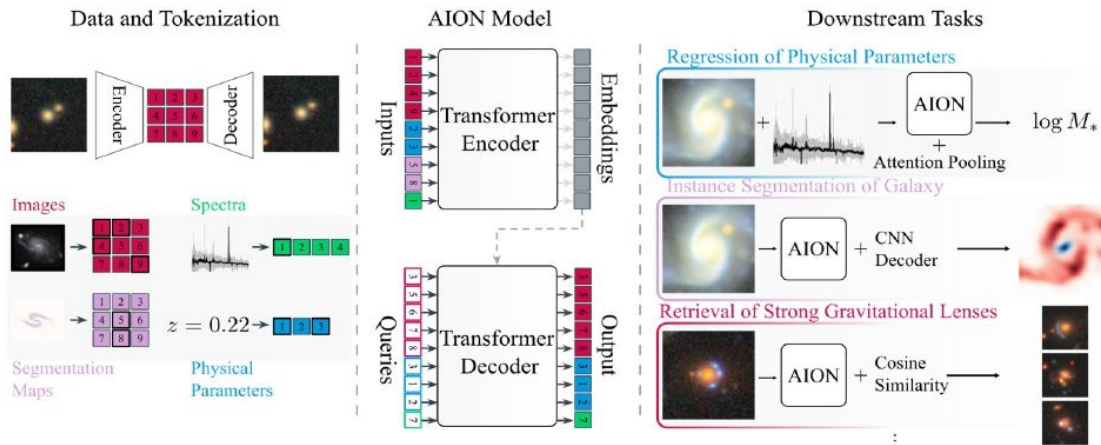
From the available inputs in a catalogue, the framework predicts missing properties and generates consistent synthetic sources for building realistic **mock catalogues**.



AION: The Generation Model



QR code
for the paper



Parker, Lanusse, Shen et al., arXiv:2510.17960

- **39 modalities** — images, spectra, photometry, and physical parameters from major surveys (Gaia, DESI, Legacy Survey, HSC).
- **Multimodal masked modeling** — a single transformer encoder-decoder trained with a masked-token objective across all modalities simultaneously.

- **Universal tokenization** — heterogeneous observations are discretized into a common token vocabulary before entering the model.
- **Cross-modal generation** — given any subset of observables, AION can generate any other; MockCraft uses this as its generative backbone.



Datasets & Modalities

Data Sources

Queried live via `fetch_sources()` — VizieR (astro-query) + Legacy Survey API

Gaia DR3

Positions, G/BP/RP fluxes, parallax — catalog
I/355/gaiadr3

DESI DR1

Spectroscopic redshifts, g/r/z fluxes — catalog
V/161/zcatdr1

Legacy Survey DR8 / DR10

Photo-z, morphological type — catalog VII/292
Image via Cutout API

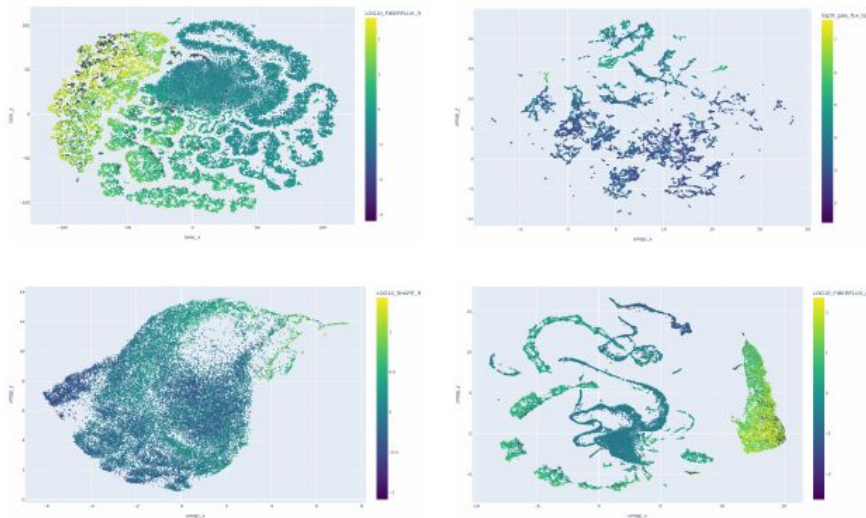
Main Modalities Used

Primary modalities worked on in this project

TYPE	MODALITIES
Image	Legacy Survey cutouts (g, r, z)
Spectrum	DESI optical, Gaia XP (BP + RP)
Photometry	Legacy Survey g/r/i/z, Gaia G/BP/RP
Scalar	Redshift z, morphology r/e1/e2



Latent Space



UMAP and t-SNE projections coloured by physical properties from Gaia, DESI and Legacy Survey metadata.

Validation strategy

Thousands of multimodal generation cases were run using Gaia, Legacy Survey and DESI datasets available on Hugging Face.

The objective was to study how AION positions galaxies and stars in the latent embedding space after generation.

Spearman correlation was used to compare pairwise distances in the original latent. Projections were coloured by physical properties available in the datasets, making it possible to identify which quantities are well organized by the learned embeddings.

All latent-space representations are exported as interactive HTML files, allowing the selected physical-property colouring to be changed and evaluated dynamically.



What is a Latent Space? UMAP & t-SNE

What is a latent space?

AION encodes every source : star or galaxy into a **high-dimensional vector** (the "embedding") that summarises all its observed properties: fluxes, spectrum shape, morphology, redshift.

Think of it as a compact description of a source, the way a stellar spectrum is a compact description of its atmosphere — but here it is learned automatically by the neural network, not prescribed by physics.

UMAP vs t-SNE — which to use and why

Property	UMAP	t-SNE
Global structure	Preserved — clusters and inter-cluster distances are meaningful	⚠ Distorted — only local neighbourhoods are reliable
Local structure	Good	Excellent, very sharp cluster boundaries
Speed	Faster on large datasets (millions of sources)	✗ Slow on large N
Stability	Deterministic with fixed seed	⚠ Sensitive to perplexity parameter
Use in MockCraft	Map large multi-survey samples (Gaia + DESI + Legacy)	Zoom into local structure: redshift gradient, star/galaxy separation

The dimensionality problem

AION's latent vectors have **hundreds of dimensions**, impossible to visualise directly.

UMAP (Uniform Manifold Approximation) and **t-SNE** (t-distributed Stochastic Neighbor Embedding) are two standard algorithms that project these embeddings down to 2D while **preserving the neighbourhood structure**: sources that are similar in the high-dimensional space land close together in the 2D map.

Schematic: what a UMAP of AION embeddings looks like



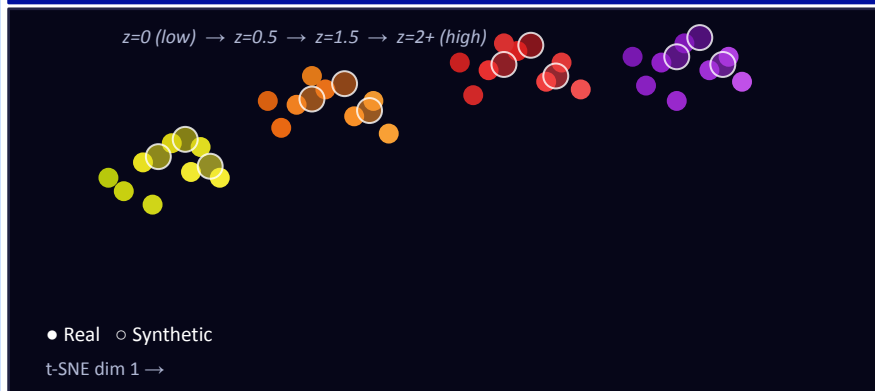
Latent Space Evaluation - Galaxies & Stars Separately

Does AION learn physically meaningful representations and are real and synthetic sources indistinguishable?

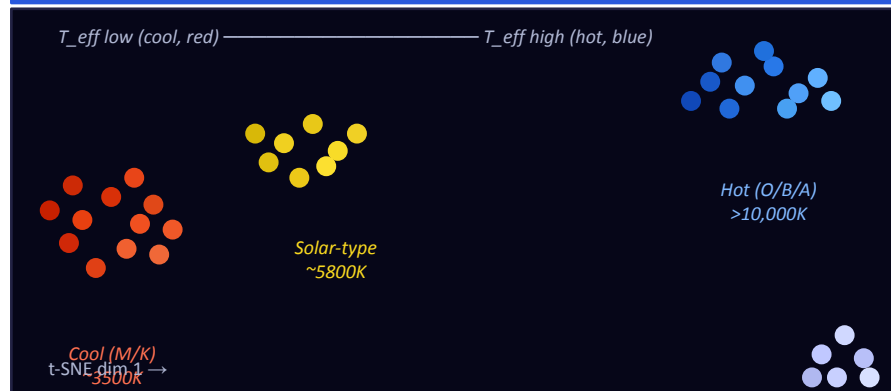
Why evaluate galaxies and stars separately?

Stars and galaxies are physically completely different objects: different spectra shapes, different physical parameters, different redshift distributions (stars have $z \approx 0$, galaxies span $z=0-3+$). Mixing them in the same validation would hide failures. A model that is excellent for galaxies and terrible for stars would still look good in aggregate.

Galaxies — t-SNE & UMAP (Gaia DR3 + Legacy Survey + DESI DR1)



Stars — t-SNE & UMAP (Gaia DR3 XP spectra + photometry)



Quantitative validation — applied identically to galaxies and stars

Cosine similarity (spectra)

Synthetic $\approx$ real $\times$ real
Median ~ 0.92 for both populations

Spearman rank correlation

Pairwise latent distances correlated between real-real and real-synthetic pairs $\rightarrow$ geometry preserved

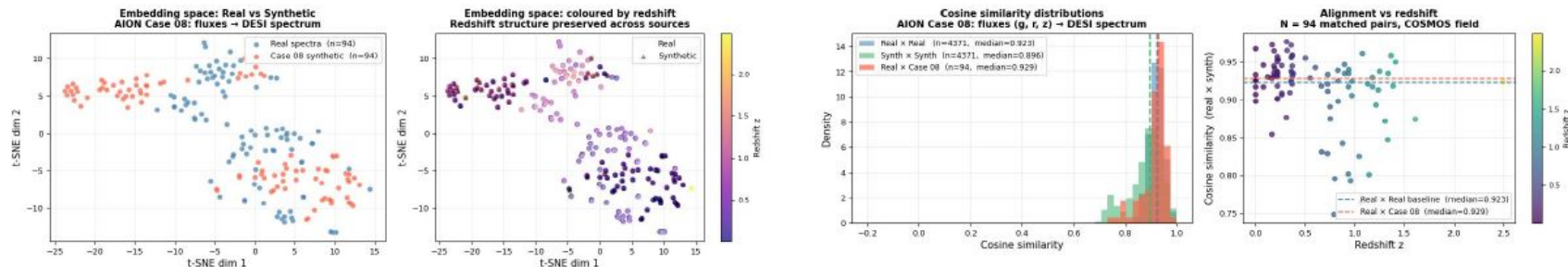
t-SNE structure by redshift

Galaxy manifold organises smoothly along z , a physical gradient the model was not explicitly trained to reproduce

Interactive HTML exports

All 2D projections exported as interactive files — any physical property can be chosen as the colour axis dynamically

Are Generated Sources Physically Realistic?



Latent-space structure

Real and synthetic sources occupy comparable regions.

Similarity and physical alignment

Synthetic spectra show cosine similarities close to the real-real.

- **Stars and galaxies evaluated separately**
- **Cosine similarity:** synthetic $\approx$ real $\times$ real distribution, with median ~ 0.92
- **t-SNE structure preserved** and organized by redshift



How Does It Work?

MockCraft API

```
from mockcraft import SourceGenerator, fetch_sources

# 1. Fetch real catalogue sources
sources = fetch_sources(surveys=["legacy"], region="cosmos", columns=["ra", "dec", "redshift", "flux_g", "flux_r", "flux_z"],
max_sources=100)

# 2. Initialize the AION-based generator
gen = SourceGenerator(model="aion", seed=42)

# 3. Generate a synthetic multimodal source
mock = gen.generate(
    inputs={"flux_g": 27.2, "flux_r": 54.5, "flux_z": 90.7},
    outputs=["spectrum", "image", "redshift"],)
```

Output

```
[SourceGenerator] Loading polymathic-ai/aion-base on cuda ... Ready.
```

```
GeneratedSource(spectrum.shape=(8704,), image.shape=(3, 96, 96), redshift=4.117302)
sources.head() -> ra=150.112282, dec=2.005249, z=0.165309, flux_g/r/z=27.1536/54.5284/90.6526
```



AI × Astrophysics

Data Explosion

Euclid, Rubin LSST, SKA will generate petabytes annually. Classical pipelines cannot scale — AI is not optional, it is essential.

Complex Pattern Recognition

From faint PNe to overlapping spectral orders, astrophysical signals are challenging. Deep learning exceeds human performance at scale.

Missing Data & Simulation

Catalogues are systematically incomplete. Generative models can fill gaps, augment datasets and support simulation pipelines.

Collaborative Framework

CentraleDigitalLab brings AI engineering to astrophysics. Students contribute modern, reproducible AI pipelines while researchers provide scientific context.

Conclusion & Perspectives

Three projects, three paradigms:

Object detection (PNe/YOLO) · Signal separation (Euclid/DISPERS) · Generative modelling (MockCraft)

A reproducible, collaborative model:

CentraleDigitalLab bridges AI engineering and scientific research, accelerating adoption of state-of-the-art methods in astrophysics.

Open science:

Code, datasets and models are released publicly to maximise reuse by the astrophysics community.

Towards the AI-native observatory:

Foundation models and generative AI will become core infrastructure for simulation, calibration and discovery in future surveys.