Class-Conditioned Consensus Loss for Multi-Expert Annotation of Low Surface Brightness structures

Renaud Vancoellie[1], Adeline Paiement[1], Pierre Alain Duc[2]

[1] Laboratoire d'Informatique et des Système, Université de Toulon
[2] Observatoire de Strasbourg, Université de Strasbourg

June 22th, 2026

Tidal features - Why them

- **Fossil record of mergers**

Produced by gravitational stripping during galaxy interactions.

- **Reconstructing merger histories**

Essential tracers galaxies assembly over time.

- **Testing ΛCDM**

Test galaxy interaction rates predictions.

Fig.1: *NGC0474, MATLAS Survey, true color*

Tidal features - Detection challenge

- **Below the noise floor**

Surface brightness μ_r ~ 26–30 mag/arcsec², at or under the sky background of most surveys.

- **No physical edge**

Fade continuously into the sky — perceptual decision.

- **Survey scale**

Expert visual inspection doesn't scales to modern wide-field surveys.

Fig.2: *NGC0474, MATLAS Survey*

MATLAS -

The Survey:
~500 dual-band (g,r) images
~29 mag/arcsec² limiting SB

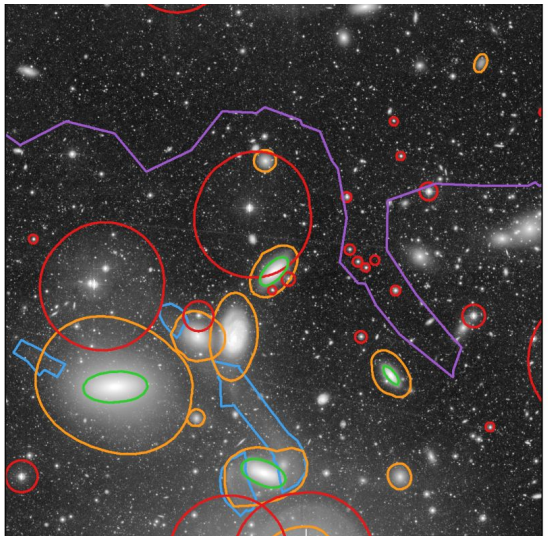


Fig.3: Annotation example

THE ANNOTATOR POOL:
Variable size K — each image is annotated independently by a different number of raters, with no communication between them.

Expert astronomers
Domain specialists (2)

Non-experts
Trained/PhD (2)

Five morphological classes:

Tidal Structures — primary target — streams, tails, plumes ~150
Ghosted Halos — optical reflection artifacts ~1 800
Galaxy — main stellar body ~400
Inner Structures — compact high-SB, e.g. nucleus ~300
High Background — cirrus, dust, sky residuals ~200

Tidal features - Inter annotator disagreement

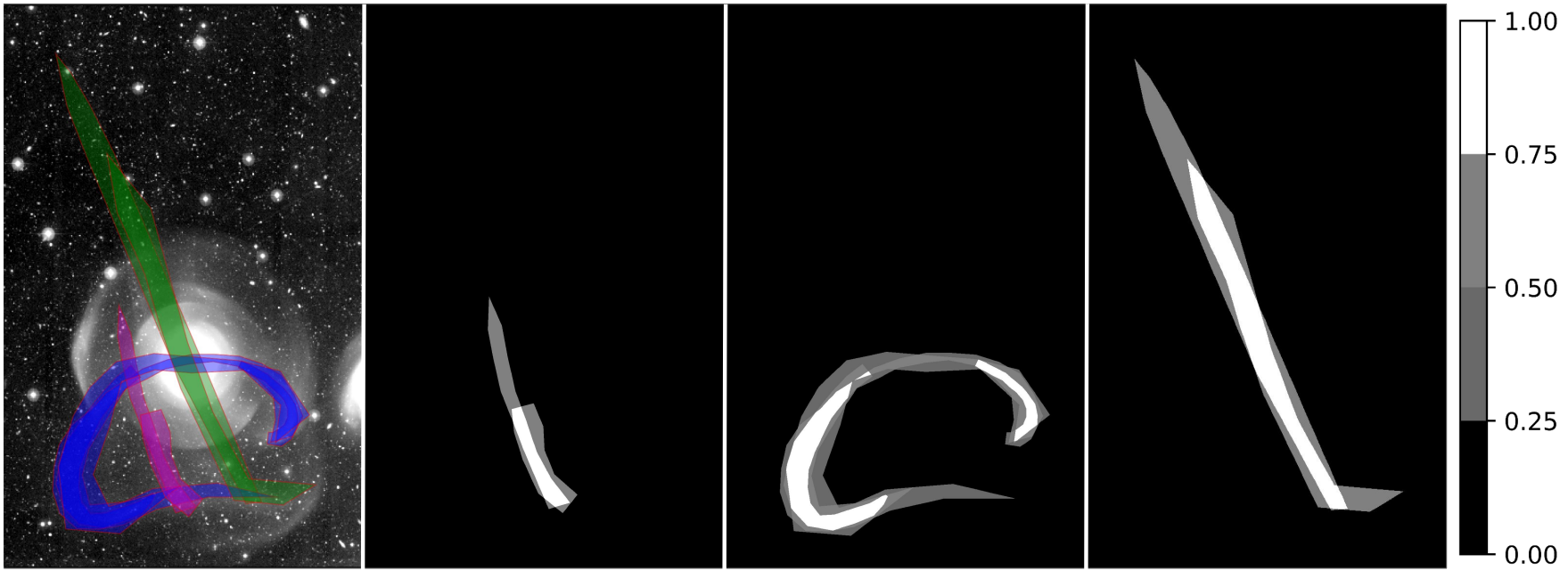


Fig.4: *NGC0474 image and the three identified structures.*

Tidal features - Inter annotator disagreement

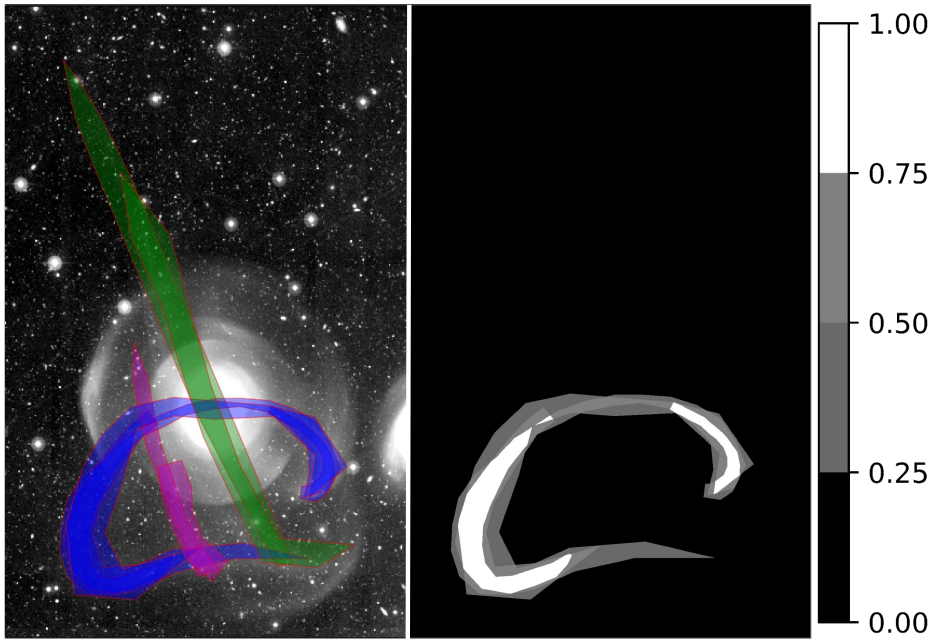


Fig.4: *NGC0474 image and the three identified structures.*

Structured, not random

Disagreement is spatially organised and concentrated on the most ambiguous regions of the image.

Astronomers rarely agree on where a tidal feature starts/ends.

How is this problem address in other field ?

Medical solution - Silva et Oliveira 2021

- **Uniform soft average**

Annotator masks are averaged with equal weights into a single soft label.

- **Cross-entropy**

Standard binary cross-entropy on every pixel.

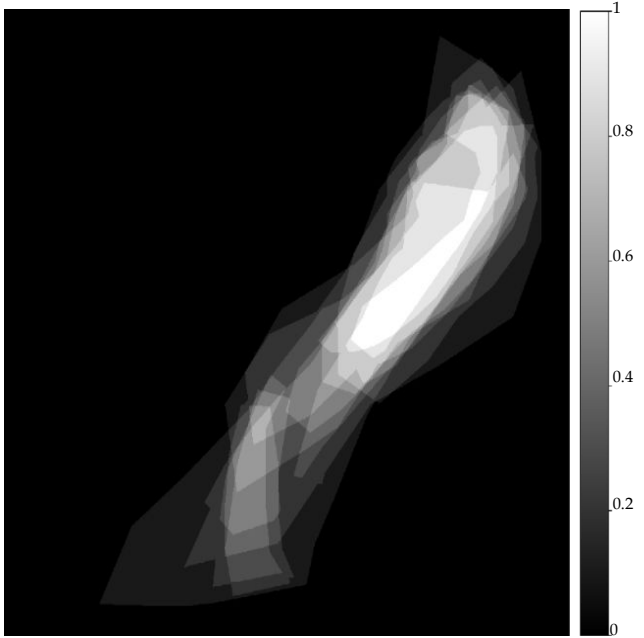


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

how far can plain soft labels alone take you?

Medical solution - Felfeliyan et al. 2022

- **Uniform soft average**

Annotator masks are averaged with equal weights into a single soft label.

- **Intensity-based boost**

High-intensity pixels inside the mask are pushed to 1; mask neighbours get a partial weight.

- **Noise-tolerant loss**

A Normalised Active–Passive Loss (NAPL), provably robust to label noise, replaces plain cross-entropy.

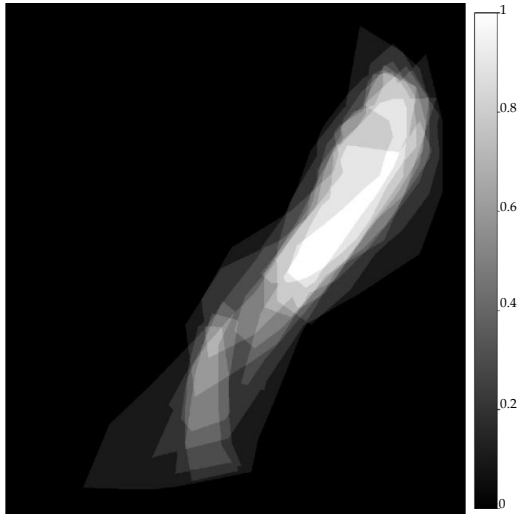


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

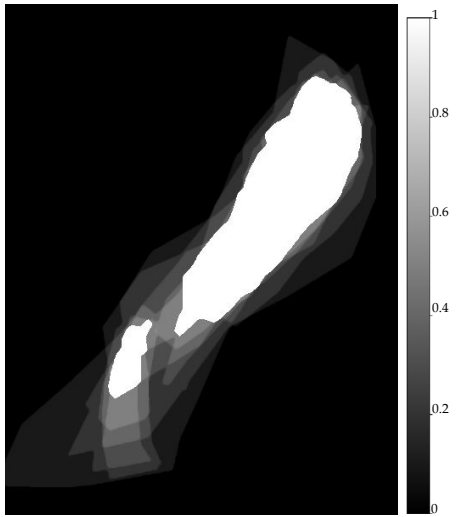


Fig.6: *Fellfeliyan method representation*

Boost confidence for high intensity pixels.

Medical solution - Zhang et al. 2023

- **Per-pixel variance map**

Inter-rater variance is computed at each pixel, quantifying where annotators disagree.

- **Gated supervision**

Standard supervision is applied only on low-uncertainty (high-agreement) pixels.

- **Consistency regularisation**

On high-uncertainty pixels, predictions are forced to be consistent under augmentation instead.

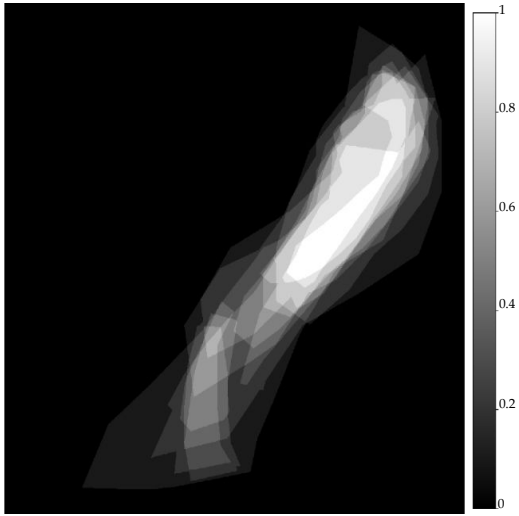


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

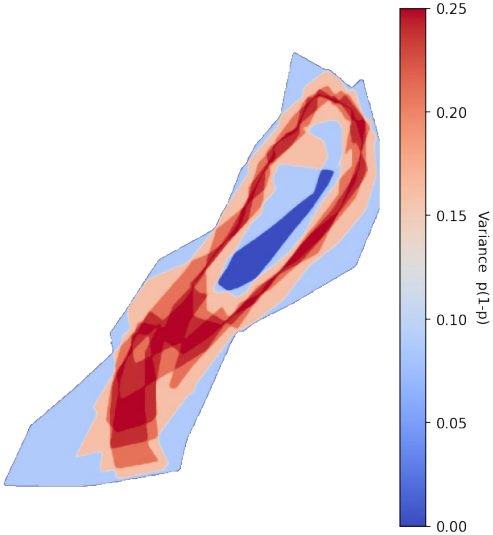


Fig.7: *Zhang variance map*

Down-weights ambiguous boundaries rather than forcing the model to fit conflicting labels.

Inter-annotator disagreement -

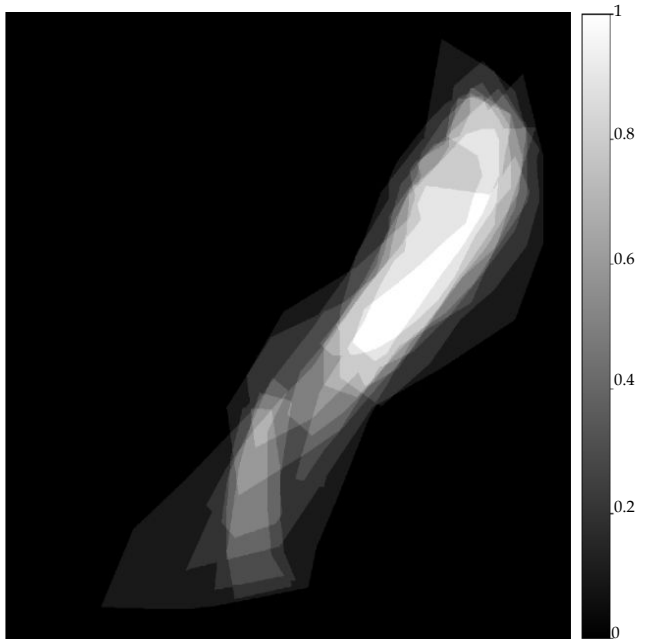


Fig.4: *15 annotations made by 4 annotators on a tidal tail.*

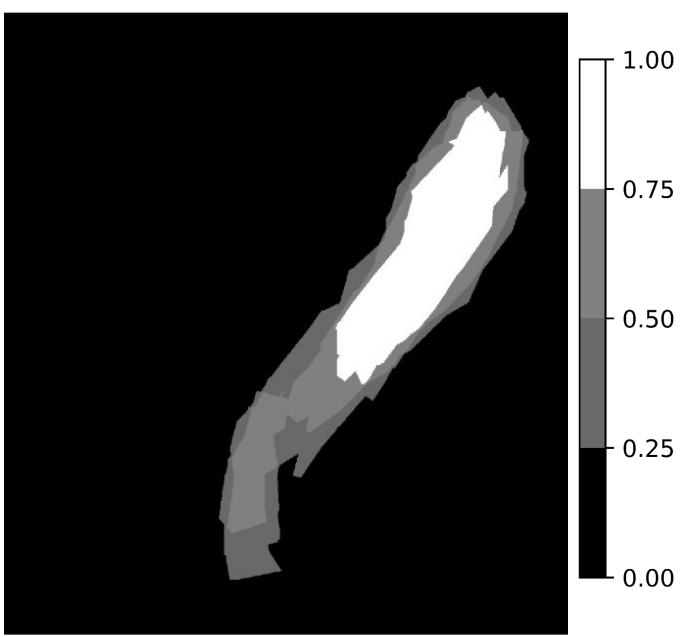
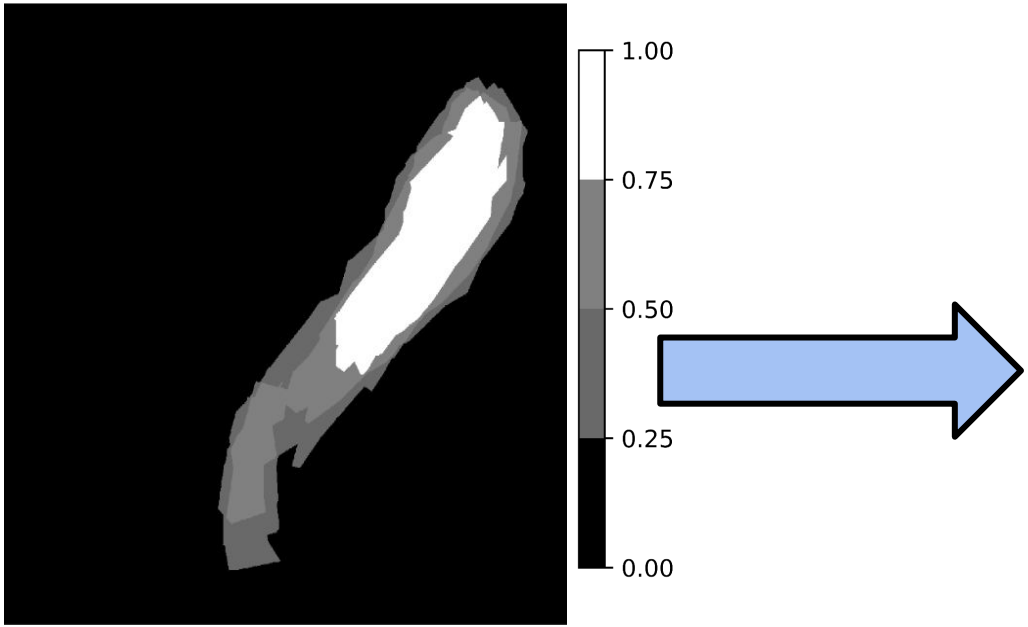


Fig.8: *Example definition of area of confidence levels.*

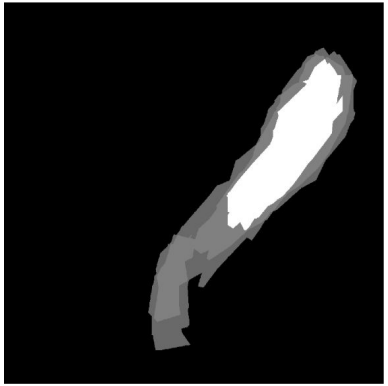
Inter-annotator consensus -



Pixel value	Structure presence
0 to 0.25	No
0.25 to 0.5	??
0.5 to 0.75	Fairly sure
0.75 to 1	Yes

Fig.8: *Example definition of area of confidence levels.*

Consensus Loss -



Pixel value	Structure presence
0 to 0.25	No
0.25 to 0.5	??
0.5 to 0.75	Fairly sure
0.75 to 1	Yes

Weight the loss function based on the confidence level in the “ground truth”:

$$L_c = \begin{cases} \beta.FL(p_t) & \text{if } y \geq 0.75 \text{ or } y \leq 0.25 \\ FL(p_t) & \text{if } 0.5 \leq y \leq 0.75 \\ 0 & \text{if } 0.25 \leq y \leq 0.5 \end{cases}$$

Ground Truth -

- At least one expert (sola/paduc) → union of the expert masks
- No expert but ≥2 non-experts → intersection of the non-expert masks
- No expert and only 1 non-expert → discarded as a likely false positive

Fig.9: *Example definition of ground truth*

MAP@50 -

Configuration	ES	G	IS	GH	HB	mAP50
Ours						
All, Union	0.094	0.619	0.742	0.509	0.656	0.524
All, Intersection	0.056	0.319	0.514	0.513	0.558	0.392
All, Consensus	**0.213**	0.603	0.814	**0.522**	0.674	0.565
Expert×2, Consensus	0.197	**0.648**	**0.838**	0.510	0.670	**0.573**
Expert only, Consensus	0.089	0.554	0.725	0.500	0.677	0.509
Non-expert only, Consensus	0.108	0.618	0.810	0.516	0.686	0.547
Medical-imaging baselines						
Silva & Oliveira [15]	0.128	0.647	0.814	0.524	**0.689**	0.560
Zhang et al. [21]	0.131	0.616	0.854	0.510	0.672	0.557
Felfeliyan et al. [5]	0.005	0.247	0.517	0.184	0.256	0.242

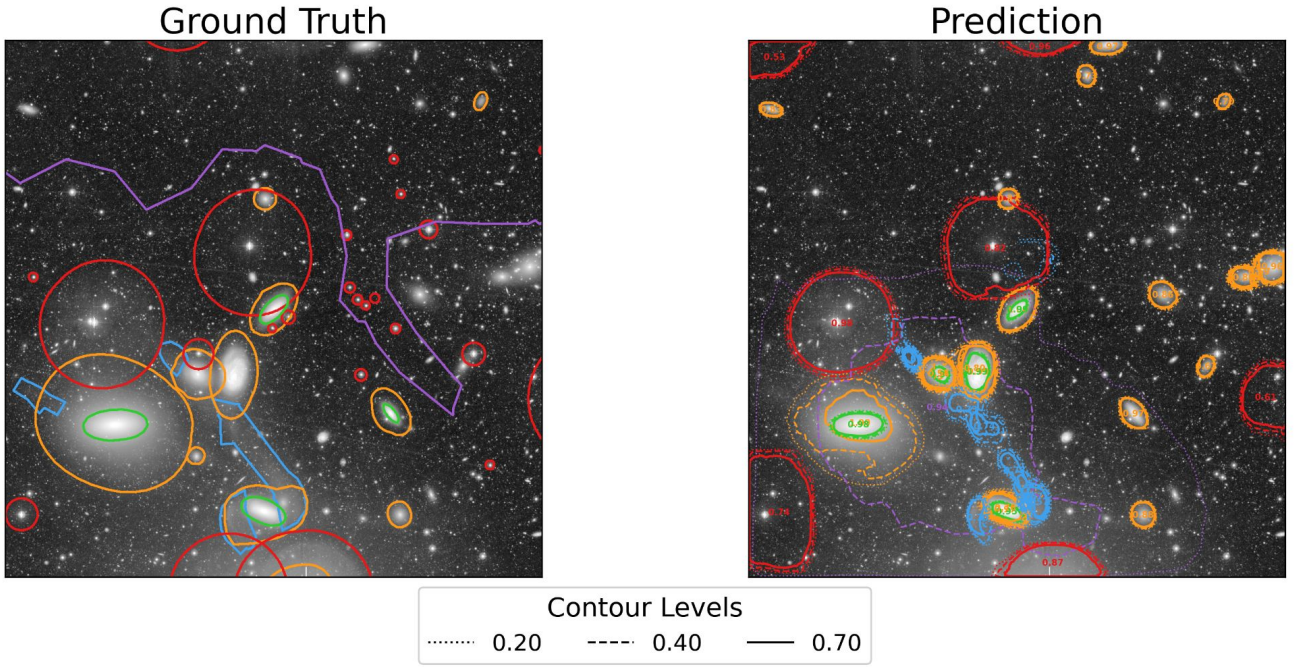
How many objects did you correctly find, and how well ?

DICE@50 -

Configuration	ES	G	IS	GH	HB	All
Ours						
All, Union	0.618	0.723	0.760	0.794	0.830	0.740
All, Consensus	**0.781**	**0.763**	0.731	**0.812**	0.897	**0.842**
Expert×2, Consensus	0.773	0.757	**0.742**	0.807	0.893	0.835
Expert only, Consensus	0.529	0.714	0.748	0.799	0.841	0.745
Non-expert only, Consensus	0.416	0.673	0.718	0.734	0.792	0.697
Medical-imaging baselines						
Silva & Oliveira [15]	0.678	0.754	0.729	0.803	0.904	0.0.769
Zhang et al. [21]	0.769	0.820	0.810	0.847	0.902	0.838

When you find an object, how well did you do it ?

Pratical example -



Conclusion :

- Hard to detect, **precise once found**
- **Large morphological variability** within Tidal structures, combined with a small dataset, drives the detection gap.
- Consensus-conditioned supervision **improves results overall**.

Prospect :

- Apply the method to **deeper surveys such as Euclid**, where signal-to-noise is better than MATLAS.
- Transferred to **medical data** (QUBIQ): mAP50 and Dice are high, confirming the method **works on easier datasets**.
- On QUBIQ it continues to outperform the competing multi-annotator methods — **not just domain tuning**.

Supplementary