Inter-annotator consensus: Towards Automated Detection of Low Surface Brightness Features in Large-Scale Surveys

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

MACHINE LEARNING FOR ASTROPHYSICS
3RD EDITION VALLETTA, 31 AUG - 4 SEP, 2026

September 1st, 2026

ML4Astro 2026 3rd edition: Inter-annotator consensus: Mr. Vancoellié renaud.vancoellie@gmail.com

Tidal features -

Why them

- **Fossil record of mergers**
- **Reconstructing merger histories**
- **Testing ΛCDM**

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

ML4Astro 2026 3rd edition: Inter-annotator consensus: Mr. Vancoellié renaud.vancoellie@gmail.com

Tidal features -

Why them

- **Fossil record of mergers**
- **Reconstructing merger histories**
- **Testing ΛCDM**

Detection challenge

- **At/Below the noise floor**
- **No physical edge**
- **Survey scale**

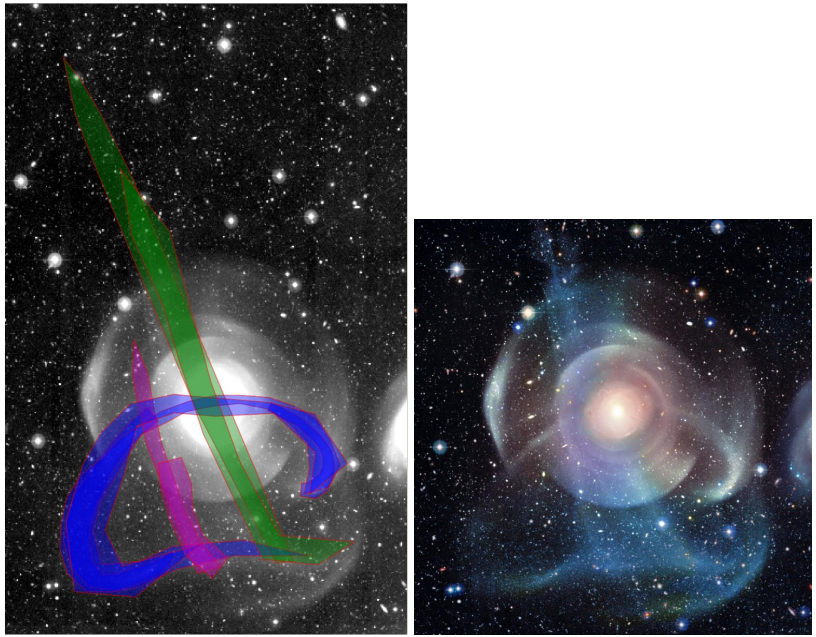


Fig.2: *NGC0474 image and the three identified structures.*

Tidal features -

Why them

- **Fossil record of mergers**
- **Reconstructing merger histories**
- **Testing ΛCDM**

Detection challenge

- **At/Below the noise floor**
- **No physical edge**
- **Survey scale**

Inter annotator disagreement

- **Astronomers rarely agree on where a tidal feature starts/ends.**
- **Structured, not random**

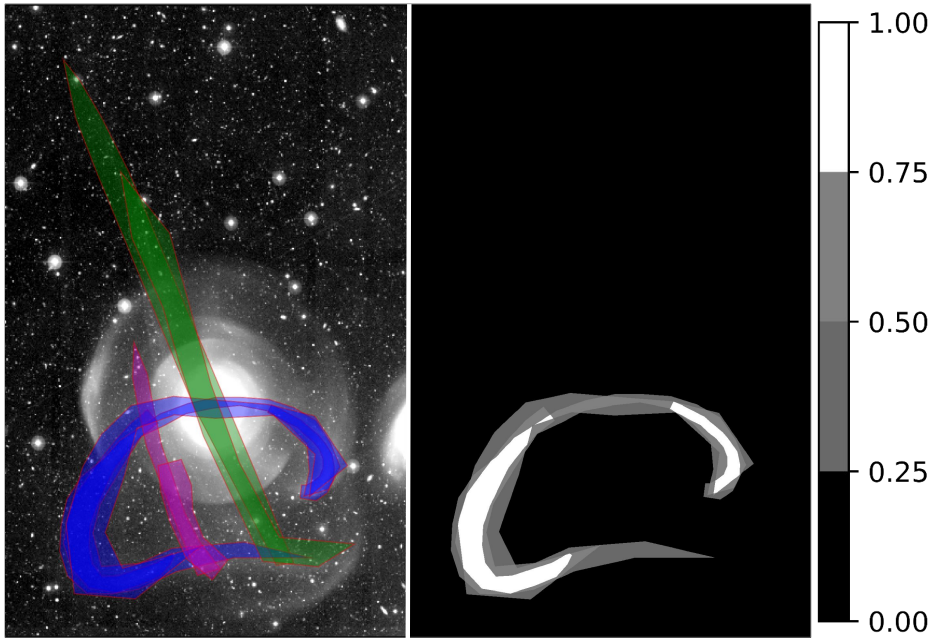
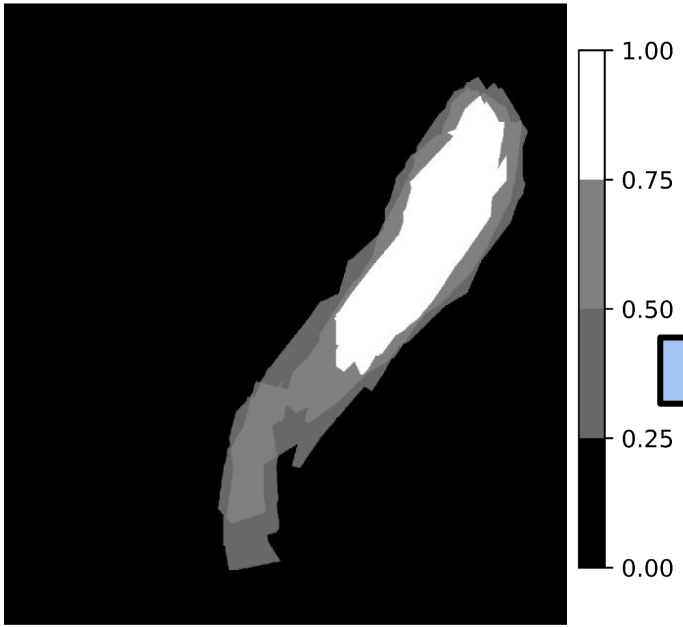


Fig.3: *NGC0474 image and the three identified structures.*

SOTA Methods -

		Multiple Annotator	Segmentation task
Astrophysic	Richards, et al. *Multi-scale gridded gabor attention* 2022	✗	✓
	Walmsley et al. *Galaxy zoo decals:* 2022	✓	✗
Medical	Zhang, et al. *A soft label method for medical image segmentation with multirater annotations.* 2023	✓	✓
	Felfeliyan, et al. *pervised medical image segmentation with soft labels and noise robust loss.* 2022	✓	✓
	Silva and Oliveira *Using soft labels to model uncertainty in medical image segmentation* 2021	✓	✓

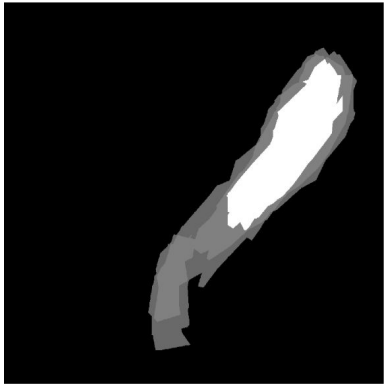
Inter-annotator consensus -



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.4: *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":

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

Dataset - MATLAS, Astrophysical dataset

MNIST - Synthetic dataset
QUBIQ - Medical dataset

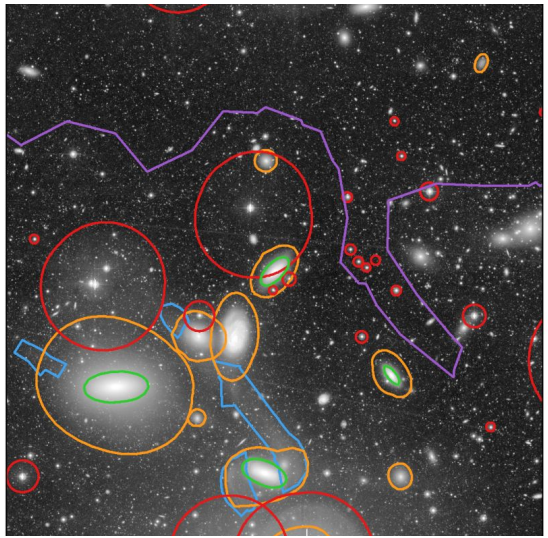


Fig.5: Annotation example

THE ANNOTATOR POOL:
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

ML4Astro 2026 3rd edition: Inter-annotator consensus: Mr. Vancoellié renaud.vancoellie@gmail.com

MATLAS - Astrophysical data

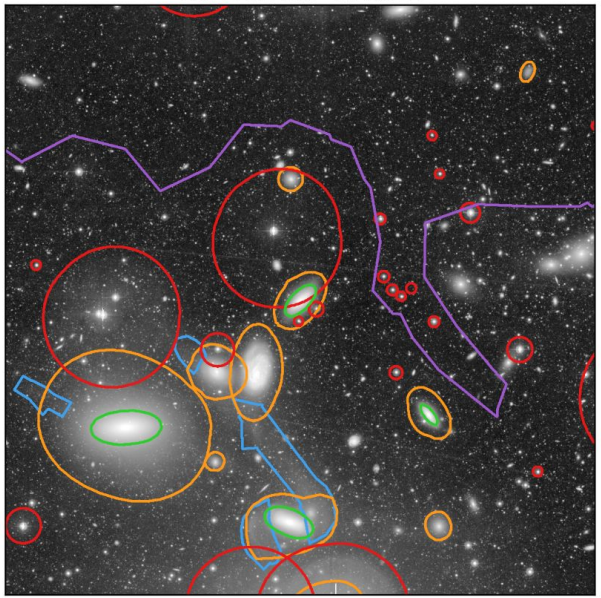
Method	ES	G	IS	GH	HB
Proposed, all raters	**0.778**	0.836	0.797	**0.892**	0.854
Proposed, expert×2	0.764	0.804	**0.860**	0.883	0.787
Proposed, expert only	0.769	0.864	0.819	0.875	0.824
Proposed, non-expert only	0.764	0.802	0.817	0.883	0.860
Union	0.752	**0.873**	0.829	0.890	**0.894**
Intersection	0.759	0.816	0.802	0.888	0.864
Silva & Oliveira [14]	0.748	0.809	0.801	0.868	0.847
Zhang *et al.* [20]	0.774	0.782	0.789	0.832	0.818
Felfeliyan *et al.* [4]	0.708	0.769	0.800	0.735	0.731

Table.1: *Dice score per MATLAS class. Best per column in* ***bold****.*

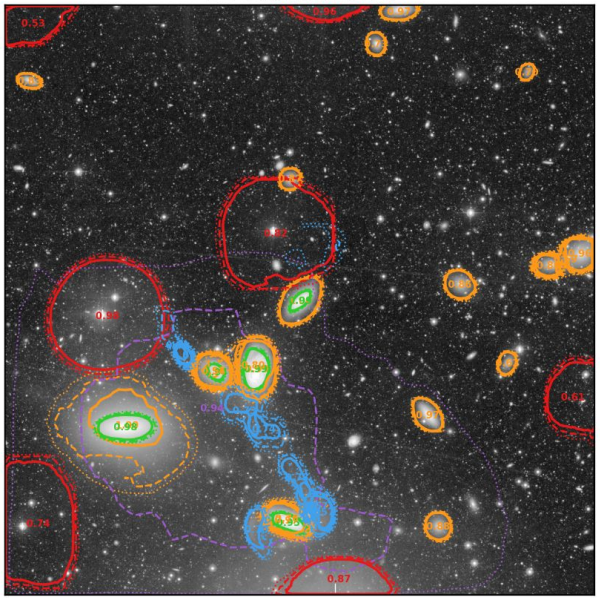
ML4Astro 2026 3rd edition: Inter-annotator consensus: Mr. Vancoellié renaud.vancoellie@gmail.com

Practical example -

Ground Truth



Prediction



Contour Levels

0.20 0.40 0.70

Elongated Structures | Galaxy | Inner Structure | Ghosted Halo | High Background

Conclusion -

- Depending on the **number of annotation** and **agreement** between **annotators, different method** should be used for optimal result

- When **low** number of **annotations**, **Non-Expert improve** the result with **our consensus methods**
- When **high** number of **annotations**, **Only-Expert** should be taken into account to minimize noise
- When **high Disagreement**, **Non-Expert improve** the result with **our methods**

Dataset - MNIST, Synthetic dataset

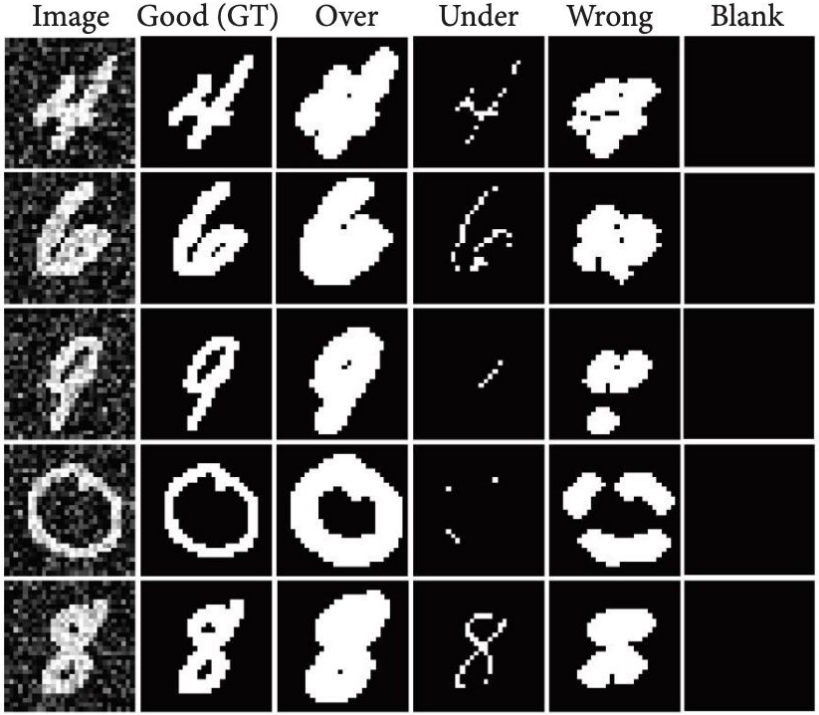
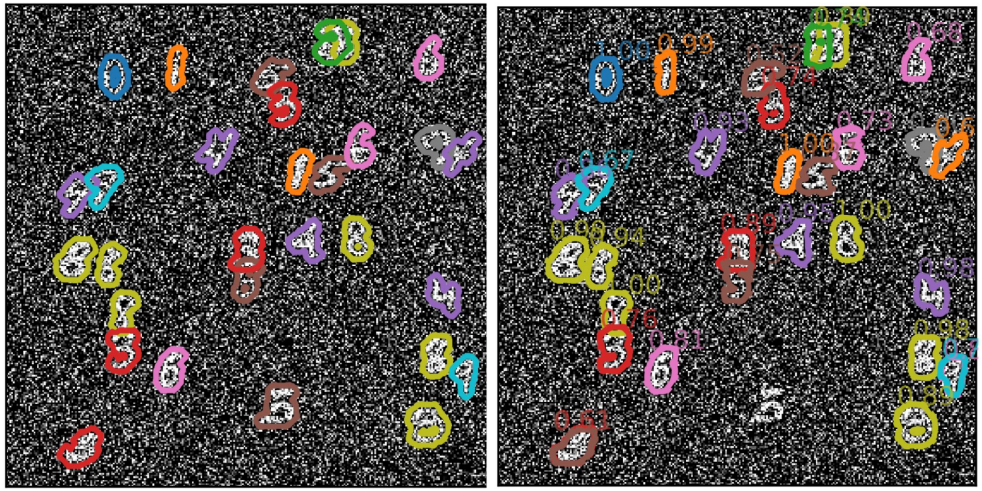
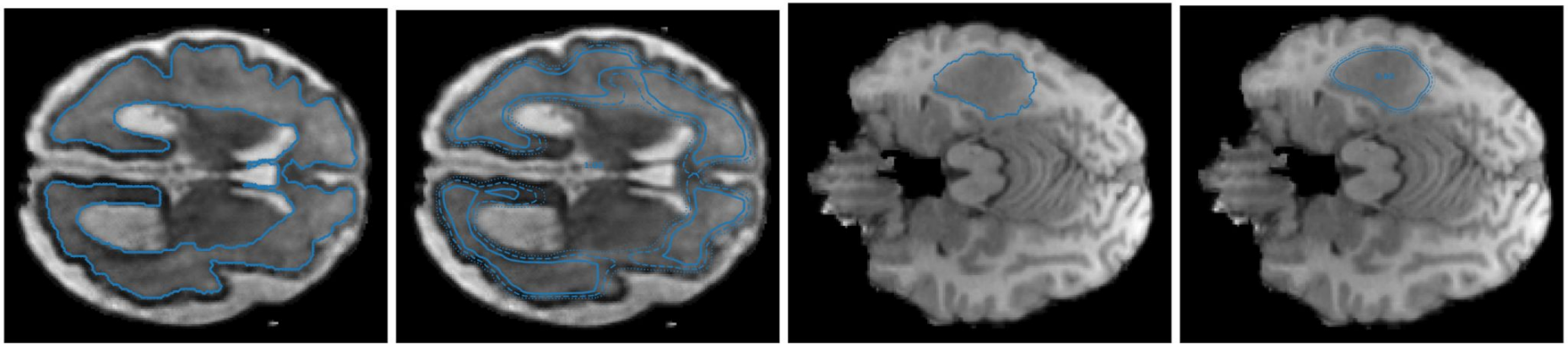


Fig.5: Zhang et al. 2023



Dataset - QUBIQ, Medical dataset



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.

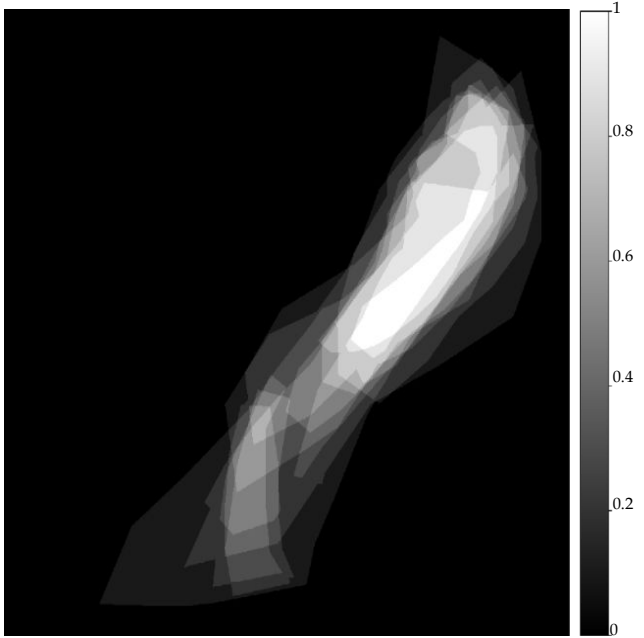


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

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.

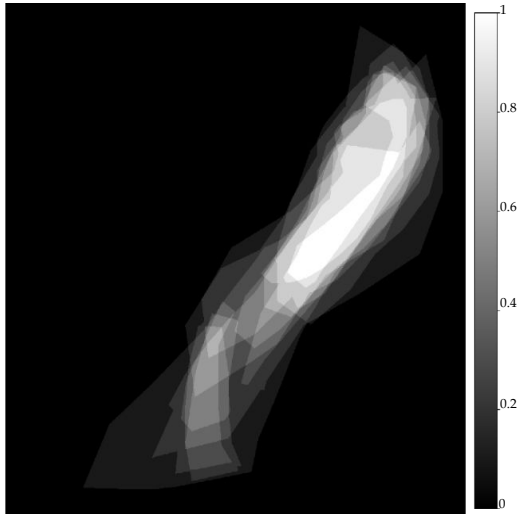


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

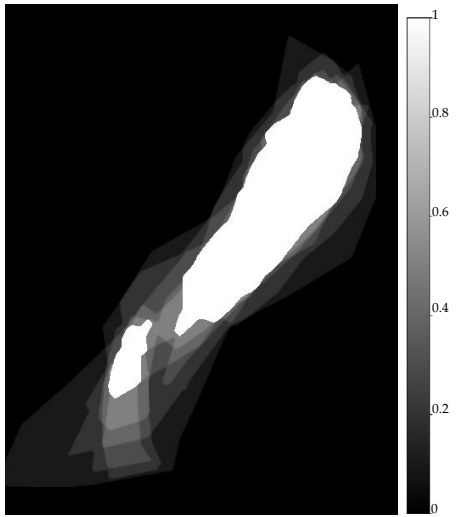


Fig.6: *Fellfeliyan method representation*

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.

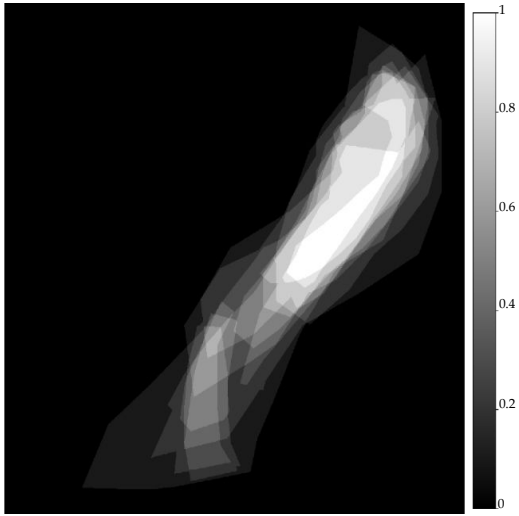


Fig.5: *15 annotations made by 4 annotators on a tidal tail.*

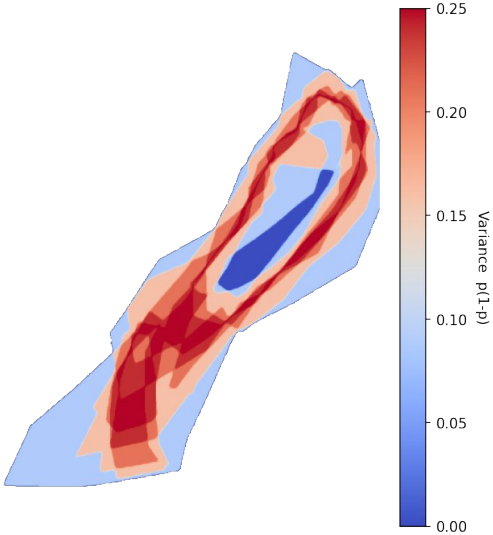


Fig.7: *Zhang variance map*

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

MNIST - Synthetic data

Method	AP segm	Dice
Proposed	**0.918**	**0.932**
Union	0.877	0.899
Intersection	0.618	0.649
Silva & Oliveira [14]	0.916	0.921
Zhang *et al.* [20]	0.748	0.769
Felfeliyan *et al.* [4]	0.837	0.856

Table 2. Results on MNIST. Best per column in **bold**.

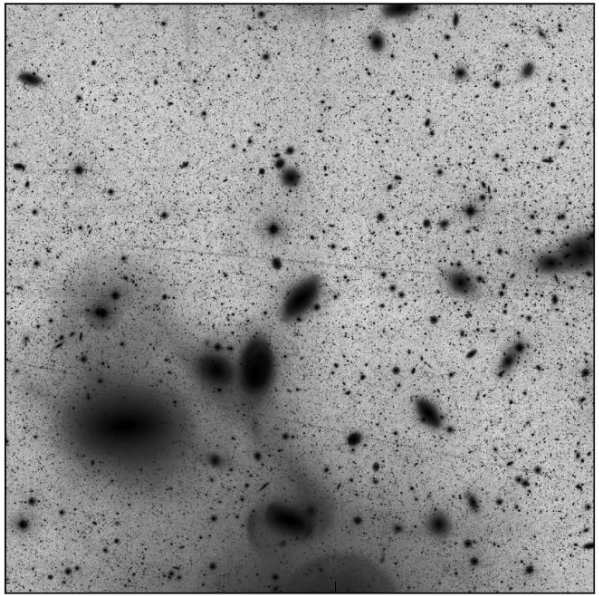
QUBIQ - Medical data

Method	Brain-growth Dice	Brain-tumour AP segm	Brain-tumour Dice
Proposed	**0.744**	0.841	0.869
Union	0.741	0.848	0.862
Intersection	0.724	0.806	**0.875**
Silva & Oliveira [14]	0.737	0.839	0.872
Zhang *et al.* [20]	0.732	0.854	0.860
Felfeliyan *et al.* [4]	0.720	**0.873**	0.867

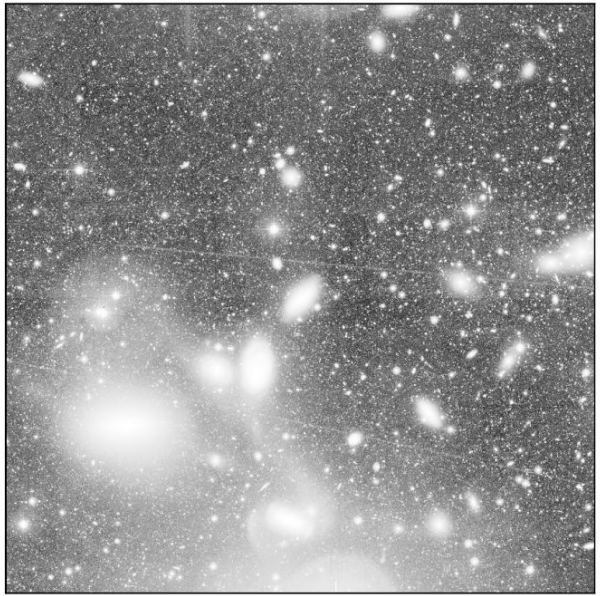
Table 3. QUBIQ results. Best per column in **bold**.

Supplementary

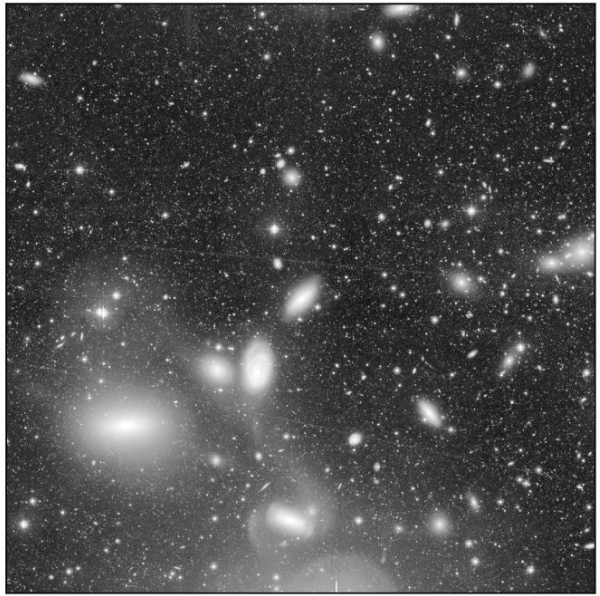
Scale 0



Scale 1



Scale 2



Inter-annotator disagreement -

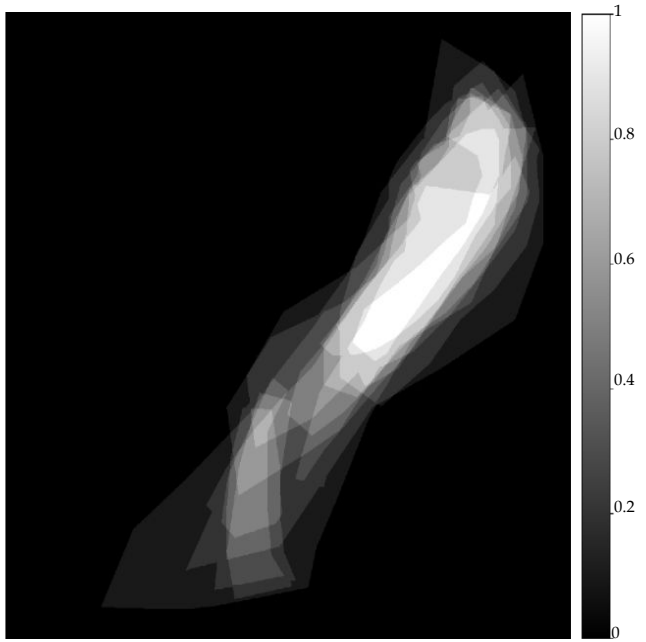


Fig.4: 15 annotations made by 4 annotators on a tidal tail.

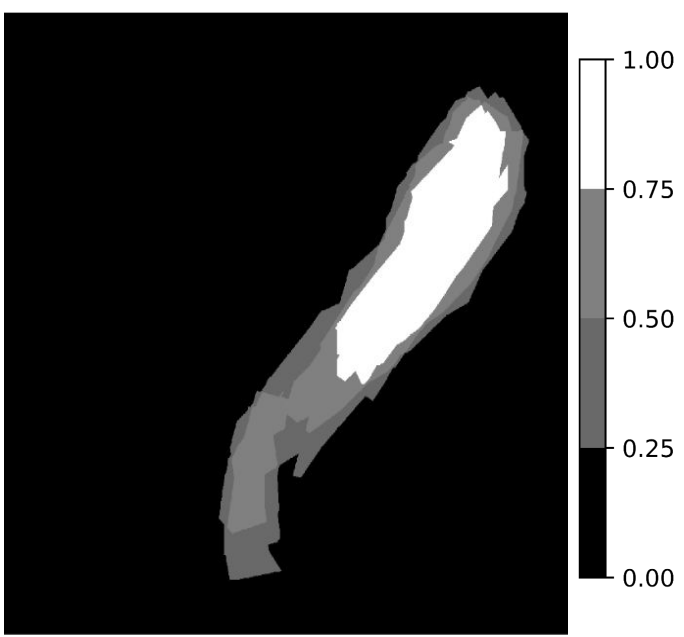


Fig.8: Example definition of area of confidence levels.

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*