Physics-informed deep neural network for characterising galaxy morphology

Adeline Paiement, Pierre-Alain Duc, Felix Richards, Elisabeth Sola, Renaud Vancoellié, Worachit Ketrungsri, Sreevarsha Sreejith

LIS Laboratoire d'Informatique & Systèmes, Université de Toulon


Observatoire astronomique de Strasbourg


Swansea University Prifysgol Abertawe


Observatoire astronomique de Strasbourg
ioa


LIS Laboratoire d'Informatique & Systèmes, Université de Toulon


LIS Laboratoire d'Informatique & Systèmes, Université de Toulon


Brookhaven National Laboratory

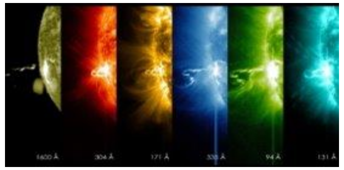
Outline

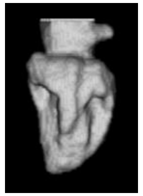
1. Quick introduction – Who am I?

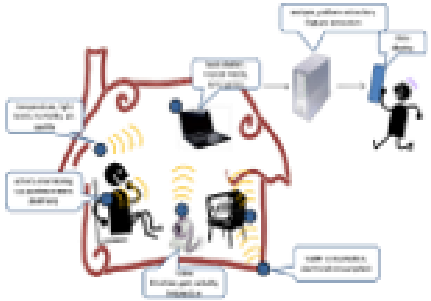
2. Characterising galaxy morphology – Context and challenges

3. Physics-informed deep learning method – Our (in progress) solution

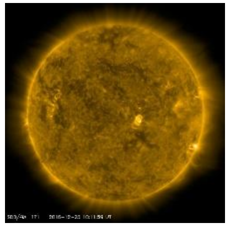
Background

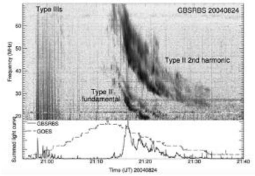


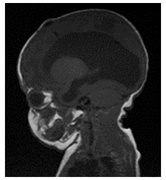




sphere







2008 | 2013 | 2016 | 2018

Diplôme d'ingénieur + MSc astrophysics → PhD in computer vision → PostDoc in machine learning → Research lecturer (eq. MCF) → Maître de conférences

ECOLE D'INGENIEURS TELECOM PHYSIQUE STRASBOURG

Observatoire astronomique de Strasbourg

SCISYS

INSTITUTE FOR ASTRONOMY UNIVERSITY OF HAWAII

University of BRISTOL

Swansea University Prifysgol Abertawe

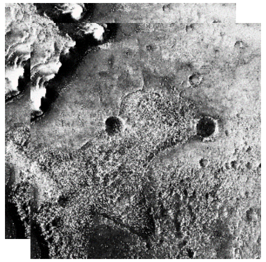
LIS LABORATOIRE D'INFORMATIQUE & SYSTEMES

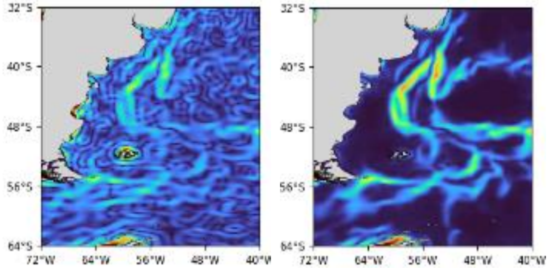
UNIVERSITÉ DE TOULON

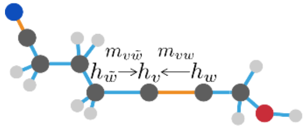
My research topic

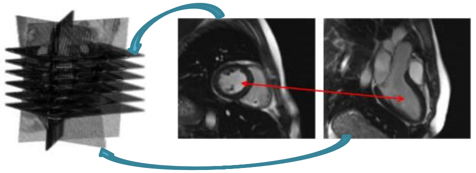
Designing domain-informed machine learning and deep learning methods

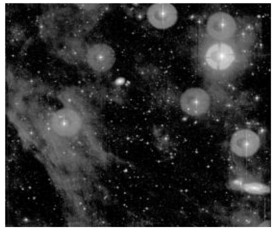
for science data analysis

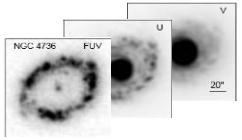


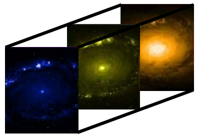


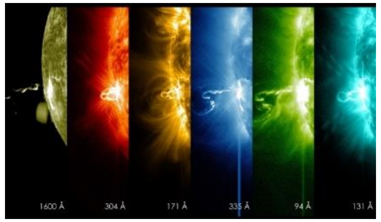


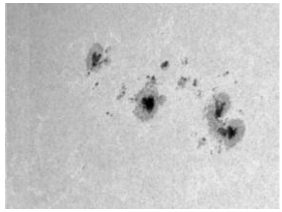


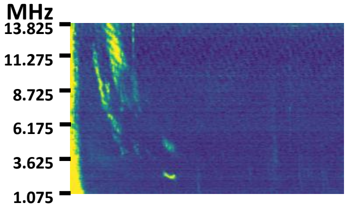












My research topic

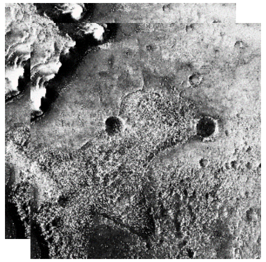
Similar tasks across applications

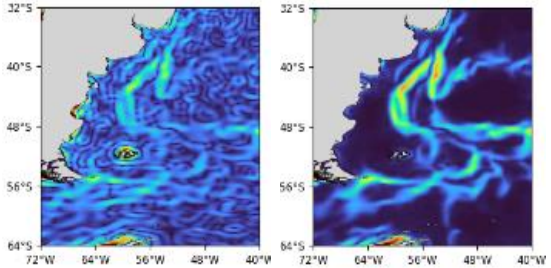
1) Spatial analysis:

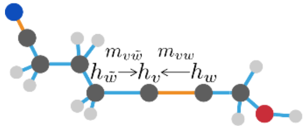
- Localisation
- Characterisation: classification
- Characterisation: properties of 2D and 3D geometry

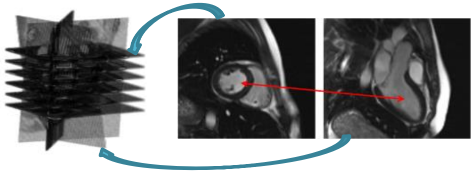
2) Temporal analysis: evolution modelling

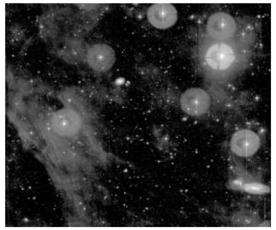
- Characterisation of movement
- Prediction

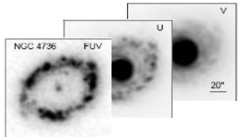


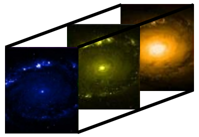


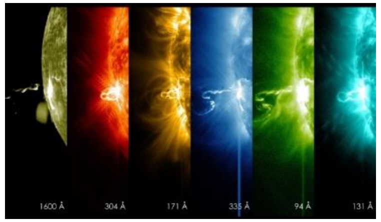


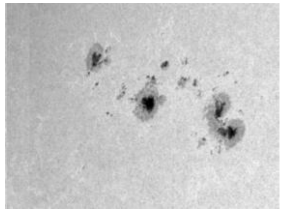


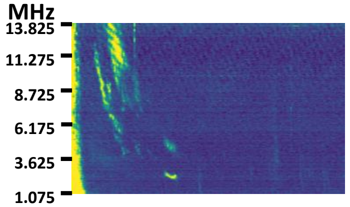












My research topic

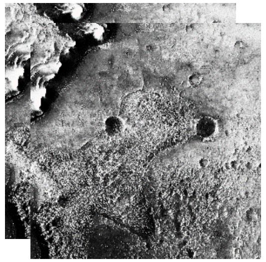
Similar challenges across applications

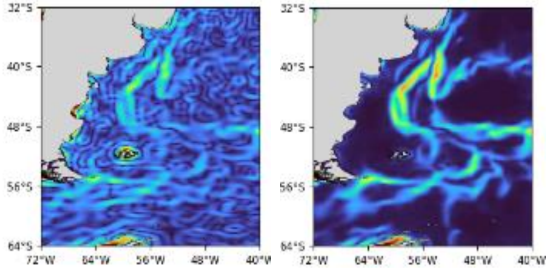
1) Multimodal data

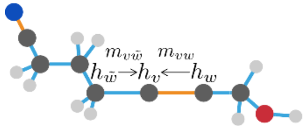
- Different appearances
- Different resolutions
- Misalignments
- Fusion of information

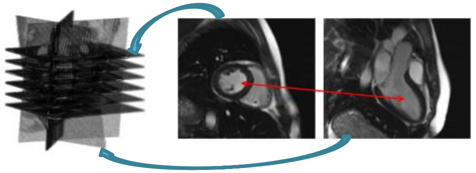
2) Scientific vs. natural images/data

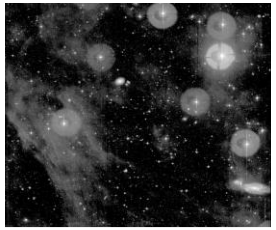
- Data quality (dynamic range, contrast, noise…)
- Expert interpretation
- Ground-truth availability

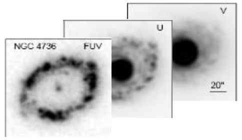


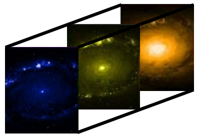


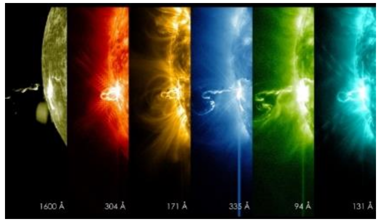


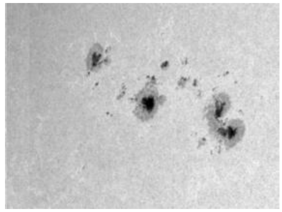


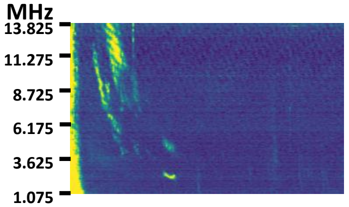












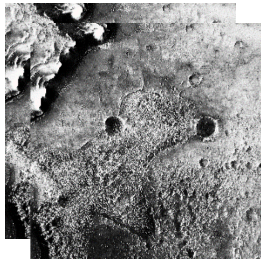
My research topic

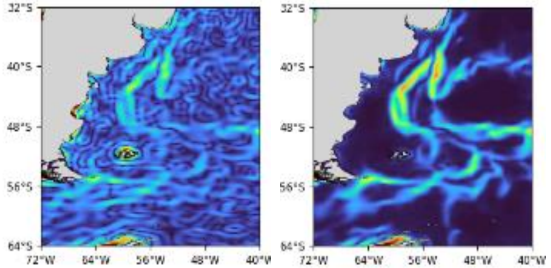
Similar requirements across applications

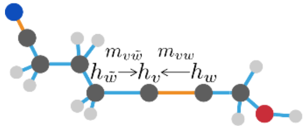
1) Robustness
   - To small annotated training sets
   - Physical validity of solutions

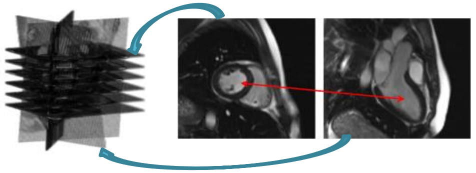
2) Interpretability

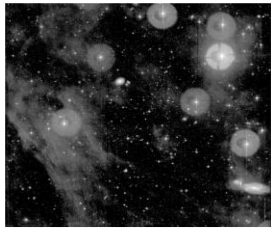
→ **Domain-informed analysis**

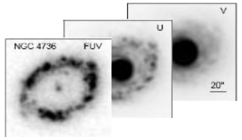


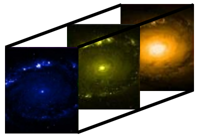


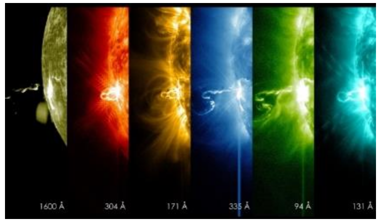


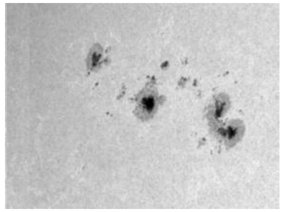


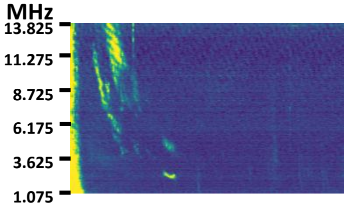










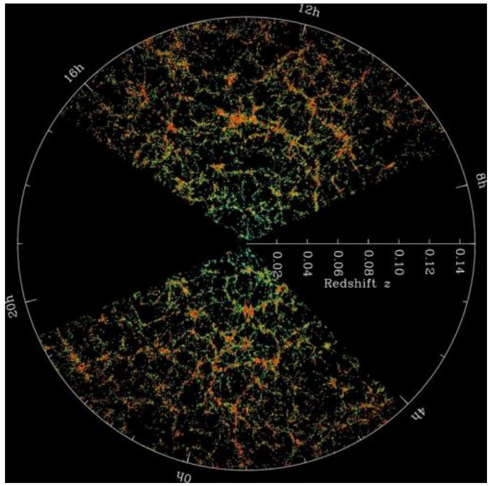


Characterising galaxy morphology

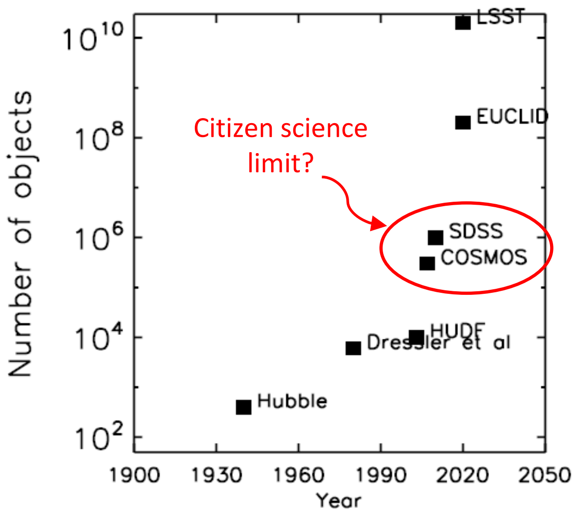
CONTEXT AND CHALLENGES

Big data call for automatic recognition algorithms

- Big data from astronomical surveys
  - Sloan Digital Sky Survey (SSDS): 5 optical wavelengths, 35% of the sky, 500 million objects, 200 GB of data per night
  - Wide-field Infrared Survey Explorer (WISE): 4 infrared wavelengths, 99% of the sky 8 times!
  - CANDELS: optical and infrared wavelengths, deep sky, more than 250 000 galaxies
  - Dark Energy Survey (DES): over 300 million galaxies, statistics on visual shear for study of the weak gravitational lensing effect
  - … and many more past and future surveys, such as Euclid, NEOCam, and LSST
- Citizen science has helped analysing data in the past, but now reaches a limit
- **Automated recognition algorithms get promising results thanks to deep learning**



2D cut in the 3D map of SDSS



Analysing galaxy morphology

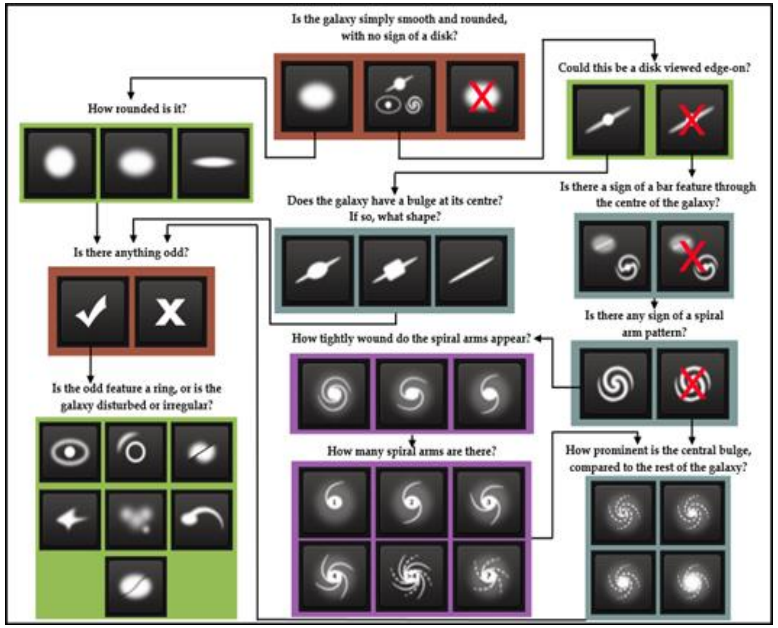
Different levels of analysis:

A. Classification of morphology types

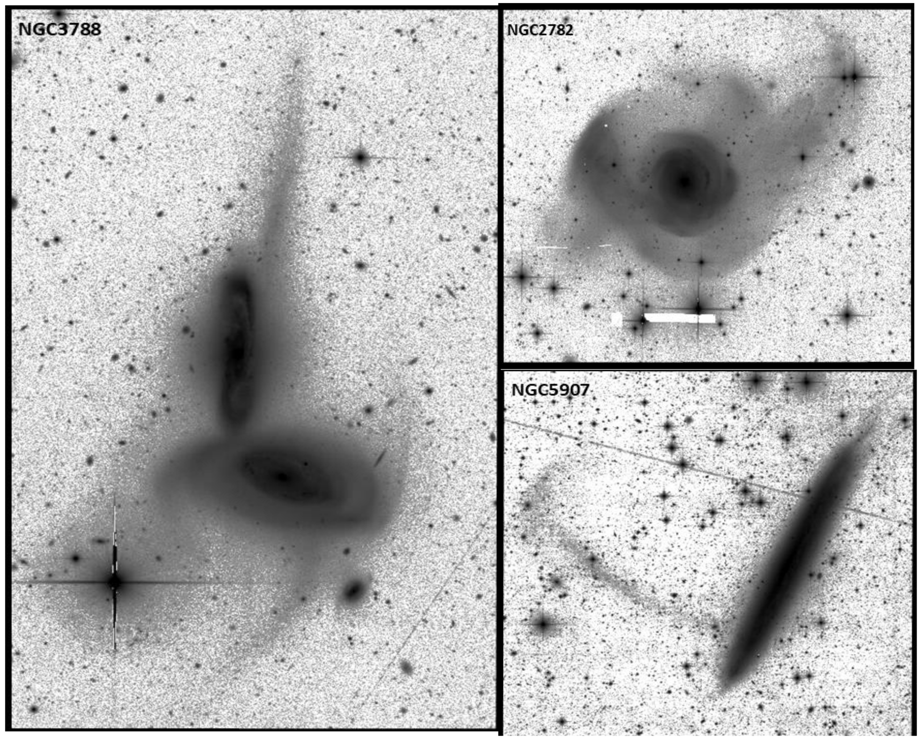
B. Classification/regression of morphology parameters

C. Identification of low brightness collisional debris

→ insights into the galaxy's evolution history



Galaxy Zoo model

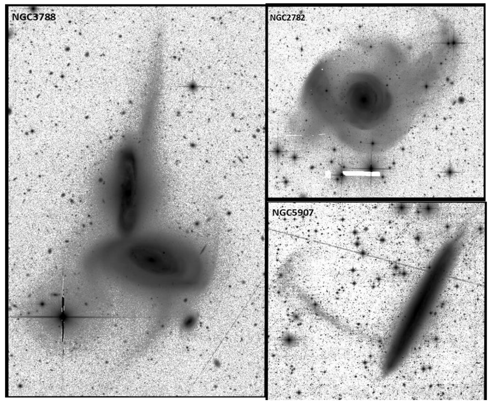


Examples of tidal features in CFIS images.

Identification of low brightness collisional debris

Here also, different levels of analysis:

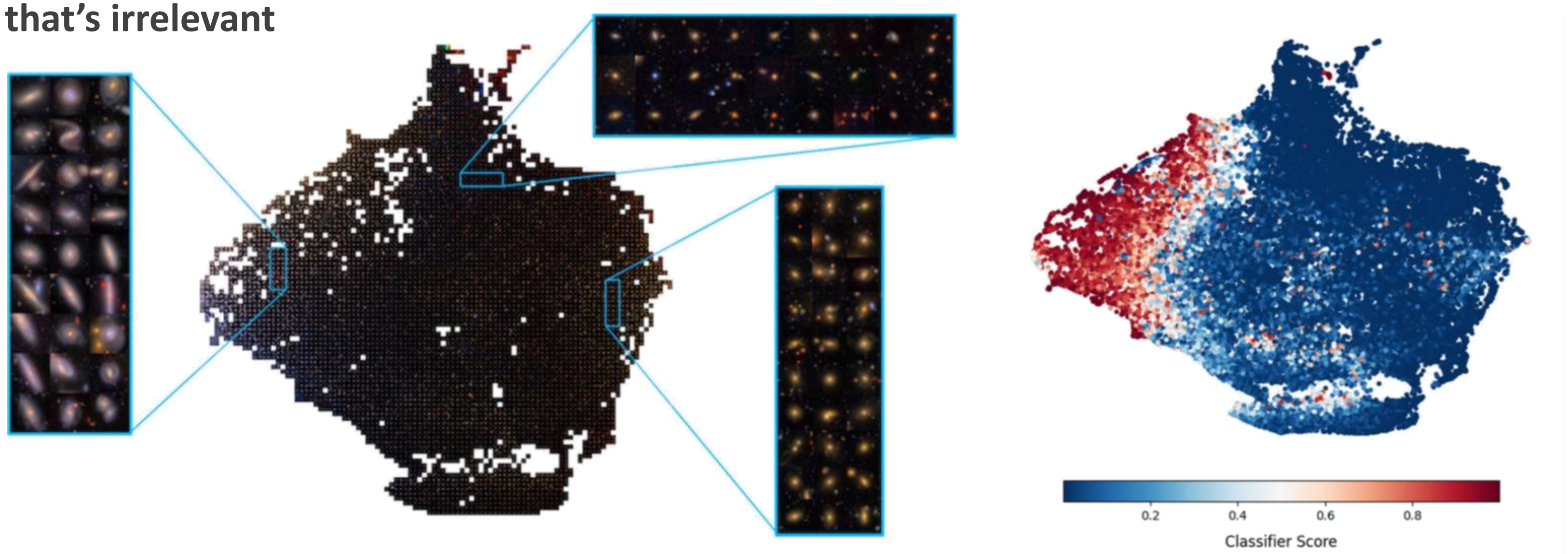
A. Classification of **presence vs. absence** of *undifferentiated* collisional debris
B. Classification of presence vs. absence of *specific types* of collisional debris
C. **Fine localisation** of collisional debris
   a) Bounding box detection
   b) Segmentation: pixel-wise localisation



Examples of tidal features in CFIS images.

Danger of a too simple classification approach:

The DL algorithm may model something that's irrelevant

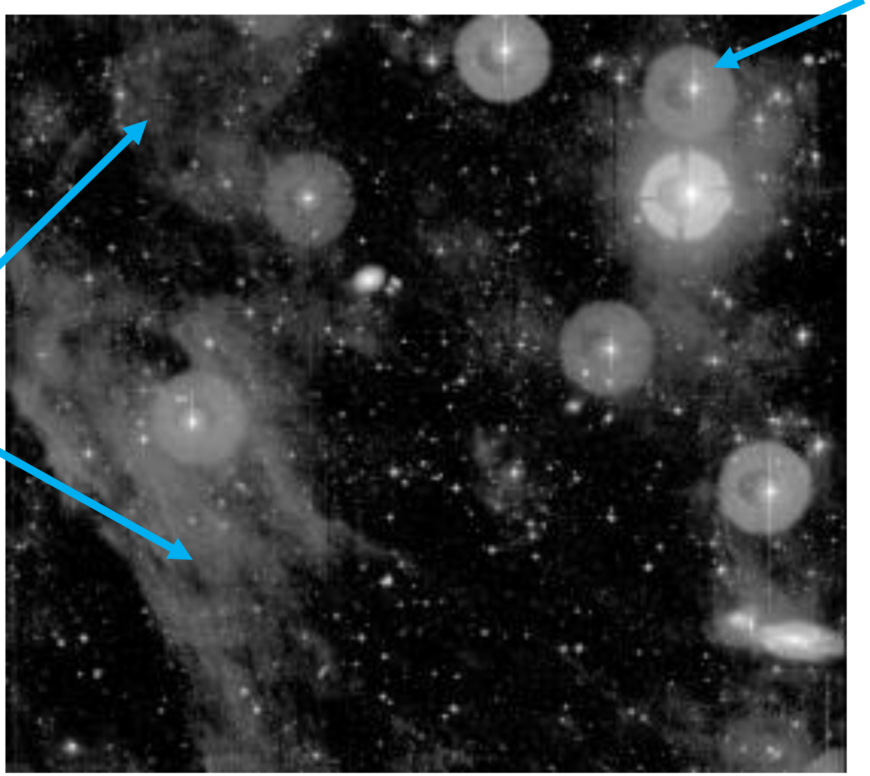


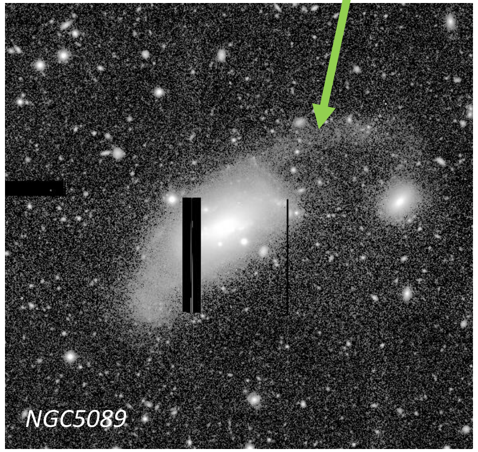
A. Desmons, S. Brough, F. Lanusse: Detecting Galaxy Tidal Features Using Self-Supervised Representation Learning. arXiv e-prints Jul 2023

Analysing galaxies in noisy & crowded images

Imaging reveals low surface brightness structures…

… but also dust clouds (cirrus) and imaging artefacts





Images from the MATLAS survey (Mass Assembly of early-Type GaLAxies with their fine Structures), CFHT MegaCam instrument

Physics-informed deep learning method

OUR (IN PROGRESS) SOLUTION

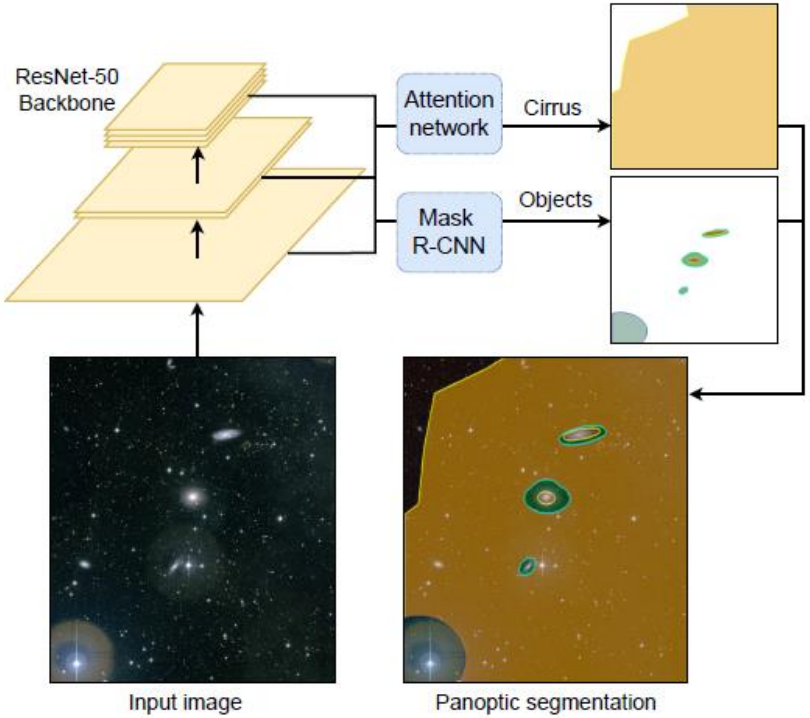
Outline of our approach

Fine-grained localisation (segmentation) of objects

Combined detection and segmentation of:

- galactic features
- image contaminants

→ Helps distinguish between tidal features and cirrus



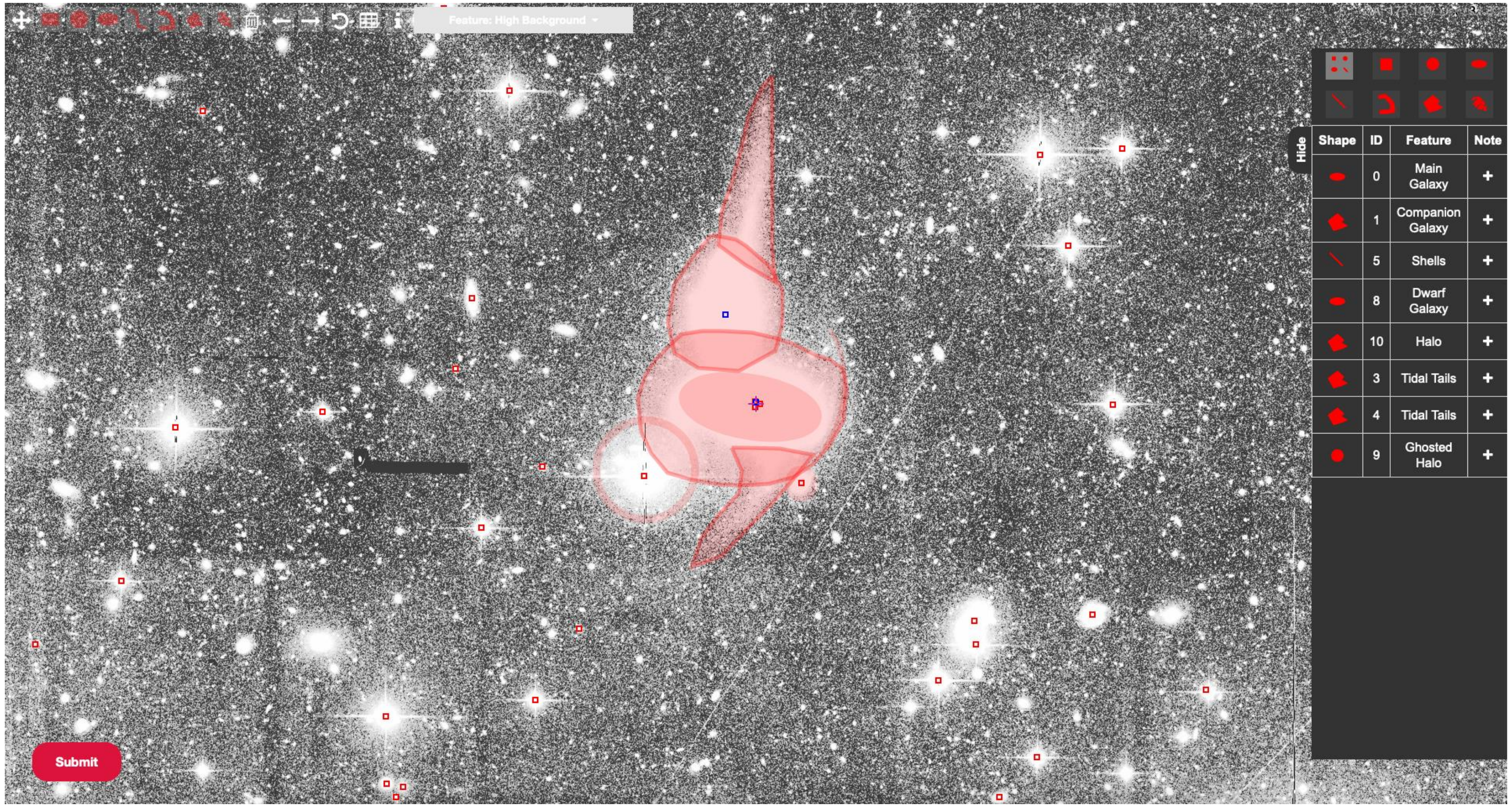
Setup of preliminary results [1]:

- New neural network input layer that is sensitive to low brightness structures [1]
- New neural network architecture that is sensitive to the oriented textures of cirrus [2,3]
- Dataset annotation using home-made tool [4]

[1] F. Richards, A. Paiement, X. Xie, E. Sola, P.-A. Duc: Panoptic Segmentation of Galactic Structures in LSB Images. *International Conference on Machine Vision Applications (MVA)*, 2023
[2] F. Richards, E. Sola, A. Paiement, X. Xie, P.-A. Duc: Multi-scale gridded Gabor attention for cirrus segmentation. *IEEE International Conference on Image Processing (ICIP)*, 2022
[3] F. Richards, A. Paiement, X. Xie, P.-A. Duc: Learnable Gabor modulated complex-valued networks for orientation robustness. Under review with *Image and Vision Computing*, 2023
[4] E. Sola, P.-A. Duc, F. Richards, A. Paiement, M. Urbano, J. Klehammer, M. Bílek, J.-C. Cuillandre, S. Gwyn, A. McConnachie: Characterization of Low Surface Brightness structures in annotated deep images, A&A, 662, A124, 2022

Step 1: Creating training data

An annotation tool for LSB structures



- **Online** tool to easily annotate and classify LSB structures in deep images
- Based on Aladin Lite
- Goal: **Draw** with precision the **shapes** of LSB structures

Annotation process

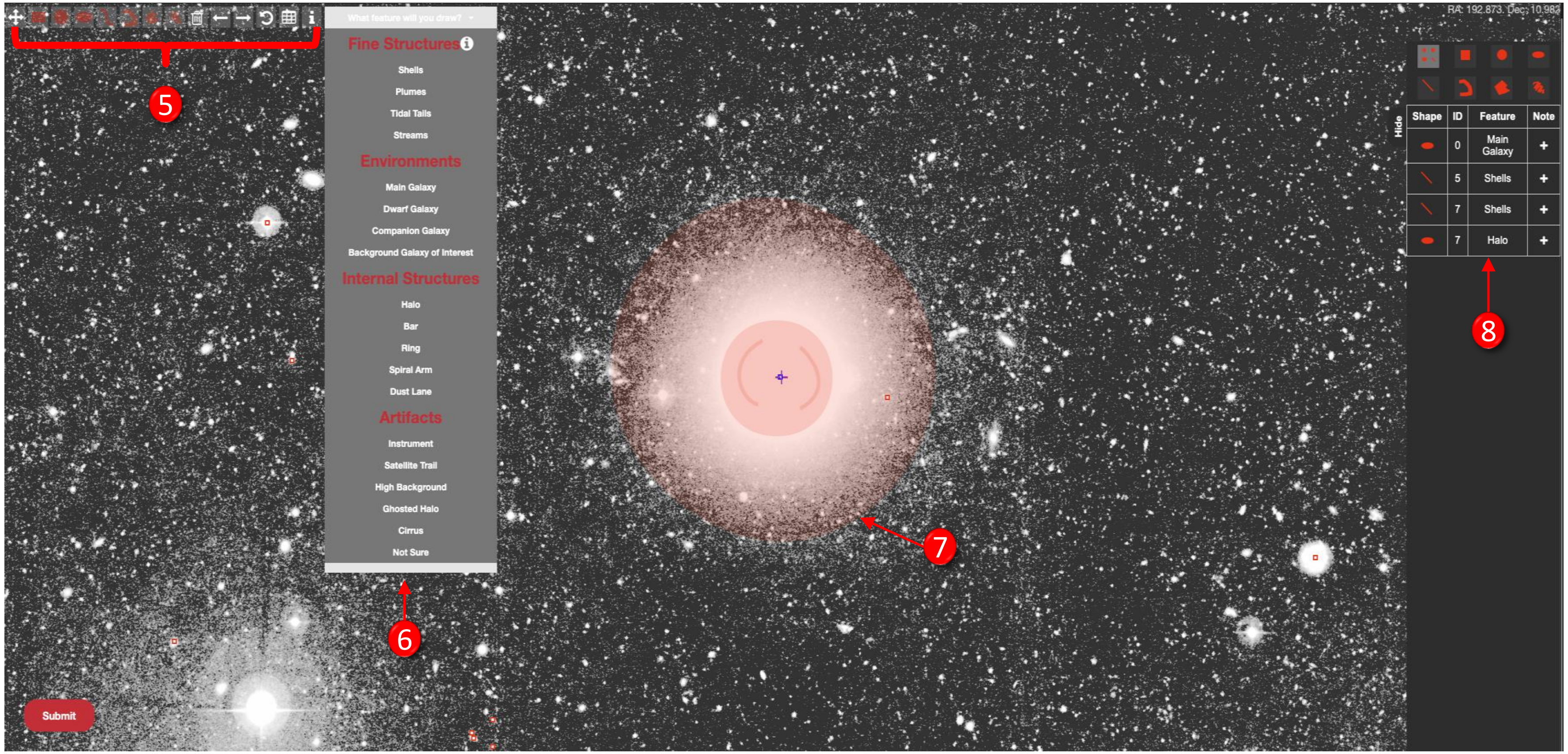
- Annotate features :

- Center
- Halo
- Tidal tails
- Streams
- Shells
- Companion

- Annotate pollutants :

- Ghosted halo
- High background
- Cirrus

→ The annotations are stored in a **database**

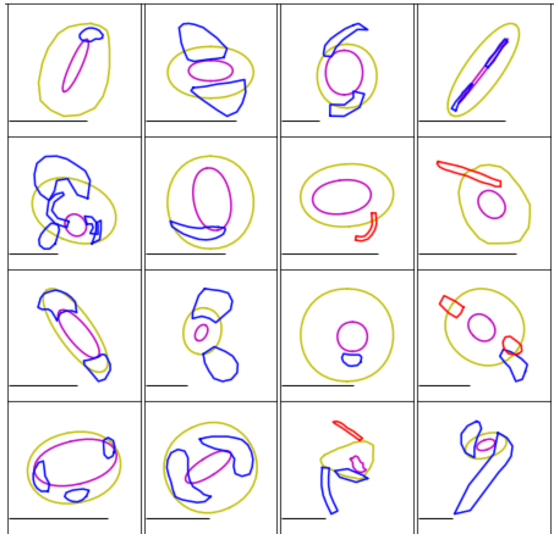


Drawing mode: 5: Drawing shapes; 6: Label selection; 7: Annotation already drawn; 8: Summary table

Analysis tools and measurements

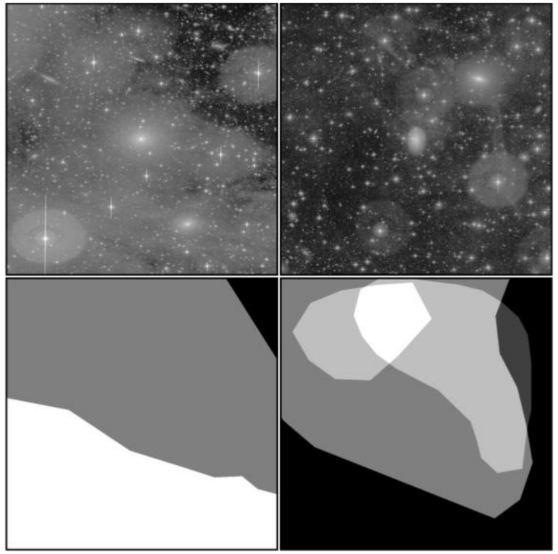
- **Several users**: need to take their contribution into account
- **Quantitative** measurements from annotation database: area, width, surface brightness
- **Thumbnails**: simple visualization the complex shapes of LSB structures

Thumbnails (Sola et al 2022)



- Ambiguous cirrus boundaries cause annotator **disagreement**

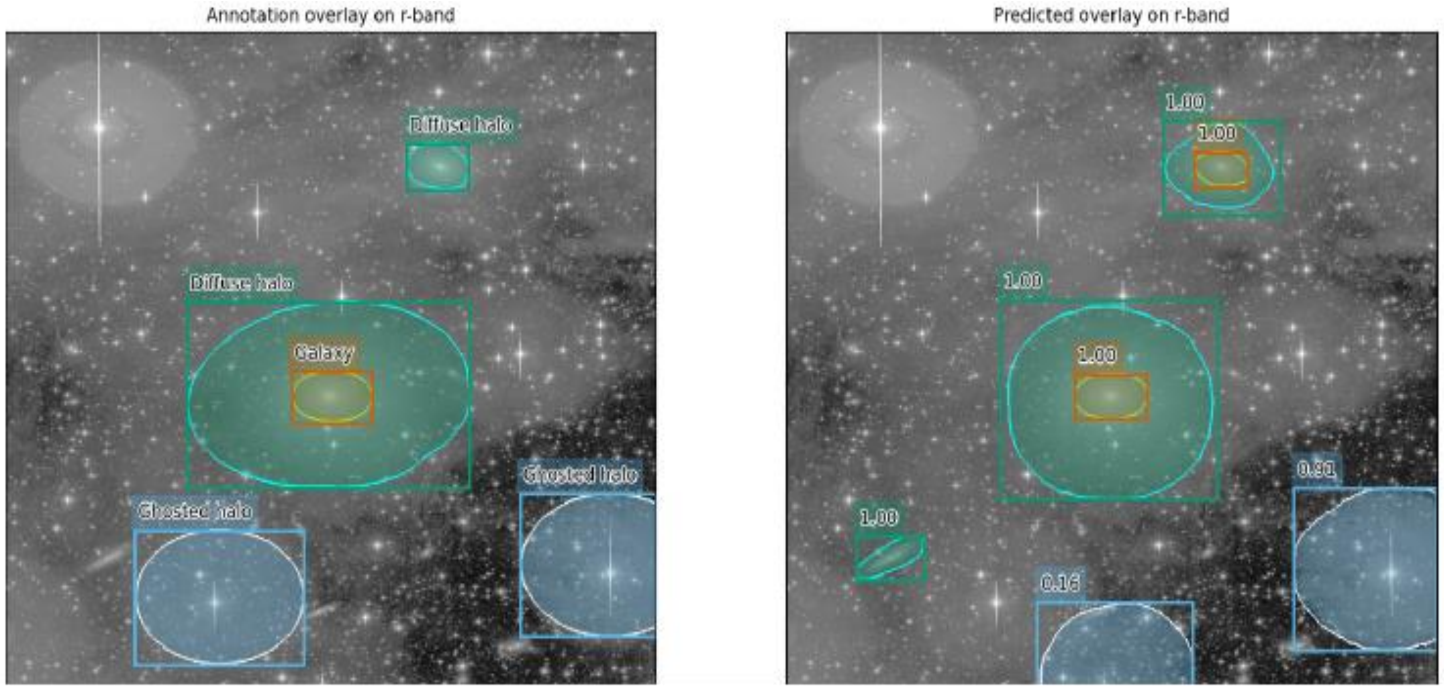
Uncertain cirrus segmentation labels (Richards et al 2022)



- Annotation **dataset** that can be used for **deep learning** algorithms
    - 186 MATLAS LSB images (6000px², two spectral channels)
    - On average 1.7 (std 0.9) galaxies annotated per image

Semi-automatic augmentation of the dataset

Human in the loop training

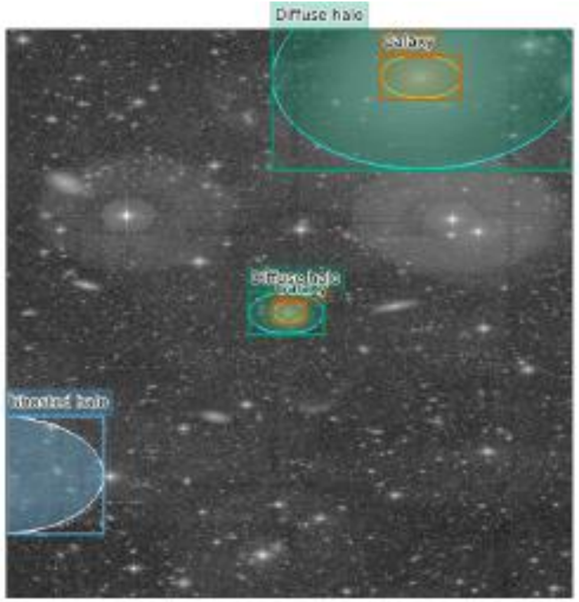


1. Manually select true detections that were not manually labelled
2. Re-train with more complete dataset
3. Repeat

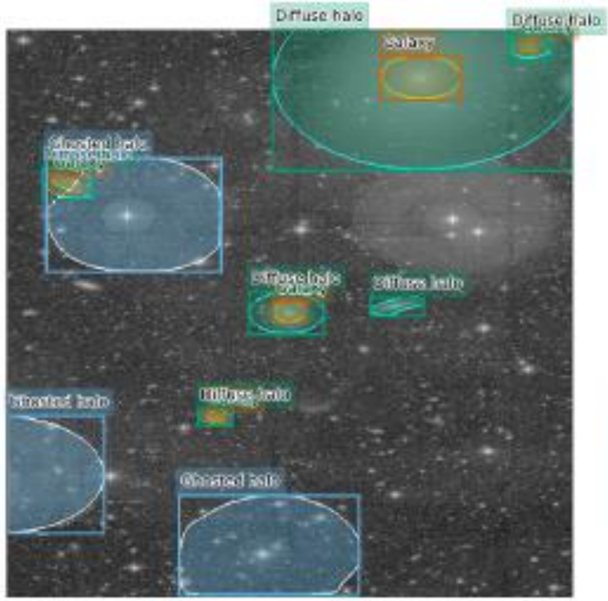
Semi-automatic augmentation of the dataset

Human in the loop training

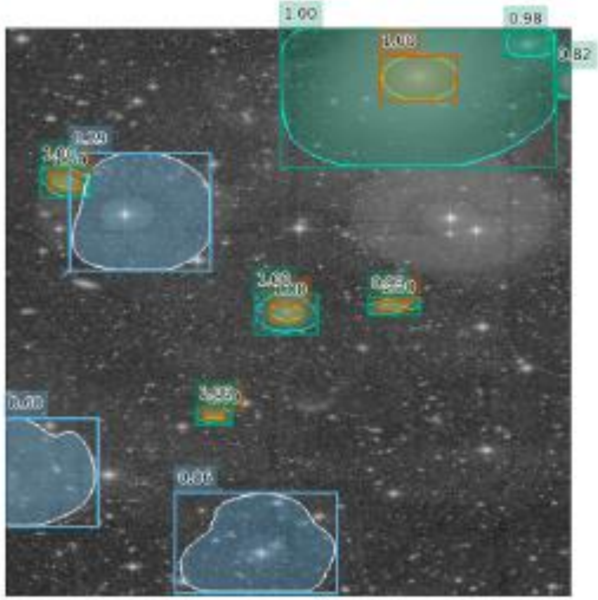
Example of results



Manual labels



Manual + HITL labels



Detections by the neural network

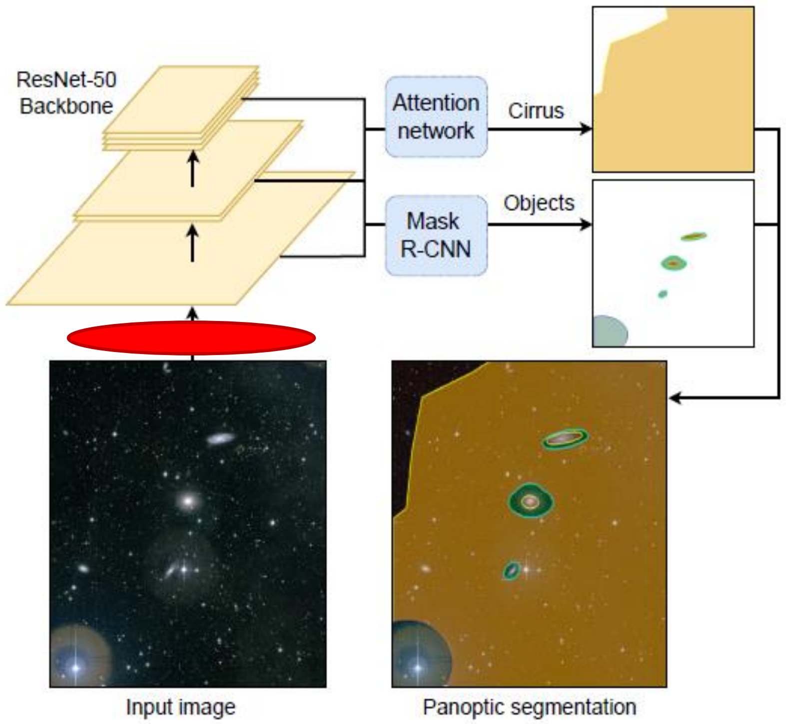
Red: galaxy
Green: extended halo
Dark blue: elongated tidal features (streams and tidal tails)
Light blue: ghosted halo (contaminant)

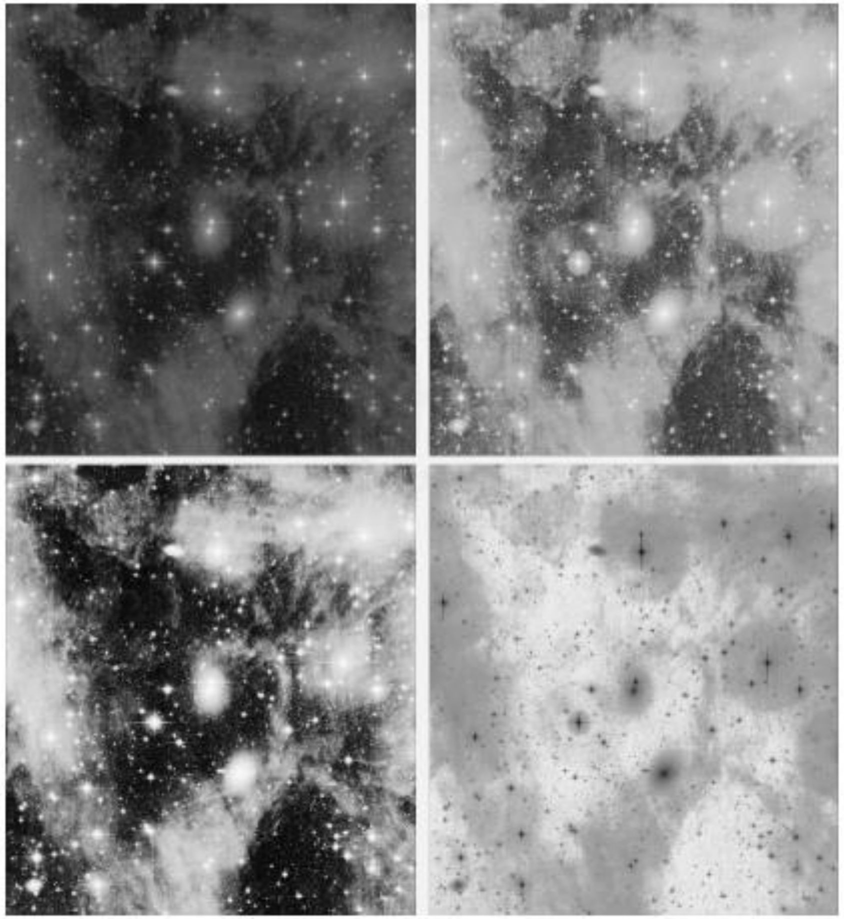
Step 2: Adapting to low surface brightness images

A new pre-processing layer that adaptively scales image intensity

$$X_s = \operatorname{arcsinh}(aX + b), \text{ where } a, b \in \mathbb{R} \text{ are learned}$$

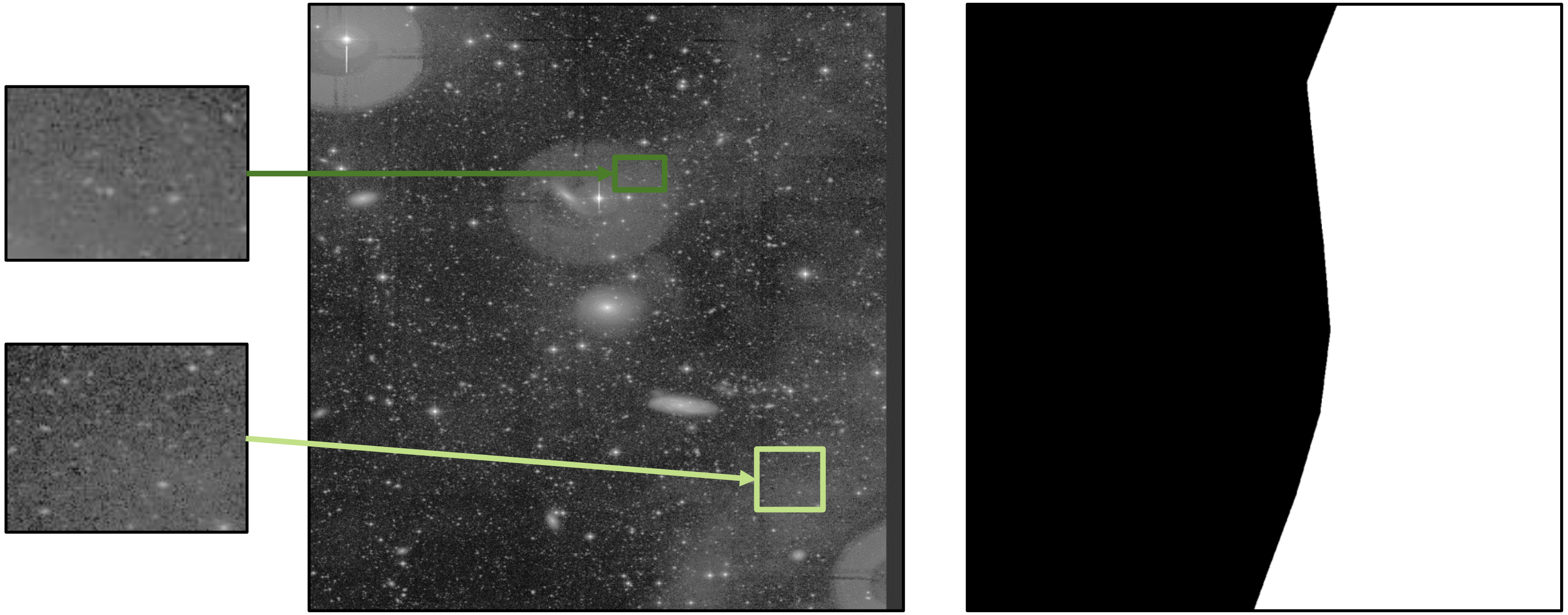
- Discovers the portions of the image's dynamic range that are relevant to identify and distinguish cirrus and tidal features





Step 3: Segmenting (out) cirrus contaminants

Precise segmentation of such structures requires ample global context alongside understanding of textural patterns

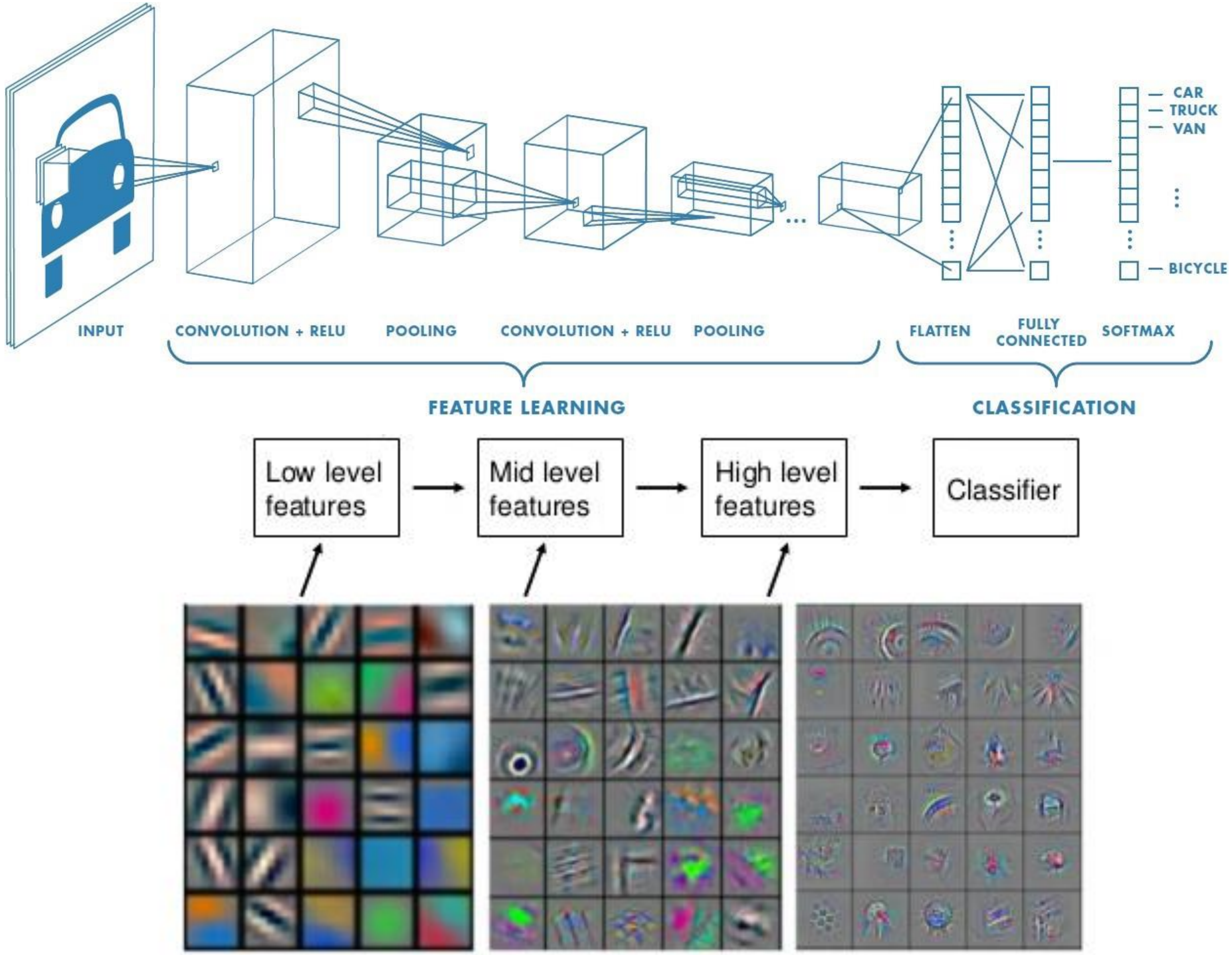


Comparison of localised regions (left), cirrus segmentation label (right)

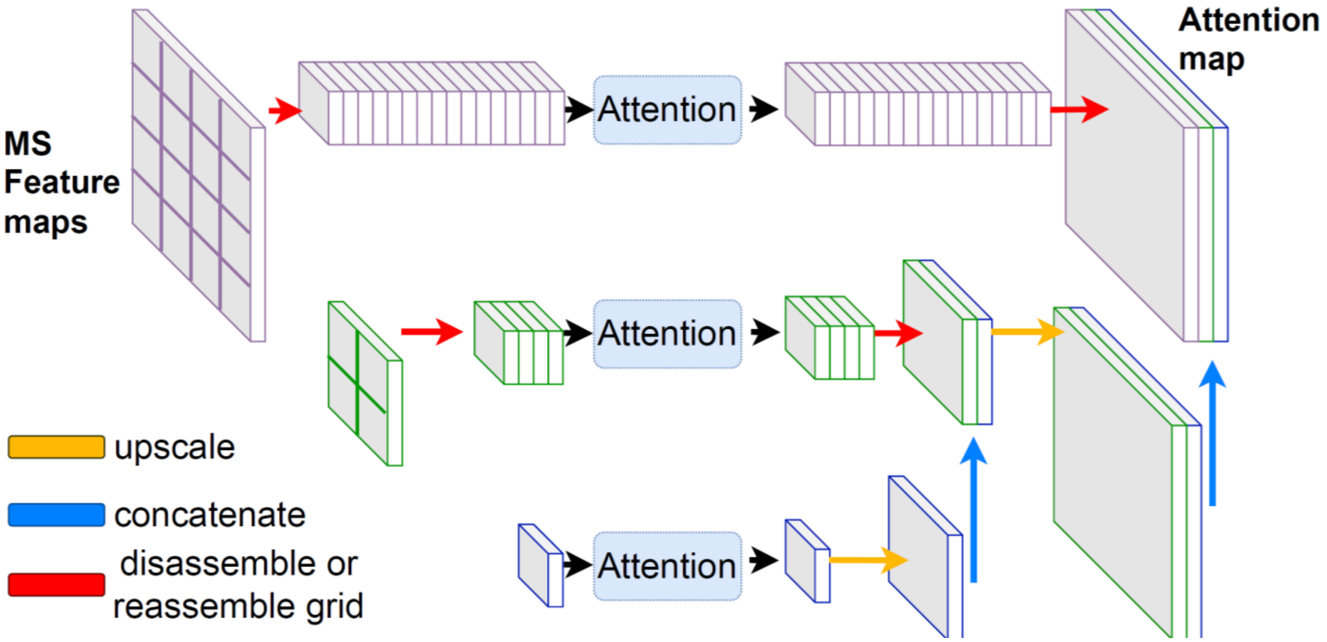
Need for purpose-designed deep neural networks

Problems:

- CNNs specialise in local textural patterns
- CNNs are not inherently sensitive to the (large-scale) orientation of texture



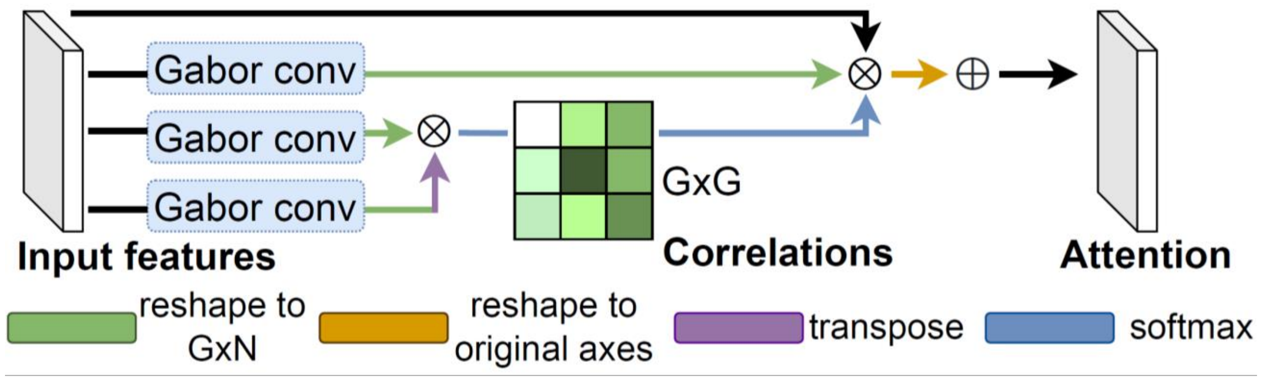
Gridded multi-scale attention



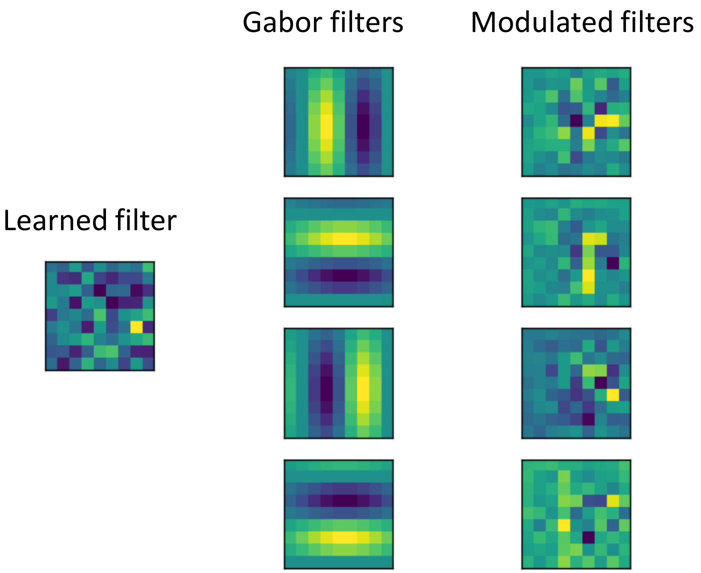
- Multi-scale analysis → Local and global context is assessed
- Multiple branches each handle a spatial scale and comprise of a separate attention module

- Additional benefit: computationally efficient for large images

Orientation-sensitive attention module



- We utilise Gabor modulated convolutions to generate image descriptors based on different angles.
- Attention is then computed across these angles, measuring correlations between orientation dependent descriptors.



Learned convolutional filter modulated with four Gabor filters of varying orientation.

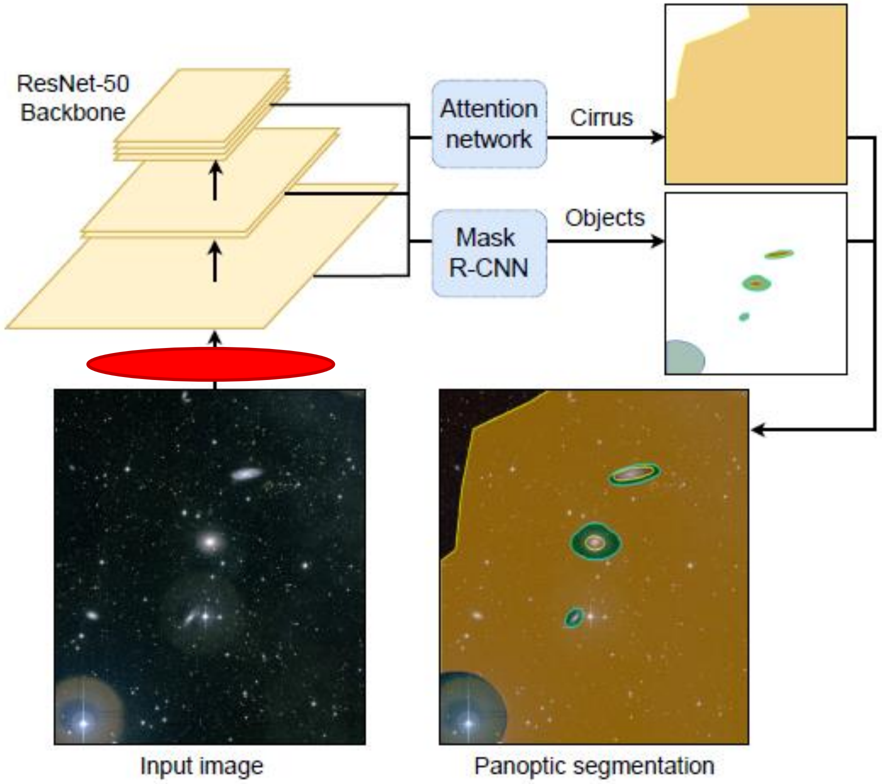
Final purpose-designed deep learning model

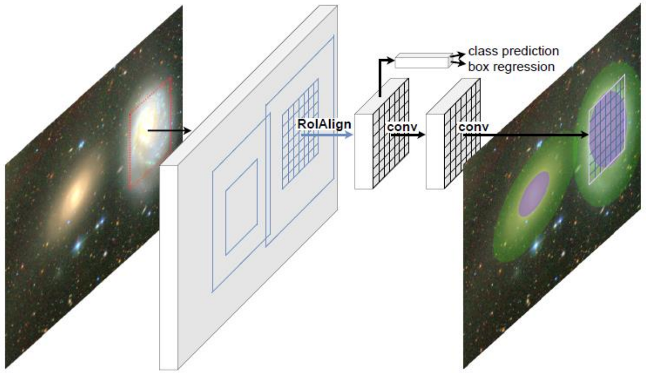
Multi-task detection and segmentation of:

- galactic features
- image contaminants

Adaptively scales image intensity

Exploits the oriented texture of dust clouds to improve their detection





Mask-RCNN

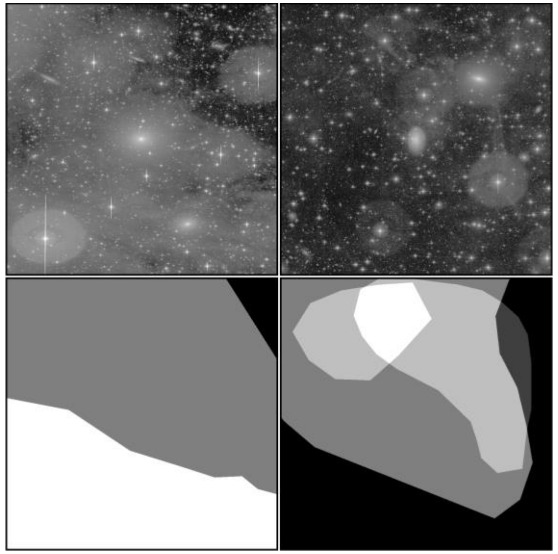
Step 4: Training with uncertain labels

Consensus between annotators is often not perfect

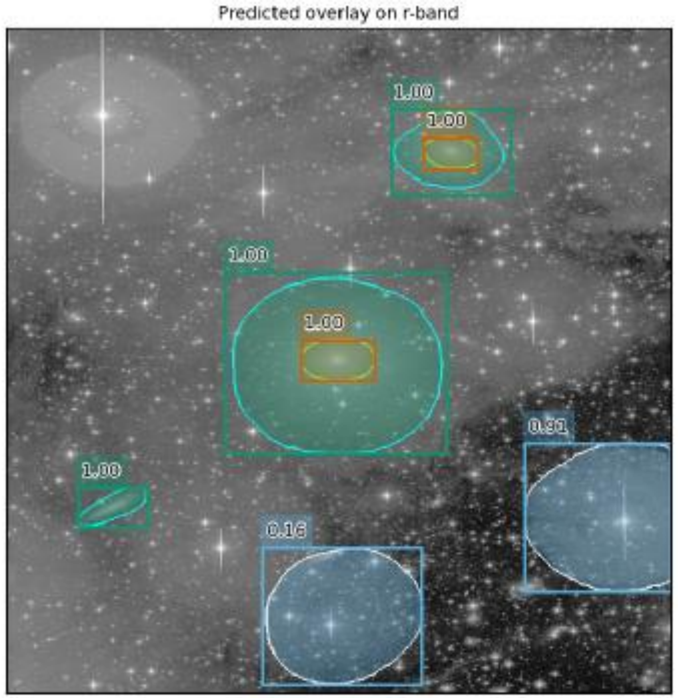
Labels may be considered as probabilistic:

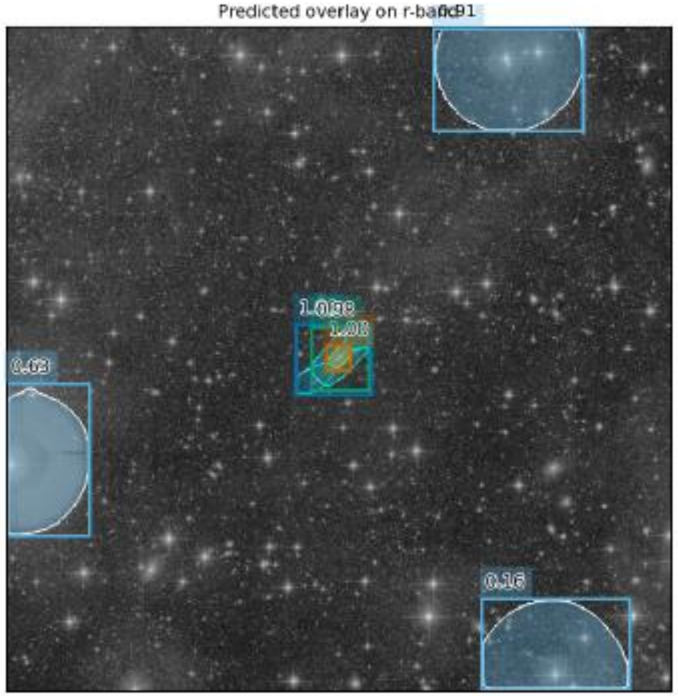
- Uncertain targets are ignored
- Super majority consensuses are prioritised with a boosting coefficient $\beta = 1.25$.
- Focal loss L_f encourages the model to focus on difficult examples.

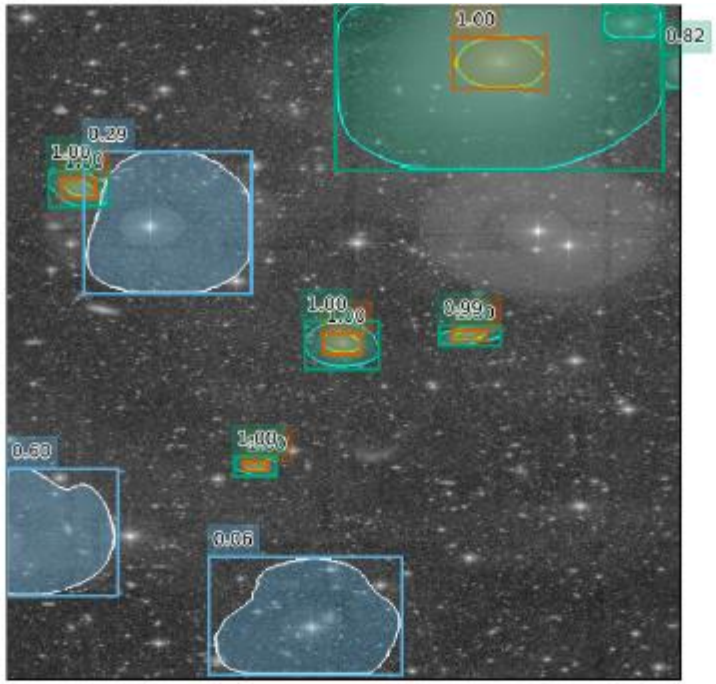
$$L_{\text{SML}} = \begin{cases} \beta \cdot L_f(x, y) & \text{if } y \geq 0.75. \\ L_f(x, y) & \text{if } 0.5 \leq y < 0.75. \\ 0 & \text{if } 0.25 < y < 0.5. \\ L_f(x, y) & \text{otherwise.} \end{cases}$$



Example results



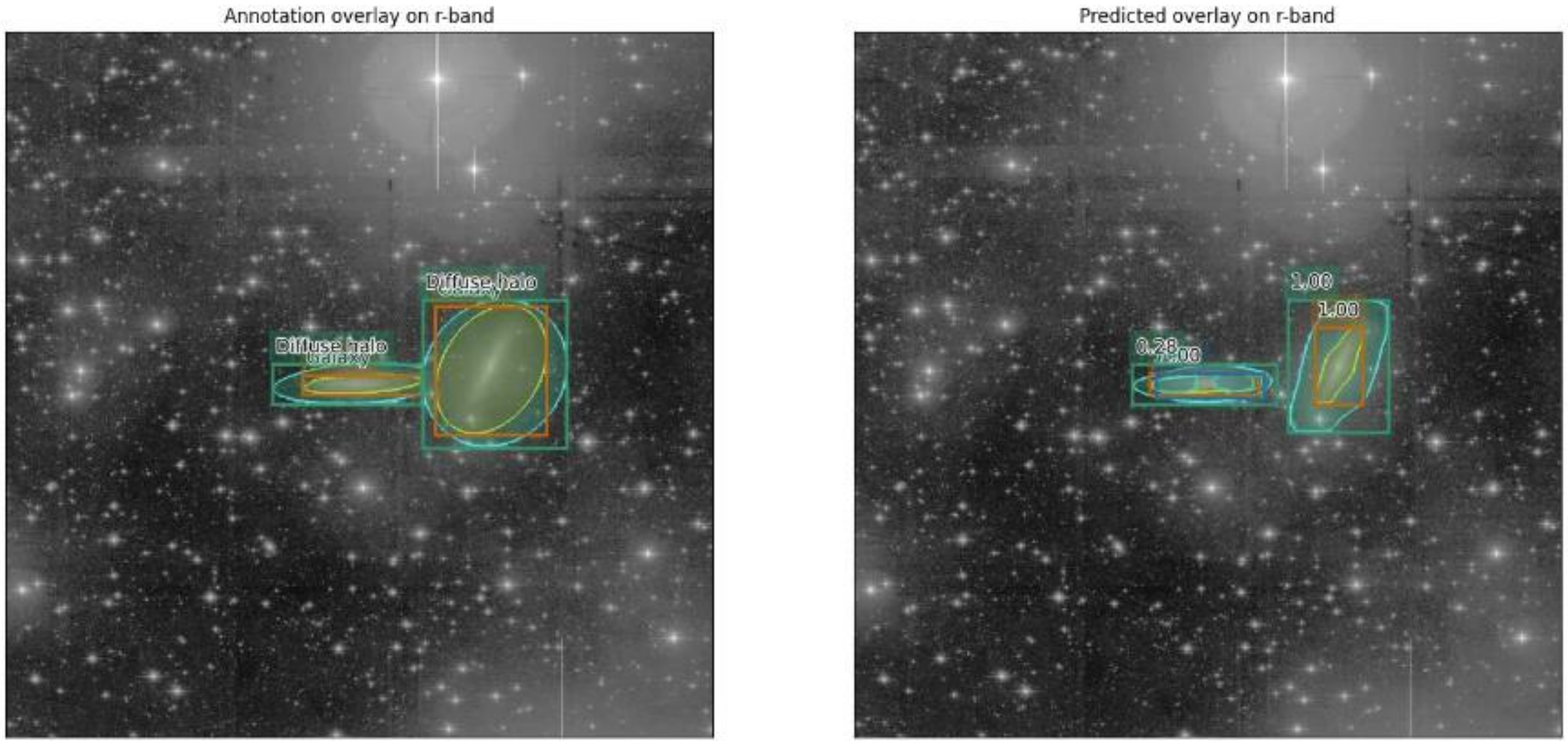




Red: galaxy
Green: extended halo
Dark blue: elongated tidal features
Light blue: ghosted halo (contaminant)

Example results

Refined annotations?

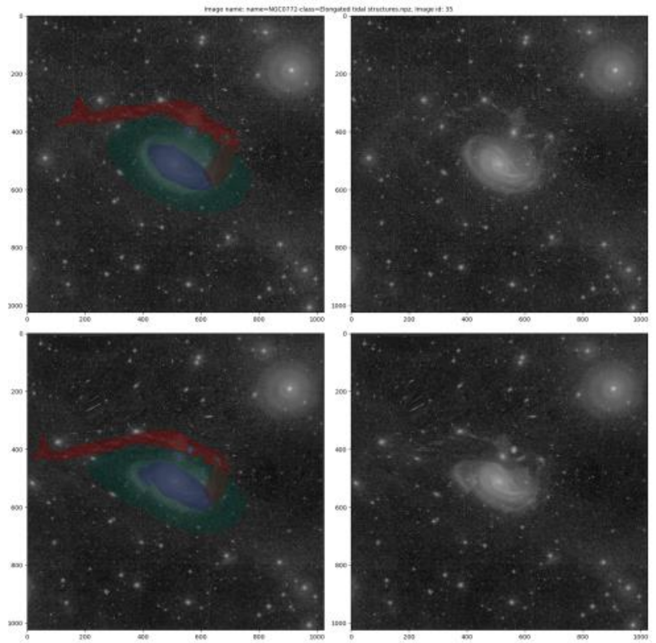


Next steps, work in progress

1. Increase and balance the size of the dataset
   - Annotation of more images
   - Data augmentation

Original

Deformed



Training set

- 'Diffuse halo': 144 images (80%)
- 'Galaxy': 147 images (79.89%)
- 'Elongated tidal structures': 45 images (81.81%)
- 'Ghosted halo': 117 images (80.68

Testing set

- 'Diffuse halo': 36 images (20%)
- 'Galaxy': 37 images (20.1%)
- 'Elongated tidal structures': 10 images (18.18%)
- 'Ghosted halo': 28 images (19.31%)

2. Structured analysis to exploit known relations
   - Correlated attributes
   - Logical constraints of relative locations

→ Hierarchical loss function: $p(A, B) = p(B|A) \cdot p(A)$

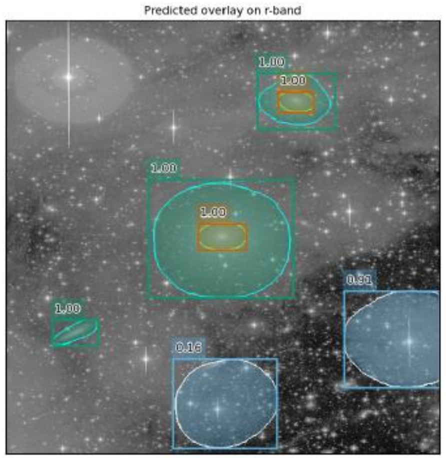
Visible

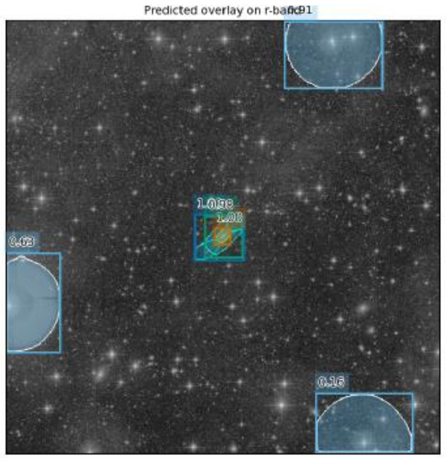
Value

Probability of the attribute value: softmax loss

Probability that attribute is visible: logistic loss

   - Detection of dwarf companions
     - Comparing apparent resolutions to determine depth relationships





Thank you for your attention