

Domain-informed analysis of astrophysics data

Adeline Paiement, LIS, Université de
Toulon

Introducing myself...



Swansea
University
Prifysgol
Abertawe

2016-2018 : Research lecturer

Department of Computer Science,
Visual Computing research group



© 2008 Microsoft Research Corporation. All rights reserved.

2013-2016 : Postdoc

- Machine learning
- Internet-of-things



University of
BRISTOL

2009-2013 : PhD

- Computer vision
- Medical image analysis



SCISYS

2008-2009 : Developer

- Robotics
- Computer vision
- Mars exploration rover



Observatoire astronomique
de Strasbourg

2008 : Engineering degree +
astrophysics MSc



UNIVERSITÉ
DE TOULON

2018 - : Maître de conférences

Laboratoire d'Informatique et des Systèmes
(LIS), pôle Science des Données



Physics data analysis

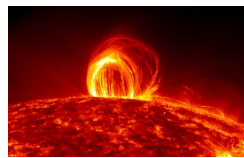
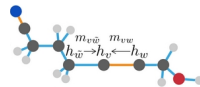
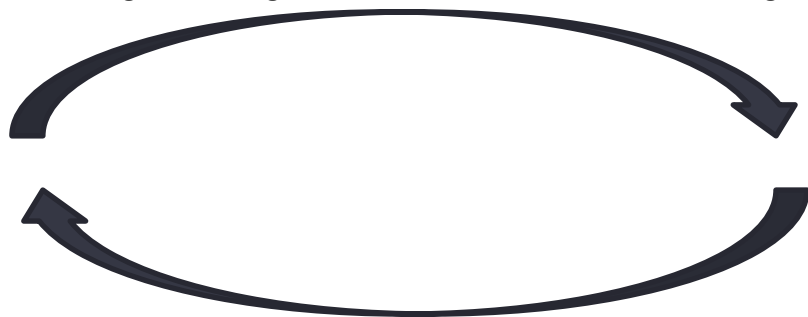
Specificities and requirements

AI and physics synergy

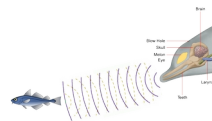
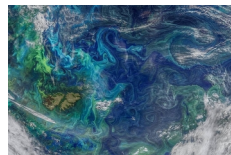
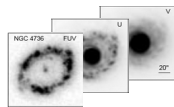
New tools:

- Analysing observations: large and heterogeneous data
- Modelling: emulating numerical simulations, data modelling

Machine and deep learning
Computer vision
Data science



Physics
Astrophysics

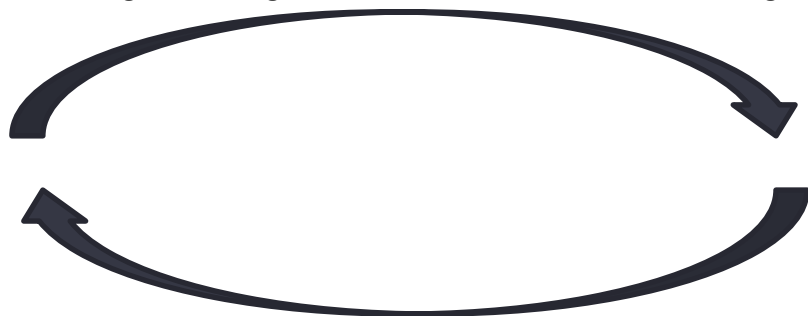


AI and physics synergy

New tools:

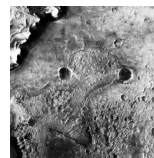
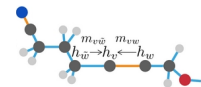
- Analysing observations: large and heterogeneous data
- Modelling: emulating numerical simulations, data modelling

Machine and deep learning
Computer vision
Data science

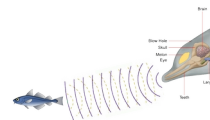
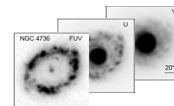


New stimulating challenges:

- New data and tasks
- Integration of domain knowledge



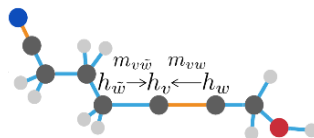
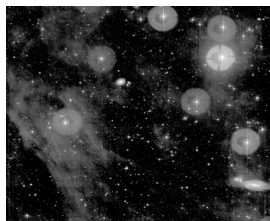
Physics
Astrophysics



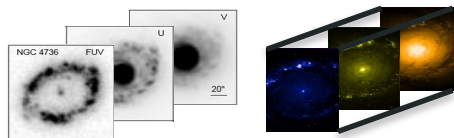
Various possible tasks

Spatial analysis:

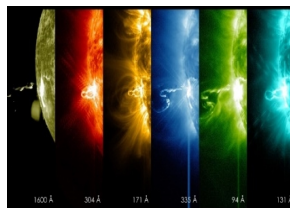
- Localisation
- Characterisation: classification
- Characterisation: properties of 2D and 3D geometry



Quantum chemistry



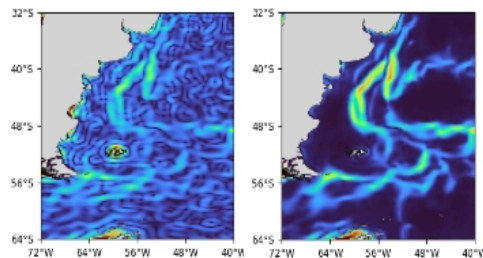
Galaxy evolution



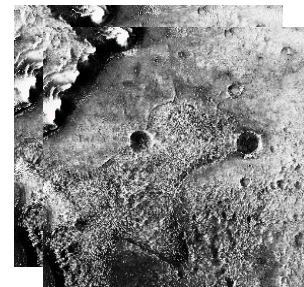
Solar activity

Temporal analysis: evolution modelling

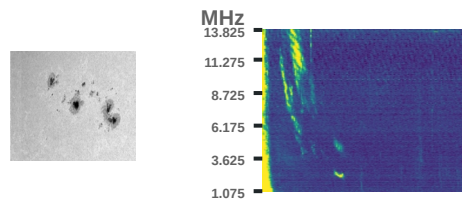
- Characterisation of evolution
- Prediction



Oceanography



Martian geology



Challenging data

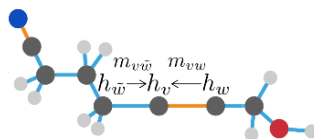
Multimodal data

- Fusion of information
- Different resolutions
- Alignment (registration)

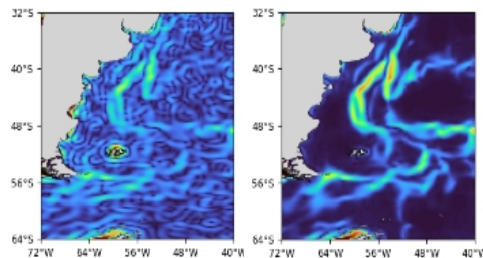
Scientific vs. natural images/data

- Data quality
- Expert interpretation
- Ground-truth availability

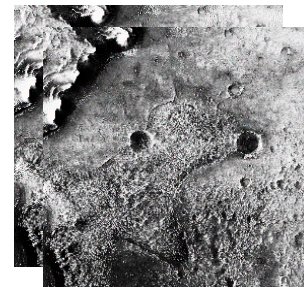
→ **Purpose-designed algorithms**



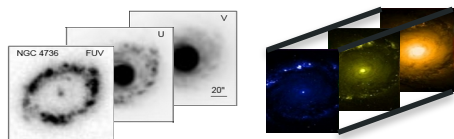
Quantum chemistry



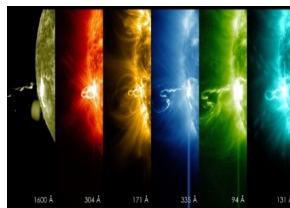
Oceanography



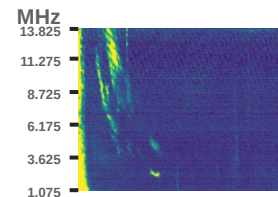
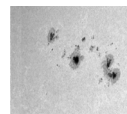
Martian geology



Galaxy evolution



Solar activity



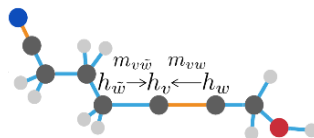
Requirements for science applications

Robustness

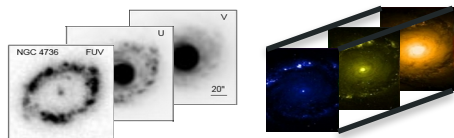
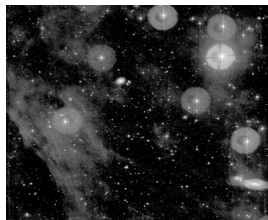
- to small annotated training sets
- Physical validity of solutions

Interpretability

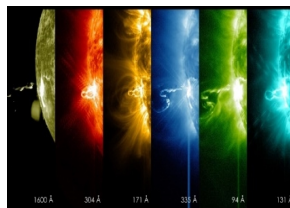
→ **Domain-informed analysis**



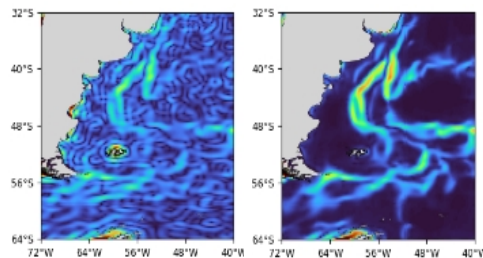
Quantum chemistry



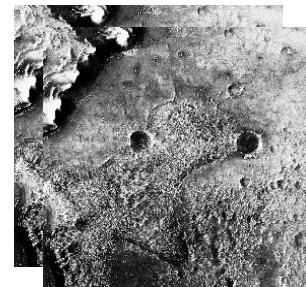
Galaxy evolution



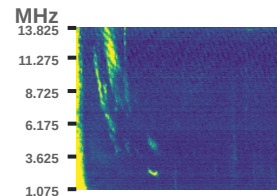
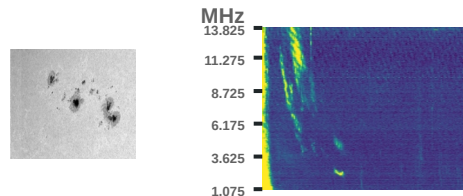
Solar activity



Oceanography



Martian geology



Domain informed-analysis

Current main paradigm:

Deep learning

Large learning models (foundation models)

- High **complexity** + large **datasets**
- Flexible
- Resource hungry → GAFAM

Large **variety** of available knowledge

- On signals & data
- On studied phenomena
- Equations?

Invariances,
uncertainties,
(ir)relevant information,
sub-tasks, etc...

Proposed paradigm:

Accounting for existing domain knowledge

- Adapting networks' architectures
- Adapting their training process
- No universal solution: based on available knowledge

type

I. Main current works in astrophysics

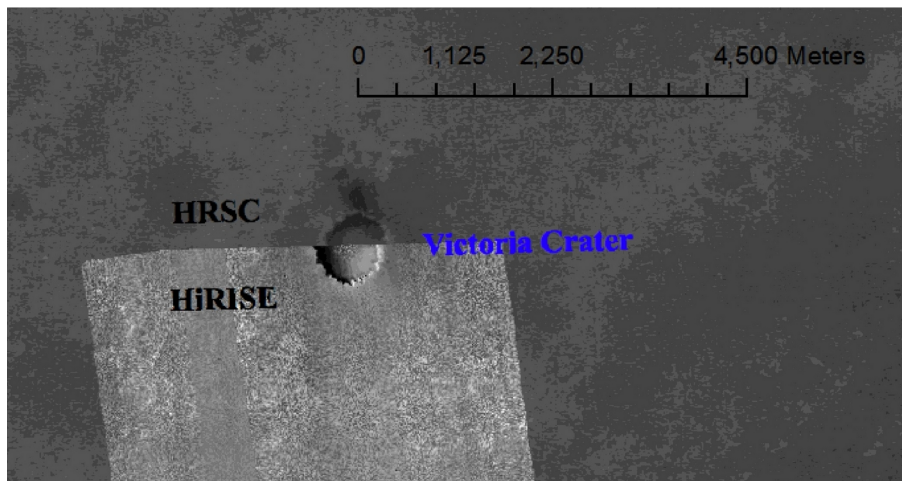
II. Dealing with sparse and inconsistent data annotations

- Fusion and modelling from Martian remote sensing data
- Characterisation of galaxy morphology and collision debris
- Monitoring of solar activity



Spatial misalignments in Martian data

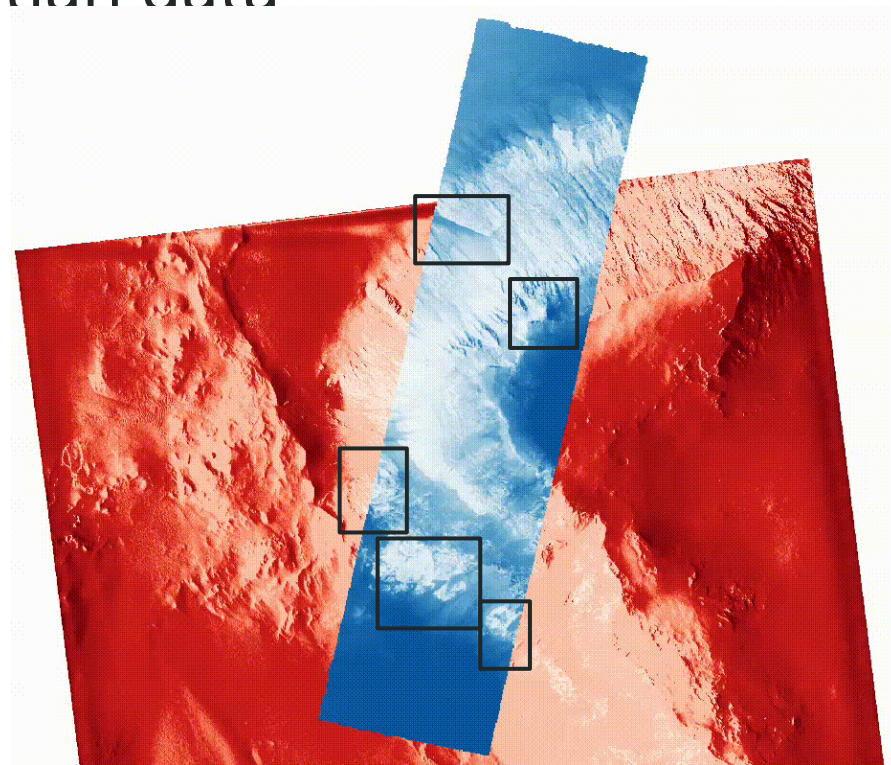
Multi-sensors



@Tao et al. 2016

Common challenges:

- lack of salient features
- heterogeneous sampling

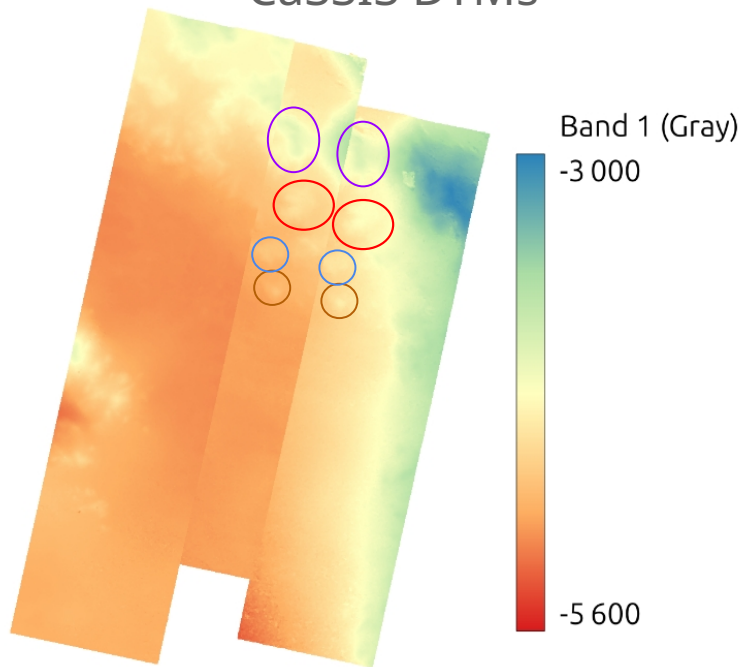


CaSSIS and CTX

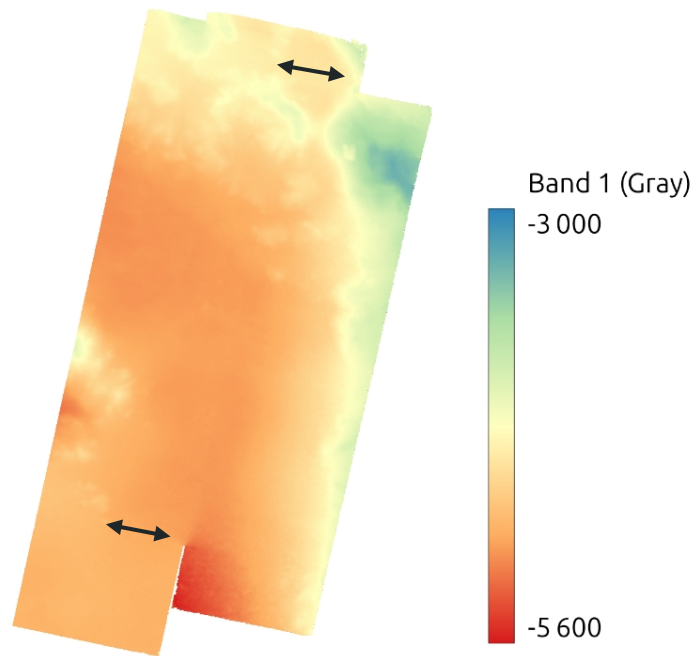
Spatial misalignments in Martian data

Same sensor

CaSSIS DTMs

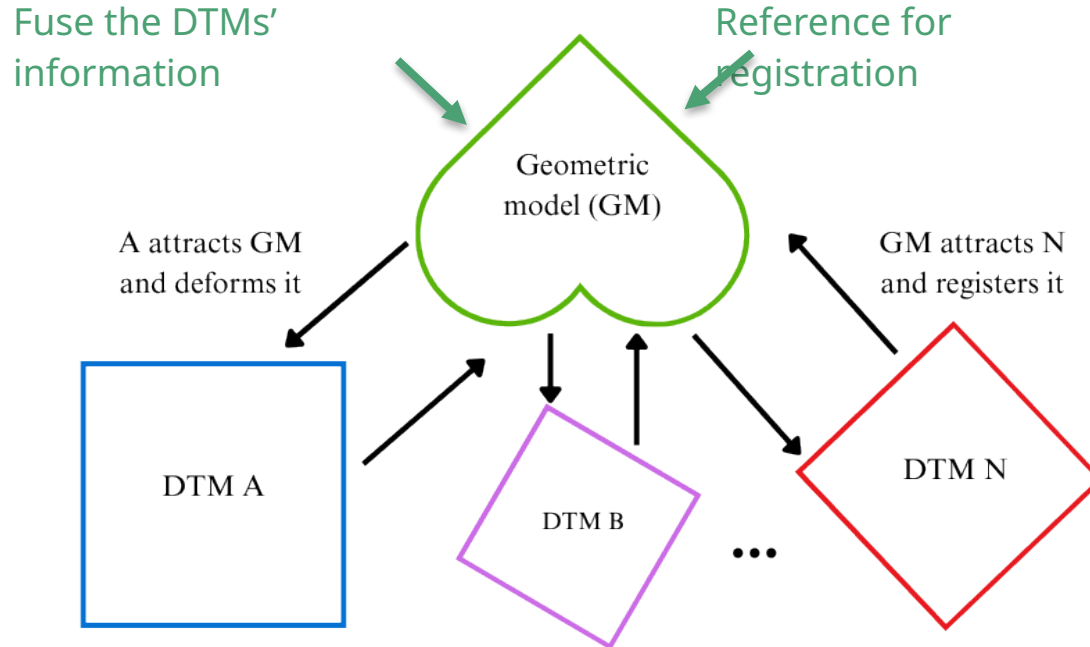


Aligned



Joint registration and fusion of DTMs

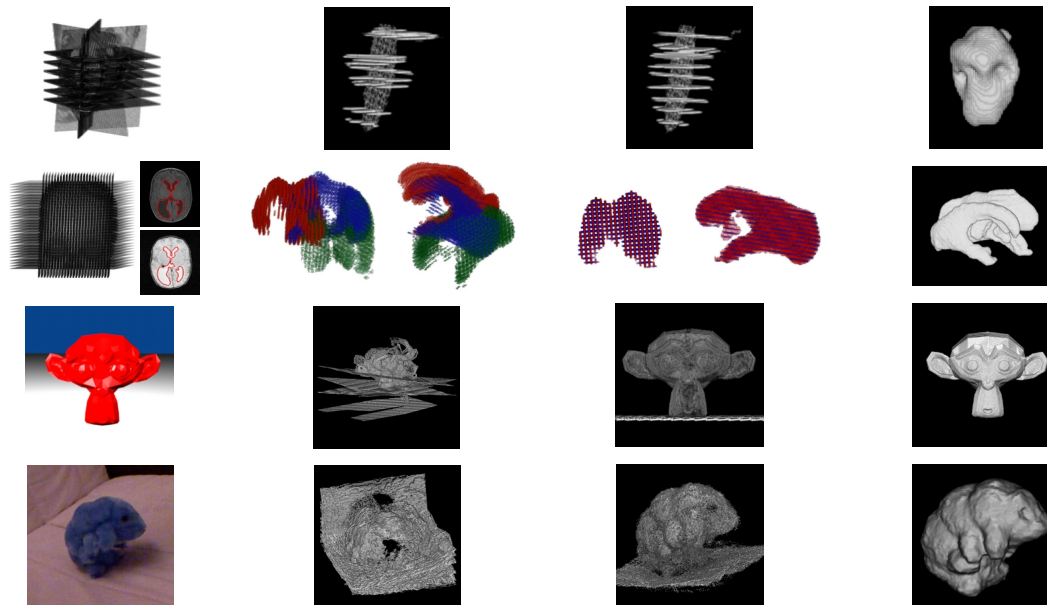
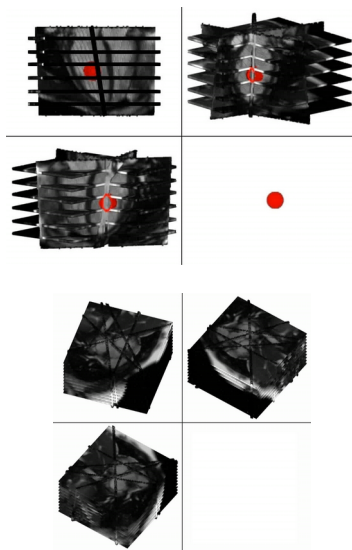
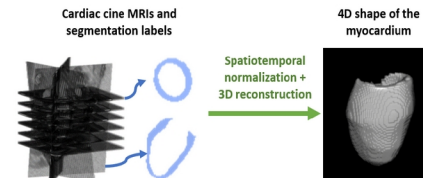
Digital Terrain Models → use 3D information



Fused 3D geometric model

Illustration with easier volumetric data

Modelling from multimodal data with heterogeneous resolutions, misalignments, and gaps



Fused 3D geometric model

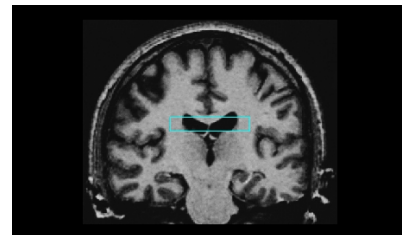
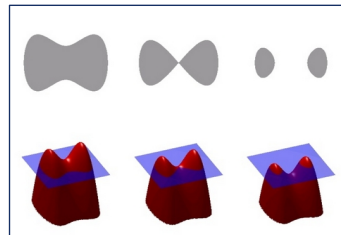
Joint registration, segmentation, and interpolation

- Segmentation: based on the level set method

$$\frac{d\Phi}{dt}(x) = F(x) |\nabla \Phi(x)|$$

Level set function

Speed of contour



- Interpolation: mathematical reformulation of the level set method

- Registration: use 3D contour information

Fused 3D geometric model

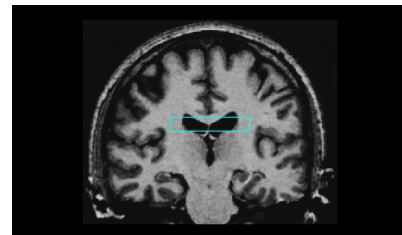
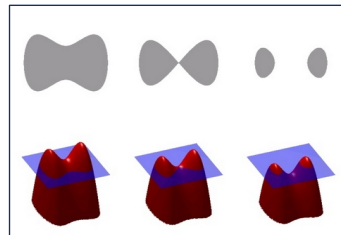
Joint registration, segmentation, and interpolation

- Segmentation: based on the level set method

$$\frac{d\Phi}{dt}(x) = F(x) |\nabla \Phi(x)|$$

Level set function

Speed of contour

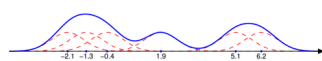
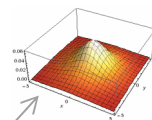


- Interpolation: mathematical reformulation of the level set method

$$\Phi(\mathbf{x}) = \sum_{i=1}^N \alpha_i \cdot \Psi_i(\mathbf{x})$$

Interpolation coef.

Radial basis function



$$\frac{d\alpha_i}{dt} = \sum_c F(\mathbf{x}) \Psi_i(\mathbf{x}) = (F * \Psi)(\mathbf{x}_i)$$

- Registration: use 3D contour information

Fused 3D geometric model

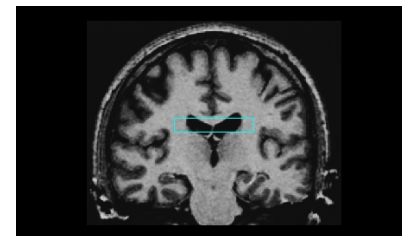
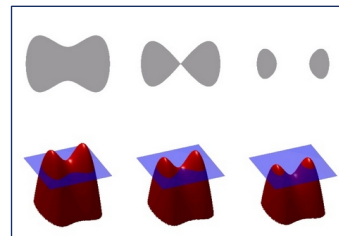
Joint registration, segmentation, and interpolation

- Segmentation: based on the level set method

$$\frac{d\Phi}{dt}(x) = F(x) |\nabla \Phi(x)|$$

Level set function

Speed of contour

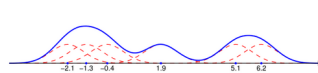
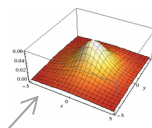


- Interpolation: mathematical reformulation of the level set method

$$\Phi(\mathbf{x}) = \sum_{i=1}^N \alpha_i \cdot \Psi_i(\mathbf{x})$$

Interpolation coef.

Radial basis function

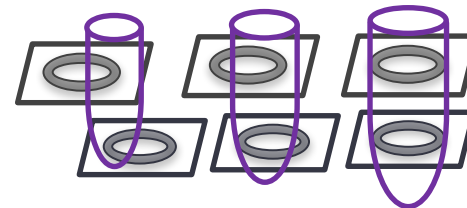
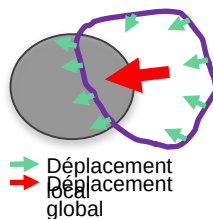


$$\frac{d\alpha_i}{dt} = \sum_c F(\mathbf{x}) \Psi_i(\mathbf{x}) = (F * \Psi)(\mathbf{x}_i)$$

- Registration: use 3D contour information

$$\frac{\partial g_i^n}{\partial t} = \int F_n \circ g^n \left\langle \frac{\partial g^n}{\partial g_i^n}, \vec{N} \right\rangle$$

Image (rigid) transformation



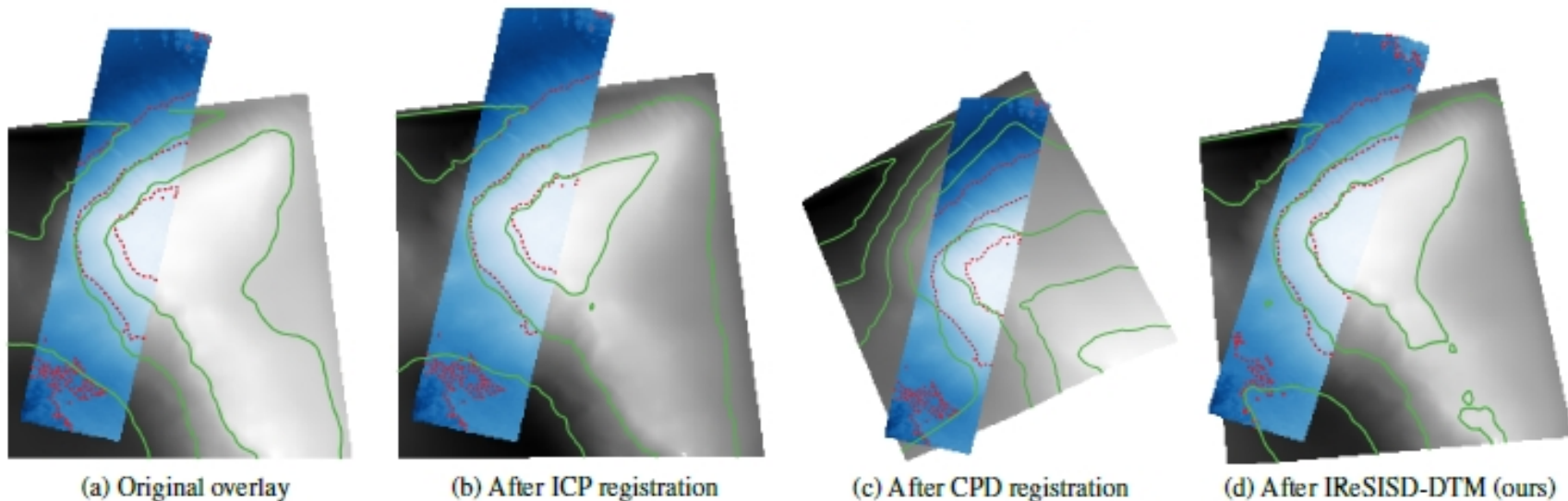
Some early results

PhD project of Kevin Kheng and Loïs Brun

Collaboration with

- IPAG
 - INAF - Osservatorio Astronomico di Padova
- Funder: Region Sud

CaSSIS (blue, 12 m/pixel) and CTX (gray, 6 m/pixel) DTMs



Perspectives: terrain characterization through spectral unmixing and micro-texture classification

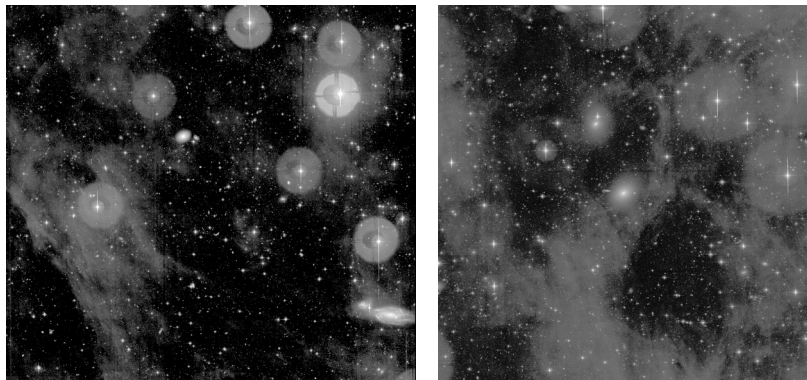
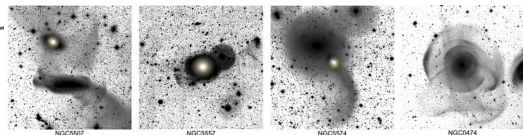
ICP: K. S. Arun, T. S. Huang, and S. D. Blostein. **Least-squares fitting of two 3-D point sets**. IEEE TPAMI, 1987

CPD: A. Myronenko and X. Song. **Point-set registration: Coherent point drift**. IEEE TPAMI, 2010

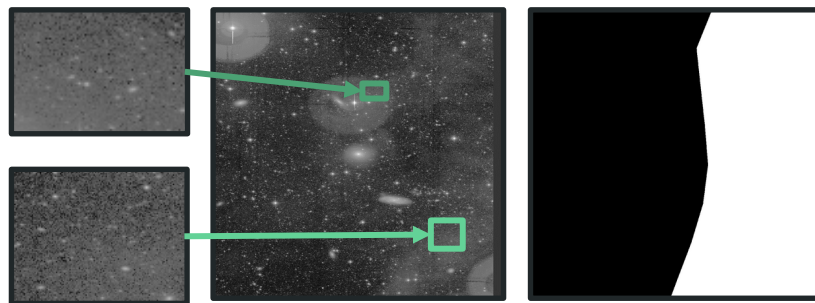
IReSISD-DTM: L. Brun, S. Douté, A. Palement. **3D registration of remote sensing data for planetary exploration**. CVPR-W 3D4Science, under review

Characterisation of galaxy morphology and collision debris

Analysing galaxy morphology in the presence of image contaminar



Example of MATLAS CFHT images contaminated with **cirrus** and **ghost halos**



Challenge: identifying contaminants requires:

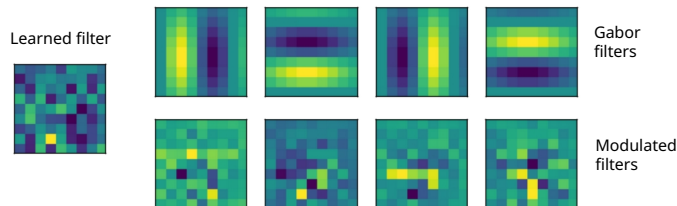
1. Global context
- +
2. Understanding textural patterns

Understanding textural patterns

Constraining learnt representations

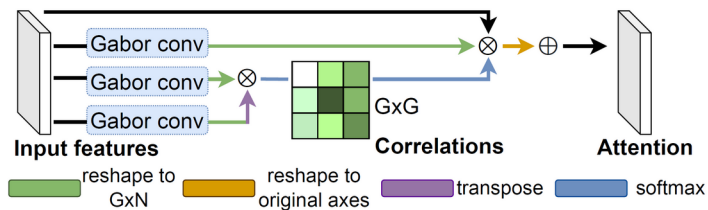
- Features with orientation sensitivity [1]

Learned convolutional filter modulated with Gabor filters of various (learnt) orientations

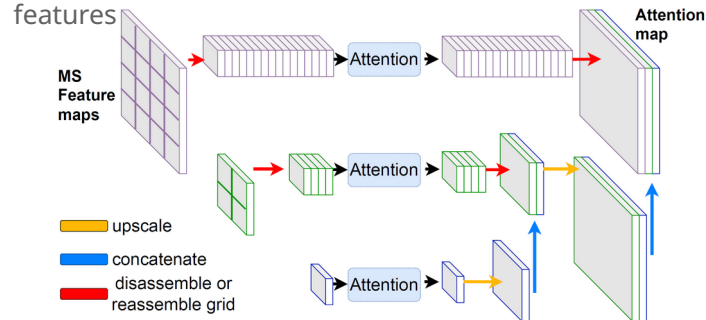


Exploiting relationships between learnt representations [2]

- Attention that considers correlations between orientations
- Exploiting short- and long-range dependencies between features



- Combined with spatial attention (correlations between locations) and channel attention (correlations between features)

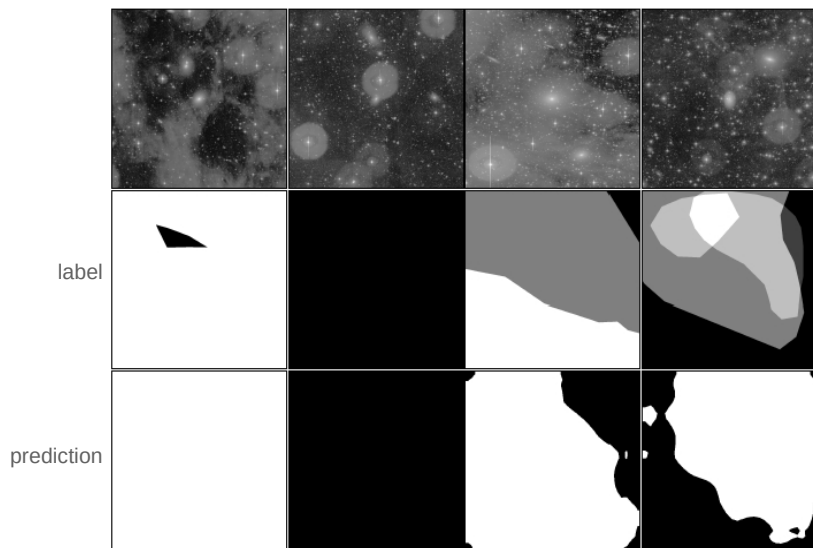


[1] F. Richards, A. Paiement, X. Xie, E. Sola, P.-A. Duc: **Learnable Gabor modulated complex-valued networks for orientation robustness**. Under review with *IET Computer Vision*

[2] F. Richards, E. Sola, A. Paiement, X. Xie, P.-A. Duc: **Multi-scale gridded Gabor attention for cirrus segmentation**. *ICIP*, 2022

Identifying contaminants

Detecting cirrus contamination in images of galaxies



Cirrus segmentation IoU:

	simple backbone	
	All images	Cirrus only
Control	0.469(0.005)	0.814(0.013)
SML	0.483(0.006)	0.822(0.013)
SML+Gabor	0.497(0.007)	0.861(0.015)
SML+Grid	0.542(0.003)	0.886(0.016)
All	0.548(0.003)	0.892(0.013)
U-Net [17]	0.381(0.128)	0.685(0.096)
LGCN [16]	0.414(0.049)	0.741(0.033)

ResNet50 backbone

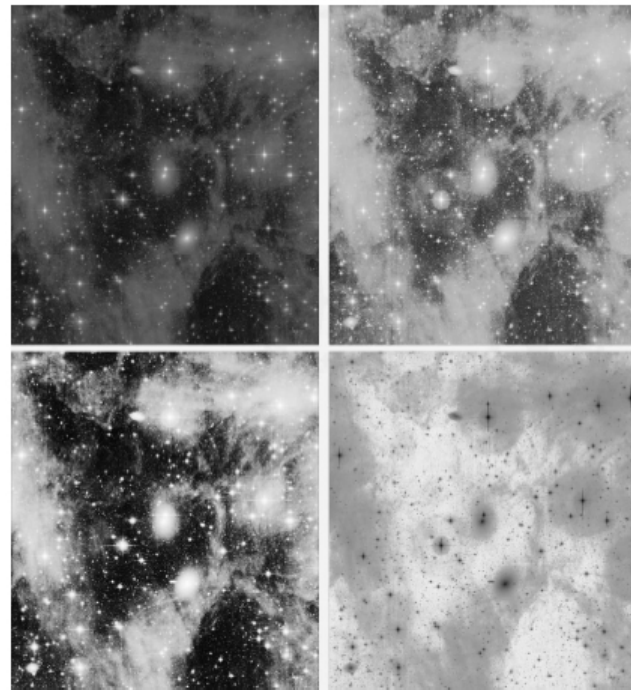
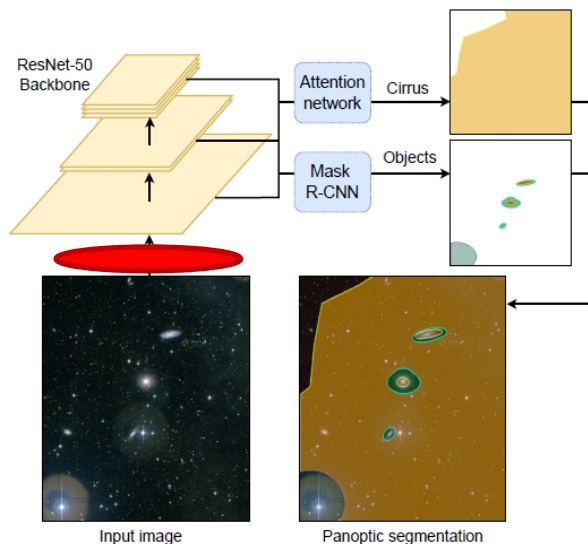
IoU	0.790
Dice	0.814

Adapting to low surface brightness images

A new pre-processing layer that adaptively scales image

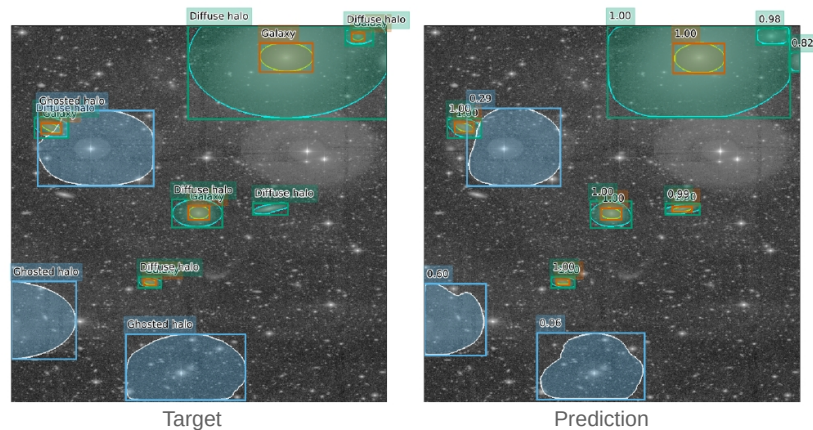
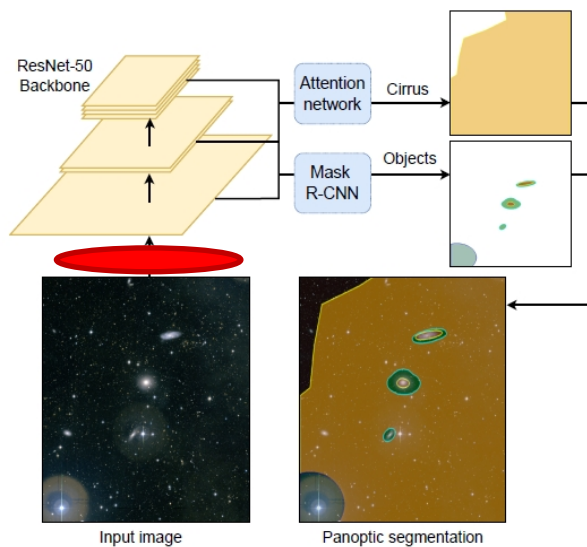
into $X_s = \text{arcsinh}(aX + b)$, where $a, b \in \mathbb{R}$ are learned

Discovers the portions of the image's dynamic range that are relevant to identifying tidal features



Framework for combined detections

- Sharing features between related tasks



		Galaxy	Diffuse halo	Tidal structures	Ghosted halo	All
Panoptic	AP ₅₀	0.782	0.788	0.000	0.658	0.543
	AP ₇₅	0.217	0.411	0.000	0.658	0.330
Instance	AP ₅₀	0.778	0.754	0.000	0.522	0.514
	AP ₇₅	0.210	0.384	0.000	0.490	0.271

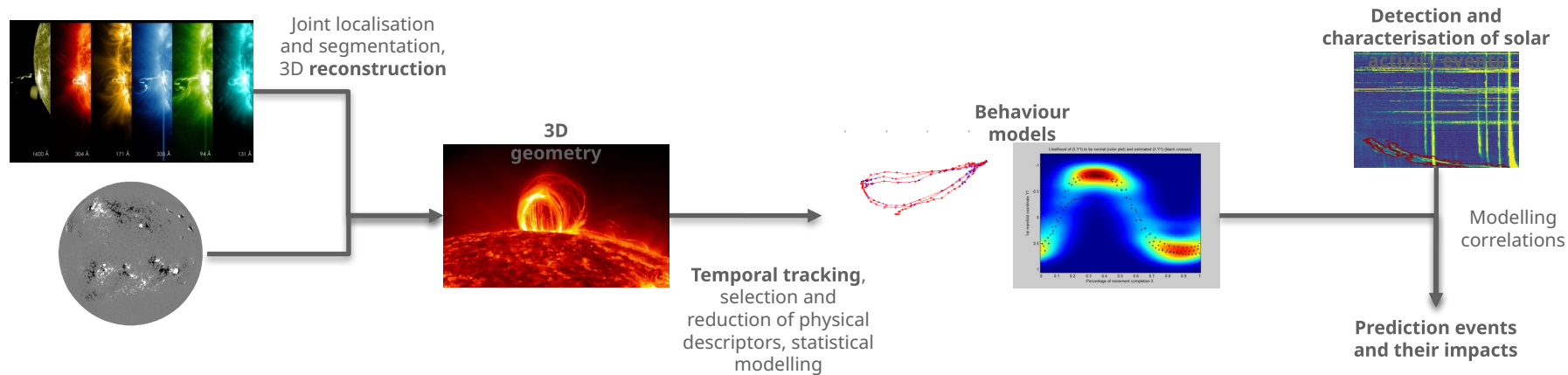
- Better detections of galactic structures
- Better segmentation of cirrus (IoU 0.75 → 0.86)

Ongoing and future work:

- More work on tidal structures, focus on dwarf galaxies
- Euclid data

Monitoring solar activity

Overview of long-term project



PhD project of Majedaldein Almahasneh and Maxime Gressier

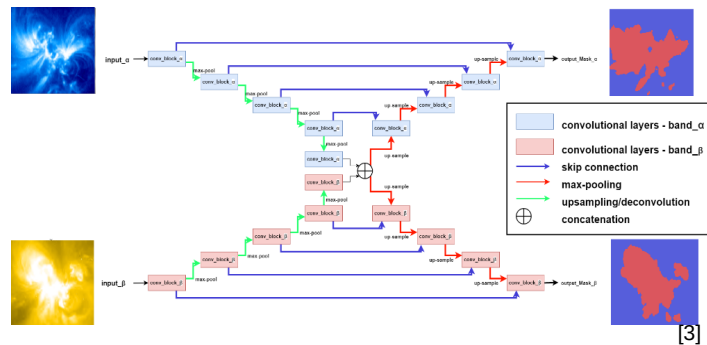
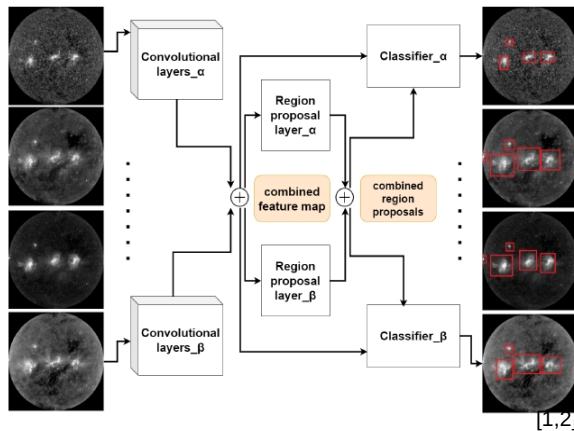
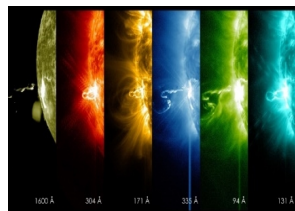
Mres project of Joseph Jenkins

Collaboration with

- Observatoire de Paris
- NICT, Japan

Funder: ANR

Detecting 3D active regions from combined 2D image slices



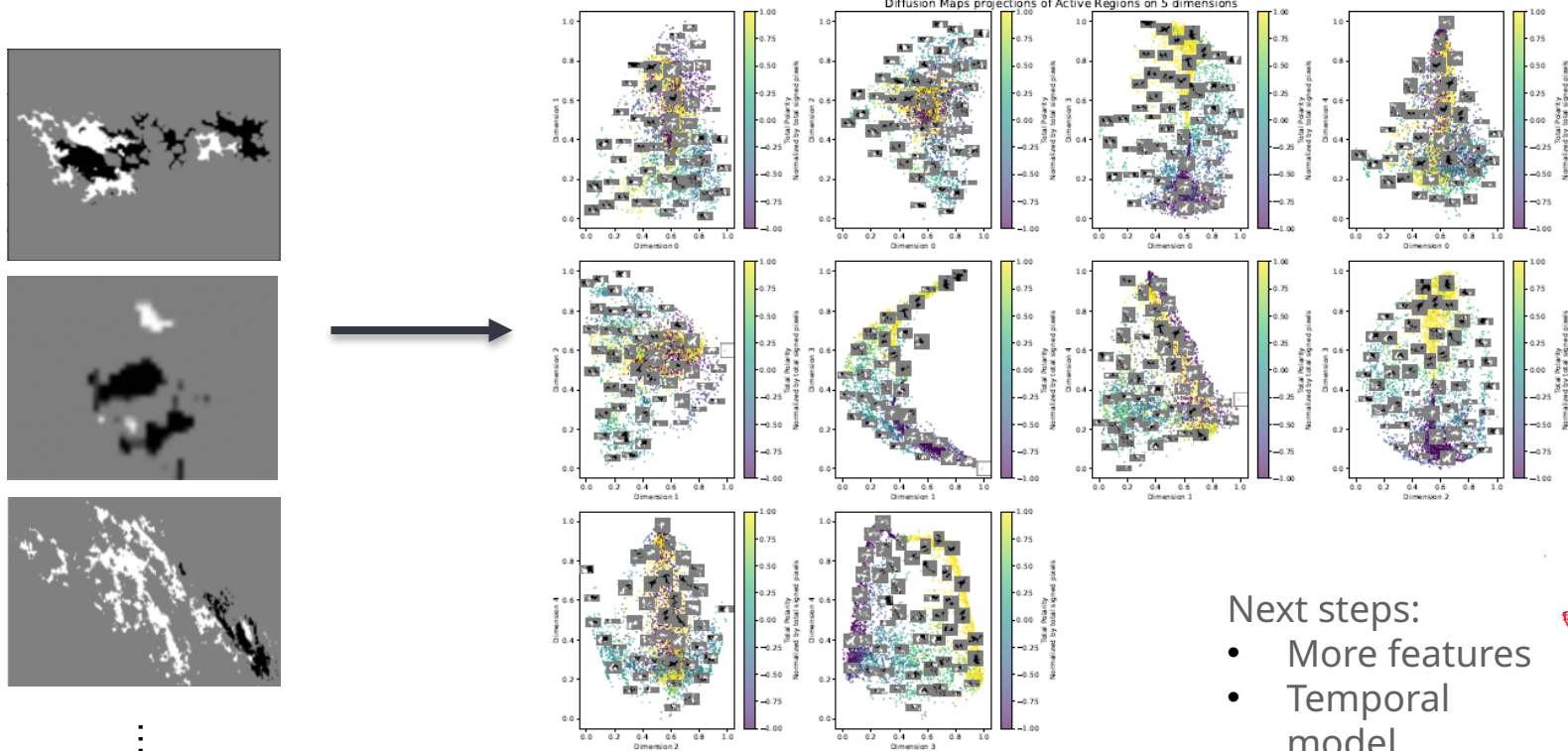
[1] Almahasneh, Paiement, Xie, Aboudarham: Active region detection in multi-spectral solar images, ICPRAM 2021

[2] Almahasneh, Paiement, Xie, Aboudarham: MSMT-CNN for solar active region detection with multi-spectral analysis, Springer Nature Computer Science 2021

[3] Almahasneh, Paiement, Xie, Aboudarham: MLMT-CNN for object detection and segmentation in multi-layer and multi-spectral images, Machine Vision Applications 2021

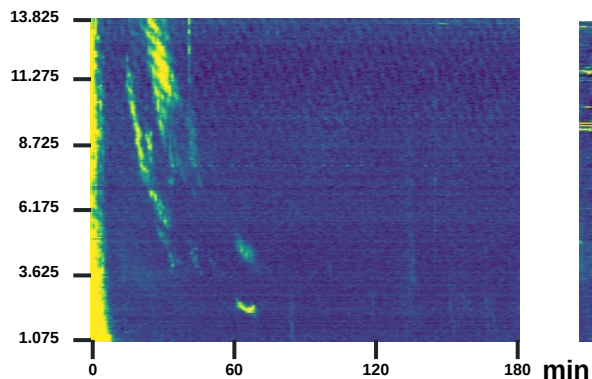
Modelling the temporal evolution of active regions

Simplified representation of the magnetic polarity geometry

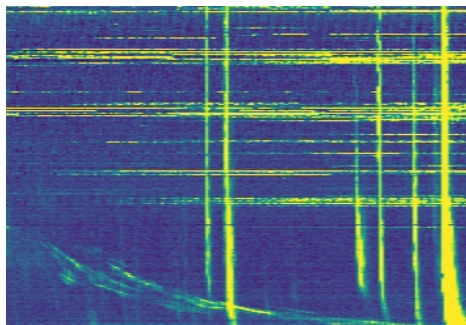


Detecting type II solar radio bursts

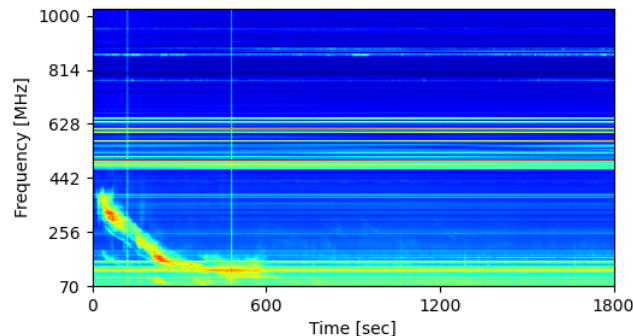
MHz



Wind/WAVES



Yamagawa



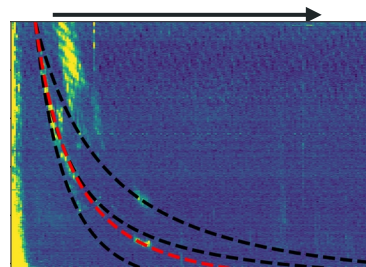
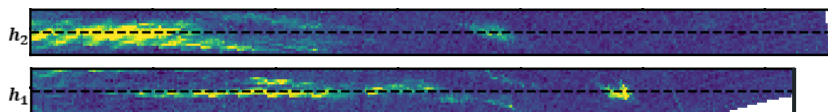
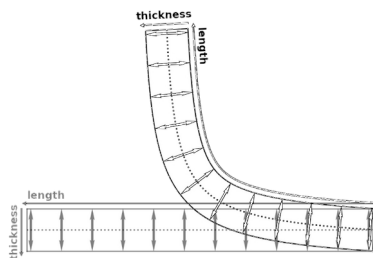
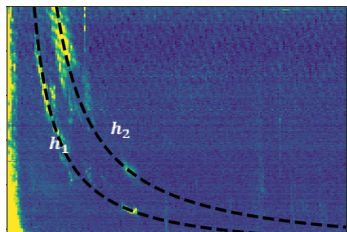
Challenges:

- Overlap of other signals
- **High variance in appearance:**
 - Possible presence of 1 or more harmonic(s)
 - **Dependency** between **drift rate** and **frequency range** [1]: $-\frac{df}{dt} = af^k$

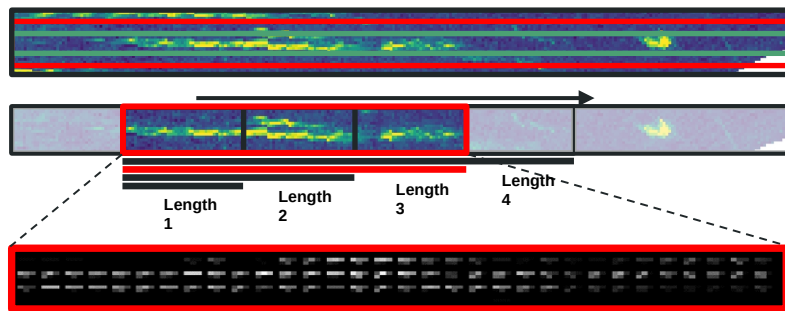
Domain knowledge!

Simplifying representations: drift rate invariance

Curvature-normalising ROI for (sliding window) detection



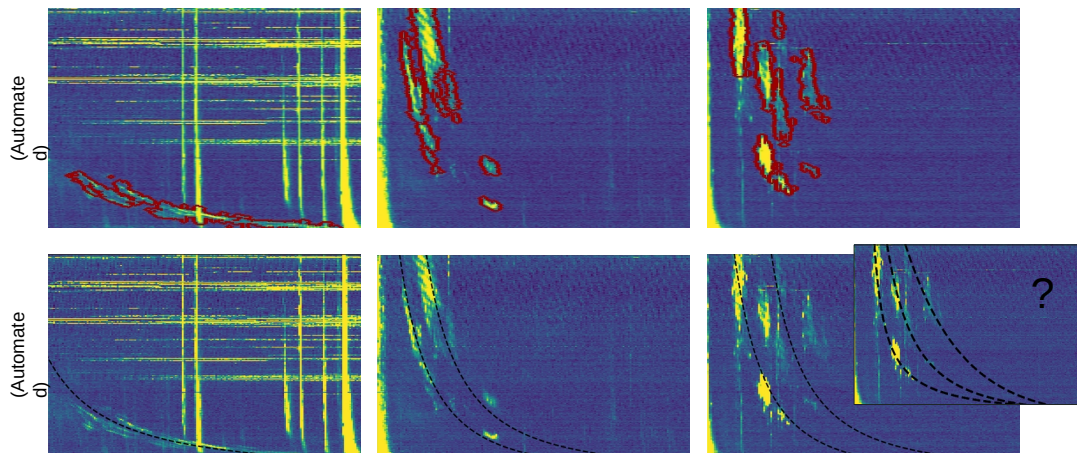
4 choices of drift rate
3 choices of thickness
4 choices of length
→ 48 ROIs tested at each location



HOG descriptor for logistic regression classification

Simplifying representations: drift rate invariance

Type II solar radio bursts



	Classical ROI	Proposed ROI
Precision	0.436 – 0.467 / 0.483 / 0.388	0.694 – 0.700 / 0.747 / 0.653
Recall	0.730 – 0.714 / 0.787 / 0.689	0.725 – 0.714 / 0.730 / 0.726
F-score	0.546 – 0.565 / 0.599 / 0.496	0.709 – 0.707 / 0.738 / 0.688
IoU per det. event	0.248 – 0.267 / 0.232 / 0.255	0.387 – 0.369 / 0.389 / 0.393
Global IoU	0.134 – 0.136 / 0.146 / 0.124	0.282 – 0.249 / 0.312 / 0.272

Perspectives (ongoing work)

- Estimation of burst parameters
- Application to Yamagawa data
- Adaptation to a deep learning framework

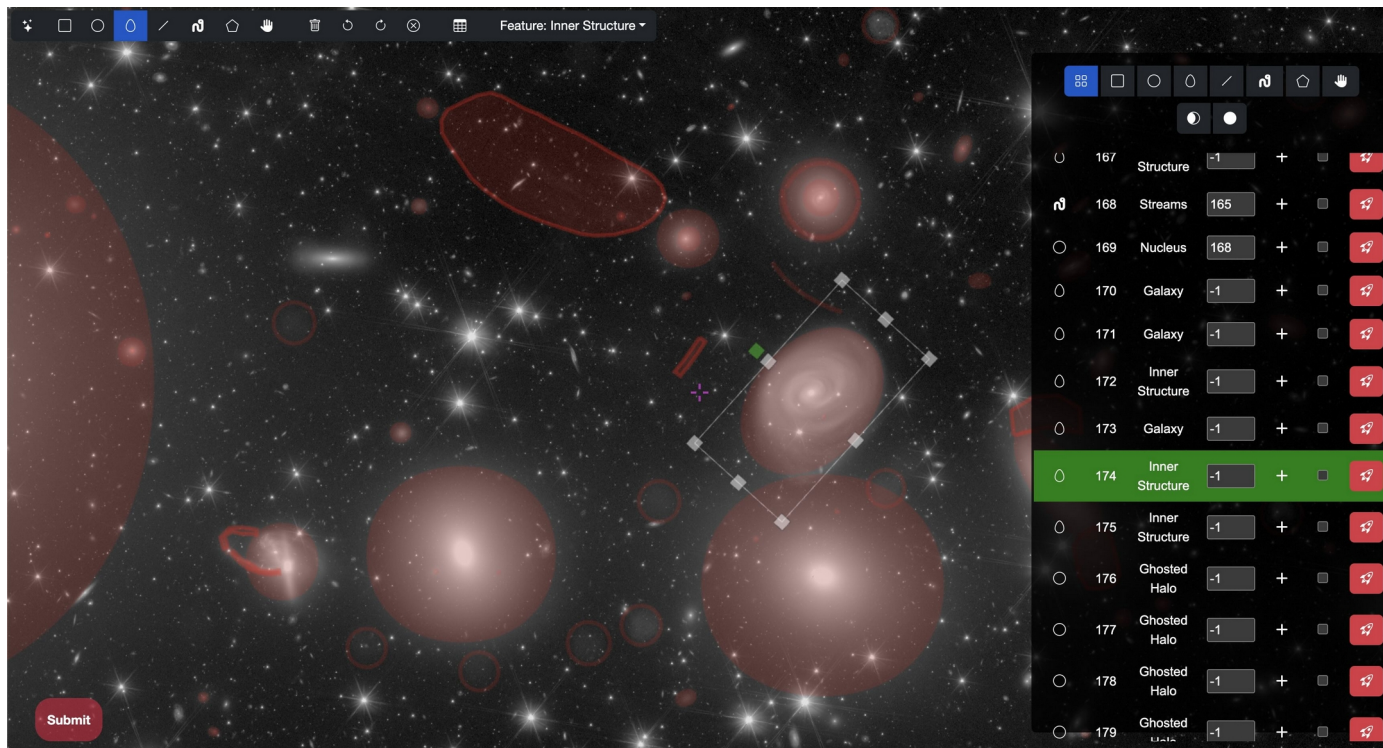
Challenges in supervised training

dealing with sparse
and inconsistent
data annotations

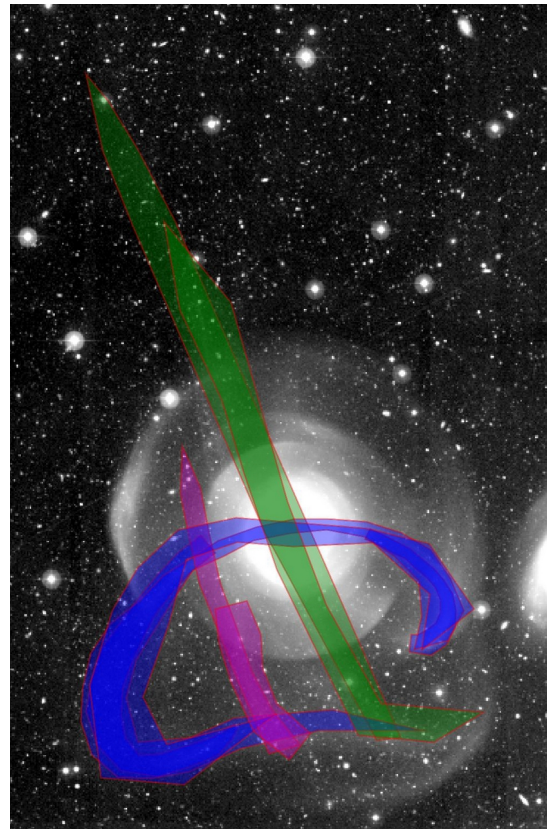
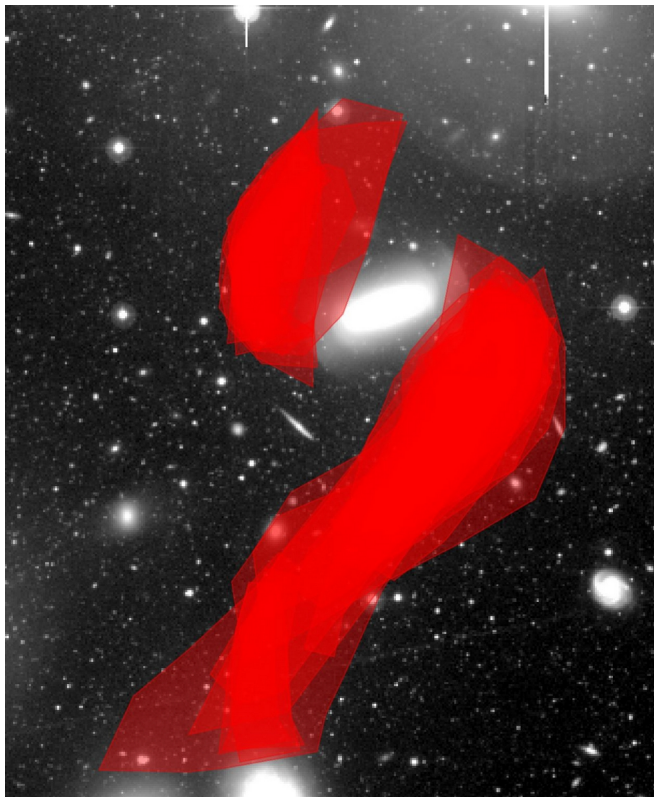
- Consensus between annotators
- Bootstrapping the learning
- Data augmentation



Manual annotations of images of galaxies



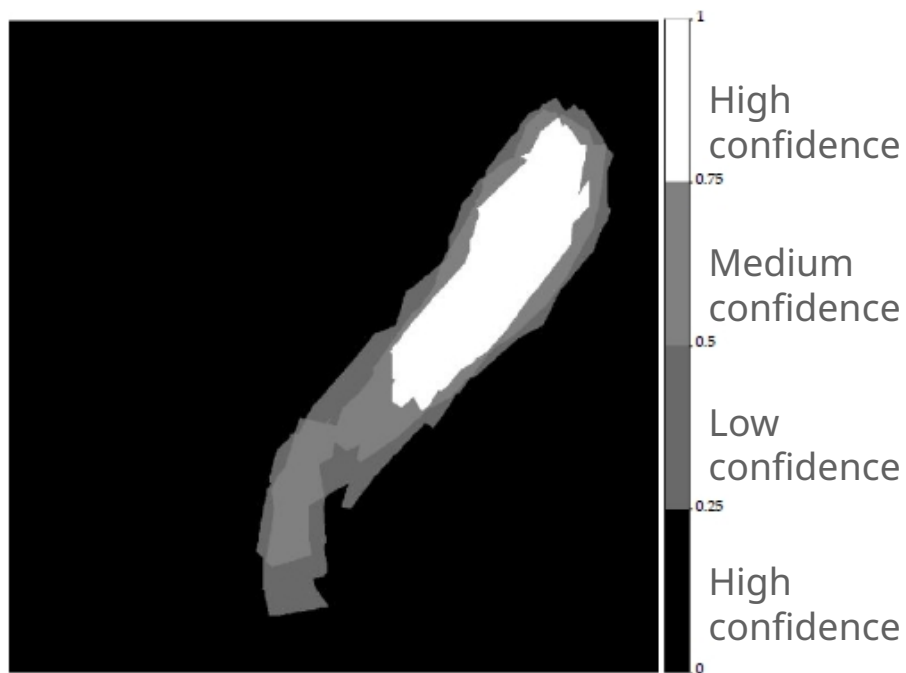
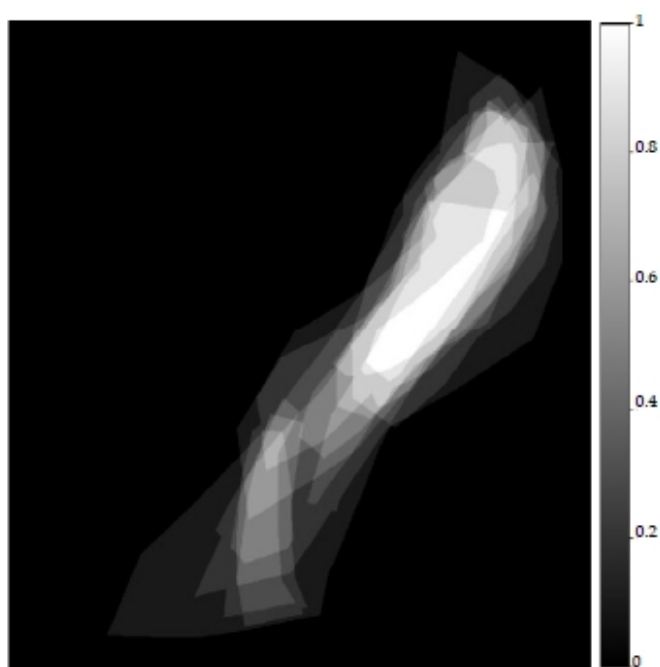
(Lack of) consensus between annotators



Training on a consensus between annotators

Different regimes indicate different levels of confidence

- Modulate the penalty during training (loss function)

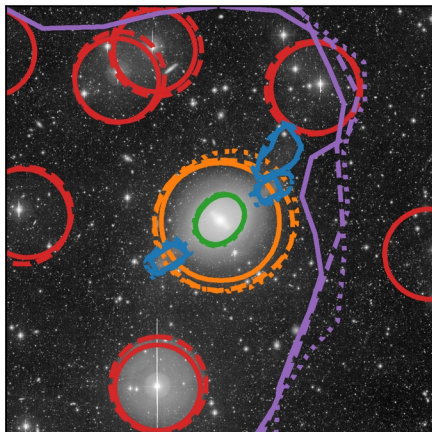


Some early results

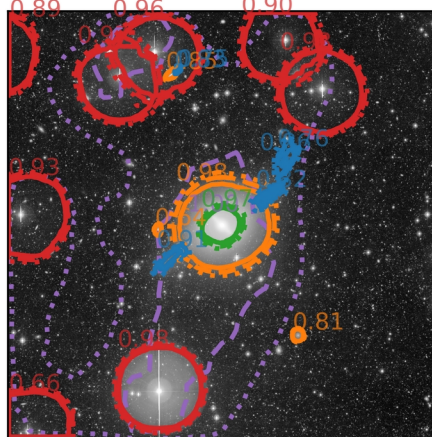
Contour Levels
..... 0.20 - - - 0.40 ——— 0.70

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

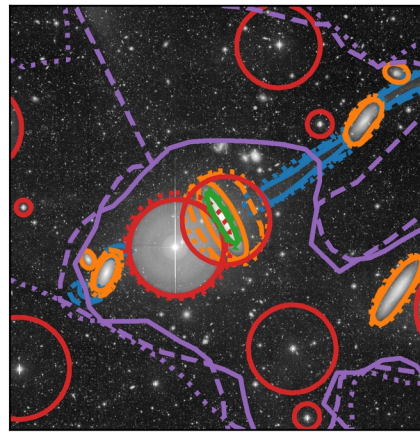
Ground Truth



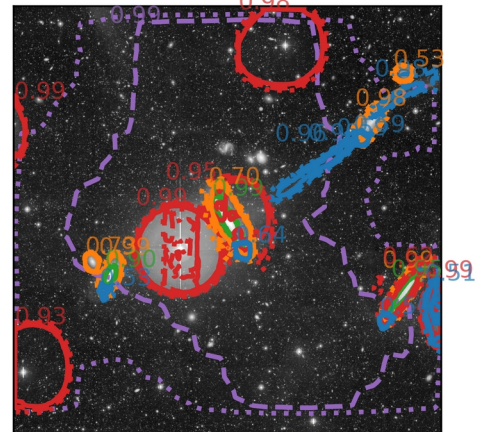
Prediction



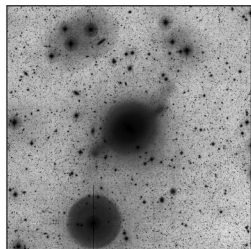
Ground Truth



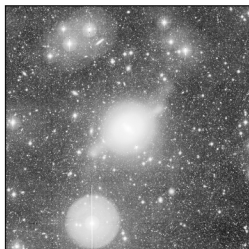
Prediction



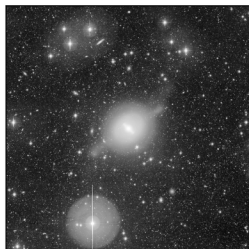
Scale 0



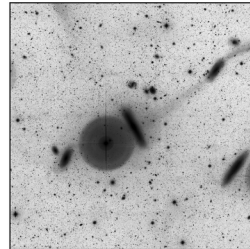
Scale 1



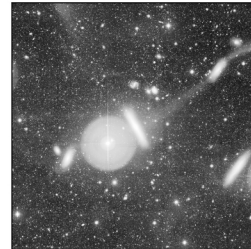
Scale 2



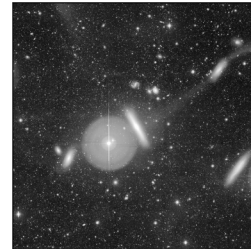
Scale 0



Scale 1



Scale 2

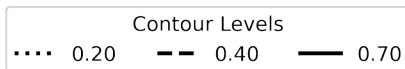
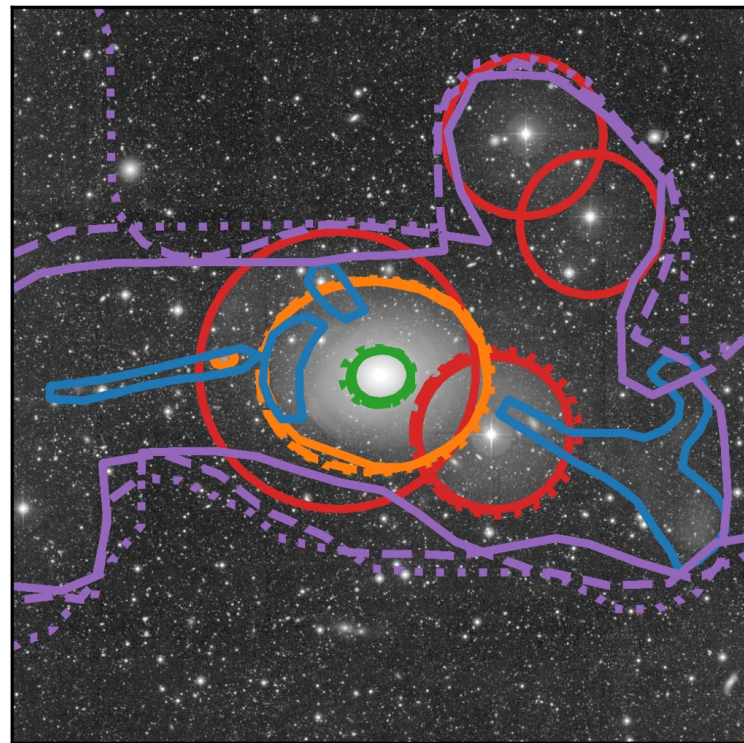


Using the expertise level of annotators

The most challenging structures are only found by exper

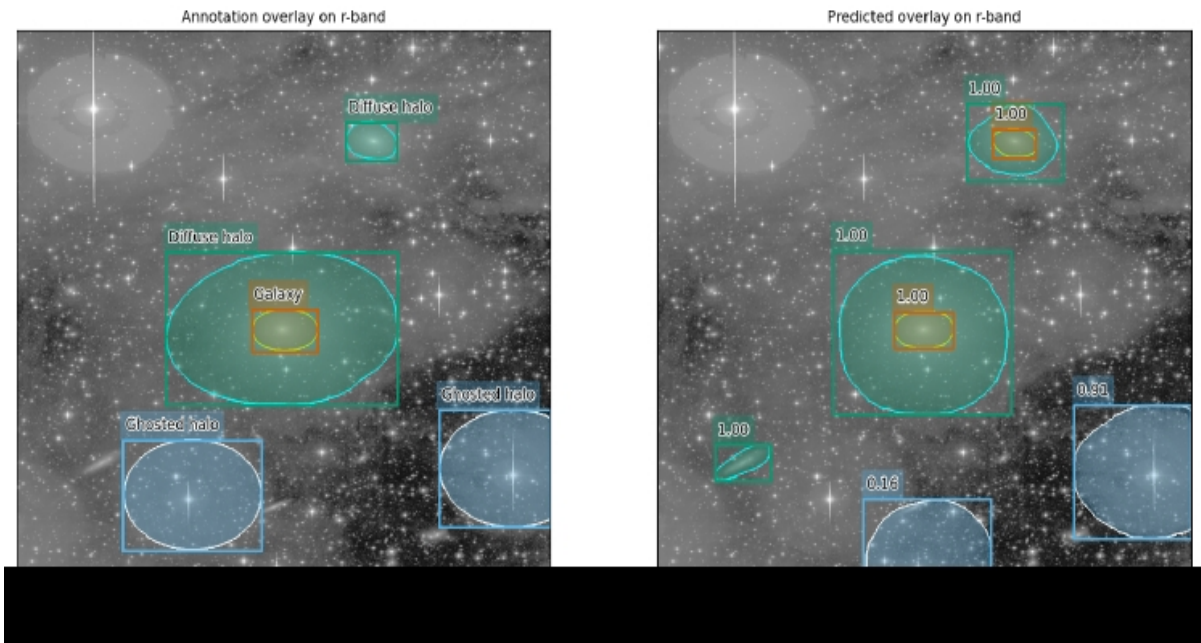
- Highlight these structures during training
- **Hard example mining**

... in progress ...



Incomplete annotation of the data

Detections of non-annotated objects are not necessarily false!

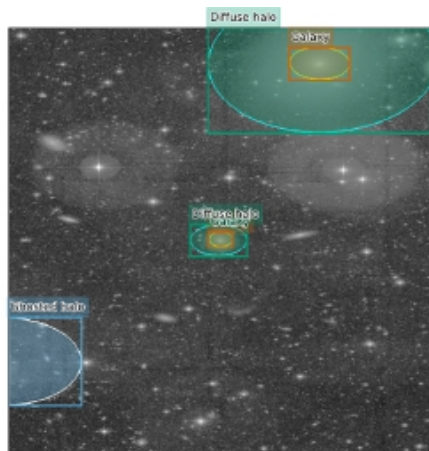


Red: galaxy
Green: extended halo
Dark blue: elongated tidal features (streams and tidal tails)
Light blue: ghosted halo (contaminant)

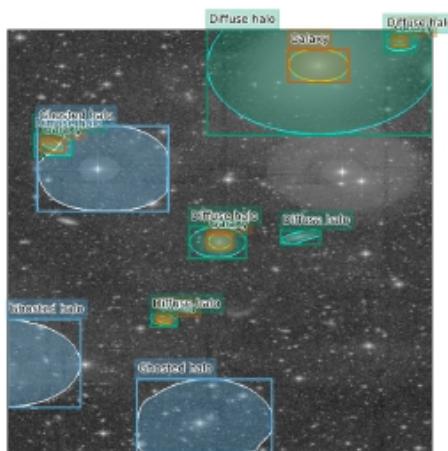
Semi-automatic completion of the dataset

Human in the loop training:

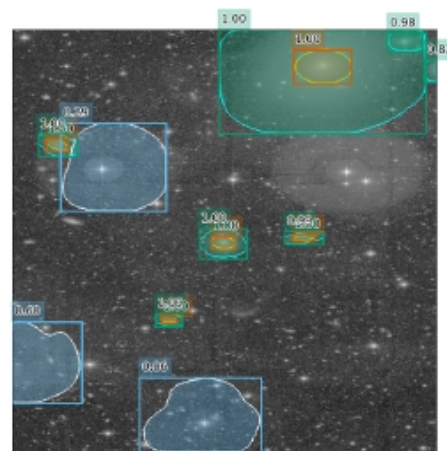
1. Manually select true detections that were not manually labelled
2. Re-train with more complete dataset
3. Repeat



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)

Bootstrapping the learning

Detailed labels are hard & time-consuming to obtain

Bootstrapping strategy:

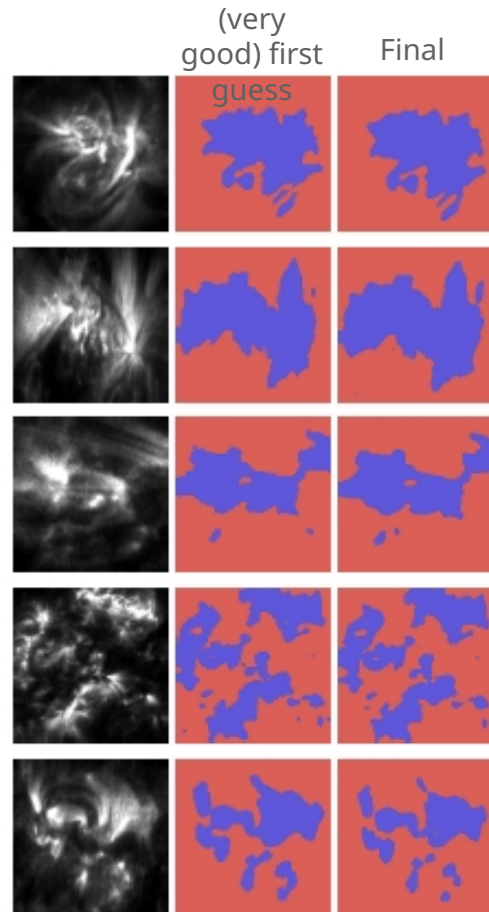
Strat with approximate labels:

- bounding box in the centre of the object
- Intensity threshold
- etc.

Train

Use the predictions as new labels

Repeat



Dealing with rare classes

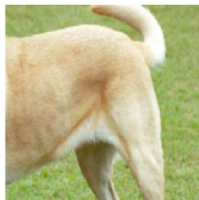
Various strategies exist, including

- Balancing the learning
- Mining rare samples
- Augmenting the data

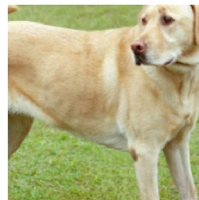
$$\text{BalancedCrossEntropy} = - \sum_{i=1}^{i=n} \alpha_i \log_b(p_i)$$



(a) Original



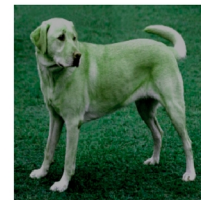
(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate {90°, 180°, 270°}



(g) Cutout



(h) Gaussian noise



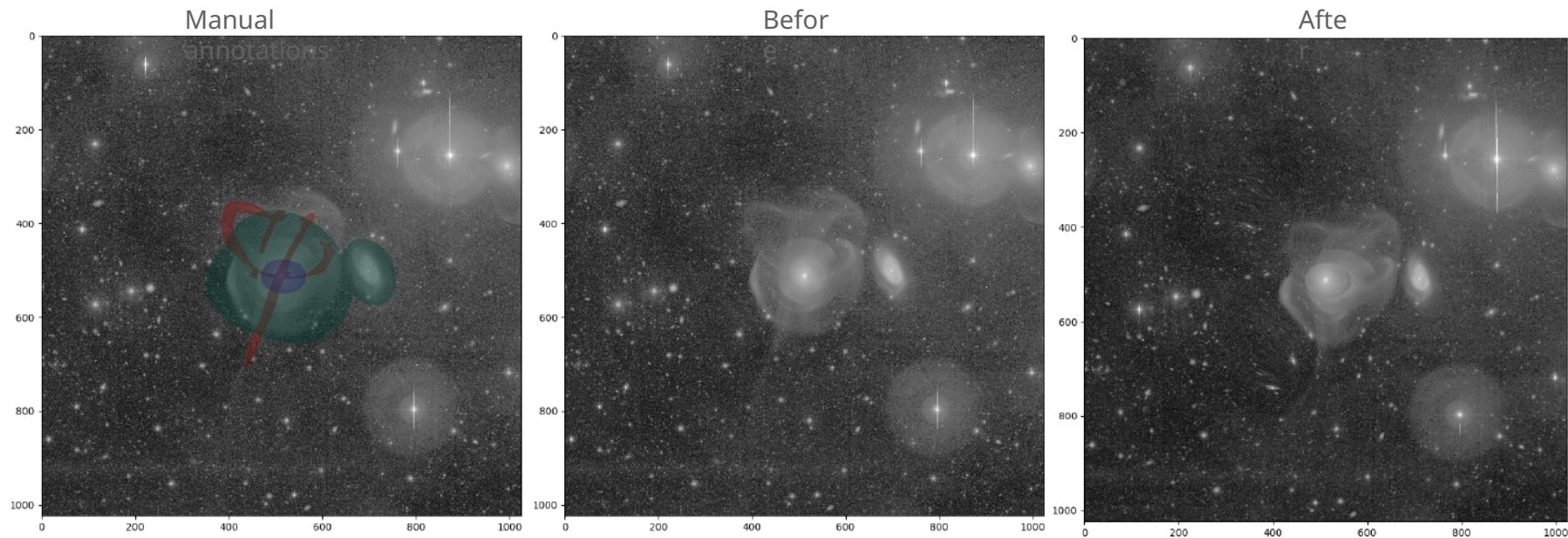
(i) Gaussian blur



(j) Sobel filtering

Class-specific data augmentations

Creating new instances through deformations



In progress: in-painting of the background for larger deformations

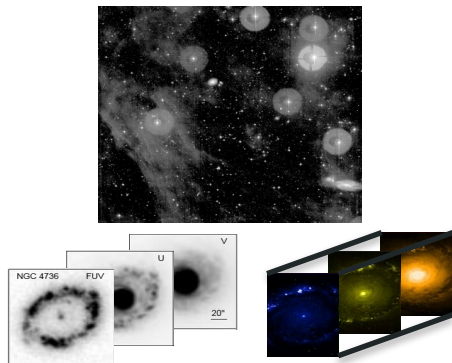
Conclusion and perspectives

Analysing astrophysics data

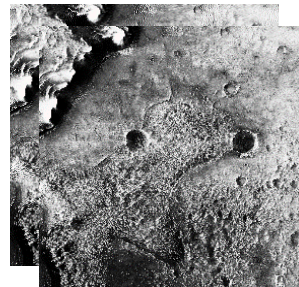
Computer vision, machine learning and deep learning → Exciting results

Going beyond off-the-shelf models:

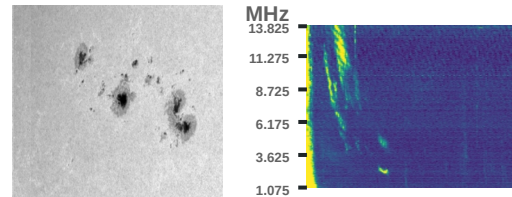
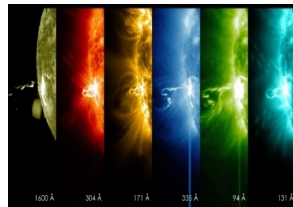
- Domain-informed methods
- Case-by-case strategies & implementations
- Addressing the data annotation challenges



Galaxy evolution



Martian geology



Solar activity

Thanks for listening!
