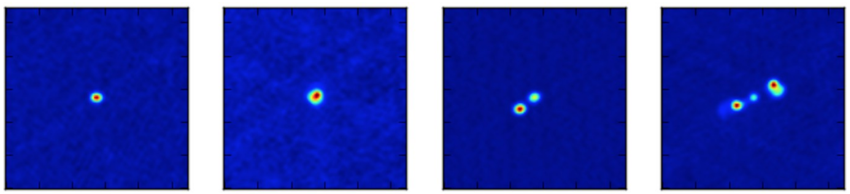
Deep learning in radio astronomy

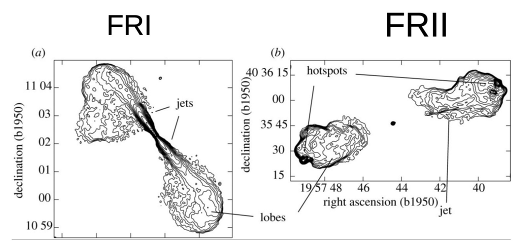
Vesna Lukic

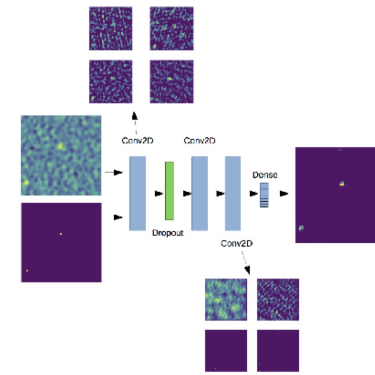
Vesna Lukic

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

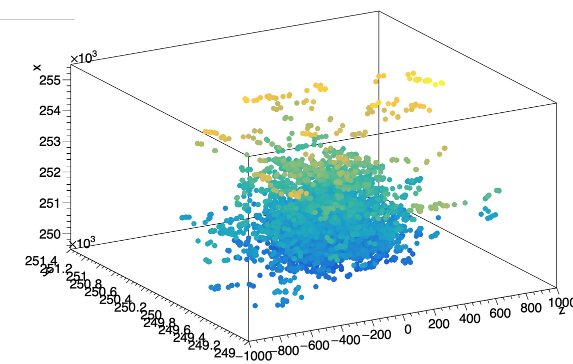
Previously: PhD at the University of Hamburg, at Hamburg Observatory, supervised by Marcus Brüggen. Looked at classifying radio galaxy morphologies, as well as source finding



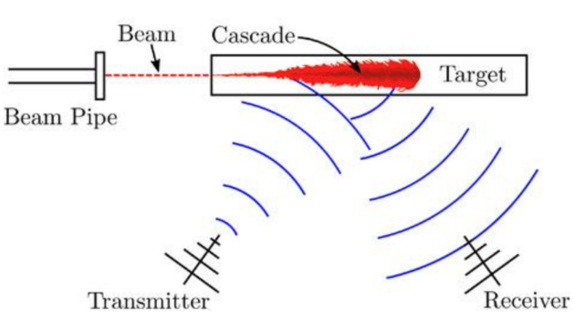


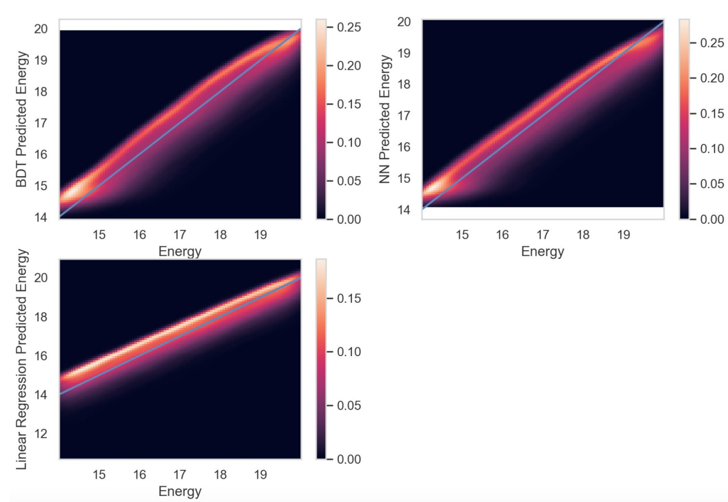


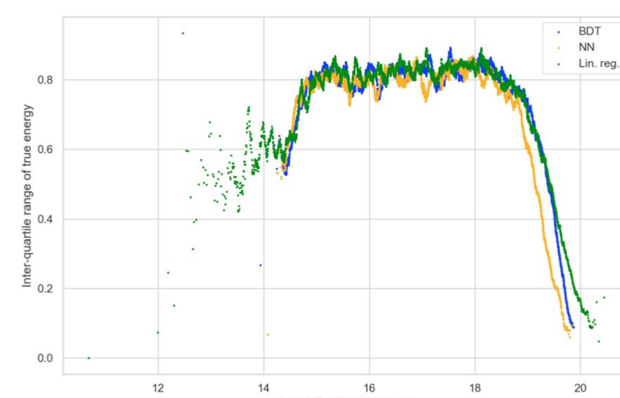
Currently: Postdoc in Krijn De Vries group, working on energy construction and resolution for the Radar Echo Telescope (RET)



VRIJE UNIVERSITEIT BRUSSEL

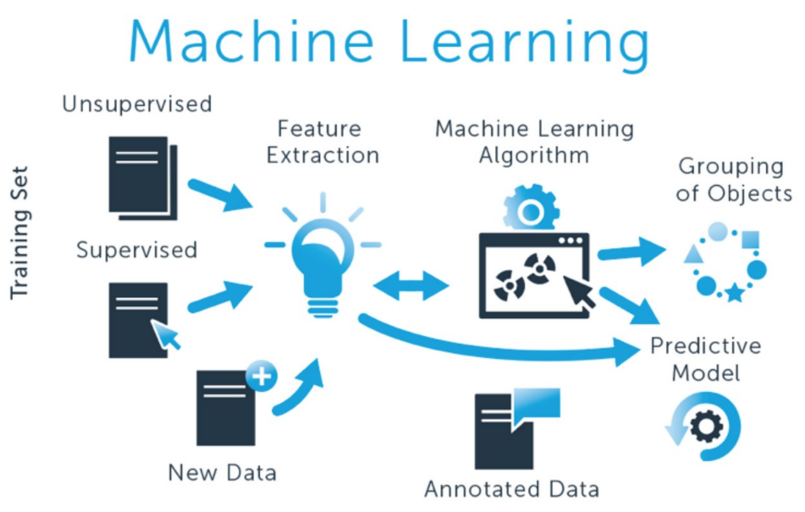


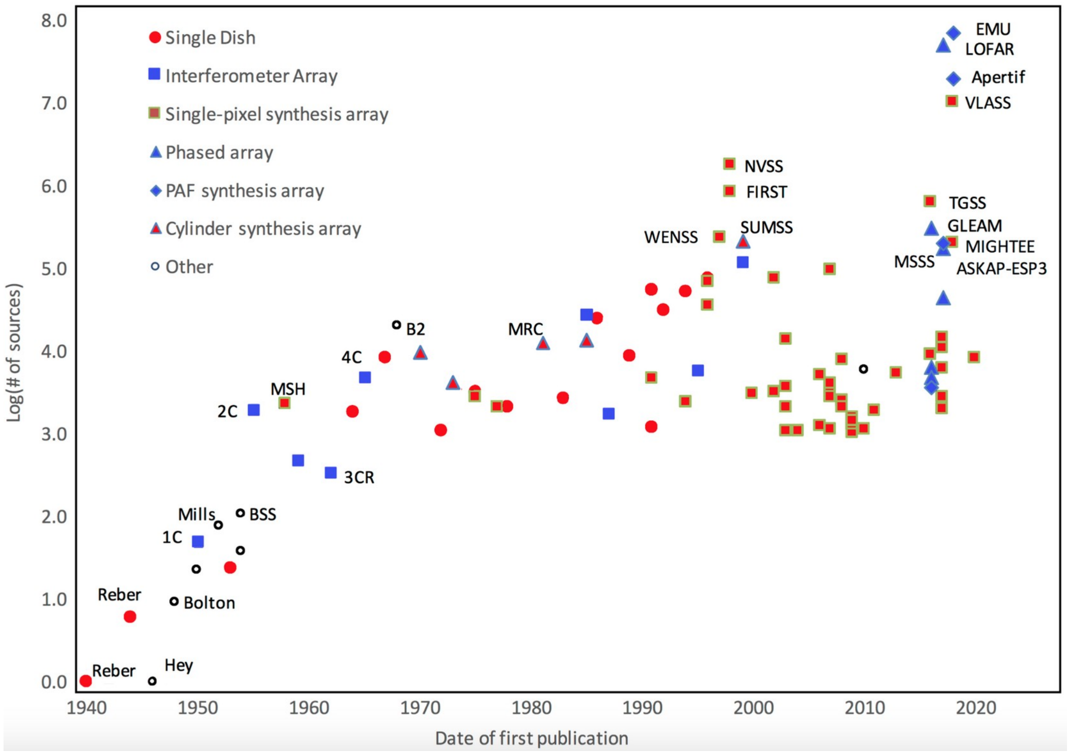




Outline

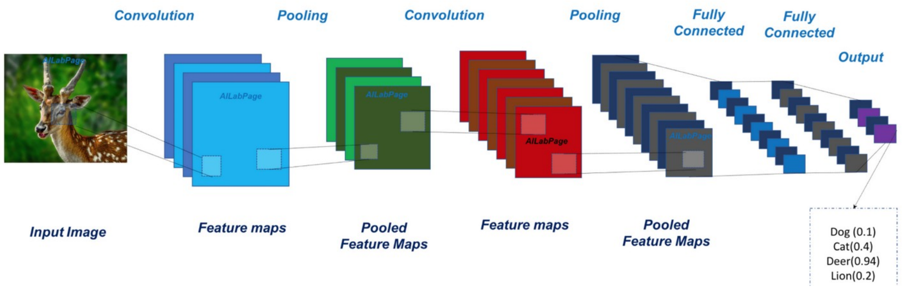
- Classifying astronomical sources is important for science results
  - AGN and SFGs are fundamentally different classes of objects
  - Different source types influence evolution and constitution of the Universe in different ways
- Surveys are detecting higher quality sources
  - Machine learning, citizen science

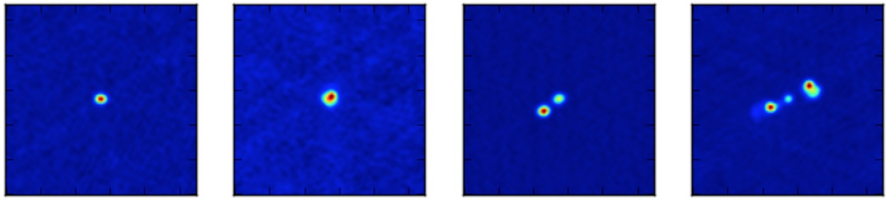




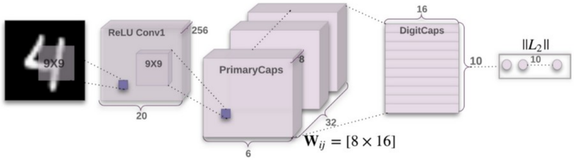
Outline

- We addressed these problems using deep learning (DL)
  - Classifying radio sources by numbers of components

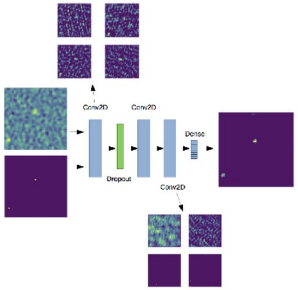




  - Comparing two approaches in classifying between radio galaxy classes

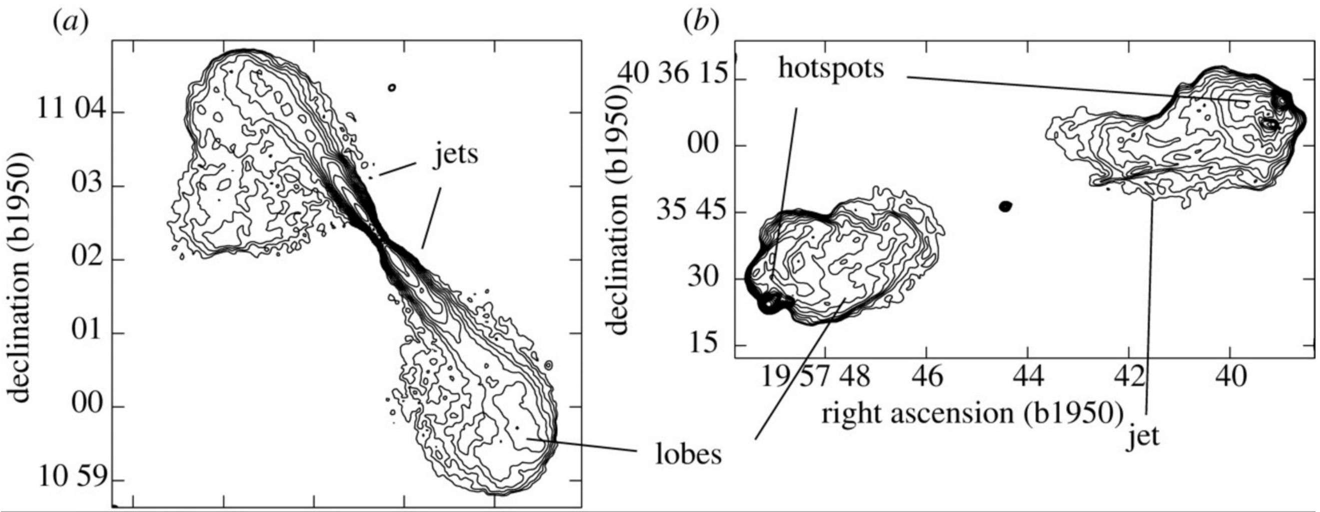


  - Investigating whether a ConvNet can be used to find radio sources

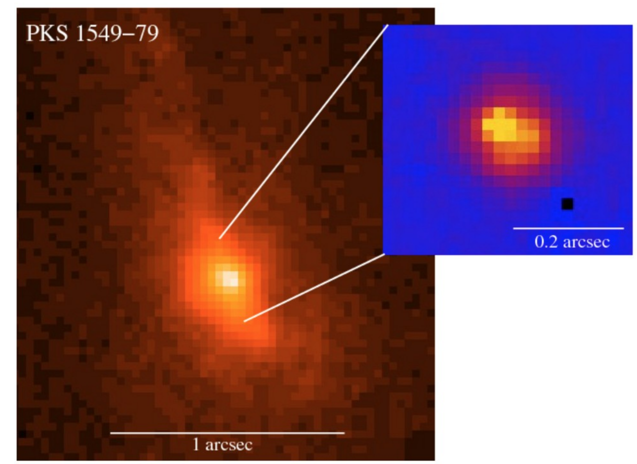


FRI

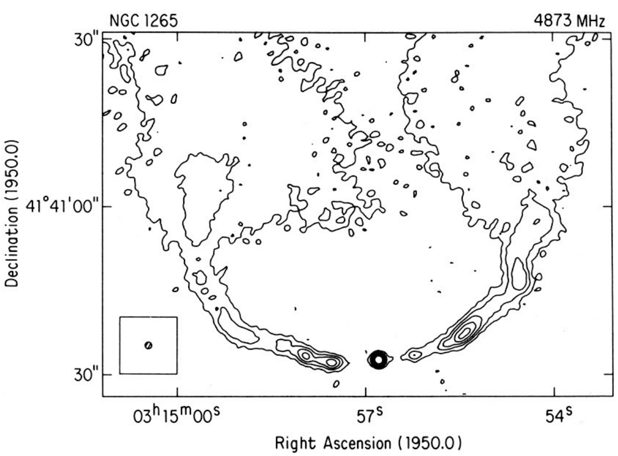
FRII



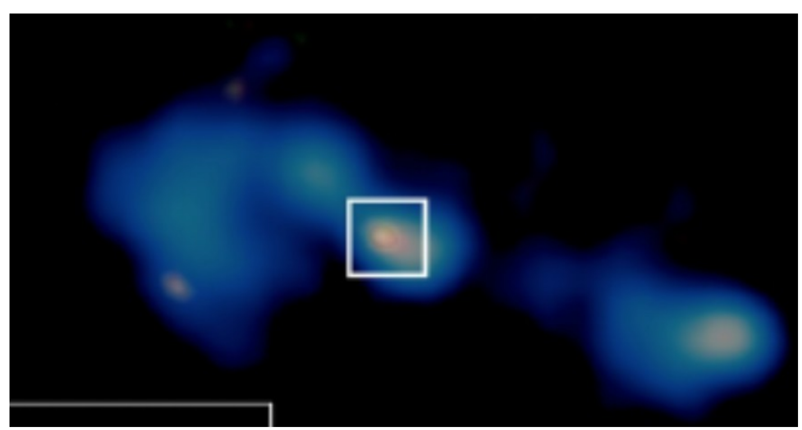
Compact



Bent-tailed



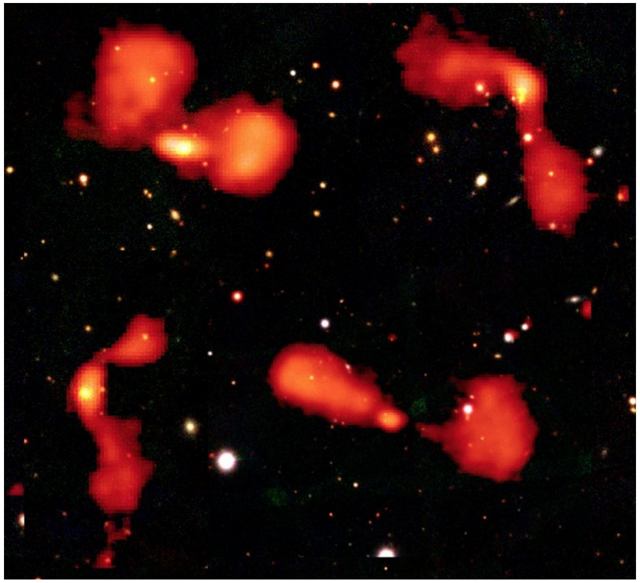
Hybrid

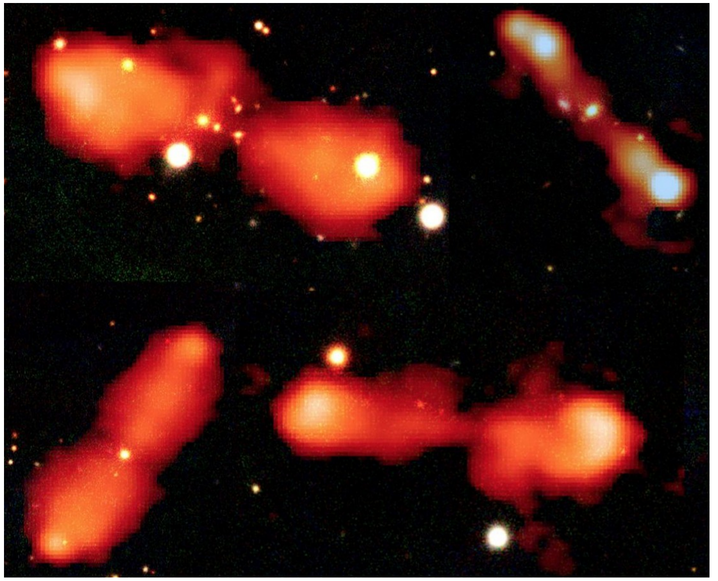


Star-forming galaxy

Background

FRI



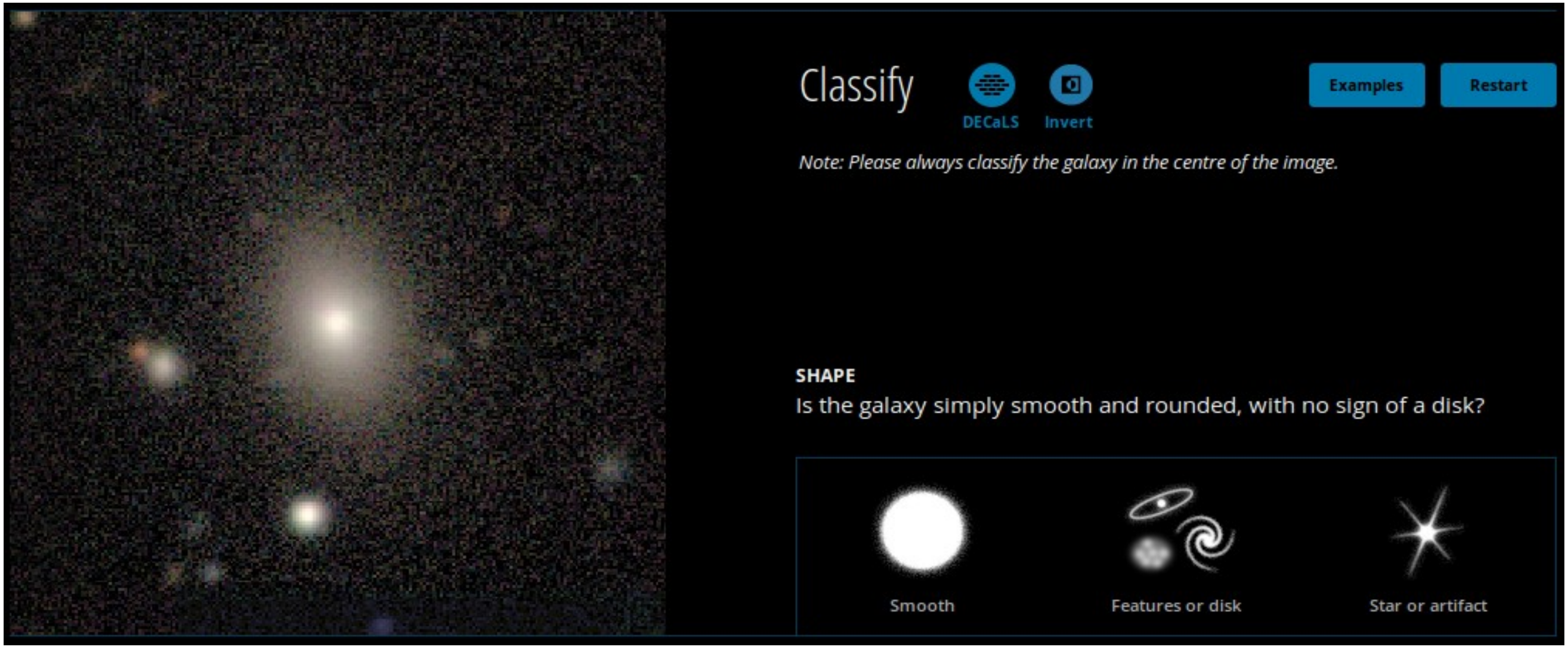


FRII

- The radio galaxy morphology reveals important properties
  - The surrounding environment of the radio galaxy
  - Almost everything we know about jets relates to their morphology and luminosity
  - Location of brightest parts of the emission, how the source type influences the immediate environment

Machine learning in astronomical images

- Citizen scientists asked to describe optical galaxy morphologies

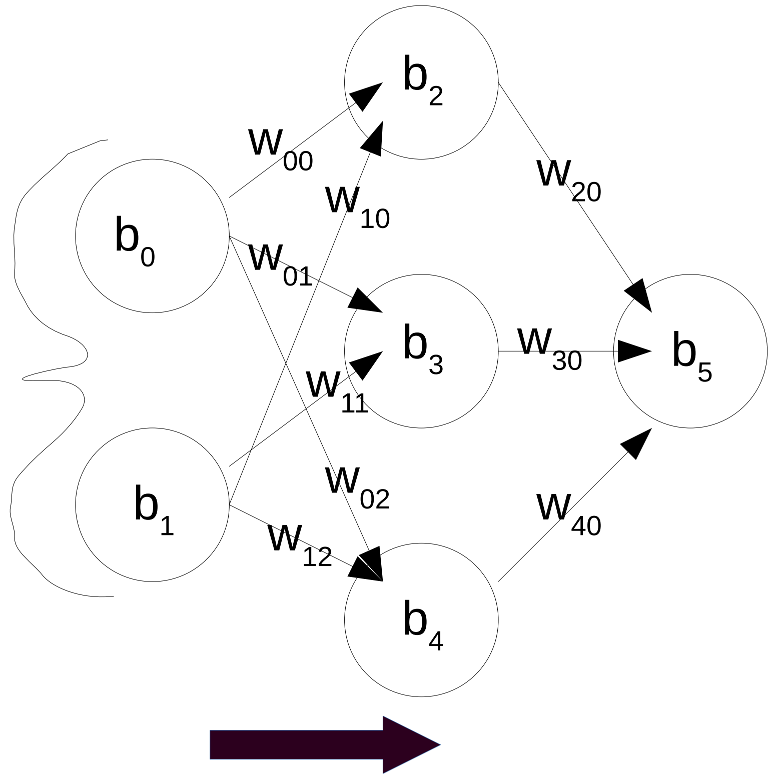


- Determine the probability that a galaxy belongs in a particular class (regression problem)
- Winning solution used **convolutional neural networks**

Neural network basics

1. Input x

This sets input layer activations



3. Error in final layer

δ^L

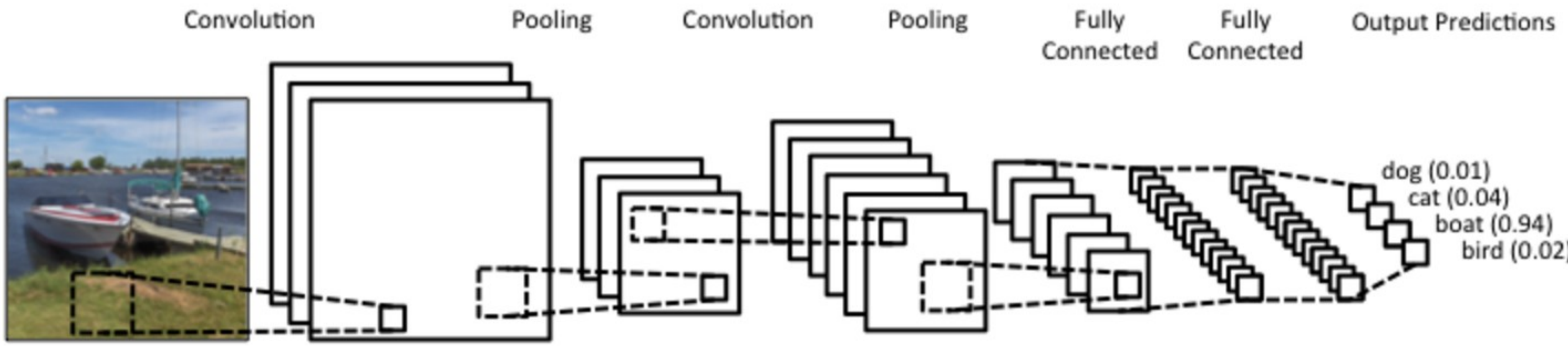
Backpropagate this error back through all layers

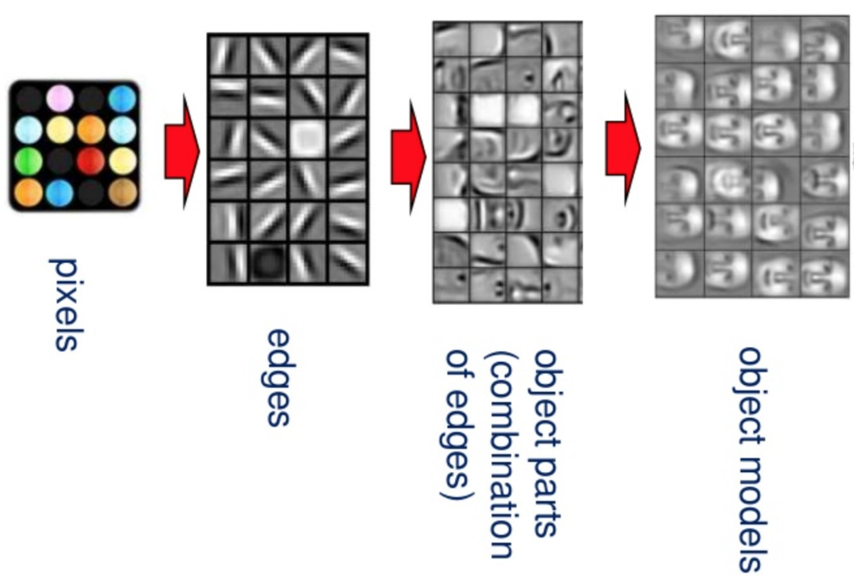
2. Feedforward through the network

4. Calculate $\frac{\partial C}{\partial w}$ $\frac{\partial C}{\partial b}$

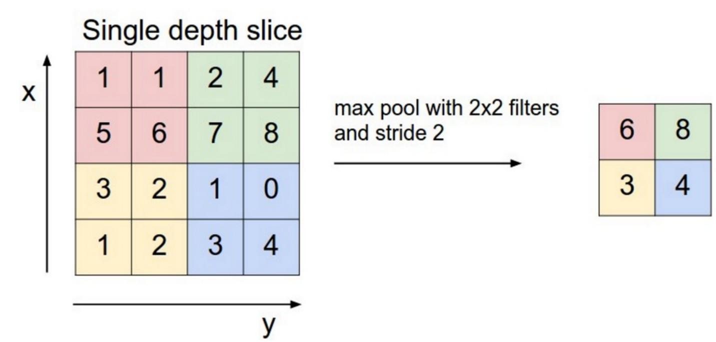
5. Update w and b

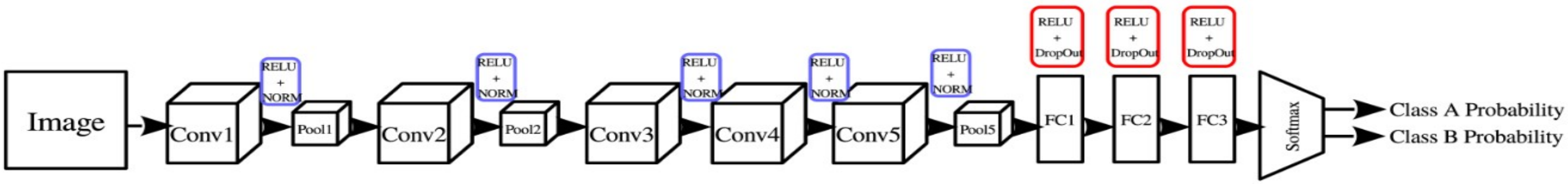
Convolutional neural networks

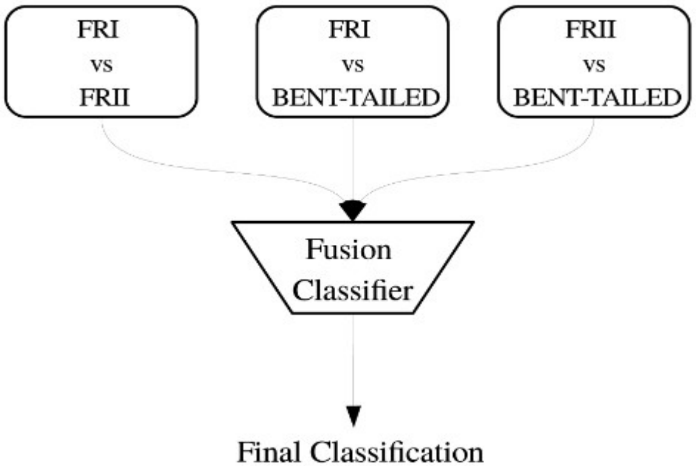




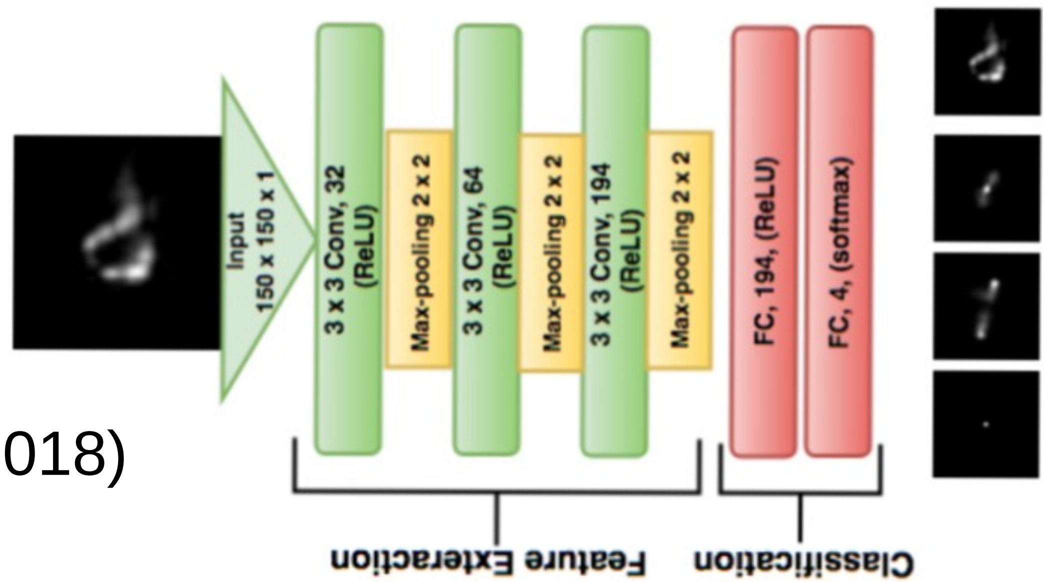
Pooling



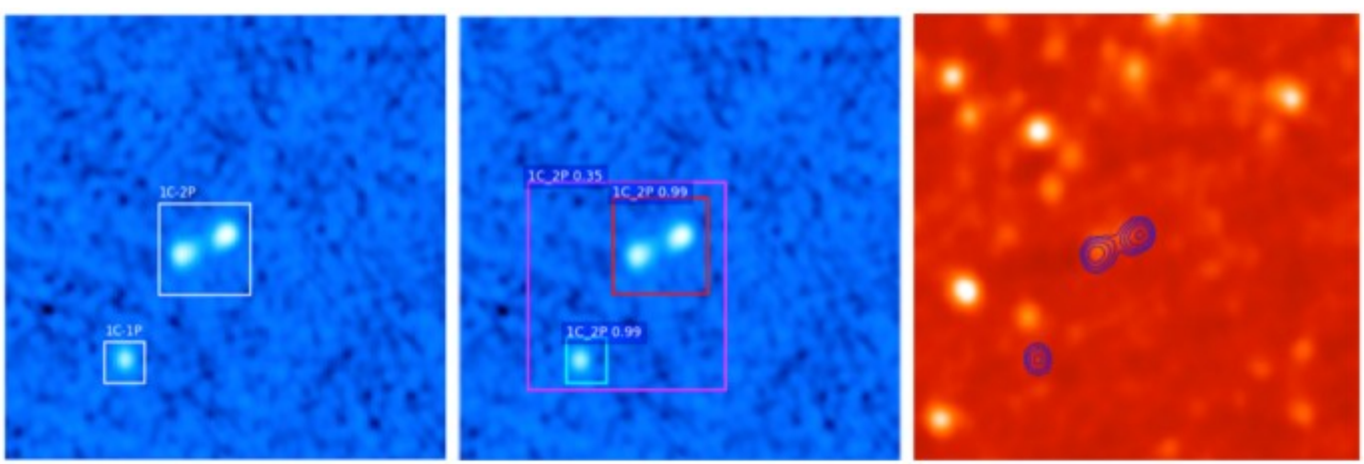




Aniyan & Thorat (2017)



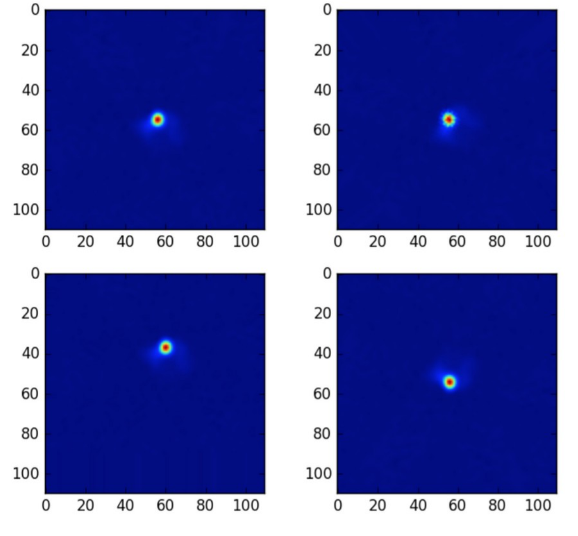
Alhassan, Taylor & Vaccari (2018)



Wu…, Lukic…, et al. (2018)

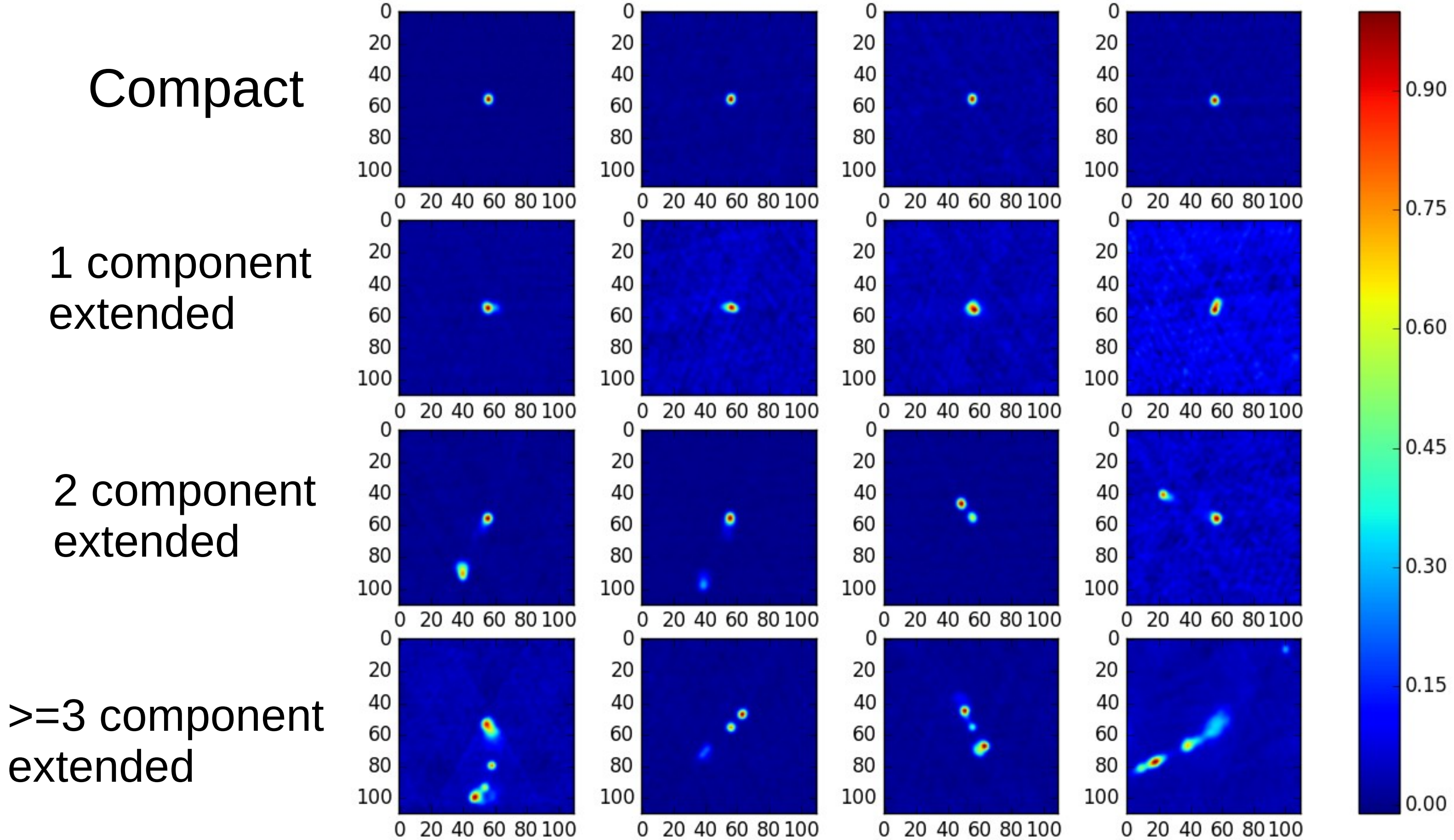
Radio Galaxy Zoo (RGZ)
Compact and Extended source classification

- Lukic et al. 2018 (published in MNRAS)
- Image data of >200,000 galaxies, no labels
- Single channel, 132x132 pixels
- Images contain different numbers of components
- Used the Python Blob Detector and Source Finder (PyBDSF) to help organise the data
- Generated more images using translation, rotation and flipping



Four-class problem

- Applied a 3 Conv + 2 dense layer CNN set-up to the four-class problem, classification accuracy > 93%



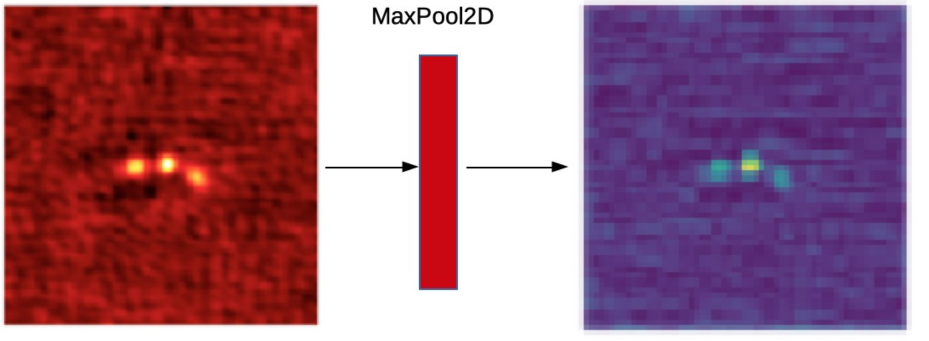
Cross-check with DR1 of RGZ

- DR1 – citizen scientists which components belonged to which sources
  - Labels: ‘Number of components’ and ‘Number of peaks’
- Test classification accuracy > 94 %
  - Influenced by high numbers of compact/single-component extended sources
  - Higher # components → worsened performance

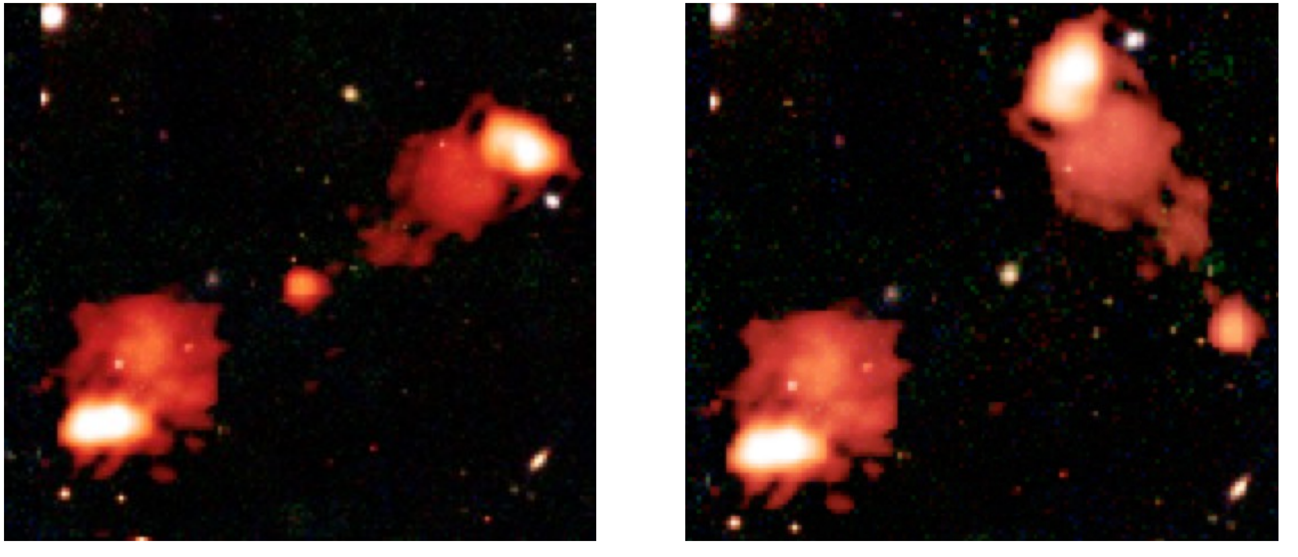
	Precision (per cent)	Recall (per cent)
Compact/single-extended	97.5	96.9
Two-component extended	88.0	89.5
Multiple-component extended	53.4	58.5

Drawback of CNNs

- Relative location of features within image is not preserved, due to pooling operation

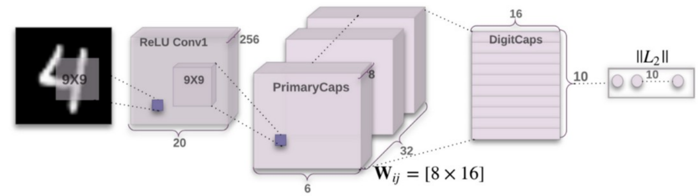


- Lack of rotational invariance

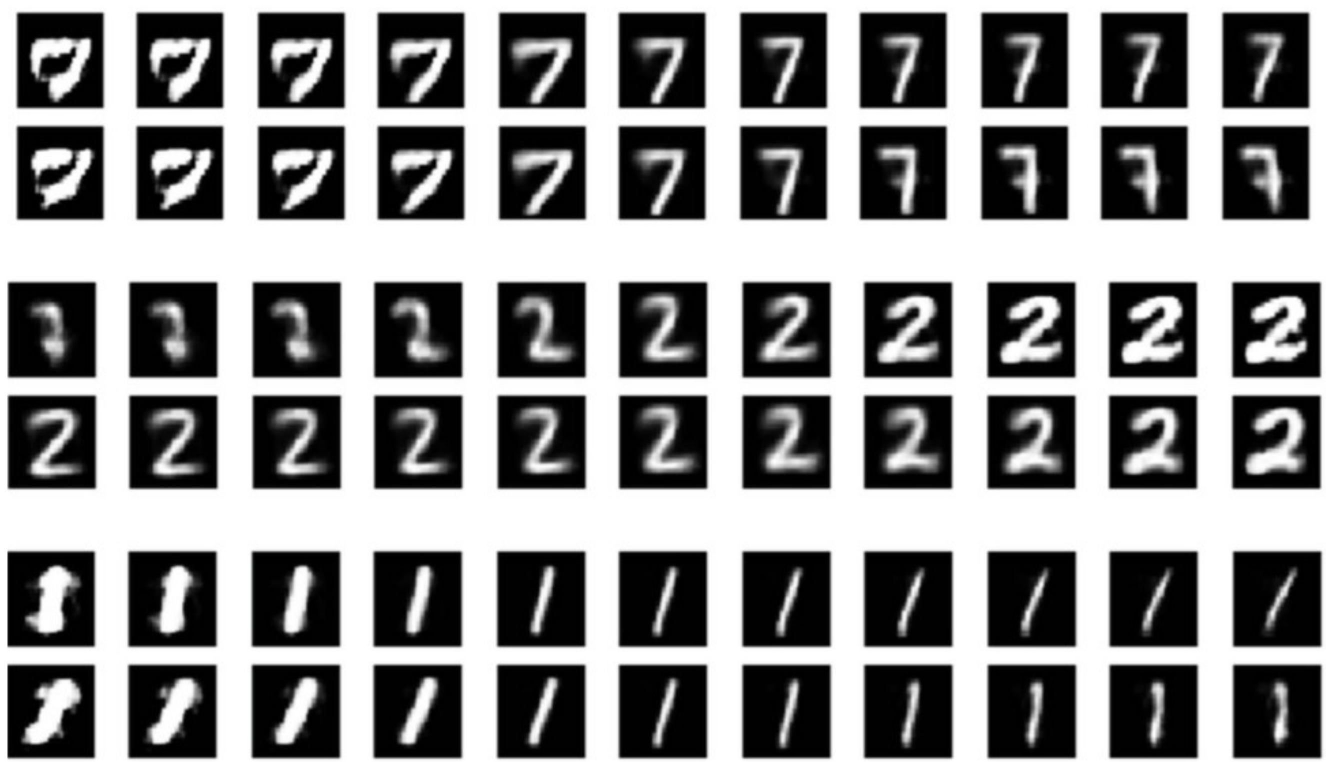


Capsule networks

- Designed to preserve hierarchical relationships in images (Sabour, Frosst, Hinton (2017))



- A capsule consists of a group of neurons that attempt to extract possible variations of the subject in the image (e.g. Thickness and deformation)

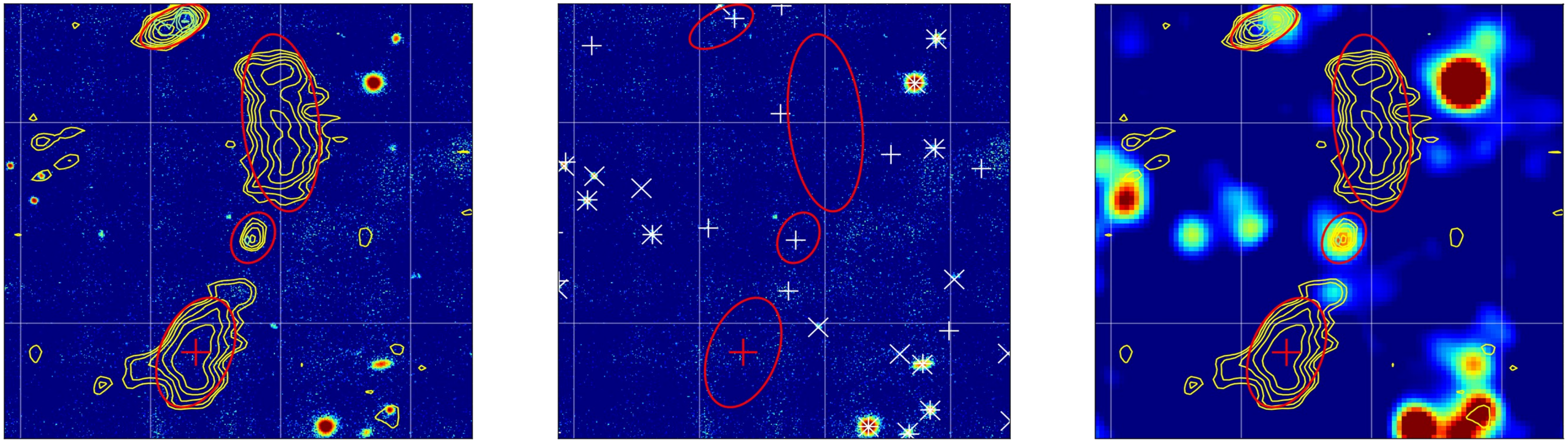


Capsule versus CNNs

- Morphological classification of radio galaxies: Capsule networks versus Convolutional Neural Networks (V. Lukic, M. Brüggen, B.Mingo, J.H. Croston, G. Kasieczka, P.N. Best). Published in MNRAS
- Sources from the LOFAR LoTSS HetDex field

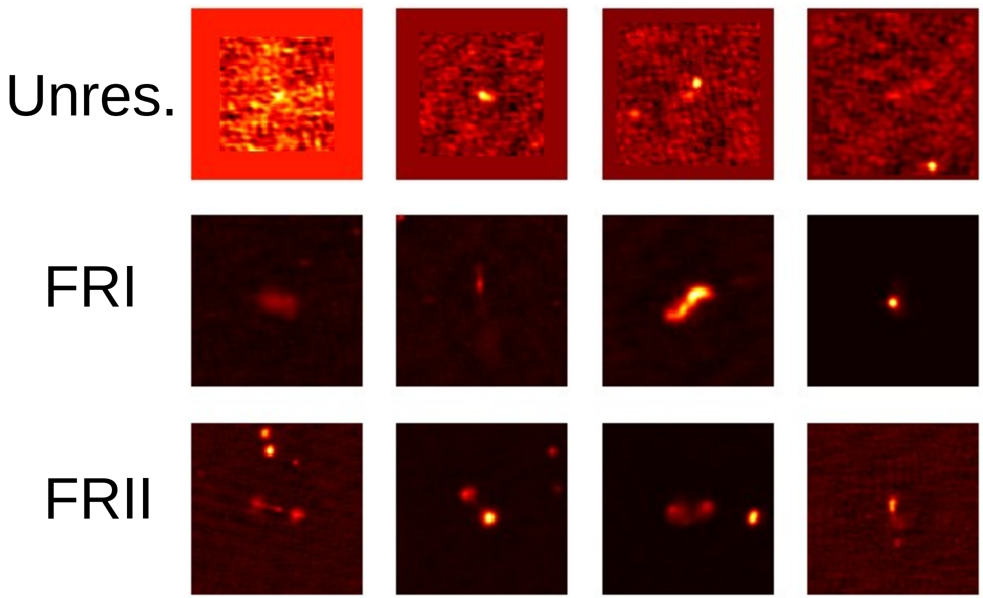
Cross-ID

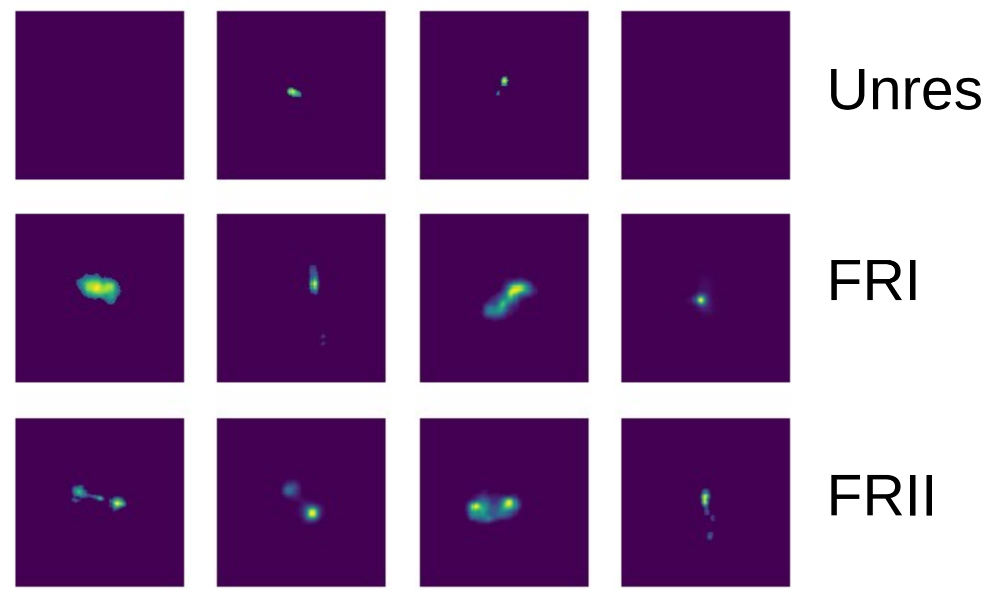
- Cross-identification of radio sources with optical source
  - Sources < 15 arcsec : Maximum likelihood technique
  - Sources > 15 arcsec : Inspected by expert astronomers



Details of images

- 2901 images with classifications:
  - Fits file cutouts and 4rms sigma-clipped numpy arrays
  - Unresolved, FRI, FRII





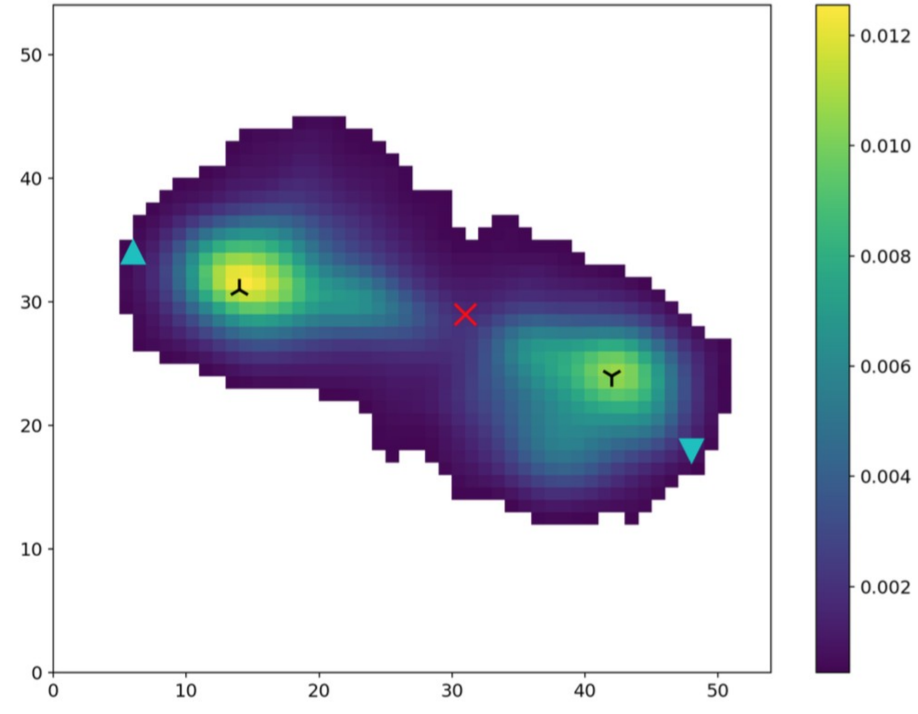
Details of images

- Labels generated using automated technique on 4rms images, FRIs and FRIIs visually cross-checked

Class	# Original	# Augmented	# Total
Unresolved	1457	4371	5828
FRI	984	5904	6888
FRII	460	2760	3220
Total	2901	13035	15936

FRI: d1/Maxd1 < 0.5 and d2/Maxd2 < 0.5

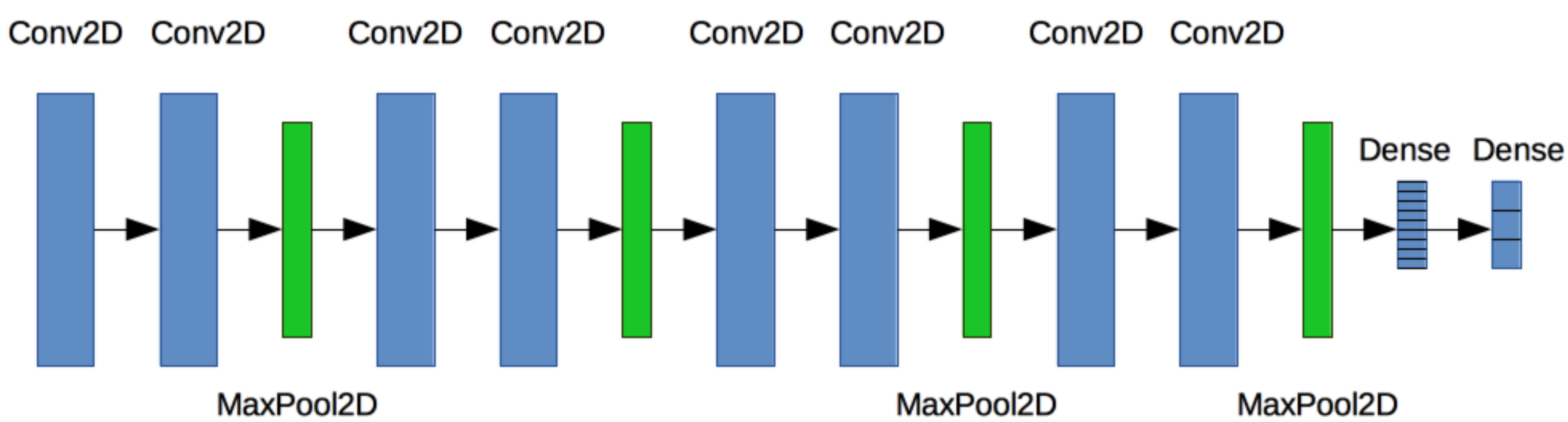
FRII: d1/Maxd1 > 0.5 and d2/Maxd2 > 0.5



Datasets and architectures used

- Original (+ augmented) fits images and 4rms numpy arrays
- 79% training and 21% for validation and testing
- 4- and 8- layer convolutional network, capsule network variations

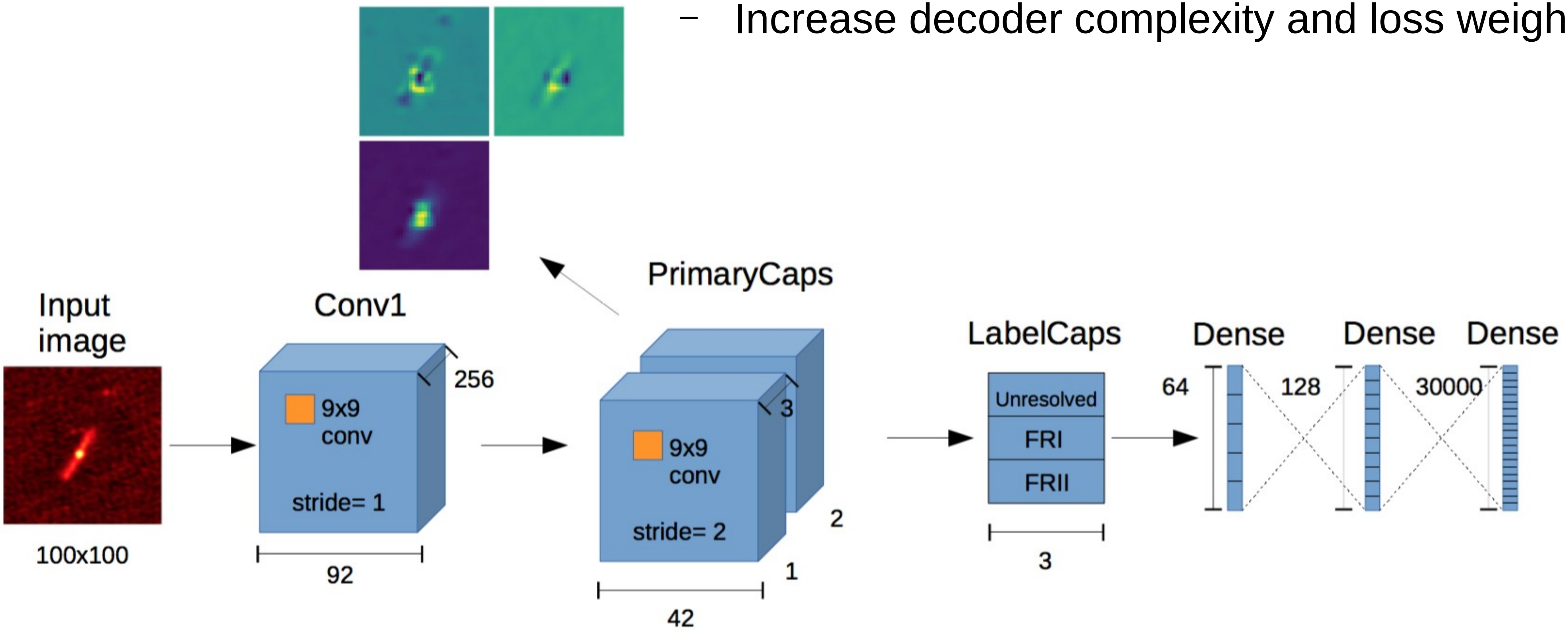
ConvNet-8

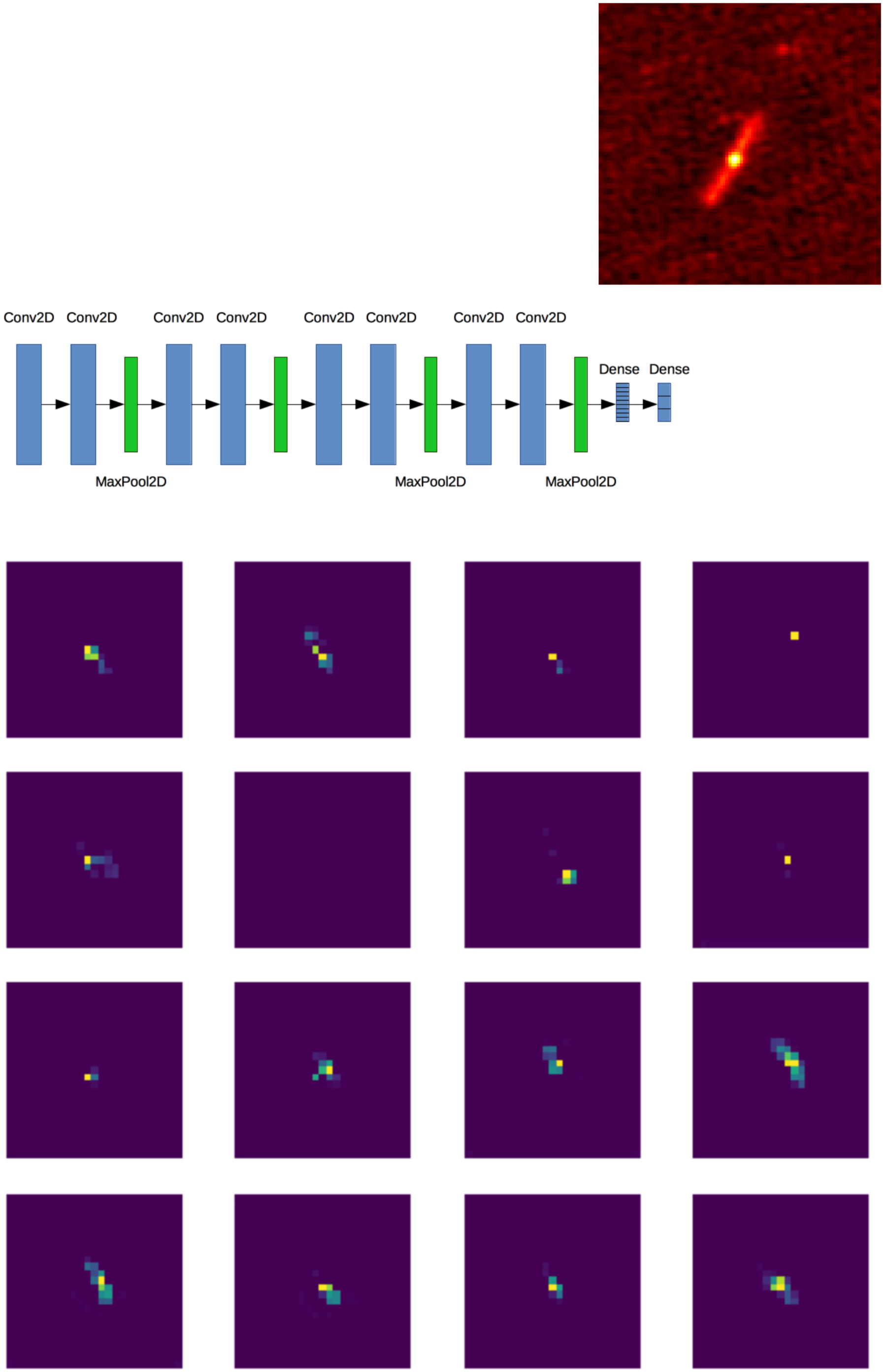


ConvNet-4

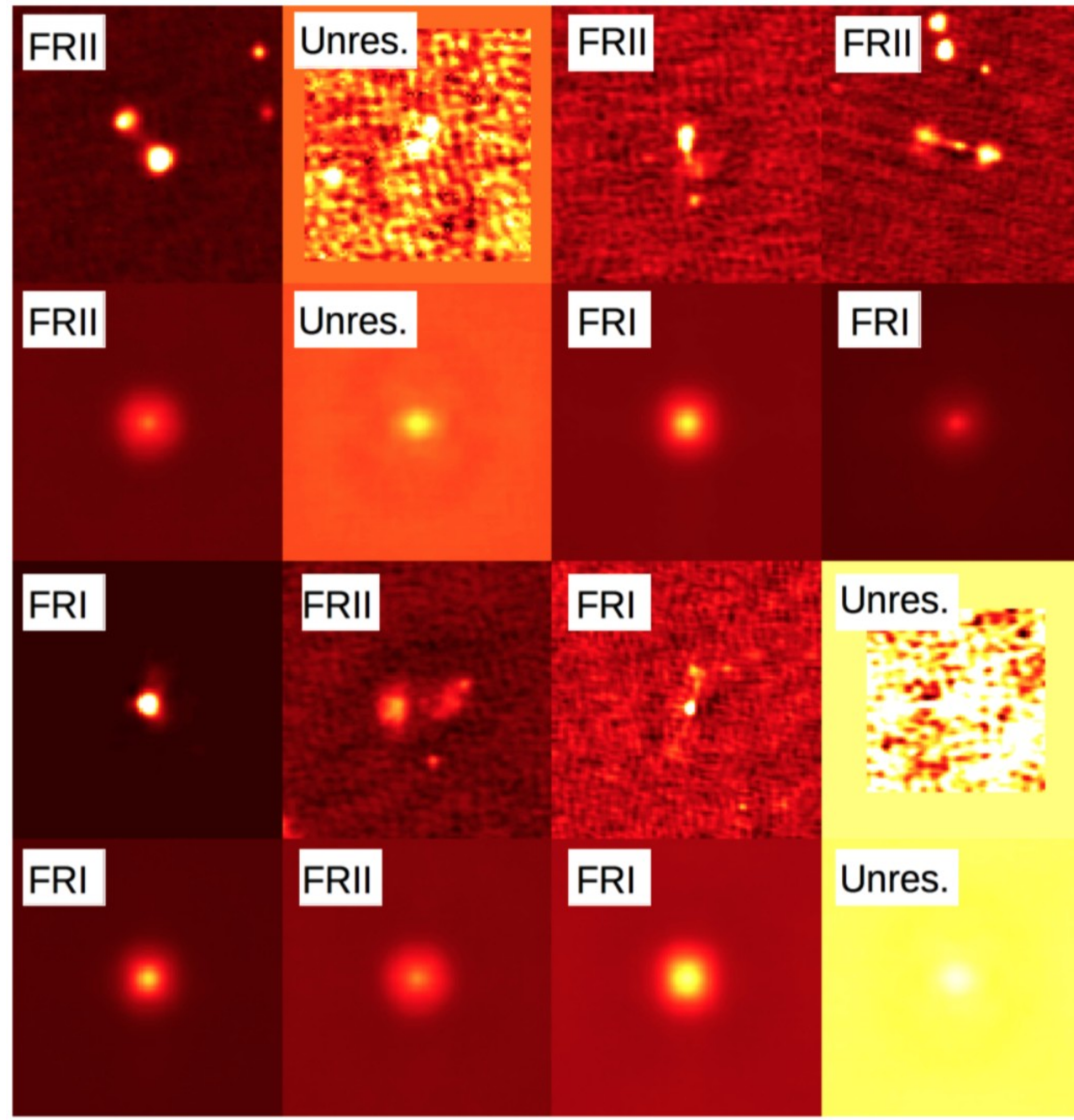
Capsule networks

- Default
- Increase size of filters and stride
- Increase decoder complexity and loss weight





4th conv layer features



CapsNet reconstructions

Results

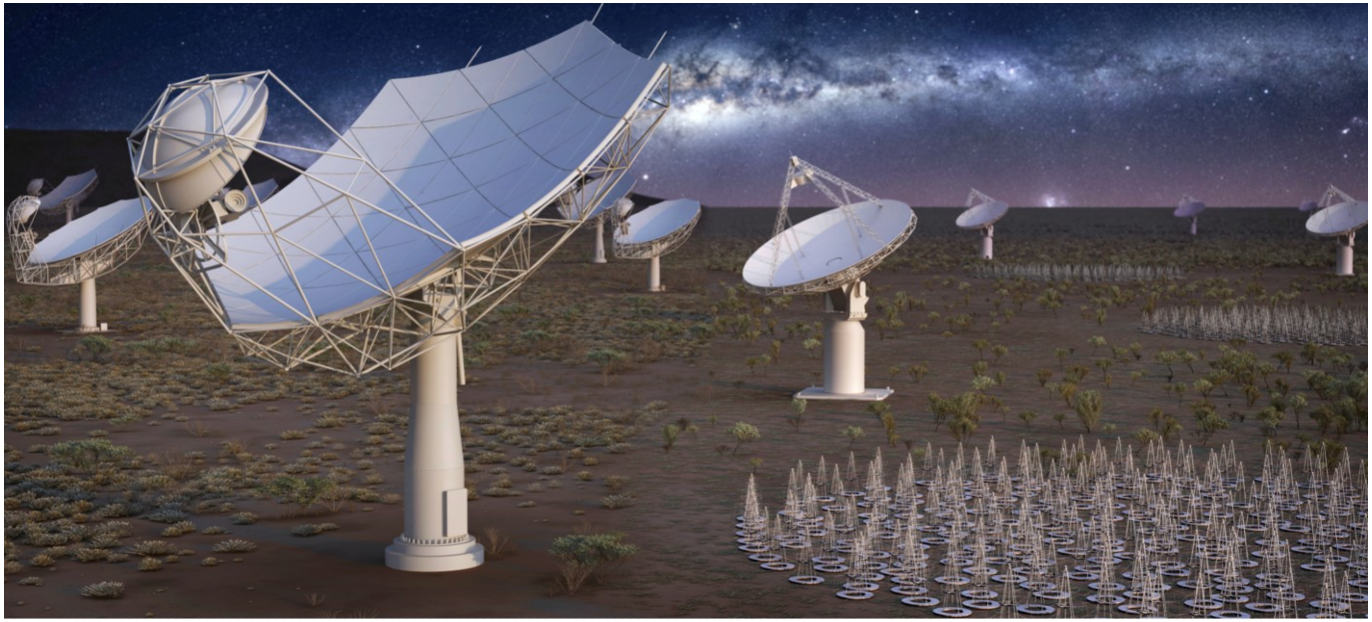
- ConvNet-4 and ConvNet-8 achieve 93.3% and 94.3% average precision respectively. CapsNet attains 89.7%. Transfer learning achieves 94.4%
- Best results across all models are obtained using 4rms masked numpy arrays
- The ConvNet architectures always outperform CapsNet

Possible reasons for performance

- Capsule network does not cope as successfully, perhaps due to preserving all features
- Pooling operation in convolutional networks appears to be advantageous
  - Pooling may help remove undesirable features, allows more degrees of freedom for morphology
- Capsule network might need more original training images to work better

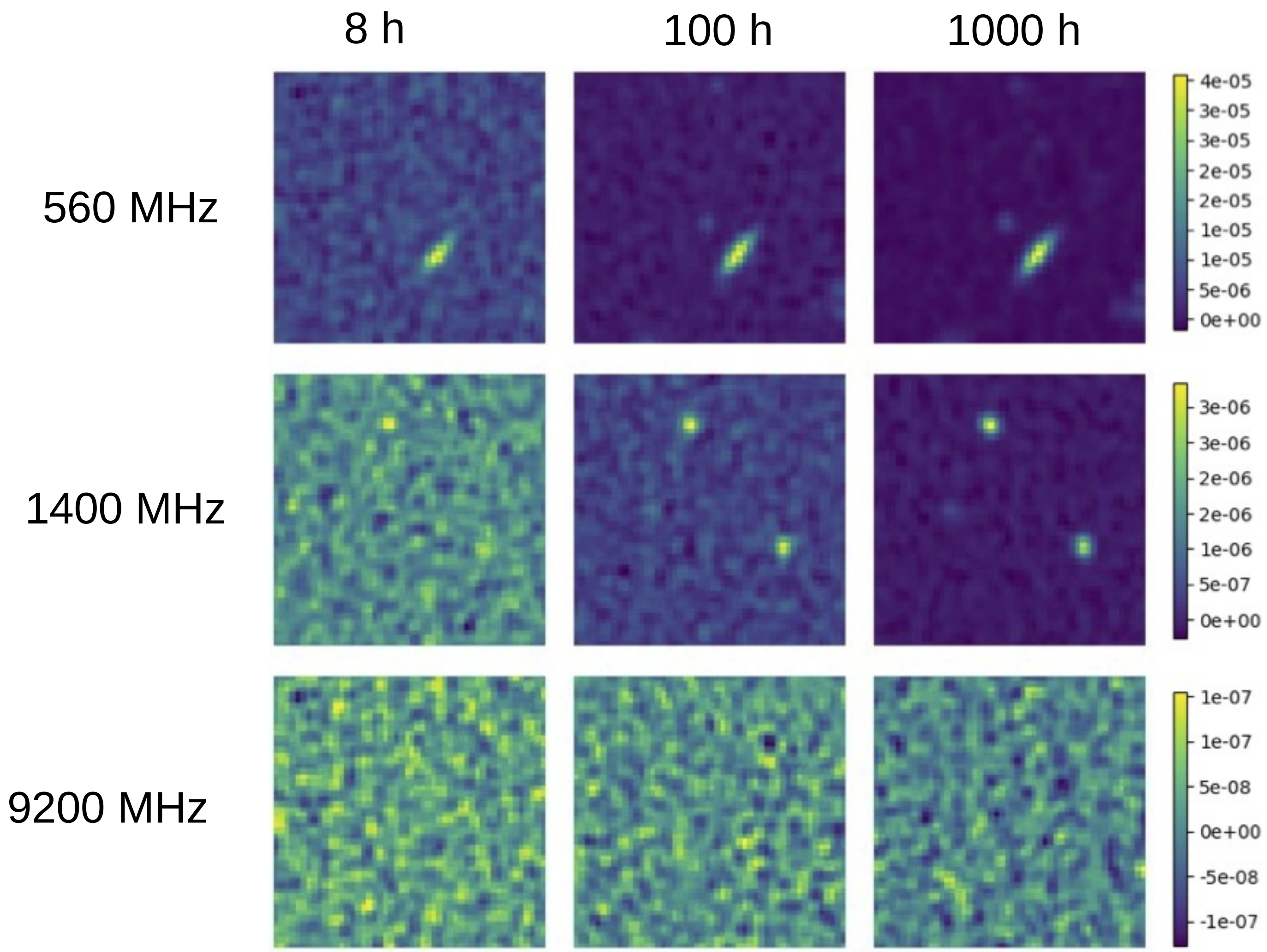
Application to source-finding in the SKA

- Lukic, de Gasperin, Bruggen (2019), *Galaxies*
- Square kilometer array (SKA) is the worlds largest radio telescope
  - >1 square kilometer of collecting area
  - Will discover up to 500 million sources
  - Science data challenge (SDC1)



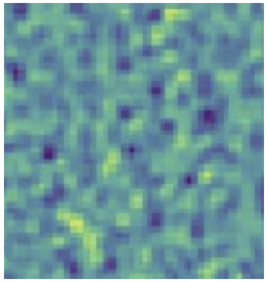
Simulated SKA data

- Sources classified as steep- and flat- spectrum AGN, SFGs
- 4000x4000 pixel training area

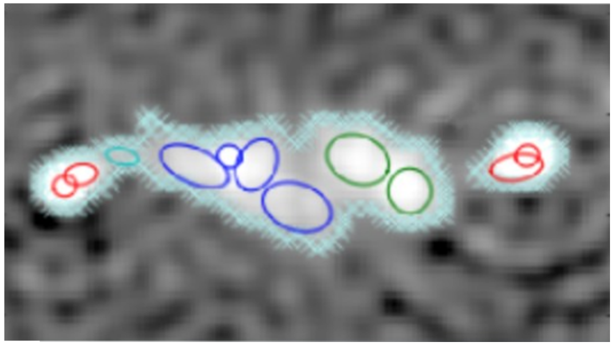


Source-finding and challenges

- A source is defined as a collection of pixels above some value
  - Correlated noise in radio

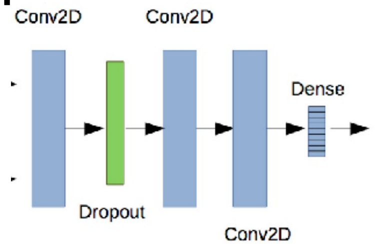


- At lower SNRs there is more difficulty in grouping the pixels belonging to a particular source
- Fitting Gaussians to sources is a common method – we used PyBDSF

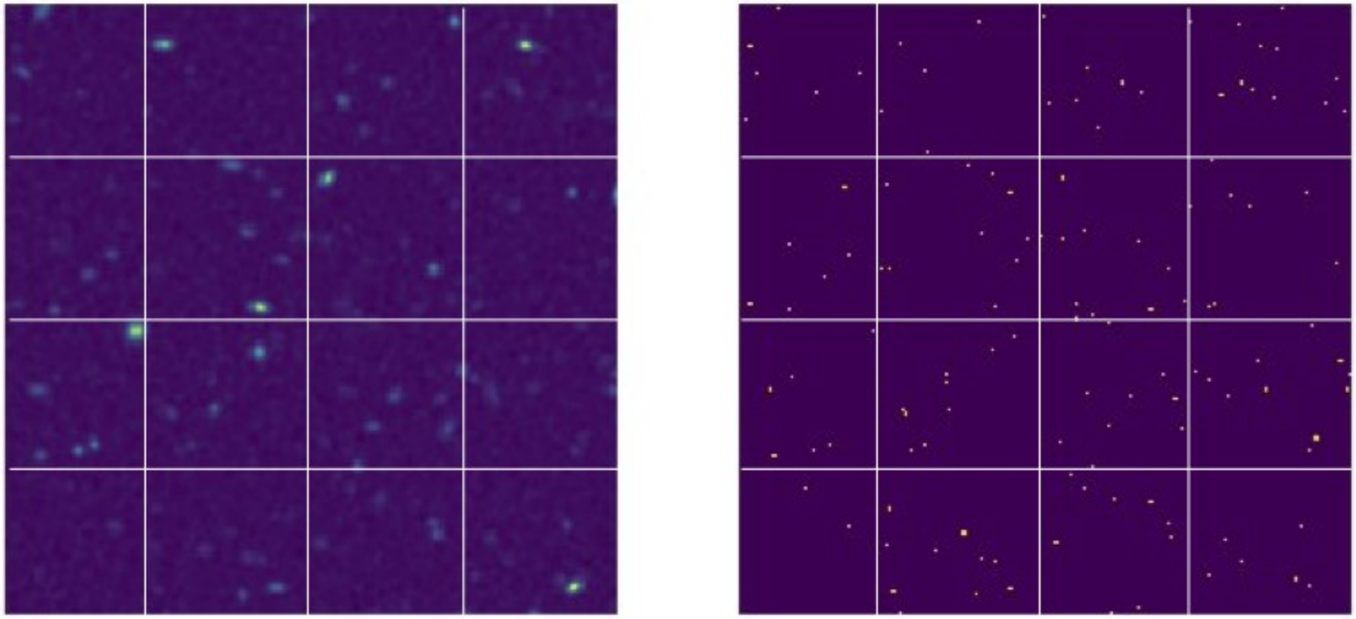


ConvoSource

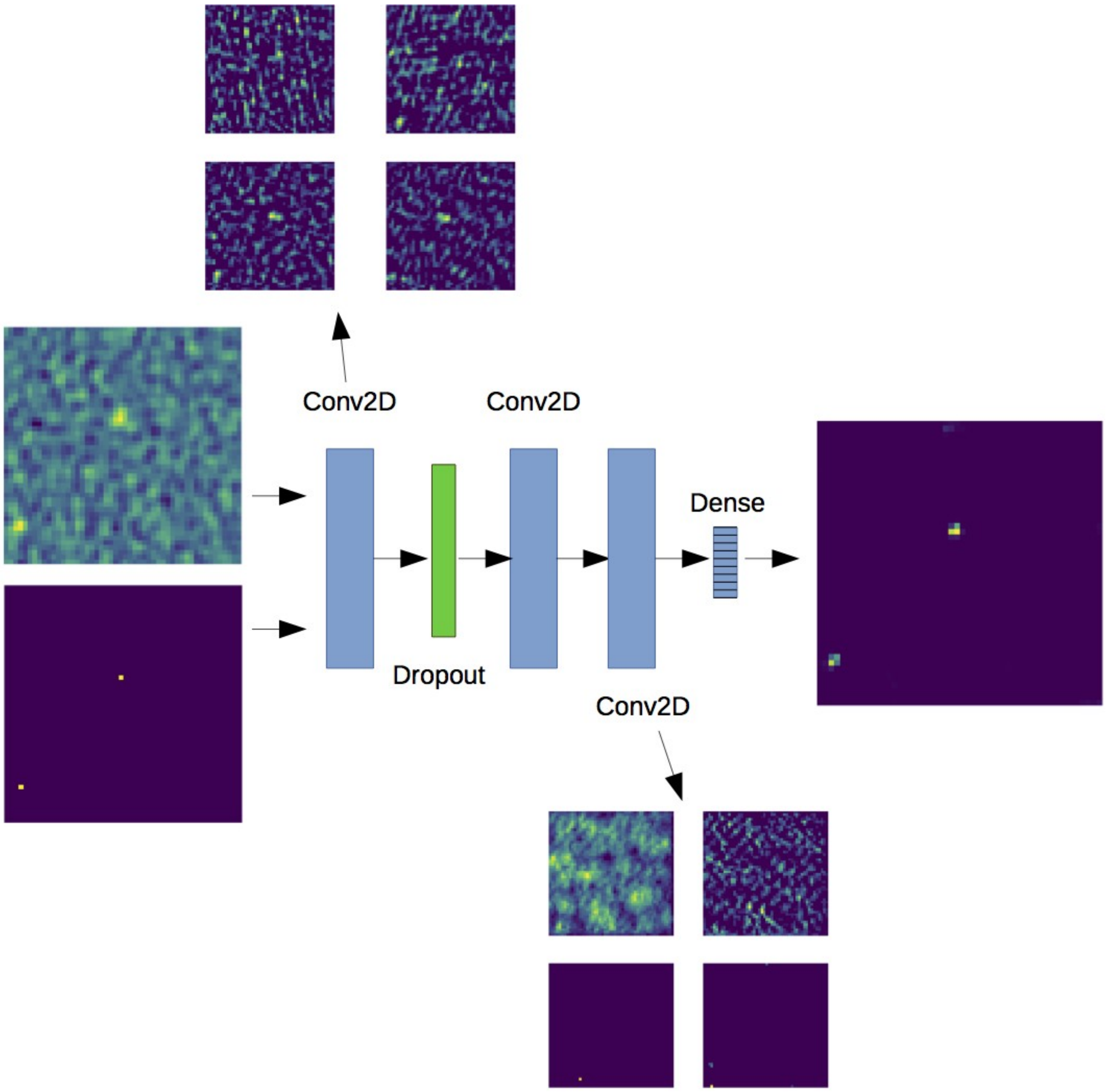
- We developed ConvoSource
  - CNN, real maps → solution maps



- Image augmentation feature may boost performance
- In Lukic et al. (2019), we created 50x50 pixel maps, spaced 50 pixels apart



ConvoSource architecture



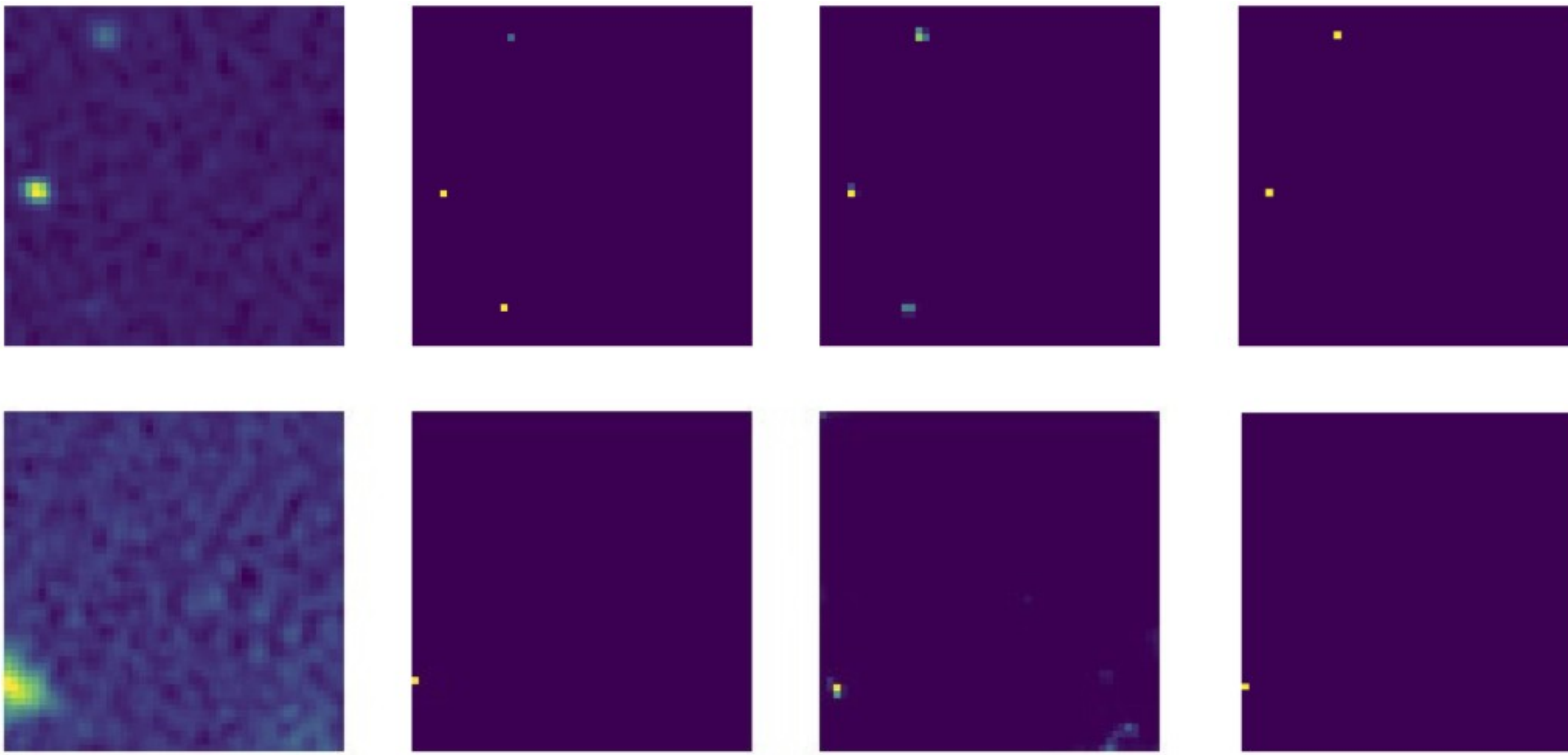
Results

- Source-finders compared using the F1 score
  - Summary of precision and recall
- Lower SNRs: ConvoSource better in recovery of SFGs, PyBDSF better in recovery of SS and FS sources
- The opposite effect is seen at higher SNR
  - The SNR, frequency and exposure time determines whether PyBDSF or ConvoSource will perform better

SNR=2 example

Eg. 1

Eg. 2



Real image | Source locations | ConvoSource | PyBDSF

ConvoSource summary

- ConvoSource
  - outputs pixel values with range 0 to 1
  - sometimes outputs sources spread over several pixels
  - detects more true positives but also more false positives

Conclusions

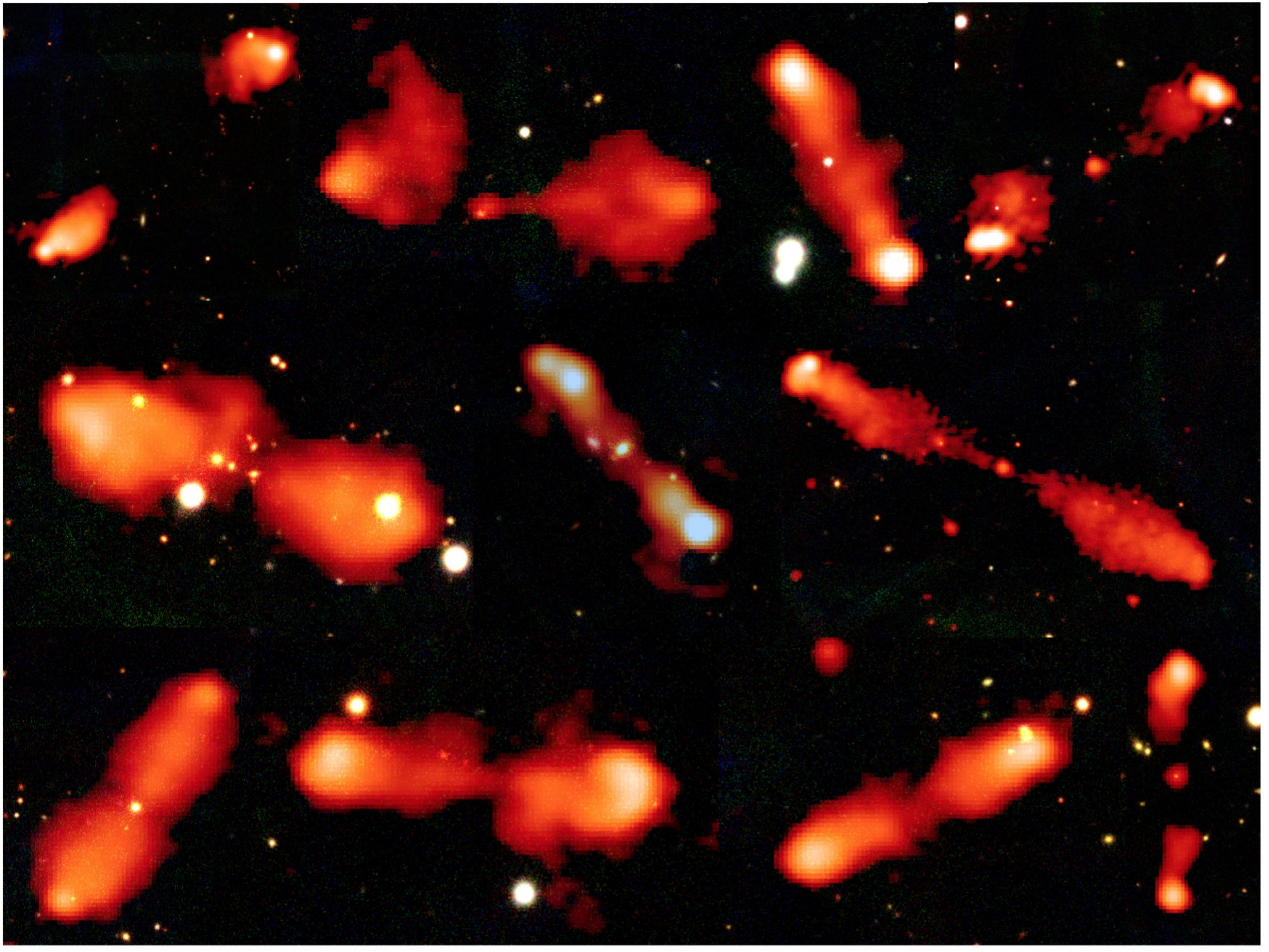
- Machine learning techniques are of increasing importance in astronomical applications
- We have shown it is possible to classify sources according to
  - Number of components
  - Fanaroff-Riley class
- We have shown that a deep learning method can be a competitive source-finder
  - Convolutional neural network
- ConvNets outperform CapsNets given the dataset type and size

Supplementary material

Compact and extended radio sources

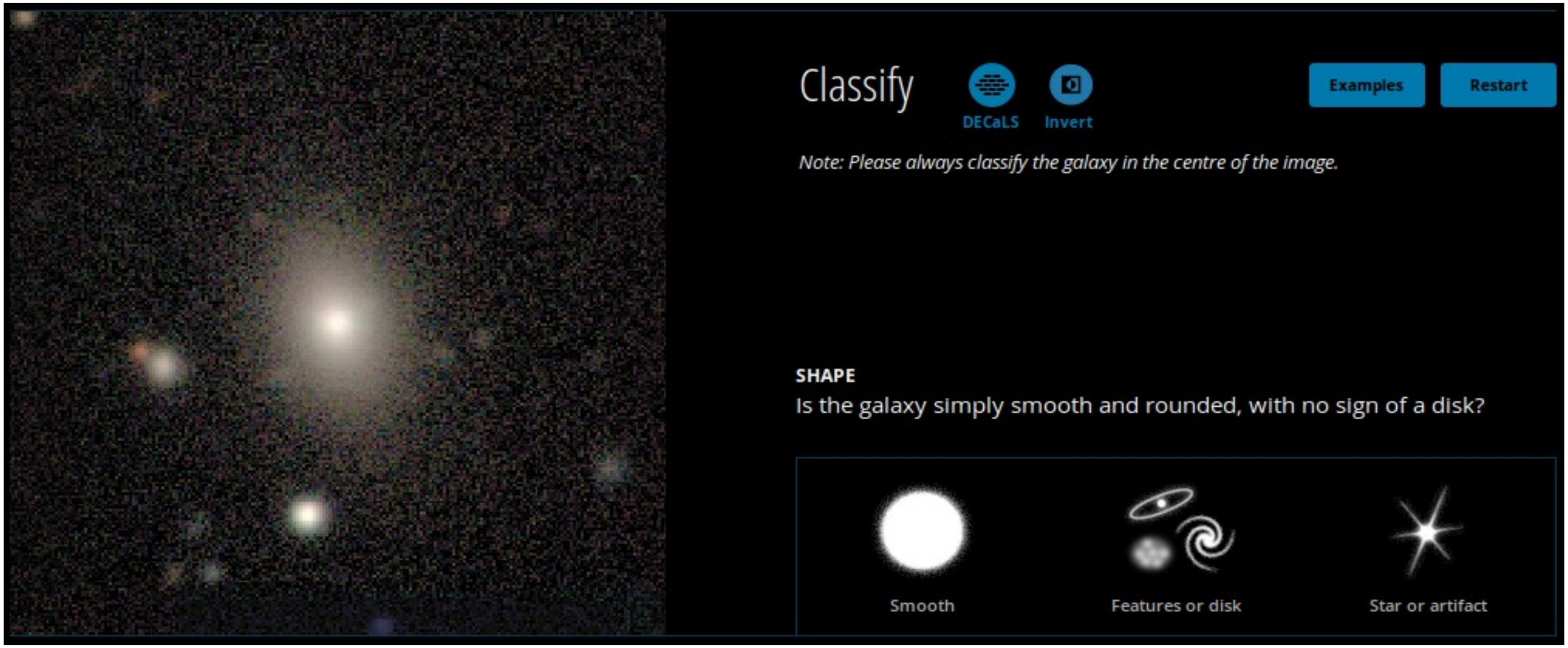
- Focus on classifying between types of radio-loud AGN
- Compact, point-like sources
  - Unresolved by the telescope
  - Generally simple morphologies
  - Some extended sources can be compact
- Extended sources
  - Resolved by the telescope
  - Broader range of morphologies

FRII possible morphologies



Machine learning in astronomical images

- Citizen scientists asked to describe optical galaxy morphologies



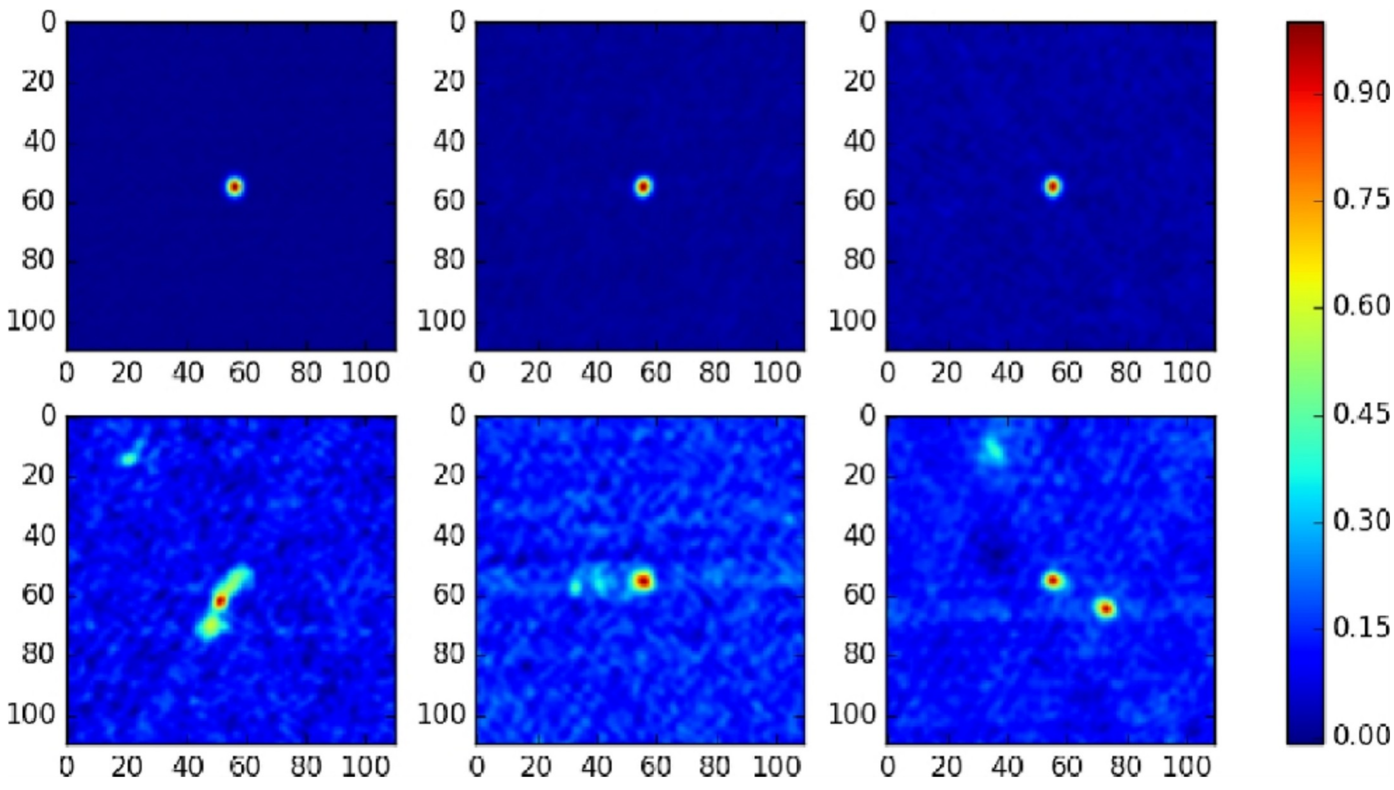
- Determine the probability that a galaxy belongs in a particular class (regression problem)
- Winning solution used **convolutional neural networks**

Two-class problem

- First manually tuned a CNN to distinguish between two morphological extremes

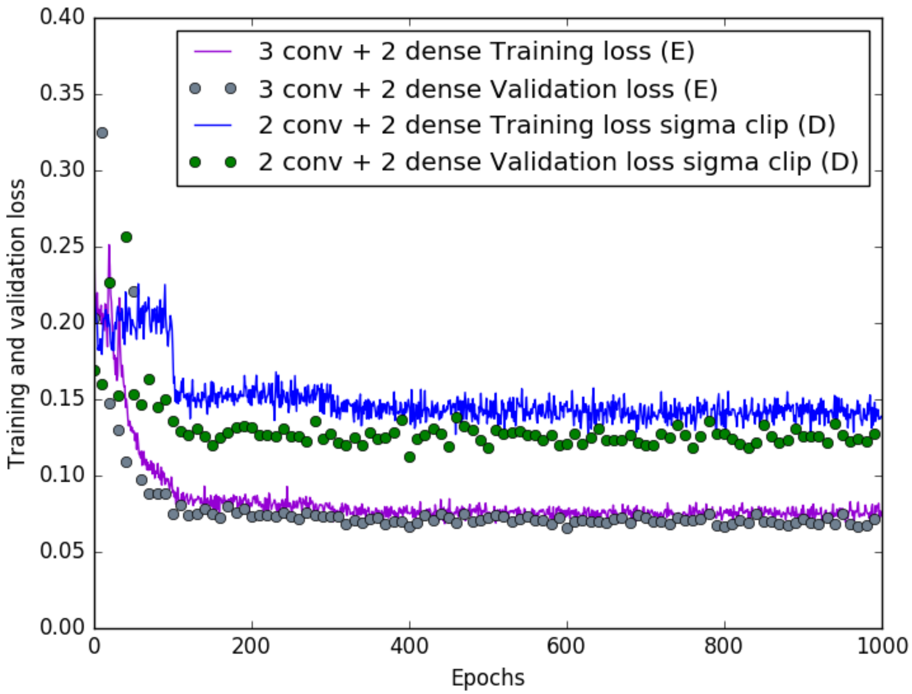
Compact

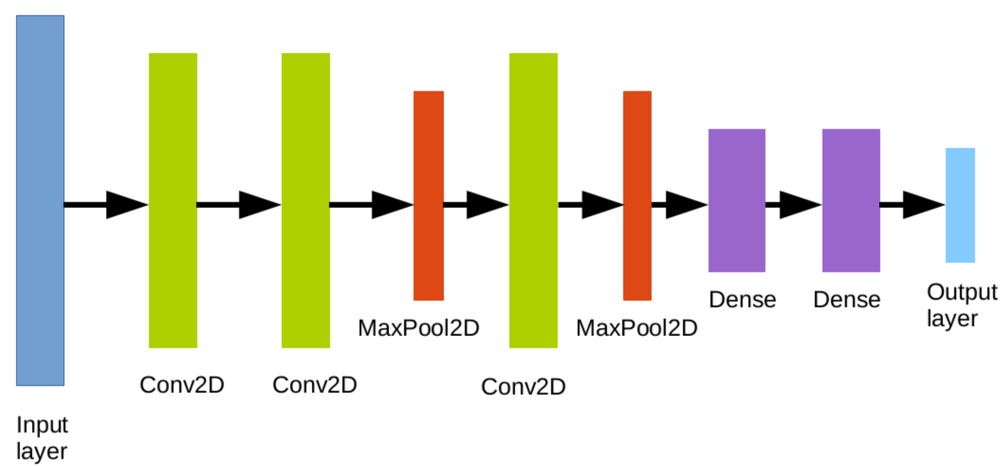
>=3 component extended



Results for two classes

- 3 conv + 2 dense architecture optimal
- Trained with original and augmented images
- Classification accuracv > 97 %



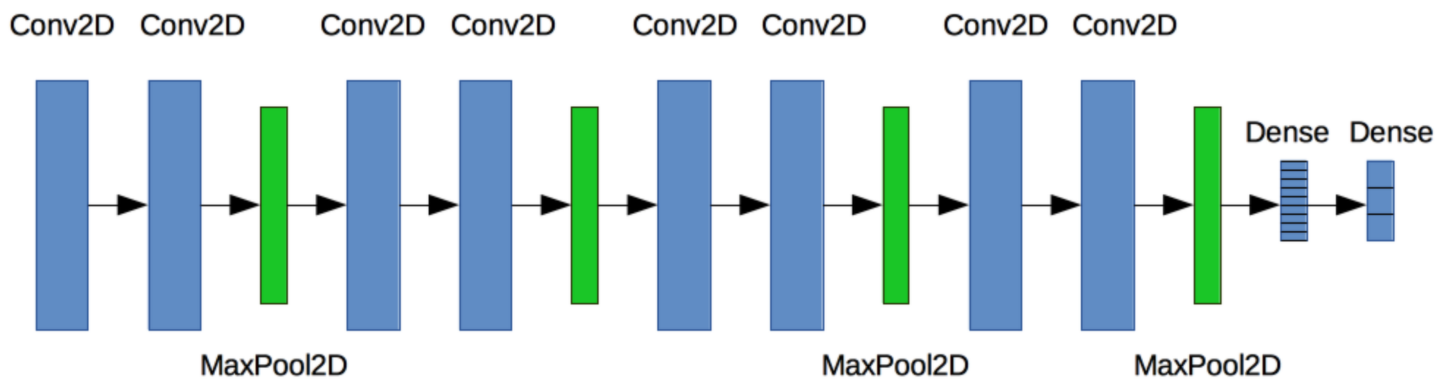


Lasagne network parameters

- Tuned neural network parameters manually to find optimal setup
  - Batch size 8
  - Learning rate set to 0.001 at start, reduced by factor of 10 at four points during training
- Categorical cross-entropy cost function
- Train for 1000 epochs
- Mini-batch Stochastic Gradient Descent (SGD) with Nesterov momentum 0.9 and weight decay of 0

Convolutional neural networks

- ConvNet-4 and ConvNet-8



- Keras library
- Learning rate 0.001, cross-entropy cost function
- Train for 50 epochs, batch size of 100
- Adam optimizer

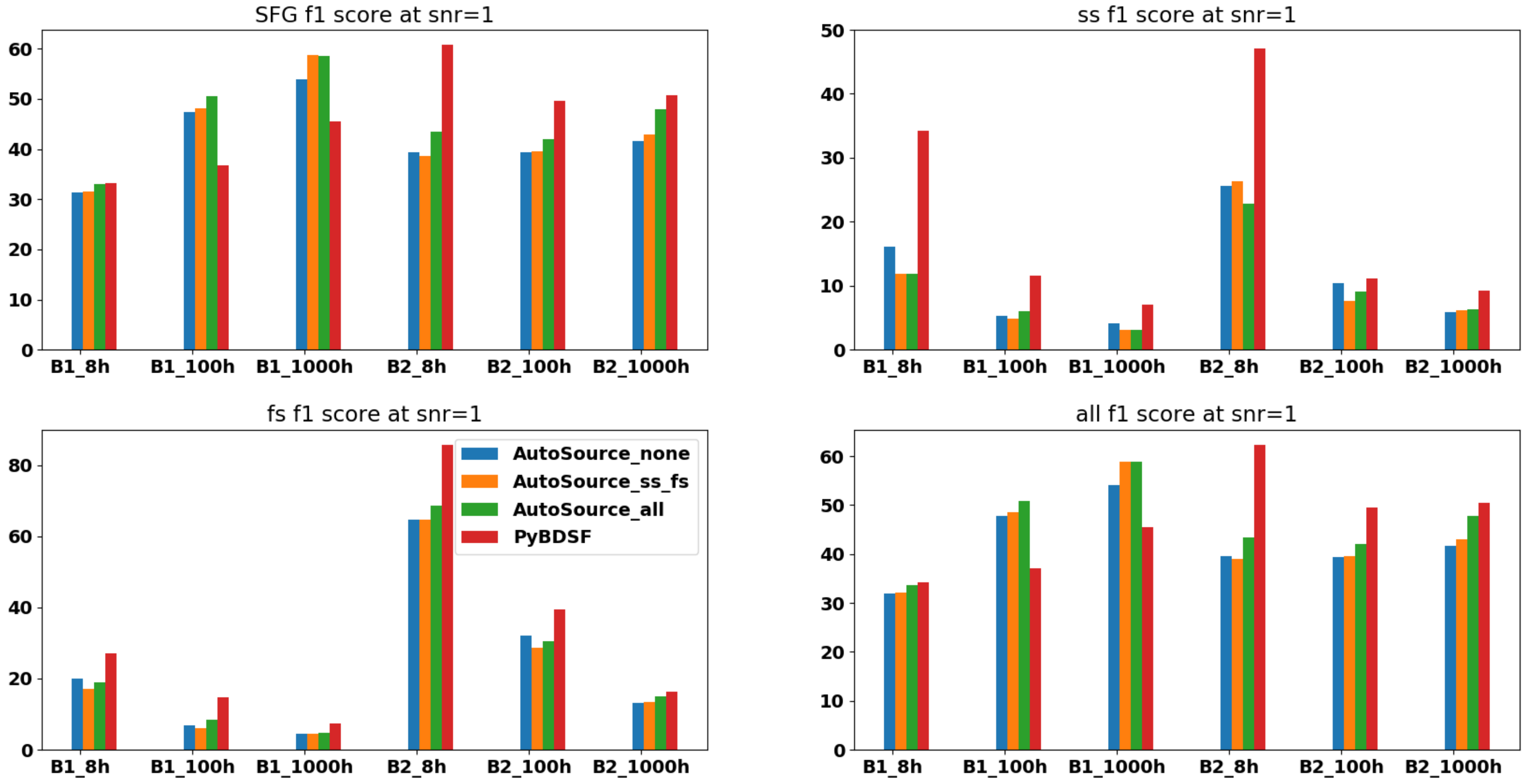
Source-finding based on deep learning

- CosmoDeep (Gheller et al. 2018) detects extended extragalactic radio sources (cluster of galaxies,filaments)
- ClaRAN (Wu et al. 2018) detects individual radio sources in an image and classify according to the number of peaks and components
- DeepSource (Sadr et al. 2019) presents a deep learning algorithm to find point sources in simulated images
- ConvoSource (Lukic et al. 2019) – the first application of a CNN to source-finding, across point and extended sources

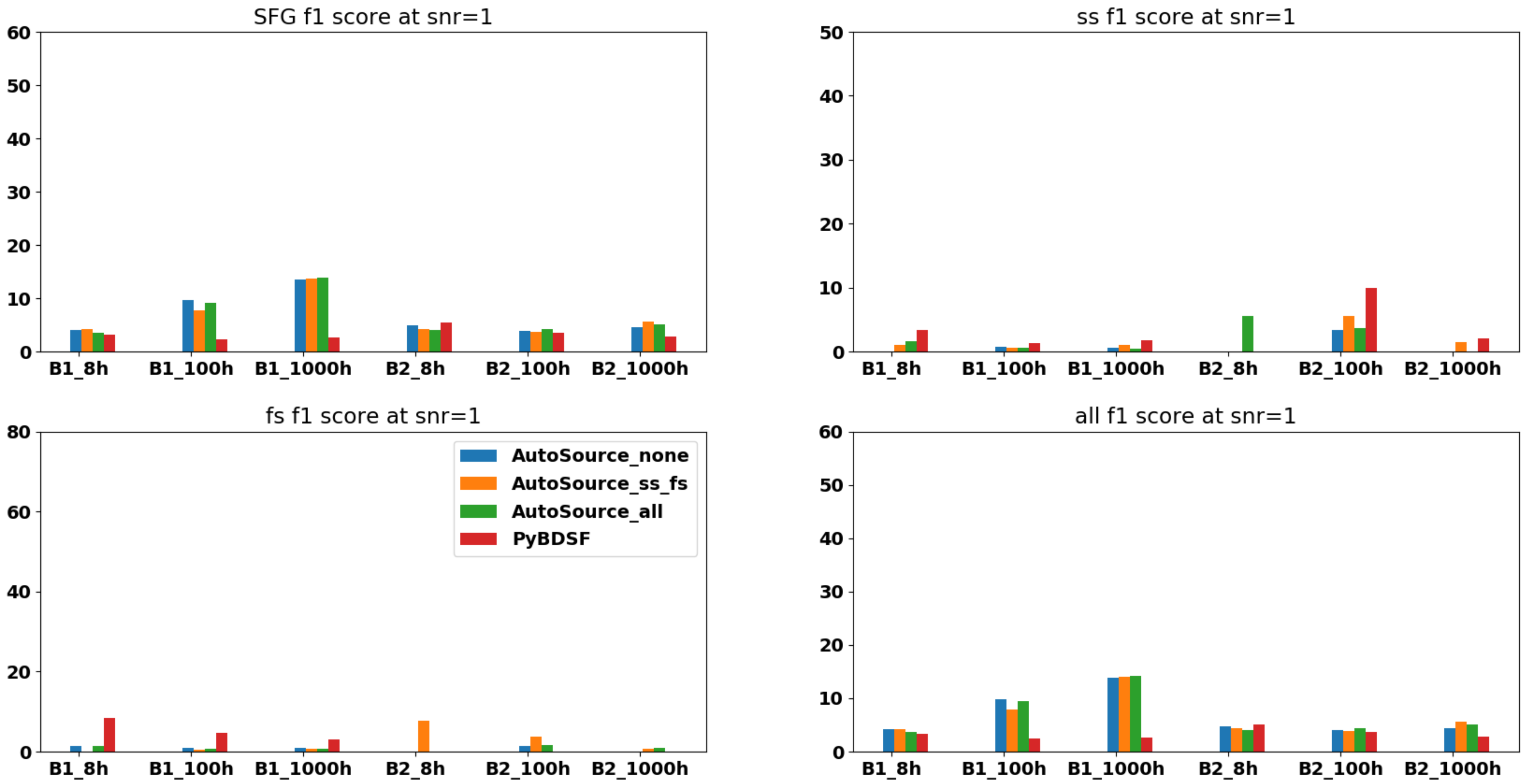
AutoSource parameters

- Keras library
- 5120 (80%) original training images, 1280 (20%) for testing
- Early stopping with patience of 5 epochs
- 16, 32 and 64 filters, with a filter size of 7, 5 and 3 in the first, second and third convolutional layers
- A dropout layer with dropout fraction of 0.25
- Stride of 1 pixel
- Batch size of 128
- We use the Adadelta optimiser with a default learning rate of 1.0, decay of 0 and a rho of 0.99.

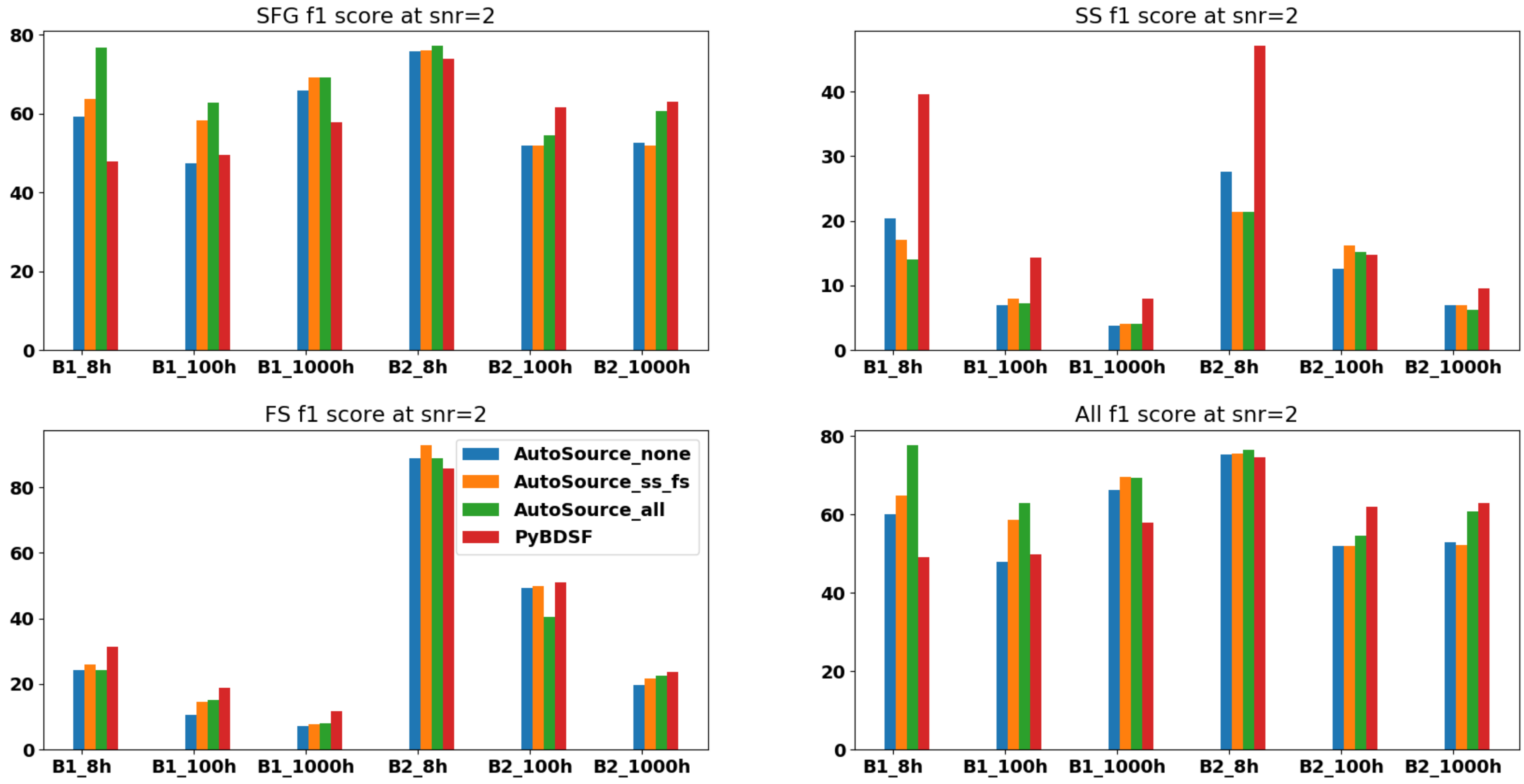
SNR=1


SFG f1 score at snr=1
ss f1 score at snr=1
fs f1 score at snr=1
all f1 score at snr=1
B1_8h
B1_100h
B1_1000h
B2_8h
B2_100h
B2_1000h
AutoSource_none
AutoSource_ss_fs
AutoSource_all
PyBDSF

SNR=1 randomised

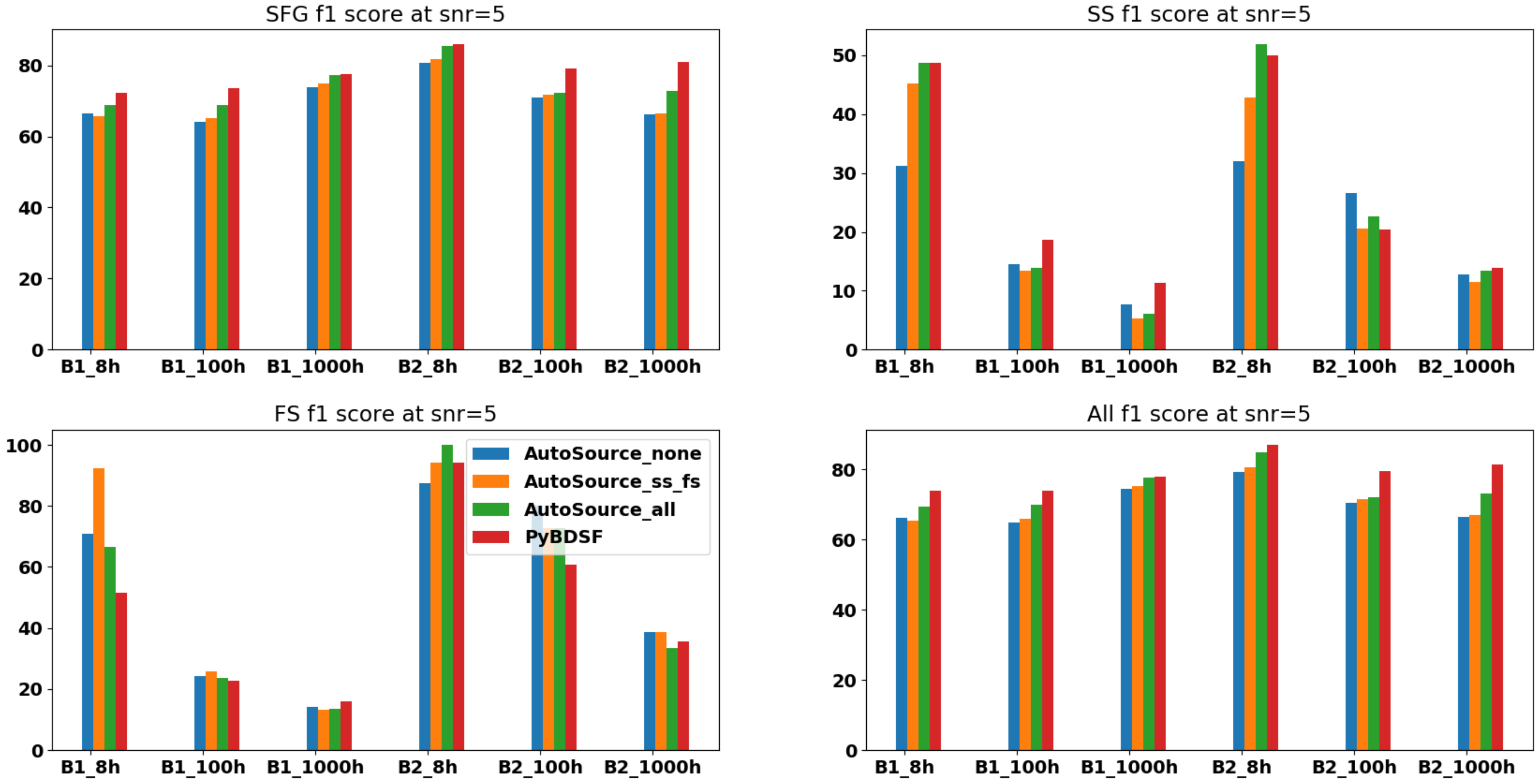


SNR=2

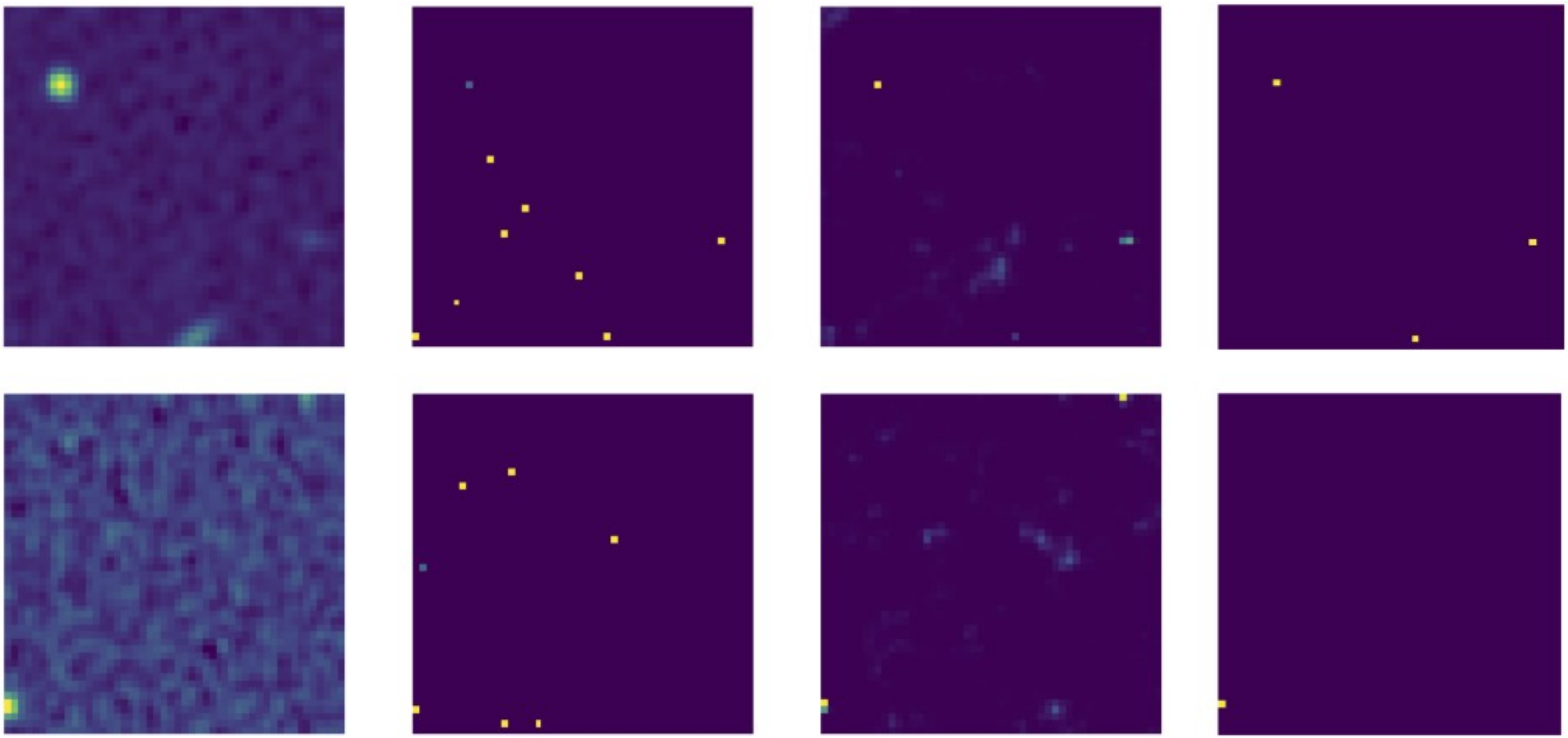


SFG f1 score at snr=2
SS f1 score at snr=2
FS f1 score at snr=2
All f1 score at snr=2
B1_8h
B1_100h
B1_1000h
B2_8h
B2_100h
B2_1000h
AutoSource_none
AutoSource_ss_fs
AutoSource_all
PyBDSF

SNR=5


SFG f1 score at snr=5
SS f1 score at snr=5
FS f1 score at snr=5
All f1 score at snr=5
AutoSource_none
AutoSource_ss_fs
AutoSource_all
PyBDSF
B1_8h
B1_100h
B1_1000h
B2_8h
B2_100h
B2_1000h

SNR=1 example



Real image | Source locations | AutoSource | PyBDSF