Galaxy detection and identification using deep learning and data augmentation

R.E. González, R.P. Muñoz, C.A. Hernández

Presented by Pongsapas Watana

25-May-2021

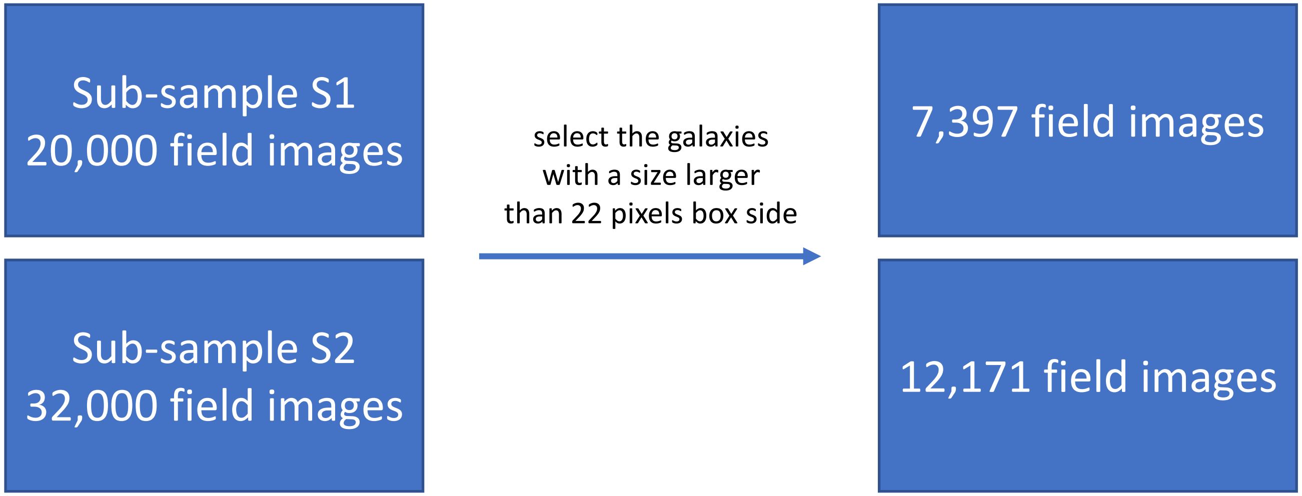
Introduction

- The authors presented a method for galaxy classification and identification with a data augmentation procedure with **AstroCV**, a computer vision library for processing and analyzing big astronomical datasets
- Data sources were **Galaxy Zoo** and **Sloan Digital Sky Survey (SDSS)**
- Object detection was performed by YOLO method

Dataset

Classification consists of five categories:

1. Elliptical
2. Spiral
3. Edge-on
4. Star/Do not know
5. Merger



Dataset

Table 1
Annotations in different subsamples.

Dataset	Elliptical	Spiral	Edge-on	DK	Merge	Total	Number images
Training S1	10 366	4535	4598	223	381	20 103	6 458
Validation S1	1 261	714	723	27	45	2 770	921
Training S2	18 030	7828	7910	350	648	34 766	11 010
Validation S2	2 119	856	873	36	82	3 966	1 161
Custom	705	401	462	474	135	2 177	87

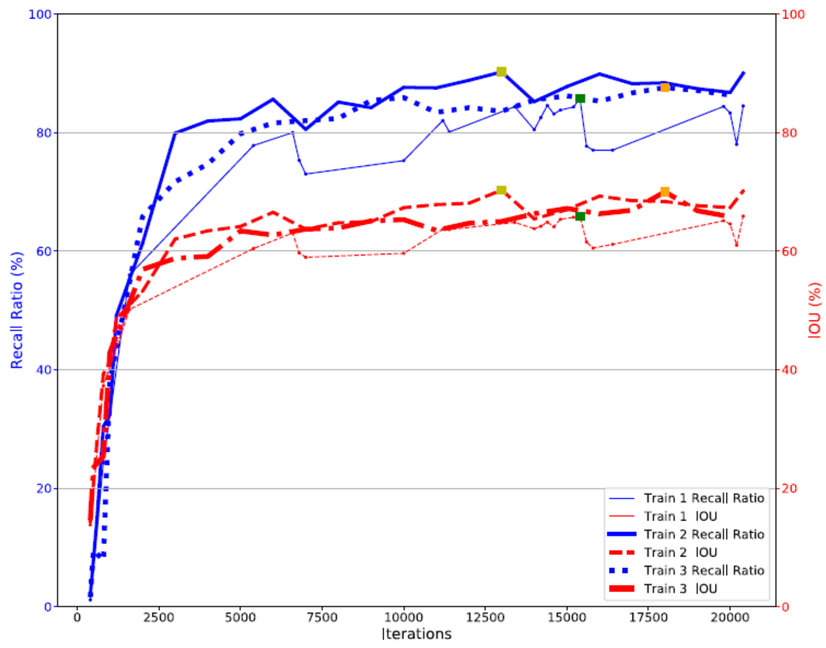
Table 2
Training sets.

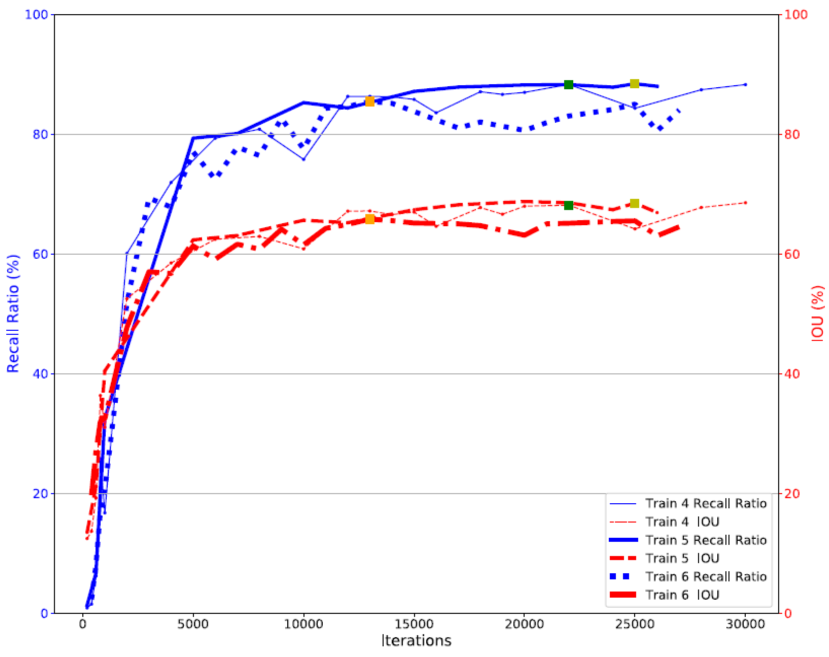
Name	Dataset	Filters	Images
T1	S1	L	6 458
T2	S1	LH	6 458
T3	S2	L	11 010
T4	S2	L+LH+S+SH+Q	55 050
T5	S2	LH+SH	22 020
T6	S2	L+LH+S+SH+Q	32 290

(L = Lupton, LH = Lupton high contrast, S = sinh, SH = sinh high contrast, Q = sqrt; C = custom Hubble sample.)

Training and data augmentation

- Create more robust model by applying color-scale/filter augmentation with common scaling methods: *sinh, asinh, sqrt functions, and Lupton method*
- Trained the model with **YOLO**





T2 returned better results since higher contrast and brightness images built from FITS images gave more information to the training. Especially, in disk galaxies where disk can be seen clearer

Recall and IOU remain similar. However, this time detection and classification is more robust against different contrast stretching functions, instruments or reduction processes.

Results

Table 3
Confusion matrix for galaxy classification using T2. Predicted values (columns) vs. actual values (rows).

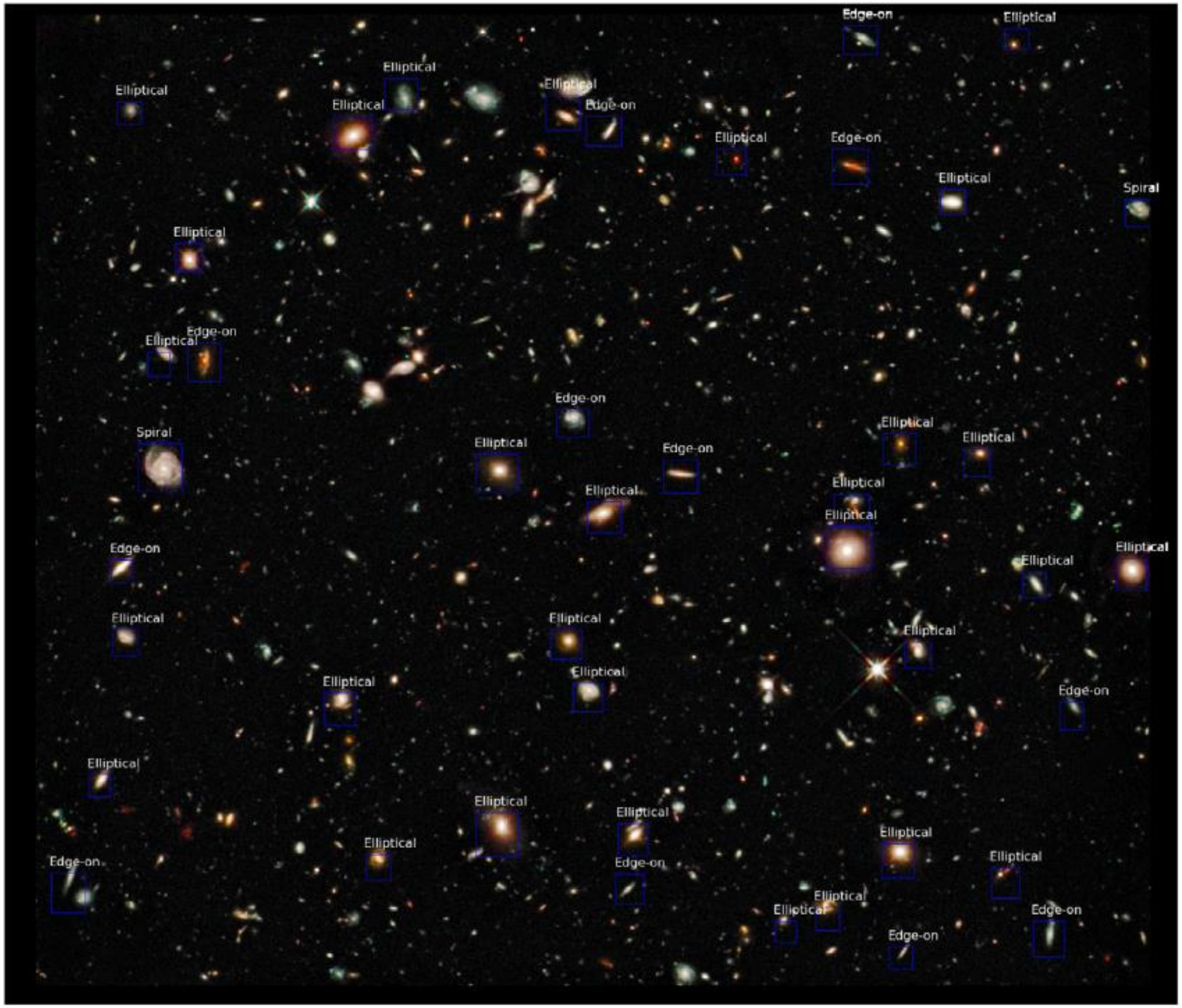
n = 2756	Elliptical	Spiral	Edge-on	DK	Merge	Recall
Elliptical	1172	33	57	1	2	0.92
Spiral	176	469	69	0	3	0.65
Edge-on	96	60	554	0	3	0.78
DK	6	9	3	3	0	0.14
Merge	26	2	3	0	9	0.23
Precision	0.79	0.82	0.81	0.75	0.53	Accuracy = 0.80

Table 4
Confusion matrix for galaxy classification using T4. Predicted values (columns) vs. actual values (rows).

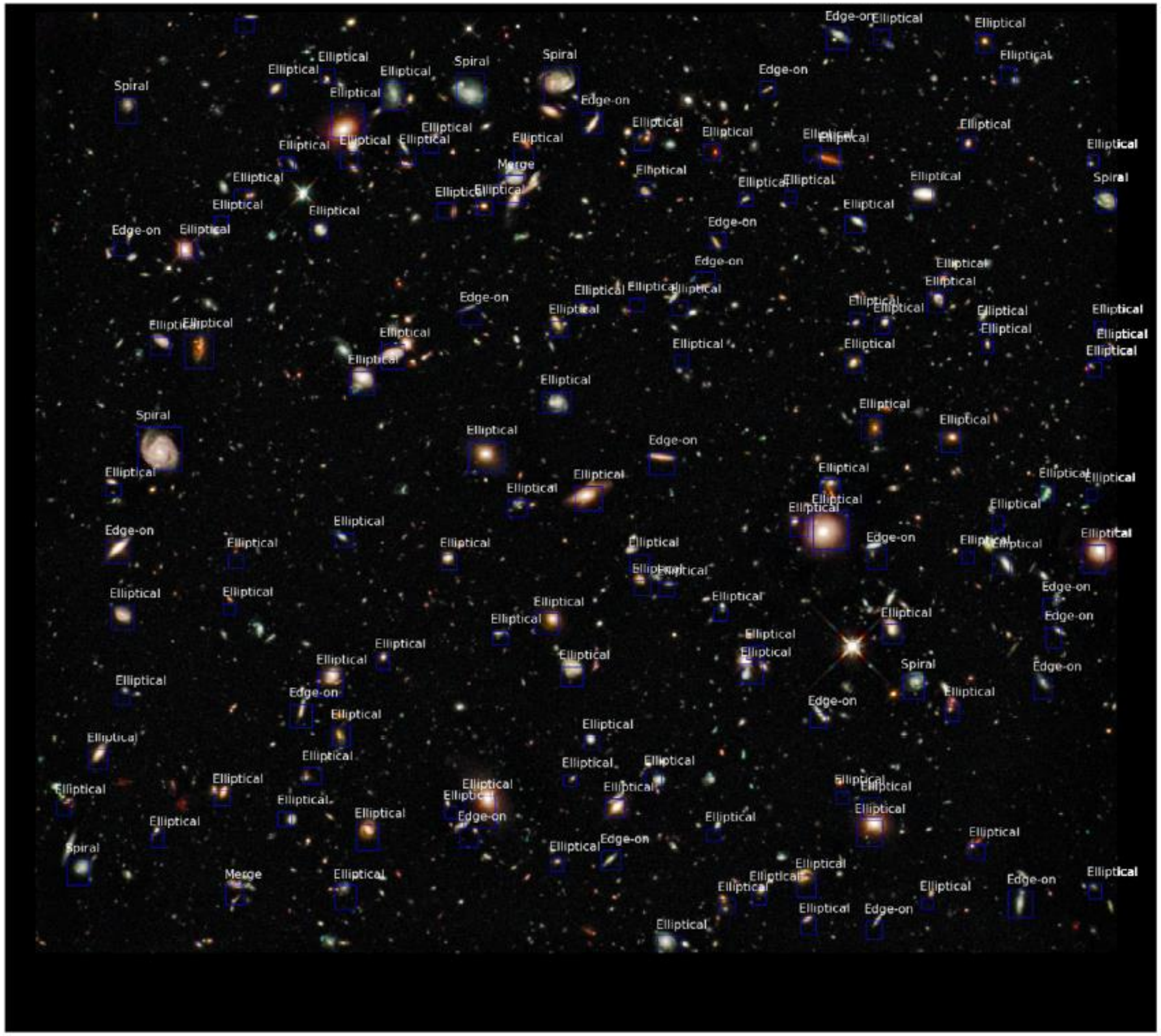
n = 19246	Elliptical	Spiral	Edge-on	DK	Merge	Recall
Elliptical	9525	476	257	0	3	0.93
Spiral	940	2991	305	0	6	0.71
Edge-on	688	499	3063	0	4	0.72
DK	39	36	24	6	4	0.06
Merge	170	129	44	0	37	0.10
Precision	0.84	0.72	0.83	1.0	0.69	Accuracy = 0.81

Table 5
Detection accuracy for different training sets validated with the custom dataset.

Training set	Recall ratio	IOU ratio
T1	0.120	0.236
T2	0.183	0.296
T3	0.192	0.292
T4	0.404	0.423
T5	0.271	0.336
T6	0.364	0.403



Hubble deep field image: T2 (44 detections)



Hubble deep field image: Augmented training T4 (140 detections)