SF2A 2025
A. Siebert, K. Baillié, M. Béthermin, F. Cantalloube, E. Josselin, N. Lagarde, J. Malzac, J. Richard, L. Selliez, O. Venot (eds)

YOLO-CL : DEEP MACHINE LEARNING FOR GALAXY CLUSTER DETECTION WITH LSST

T. V. Phat[1], S. Mei[1,], M. Aguena[2] and S. Ilić[3]

Abstract. Galaxy clusters are powerful and key cosmological probes that trace the evolution of large-scale structures of the universe and inform models of cosmic evolution. The Vera Rubin Legacy Survey of Space and Time (LSST) will provide deep, wide-field optical imaging, enabling the detection of thousands of galaxy clusters up to high redshifts. We are adapting the YOLO-CL convolutional neural network to detect galaxy clusters in LSST survey images. Grishin et al. (2023, 2025)demonstrated that YOLO-CL produces cluster catalogs that are highly complete and pure using SDSS observations and Rubin/LSST DC2 (LSST Dark Energy Science Collaboration (LSST DESC) et al. 2021) simulated images of individual clusters, outperforming traditional detection methods based on photometric catalogs. Future work will involve optimizing the model's performance under survey-like conditions and ensuring robust cluster selection functions across redshift and mass ranges. This effort aims to contribute a scalable, high-fidelity cluster detection pipeline for next-generation cosmological analyses.

Keywords: Machine Learning, Galaxy Clusters, LSST, SDSS, YOLO-CL

1 Introduction

Galaxy Clusters are the largest known gravitationally bound structures in the Universe, comprised of hundreds or even thousands of galaxies held together by gravitational forces exerted by dark matter. Due to their massive size, their abundance and spatial distribution are extremely sensitive to the underlying cosmology which makes them key cosmological probes to constrain parameters such as matter density Ω_m, the amplitude of matter density fluctuations σ_8 and the growth rate of cosmic structures over time. They are also excellent tracers of Large Scale Structures such as the Cosmic Web and offer insights into the inner-workings of galaxy evolution as well as the environmental effects that influences cluster members themselves. All the while, traditional galaxy detection methods rely heavily on a multitude of intermediate catalogs (from galaxy photometry to photometric redshift), each of which introduces noise as well as model bias. With the advent and rapid optimization of machine learning, we can observe an increasing amount of applications in the field of Astrophysics. The very subject of this project is one such example : YOLO-CL, a deep machine learning, object detection algorithm specializing in galaxy clusters detectionRedmon et al. (2015); Grishin et al. (2023, 2025). This model aims to sidesteps the constraints of previous traditional methods by operating directly on color-calibrated images, meaning that the model learns patterns directly from the data rather than imposing model-based assumptions. And with upcoming large surveys like the Vera Rubin Telescope Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al. 2009; Ivezić et al. 2019), Euclid (Laureijs et al. 2011) and Nancy Roman Space Telescope (Eifler et al. 2021), we expect to observe millions of groups and clusters of galaxies across a wide redshift range. But to extract cosmological information from them, we will need cluster catalogs that are both highly complete and pure. That is why cluster detection remains a central challenge in survey cosmology and why YOLO-CL, a machine learning approach is an interesting avenue to explore.

[1] APC Laboratory (IN2P3), 75013 Paris, France
[2] Obervatory of Trieste, Italy
[3] IJCLab, Orsay, France

2 YOLO-CL

2.1 The model

YOLO-CL is a deep machine learning method optimized for galaxy clusters detection by Stéphane Ilic. It is based on YOLOv3 (Redmon & Farhadi 2018) which is a real-time object detection algorithm, with a Convolutional Neural Network architecture (CNN). In a typical CNN configuration, the model is inputted an image, it extracts features such as shapes, edges, forms and textures through a multitude of convolutional and pooling layers producing feature maps. These pieces of information are then flattened and processed through a fully-connected layer where the classification step is performed, outputting a probabilistic distribution of what the model thinks was in the input image. In our situation, things are a bit different where the only class we are interested in are galaxy clusters, so technically there is only 1 class, making the classification step simpler. However, what makes YOLO-CL and YOLOv3 an objection detection algorithm is the extra step of bounding box regression that is performed. How all of this is done is as follows: the input image is divided into a grid and for each cell the model predicts whether a galaxy cluster is present, where it is (via bounding box coordinates), and how confident the model is. Just like a regular CNN, this happens in one forward pass, but instead of one label, YOLO-CL gets multiple localized predictions - each with a position, a score, and a label. When training a machine learning model like YOLO-CL, it is necessary to minimize as best as possible its loss function which is a mathematical function that measures how wrong a model's predictions are compared to the true values. In our case, that function is the following (Redmon et al. 2015; Grishin et al. 2023) : $L = L_{obj} + L_{bbox} + L_{class}$ with L_{obj} representing the "objectness loss", penalizing the model if it wrongly predicts whether there is a cluster or not. L_{bbox} is the loss of the bounding box, it penalizes the model if the predicted bounding box does not overlap with the ground truths. Grishin et al. (2023) and Grishin et al. (2025) use the generalized Intersection over Union (gIOU) loss (Rezatofighi et al. 2019) which takes into account, in the case where there is no overlap at all, the distance which separates the two non-overlapping boxes in order obtain additional optimizing information. And finally, L_{class} is the "classification loss" but as previously stated, the YOLO-CL case is simpler since we are only interested in the class "cluster" all other classes have been removed and thus L_{class} is inconsequential.

2.2 The training and validation data

2.2.1 SDSS

We presented results from Grishin et al. (2023) and Grishin et al. (2025). The first study case of the model was with the Sloan Digital Sky Survey's (SDSS), using as training sample a cluster catalog produced with the red-sequence Matched-filter Probabilistic Percolation (redMaPPer) algorithm. In particular, they use the redMapper Data Release 8 (DR8) catalog from Rykoff et al. (2014). redMaPPer is a red sequence cluster finder designed for large photometric surveys. The catalog of 26,111 clusters over the redshift range z $\in$ [0.08, 0.55] was built on $\sim$10,000 square degrees of the SDSS DR8 data release. After some more filtering (to avoid troublesome areas such as edges) they obtain a training and validation sample of 24 406 redmapper clusters, of which half is used for training and the other half, combined with blank fields for validation. In Fig. 1 we show the resulting performance of the model on resampled SDSS images.

To validate the YOLO-CL performance independently of redMaPPer, Grishin et al. (2023) applied the network to galaxy clusters detected by X-ray emission and published in the MCXC2021 catalog (Piffaretti et al. 2011). As a comparison, they show the efficiency of redMaPPer on the same MCXC2021 cluster sample. These results are shown in Fig. 2.

2.2.2 DESC DC2*

In preparation for the upcoming LSST data, Grishin et al. (2025) adapted and applied the model to simulations. For that, They apply a machine learning technique for training called "transfer learning" where a model that was trained on one dataset is adapted to a different, but related, task or dataset. The idea is that the model

*Data Challenge 2. It refers to a large-scale simulation effort led by the LSST Dark Energy Science Collaboration (DESC) to produce realistic synthetic LSST observations. The main goals of DC2 is to 1 : test and validate LSST pipelines and algorithms (such as galaxy cluster detection, photometric redshift estimation, weak lensing analysis, etc.); 2 : provide a benchmark for machine learning approaches like YOLO-CL; and 3 : simulate observations based on realistic galaxy distributions, including dark matter halos, with synthetic imaging processed using the LSST Science Pipelines.

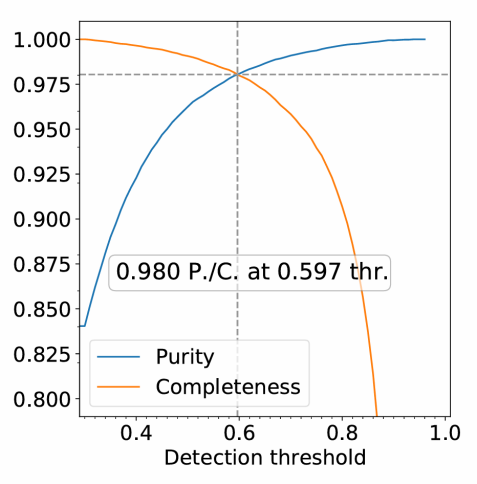

Fig. 1. From Grishin et al. (2023) : "The YOLO-CL cluster catalog completeness and purity as a function of the detection threshold when using 1024x1024 resampled SDDS images. As observed, we obtain a completeness and purity of 95% at an optimal threshold of 60% giving YOLO-CL an overall performance of rather high quality."

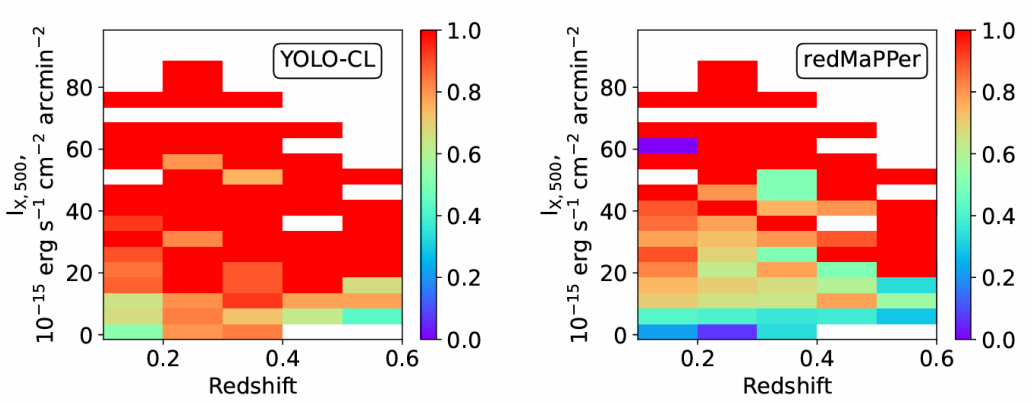

Fig. 2. From Grishin et al. (2023) : "The YOLO-CL and redMapper MCXC2021 cluster detection completeness as a function of redshift and mean X-ray surface brightness. As observed from this comparison,YOLO-CL outperforms RedMaPPer when comparing detections completeness of the same MCXC2021 clusters, showcasing a flat selection function with X-ray surface brightness"

has already learned useful visual patterns - like textures, edges, or in our case, what clusters tend to look like - from one domain, and they don't want to throw all that knowledge away when they move to a new one. In that case, they trained YOLO-CL first on real SDSS images of galaxy clusters (24 406 SDSS images). These clusters are real, visually rich, and provide strong supervised labels. Then, they fine-tuned the same model on an hybrid sample of SDSS and simulated LSST DC2 images (2342 images), which are deeper and noisier, but fewer in quantity. More specifically, training and validation each were performed respectively with combined halves of SDSS and LSST DC with the validation step having additional blank fields added as well. They obtain a Completeness and purity of 94% (Fig. 3) for halos with $M_{200c} > 10^{14} M_{\odot}$ at $z \lesssim 1$.

With a more detailed look in Fig. 3 we can observe an even better performance for masses $M_{200c} > 10^{14.6} M_{\odot}$ at $0.2 < z < 0.8$. Something noteworthy as well are the false-positives (see Fig. 4) : they obtained a false positive rate of $\sim 6\%$ but nearly all false positives matched halos with masses between $10^{13.4}$ and $10^{14} M_{\odot}$. While some detections are labeled as false positives because their masses fall below the $10^{14} M_{\odot}$ training cutoff, most of them are actually real halos just below the threshold - they fall within the expected scatter and often look visually indistinguishable from lower-mass clusters. So these are not random noise detections - they are physically meaningful low-mass structures that simply fall below the catalog's official threshold. The main takeaways is that even when YOLO-CL detects a false positive, it tends to pick up real overdensities - mostly low-mass

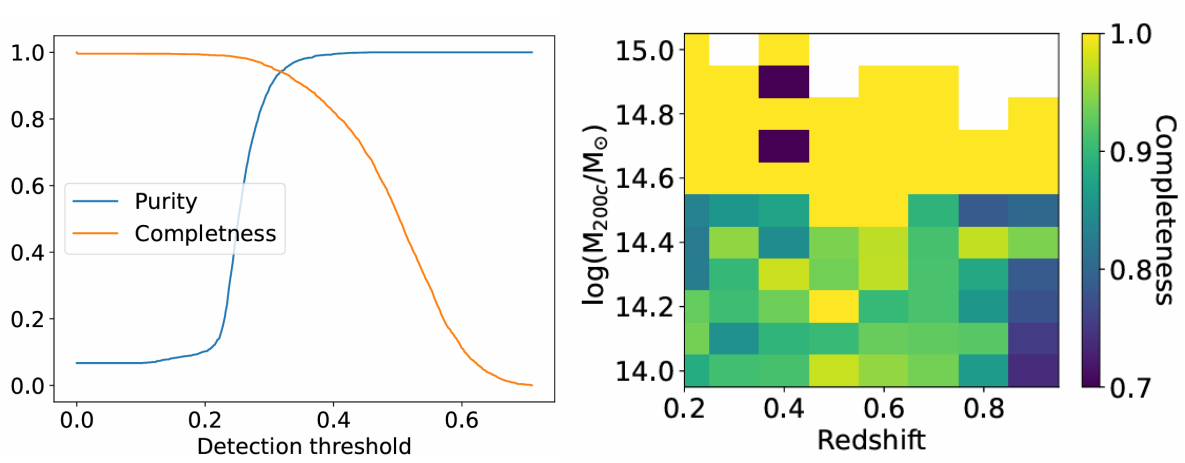

Fig. 3. Left: From Grishin et al. (2025) : "Purity and Completeness of the YOLO-CL DC2 detection catalogs for 1024x1024 images as a function of the detection threshold. The optimal value of their intersection is 94%." **Right:** From Grishin et al. (2025) : "YOLO-CL DC2 detection completeness as a function of redshift and halo mass. The colored colored vertical bars show the scale for the completeness. The YOLO-CL selection is almost flat as a function of redshift up to $z \sim 0.8$ when we consider the halo mass. The catalog is $\sim 100\%$ complete for $M_{200c} \gtrsim 10^{14.6} M_{\odot}$ at all redshifts."

galaxy groups - that are just below the official catalog thresholds. This means that the model's false positives aren't noise, but rather astrophysically plausible structures, making YOLO-CL a powerful exploratory tool in addition to a precision detector.

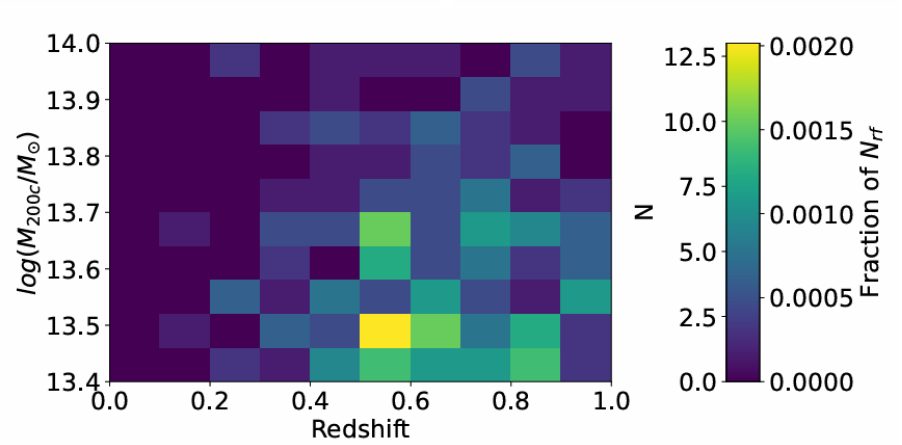

Fig. 4. From Grishin et al. (2025) : "Distribution of the ration of YOLO-CL DC2 false positive detections to the total number of random fields N_{rf} as a function of halo mass and redshift. The scale on the right indicates the number of false positive detections N, and the ratio of N to the total number of random fields N_{rf}. The total number of YOLO-CL random fields is 6,451."

3 Conclusions

The YOLO-CL Neural Network not only demonstrates a high completeness and purity when detecting galaxy clusters in the SDSS footprints with a performance of 98% in completeness and purity but also when comparing to older traditional methods such as RedMapper where YOLO-CL detected ∼98% of the X-ray detected clusters (MCXC2021) with $I_{X,500} \gtrsim 20 \times 10^{-15} erg/s/cm^2/arcmin^2$ at $0.2 \lesssim z \lesssim 0.6$ and ∼100% of the MCXC2021 clusters with $I_{X,500} \gtrsim 30 \times 10^{-15} erg/s/cm^2/arcmin^2$ at $z \gtrsim 0.3$.

In preparation for upcoming large scale surveys such as LSST and Euclid, the model has also been trained

on a hybrid SDSS and simulated LSST DESC DC2 dataset, enabling the model to keep a robust foundation with real SDSS images while adapting to "LSST-like". Obtaining an optimal 94% Completeness and Purity, an almost flat selection function of redshift up to $z \sim 0.8$ and is even 100% complete for $M_{200c} \gtrsim 10^{14.6} M_{\odot}$ between $0.2 < z < 0.8$. These are promising results, even though they were obtained in ideal conditions : "targeted mode", meaning that currently, YOLO-CL was applied to images that were cut around the simulated clusters. My Ph.D. thesis research consists in optimizing the model for real "survey-like" conditions(what happens when we do not know where the clusters are?) that will be adapted to LSST surveys and possibly combined LSST and Euclid data. Something that could be interesting to explore as well in the future is to improve the network to include more information on the cluster such as redshifts, members and mass. In short: (1) YOLO-CL is a fast, robust and accurate machine learning algorithm adapted for galaxy clusters detection; (2) it outperforms or matches traditional methods; (3) its false positives are still interesting massive objects to be analyzed; (4) we are working on further improvements and optimizations for large scale and survey-like condition to be made; (5) it reveals a promising future for YOLO applications in Astrophysics.

Thank you to the organizers of the SF2A 2025!

References

Eifler, T., Miyatake, H., Krause, E., et al. 2021, MNRAS, 507, 1746

Grishin, E., Contigiani, O., Aguena, M., et al. 2025, A&A, 695, A246

Grishin, K., Mei, S., & Ilić, S. 2023, A&A, 674, A76

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

LSST Dark Energy Science Collaboration (LSST DESC), Abolfathi, B., Alonso, D., et al. 2021, ApJS, 253, 31

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv e-prints, arXiv:0912.0201

Piffaretti, R., Arnaud, M., Pratt, G. W., Pointecouteau, E., & Melin, J.-B. 2011, A&A, 534, A109

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2015, arXiv e-prints

Redmon, J. & Farhadi, A. 2018, arXiv e-prints, arXiv:1804.02767

Rezatofighi, H., Tsoi, N., Gwak, J., et al. 2019, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 658–666

Rykoff, E. S., Rozo, E., Koester, B. P., & et al. 2014, ApJ, 785, 104