YOLO-CL : Deep Machine Learning for Galaxy Cluster detection with LSST (DESC projects [428] and [429])

Tran Vinh Phat, PhD student, APC/IN2P3/University Paris Cité
Simona Mei (APC, FR), Michel Aguena (Obs. Trieste, IT) ,
Stephane Ilic (IJCLab, FR)

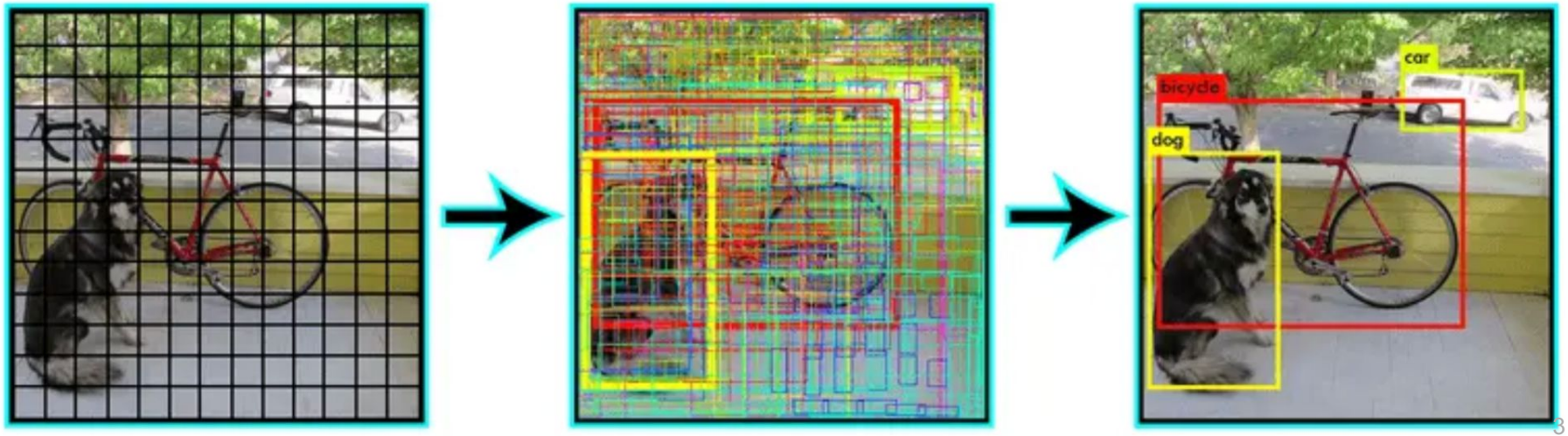
Introduction

- Galaxy Clusters Detections
  => Largest gravitationally bound system
  => Key cosmological probes (matter density, structure evolution, etc…)
  => Excellent tracers of LSS
  => Offer insights to the inner workings of galaxies

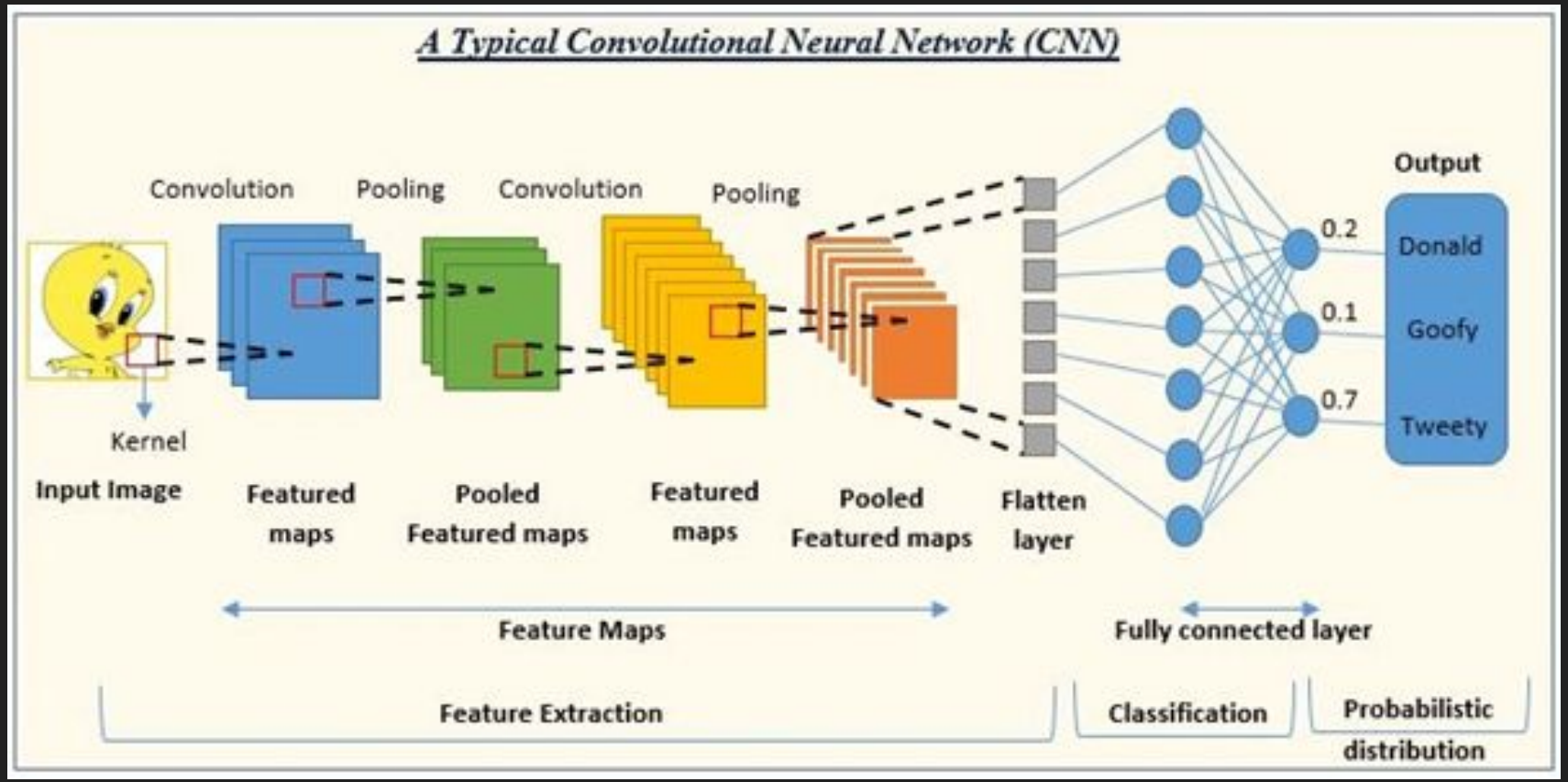
- Traditional Methods
  → Typically rely on galaxy catalogs (e.g. colors, photometric redshifts)
  → These approaches are subject to biases and systematic uncertainties

What is YOLO-CL ?

- YOLOv3 (Redmon 2015; Redmon & Fahradi 2018)
- Convolutional Neural-Network (CNN)
- Modified to detect galaxy clusters in color images -> YOLO-CL (Grishin, Mei, Ilic 2023)
- One-shot detection (single forward pass for bounding box and classification)



CNN

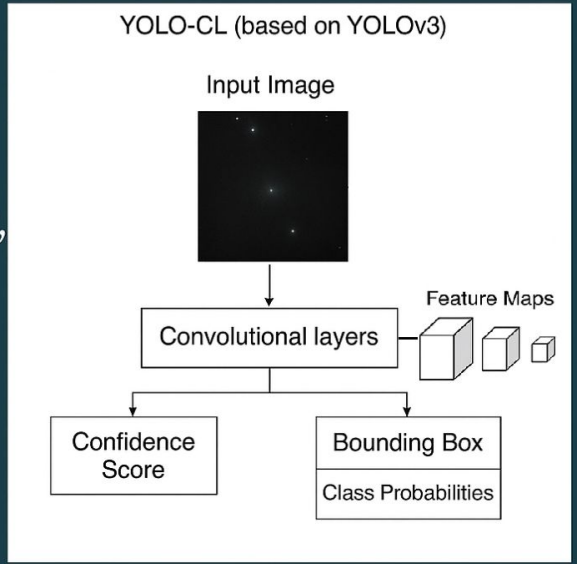

A Typical Convolutional Neural Network (CNN)
Convolution
Pooling
Convolution
Pooling
Output
0.2
0.1
0.7
Donald
Goofy
Tweety
Kernel
Input Image
Featured maps
Pooled Featured maps
Featured maps
Pooled Featured maps
Flatten layer
Feature Maps
Fully connected layer
Feature Extraction
Classification
Probabilistic distribution

YOLO-CL Architecture

Grishin, Mei, Ilic 2023

What is YOLO-CL?

- YOLO-CL = YOLOv3 + custom modifications for astrophysical image classification
- Architecture based on YOLOv3, a real-time object detector
- Key ML advantage: Operates directly on RGB composite images without needing photometry or redshifts



$$\mathcal{L} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{class}}$$

Training Dataset for SDSS

- SDSS: 24,406 redMaPPer clusters => training and validation each with combined half sample + blank fields for validation

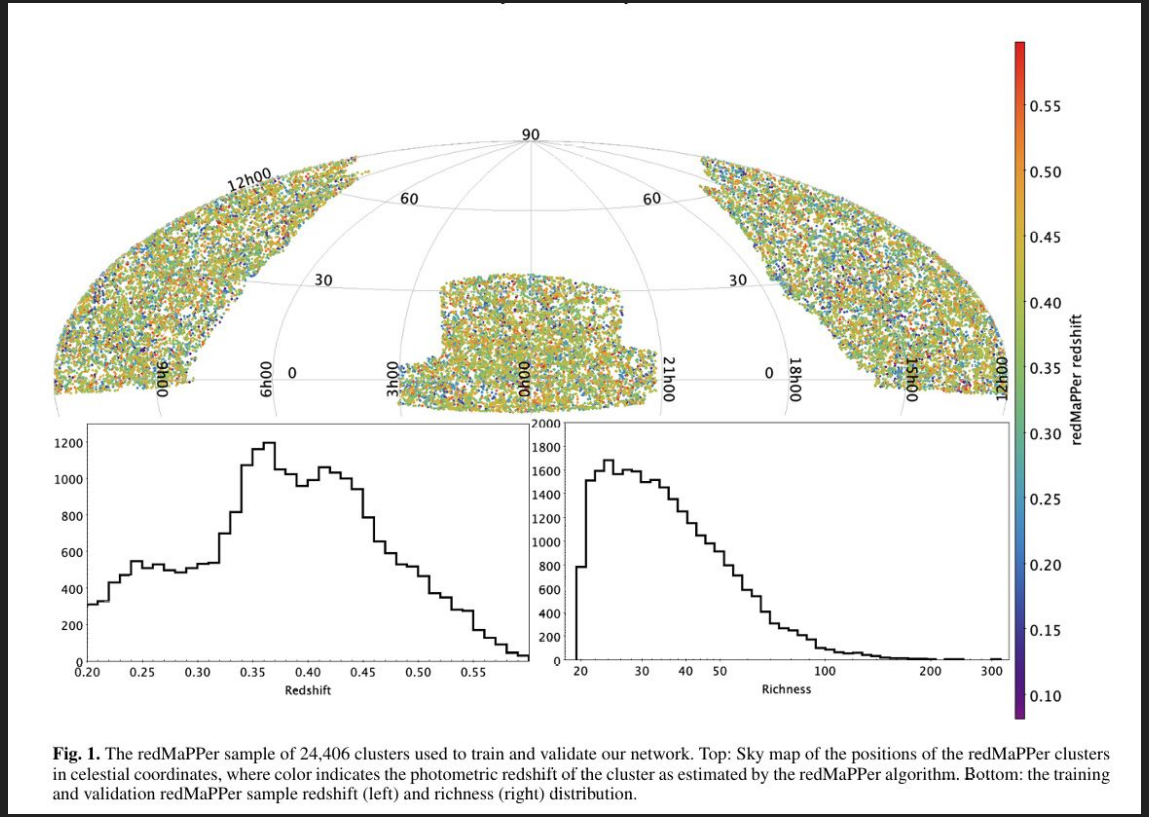
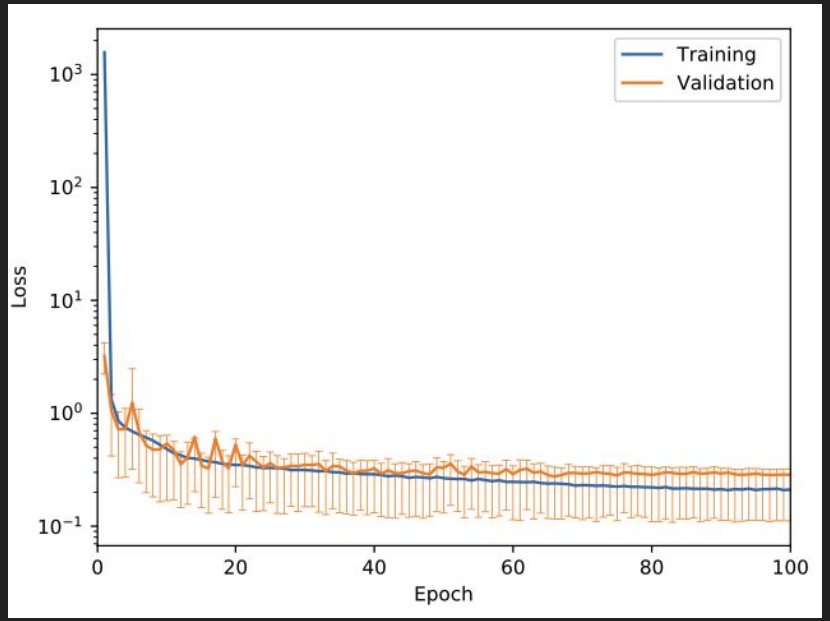


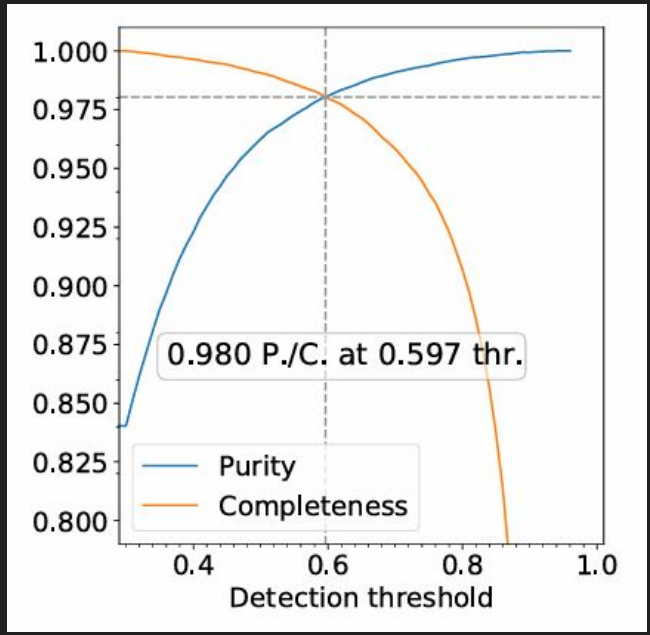
Fig. 1. The redMaPPer sample of 24,406 clusters used to train and validate our network. Top: Sky map of the positions of the redMaPPer clusters in celestial coordinates, where color indicates the photometric redshift of the cluster as estimated by the redMaPPer algorithm. Bottom: the training and validation redMaPPer sample redshift (left) and richness (right) distribution.

Grishin, Mei, Ilic 2023; Grishin et al. 2025

Performance (SDSS)

- YOLO-CL SDSS Completeness and Purity: 98%

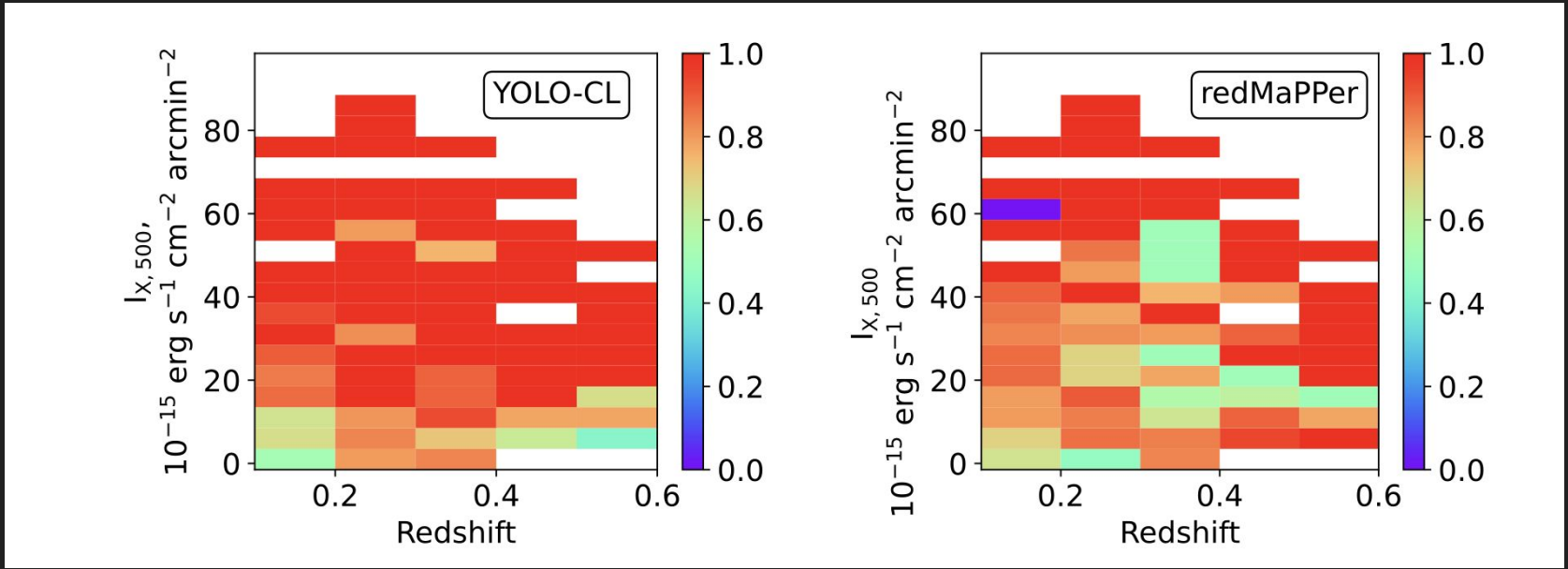




Grishin, Mei, Ilic 2023

Performance (SDSS)

- YOLO-CL outperforms RedMaPPer when comparing detections of the same MCXC2021 clusters, showcasing a flat selection function with X-ray surface brightness



Grishin, Mei, Ilic 2023

Training Dataset for application on LSST DC2 simulations

- **Transfer learning approach = SDSS + DC2 clusters**
- Hybrid training with SDSS and DC2 simulations:
  24,406 SDSS redMaPPer clusters + 2342 LSST DC2 simulated halos
  => training with combined half sample of each dataset
  => validation, with the other combined halves,
  alongside blank fields on DC2

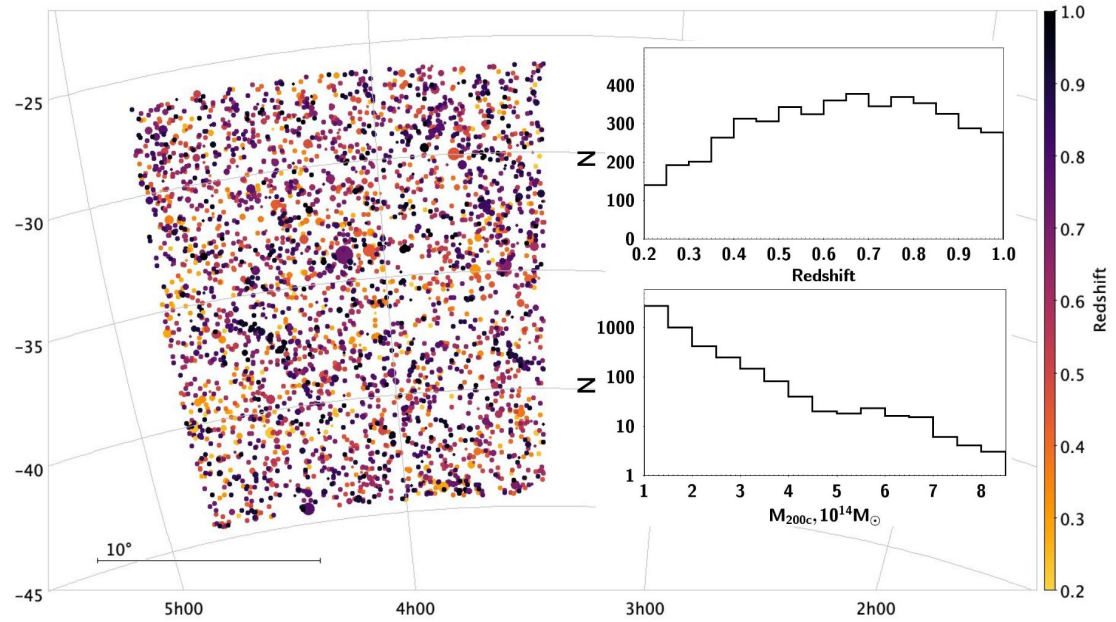
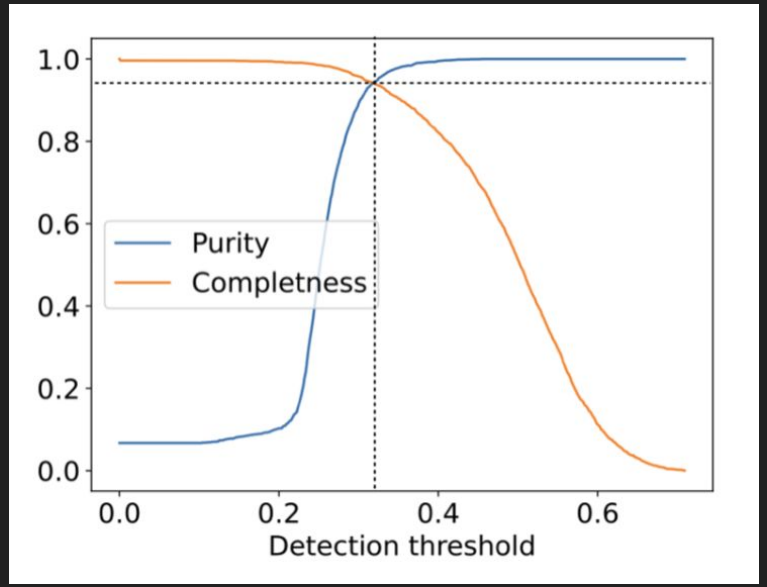


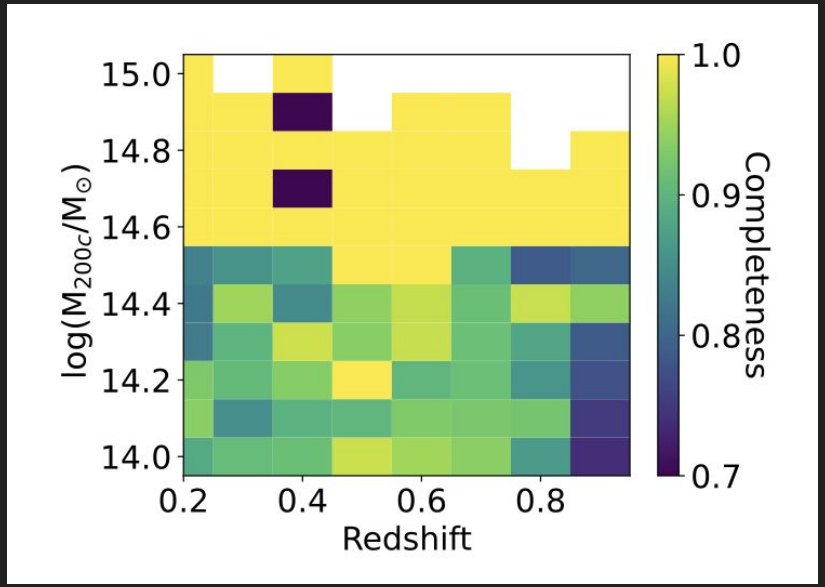
Fig. 1. Sky map with the positions of the 2,342 total CosmoDC2 clusters with $M_{200c} > 10^{14} M_{\odot}$ that we used for the YOLO-CL training and validation. Larger circle sizes indicate larger masses, and redshift is coded by color, as is indicated in the right bar). In the insert are the dark matter halo redshift and mass distributions.

Grishin, Mei, Ilic 2023, Grishin et al. 2025

Performance for LSST DC2

- Completeness and purity: 94% for halos with $M_{200c} > 10^{14}\ M_{\odot}$ at $z \lesssim 1$.
- 100% Completeness for $M > 10^{14.6}\ M_{\odot}$ at $0.2 < z < 0.8$

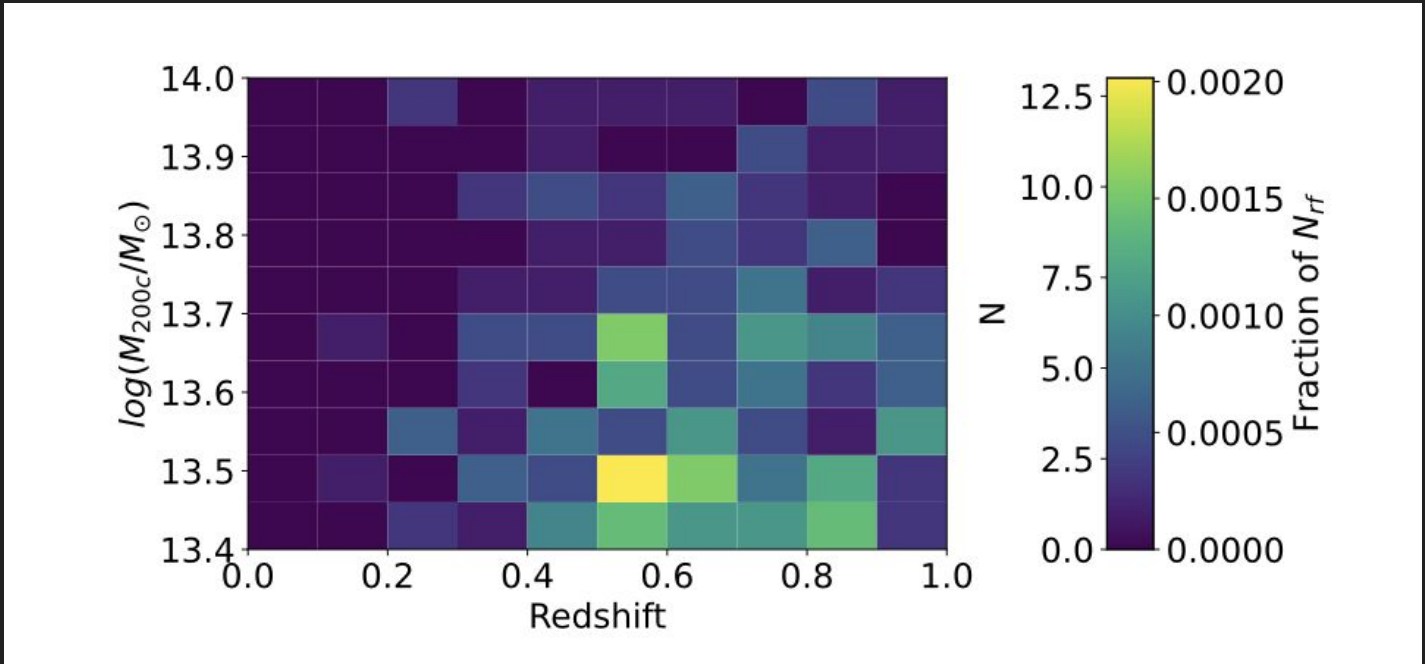




Grishin et al. 2025

FALSE POSITIVE DC2 (6% catalog contaminant)

- False positive on the entire sample (6%) are halos with masses $10^{13.4}\, M_\odot \lesssim M_{200c} \lesssim 10^{14}\, M_\odot$ corresponding to galaxy groups.



Grishin et al. 2025

Advantages of YOLO-CL

- No photometric redshift and galaxy catalog or stellar masking needed
- Robust to systematics and the presence of crowded star fields and artifacts
- Fast inference (single forward pass)
- Scalable to large-area surveys

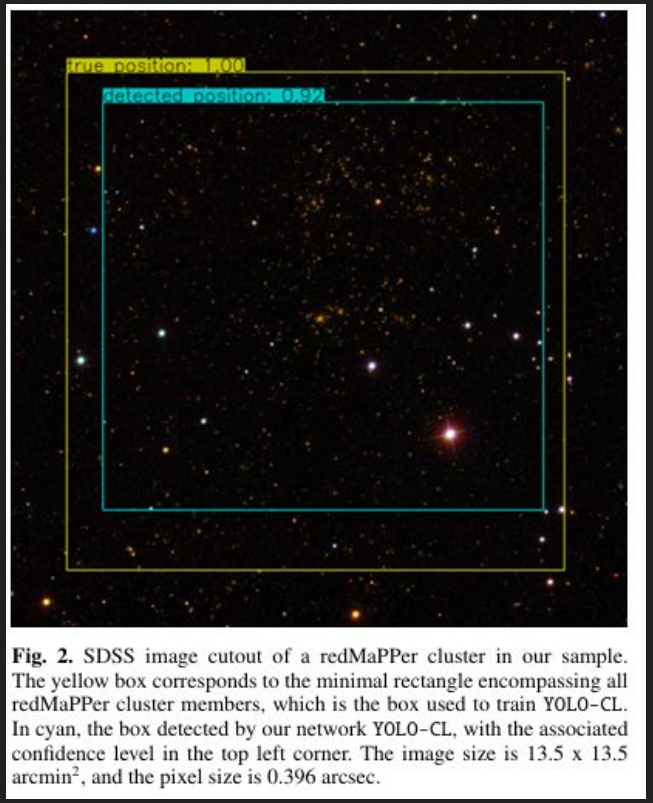


Fig. 2. SDSS image cutout of a redMaPPer cluster in our sample. The yellow box corresponds to the minimal rectangle encompassing all redMaPPer cluster members, which is the box used to train YOLO-CL. In cyan, the box detected by our network YOLO-CL, with the associated confidence level in the top left corner. The image size is 13.5 x 13.5 arcmin^2, and the pixel size is 0.396 arcsec.

Current status & Future steps and challenges

- Currently only tested in “*targeted mode*” meaning that YOLO-CL was applied to images that were cut around the simulated clusters: what happens when we do not know where the clusters are?

- My goal to optimize YOLO-CL for “Survey mode” in order to apply the model to LSST (DESC projet [428]) and combined LSST and EUCLID images (DESC projet [429]) => based on the latest version of the network developed by Michel Aguena.

- I will also improve the network to include more information on the cluster such as redshifts, members and mass.

Summary

- YOLO-CL is fast, robust, and accurate -> Image-based Machine Learning = promising future
- Outperforms or matches traditional algorithms
- False Positives in DC2 simulations are massive groups within the mass uncertainty in observations = interesting massive objects to be analyzed
- Optimization of “*survey mode*” for large-scale surveys LSST and LSST/EUCLID

Questions ?
Thank You !

Extra Slides

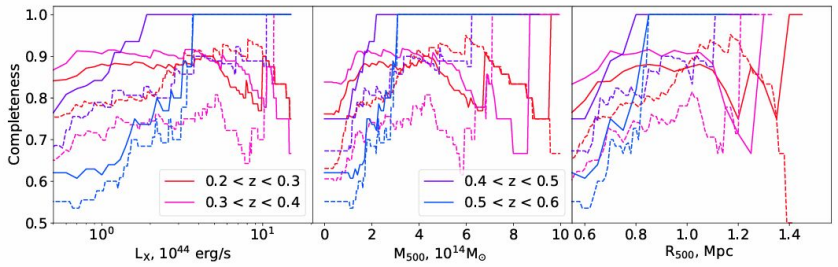


Fig. 9. The YOLO-CL (continuous lines) and redMaPPer (dashed lines) cluster detection completeness above a given X-ray luminosity, L_X (left panel), M_{500} (middle panel) and R_{500} (right panel). YOLO-CL recovers all clusters at $L_X \gtrsim 1-3\times10^{44}$ erg/s, $M_{500} \gtrsim 2-3\times10^{14}M_\odot$, $R_{500} \gtrsim 0.75-0.8$ Mpc and $z \gtrsim 0.4$. At lower luminosity, mass, radius and redshift, its performance is worse. The redMaPPer algorithm recovers all clusters at at $L_X \gtrsim 3-9\times10^{44}$ erg/s, $M_{500} \gtrsim 2-6\times10^{14}M_\odot$, $R_{500} \gtrsim 0.8-1.2$ Mpc and $z \gtrsim 0.4$. At high redshifts both YOLO-CL and redMaPPer demonstrate similar performance that is limited by the SDSS depth.

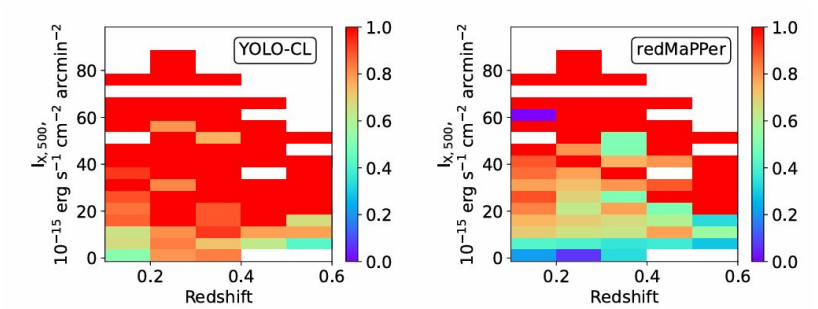


Fig. 10. The YOLO-CL and redMaPPer MCXC2021 cluster detection completeness as a function of redshift and mean X-ray surface brightness. Left: YOLO-CL detects $\sim$ 98% of the MCXC2021 clusters with $I_{X,500} \gtrsim 20\times10^{-15}$ erg/s/cm^2/arcmin^2 at $0.2 \lesssim z \lesssim 0.6$ and $\sim$ 100% of the MCXC2021 clusters with $I_{X,500} \gtrsim 30\times10^{-15}$ erg/s/cm^2/arcmin^2 and $z \gtrsim 0.3$. Right: redMaPPer detects $\sim$ 98% of the MCXC2021 clusters with $I_{X,500} \gtrsim 55\times10^{-15}$ erg/s/cm^2/arcmin^2 at $0.2 \lesssim z \lesssim 0.6$ and $\sim$ 100% of the MCXC2021 clusters with $I_{X,500} \gtrsim 20\times10^{-15}$ erg/s/cm^2/arcmin^2 at $0.5 \lesssim z \lesssim 0.6$. On the right of each figure is the completeness scale. From this comparison, YOLO-CL is more complete than redMaPPer in detecting MCXC2021 clusters.

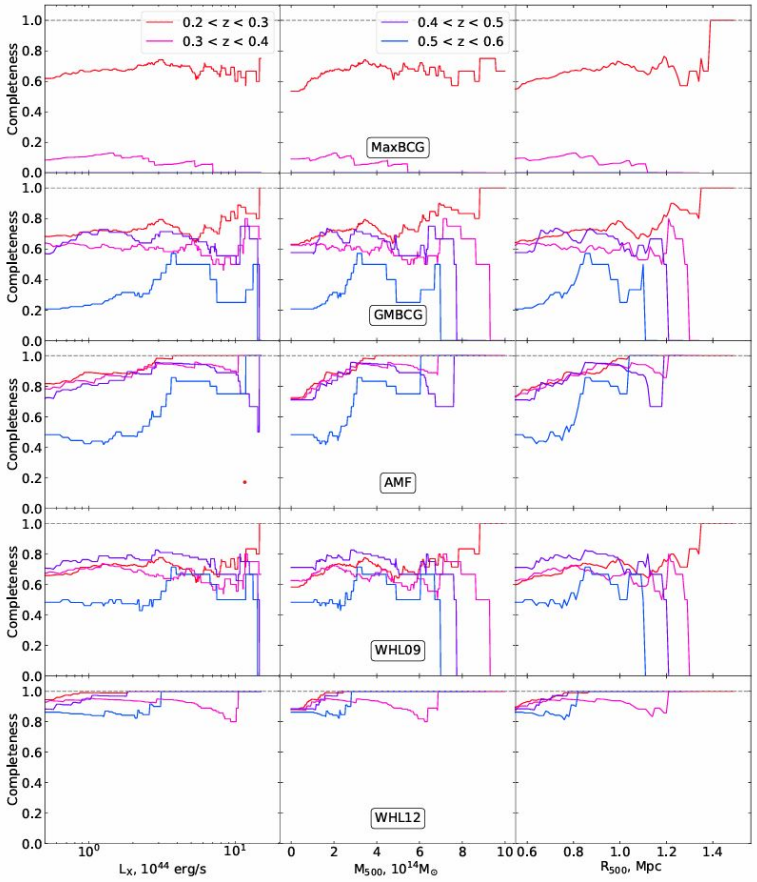


Fig. 11. The fraction of MCXC2021 clusters recovered by tradition cluster detection methods in the SDSS (see text) from top to bottom: MaxBCG, GMBCG, AMF, WHL09 and WHL12. The details of cluster recovery in each case are detailed in the text. In all cases, except WHL12, their completeness is worse than that reached by redMaPPer and YOLO-CL. These results, compared with Fig. 8 outline the high performance of our YOLO-CL with respect to traditional cluster detection methods in the SDSS.

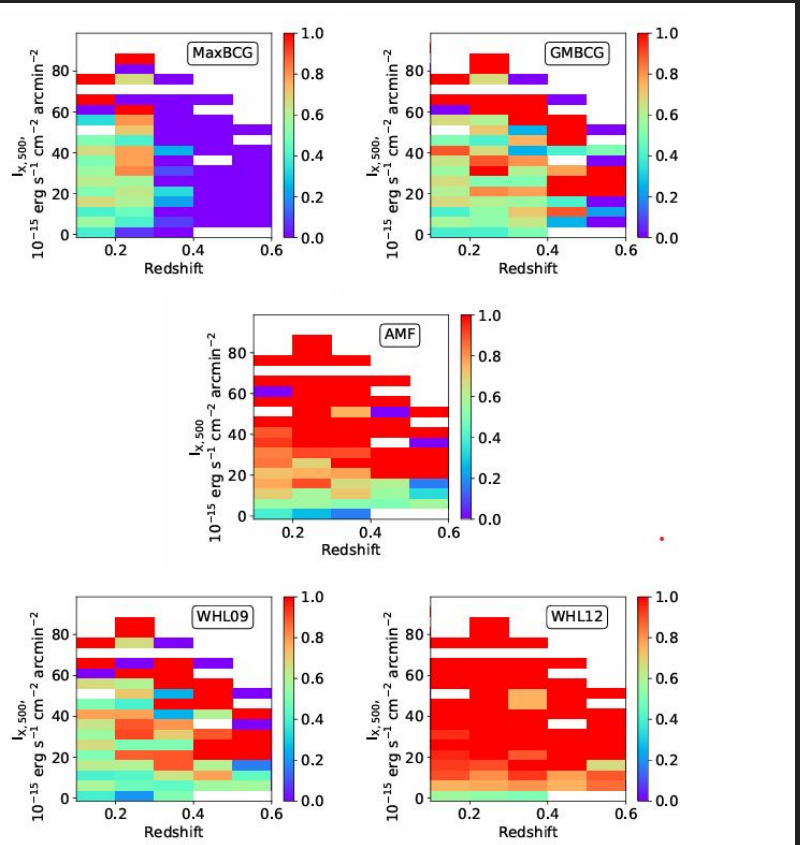
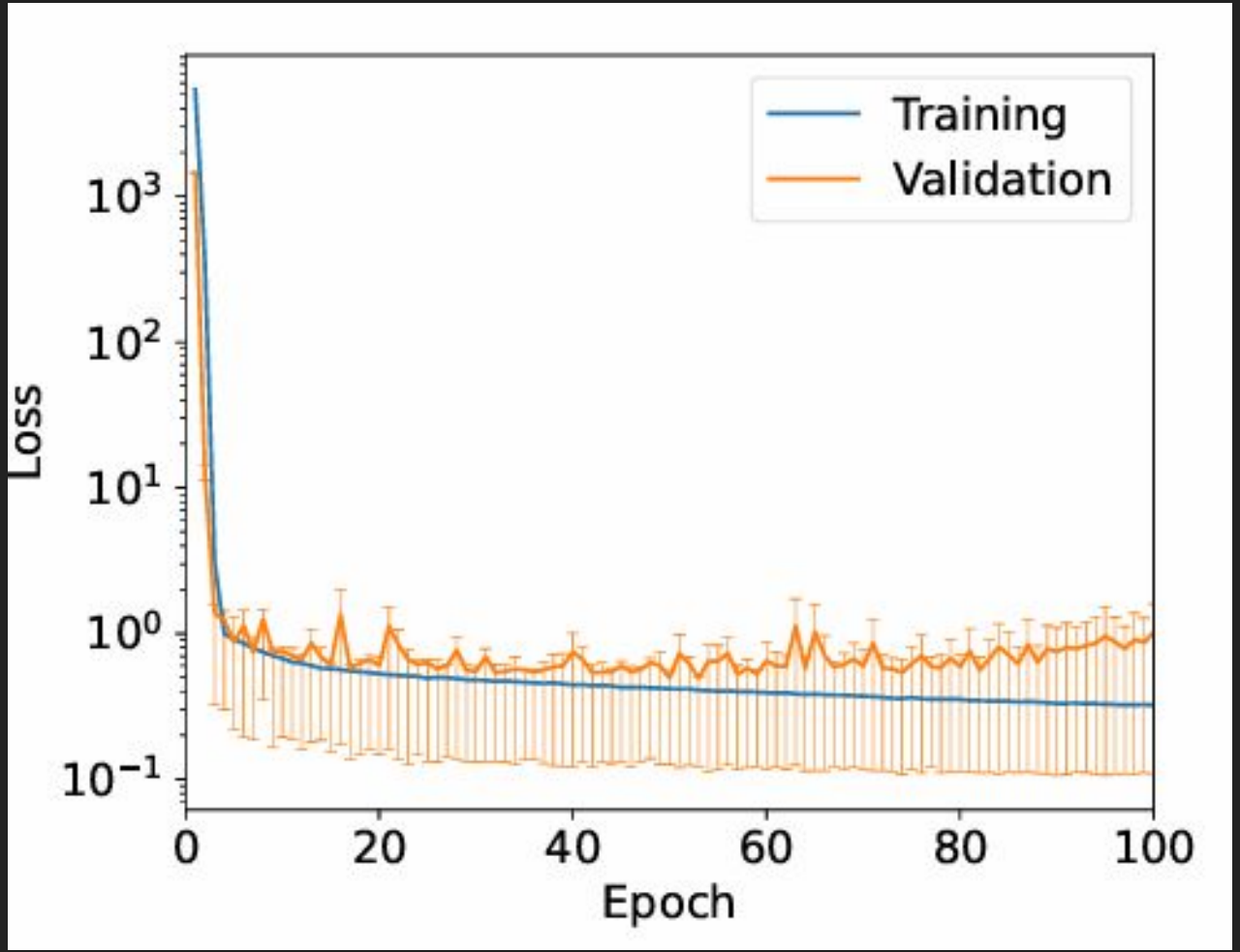


Fig. 12. The completeness of the MaxBCG, GMBCG and AMF galaxy cluster catalogs as a function of redshift and mean X-ray surface brightness, $I_{X,500}$. On the right of each figure is the completeness scale. All traditional cluster detection algorithms applied to SDSS are less complete than redMaPPer and YOLO-CL (see text), except AMF, which has a performance similar to redMapPPer, and WHL12, which has a performance similar to YOLO-CL.


Training
Validation
Loss
10^3
10^2
10^1
10^0
10^{-1}
0
20
40
60
80
100
Epoch

When training a YOLO network using a set of images (with their associated "true" bounding boxes), we optimise the following multi-part loss function $\mathcal{L}$ (Redmon et al. 2015):

$$\mathcal{L} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{class}} \,. \tag{2}$$

The first term of Eq. (2) is the "bounding box loss":

$$\begin{aligned}\mathcal{L}_{\text{bbox}} = \alpha_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\ + \alpha_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]\end{aligned} \tag{3}$$

where the (x, y) coordinates represent the center of the box relative to the bounds of the grid cell, and w and h are the width and height of the box. The symbol $\mathbb{1}_i^{\text{obj}}$ denotes if an object appears in cell i and $\mathbb{1}_{ij}^{\text{obj}}$ denotes that the jth bounding box predictor in cell i is "responsible" for that prediction. In these equations, and those below, the variables with a hat over them are the "true values" that the network is learning.

The second term is the "objectness loss":

$$\mathcal{L}_{\text{obj}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 + \alpha_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \tag{4}$$

where C represents the conditional class probability. Finally, the last term represents the "classification loss":

$$\mathcal{L}_{\text{class}} = \sum_{i=0}^{S^2} \mathbb{1}_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \tag{5}$$

where the $p_i(c)$ correspond to the probabilities to belong to a certain class i. The α_{coord} and the α_{noobj} coefficients appearing in the previous formulas can be changed to give more weight to certain components of the total loss. We choose to set $\alpha_{coord} = \alpha_{noobj} = 1$

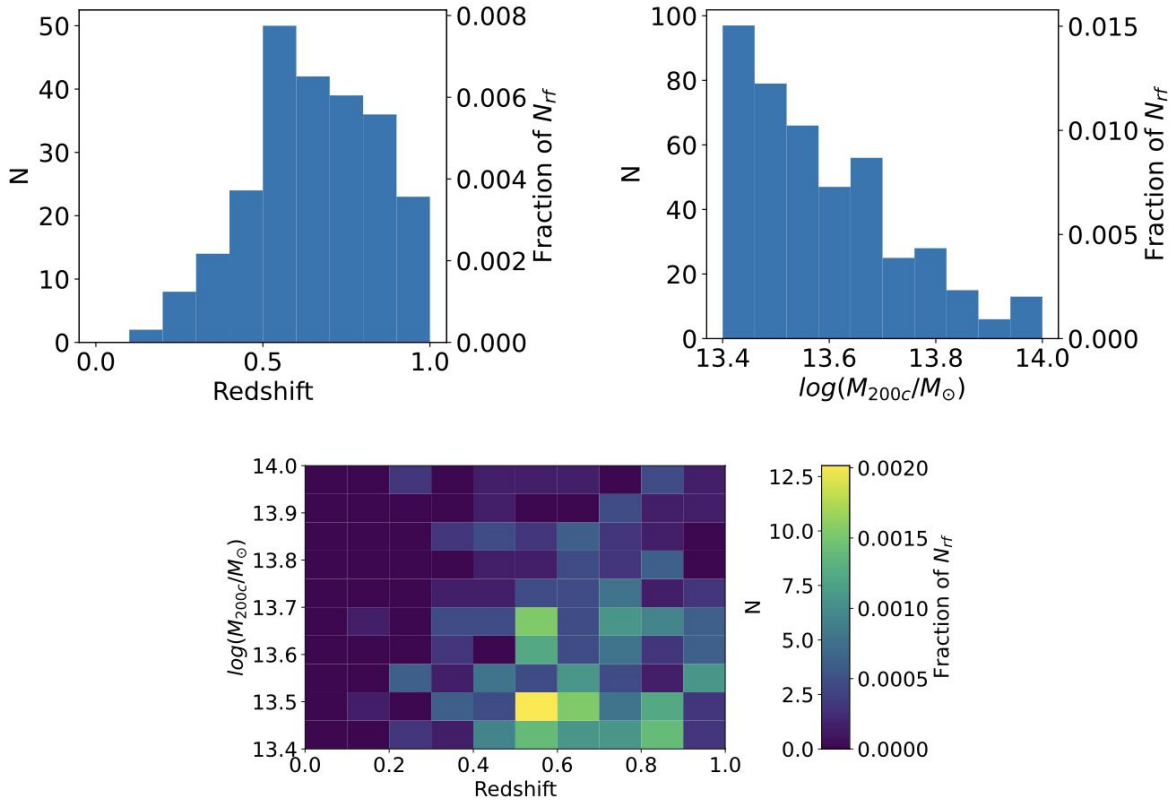
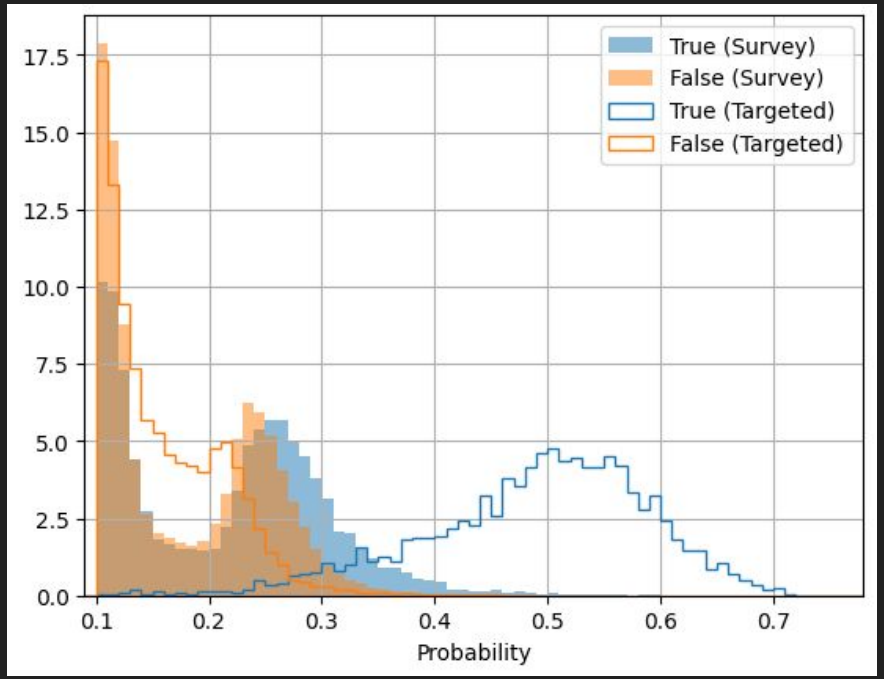


Fig. 6. Distribution of the ratio of YOLO-CL DC2 false positive detections to the total number of random fields, N_{rf}, as a function of halo mass and redshift (Top) and both (Bottom). In the bottom panel, the scale on the right indicates the number of false positive detections, N, and the ratio of N to the total number of random fields, N_{rf}. The total number of YOLO-CL random fields is 6451.

Survey detections

Probabilities for true detections are underestimated:



Boxes are ok: