Picking Particles in Cryo-EM Images without Knowing Particle Size

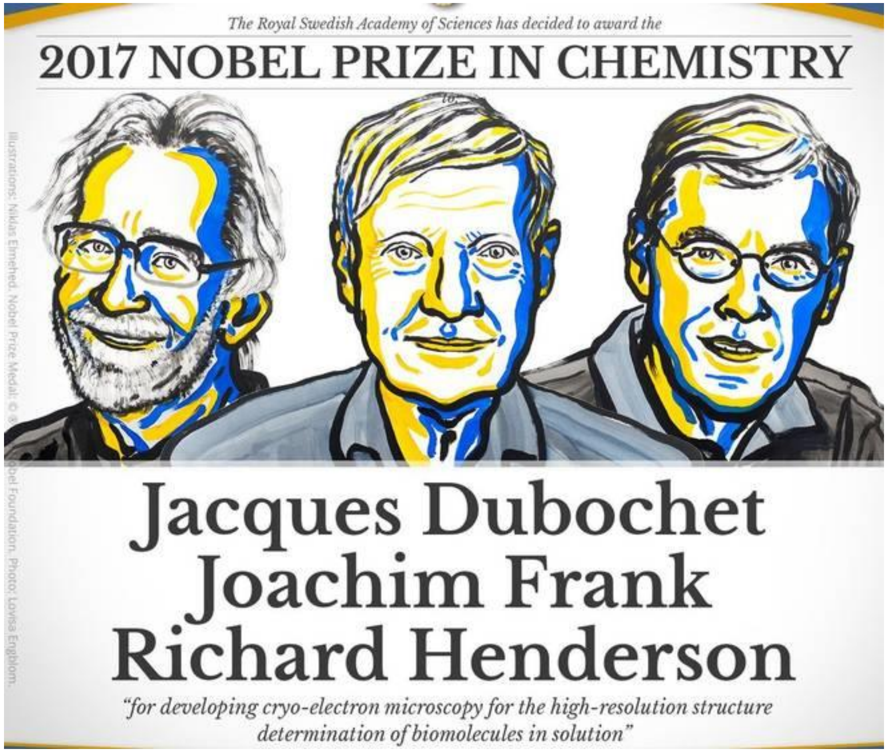
Yuewei Lin[1], Xiaoning Li[2], Qun Liu[1], Shinjae Yoo[1]

1. Brookhaven National Laboratory
2. Stony Brook University

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Cryogenic Electron Microscopy (cryo-EM)


The Royal Swedish Academy of Sciences has decided to award the
2017 NOBEL PRIZE IN CHEMISTRY
to
Jacques Dubochet
Joachim Frank
Richard Henderson
"for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"

Illustrations: Niklas Elmehed. Nobel Prize Medal: © ® Nobel Foundation. Photo: Lovisa Engblom.

In 2017, the Nobel Prize in Chemistry --

“for developing **cryo-electron microscopy** for the high-resolution structure determination of biomolecules in solution.”

How cryo-EM works

Each 2D image is a projection of its the 3D shape

– Each pixel value in the 2D image is the sum of the values along the line (along the direction of the electron beam) through the 3D sample

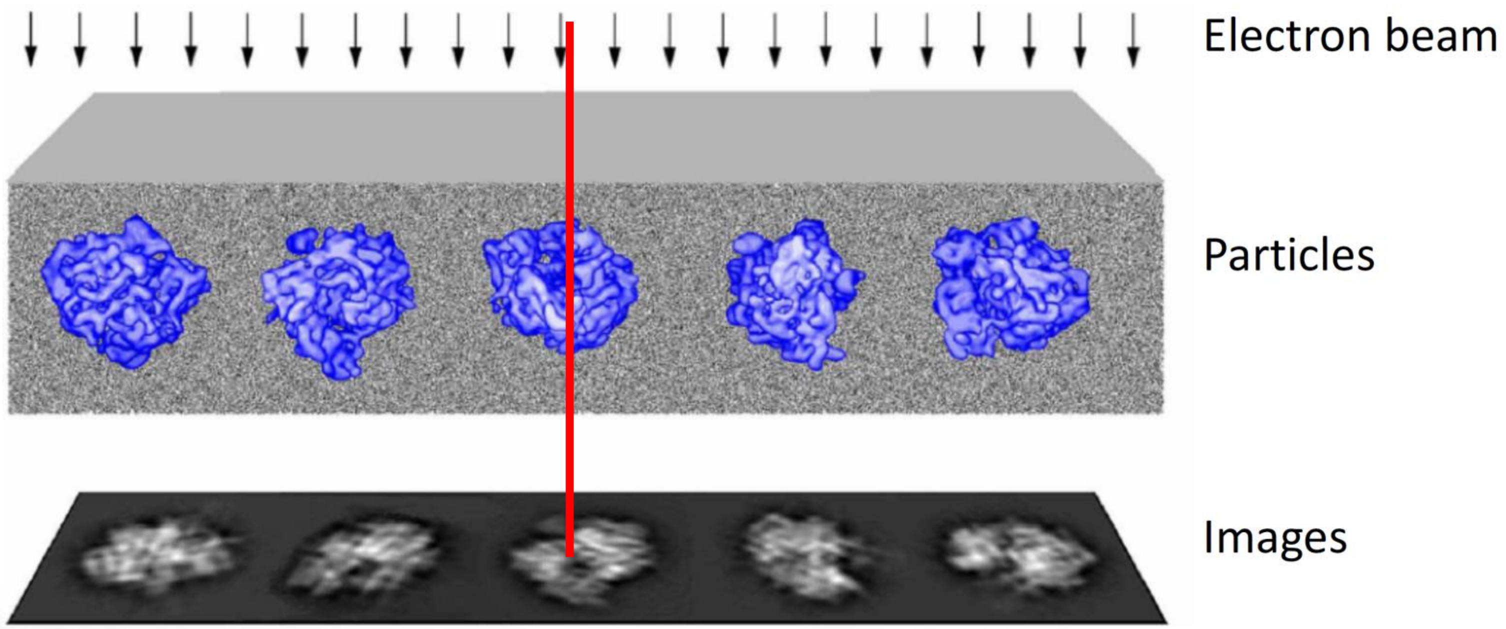


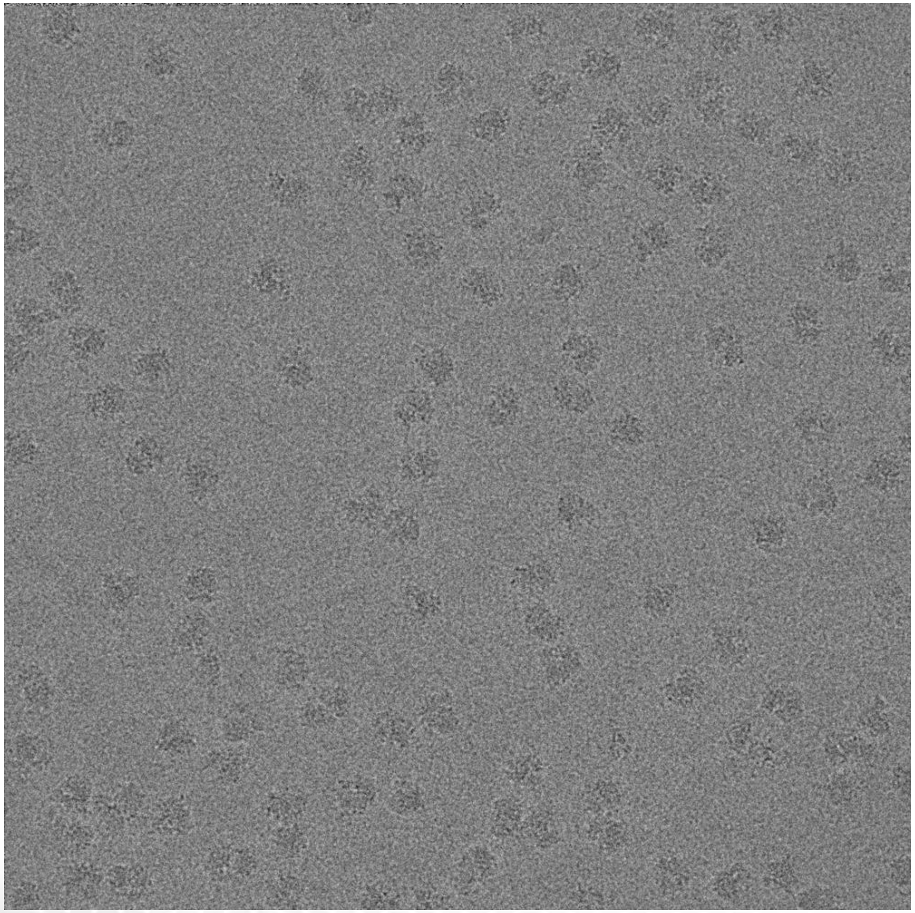
Image from http://biomachina.org/courses/structures/091.pdf

Computational steps in cryo-EM

Aims to reconstruct the structure of a "particle": single molecule (e.g., protein) or composed of many molecules (e.g., a ribosome)

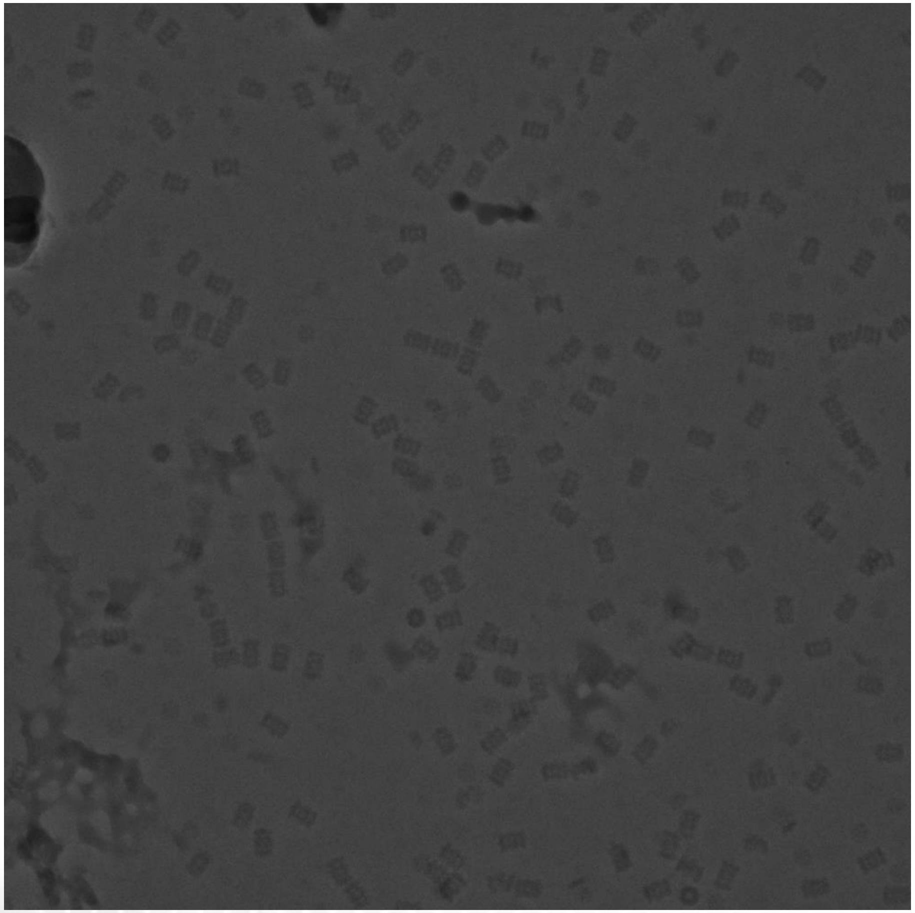
- Spreading identical particles, and image them using an EM. Particles are positioned with different orientations
- Picking as many as particles in micrographs
- 3D structure reconstruction using 2D particles

Challenges



- The micrographs usually have very low signal-to-noise ratio (SNR)

Challenges

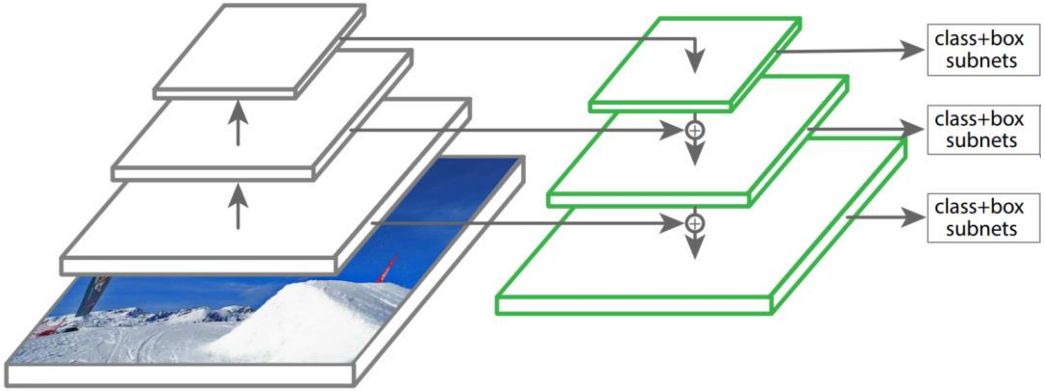


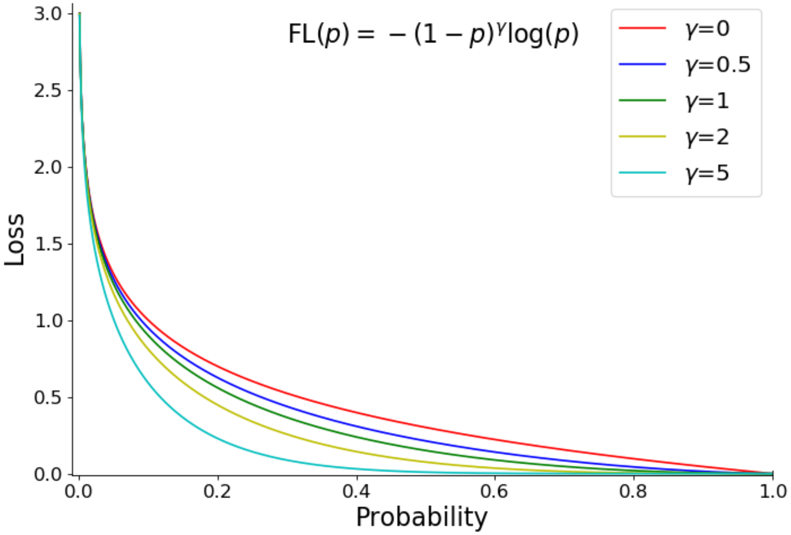
- The micrographs usually have very low signal-to-noise ratio (SNR)
- Other interferences, such as ice contamination, background noise, amorphous carbon and particle overlap.
- High-resolution reconstruction requires extensive particles identification (> 100,000).

Limitation of the existing methods

- Traditional template based models
  - Very slow (minutes per micrograph)
  - Need to know particle size
  - Sensitive to noise
- Deep neural network (DNN) based models
  - Slow (seconds to tens of seconds per micrograph)
  - Need to know particle size

Advanced DNN based object detection model -- RetinaNet





Structure:

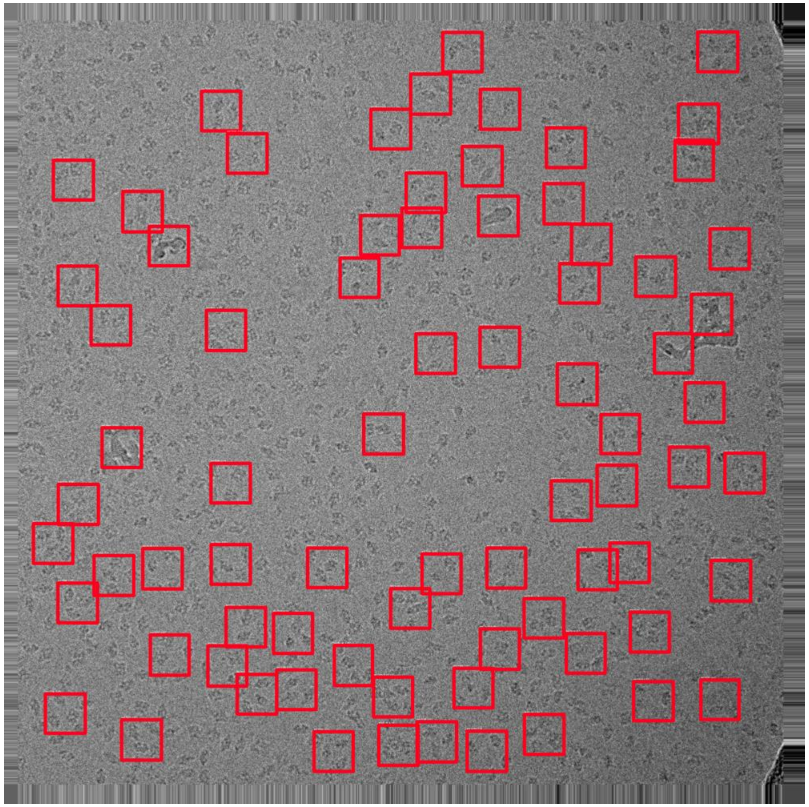
- Resnet for encoder
- Feature pyramid network for multi-scale feature extraction
- Subnets for class and box regression in each scale

Focal loss:

- Penalizes less to easy samples
- Deals with class imbalance

Fail to apply across domains

- All the ground truth particles are the same size
- Model tends to pick particles with the same size as ground truth
- It impedes the trained model to apply on different datasets

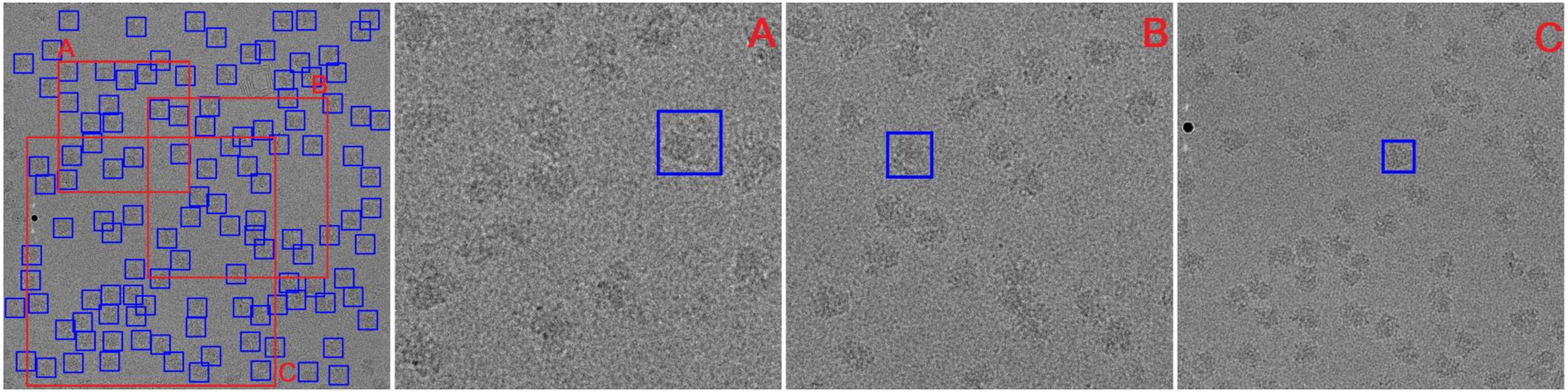


Trained on particle size 200
Tested on particle size 120

Data augmentation with diverse size of particles

Utilizing different sizes of particles as training data

- Cropping patches with random positions and sizes
- Re-sizing them to be the same size (e.g., 1000×1000).



4096×4096 1388×1388 1905×1905 2643×2643

The key to make the model picking particles without knowing the size!

Prior knowledge – size consistency

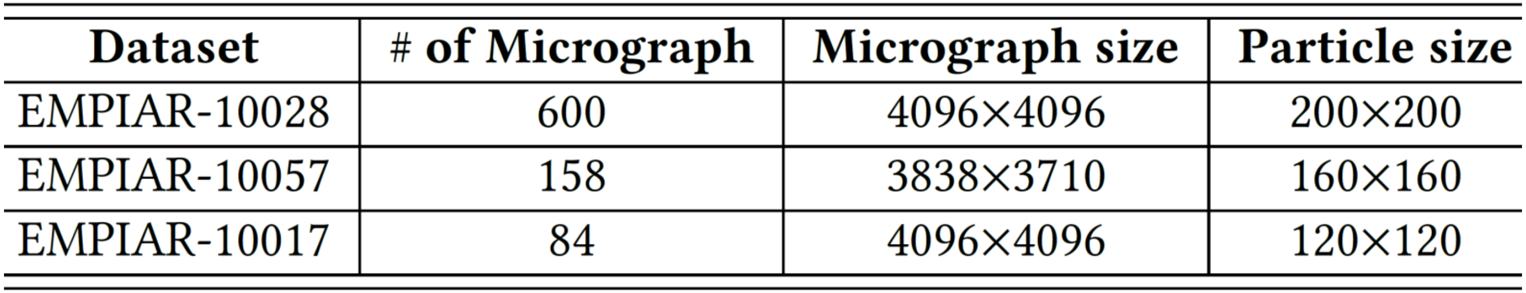
Single particle picking assumption: all the particles should have the same or very similar sizes.

Adding size consistency in the total loss to penalty the particles with sizes significantly different from majority of particles. :

$$\mathcal{L}_{sc} = \frac{1}{N}\sum_{i}(s_i - \sum_{i}\frac{1}{N}s_i)^2$$

s_i denotes the size of the *i*th predicted particle, *N* denotes the total number of predicted particles.

Data & experiment setting



Dataset	# of Micrograph	Micrograph size	Particle size
EMPIAR-10028	600	4096×4096	200×200
EMPIAR-10057	158	3838×3710	160×160
EMPIAR-10017	84	4096×4096	120×120

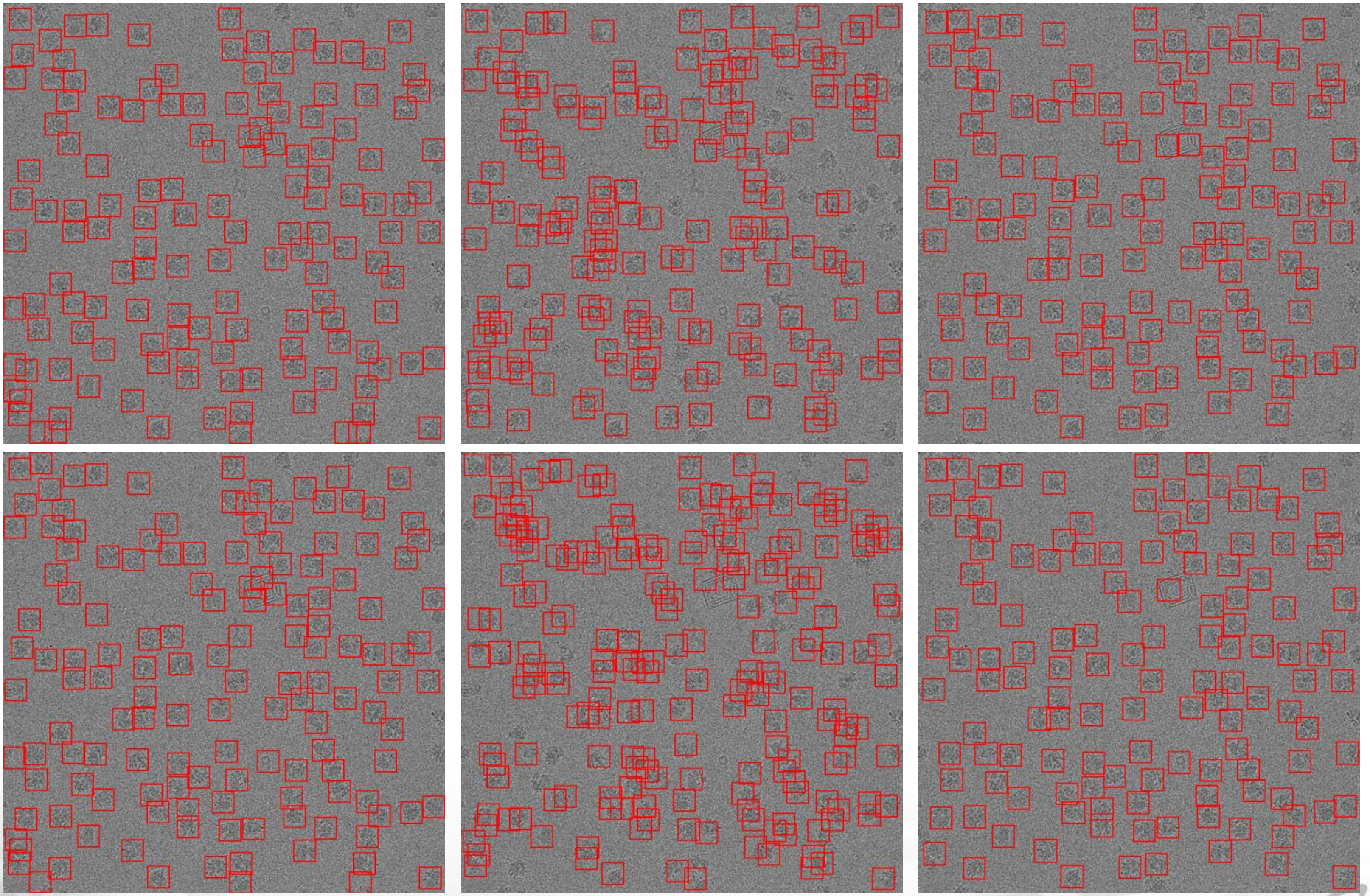
Public datasets: https://www.ebi.ac.uk/pdbe/emdb/empiar/

- **Different training amounts:**
  - 20 micrographs for training
  - 50 micrographs for training
- **Different training/testing settings:**
  - Single domain (dataset): train and test on the same domain
  - Cross domain (dataset): train on domain(s), test on different domain(s)

Metrics

- *True positive* is the correct predictions
- ***Precision*** = # of true positive / # of predictions
- ***Recall*** = # of true positive / # of ground truth
- ***F-measure*** = 2 × Prec. × Rec. / (Prec. + Rec.)
- ***Average precision (AP):*** *a*djusting prediction confidence score threshold will lead multiple precision/recall pairs, ***AP*** is basically the area under precision-recall curve.

Qualitative results in 10028 -- single domain

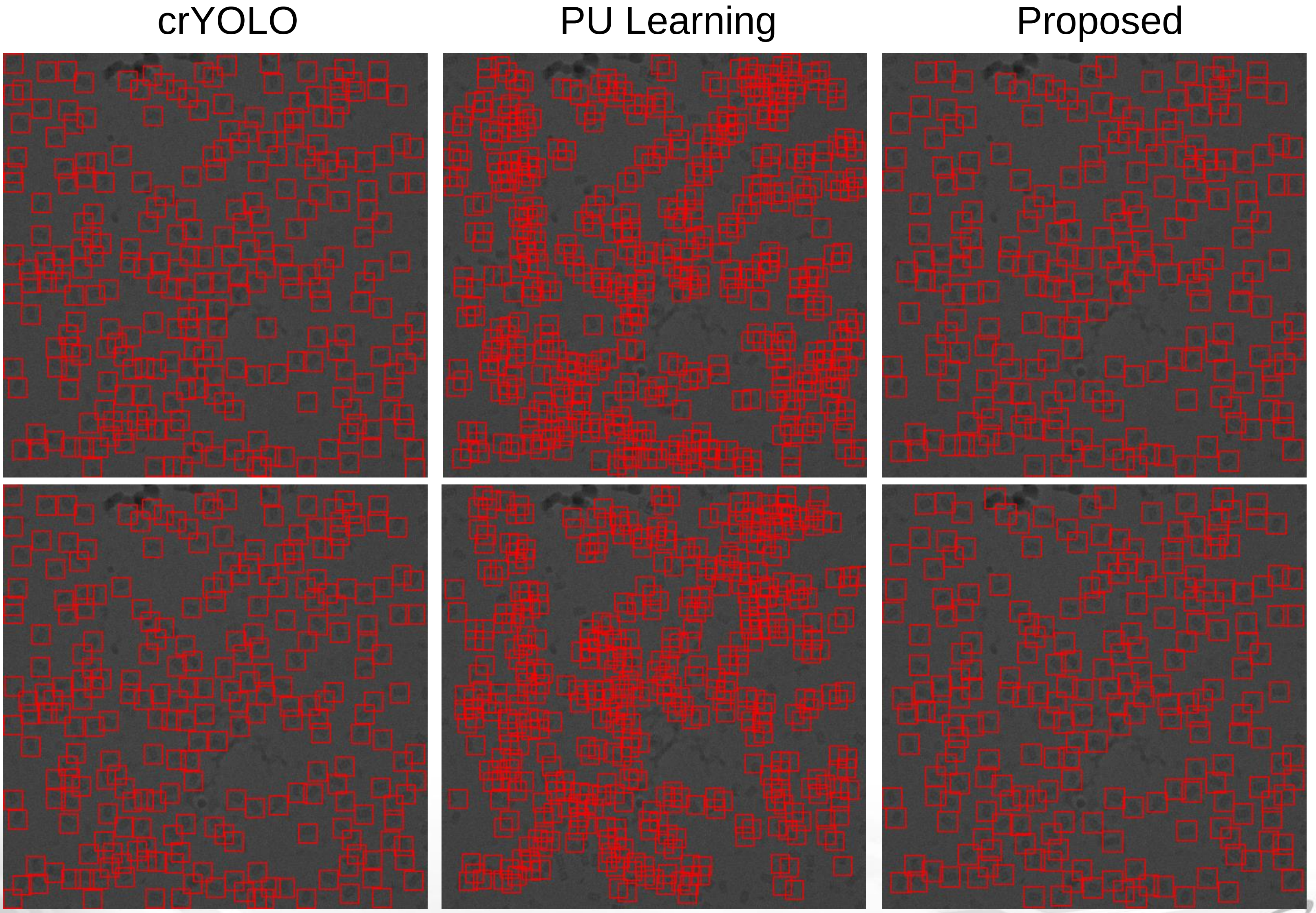


crYOLO

PU Learning

Proposed

Qualitative results in 10057 -- single domain

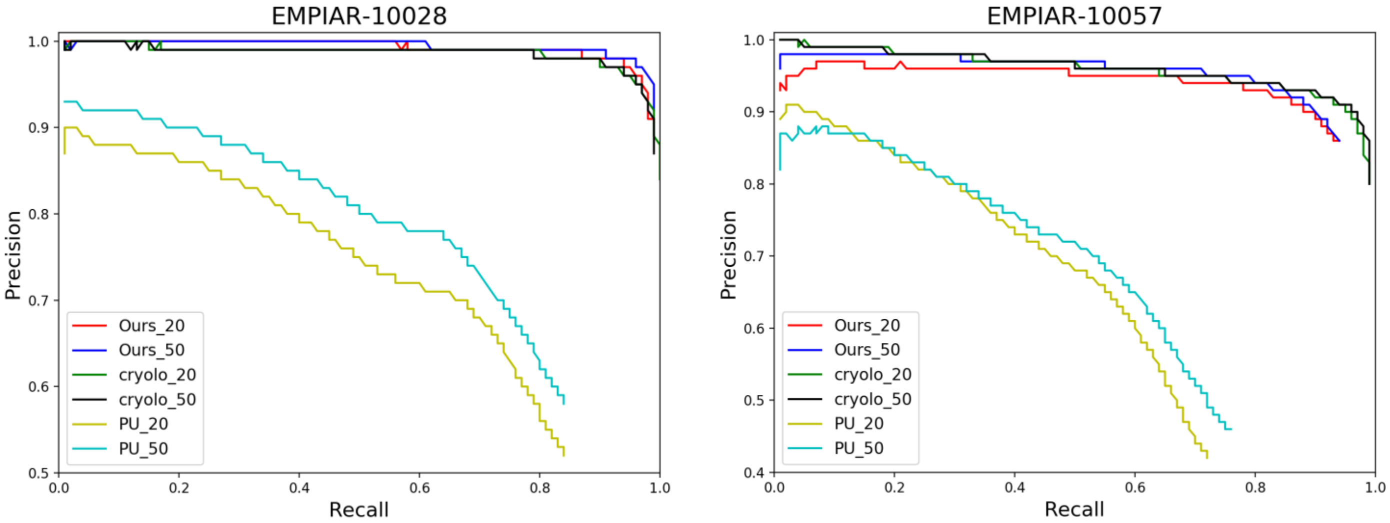


crYOLO

PU Learning

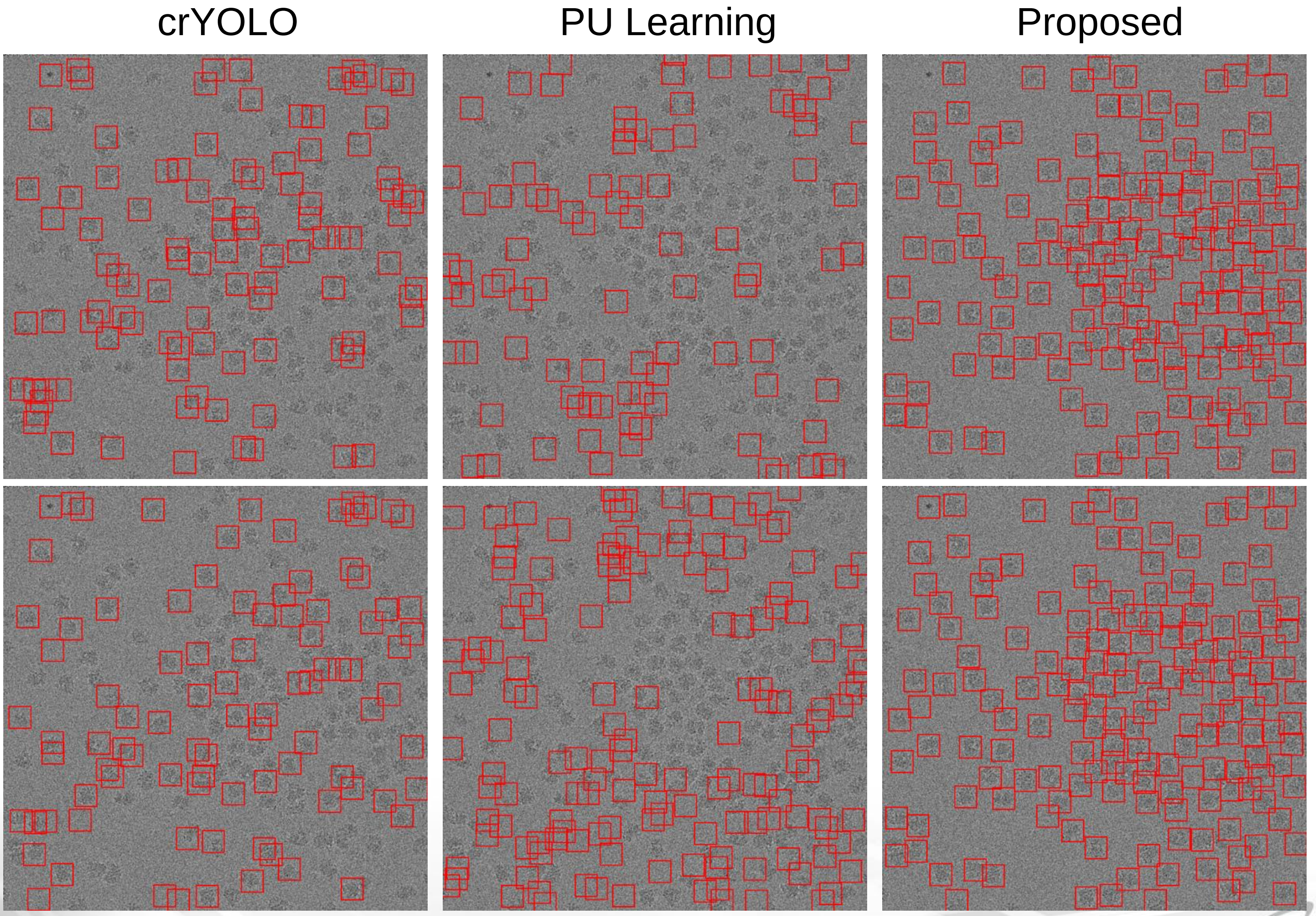
Proposed

Quantitative results -- single domain



Dataset	Training amount	Model	AP	Precision	Recall	F-measure
EMPIAR-10028	w/50 micrographs	PU Learning	0.6904	0.5821	0.8411	0.6880
		crYOLO	0.9825	0.8659	**0.9952**	0.9260
		Ours	**0.9859**	**0.9206**	0.9925	**0.9552**
	w/20 micrographs	PU Learning	0.6505	0.5181	0.8434	0.6419
		crYOLO	**0.9843**	0.8352	**0.9973**	0.9091
		Ours	0.9783	**0.9043**	0.9863	**0.9436**
EMPIAR-10057	w/50 micrographs	PU Learning	0.5635	0.4587	0.7566	0.5711
		crYOLO	**0.9554**	0.8005	**0.9941**	0.8868
		Ours	0.9320	**0.8619**	0.9379	**0.8983**
	w/20 micrographs	PU Learning	0.5334	0.4203	0.7209	0.5310
		crYOLO	**0.9520**	0.7978	**0.9914**	0.8841
		Ours	0.9225	**0.8556**	0.9403	**0.8960**

Qualitative results in 10028 -- cross domain

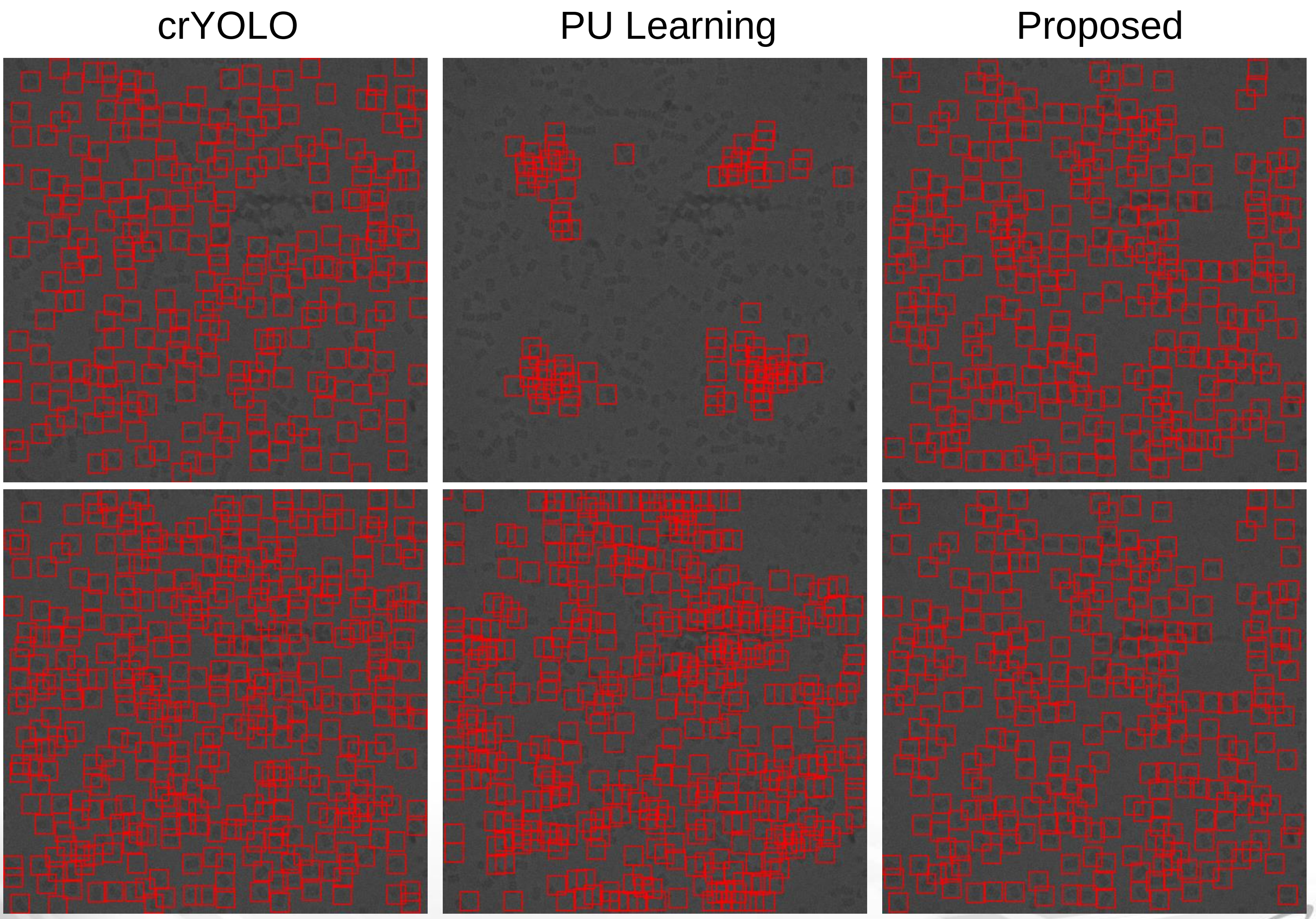


crYOLO

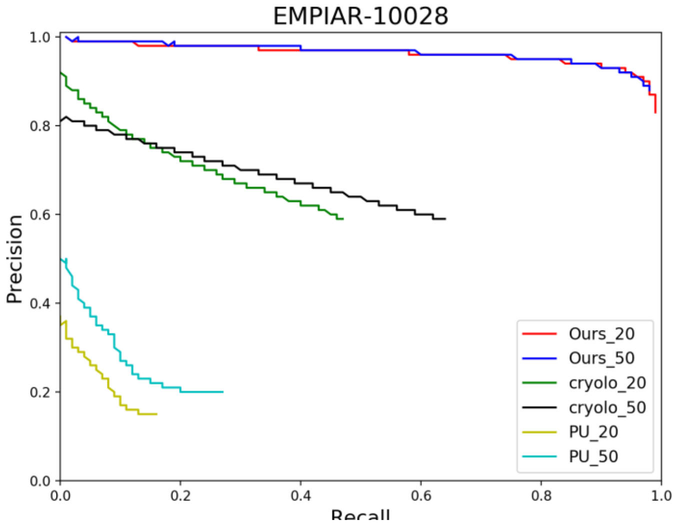
PU Learning

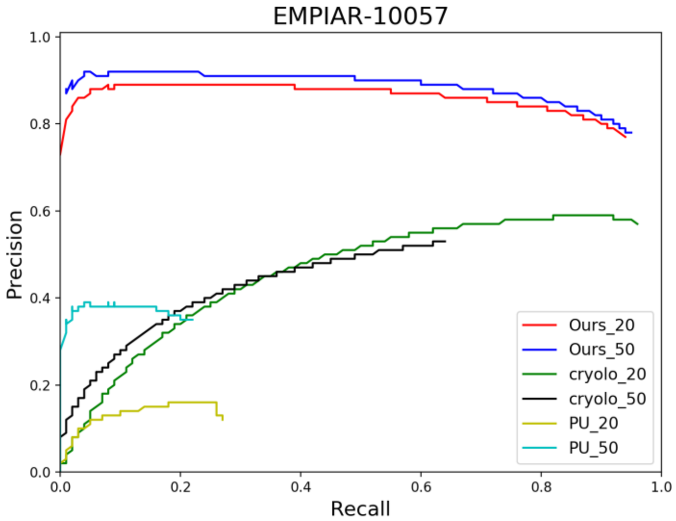
Proposed

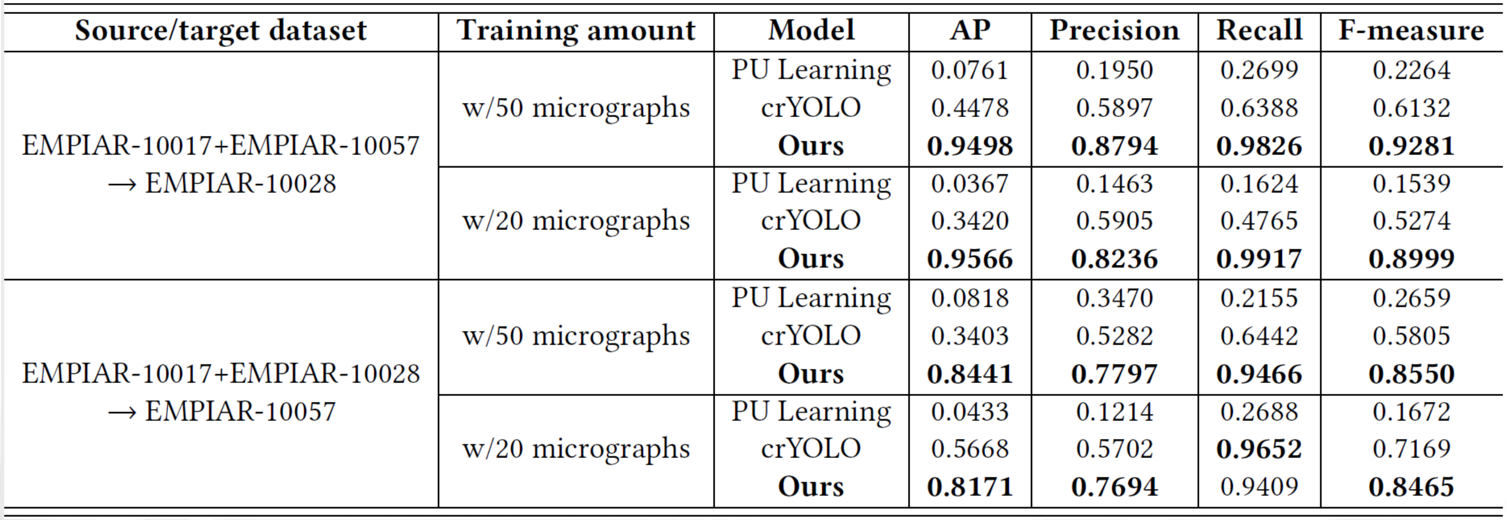
Qualitative results in 10057 -- cross domain



Quantitative results -- cross domain



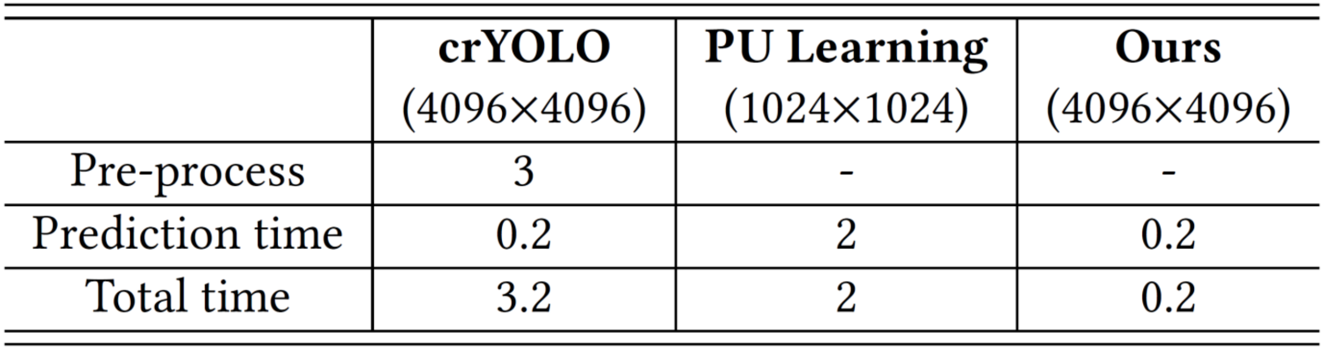




Source/target dataset	Training amount	Model	AP	Precision	Recall	F-measure
EMPIAR-10017+EMPIAR-10057 → EMPIAR-10028	w/50 micrographs	PU Learning	0.0761	0.1950	0.2699	0.2264
		crYOLO	0.4478	0.5897	0.6388	0.6132
		Ours	**0.9498**	**0.8794**	**0.9826**	**0.9281**
	w/20 micrographs	PU Learning	0.0367	0.1463	0.1624	0.1539
		crYOLO	0.3420	0.5905	0.4765	0.5274
		Ours	**0.9566**	**0.8236**	**0.9917**	**0.8999**
EMPIAR-10017+EMPIAR-10028 → EMPIAR-10057	w/50 micrographs	PU Learning	0.0818	0.3470	0.2155	0.2659
		crYOLO	0.3403	0.5282	0.6442	0.5805
		Ours	**0.8441**	**0.7797**	**0.9466**	**0.8550**
	w/20 micrographs	PU Learning	0.0433	0.1214	0.2688	0.1672
		crYOLO	0.5668	0.5702	**0.9652**	0.7169
		Ours	**0.8171**	**0.7694**	0.9409	**0.8465**

Efficiency

- CPU: Intel(R) Core(TM) i7-7800X 3.5GHz
- GPU: Nvidia GeForce GTX 1080 Ti



	crYOLO (4096×4096)	**PU Learning** (1024×1024)	**Ours** (4096×4096)
Pre-process	3	-	-
Prediction time	0.2	2	0.2
Total time	3.2	2	0.2

Conclusion & Discussion

The proposed method

- Picking particles without knowing the particle size
- Reasonably good on picking particles across domains
- Fast (0.2s per micrograph)

Future improvements

- Advanced domain adaptation techniques
- Train on more diverse of data(sets)

Future applications

- Object of interests detection in any electron microscopy images, such as transmission EM (TEM), scanning EM (SEM), etc.