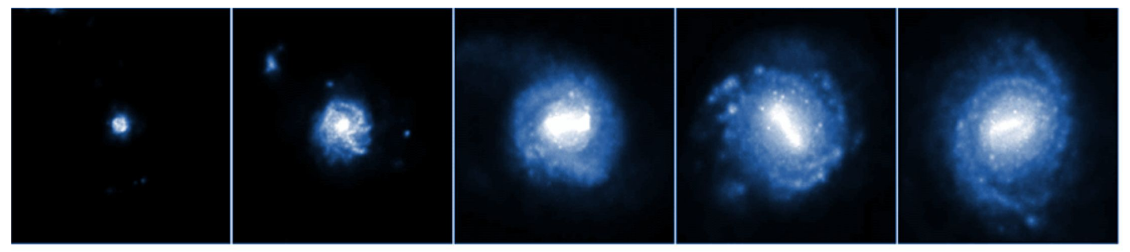
Forming Galaxies using Machine Learning

Xinyue Zhang , Yanfang Wang, Yueqiu Sun, Wei Zhang
Center for Data Science, New York University

Introduction

In Cosmology, Computer simulations is vital in understanding how the primordial universe evolved into the universe we observe today.

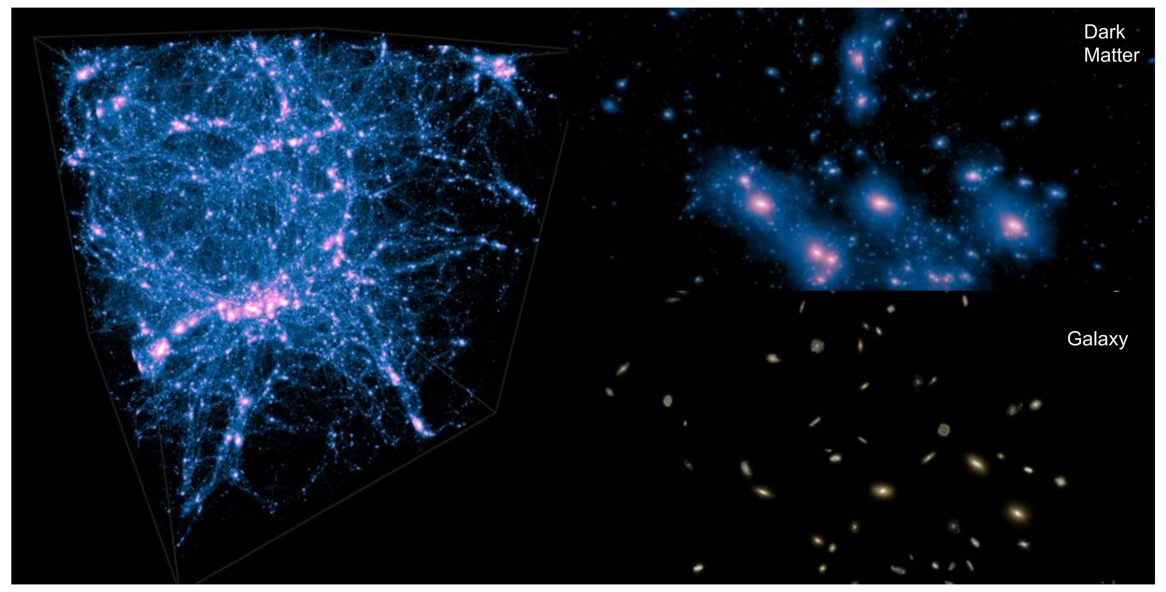


Example galaxy evolve in time from left to right, from when the universe was a quarter its current age, to the present.

However, simulations of the baryonic component require vast computation power and time. To combat this challenge, we propose to map the simulations of the dark matter to the simulations of galaxies through convolutional neural networks instead of physical laws.

Data Preparation

Data for this project came from Illustris simulation, the most accurate cosmology simulation that is available public. Redshift=0 simulations are deployed. The dark matter simulation and full hydrodynamic simulation are on the same time step and correspond to each other.



Visualization of Illustris dark matter simulation at reshift=0 (left)
Zoom-in visualization of corresponding dark matters and galaxies. (Right)

Whole simulation was divided into 1024×1024×1024 grid. Then the number of galaxies in each grid cell was counted to create the density image p(x).

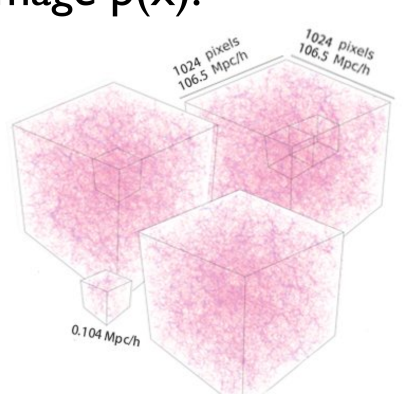


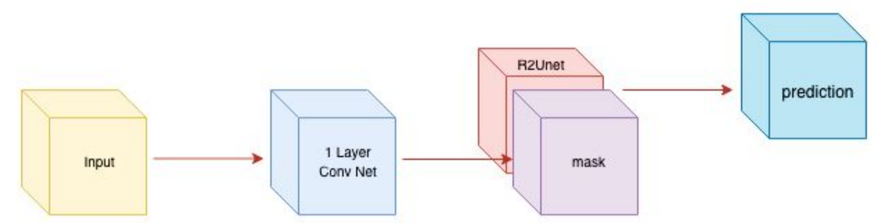
Illustration of grid representation of whole simulations

In order to correct for a spatially varying galaxy selection function, we normalize the observed galaxy density p(x) to a dimensionless over-density.

$$\delta(x) = \frac{\rho(x) - \bar{\rho}(x)}{\bar{\rho}(x)}$$

Model Structure

Based on the sparsity and imbalance of our dataset, we deployed two phase training. This allow us to refine our result using Unet after a high recall convolution classification Layer.



Two Phase network Illustration

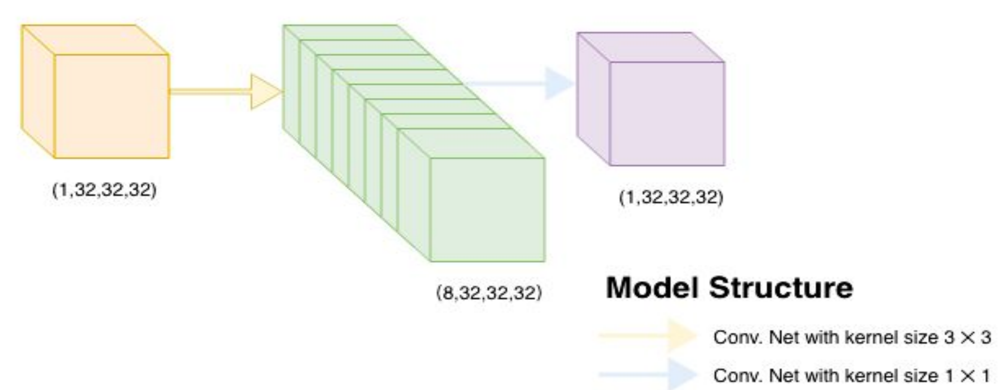


Illustration of step one of our two phase network:
a simple one-layer-convolution was used to capture the monotonically relation between the density of dark matters and galaxies, and generating a rough distribution of where galaxies are possibly located

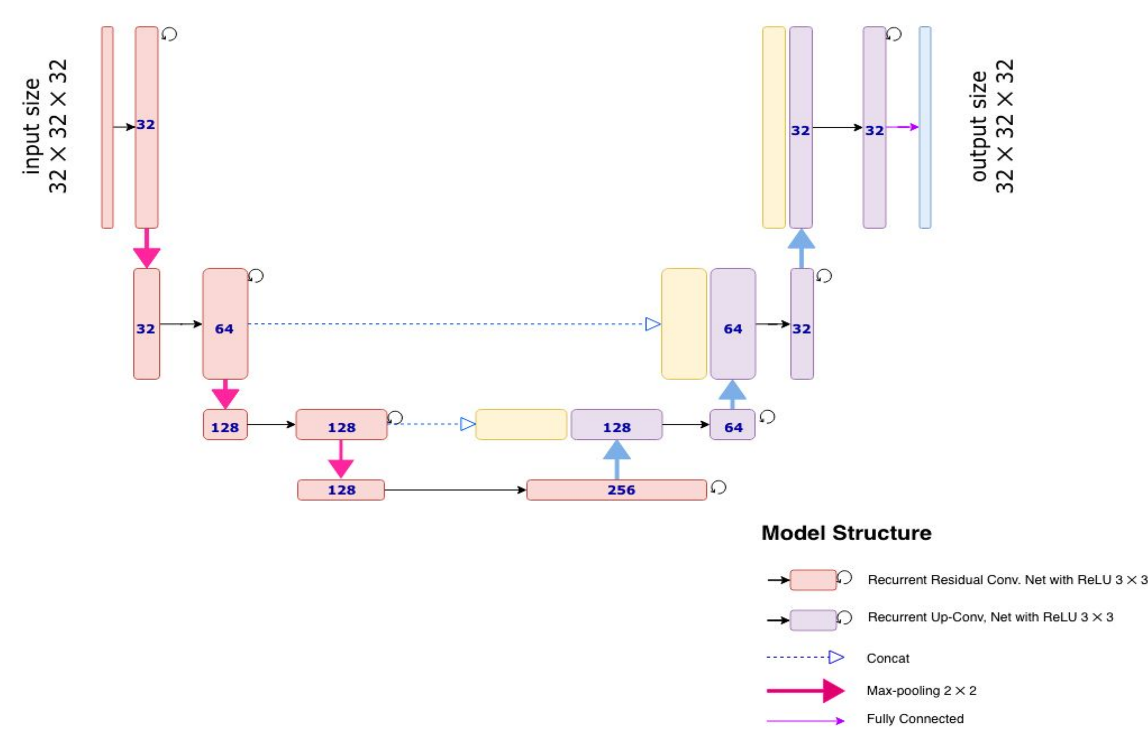


Illustration of step two of two phase netwrok:
Recurrent Residual U-Net (R2Unet) structure to predict based on previous classification mask

To further address the issue of image sparsity and Galaxy clustering, we also introduced two specifically designed loss function for our network. The idea is to penalize on the unbalanced prediction.

CounterBlobLoss:

$$Loss_{blob} = \sum_i \sum_j (1 - (n_i - n_i^{N_j}))^2 * p_i * p_i^{N_j}$$

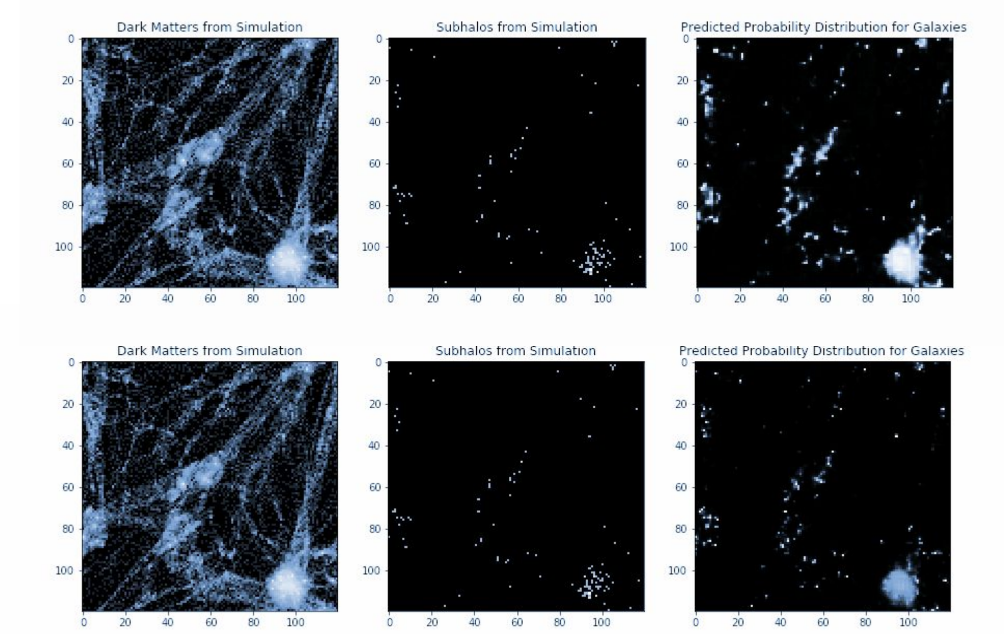
WeightedNNLoss:

$$Loss_{weighted} = \sum_{i,t_i=0} (n_i - t_i)^2 + w * \sum_{i,t_i=1} (n_i - t_i)^2$$

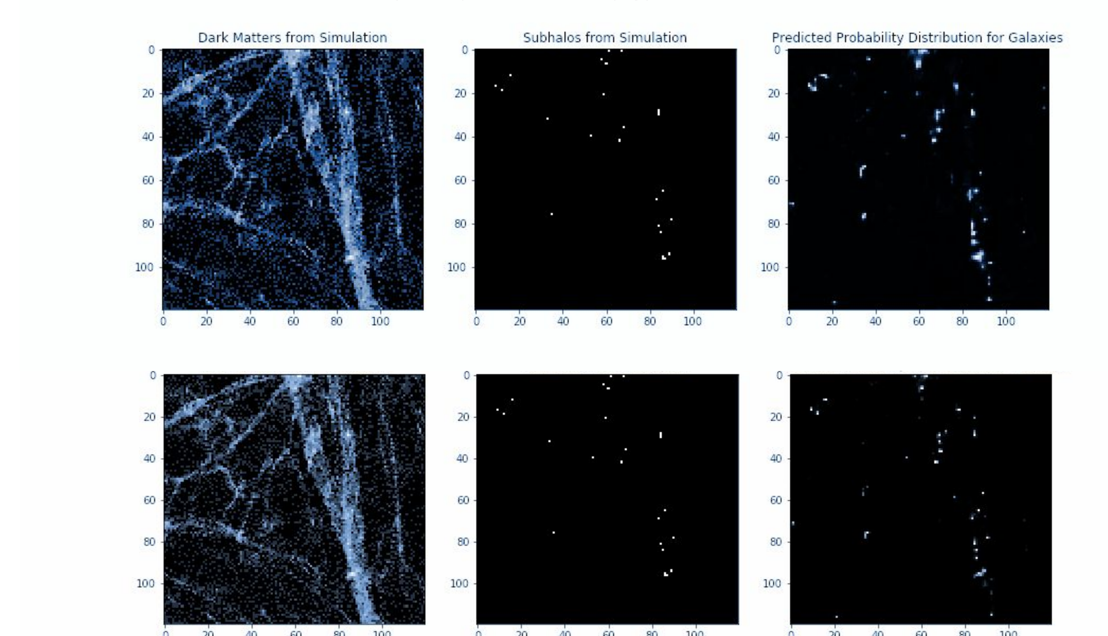
Result

Model	Configuration	Accuracy	Recall	Precision
R2Unet	Loss Weight: 1	99.72	40.49	70.54
R2Unet	Loss Weight: 5	99.52	63.17	41.91
R2Unet	Loss Weight: 10	99.29	74.8	32.42
R2Unet	Loss Weight: 25	98.8	84.31	21.05
R2Unet with Counterblob loss	Loss Weight: 5, wblob=1	99.66	49.27	52.42
R2Unet with Counterblob loss	Loss Weight: 5, wblob=10	99.72	34.73	63.76

Table of Trade-off Result Using Different Weights on Loss function



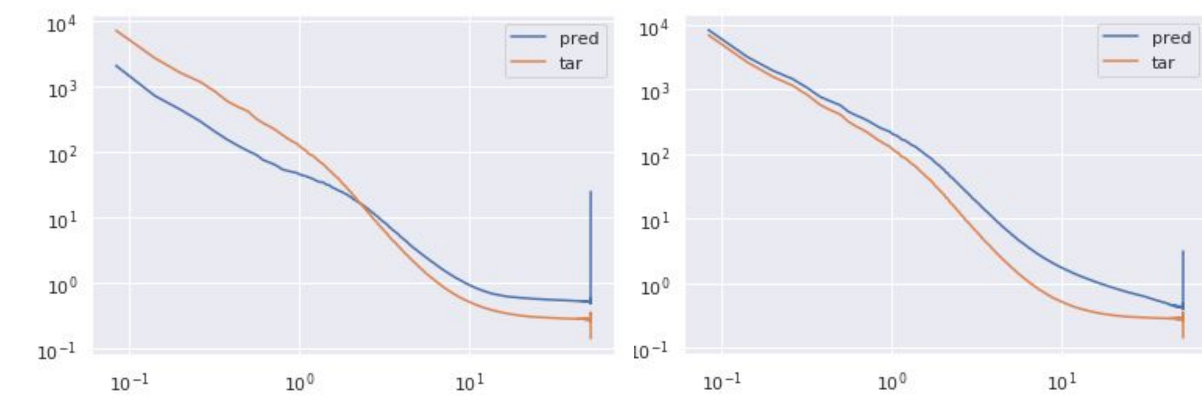
A pair of Sliced Visualization of Galaxy density prediction in a 2D box of 12.5 Mpc*12.5 Mpc
With (Down) and without (Up) Counter Blob Loss



A pair of Sliced Visualization of Galaxy density prediction in a 2D box of 12.5 Mpc*12.5 Mpc
With (Down) and without (Up) Counter Blob Loss

The model prediction on galaxy density had gain high recall compared to the target. The model is working well when predicting whether certain place exist galaxies, as illustrated in figure.

The 3D monopole power spectrum of predicted and simulated galaxy density fields are displayed below.



Monopole Power spectrum log-log comparison between testing prediction and real simulation
Using one layer conv as first phase (left), and using R2Unet as first phase (right)

Where the power spectrum is calculated by taking the Fast Fourier Transform of density fields.

$$\delta(k) = \int \delta(r) e^{ik.r} d^3 r$$

$$\delta(r) = \int \delta(k) e^{-ik.r} \frac{d^3 k}{(2\pi)^3}$$

$$P(k_1, k_2) = \frac{1}{(2\pi)^3} \langle \delta(k_1), \delta(k_2) \rangle$$

Future Work

Planning future works includes incorporate power spectrum similarity into loss function during training, evaluate model applicability on different times of universe (different redshifts) together with different attributes of galaxies, and studying convolution filters to discover physical rules.

References

[1] Ravanbakhsh, Siamak, et al. "Estimating Cosmological Parameters from the Dark Matter Distribution." ICML. 2016.
[2] Ribli, Dezső, Bálint Ármin Pataki, and István Csabai. "Learning from deep learning: better cosmological parameter inference from weak lensing maps." arXiv preprint arXiv:1806.05995 (2018).
[3] Chen, Yen-Chi, et al. "Investigating galaxy-filament alignments in hydrodynamic simulations using density ridges." Monthly Notices of the Royal Astronomical Society 454.3 (2015): 3341-3350.
[4] Alom, Md Zahangir, et al. "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation." arXiv preprint arXiv:1802.06955 (2018).
[5] Vogelsberger M. et al., 2014b, MNRAS, 444, 1518, arXiv:1405.2921

Acknowledgement

- This work was supported and supervised by the Flatiron Institute, Simons Foundation.
- Special thanks to Professor Shirley Ho, Siyu He, Gabriella Contardo and all members of the Cosmology X Data Science group at Flatiron Institute.