Forming Galaxies using Machine Learning

Xinyue Zhang [1] **Yanfang Wang** [1] **Yueqiu Sun** [1] **Wei Zhang** [1]

z

Abstract

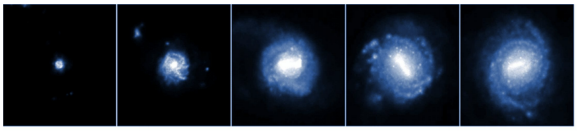
This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

The Big Bang Theory states that from 11 to 15 billion years ago, the universe was only the size of an atom. Then it suddenly starts to expand at an incredible rate. The protons and neutrons started to form gas and the gas began to from stars and galaxies. One thing that is significant to the research in how the universe evolve is the dark matter which make up mass of the universe and cannot be directly detected. This dark matter forms the skeleton on which galaxies form, evolve, and merge. In other words, the growth, internal properties, and spatial distribution of galaxies are likely to be closely connected to the growth, internal properties, and spatial distribution of dark matter halos.

Computer simulation plays an important part in the research in cosmology. Because of the important role the dark matter plays in the formation of the galaxies, computer simulation of dark matter is widely implemented and achieved great success. However, the dark matter simulation is limited in a way that it cannot predict the distribution of galaxies made up of normal matter(stars, gas, etc.). This limitation make it hard to directly connect the simulation to the observations.

The hydrodynamic simulation overcome the limitation by tracking the formation of stars and black holes, the motion of gas and the expansion of the universe. The simulation contains thousands of galaxies that agree with the observation in the real universe. However, simulations of the baryonic component require vast computation power and time, making the wide adoption of such practice unrealistic.



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

Halo Occupation Distribution (HOD) is a function of M, the mass of a dark matter halo. It describes how the distribution of the galaxies is related to the distribution of the dark matter, therefore providing us a way to populate the universe using the dark matter simulation. However, this method comes with its own limitation: parameters of the function needs to be tuned to generate the simulation of the universe. And the parameter tuning process is long and painful.

Convolutional neural networks has been shown to successfully extract information from a variety of computer vision tasks, such as image classification, detection and segmentation. CNN has been primarily applied to 2D images. In our paper, we explore the use of CNN to perform the mapping from 3D dark matter simulation to the full hydrodynamic simulation. The problem can be formulated as a supervised learning problem in which the input is the dark matter simulation and the target is the full hydrodynamic simulation. In CNN, trainable 3D filters and local pooling operations are applied to the input to capture local information and extract a hierarchical of increasingly complex features. The features can be used to predict the target simulation parameters including galaxies count, mass and etc.

2. Relevant Research

In the application of cosmology, CNN has been demonstrated to gives significantly better estimates of ω_8 and σ_8 cosmological parameters from simulated convergence maps than the state of art methods and also is free of systematic bias(Ribli et al., 2018). (Ravanbakhsh et al., 2016) estimating cosmology parameters from the volumetric representation of dark-matter simulations using 3D convolutional networks with high accuracy. (Kamdar et al., 2016) demonstrated ML as a promising and a signicantly more

*Equal contribution [1]Center for Data Science, New York University. Correspondence to: Xinyue Zhang <xz2139@nyu.edu>, Wei Zhang <wz1218@nyu.edu>, Yanfang Wang <yw1007@nyu.edu>, Yueqiu Sun <ys3202@nyu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

computationally efficient tool to study small-scale structure formation and presented a machine learning (ML) framework to study galaxy formation in the backdrop of a hydrodynamical simulation.

Semantic Segmentation aims to predicted the class for each pixel in the image and a one of the key problems in computer vision. Deep convolution neural networks are proved to achieve great performance in Semantic Segmentation task. (Simonyan & Zisserman, 2014) U-Net is one of the most important model in image segmentation.(Ronneberger et al., 2015a) The advantages of The U-Net model is 1. The maxpool layer in each level significantly reduce the number of parameters by downsizing the input image. 2. The model can be trained end to end 3. The model can be trained using very few images 4. The model is fast. Segmentation of a 512x512 image takes less than a second on a recent GPU. 5. The model achieve very good performance. It won the ISBI cell tracking challenge 2015 for segmentation of neuronal structures in electron microscopic stacks.

(Drozdzal et al., 2016) point out the importance of using skip-connections in a deep neural network. Skip connection in U-net achieve great results in image segmentation task by preserving spatial information lost in the encoder.

Here, we present a first attempt at using end to end convolution neural networks to predict hydrodynamic simulation directly from the distribution of dark matter. In the following, Section 3 introduce the data and describes our data preparation process. Section 4 introduces the methodology of One-layer convolution, U-net, R2Unet and two phase models. Section 5 presents quantitative results and visualize our prediction hydrodynamic simulation and compare it to the target hydrodynamic simulation. Section 6 concludes the paper. Section 7 propose some future work.

3. Data and its Preparation

In this project, we use the dark matter and subhalo simulations from illustris data (Vogelsberger et al., 2014), which has particle property information at a box size of 106.5 Mpc. The cosmological parameters used in the simulation is adopted from WMAP9 (Nine-Year Wilkinson Microwave Anisotropy Probe Observations) results. More speepcificly, Ω_m = 0.2726, Ω_Λ = 0.7274, Ω_b = 0.0456, σ_8 = 0.809, n_s = 0.963,H_0 = 100 h kms^1Mpc^1 and h = 0.704. (Hinshaw et al., 2012).

We use the level-1 simulations within illustris, which is the simulation with highest resolution among all. It initially consists of 6,028,568,000 hydrodynamic cells and the same number of Dark Matter particles. The snapshot we selected is captured when redshift equals 0 with the age of universe at 13.752 Gigayears, which represent the current universe. At this time step, the simulation contains 5280615062 gas resolution elements, 595243070 stellar particles, and 32552 supermassive black hole particles. The number of Dark Matter within this snapshot is 6028568000, and the number of galaxies is 4366546.

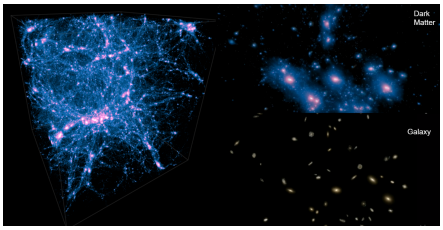


Figure 1. Visualization of Illustris dark matter simulation at reshift=0 (left), and Zoom-in visualization of corresponding dark matters and galaxies. (Right)

Following the extraction of positional information on galaxies and dark matters. We divide the whole simulation box into 1024*1024*1024 grid. The count on numbers of dark matters and galaxies in each sub-box was used to create the density grids. Within the grid, the range of particles within a cell is from 0 to 747865 for dark matters and 0 to 10 for galaxies. And the percentage of non-zero cells is 44.99% for dark matters and 00.37% for galaxies. Some Visual Representation of the data can be found in figure 2.

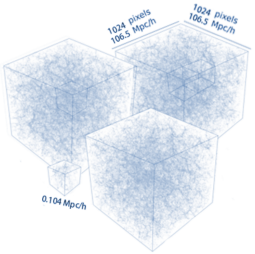


Figure 2. Illustration of grid representation of whole simulations

Since any galaxy survey can only observe a particular part of the Universe, consisting of an angular mask of the area observed, and a radial distribution of galaxies. On top of the particle densities, in order to correct for a spatially varying galaxy selection function, we normalize the observed galaxy and dark matter density p(x) to a dimensionless over-density field displayed in equation 1, where $p(x)$ is the expected mean density. This corrected field is more related with gravitational velocities and hence has more relevance with the formation of galaxies.

$$\delta(x) = \frac{\rho(x) - \bar{\rho}(x)}{\bar{\rho}(x)} \tag{1}$$

During modeling, the density data were separated into corresponding sub-boxes of size 32*32*32 and used as independent training points. There are a total of 32768 unique boxes being used during this project. And because of the homogeneous of particles in Universe. The top 71.8% of all boxes are used for training, the following 9.3% are used for validation and then the bottom 18.7% are used for testing.

4. Models

For this section, we are going to discuss the models that we tried throughout the project, including one-layer convolution, U-Net and Recurrent Residual U-Net. We also develop a two-phase training model, composed of a classification mask and a regression prediction, to get the final prediction.

4.1. One-layer Convolution

The first model that we tried is a simple one-layer convolution neural network. The neural network is composed of a convolution with 8 kernels, each with size 3×3, followed by a ReLU6 non-linearity, and then another 1×1 convolution followed by ReLU. The detailed structure is illustrated in Fig 4. The first 3×3 convolution extract local features, and the second 1×1 convolution acts as a fully-connected layer, to predict galaxy number from these local features.

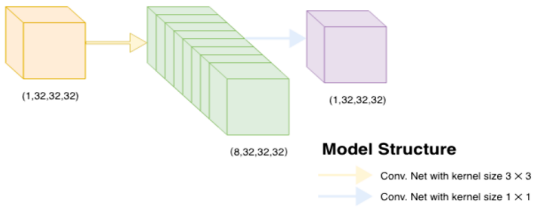


Figure 3. One Layer Convolution Neural Network

The inspiration for the model comes from our assumption about how galaxy is formed. The assumption is that, in a certain region, the more dark matter there is, the more likely that they would come into each other, collide, and finally foster a galaxy. This assumption implies a local correlation between the number of dark matter particles and number of galaxies, which resonates with the assumption we have in convolutional neural nets applications.

4.2. U-Net

Of course our assumption about how galaxy is formed is relative naive. Besides the local information, information from larger scale is also helpful in prediction, since a galaxy do get influenced also from distant galaxies. These addition in assumption propels us to also try U-net, which is a model that can take both local structure and global structure into account.

U-Net as deep convolution neural network first proposed in (Ronneberger et al., 2015b) working with image segmentation. The network first reduced the spatial information as the increment of feature information with conventional neural nets followed by ReLU and MaxPooling, then gradually recover to the target size by level with up convectional layer. At each level of expansive pathway, the concatenation of high-resolution feature from the previous step was performed, as illustrated in (Figure 4). This architecture is reasonable to use for our task as it is in coincidence with of purpose of feature extraction and image reconstruction on 3D representations.

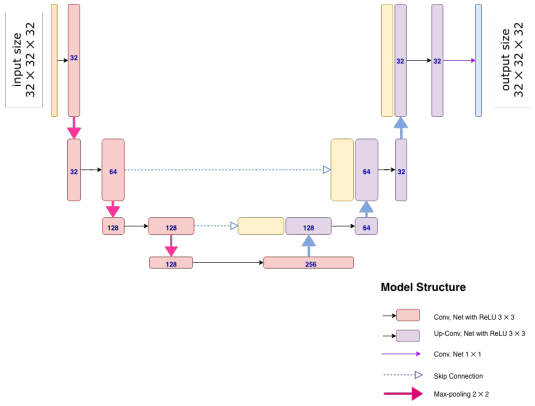


Figure 4. U-Net

In addition to the original architecture, change has been made to yield better galaxy prediction result. At first, the prediction from the model showed over-weighted similarity between itself and the input. Thus, the last skip-connection (the highest level) has been removed from the model to discourage it from simply copying information from high resolution feature.

4.3. Recurrent Residual U-Net (R2Unet)

Recurrent Residual U-Net proposed in (Alom et al., 2018) was an upgrade from U-Net with deeper structure. Except the skip connection mechanism, there is one Recurrent Residual Convectional Neural Net (R2CNN) block at each level for both contracting path and expansive path. Within the R2CNN block, it contains 2 recurrent convolution layers, and the residual connection mechanism has been applied to the input of the output from the last recurrent convolution layer.

In each recurrent convolution layer, the initial state of the input followed with t times convolution layers. At each time step, the output was element-wise summed up with the input

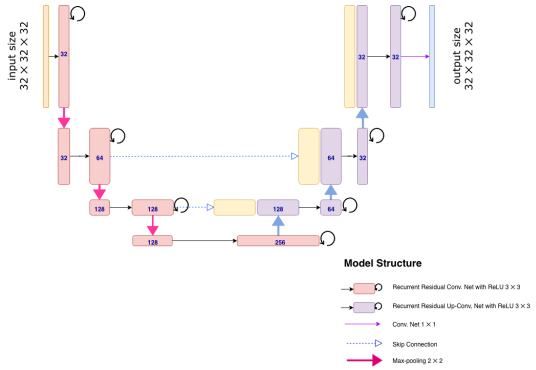


Figure 5. Recurrent Residual U-Net

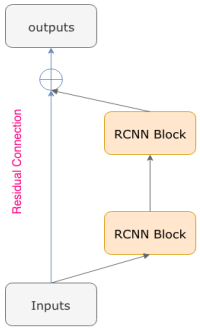


Figure 6. Recurrent Residual Convolution Block (R2CNN)

as the new input to next time step. The recurrent convolution block helps to accumulate the feature from different time step to accumulate the feature representation.

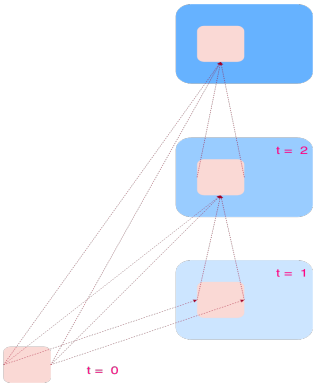


Figure 7. Recurrent Convolution Block t = 2

4.4. Two Phase Models

The high sparsity in our simulation dataset (99.6% of whole pixels does not contain any galaxies) adds a great difficulty to our training. Because of the high sparsity, even if our model predicts 0 galaxy everywhere, we would still get a high accuracy. In order to overcome this problem, we modified both our loss function and model structure.

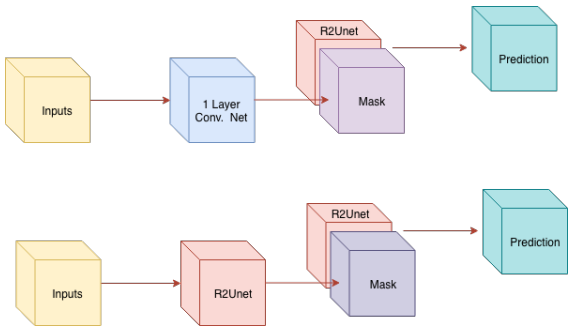


Figure 8. Two Phase Model Structure

Instead of the regular l2 loss, we use a weighted l2 loss function, which has a form of

$$L_{weighted} = \sum_{i:t_i=0} (p_i - t_i)^2 + w \times \sum_{i:t_i>0} (p_i - t_i)^2$$

where p_i is our prediction of the number of galaxies in position i, and t_i is the groundtruth number of galaxies in position i. w is the relative weight of how much we should penalize more if the model predict a position with galaxies to be 0, compared to if the model predict galaxy existence in a position with no galaxy in the target. The higher w is, the higher the recall for our model, and at the same time precision would be lower.

We also modified our model structure. Instead of using one model to do the final prediction, we divide the prediction into two phases. In the first phase, we train a model to predict what are the locations that has a higher probability of having galaxies. We use this model as a mask, and train another model separately to do the prediction on the number of galaxies. The final prediction is the prediction from the second model, with the locations' corresponding probability of having a galaxy, which is the output from our mask model, greater than a certain threshold. The model structure is illustrated in Fig 3. We hope that by dividing the prediction into two phases, our model is be able to learn to how gradually refine its predictions, which is easier than learning everything at once.

In our experimentation, we tried two settings of the two-phase model: one with a one-layer convolution mask model and R2Unet prediction model, another with both mask and prediction model be R2Unet. The prediction is evaluated in the power spectrum in Fig 6.

4.5. Counter-Blob Loss

Following the two phase training, we first train a classification model, to predict how likely a position has a galaxy. From the visualizations, we find that the model tend to predict a blob of areas to have high probabilities, which means it predict every position within the blob to have a large probability. This blobby prediction is not desired compared to the real target. In order to remove this blob, we come up with a counter-blob loss, which has the form of

$$L_{blob} = \sum_i \sum_{j \in N_i} (1 - (p_i - p_j))^2 p_i p_j$$

where p_i and p_j are the probabilities of galaxy existence in position i and j. N_i denote all neighboring position for position i. This loss would push for more differences for adjacent positions, which removes the blob, because blob is caused when all neighboring positions have the same high probability of having galaxy. We also want to penalize more when the probability is large, for example, we want to push for more difference between two adjacent positions with probability 0.9, 0.9, compared to two positions with probability 0.1, 0.1.

We combine this loss together with weighted cross entropy loss to train the first phase mask model. The result can be seen in Fig 7.

5. Result

5.1. Quantitative Result

Evaluating the result, the weighted L2 and counter blob loss had been stunningly effective. Different weight on L2 loss give different precision and recalls when evaluating the result density as a binary field. Adjusting weights will give us a trade-off between recall and the precision, as seen in (Table 1). R2Unet shows similar accuracy comparing to Original Unet, but increased in performance on both Recalls and Precisions.

Counter-blob loss also boosted the performance by increasing precision and the overall accuracy.

Table 1. Trade-off result using different weights on Loss Function

Model	Configuration	Accuracy	Recall	Precision
One-Layer Conv	Loss Weight: 300	93.8	98.8	6.3
Unet	Loss Weight: 5	99.6	59.8	39.2
Unet	Loss Weight: 10	99.4	69.3	29.7
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

The inclusion of Counter-blob loss was also being effective in reducing the incorrect galaxy clustering appeared in prediction. As seen in figure 9 and 10, the prediction around galaxy clusters appeared in high density fields of the original sources was significantly decreased in size and appearance.

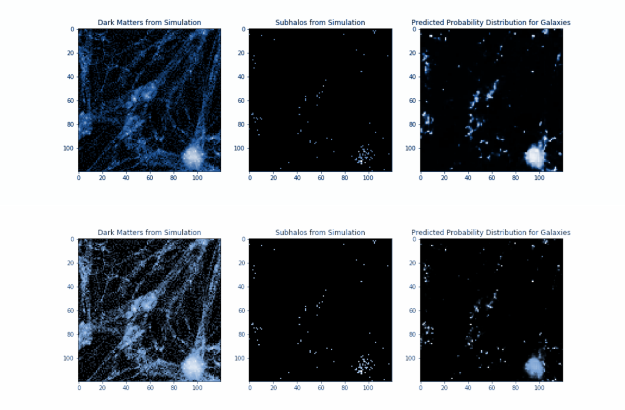


Figure 9. Source, Target and Prediction Visualization For sliced simulation with (Bottom) and without (Top) Counter-Blob Loss

5.2. Visualization

As figure 11 shows, our model effectively learned the position of the galaxies when the distribution of galaxies and dark matters are relevantly sparse. However, when its compact, it shows that our model could not predict the exact position. One of the reason is the imbalance of our dataset. As described in the above section, we tried to minimize this effect by enforcing weight and build the loss function to penalize the blobs within the prediction. In all, it approximately reflected the density distribution of the galaxies.

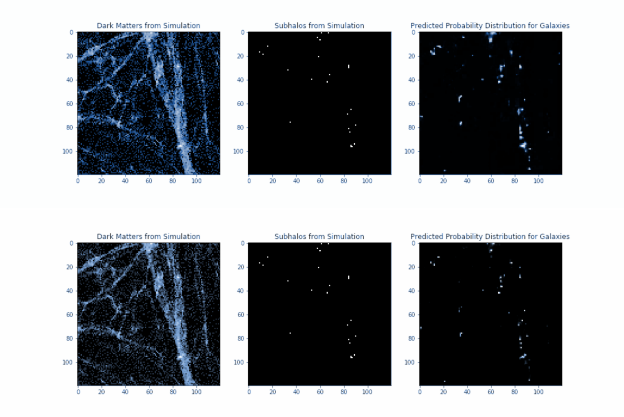


Figure 10. Source, Target and Prediction Visualization For sliced simulation with (Bottom) and without (Top) Counter-Blob Loss

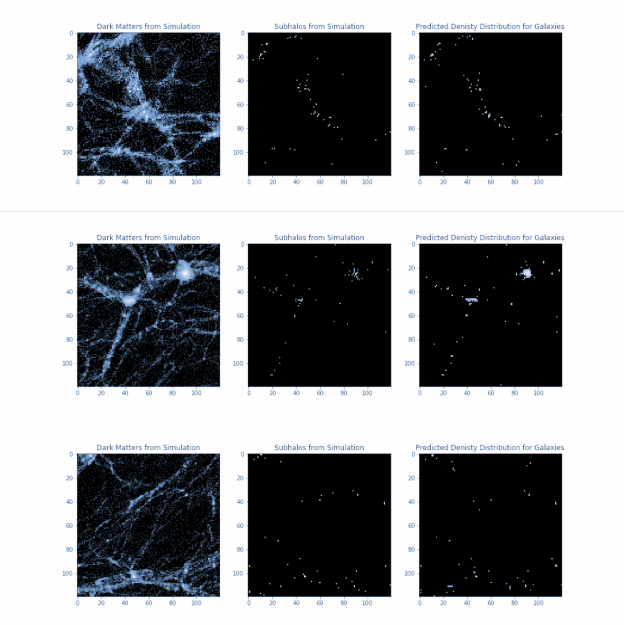


Figure 11. Two Phase Result

5.3. Power Spectrum

The 2-point correlation function as used in astrophysics describes one way in which the actual spatial or the angular distribution of galaxies deviates from a simple Poisson distribution. The power spectrum is the Fourier transform of the correlation function. The spatial two-point or function is defined as the excess probability of finding a pair of galaxies, compared with that expected for a random distribution. And the power spectrum was related to two point correlation by equation 2.

$$\xi(r) = \frac{1}{2\pi^2} \int dk\, k^2 P(k) \frac{\sin(kr)}{kr} \tag{2}$$

Hence power spectrum can be considered an evaluation metric for the statistical properties of galaxy density field.

The 3d Monopole power spectrum of our predicted density and actual simulated galaxy density from are displayed below. Where the reference line on noise is calculated by (Equation 3)

$$n = \log \frac{(*Galaxies)}{Volume} \tag{3}$$

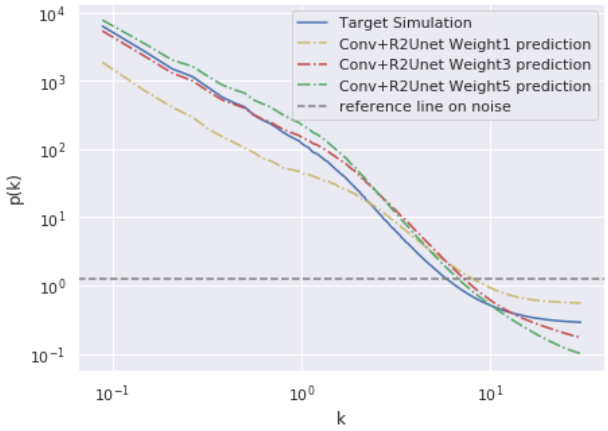


Figure 12. Monopole Power spectrum log-log comparison between testing prediction and real simulation for both model with one layer conv as first phase and R2Unet as first phase

Where the power spectrum in this case is calculated by taking the Fast Fourier Transform on density fields.

$$\delta(k) = \int \delta(r) e^{ik.r} d^3r \tag{4}$$

$$\delta(r) = \int \delta(k) e^{-ik.r} \frac{d^3k}{(2\pi)^3} \tag{5}$$

$$P(k_1, k_2) = \frac{1}{(2\pi)^3} \langle \delta(k_1), \delta(k_2) \rangle \tag{6}$$

We can see from the power spectrum comparison that the two model predicts better in terms of long-term frequencies. But still have some divergence when dealing with short-term frequencies. This is consistent with our assumption that noises and localized correlations are harder to predict solely based on dark matter density.

6. Conclusion

As we illustrated above, deep neural network, especially Residual Recurrent U-net has the better performance to predict from dark matter distribution to galaxies density distribution than the shallow neural network models. In addition, the new created loss function, "counterblob" loss, works well to penalize the blobs appeared in the galaxy density distribution. Further more, as the result shown in power spectrum plot, the proposed "Two Phase Training" eventually generated the quite satisfied density distribution for galaxies. In all, we prove the applicability of Machine Learning models in the task of Simulating Galaxies.

7. Future Work

Planning future works includes incorporate power spectrum computation into loss function during training. This will allow us to optimize the model based on distribution similarity, which might helps to capture more underlying correlation while training.

We also plan on evaluate model applicability on different times of universe, and seek for generalization. Adding velocity and merge tree information is also another direction to further boost the model performance.

Acknowledgements

This work was supported and supervised by Prof. Shirley Ho, Siyu He, Gabriella Contardo. Thanks to everyone in Center of Data Science at New York University and all members of the Cosmology X Data Science group at Flatiron Institute.

References

Alom, Md. Zahangir, Hasan, Mahmudul, Yakopcic, Chris, Taha, Tarek M., and Asari, Vijayan K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *CoRR*, abs/1802.06955, 2018. URL http://arxiv.org/abs/1802.06955.

Drozdzal, Michal, Vorontsov, Eugene, Chartrand, Gabriel, Kadoury, Samuel, and Pal, Chris. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187. Springer, 2016.

Hinshaw, G., Larson, D., Komatsu, E., Spergel, D. N., Bennett, C. L., Dunkley, J., Nolta, M. R., Halpern, M., Hill, R. S., Odegard, N., Page, L., Smith, K. M., Weiland, J. L., Gold, B., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., Wollack, E., and Wright, E. L. Nine-year wilkinson microwave anisotropy probe (wmap) observations: Cosmological parameter results. 2012. doi: 10.1088/0067-0049/208/2/19.

Kamdar, Harshil M, Turk, Matthew J, and Brunner, Robert J. Machine learning and cosmological simulations–ii. hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 457(2):1162–1179, 2016.

Ravanbakhsh, Siamak, Oliva, Junier B, Fromenteau, Sebastian, Price, Layne, Ho, Shirley, Schneider, Jeff G, and Póczos, Barnabás. Estimating cosmological parameters from the dark matter distribution. In *ICML*, pp. 2407–2416, 2016.

Ribli, Dezső, Pataki, Bálint Ármin, and Csabai, István. Learning from deep learning: better cosmological parameter inference from weak lensing maps. *arXiv preprint arXiv:1806.05995*, 2018.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015a.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation, 2015b.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vogelsberger, Mark, Genel, Shy, Springel, Volker, Torrey, Paul, Sijacki, Debora, Xu, Dandan, Snyder, Gregory F., Nelson, Dylan, and Hernquist, Lars. Introducing the illustris project: Simulating the coevolution of dark and visible matter in the universe. 2014. doi: 10.1093/mnras/stu1536.