Inpainting Dark Matter from Galaxy Halo Maps

Celine, Lauren, May, Raul, Xing, and Yuki
ULAB Group 3 — Mentored by Cooper

December 18, 2023

Abstract

We plan to create a machine learning model which can predict the large-scale distribution of dark matter from distributions of galaxies. We will be adapting deep-learning techniques in image inpainting, commonly used for tasks such as photo colorization and detail reconstruction, to inpaint dark matter distributions onto three-dimensional galaxy halo data. We hope improvements to the accuracy and resolution of dark matter inference will better test the correlation of galaxy maps against other surveys on the large-scale structure, such as weak-lensing and Lyman-α maps.

1 Introduction

The large-scale 3D distribution of dark matter (DM) in the universe is difficult to experimentally determine. Producing a faint signal from weak gravitational lensing, this distribution is primarily inferred from cosmological surveys of baryonic matter. For massive, gravitationally-bound systems of dark matter (known as DM halos), owing to their role in the formation of galaxies and galaxy clusters, their masses and positions can be strongly correlated with galaxy luminosities and positions measured in redshift catalogues such as 2df-SDSS, DES, and DESI. (Citation needed) On the other hand, despite constituting a much larger mass fraction than DM halos, extragalactic dark matter is much more difficult to detect. Most commonly found in cosmic filaments stretched between galaxy clusters, extragalactic DM is associated with the sparse, low-luminosity intergalactic medium (IGM). Measurements of their distribution relies on observations of the 21-centimeter line and Lyman-α forest, such as (What surveys?).[1]

As many more of these DM surveys come online in the next decades, with larger and larger datasets, it is vital to develop methods which can cross-verify surveys of different kinds against each other. In particular, an accurate algorithm for predicting the distribution of extragalactic DM from the distribution of galaxies will be able to test the correspondence between galaxy halo and extragalactic DM surveys. It allows correlating the predictions of DM from, for example, measurements of weak-lensing and the Lyman-α forest against the large dataset of galaxy maps. Such an algorithm will thus improve our observational understanding of dark matter structure at the largest scales.

Therefore, we plan to develop a deep-learning neural network which can infer the large-scale distribution of dark matter from the locations, masses, and effective radii of dark matter halos. Due to the lack of observational data for the former, we will be training the neural network from DM hydrodynamic simulations. These simulations output both three-dimensional DM density fields as well as any gravitationally-bound halos identified within them.

After thorough training and statistical validation, we hope to supply our model with observational data on galaxy halos and generate predictions for the distribution of dark-matter in the local universe.

2 Prior Work

Recently, image inpainting models have been applied to cosmology to add baryon distribution data to dark matter maps. These works are able to achieved remarkably fast and accurate predictions for the distribution of gas and galaxies from dark-matter-only simulations.

We hope to explore the inverse process: given a list of simulated galaxy halo locations and masses, our model aims to inpaint a distribution of dark matter.

[2]

[3]

3 Methods

We plan to use U-Net, a deep image-to-image convolutional neural network architecture with many uses in processing and generating high-dimensional data. In particular, we wish to leverage its image inpainting capabilities to populate a voxelized distribution of galaxy halo masses and velocities with predictions of dark matter masses and velocities.

The supervised learning data will be from traditional cosmological simulations, with loss functions based on direct correlation and other statistics. We may incorporate adversarial training (i.e. GANs) with baryon inpainting algorithms if there is sufficient time.

After training, we will apply the model to real-world galaxy distribution data from redshift surveys such as SDSS and DESI and compare the predicted dark matter distribution to diffuse gas distributions from Lyman-alpha forest surveys such as BOSS and 2dF-SDSS.

We plan to implement this in Python with wide usage of the NumPy and PyTorch libraries. Most work will be done on Jupyter notebooks to facilitate online collaboration through Google Colab. Computationally intensive neural network training will be performed on Google Colab's Nvidia A100 chips.

3.1 Dataset

Our training data will be derived from large-scale hydrodynamic simulations published by the IllustrisTNG project. Specifically, we will be using the halo trees and matter density fields generated from IllustrisTNG's dark-matter-only runs.

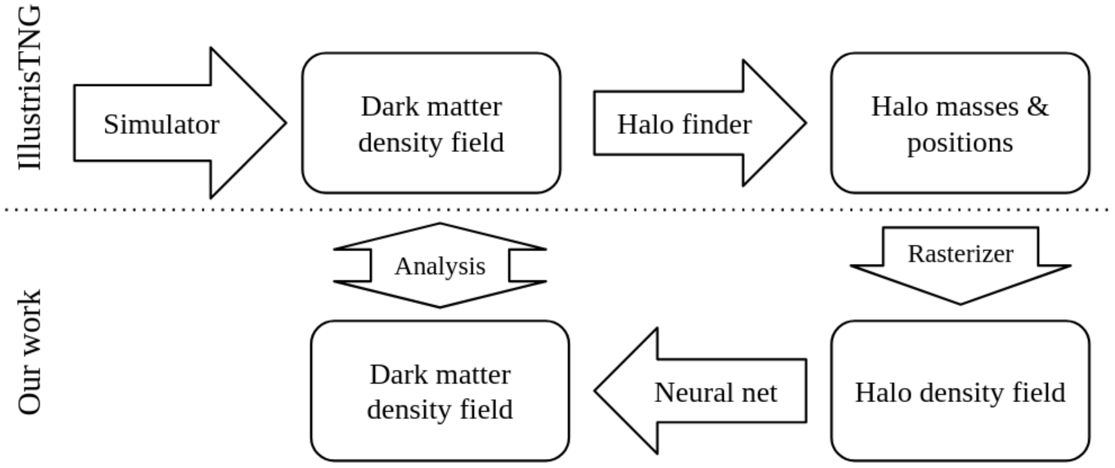

Figure 1: Schematic of our methodology.

IllustrisTNG dark-matter-only runs simulates the gravitational interactions between cold dark matter particles (CDM). Outputs 100 time-slices, though we are not working across time-scales.

For every time-slice, IllustrisTNG runs a recursive friend-of-a-friend (FoF) search for gravitationally bound halos. This generates its halos trees. Here, each halo is associated with a total mass GroupMass, along with various comoving radii Group_R_Crit$N$ inside which the halo has a mean density N times the universe's critical density ρ_c.

The HD5 file format will be parsed with the H5Py library.

3.2 Preprocessing

3.2.1 Halo tree rasterization

The tree structure of the IllustrisTNG halo data is not directly compatible with the grid structure of U-Net inputs. Thus, a rasterization stage must first be performed to stack the individual halos onto a single field. There are a number of approaches possible here; Motivated to reduce the workload of the neural net, we will rasterize each halo as a physically realistic estimate of its mass density distribution. This way, the neural net can bypass some initial conditions and focus on inferring perturbative interactions between halos.

The mass density estimate will be generated from the Einasto profile. While the Navarro–Frenk–White profile has simpler parameters, it ignores baryon interactions and has a singularity at $r = 0$, making sampling and normalization difficult.[4] But though better-behaving, the Einasto profile needs some modifications to be compatible with the IllustrisTNG halo parameters. This is documented in Appendix A.

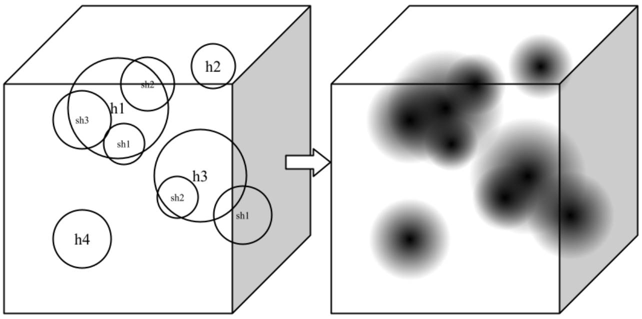

Figure 2: Rasterizing halos and subhalos by sampling their density profiles.

The input to the neural network will thus be a three-dimensional array of floating-point values, matching the resolution of the output density field, with each voxel summing all halo densities evaluated at its position from the modified Einasto profile in Appendix A.

3.2.2 Dark matter density field

Minor differences in format

3.3 U-Net

U-Nets are a type of deep learning network that function via a U-shaped encoder-decoder system. They utilize a multilayered convolutional neural network to interpret images, primarily by segmenting them. Some of its most basic functions are to colorize and enhance images, which is particularly useful in our project as the process of inpainting is at its core "colorizing" the dark matter on top of images of galaxy halos.

More specifically, we wish to leverage its image inpainting capabilities to populate a voxelized distribution of galaxy halo masses and velocities with predictions of dark matter masses and velocities. The supervised learning data will be from traditional cosmological simulations such as IllustrisTNG, with loss functions based on direct correlation and other statistics.

L1, L2, statistical tests.

3.4 Analysis

We intend on applying our model to real world galaxy distribution data from redshift surveys such as SDSS and DESI, and subsequently comparing our predicted dark matter distribution to diffuse gas distributions obtained from Lyman-alpha forest surveys such as BOSS and 2df-SDSS.

In particular, we will examine grid-cell-level occupancy in order to determine whether or not our predicted dark matter clumps are in the correct location. We will also examine mean squared error, which is indicative of whether or not the predicted density of the dark matter clumps is correct. Finally, we will also look at power spectra to determine whether or not our predicted dark matter clumps are the correct size.

Power spectrum analysis will be done with the PyLians library (https://pylians3.readthedocs.io/en/master/)

3.5 GAN

Apart from U-net architecture, we will also be able to utilize GAN's which consist of two key separate neural networks: the Generator and the Discriminator. The generator creates data resembling the input data distributions while the Discriminator evaluates this created data against the real data set. This results in a continuous adversarial process between the two networks, which leads to the generation and refinement of data. Which on a basic level work as a judge and counterfeiter, with the judge being the Discriminator and the Counterfeiter being the Generator.

Due to this structure, GANs are continuously improving which is helpful with our project of translating from Halo Density Fields to Dark Matter Density Fields.

...

4 Timeline

Dividing this project into smaller milestones, we will start with the 2D static mass-field-to-mass-field translation case, progress to 3D, and then incorporate velocity fields. Each step will increase the programming complexity and computational power. Our training data (IllustrisTNG) and untrained machine learning model (U-Net) are both open-source, so there should be no issues in access. We foresee some technical challenges in getting U-Net to run on our variety of computers, but given its popularity, we hope we will be able to quickly resolve them.

4.1 Initial targets

4.2 Possible roadblocks

4.3 Aspirations

GAN.

Appendix A: Halo profile

The Einasto model for the halo density profile is originally parameterized in [4] by the half-mass radius r_e, half-mass density $\rho_e \equiv \rho(x) : r = r_e$, and shape profile n as

$$\rho(x) = \rho_e \exp(d_n - x), \qquad ([4]\ 20)$$

where the normalization coefficient d_n satisfies $\Gamma(3n) = 2\gamma(3n, d_n)$, and the fractional radius $x = d_n(r/r_e)^{1/n}$. Γ and γ are the complete and incomplete gamma functions respectively.

Integrating ([4] 20) gives the total mass

$$M = \frac{4\pi n r_e^3 \Gamma(3n) \exp(d_n)}{d_n^{3n}} \rho_e. \qquad ([4]\ 22)$$

and the enclosed mass profile

$$m(x) = \frac{\gamma(3n, x)}{\Gamma(3n)} M, \qquad ([4]\ 22)$$

Although IllustrisTNG's FoF halo data include each halo's total mass M and several threshold radii (such as R_{Crit200} and R_{Crit500}), it does not have direct equivalents of n, r_e and ρ_e. While we can easily rewrite $\rho(x)$'s ρ_e in terms of M:

$$\rho(x) = \frac{M d_n^{3n}}{4\pi n r_e^3 \Gamma(3n)} \exp(-x).$$

reparameterizing n and r_e is less trivial—in fact, with our limited information, we cannot constrain both simultaneously.

However, for our purposes, the density profile serves as only an initial estimate. So for simplicity, we fix $n \equiv 6$, near the mean n_{Ein} for galaxy-sized halos in ([4] Table 1). This lets us constrain r_e by fitting the logarithmic slope between the densities at two given threshold radii:

$$\begin{aligned} r_e &= r\left(-\frac{d_n}{nr}\frac{dr}{d[\ln\rho(r)]}\right)^n && \text{([5] 22)} \\ &\approx \overline{r}\left(-\frac{d_n}{n\overline{r}}\frac{r_2 - r_1}{\ln[\rho_2/\rho_1]}\right)^n. \end{aligned}$$

We have thus parameterized $\rho(x)$ in terms of r_1, r_2, ρ_1, ρ_2, and M — corresponding to the following fields in the IllustrisTNG dataset:

Parameter	Formula [6]
r_1	$h\cdot$GROUP_R_CRIT200
r_2	$h\cdot$GROUP_R_CRIT500
ρ_1	$200\rho_c$
ρ_2	$500\rho_c$
M	GROUPMASS

where the critical density of the universe $\rho_c = \frac{3H^2}{8\pi G}$.

References

[1] M. McQuinn, "The evolution of the intergalactic medium," *Annual Review of Astronomy and Astrophysics*, vol. 54, no. 1, pp. 313–362, 2016.

[2] P. Ganeshaiah Veena, R. Lilow, and A. Nusser, "Large-scale density and velocity field reconstructions with neural networks," *MNRAS*, vol. 522, pp. 5291–5307, 04 2023.

[3] F. G. Mohammad, F. Villaescusa-Navarro, S. Genel, D. Anglés-Alcázar, and M. Vogelsberger, "Inpainting Hydrodynamical Maps with Deep Learning," *ApJ*, vol. 941, p. 132, Dec. 2022.

[4] D. Merritt, A. W. Graham, B. Moore, J. Diemand, and B. Terzić, "Empirical Models for Dark Matter Halos. I. Nonparametric Construction of Density Profiles and Comparison with Parametric Models," *ApJ*, vol. 132, pp. 2685–2700, Dec. 2006.

[5] A. W. Graham, D. Merritt, B. Moore, J. Diemand, and B. Terzić, "Empirical models for dark matter halos. ii. inner profile slopes, dynamical profiles, and ρ/σ^3," *The Astronomical Journal*, vol. 132, p. 2701, nov 2006.

[6] The TNG Collaboration, "Data specifications." https://www.tng-project.org/data/docs/specifications/, 2023.