Super-resolving Dark Matter Halos using Generative Deep Learning

David Schaurecker*
Institute for Particle Physics and Astrophysics
ETH Zurich
Zurich, Switzerland
dschaurecker@gmail.com

Yin Li
Center for Computational Astrophysics
Flatiron Institute - Simons Foundation
New York City, USA
eelregit@gmail.com

Shirley Ho
Center for Computational Astrophysics
Flatiron Institute - Simons Foundation
New York City, USA
shirleyho@flatironinstitute.org

Jeremy Tinker
Center for Cosmology and Particle Physics
New York University
New York City, USA
jlt12@nyu.edu

Abstract

Generative deep learning methods built upon Convolutional Neural Networks (CNNs) provide great tools for predicting non-linear structure in cosmology. In this work we predict high resolution dark matter halos from large scale, low resolution dark matter only simulations. This is achieved by mapping lower resolution to higher resolution density fields of simulations sharing the same cosmology, initial conditions and box-sizes. To resolve structure down to a factor of 8 increase in mass resolution, we use a variation of U-Net with a conditional Generative Adversarial Network (GAN), generating output that visually and statistically matches the high resolution target extremely well. This suggests that our method can be used to create high resolution density output over Gpc/h box-sizes from low resolution simulations with negligible computational effort.

1 Introduction

Cosmological galaxy surveys, both current and planned, span increasingly large volumes, e.g. DESI [1], EUCLID [2], DES [3] and LSST [4]. In the previous decade, surveys such as the Baryon Oscillation Spectroscopic Survey [5] have used bright red galaxies as tracers of large scale structure. To make larger maps, the current and next generation of surveys rely on utilizing fainter galaxies. Spectroscopic surveys specifically are planning to observe galaxies targeted by the presence of star formation within them. Such galaxies, some of them Emission Line Galaxies (ELGs; [6]), are the cosmological workhorse of the next decade. An important issue with creating bigger large-scale structure with smaller galaxies is being able to simulate such maps computationally. ELGs reside in halos at or below the mass scale of our own Milky Way [7; 8]. Thus, simulations of such galaxy samples require not just large volume, but also higher mass resolution than previously needed. A single simulation with equal volume of the DESI ELG sample that properly resolves all halos in which ELGs could form would require upwards of 9500^3 particles, assuming a volume of 122 Gpc and a resolution of halos at around 1000 particles for 10^{11} $M_\odot$ halos. We often need a large suite of such simulations to explore various parameters or build up statistics. This will quickly become computationally too expensive to simulate using conventional methods.

*Corresponding author.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

A promising solution to this problem is the use of deep learning networks. Various works in recent years applied convolutional neural networks to cosmological applications successfully, generally increasing small scale structure information (e.g. galaxy distributions in [9], mapping dark matter to galaxies in [9; 10], extracting cosmological information while marginalizing over baryonic effects in [11], or for general simulation output [12; 13; 14]. In the case of creating cosmological simulations spanning large volumes, a simple first approach is to use dark matter only (DM-only) density fields.

In this paper, we propose a novel method to predict the redshift $z = 0$ density field of a high resolution simulation, from the $z = 0$ density field of a low resolution simulation of the same volume, run using the same cosmology and initial conditions. High and low resolution here refers to the simulation's particle mass and the number of particles used inside the simulation volume. The halos and subhalos in this high-resolution output can be populated with galaxies in any number of ways, as required by the specific goal of the mock galaxy catalog (see [15] for a review on the galaxy-halo connection).

This paper provides a first proof of work that uses density fields to super resolve dark matter halos. Previous works either focus on other statistics such as void abundance [14], or use particle displacement fields that assume lattice pre-initial conditions (pre-ICs, the Lagrangian particle positions) [16] and therefore cannot be applied to many of the state-of-the-art simulations using glass pre-ICs.

In particular, our presented method can be applied to dark matter simulations with glass initial conditions [17] (here after: IC) while previous super-resolution techniques such as [18] cannot. Many state-of-the-art simulations use glass IC, such as the Illustris project [19] while there is currently no satisfactory way to interpolate those glass IC particles onto a grid in order to compute their displacement field needed by previous super-resolution methods such as [16; 18]. Rather than wasting computational resources on new costly simulations for super-resolution tasks, we develop the alternative Eulerian method to exploit the existing state-of-the-art cosmological simulations initialized using glass IC.

2 Data and Method

2.1 Simulations and Data

In this work, we use redshift $z = 0$ data from the Illustris-2-Dark (high-res) and Illustris-3-Dark (low-res) dark matter only simulations [19] for training and testing. Both simulations span the same volume of $(75\ \text{Mpc/h})^3$, use the same cosmological parameters, the same glass-tiled IC and vary only in the number of simulated particles and their mass.

In our proposed method, we do however use grids to create density grids, but do not need to create a displacement field (which requires having at least one particle per pixel in order to find its displacement field).

Our deep learning method utilizes convolution layers, which are typically applied to imaging data that has translational symmetry. Convolutional neural networks (CNNs) generally work well with equidistantly gridded n-dimensional data. In our case, a fixed spatial resolution is perfect for a CNN. We prepare dark matter density maps, with a 2048^3 grid, creating a number count density field, where each 3D voxel of side-length ≈ 36.6 kpc/h contains information about the number of particles that lie inside that small region. Each voxel's field-value is obtained by TSC interpolating the Illustris catalogs onto a mesh and then log1p-normalizing, reducing the feature distribution ranges. Each of the data sets is then subsequently divided into eight sub-cubes of equal size (37.5 Mpc/h on the side), from which six are used for training, one for validation, and the last one for testing. This results in the low resolution input (Illustris-3-Dark) and high resolution target (Illustris-2-Dark) simulations being divided into 32768 3-dimensional 64^3 voxel cubes, of which an eighth is used for testing.

2.2 Method

At the 36.6 kpc/h pixel scale, dark matter clusters in the low and high resolution simulations are spatially shifted by a significant amount from each other. This is due to the addition of high-frequency modes in the high-resolution simulation. This makes the simpler training approach of supervised learning (e.g. using a mean squared error loss function) impossible, as the model does not have enough information to predict this shift and will simply blur the output. Recent works (e.g. on displacement fields not density fields [16]) showed that an unsupervised approach to deep learning

is a possible pathway forward given the limitation described here. We apply the method of GAN training [20] with a few modifications, i.e. we simultaneously train a generator (G) and discriminator (D) to generate simulation output and discriminate between real and artificial data. Nevertheless, in this work G needs to generate output, not from a randomly sampled input distribution, but from the low resolution training cubes and from white noise fields in the noise layers. In order to help the discriminator differentiate between real and fake data, the low-res input is concatenated to the generated output and high-res target before passing the data through D. The training process thus is no longer purely unsupervised, and our novel method is more accurately labeled as conditional GAN, an adapted version of the standard approach first introduced by [21], which utilized a different architecture.

This results in the following loss functions for the discriminator and generator in this work:

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_{\hat{x},x}\left[\log D(\hat{x},x)\right] \\ & -\mathbb{E}_{\hat{x},z}\left[\log(1-D(\hat{x},G(\hat{x},z)))\right] \\ & -\gamma\,\mathbb{E}_{\hat{x},x}[\|\nabla D(\hat{x},x)\|^2] \end{aligned} \tag{1}$$

$$\mathcal{L}_G = -\mathbb{E}_{\hat{x},z}\left[\log(D(\hat{x},G(\hat{x},z)))\right], \tag{2}$$

where x is sampled from the high-res distribution $p_{\text{high-res}}$, $\hat{x}$ is sampled from the low-res distribution $p_{\text{low-res}}$ and z is sampled from a random white noise distribution p_z. The third term in equation 1 is the R_1 regularization (introduced in [22]), which is applied for real data only. The chosen penalty weight $\gamma = 5$ is set constant.

2.3 Models and Training

Training is not too sensitive on the exact discriminator architecture, as long as the main idea of progressively extracting higher-level features by down-sampling convolutions is followed.

The exact generator architecture however, is extremely important to a successful training process. The model, is a shallower adaptation of the often used U-Net [12; 23; 24; 25], utilizing heavily on CNNs' translational equivariance which is especially useful in cosmological applications. The most important change from a conventional U-Net is that we replace the transposed upsampling convolution by a tri-linear interpolation followed by a usual convolution. We also add noise layers to help the generator predict the highly non-linear structure of the high-res target. Please refer to the data availability section for more details on the exact architecture.

3 Results and Conclusion

The visual differences between the low-res input and generated output / high-res target simulation become clear immediately in figure 1, especially when looking at the smaller mass halos that are missing in the low-res simulation data. As expected, the generated output matches the high-res simulation on large scales thanks to the conditional GAN training, while on small scales the generated fine structures are different from the ones in the simulated projection but look statistically consistent. For quantitative comparisons, we need to look at various matter and halo statistics which provide a more solid and consistent way to quantify model performances.

As the resolution of both low-res and high-res simulation is already very high, the most common statistics (i.e. power spectrum, particle two-point-correlation function, etc.) don't show differences between the two. This of course can consequently not be used to show the model's performance. Importantly though, the dark-matter halo two-point function (or auto-correlation function) clearly proves the model's performance by comparing statistical clustering of FoF (friends-of-friends) halos produced from the corresponding particle catalogs.

The halo auto-correlation function shown in figure 2 is calculated by counting halo pairs in binned distance-regions by using the halo's central positions $\mathbf{x_i}$, providing an estimate for the excess probability of finding halo pairs at a given spatial distance x:

$$\xi_2(x) \equiv \langle \delta(\mathbf{x_1})\delta(\mathbf{x_2})\rangle\,, \quad x = |\mathbf{x_1} - \mathbf{x_2}| \tag{3}$$

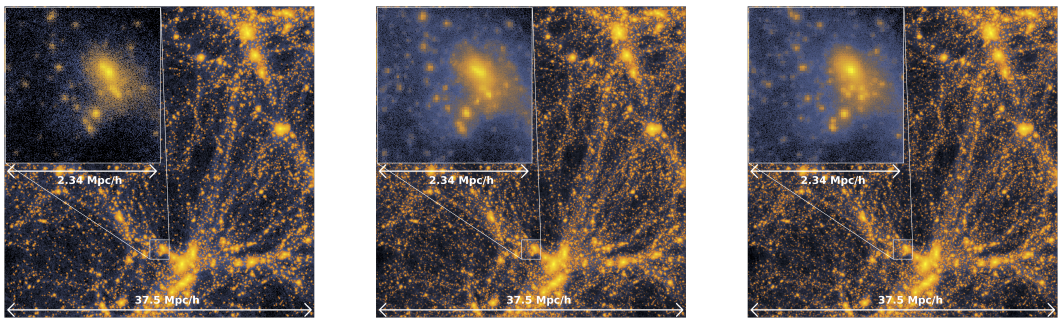

Figure 1: *left:* low-res catalog, *middle:* generated output catalog, *right:* high-res catalog. Plotted are projections on the z-axes of each catalog's particles, where FoF halo particles, are highlighted in orange. The generated output is indistinguishable from the high-res plot by eye on large scales. At smaller scales, the shapes and positions of small halos look statistically consistent but vary a bit between the two as expected, whereas they are completely missing in the low-res plot. The depicted testing-region box is not periodic as it only spans one eighth in volume of the entire corresponding Illustris simulation, which spans $(75\,\mathrm{Mpc}/h)^3$.

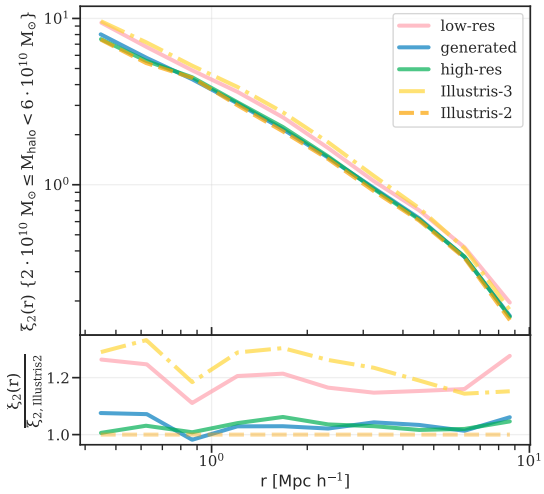

Figure 2: The halo two-point function comparison between testing data and "true" un-pixelated Illustris simulation catalog FoF halos of the testing region, calculated using the Landy-Szalay estimator. *low- and high-res:* training data, *Illustris-3 and -2:* unpixelized "true" simulation data, *generated:* network output.

Importantly for halos between $2 \cdot 10^{10}$ $M_\odot$ and $6 \cdot 10^{10}$ $M_\odot$, a clear difference between the low resolution and high resolution simulation can be seen. The network manages to match the high resolution simulation perfectly, proving its performance of producing particles affected by extremely small-scale clustering physics and not simply up-scaling low-res density fields. It also matches the two-point functions for higher mass halos, which is to be expected as those halos are larger in size and easier to predict.

This work pushes the resolution limit (of for example [18; 26], while also being more generally applicable to state-of-the-art DM-only simulations than e.g. displacement field training solutions, as it only relies on knowing the particle positions at snapshot $z = 0$ and nothing else. It provides an important first step to building a full suite going from low-res dark matter simulations, to ultimately producing ELG mock catalogs over Gpc in scale, by predicting new halos at mass ranges relevant to ELGs with almost negligible computational effort. To populate halos with ELGs, subhalos need to be accurately identified from the generated output. This requires an even smaller pixel-size and thus more training data for a given simulation volume, which will be part of future work.

We generate high-res density fields by only investing the CPU time it takes to run the low-res simulation instead of actually running the entire high-res simulation, as testing only took around 26 CPU hours (Intel Xeon Platinum 8268 24C 205W 2.9GHz Processor) for this paper's result. Training time on GPUs is hereby neglected, as finding the correct architecture, hyperparameters, etc., plays an important role in the model's training process and is impossible to quantify accurately. It only partly consists of the final 600 epoch training run, which eventually produced this work's trained model.

This result is also naturally scaleable to significantly bigger testing regions by simply passing more low-res input through the trained model.

Broader Impact

This work successfully shows the capabilities of conditional generative deep learning methods inside the field of cosmology. We showed that it is possible to accurately predict small scale highly non-linear dark matter clustering given a lower-resolution simulation with negligible computational effort. Within the astronomy community, this work will be helpful as a publicly available method to enhance the resolution of computationally expensive state-of-the-art large scale glass-IC dark matter only simulations. This pipeline provides an important first step towards producing ELG mock catalogs from low-resolution simulations, not only saving a lot of CPU-time, but making previously impossible ELG mock catalogs at scales relevant to observations, even achievable in future work.

In a broader context, we show that generative adverserial deep learning methods similar to this one, might also be used to map other general low-res density fields to a higher-res target, improving any given arbitrary simulation of fields.

References

[1] DESI Collaboration, A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L.E. Allen et al., *The desi experiment part i: Science,targeting, and survey design*, 2016.

[2] R. Laureijs, J. Amiaux, S. Arduini, J.L. Augueres, J. Brinchmann, R. Cole et al., *Euclid definition study report*, 2011.

[3] DES Collaboration, T.M.C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves et al., *Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing*, 2021.

[4] Z. Ivezic, S.M. Kahn, J.A. Tyson, B. Abel, E. Acosta, R. Allsman et al., *Lsst: From science drivers to reference design and anticipated data products*, *The Astrophysical Journal* **873** (2019) 111.

[5] K.S. Dawson, D.J. Schlegel, C.P. Ahn, S.F. Anderson, E. Aubourg, S. Bailey et al., *The baryon oscillation spectroscopic survey of sdss-iii*, *The Astronomical Journal* **145** (2012) 10.

[6] A. Raichoor, J. Comparat, T. Delubac, J.-P. Kneib, C. Yèche, K.S. Dawson et al., *The sdss-iv extended baryon oscillation spectroscopic survey: final emission line galaxy target selection*, *Monthly Notices of the Royal Astronomical Society* **471** (2017) 3955–3973.

[7] H. Guo, X. Yang, A. Raichoor, Z. Zheng, J. Comparat, V. Gonzalez-Perez et al., *Evolution of star-forming galaxies from z = 0.7 to 1.2 with eBOSS emission-line galaxies*, *The Astrophysical Journal* **871** (2019) 147.

[8] V. Gonzalez-Perez, J. Comparat, P. Norberg, C.M. Baugh, S. Contreras, C. Lacey et al., *The host dark matter haloes of [o ii] emitters at 0.5 < z < 1.5*, *Monthly Notices of the Royal Astronomical Society* **474** (2017) 4024–4038.

[9] J.H.T. Yip, X. Zhang, Y. Wang, W. Zhang, Y. Sun, G. Contardo et al., *From dark matter to galaxies with convolutional neural networks*, 2019.

[10] N. Kasmanoff, F. Villaescusa-Navarro, J. Tinker and S. Ho, *dm2gal: Mapping dark matter to galaxies with neural networks*, 2020.

[11] F. Villaescusa-Navarro, B.D. Wandelt, D. Anglés-Alcázar, S. Genel, J.M.Z. Mantilla, S. Ho et al., *Neural networks as optimal estimators to marginalize over baryonic effects*, 2020.

[12] S. He, Y. Li, Y. Feng, S. Ho, S. Ravanbakhsh, W. Chen et al., *Learning to predict the cosmological structure formation*, *Proceedings of the National Academy of Sciences of the United States of America* **116** (2019) 13825 [1811.06533].

[13] R.A. de Oliveira, Y. Li, F. Villaescusa-Navarro, S. Ho and D.N. Spergel, *Fast and accurate non-linear predictions of universes with deep learning*, 2020.

[14] D. Kodi Ramanah, T. Charnock, F. Villaescusa-Navarro and B.D. Wandelt, *Super-resolution emulator of cosmological simulations using deep physical models*, *Monthly Notices of the Royal Astronomical Society* **495** (2020) 4227–4236.

[15] R.H. Wechsler and J.L. Tinker, *The connection between galaxies and their dark matter halos*, *Annual Review of Astronomy and Astrophysics* **56** (2018) 435–487.

[16] Y. Li, Y. Ni, R.A.C. Croft, T. Di Matteo, S. Bird and Y. Feng, *Ai-assisted superresolution cosmological simulations*, *Proceedings of the National Academy of Sciences* **118** (2021) [https://www.pnas.org/content/118/19/e2022038118.full.pdf].

[17] S.D.M. White, *Formation and Evolution of Galaxies*, in *Cosmology and Large Scale Structure*, R. Schaeffer, J. Silk, M. Spiro and J. Zinn-Justin, eds., p. 349, Jan., 1996.

[18] Y. Ni, Y. Li, P. Lachance, R.A.C. Croft, T.D. Matteo, S. Bird et al., *Ai-assisted super-resolution cosmological simulations ii: Halo substructures, velocities and higher order statistics*, 2021.

[19] M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu et al., *Introducing the illustris project: Simulating the coevolution of dark and visible matter in the universe*, *Monthly Notices of the Royal Astronomical Society* **444** (2014) 1518 [1405.2921].

[20] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al., *Generative Adversarial Networks*, *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019* (2014) 3063 [1406.2661].

[21] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014.

[22] L. Mescheder, A. Geiger and S. Nowozin, *Which training methods for GANs do actually converge?*, *35th International Conference on Machine Learning, ICML 2018* **8** (2018) 5589 [1801.04406].

[23] O. Ronneberger, P. Fischer and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351** (2015) 234 [1505.04597].

[24] E. Giusarma, M. Reyes Hurtado, F. Villaescusa-Navarro, S. He, S. Ho and C. Hahn, *Learning neutrino effects in Cosmology with Convolutional Neural Networks*, *arXiv e-prints* (2019) arXiv:1910.04255 [1910.04255].

[25] C. Chen, Y. Li, F. Villaescusa-Navarro, S. Ho and A. Pullen, *Learning the evolution of the universe in n-body simulations*, 2020.

[26] D. Kodi Ramanah, T. Charnock and G. Lavaux, *Painting halos from cosmic density fields of dark matter with physically motivated neural networks*, *Physical Review D* **100** (2019) .

[27] M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu et al., *Introducing the illustris project: Simulating the coevolution of dark and visible matter in the universe*, *Monthly Notices of the Royal Astronomical Society* **444** (2014) 1518 [1405.2921].

[28] N. Hand, Y. Feng, F. Beutler, Y. Li, C. Modi, U. Seljak et al., *nbodykit: An open-source, massively parallel toolkit for large-scale structure*, *The Astronomical Journal* **156** (2018) 160.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See section 3.
   (b) Did you describe the limitations of your work? [Yes] The limitations of the pixelization where described shortly at the end of section 3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The field of cosmology (just as most of physics) does not have a direct impact on society.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We did not include theoretical results.
   (b) Did you include complete proofs of all theoretical results? [N/A] We did not include theoretical results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See section A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See section A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We ran the training with a different seed and got almost identical results. Training used up quite many GPU nodes, thus our resources where limited.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We only included the CPU time it took to produce the output. Training on GPU nodes varied a lot given the machines we had access to.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section A and the Illustris project [27] for training data.
   (b) Did you mention the license of the assets? [Yes] See appendix A.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See appendix A.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See appendix A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data used/curated does not contain information that could identify the authors or offensive content, as the relevant provided Github link is anonymized.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Does not apply.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Does not apply.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Does not apply.

A Data availability

All code and models used and created in this work are publicly available on GitHub under the following url: `https://github.com/dschaurecker/dl_halo`. Please refer to the ReadMe file in the `dl_halo` repository for a very in-depth guide through the code and the process of training and testing. The training code builds upon Yin Li's `map2map` code repository, available on GitHub as well (`https://github.com/eelregit/map2map`). It allows for general training of n arbitrary input fields to n arbitrary output fields using custom models, normalizations and loss functions. Furthermore some utilities from `nbodykit` an "Open-source, Massively Parallel Toolkit for Large-scale Structure" [28] were used in pre-processing and statistical evaluation of this work.