Emulating cosmological multifields with generative adversarial networks

Sambatra Andrianomena
South African Radio Astronomy Observatory
Cape Town, 7925
Department of Physics & Astronomy
University of the Western Cape
Cape Town 7535
andrianomena@gmail.com

Francisco Villaescusa-Navarro
Center for Computational Astrophysics
Flatiron Institute
New York, NY 10010
villaescusa.francisco@gmail.com

Sultan Hassan
Center for Computational Astrophysics
Flatiron Institute
New York, NY 10010
Department of Astrophysical Sciences
Princeton University, Peyton Hall
Princeton, NJ, 08544
Department of Physics & Astronomy
University of the Western Cape
Cape Town 7535
shassan@flatironinstitute.org
NHFP Hubble Fellow

Abstract

We explore the possibility of using deep learning to generate multifield images from state-of-the-art hydrodynamic simulations of the CAMELS project. We use a generative adversarial network to generate images with three different channels that represent gas density (Mgas), neutral hydrogen density (HI), and magnetic field amplitudes (B). The quality of each map in each example generated by the model looks very promising. The GAN considered in this study is able to generate maps whose mean and standard deviation of the probability density distribution of the pixels are consistent with those of the maps from the training data. The mean and standard deviation of the auto power spectra of the generated maps of each field agree well with those computed from the maps of IllustrisTNG. Moreover, the cross-correlations between fields in all instances produced by the emulator are in good agreement with those of the dataset. This implies that all three maps in each output of the generator encode the same underlying cosmology and astrophysics.

1 Introduction

Multiwavelength astronomy has been identified as a priority in the 2020 decadal survey. Current and upcoming missions will survey the Universe at different wavelengths, from X-rays to radio: e.g. e-Rosita, SKA, Euclid, DESI, CMB-S4, Rubin and Roman Observatories, HIRAX, CHIME, Spherex. The data from these missions is expected to help us improve our knowledge on both galaxy formation and cosmology. In order to maximize the scientific outcome of these missions, accurate theoretical predictions are needed for different physical fields (different wavelengths). Hydrodynamic simulations are the most sophisticated tools that can be used to study and model this. Their main

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

drawback is their computational cost, that being very large, limits the volume and resolution of these simulations [1]. Being able to speed up these simulations is critical in order to perform a large variety of tasks: from parameter inference to model discrimination. Previous studies aimed at accelerating various cosmological simulations of different fields using generative models [2, 3, 4, 5, 6, 7, 8]. In this work we study the possibility of using generative adversarial networks (GANs) to generate multifield images –2D maps where every channel corresponds to a different physical field– that have the desired statistical properties for each field individually but also their cross-correlations. This is crucial in order to extract most of information from different large scale surveys. We note that while the generation of multifield images has been performed in the past [see e.g. 9] this work contributes to that direction by using a more sophisticated model and a richer and more complex dataset.

2 Methods

2.1 Data

We make use of images from the CAMELS Multifield Dataset (CMD)[10]. CMD contains hundreds of thousands of 2D maps generated from the output of state-of-the art hydrodynamic simulations of CAMELS [11] for 13 different physical fields, from dark matter to magnetic fields. Every image represents a region of dimensions $25 \times 25 \times 5$ $(h^{-1}\mathrm{Mpc})^3$ at $z = 0$. In this work we use images representing three different fields –gas mass density (Mgas), neutral hydrogen density (HI), and magnetic fields magnitude (B)– generated from the IllustrisTNG LH set [see 11, 12, for details]. Every field contains 15,000 maps, each of them having 256×256 pixels, and the images are characterized by six numbers: two cosmological parameters (Ω_m and σ_8) and four astrophysical parameters (A_SN1, A_SN1, A_AGN1, A_AGN2) that characterize the efficiency of supernova and AGN feedback. The multifields are produced by stacking 3 images that represent the same physical region but represent the different fields: Mgas, HI, and B. Thus, our images have 256×256 pixels and contain three channels.

2.2 Model and training procedure

We consider the Generative Adversarial Network (GAN) described in [13], who introduced some novelties such that the generative model is capable of synthesizing high-resolution images while being trained with a relatively small number of examples. They prescribed a modified version of skip-layer connection [14], named *Skip-Layer Excitation* (SLE), which, like the original skip-layer, allows the gradient flow to be preserved through the layers. Unlike the skip connection in [14], SLE does not require the two inputs to have the same dimension as the output results from channel-wise multiplications between the two inputs, according to [13]

$$\mathbf{y} = \mathcal{F}(\mathbf{x}_\mathrm{low}, \mathbf{W}) \cdot \mathbf{x}_\mathrm{high}, \tag{1}$$

where $\mathbf{x}_\mathrm{low}$ and $\mathbf{x}_\mathrm{high}$ are two inputs which are feature maps with lower and higher resolution respectively, $\mathcal{F}$ denotes a module with learnable weights $\mathbf{W}$ that operates on $\mathbf{x}_\mathrm{low}$. The generator *G* in our model contains three SLE modules that perform channel-wise multiplications of maps at 4×4, 8×8 and 16×16 resolutions with maps of 64×64, 128×128, and 256×256 resolutions respectively.

Following the prescription in [13], the discriminator *D*, treated as an encoder, is trained with three decoders in order to enhance its ability to extract the salient features from the input maps. The auto-encoder, which consists of *D* and the decoders, is optimized using only real images by minimizing the reconstruction loss [13]

$$\mathcal{L}_\mathrm{recons} = \mathbb{E}_{\mathbf{f}\sim D(x)}[||\mathcal{G}(\mathbf{f}) - \mathcal{T}(x)||], \tag{2}$$

where $\mathbf{f}$ denotes intermediate feature maps with resolutions 8×8 and 16×16 from *D*, $\mathcal{G}$ is the decoding process and $\mathcal{T}$ represents some transformations on the real images, namely cropping and downsampling, on the real images x. For the adversarial training, we have that [13]

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_\mathrm{data}(x)}[\min(0, -1 + D(x))] - \mathbb{E}_{z\sim p_z(z)}[\min(0, -1 - D(G(z)))] + \mathcal{L}_\mathrm{recons} \tag{3}$$

as the loss function for *D* and

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z(z)}[D(G(z))], \tag{4}$$

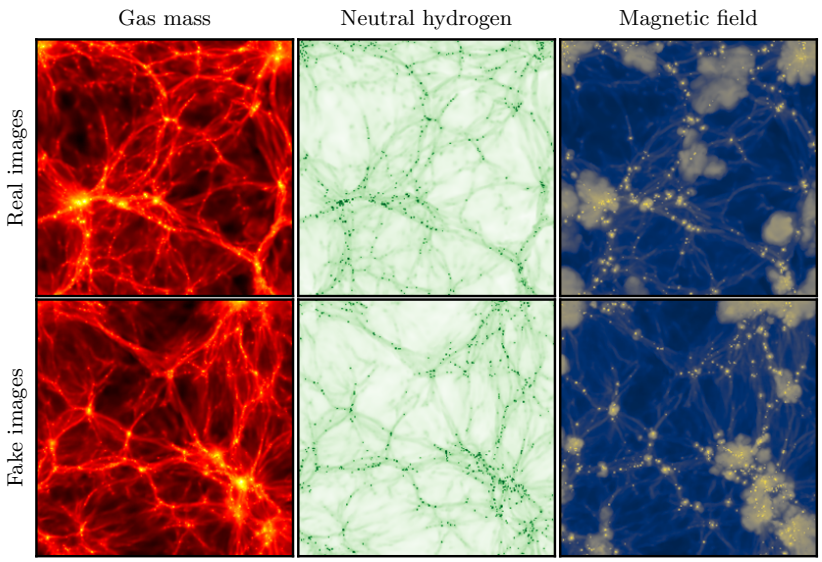

Figure 1: The top row shows images of gas mass density (left), neutral hydrogen mass density (middle), and magnetic fields strength (right) from a state-of-the hydrodynamic simulation from the CAMELS project. Note that three images represent the same region in space. The bottom row shows the same but from the our GAN model. From visual inspection we can see that the morphological features from the different fields are very well reproduced by the model.

the loss for *G*. It is worth noting that z, which is a vector from latent space, is drawn from a standard normal distribution. Our implementation is based on this. We emphasize that our model is currently not able to perform the image generation conditioned on the value of the cosmological and astrophysical parameters.

For training we set the learning rate to 0.0002 and employ the Adam optimizer. The model is trained with 10,000 multifield images for 700 epochs, which take about 12 minutes each, in batches of 8 instances on a NVIDIA GeForce GTX 1080 Ti.

3 Results

Fig. 1 shows the three channels from a real (top row) and generated (bottom row) multifield image. From visual inspection we can see that the morphological features associated to each field, from voids, filaments, halos, and bubbles, are generated with good precision. We now make use of several summary statistics to quantify the agreement between the statistical properties of the true and generated multifield images.

We first consider the probability distribution function of the pixel values of the different images. We compute the $\mu_{\rm PDF}$ and standard deviation $\sigma_{\rm PDF}$ as a function of pixel value for each field using 1,500+1,500 unseen+generated multifield images. Results are shown in Fig. 2. We find that our GAN is able to encode the key features of each field, as suggested by the overall good agreement between the $\mu_{\rm PDF}$ of real and fake maps in each channel (see top panel in Figure2). The ability of the generator to capture the variability of the pixel distribution of each field suggests that it has learned a robust lower dimension representation of the data.

Next, we quantify the agreement between clustering properties of the real and generated images by computing their auto- and cross-power spectra using PYLIANS [15]. For the 1,500+1,500 real and generated images we compute their auto-power spectrum for each channel and their cross-power spectrum between channels. We show the mean and standard deviation of the results for each wavenumber in Fig. 3 top plot in each panel, whereas the relative difference between the two power spectra at each k-mode $\left(\frac{P_{\rm fake}(k)}{P_{\rm real}(k)}-1\right)$ is shown at the bottom plot in each panel (solid green). As can

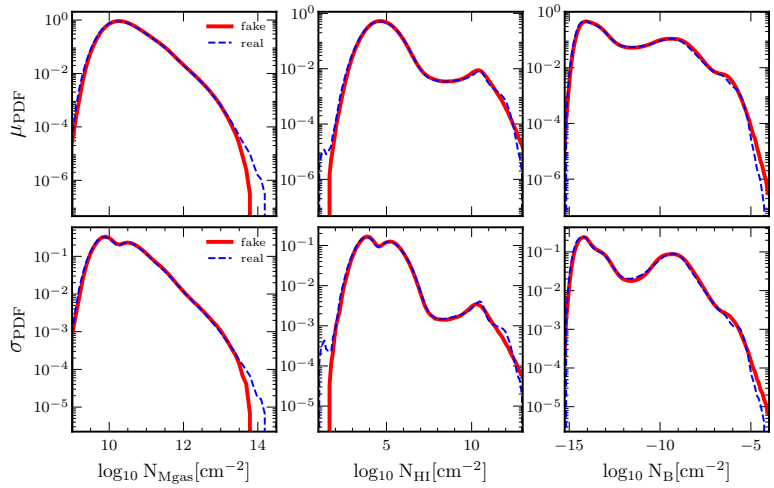

Figure 2: We have computed the probability distribution function (PDF) for 1,500 unseen real images (red) plus 1,500 generated images (blue) for each channel (field). This plot shows the mean (top) and standard deviation (bottom) of the PDF as a function pixel value for the three different fields. The agreement between the results of the true and generated images is good overall, however some statistical differences at the tails of the distributions can be noticed.

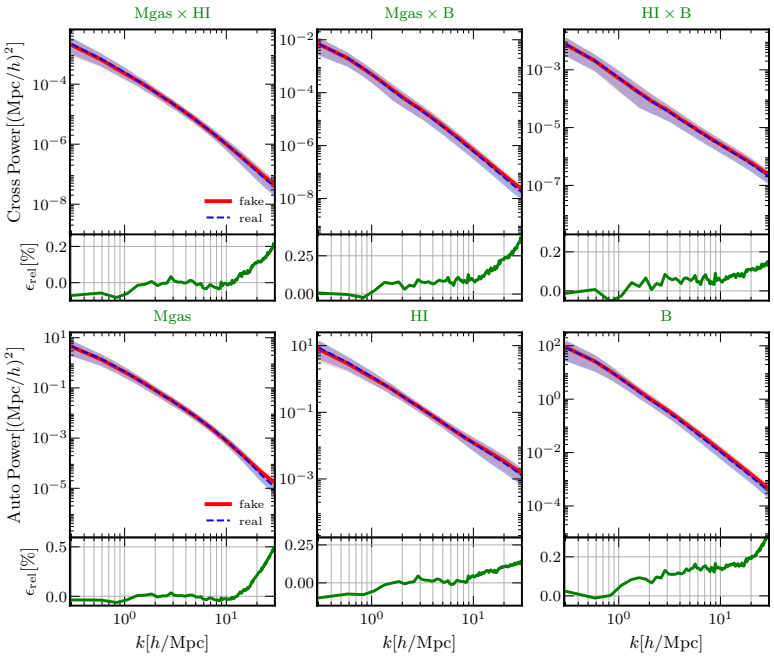

Figure 3: Same as Fig. 2 but for the cross-power spectrum (top row) and auto-power spectrum (bottom row). Relative error between the two power spectra (fake and real), denoted by solid green, is presented at the bottom part of each panel. As can be seen, the statistical agreement between the true and generated images is relatively good for this statistics. The title of each panel displays the considered field(s).

be seen, although the relatively larger difference on small scales ($k > 10\,h/\mathrm{Mpc}$), the model is able to reproduce the overall properties of the auto- and cross-power spectra from the real images very well, including their scatter induced by cosmic variance and changes in cosmology and astrophysics.

The above statistical tests indirectly point towards the fact that our model is not affected by mode collapse, an issue GANs are known to suffer from, making their training challenging. In order to further test this, we have created multifield images from different latent-space points, and by interpolating among them (in latent-space) we find a smooth transition between the images and their channels.

4 Conclusions and future work

We have used a generative model to produce multifield images with very similar statistical properties (as quantified by pdfs and power spectra) orders of magnitude faster than the ones produced from the very computationally demanding state-of-the-art hydrodynamic simulations. The agreement in the cross-power spectra indicate the models is able to preserve Fourier phases (on large scales) across channels, and therefore represents a great tool for multiwavelength studies. We note that although our model seems able to capture the effects of cosmic variance and variations in cosmology and astrophysics, it is currently unable to condition the image generation on the value of those parameters (cosmological and astrophysical). In future work we will extend our model to be able to perform this task.

5 Broader impact

Multiwavelength astronomy was recommended as a priority in the recent decadal survey in 2022, given its potential to improve our understanding on galaxy formation and cosmology. We have used a generative model that is able to produce multifield images from different physical fields that is orders of magnitude faster than the method used to generate the true images. Given the good statistical properties of the generated images, they can be used for a variety of task, from pipeline design to studies involving cross-correlations. Our model is general and can be applied to other datasets, such as images of galaxy fields to X-rays and SZ maps. We believe our model, and further developments on it, has the potential to play an important role in the multiwavelength astronomy era.

6 Acknowledgments

SA acknowledges financial support from the South African Radio Astronomy Observatory (SARAO). FVN and SH acknowledge support provided by the Simons Foundation. SH also acknowledges support for Program number HST-HF2-51507 provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, incorporated, under NASA contract NAS5-26555.

References

[1] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation. *Nature Reviews Physics*, 2(1):42–66, January 2020.

[2] Mustafa Mustafa, Deborah Bard, Wahid Bhimji, et al. Cosmogan: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, 6(1):1–13, 2019.

[3] Juan Zamudio-Fernandez, Atakan Okan, Francisco Villaescusa-Navarro, et al. Higan: Cosmic neutral hydrogen with generative adversarial networks. *arXiv preprint arXiv:1904.12846*, 2019.

[4] Nathanaël Perraudin, Ankit Srivastava, Aurelien Lucchi, et al. Cosmological n-body simulations: a challenge for scalable generative models. *Computational Astrophysics and Cosmology*, 6(1):1–17, 2019.

[5] Richard M Feder, Philippe Berger, and George Stein. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Physical Review D*, 102(10):103504, 2020.

[6] Andres C Rodriguez, Tomasz Kacprzak, Aurelien Lucchi, et al. Fast cosmic web simulations with generative adversarial networks. *Computational Astrophysics and Cosmology*, 5(1):1–11, 2018.

[7] Nathanaël Perraudin, Sandro Marcon, Aurelien Lucchi, and Tomasz Kacprzak. Emulation of cosmological mass maps with conditional generative adversarial networks. *Frontiers in Artificial Intelligence*, 4:673062, 2021.

[8] Olivia Curtis, Tereasa G Brainerd, and Anthony Hernandez. Cosmic voids in gan-generated maps of large-scale structure. *Astronomy and Computing*, 38:100525, 2022.

[9] Andrius Tamosiunas, Hans A Winther, Kazuya Koyama, et al. Investigating cosmological gan emulators using latent space interpolation. *Monthly Notices of the Royal Astronomical Society*, 506(2):3049–3067, 2021.

[10] Francisco Villaescusa-Navarro, Shy Genel, Daniel Angles-Alcazar, et al. The camels multifield data set: Learning the universe's fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, 2022.

[11] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, et al. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, 2021.

[12] Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, et al. The camels project: public data release. 2022.

[13] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Francisco Villaescusa-Navarro. Pylians: Python libraries for the analysis of numerical simulations. *Astrophysics Source Code Library*, pages ascl–1811, 2018.

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]
   (b) Did you describe the limitations of your work? [Yes] See Section 4

(c) Did you discuss any potential negative societal impacts of your work? [N/A]

(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]

   (b) Did you include complete proofs of all theoretical results? [N/A]

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)? [No] The data for the training is from CAMELS project which is publicly available. However the code will be made available upon request after the paper has been accepted in a peer-reviewed journal.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.1

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure. 3

   (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)? [Yes] See Section 2.1

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] We cited the Pylians library we use for post-processing [15] and inserted a link on the implementation our code is based on in Section 2.2

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 2.2

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data is publicly available

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

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]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]