Can denoising diffusion probabilistic models generate realistic astrophysical fields?

Nayantara Mudur
Department of Physics
Harvard University
Cambridge, MA, 02138
nmudur@g.harvard.edu

Douglas P. Finkbeiner
Departments of Astronomy and of Physics
Harvard University
Cambridge, MA, 02138
dfinkbeiner@cfa.harvard.edu

Abstract

Score-based generative models have emerged as alternatives to generative adversarial networks (GANs) and normalizing flows for tasks involving learning and sampling from complex image distributions. In this work we investigate the ability of these models to generate fields in two astrophysical contexts: dark matter mass density fields from cosmological simulations and images of interstellar dust. We examine the fidelity of the sampled cosmological fields relative to the true fields using three different metrics, and identify potential issues to address. We demonstrate a proof-of-concept application of the model trained on dust in denoising dust images. To our knowledge, this is the first application of this class of models to the interstellar medium.

1 Introduction

Generative models of astrophysical and cosmological fields can serve a multitude of purposes. Cosmological simulations take several thousand CPU hours to run and can only be generated for a limited set of parameters. There thus exists a demand for emulators: frameworks that can generate summary statistics [Heitmann et al., 2009] or the underlying cosmological fields [Feder et al., 2020, Jamieson et al., 2022] conditional on an input cosmological / astrophysical parameter set. This can accelerate parameter inference approaches that require evaluating likelihoods at intermediate values. In the context of observed astrophysical fields, statistical descriptions capable of capturing the non-Gaussian nature of the interstellar medium would aid component separation problems encountered in searches for the B-mode of the Cosmic Microwave Background [Remazeilles et al., 2018] and statistical regularization for interstellar dust mapping [Green et al., 2019, Leike and Enßlin, 2019].

Score-based generative models [Song et al., 2020] have witnessed a surge in interest because of findings that show that they surpass GANs in terms of image fidelity [Dhariwal and Nichol, 2021] and their ability to produce realistic images conditional on text inputs [Saharia et al., 2022]. These models learn the gradient of the probability density of the data to learn a generative model of the data distribution. A subset of this class includes denoising diffusion probabilistic models (DDPMs) [Ho et al., 2020]. Smith et al. [2022] used DDPMs to generate galaxy images. Rémy et al. [2020], Remy et al. [2022] used the denoising score matching framework to learn the non-Gaussian component of a prior on weak lensing convergence maps from simulations.

In this work we investigate two applications of DDPMs – one to simulation data products and one to images of interstellar dust. We train models to generate dark matter density fields from a simulation suite on grids of 64x64 and 128x128 pixels. We then compare five summary statistics between samples from the trained models and the real simulation fields. While we currently generate these fields unconditionally, this benchmarking is an important step toward using these models as emulators, and generating fields conditional on an input parameter vector. We then turn our attention to the real

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

sky, and train a model to generate square patches from an interstellar dust map. We use the trained model as a denoising model and examine how well it can reconstruct the underlying image given a noisy input.

2 DDPM Background

In this section, we briefly review the denoising diffusion probabilistic model formulation in Ho et al. [2020]. A DDPM consists of a forward and a reverse diffusion process, over a fixed number of time steps, T, where $\mathbf{x}_0$ is a draw from the image distribution and $\mathbf{x}_T \sim \mathcal{N}(0, 1)$. The forward diffusion process is defined according to a variance schedule $\{\beta_t\}$. Thus $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$ and $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{t'=1}^{t} 1 - \beta_{t'}$. The neural network $\epsilon_\theta(\mathbf{x}_t, t)$ parameterizes the reverse diffusion process: $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$ where $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\big(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\big)$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. A simplified loss function is minimized where $L_{t-1} = ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t))||^2$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Thus for each batch a set of timesteps t is uniformly sampled from $t \sim U[1...T]$ to minimize L_{t-1}.

3 Generative Models for Dark Matter Density Fields

Dataset The CAMELS Multifield Dataset (CMD) [Villaescusa-Navarro et al., 2022] from the Cosmology and Astrophysics with MachinE-Learning Simulations (CAMELS) dataset [Villaescusa-Navarro et al., 2021] was used for this application. The CMD includes an ensemble of thirteen two-dimensional physical fields for 1000 different simulation parameter vectors where each vector consists of 2 cosmological and 4 astrophysical feedback parameters. Every unique parameter vector has 15 samples and the full data set thus consists of 15000 samples of each physical field. We work with the log (base 10) of the cold dark matter mass density field from the IllustrisTNG hydrodynamical simulation at $z = 0$, at two grid sizes (64x64 and 128x128) binned down from the original 256x256. Each side of an image corresponds to $25h^{-1}$ Mpc. For the 64x64 model, we use fields corresponding to the first 60% of the parameters as our training data. We augment our fields with rotations and flips, to yield 54000 (9000x6) train fields. For the 128x128 model, we use fields corresponding to the first 70% of the parameters as our training data. We augment our data with rotations, flips and translational shifts (since the data has periodic boundary conditions). We thus have 252000 (10500x24) train fields. We apply a minmax transform that scales the minimum and the maximum pixel intensity of the full training set to [-1, 1].

Training Details We train two models, one at each resolution, for 60k iterations (batch updates). In both cases, we use a forward diffusion process parameterized by a linear variance schedule lying in the range $[10^{-4}, 2 \times 10^{-2}]$, T=2000, a batch size of 40 images, and the Huber loss in place of the L2 loss with the Adam optimizer [Kingma and Ba, 2014]. We used a learning rate of 5×10^{-4} for the model at 64x64 and 2×10^{-4} for the model at 128x128 and saved checkpoints every 2000 iterations, to enable sampling from multiple models. For the 64x64 case, we train models with 3 different seeds. We use code blocks and architecture from Hugging Face's The Annotated Diffusion model [Rogge and Rasul, 2022], `https://github.com/lucidrains/denoising-diffusion-pytorch` and Ho et al. [2020]. The architecture we use is similar to that in Ho et al. [2020], and consists of a U-Net [Ronneberger et al., 2015] with 4 down and up-sampling blocks consisting of 2 ResNet blocks [Zagoruyko and Komodakis, 2016], group-normalization [Wu and He, 2018], and attention [Vaswani et al., 2017, Shen et al., 2021]. We use the Weights and Biases framework [Biewald, 2020] for our experiments. The code for experiments in this paper is available at the following repository: `https://github.com/nmudur/diffusion-models-astrophysical-fields-mlps`.

Summary Statistics Four samples from the trained model at 128x128 are plotted in the top right panel of Figure 1. We consider three summary statistics – the power spectrum, the normalized pixel intensity histograms and the three Minkowski functionals in Figure 2 to gauge consistency between the sampled fields and the true image distribution. Minkowski functionals [Schmalzing et al., 1995, Schmalzing and Górski, 1998] are topological descriptors of fields sensitive to correlation functions beyond the second order and can be used as metrics to gauge how similar the statistics of the generated fields are to those of fields from the true distribution [e.g., Tamosiunas et al., 2021, Régaldo-Saint Blancard et al., 2022]. They are computed as integrals over excursion sets with pixels

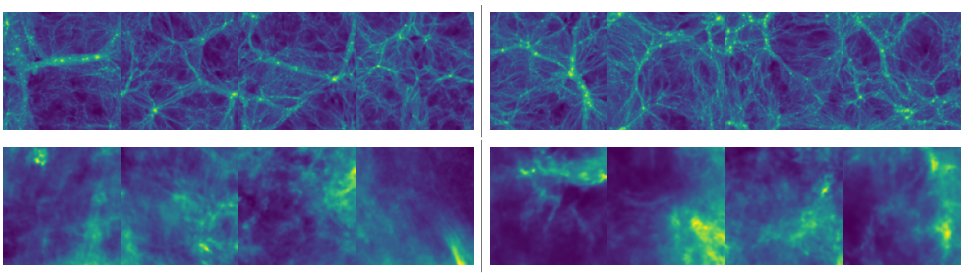

Figure 1: Four log cold dark matter mass density fields from the training data (top left) and from the sampled model (top right) at 128x128. Four samples of dust from the training data (bottom left) and from the trained model (bottom right).

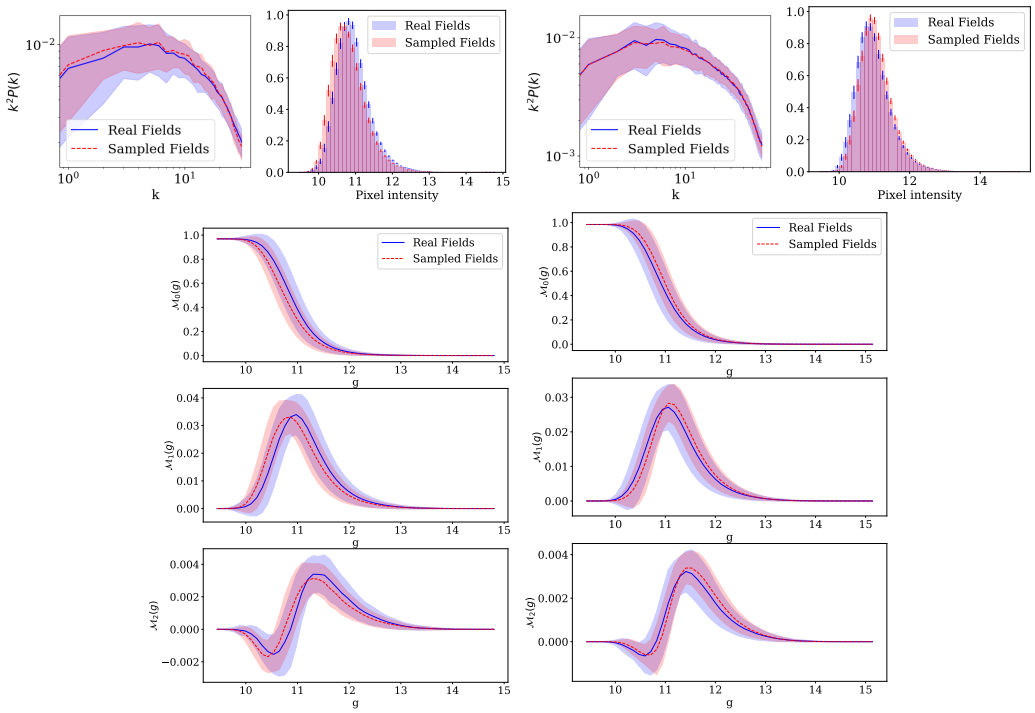

Figure 2: Power spectra, normalized pixel histograms and Minkowski functionals for 100 draws from the real fields and the trained models at 64x64 (left) and 128x128 (right). The envelopes in the power spectra and the Minkowski functionals' panels represent the standard deviation of the value of the statistic in each bin. The height of the bars in the pixel histogram is the mean height of the histogram bin and the error bar is the standard error over all 100 samples.

whose intensity is greater than a value g. In 2D, $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2$ reflect the number of pixels in the excursion set (area), the length of its boundary (perimeter) and the number of holes. We use QuantImPy [Boelens and Tchelepi, 2021] to compute these functionals.

The loss function used to train these models does not directly enforce the summary statistics of the generated images to exactly match the summary statistics of the training distribution. Thus, while the mean of the statistics of the generated samples across checkpoints typically lie within the one standard deviation envelope of the power spectra and Minkowski functionals of the true distribution, convergence in terms of the loss does not imply that the converged models are stationary with respect to the distribution of the summary statistics of the generated images. We thus draw 100 samples from each of the last 10 checkpoints of the trained diffusion models and from the real (train) dataset. Since the variability across checkpoints is similar across all three seeds for the 64x64 models, we train a single model for the 128x128 case and sample from its last 10 checkpoints. For the samples

across all 10 checkpoints, the mean absolute fractional difference in the power spectrum is typically around 10% (std. dev. 5% for 64x64) and 15% (std. dev. 10% for 128x128) and is higher at the lowest and highest $k-$bins, which are most affected by cosmic variance and noise, respectively. For both the 128x128 model and all 3 runs of the 64x64 model, at least one checkpoint with a mean (over all bins) absolute fractional difference of less than 5% could be found. As in Mustafa et al. [2019] and Tamosiunas et al. [2021] that use adherence to summary statistics as a model selection criterion, we identify the checkpoints with the lowest absolute fractional error in the power spectrum and the best agreement with the Minkowski functionals. The statistics plotted in Figure 2 correspond to these models. For the 64x64 case, we plot the statistics for the worst of the three best-case models (one for each seed). While the mean and the spread of the power spectra appear to be largely consistent for these models, the Minkowski functionals are more sensitive to differences between the true and the generated samples. We intend to explore whether these differences and the lack of convergence to a distribution stable with respect to the summary statistics can be mitigated with different design choices, or whether more fundamental changes to the method are required.

4 A Generative Model for Interstellar Medium Fields

Dataset and Training Details The dataset consists of 12482 images from the Schlegel-Finkbeiner-Davis (SFD) [Schlegel et al., 1998] map of interstellar dust extinction inferred from emission at 100 microns. We restrict ourselves to images with extinction $E_{SFD} < 3$. Each image is 128x128 and spans an area of $(6.4^\circ)^2$ on the sky. The validation set consists of images lying in Galactic longitude $0^\circ < l < 42^\circ$ and the test set (held out for future validation) consists of images lying between $200^\circ < l < 240^\circ$. All other images (roughly 69%) belong to the train distribution. The train images are augmented with rotations and flips, yielding 68048 images (8506x8). We apply a minmax transform that scales the minimum and the maximum pixel intensity of *each image* to [-1, 1]. We peg the transform to each image because images of interstellar dust span a larger dynamic range and pixel intensities in two images are much more likely to be dissimilar than for the cosmic web. We train our models for 42k iterations with a learning rate of 6×10^{-5}. All other training details are the same as in Section 3. Four sampled images are plotted in the bottom right panel of Figure 1. For both the generated cosmic web images and the dust images we see that the models are able to capture a rich variety of structure.

Denoising Tests We select a filamentary field from the validation dataset, whose standard deviation corresponds to roughly the 50^{th} percentile of the standard deviation across all images in the validation set. We add $\mathcal{N}(0, \sigma^2)$ noise such that σ is 20% of the mean of the intensity in the image. The noisy input is scaled to [-1, 1] and the timestep at which $\sqrt{1-\bar{\alpha}_t}$ is closest to the corresponding scaled sigma σ_{tr} is identified for each image. We then iteratively sample from $p_\theta(x_{t-1}|x_t)$ from $t = t_{\sigma_{tr}}$ to $t = 0$ to derive the denoised image. As a baseline, we find the corresponding Gaussian filter that would reduce the RMSE in the low-signal portion on the bottom right of the image by the same factor (3.3). Figure 3 plots the denoised images with the model and the baseline. The correlation of the residual with filaments is significantly lower with the diffusion model than with the baseline.

5 Conclusions and Future Work

In this work, we investigated whether DDPMs are able to learn and sample from the image distribution for two astrophysical fields — one from simulations and the other from real interstellar dust maps — in a 'physically meaningful' way, gauged by two different yardsticks. In the case of models trained to generate images of interstellar dust, the models are able to denoise highly filamentary images of interstellar dust while recovering underlying structure better than a smoothing baseline. Approaches such as Régaldo-Saint Blancard et al. [2022] involve learning generative models of interstellar dust using the wavelet statistics of a single image, whereas our description learns a prior using multiple images. The former approach can be useful in cases where limited training data is available while the latter approach is more likely to account for diversity and multiple modes of the image distribution. In the context of dark matter density fields from simulations, we examined three sets of summary statistics of cosmological significance. The ability to generate images with the same statistics is an important first step toward deploying these models as emulators. While DDPMs show promise in terms of our ability to find models that generate samples that are consistent up to 10% in the power

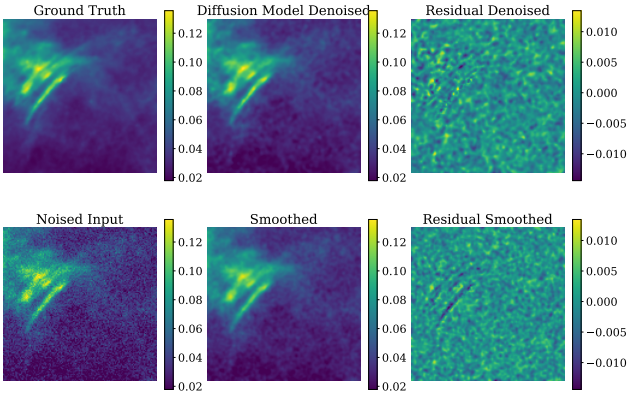

Figure 3: Ground truth, noised input, denoised image and its residual (denoised - truth), smoothed baseline and its residual (smoothed - truth).

spectrum, we intend to work on finding architectures or models with inductive biases that prioritize convergence to distributions that are stationary with respect to these summary statistics.

6 Broader Impact

Several papers have examined and improved the performance of score-based generative models for standard machine learning datasets using standardized metrics such as the Fréchet Inception Distance [Heusel et al., 2017], demonstrated their ability to generate photo-realistic images conditional on text inputs [Ramesh et al., 2022] and identified issues to work on, such as their slower sampling speed, relative to GANs. This raises the question – how can these models help accelerate science? As a first demonstration of the use of these models to generate fields from the interstellar medium, we hope that this paper can motivate both potential use cases for these models in astrophysics and cosmology as well as work on imbuing these models with physical inductive biases that make them more suitable for applications where practitioners care about recovering specific summary statistics. While our focus here has been on astrophysics, the questions we consider and the desire for high-fidelity generative models capable of sampling distributions of images faster than simulations have wider applications across the physical sciences [Kasim et al., 2021] — from the geophysical sciences to high energy physics [Paganini et al., 2018].

7 Acknowledgements

We thank Shuchin Aeron, Carolina Cuesta-Lazaro, Tanveer Karim, Andrew K. Saydjari, and Justina R. Yang for helpful discussions. This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

References

Katrin Heitmann, David Higdon, Martin White, Salman Habib, Brian J. Williams, and Christian Wagner. The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. *Astrophys. J.*, 705:156–174, 2009. doi: 10.1088/0004-637X/705/1/156.

Richard M. Feder, Philippe Berger, and George Stein. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Phys. Rev. D*, 102:103504, Nov 2020. doi: 10.1103/PhysRevD.102.103504. URL `https://link.aps.org/doi/10.1103/PhysRevD.102.103504`.

Drew Jamieson, Yin Li, Renan Alves de Oliveira, Francisco Villaescusa-Navarro, Shirley Ho, and David N Spergel. Field level neural network emulator for cosmological n-body simulations. *arXiv preprint arXiv:2206.04594*, 2022.

Mathieu Remazeilles, Anthony J Banday, Carlo Baccigalupi, S Basak, A Bonaldi, G De Zotti, J Delabrouille, C Dickinson, HK Eriksen, J Errard, et al. Exploring cosmic origins with core: B-mode component separation. *Journal of Cosmology and Astroparticle Physics*, 2018(04):023, 2018.

Gregory M Green, Edward Schlafly, Catherine Zucker, Joshua S Speagle, and Douglas Finkbeiner. A 3d dust map based on gaia, pan-starrs 1, and 2mass. *The Astrophysical Journal*, 887(1):93, 2019.

RH Leike and TA Enßlin. Charting nearby dust clouds using gaia data only. *Astronomy & Astrophysics*, 631:A32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Michael J Smith, James E Geach, Ryan A Jackson, Nikhil Arora, Connor Stone, and Stéphane Courteau. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, 2022.

Benjamin Rémy, Francois Lanusse, Zaccharie Ramzi, Jia Liu, Niall Jeffrey, and Jean-Luc Starck. Probabilistic mapping of dark matter by neural score matching. *arXiv preprint arXiv:2011.08271*, 2020.

Benjamin Remy, Francois Lanusse, Niall Jeffrey, Jean-Luc Starck, Ken Osato, and Tim Schrabback. Probabilistic mass mapping with neural score estimation. *arXiv preprint arXiv:2201.05561*, 2022.

Francisco Villaescusa-Navarro, Shy Genel, Daniel Angles-Alcazar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, 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.

Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N Spergel, Rachel S Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, et al. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Niels Rogge and Kashif Rasul. The annotated diffusion model, 2022. URL `https://huggingface.co/blog/annotated-diffusion`.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

Jens Schmalzing, Martin Kerscher, and Thomas Buchert. Minkowski functionals in cosmology. *arXiv preprint astro-ph/9508154*, 1995.

Jens Schmalzing and Krzysztof M Górski. Minkowski functionals used in the morphological analysis of cosmic microwave background anisotropy maps. *Monthly Notices of the Royal Astronomical Society*, 297(2):355–365, 1998.

Andrius Tamosiunas, Hans A Winther, Kazuya Koyama, David J Bacon, Robert C Nichol, and Ben Mawdsley. Investigating cosmological gan emulators using latent space interpolation. *Monthly Notices of the Royal Astronomical Society*, 506(2):3049–3067, 2021.

Bruno Régaldo-Saint Blancard, Erwan Allys, Constant Auclair, François Boulanger, Michael Eickenberg, François Levrier, Léo Vacher, and Sixin Zhang. Generative models of multi-channel data from a single example–application to dust emission. *arXiv e-prints*, pages arXiv–2208, 2022.

Arnout MP Boelens and Hamdi A Tchelepi. Quantimpy: Minkowski functionals and functions with python. *SoftwareX*, 16:100823, 2021.

Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Zarija Lukić, Rami Al-Rfou, and Jan M Kratochvil. Cosmogan: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, 6(1):1–13, 2019.

David J Schlegel, Douglas P Finkbeiner, and Marc Davis. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *The Astrophysical Journal*, 500(2):525, 1998.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

MF Kasim, D Watson-Parris, L Deaconu, S Oliver, P Hatfield, DH Froula, G Gregori, M Jarvis, S Khatiwala, J Korenaga, et al. Building high accuracy emulators for scientific simulations with deep neural architecture search. *Machine Learning: Science and Technology*, 3(1):015013, 2021.

Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Accelerating science with generative adversarial networks: an application to 3d particle showers in multilayer calorimeters. *Physical review letters*, 120(4):042003, 2018.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

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] One of the key limitations is the observation that the cosmic web diffusion models fail to converge to a distribution that is stationary with respect to the summary statistics we examine. We discuss this in Section 3.

(c) Did you discuss any potential negative societal impacts of your work? [N/A] While this class of generative models can be applied in a diversity of contexts, our application here does not entail any negative societal impacts.

(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)? [Yes] The datasets are both available in the public domain. The code can be made available via Github.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We discuss data splits, scaling, and other design choices in the 'Training Details' paragraphs in Sections 3 and 4, and Appendix A.1.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The variability we observed across different seeds was less than the variability across different checkpoints of the same run. We quantify the latter and discuss this in Section 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 Appendix A.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 cite them in Section 3: Training Details and Summary Statistics.

(b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

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

(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]

A Appendix

A.1 Experiments and Compute Time

The 64x64 model took 1.25h to train for 60k iterations while the 128x128 models took 13.5 hours to train for the same number of iterations. We used an NVIDIA A100 for most experiments and runs. We ran most experiments on the 64x64 model, since it took 1.25 hours to train for 60k iterations. The cumulative compute time spent on experiments was around 5 days. Sampling 10 images from the trained diffusion models took ~ 40 seconds for the 64x64 model and ~ 110 seconds for the 128x128 model on the A100. We gauged whether other hyperparameter choices were better by plotting the summary statistics of samples from the generated model (as in Figure 2).

- We tried the cosine learning schedule proposed in [Nichol and Dhariwal, 2021] and found the linear schedule to work better, for the number of iterations we experimented with.
- We chose a lower learning rate for the dust images since we found this to be more stable.
- We also tried learning rates of $[6 \times 10^{-5}, 2 \times 10^{-4}, 1 \times 10^{-3}]$ for the 64x64 model and did not find any of the other learning rates to improve performance.
- We examined the summary statistics for samples from the 64x64 runs described in Section 3 for the last 15 checkpoints and did not find significant differences in the quality of the summary statistics.