Sparse 3D Images: Point Cloud or Image methods?

Fernando Torales Acosta
ftoralesacosta@lbl.gov
Physics Division, Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA

Vinicius Mikuni
National Energy Research Scientific Computing Center
Berkeley Lab, Berkeley, CA 94720, USA

Benjamin Nachman
Physics Division, Lawrence Berkeley National Laboratory
Berkeley Institute for Data Science, University of California
Berkeley, CA 94720, USA

Miguel Arratia
Department of Physics and Astronomy, University of California
Riverside, CA 92521, USA
Thomas Jefferson National Accelerator Facility, Newport News
Virginia 23606, USA

Bishnu Karki
Department of Physics and Astronomy, University of California
Riverside, CA 92521, USA

Ryan Milton
Department of Physics and Astronomy, University of California
Riverside, CA 92521, USA

Piyush Karande
Computational Engineering Division, Lawrence Livermore National
Laboratory, Livermore CA 94550

Aaron Angerami
Nuclear and Chemical Science Division, Lawrence Livermore National
Laboratory, Livermore, CA 94550

Abstract

Score based generative models are a new class of generative models that have been shown to accurately generate high dimensional datasets. Recent advances in generative models have used images with 3D voxels to represent and model complex detector data. Point clouds, however, are likely a more natural representation for many of these data sets, particularly in calorimeters with high granularity that produce very sparse images. Point clouds preserve all of the information of the orig-

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

inal simulation, more naturally deal with sparse datasets, and can be implemented with more compact models and datasets. In this work, two state-of-the-art score based models are trained on the same set of calorimeter simulation and directly compared.

1 Introduction

Score based diffusion models have had incredible success generating high fidelity images. Recently, these techniques have been adapted for use in collider physics where images are often used represent the data measured in detectors. However, these data are unlike natural images in a number of ways, most notably in their sparsity. As such, image-based approaches pioneered in industry may not be the most effective for particle interactions.

We study this problem in the context of a specific example: Calorimeters. Since most cells in a calorimeter image are empty, a more natural representation of these data may be a point cloud. Point clouds are a set of attributes assigned to locations in space. In the calorimeter case, the attribute is energy and the location is the cell coordinates. A calorimeter point cloud would require far fewer numbers to specify than an image representation, since only cells with non-zero energy would be recorded. The main challenges for point cloud models in contrast to image-based approaches is that they must cope with variable-length outputs that respect permutation invariance. With a lag compared to image-based approaches, point cloud generative models for particle/nuclear physics applications have seen a rapid development in recent years [1, 2, 3, 4, 5, 6]. However, until recently, these models have never been applied to calorimeter simulations.

In this paper, we explore point cloud generative models applied directly to cell-level information. In other words, we take calorimeter images and compare state-of-the-art generative models that represent the same inputs as either images or (zero-suppressed) point clouds. As a case study, the two representations are compared using simulations of a high-granularity hadronic calorimeter, similar to the design planned for the ePIC detector at the future Electron-Ion Collider [7, 8, 9].

2 Deep Learning Models

Diffusion models are a class of generative neural networks that allow for stable training paired with high flexibility in the model design. Data is slowly perturbed over time using a time parameter $t \in \mathbb{R}$ that determines the perturbation level. The task of the neural network is to approximate the gradients of the log probability of the data, $\log p_{\mathrm{data}}(\mathbf{x})$, or the score function $\nabla_{\mathbf{x}} \log p_{\mathrm{data}}(\mathbf{x}) \in \mathbb{R}^D$, based on data observations $\mathbf{x} \in \mathbb{R}^D$ in the D-dimensional space. This can be approximated by a denoising score-matching strategy [10]. In the implementation used in this paper, data observations $\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})$ are perturbed using a Gaussian kernel q such that $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})$, with time-dependent parameters α and σ determining the strength of the Gaussian perturbation to be applied. In the variance-preserving setting of diffusion processes, $\sigma_t^2 = 1 - \alpha_t^2$. For the time-dependence, a cosine schedule is used such that $\alpha_t = \cos(0.5\pi t)$. The loss function to be minimized is implemented using a *velocity* parameterization:

$$\mathcal{L}_\theta = \mathbb{E}_{\epsilon,t} \left\| \mathbf{v}_t - \hat{\mathbf{v}}_{t,\theta} \right\|^2, \tag{1}$$

where the time-dependent network output with trainable parameters θ, $\hat{\mathbf{v}}_{t,\theta}$, is compared with the velocity of the perturbed data at time t, $\mathbf{v}_t \equiv \alpha_t \epsilon - \sigma_t \mathbf{x}$, with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The score function is then identified as

$$\nabla_x \log \hat{p}_\theta(\mathbf{x}_t) = -\mathbf{x}_t - \frac{\alpha_t}{\sigma_t} \hat{\mathbf{v}}_{t,\theta}(\mathbf{x}_t). \tag{2}$$

The data generation from the trained diffusion models is implemented using the DDIM sampler proposed in Ref. [11] that can be interpreted as an integration rule [12] with update rule specified by:

$$\mathbf{x}_s = \alpha_s \hat{\mathbf{x}}_\theta(\mathbf{x}_t) + \sigma_s \frac{\mathbf{x}_t - \alpha_t \hat{\mathbf{x}}_\theta(\mathbf{x}_t)}{\sigma_t}. \tag{3}$$

For a fair comparison, all diffusion models are trained using the same score-matching strategy and fixed number of 512 time steps during sampling.

The fast point cloud diffusion model (FPCD) follows [5], where a permutation equivariant estimation of the score function is obtained by the combination of a DEEPSETS [13] architecture with attention layers [14]. During the point cloud simulation, two models are also defined: one that learns the number of non-empty cells, conditioned on the initial energy of the incoming particle, and one model that learns the score function of the normalized point cloud, also conditioned on the momentum of the particle to be simulated and the number of hits to be generated.

The model trained on the image dataset (CALOSCORE) is adapted from [15] with a few modifications. Compared to the original implementation, the calorimeter simulation task is now broken down in two diffusion models: one that learns only the energy deposits in each layer of the calorimeter, conditioned only on the initial energy of the particle to be simulated, and one model that learns to generate normalized voxels per layer, conditioned on the energy deposition in each layer and the initial energy of the particle to be simulated. Additionally, the original U-NET [16] model is combined with attention layers.

3 Detector and Data Descriptions

The DD4HEP framework [17] is used to run GEANT simulations of a high-granularity iron-scintillator calorimeter based on previous designs for use in colliders [18, 8]). The sampling structure comprises of 55 $10 \times 10 \times 0.3$ cm scintillator tiles sandwiched between 2.0 cm thick steel absorber plates. The calorimeter is 1.2 meters long, with it's front set at z=3.8 m. 1.7 million events of single π^+ particles are generated with momentum $1.0 < P < 125$ GeV/c in rings within the detector (see Figure 1).

Dataset 1 and Dataset 2 used in training share the same parent GEANT simulation, such that the fast point-cloud diffusion model and the image model are trained on different representations of the same set of calorimeter showers.

Dataset 1 is stored as zero-suppressed point-cloud representation. The GEANT data is stored in files containing two datasets, *clusters* and *cells*. The cluster dataset contains the P_{Gen} of the incident pion, as well as the number of hits in the calorimeter. The cell data is comprised of a constant number of 200 cells per event. Each cell contains energy, x, y, and z coordinate values. Empty cells, or cells with deposited energy below the threshold are 0-masked and ignored during training.

Dataset 2 is created by converting the point cloud dataset into an image format. Images at the full granularity of the detector would be unrealistic for an real-world detector. For example, the detector in [8], contains only 7 readout channels along the z-dimension, no where near 55. Additionally, images at full resolution would result in an unmanageably large datasets (see Table 1), and would represent the largest calorimeter image training ever done. The calorimeter cells were therefore clustered into groups of 5 along each axis of the detector to create voxels, where $5 \times 5 \times 5$ cells = 1 voxel. Energy in each of the cells making up the voxel were summed and assignd to the final voxel's total energy. The final image format consists of 11×11 voxels. A hit in the voxelized dataset is defined as any voxel with energy deposition above threshold.

For the final comparison, generated samples from the point cloud model are voxelized using the same method for Dataset 2. All comparisons are in this image format, at the same resolution of $11 \times 11 \times 11$ voxels per image.

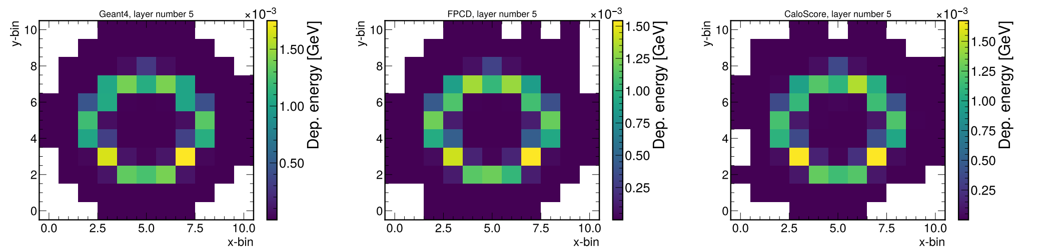
4 Results

A variety of distributions are used to evaluate the quality of the generated images. For all comparisons, the Earth mover's distance (EMD) [19], also known as the 1-Wasserstein distance [20], between generated distributions and GEANT) distributions is calculated. The EMD score a distance-like measure of the dissimilarity between two distributions. All EMD scores in Figures 2 are calculated on the final voxelized distributions.

Table 1 shows the model size, size of each dataset, and time to generate 100k calorimeter showers. The point cloud model is smaller by a factor of 4 compared to the image based model, and samples events 3 times faster. Lastly, the point cloud dataset requires over 100 times less disk space than the image format at full granularity. The AUC is obtained from a classifier trained to distinguish the samples of both models only in the voxelized image format. Both models have very good AUC,

Model	# Parameters	Disk Size (Full)	Sample Time	AUC
Image	2,572,161	1016MB (62GB)	8036.19s	0.673
Point Cloud	620,678	509 MB	2631.41s	0.726

Table 1: Comparison of model size, size of data representation on disk, generation time, and AUC of the same classifier trained to distinguish between the model and the original GEANT showers. All comparisons are done for 100k calorimeter showers. All results in the image row were obtained with $11 \times 11 \times 11$ voxel images. Disk size of the image dataset at full granularity is shown in parenthesis.



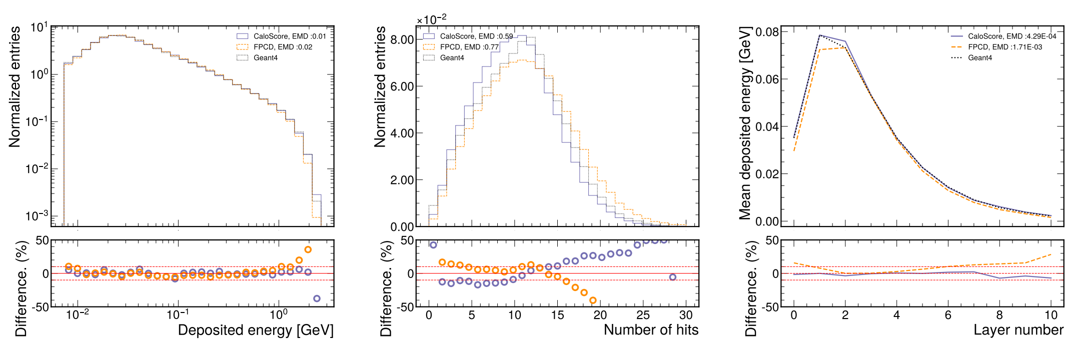
(a) (b) (c)

Figure 1: The 2-dimensional distribution of the mean deposited energy in 5th voxelized layer of the calorimeter. The first column is the original Geant simulation. The second column is the fast point-cloud based diffusion model (FPCD), and the 3rd column is the image-based model (CALOSCORE).

reasonably close to 0.5 where the classifier would be unable to distinguish between original simulation and generated samples.

Figure 1 shows a qualitative assessment of the models using the 2-dimensional distribution of the average energy deposition in a specific z-layer. All voxels with an expected energy deposition above 0 are populated in both the image and point cloud based models, with very few additional hits. The qualitative similarities in each image in Fig 1 indicate that models reproduce the various showers from the training dataset well.

Figure 2(a) compares the total energy deposited in the calorimeter. Both models are in good agreement with GEANT at small deposited energies, deviating no more than 10%. At the highest deposited



(a) (b) (c)

Figure 2: Comparison of point-cloud model (orange), image based model (grey-blue) and GEANT (black). Sum of all voxel energies is shown in (a), the total number of voxel hits is shown in (b), and the mean energy per z-layer is shown in (c). The dashed red lines in the bottom panels represent the 10% deviation interval of the generated samples from the original GEANT simulation. The earth mover's distance (EMD) between each distribution and the GEANT distribution is also shown.

energies, however, both diffusion models begin to fall away from GEANT, with the point-cloud model generating less energy, and the image based model generating slightly more energy than GEANT. Events in this region are rare, however, and statistical fluctuations begin to dominate even GEANT. Figure 2(b) comparing total number of calorimeter hits. At low number of hits, both models show good agreement with GEANT, with deviations slightly above 10%. At 15 or more hits, both models begin to deviate well past 10%, with the point cloud model oversampling the number of hits, and the image based model undersampling the number of hits.

Figure 2(c) shows the average deposited energy along z. The mean deposited energy in z-coordinates in panel (c) show both models in very good agreement with the original GEANT predictions. The z distribution shows the point cloud samples are systematically lower than the original GEANT distributions. This indicates the point cloud model would benefit from learning the energy per layer directly, as is done in the image model described Sec. 2.

5 Conclusion and Outlook

This work makes the first direct comparison between two score based generative models using either images or point clouds as representations of the same calorimeter training data. Both models perform well for most distributions, with very similar AUCs, but the image-based diffusion model invariably has a lower EMD in each comparison to GEANT. Overall, the performance of the point-cloud diffusion model is very close to the image model. This is despite the point cloud model being disadvantaged: the second model in it's architecture is not conditioned on the energy per layer, unlike the image based model. At the same time, the point cloud model offers several advantages over the image model: Vastly smaller dataset size - about 100x smaller at full resolution, individual cell hits do not need to be summed in a voxelization procedure, resulting in information loss, and 3x faster generation times.

This work establishes a benchmark for future research on generative models, offering valuable insights into the challenges of modeling hadronic showers in highly granular calorimeters using image-based techniques, while also exploring the potential of point-cloud methods. The current advantages of point clouds, in combination with improvements to close the remaining performance gap described earlier, will likely make point cloud based models a clear choice for highly granular calorimeters. This work should serve as a reference for studies utilizing future calorimeters based on the CALICE design, including those intended for use in CMS at the LHC and ePIC at the EIC.

References

[1] Raghav Kansal, Javier Duarte, Hao Su, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, Jean-Roch Vlimant, and Dimitrios Gunopulos. Particle Cloud Generation with Message Passing Generative Adversarial Networks. 6 2021.

[2] Erik Buhmann, Gregor Kasieczka, and Jesse Thaler. EPiC-GAN: Equivariant Point Cloud Generation for Particle Jets. 1 2023.

[3] Benno Käch, Dirk Krücker, Isabell Melzer-Pellmann, Moritz Scham, Simon Schnake, and Alexi Verney-Provatas. JetFlow: Generating Jets with Conditioned and Mass Constrained Normalising Flows. 11 2022.

[4] Rob Verheyen. Event Generation and Density Estimation with Surjective Normalizing Flows. 5 2022.

[5] Vinicius Mikuni, Benjamin Nachman, and Mariel Pettee. Fast Point Cloud Generation with Diffusion Models in High Energy Physics. 4 2023.

[6] Matthew Leigh, Debajyoti Sengupta, Guillaume Quétant, John Andrew Raine, Knut Zoch, and Tobias Golling. PC-JeDi: Diffusion for Particle Cloud Generation in High Energy Physics. 3 2023.

[7] R. Abdul Khalek et al. Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report. *Nucl. Phys. A*, 1026:122447, 2022.

[8] F. Bock et al. Design and Simulated Performance of Calorimetry Systems for the ECCE Detector at the Electron Ion Collider. 7 2022.

[9] Miguel Arratia et al. A high-granularity calorimeter insert based on SiPM-on-tile technology at the future Electron-Ion Collider. *Nucl. Instrum. Meth. A*, 1047:167866, 2023.

[10] Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 07 2011.

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020.

[12] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[13] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets. *arXiv e-prints*, page arXiv:1703.06114, March 2017.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[15] Vinicius Mikuni and Benjamin Nachman. Score-based generative models for calorimeter shower simulation. *Phys. Rev. D*, 106(9):092009, 2022.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[17] Markus Frank, F. Gaede, C. Grefe, and P. Mato. DD4hep: A Detector Description Toolkit for High Energy Physics Experiments. *J. Phys. Conf. Ser.*, 513:022010, 2014.

[18] Design, Construction and Commissioning of a Technological Prototype of a Highly Granular SiPM-on-tile Scintillator-Steel Hadronic Calorimeter. 9 2022.

[19] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998.

[20] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256 vol.2, 2001.