A Survey on Evaluation Metrics for Synthetic Material Micro-Structure Images from Generative Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

The evaluation of synthetic micro-structure images is an emerging problem as machine learning and materials science research have evolved together. Typical state of the art methods in evaluating synthetic images from generative models have relied on the Fréchet Inception Distance. However, this and other similar methods, are limited in the materials domain due to the unique features that characterize physically accurate micro-structures and limited dataset sizes. In this study we evaluate a variety of methods on scanning electron microscope (SEM) images of graphene-reinforced polyurethane foams. The primary objective of this paper is to report our findings with regards to the shortcomings of existing methods so as to encourage the machine learning community to consider enhancements in metrics for assessing quality of synthetic images in the material science domain.

1 Introduction

The adoption of artificial intelligence techniques within the material science community has been critical in enabling high throughput research through rapid characterization and discovery of materials. In particular, various forms of generative models have been proposed to accelerate the design of new materials [1, 2, 3, 4]. Nonetheless, the applications of traditional image based deep learning models in this domain are constrained, among other design choices, by the ability to evaluate trained models and synthetic image quality for physically accurate characteristics.

Generative networks are unsupervised learning methods that aim to learn the regularities and patterns in input data such that the model can be used to generate new examples that could have plausibly been drawn from the original dataset [5]. These models are traditionally evaluated using the Fréchet Inception Distance (FID) score, which leverages the InceptionV3 pretrained model on the ImageNet dataset. Several studies have shown features from this model to be *vision-relevant* [6]. By comparing the mean and variance of these features between synthetic and ground truth datasets, the FID score evaluates both the image quality and diversity of the synthetic images. Though plenty of works have highlighted various pitfalls of the FID score and suggested variations, it is still regarded as the standard for model evaluation. In this study, we explore the drawbacks of existing methods when working with datasets in the materials domain. In particular, we outline two main causes of concern:

1. Dataset Size: Materials scientists have worked with small & noisy datasets with great success in uncovering process-structure-property relationships [3]. However, in small data regimes, the FID score promotes over-fitting, as the distributions are minimized when matching the smaller variance of our ground truth data. Refer to Appendix A for an experimental analysis.
2. Feature Relevance: Unlike well-studied datasets (e.g. ImageNet), many of the features of a realistic micro-structure image cannot be recognized by the human eye, thus limiting the applications of a model pretrained on ImageNet as the base of the metric. Beyond

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

image quality and diversity, in the material science domain, we aim to generate valid micro-structures which can be synthesized in an experimental setting.

2 Methods

Unlike traditional machine learning metrics, the evaluation of generative models and synthetic images have been an open subject of study for many years. Many works [7] have summarized the relationships between various metrics and outlined corresponding limitations. However, much of this work has been done on large datasets of simple images where population surveys can verify results. On the contrary, verification of synthetic images in the material science domain is limited to costly experimental trials [8]. Subsequently, this limits the use of metrics using human evaluation such as those provided by Amazon Mechanical Turk [9] and HYPE [10]. Studies in the medical domain [11] have shown how standard metrics can give misleading results when applied to more complex images.

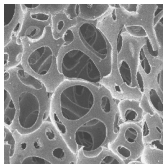
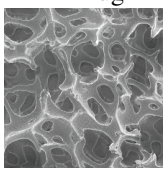
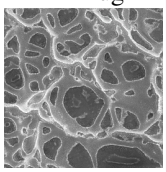
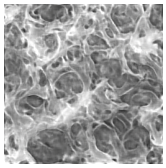
For our study, we use a dataset of SEM images of graphene-reinforced polyurethane foams (PUF) with seven unique additive amounts. The images were acquired under an SE2 secondary electron detector at fixed beam energy and working distance with 25x magnification (approx. 10.3 pix/micron). The brightness and contrast levels were held constant across all images and samples. SEM image data provides researchers with additional information about materials (e.g. connectivity of cells, cell sizes) that cannot be inferred by numerical data measurements.

There are three key properties we hope to evaluate in our synthetic micro-structure image datasets:

1. Image Quality: Are the images of high fidelity?
2. Image Diversity: Are new images (outside the training data) being formed?
3. Physically Accurate: Are the synthetic structures experimentally feasible/valid characteristics of materials (e.g. complete cell walls, inter-connected cell structures)?

Several generative models were trained using our ground truth dataset. In this study, we focus on the outputs from three particular models, summarized in Table 1 and detailed further in Appendix B. Resources were not spent to find ideal model parameters/architecture, as the focus was on comparing relative outputs of different generative models. Refer to Appendix C for a detailed visualization of the image quality and diversity in each dataset.

Table 1: Summary and sample images of the 4 datasets. State of the art methods [12, 13] were used.

	Ground Truth Dataset	Dataset A	Dataset B	Dataset C
Num. Images	2800	1000	1000	1000
Data Generation Method	Experimental	StyleGAN2 + DiffAug	StyleGAN2 + DiffAug	Variational Autoencoder
Sample Image				

Many alternatives of the FID score have been proposed and were considered during our study. These include the *Inception Score (IS)* [14] and the *Kernal Inception Distance (KID)* [15] to address concerns with the small training dataset. To differentiate between the image fidelity and diversity we also considered *Precision & Recall (P&R)* [16, 17] and *Density & Coverage* [18].

To mitigate the feature relevance concern with the FID score, various other feature extractors have been proposed, from randomly initialized weights [18] to task-specific tuned models [11]. We leveraged transfer learning methods to fine-tune an InceptionV3 model to perform a simple classification task of matching an input PUF image to one of the seven unique additive combinations. The model was trained for 10 epochs and achieved a >97% accuracy on both the training (2000 images) and test (800 images) sets. Similar to the FID, the output of the final pooling layer was extracted as the 2048-dim feature vector for calculations. We define this as the *FID Modified* score.

Table 2: Standard metrics for each of the synthetic datasets relative to the ground truth. Average of 10 trials shown.

	Dataset		
	A	B	C
FID (↓)	**60** ± **0.4**	65 ± 0.3	239 ± 1.7
FID Modified (↓)	**97** ± **1.0**	208 ± 1.7	438 ± 2.8
IS (↑)	**1.3** ± **0.0**	1.2 ± 0.0	1.2 ± 0.0
KID (↓)	**0.1** ± **0.0**	0.1 ± 0.0	0.5 ± 0.0
Precision (↑)	**0.5** ± **0.0**	0.3 ± 0.0	0.0 ± 0.0
Recall (↑)	**0.0** ± **0.0**	0.0 ± 0.0	0.0 ± 0.0
Density (↑)	**0.2** ± **0.0**	0.1 ± 0.0	0.0 ± 0.0
Coverage (↑)	**0.04** ± **0.0**	0.02 ± 0.0	0.00 ± 0.0

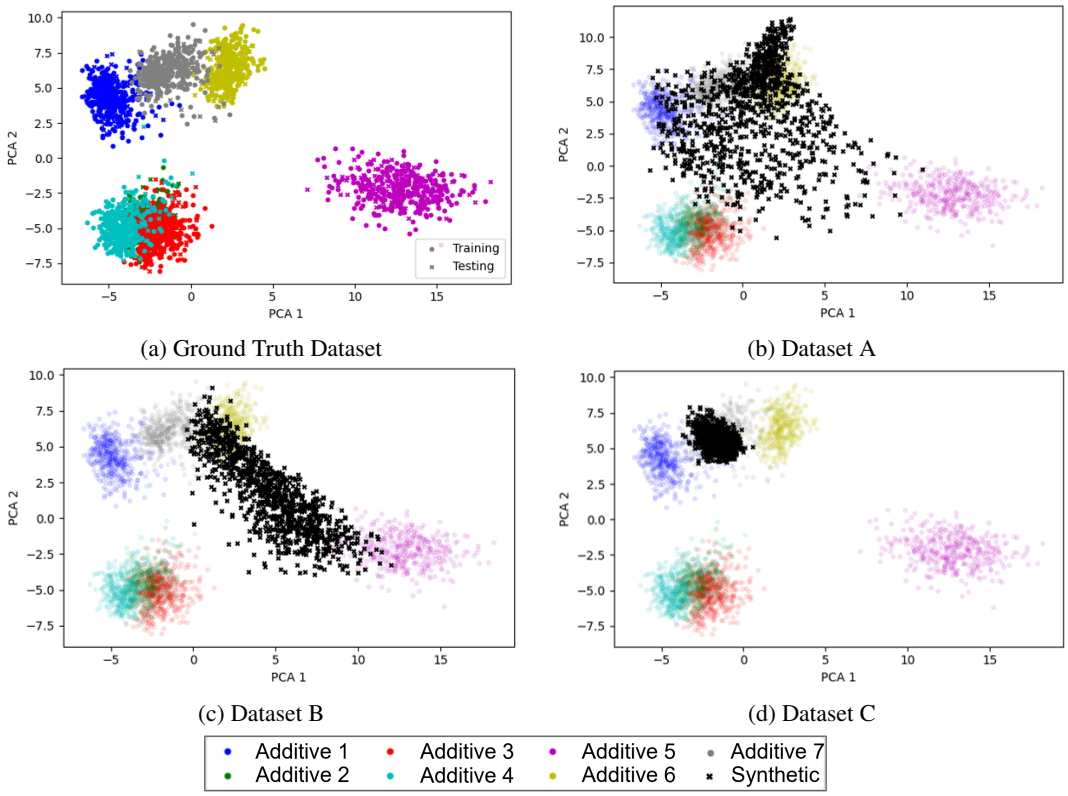
Table 3: Visual inspection of each dataset. Complete analysis is shown in Appendix C. Note: $\uparrow good, -neutral, \downarrow bad$

	Dataset		
	A	B	C
Image Fidelity	↑	↑	↓
Image Diversity	-	↓	↑
Physically Accurate	↓	-	-

We additionally considered a wide array of image processing methods, from pixel-wise histograms to Image Quality Assessment (IQA) metrics. These included BRISQUE [19], LPIPS [20, 21], Manipulative Precision Metric [22], Optimal Transport [23], along with others. A key advantage of IQA methods is the ability to score individual images on image naturalness/distortion based on deviation from natural image distributions. However, experimental results showed these metrics were not suited for our application, with all our synthetic images scoring better than ground truth images.

3 Results and Discussion

Metric results from the FID score and alternatives are summarized in Table 2. We can infer Dataset A is consistently slightly higher ranked than Dataset B, and both Datasets A & B are significantly better than Dataset C. Table 3 provides a visual interpretation of each of the datasets for each of our key scoring properties we hope to evaluate. We observe the *vision-relevant* features (scoring Dataset A the best) don't correlate well with our target of *physically-accurate*. Similarly, Figure 3 highlights



(a) Ground Truth Dataset (b) Dataset A (c) Dataset B (d) Dataset C

Figure 1: 2D PCA of feature vectors for each image using fine-tuned InceptionV3 model.

the severe over-fitting in Dataset B resulting in lack of any diversity - yet this dataset is still ranked similar to Dataset A. It is clear there are flaws in the existing methods.

From Table 2 we see our FID Modified score provides the greatest differentiation between datasets, with Dataset A being the best. However, there is an evident lack of value for the greater diversity seen in Dataset C. To further investigate, we extracted the outputs from the final pooling layer of the InceptionV3 model (finetuned on our materials dataset) to obtain 2048-dim feature vectors for each image. Using PCA, we were able to visualize these images in 2D, as seen in Figure 1. Note, for the ground truth dataset, we see the expected clusters for the seven additive types, as this is what the model was trained to distinguish. The black data points represent the synthetic images distributed on the same two principal components.

The results from Figure 1 are quantified in Table 4 as the euclidean distance between various sets of image feature vectors. For each of the four datasets, we calculated the *same* and *different* distances with respective to each additive type. The *same* distance represents the distance between each point in Dataset X classified as Additive Y (using our fine-tuned InceptionV3 model) with every point in the Ground Truth Dataset of Additive Y. Conversely, the *different* distance compares every point in Dataset X classified as Additive Y with every point in the Ground Truth Dataset not of Additive Y. Note these distances are calculated using the complete 2048-dim feature vectors.

Similar to the 2D PCA in Figure 1, Table 4 indicates there is a clear distinction between each category. For the Ground Truth Dataset, distances within the same category average 12.1 while distances to other category labels average 20.8. An ideal synthetic dataset should provide *same* distance values between these endpoints - indicating both realistic images and greater diversity than the original categories. Results from Datasets A-C support this.

Table 4: Distances between 2048-dim feature vectors of dataset images and corresponding additive types from training dataset. Any values with "-" indicate no synthetic images were generated with that category prediction for that dataset.

	Ground Truth Dataset		Dataset A		Dataset B		Dataset C	
	Same	Different	Same	Different	Same	Different	Same	Different
Additive 1	12.5	21.0	15.7	20.6	-	-	-	-
Additive 2	11.9	20.2	15.2	18.5	-	-	-	-
Additive 3	12.4	20.4	-	-	15.2	18.0	-	-
Additive 4	12.2	20.4	15.3	19.3	-	-	-	-
Additive 5	13.5	22.7	17.6	19.7	17.5	20.3	-	-
Additive 6	10.3	20.5	14.8	20.9	15.6	19.0	-	-
Additive 7	12.2	20.5	15.4	20.7	16.1	20.4	21.2	25.1

A key limitation of our method is the model fine-tuned on our ground truth dataset seems unable to distinguish image diversity when image quality is low. Figure 1 indicates Dataset C is the least diverse, while Table 4 indicates the diversity in the Additive 7 class is much higher than any other class. Additionally, each of these metrics lacks the ability to distinguish physically accurate features when scoring. Dataset C, though of low image quality, exhibits diverse sets of cell structures that still provide valuable information in the materials discovery platforms.

4 Conclusion and Future Works

We have identified and evaluated several limitations of current state of the art methods in evaluating generative models from non-traditional problem domains. In particular, we highlight the limitations from using small dataset sizes, images with features different from those in ImageNet, and the need to score based on physically accurate micro-structure formations that can be synthesized in the lab. As we progress this work, we hope to leverage additional deep learning tools (e.g. Dragonfly [24, 25]) to quantify image features such as number of cells, cell wall width and size of cells (sample visualization in Appendix D) and evaluate feasibility of synthesized images. Similarly, we plan to leverage Physics Informed Neural Networks (PINNs) to embed the knowledge of physical laws that govern our materials dataset during the learning process of the generative models [26, 27].

References

[1] Addis S Fuhr and Bobby G Sumpter. "Deep Generative Models for Materials Discovery and Machine Learning-Accelerated Innovation". In: *Frontiers in Materials* (2022), p. 182.

[2] Litao Chen et al. "Generative models for inverse design of inorganic solid materials". In: *Journal of Materials Informatics* 1.1 (2021), p. 4.

[3] Alemayehu Admasu et al. "Deep Learning Image-based Design of Graphene-reinforced Polyurethane Foams". In: (2022).

[4] Rahul Singh et al. "Physics-aware deep generative models for creating synthetic microstructures". In: *arXiv preprint arXiv:1811.09669* (2018).

[5] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[6] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[7] Ali Borji. "Pros and cons of GAN evaluation measures: New developments". In: *Computer Vision and Image Understanding* 215 (2022), p. 103329.

[8] Brian Gallagher et al. "Predicting compressive strength of consolidated molecular solids using computer vision and deep learning". In: *Materials & Design* 190 (2020), p. 108541.

[9] Andreas Rossler et al. "Faceforensics++: Learning to detect manipulated facial images". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1–11.

[10] Sharon Zhou et al. "Hype: A benchmark for human eye perceptual evaluation of generative models". In: *Advances in neural information processing systems* 32 (2019).

[11] Lorenzo Tronchin et al. "Evaluating GANs in Medical Imaging". In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Springer, 2021, pp. 112–121.

[12] Tero Karras et al. "Training Generative Adversarial Networks with Limited Data". In: *Proc. NeurIPS*. 2020.

[13] Shengyu Zhao et al. "Differentiable augmentation for data-efficient gan training". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7559–7570.

[14] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems* 29 (2016).

[15] Mikołaj Bińkowski et al. "Demystifying mmd gans". In: *arXiv preprint arXiv:1801.01401* (2018).

[16] Mehdi SM Sajjadi et al. "Assessing generative models via precision and recall". In: *Advances in neural information processing systems* 31 (2018).

[17] Tuomas Kynkäänniemi et al. "Improved precision and recall metric for assessing generative models". In: *Advances in Neural Information Processing Systems* 32 (2019).

[18] Muhammad Ferjad Naeem et al. "Reliable fidelity and diversity metrics for generative models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7176–7185.

[19] Anish Mittal, Anush K Moorthy, and Alan C Bovik. "Blind/referenceless image spatial quality evaluator". In: *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE. 2011, pp. 723–727.

[20] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[21] Keyan Ding et al. "Comparison of full-reference image quality models for optimization of image processing systems". In: *International Journal of Computer Vision* 129.4 (2021), pp. 1258–1281.

[22] Bowen Li et al. "Manigan: Text-guided image manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7880–7889.

[23] Luis Caicedo Torres, Luiz Manella Pereira, and M Hadi Amini. "A Survey on Optimal Transport for Machine Learning: Theory and Applications". In: *arXiv e-prints* (2021), arXiv–2106.

[24] Ruslana Makovetsky, Nicolas Piché, and Mike Marsh. "Dragonfly as a Platform for Easy Image-based Deep Learning Applications". In: *Microscopy and Microanalysis* 24 (Aug. 2018), pp. 532–533. DOI: 10.1017/S143192761800315X.

[25] Object Research Systems, Comet Group. *Dragonfly Software*. URL: https://www.theobjects.com/dragonfly/index.html.

[26] Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. "Artificial neural networks for solving ordinary and partial differential equations". In: *IEEE transactions on neural networks* 9.5 (1998), pp. 987–1000.

[27] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational physics* 378 (2019), pp. 686–707.

[28] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1501–1510.

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 3
   (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 code and data are proprietary. Sample images of data are shown in Appendix C
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2 and Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 2
   (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)? [N/A] Negligible resources used. As discussed in Section 2, we limited hyperparameter tuning as goal was to generate multiple synthetic datasets of different strengths/weaknesses for comparing evaluation metrics.
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 [12, 13, 25]
   (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? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Data was generated in our own experimental labs
   (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 Datasets Limitations

While AI methods require substantially more data than conventional methods, materials scientists have traditionally worked with small & noisy datasets [3], thanks in part to advancements of generative modeling in small data regimes using methods such as differential augmentation [13] and adaptive instance normalization [28]. Yet, there is minimal discussion around the impact of such small training datasets on model evaluation metrics.

The FID score is the Fréchet distance $d()$ between the Gaussian with mean (m, C) obtained from $p(.)$ and the Gaussian with mean (m_w, C_w) obtained from $p_w()$ given by the following [6]:

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}) \tag{1}$$

From Equation 1, we see the FID score compares both the mean and variance of the training dataset to the respective components of the synthetic images. The FID score is minimized (ideal) when these variables are as similar as possible. Although any number of synthetic images can be generated, the size of the training dataset can become a limitation. Smaller training datasets provide a smaller variance, thus resulting in a lower FID score when our synthetic images have a similar variance, subsequently encouraging over-fitting.

We ran several trials to highlight this difference, as seen in Table 5. The experiments were run using three datasets: CIFAR10, FFHQ* (the thumbnail version), and our internal SEM micro-structure dataset. For each trial, the dataset was randomly (without repetition) split into two groups of varying sizes. These groups were representative of a ground truth (Dataset 1) and a synthetic (Dataset 2) dataset. Note, for the micro-structure trials, Dataset 1 came from our actual ground truth dataset while Dataset 2 was randomly produced using the trained network for Dataset A. As a result, we only see trail results where the ground truth dataset is less than or equal to 1000 images.

Table 5 clearly highlights the impact of small training datasets (Dataset 1), even when both datasets are sampled from the same distribution (CIFAR10, FFHQ*). In particular, we notice that the FID score is consistenly low, no matter the problem domain, when working with large training and synthetic datasets. However, as the dataset sizes decrease, not only does the FID score diverge quickly, there is a lack of consistency between datasets. As a result, it becomes difficult to use the FID score as a standard metric to compare multiple models during training when working in problem domains with limited data. An alternative metric is needed for scoring generative models.

Table 5: The FID score is reliant on using larger dataset sizes in order to consistently minimize the FID score. As the dataset sizes get smaller, the FID score diverges at different rates depending on the particular dataset. Trial E most closely represents the size of datasets we are working with in our problem domain.

	Dataset Size		FID Score		
Trial	Dataset 1	Dataset 2	CIFAR10	FFHQ*	Our SEM Images
A	30,000	30,000	1.8	1.9	-
B	10,000	50,000	3.3	3.4	-
C	10,000	10,000	5.4	5.5	-
D	1000	50,000	33	26	62
E	**1000**	**1000**	**54**	**35**	**64**
F	100	50,000	153	94	76
G	100	100	180	109	83

B Model Architecture/Hyper-parameters

Dataset A and B were both derived from a base StyleGAN2 model [12] with Differential Augmentation [13]. The augumentations applied were color, cutout, and translation. Dataset A was trained for 280 Kimg and Dataset B for 240 Kimg. Both models were randomly initialized with weights before training. Dataset C is derived from a ResNet18 based Variational Autoencoder.

C Image Analysis

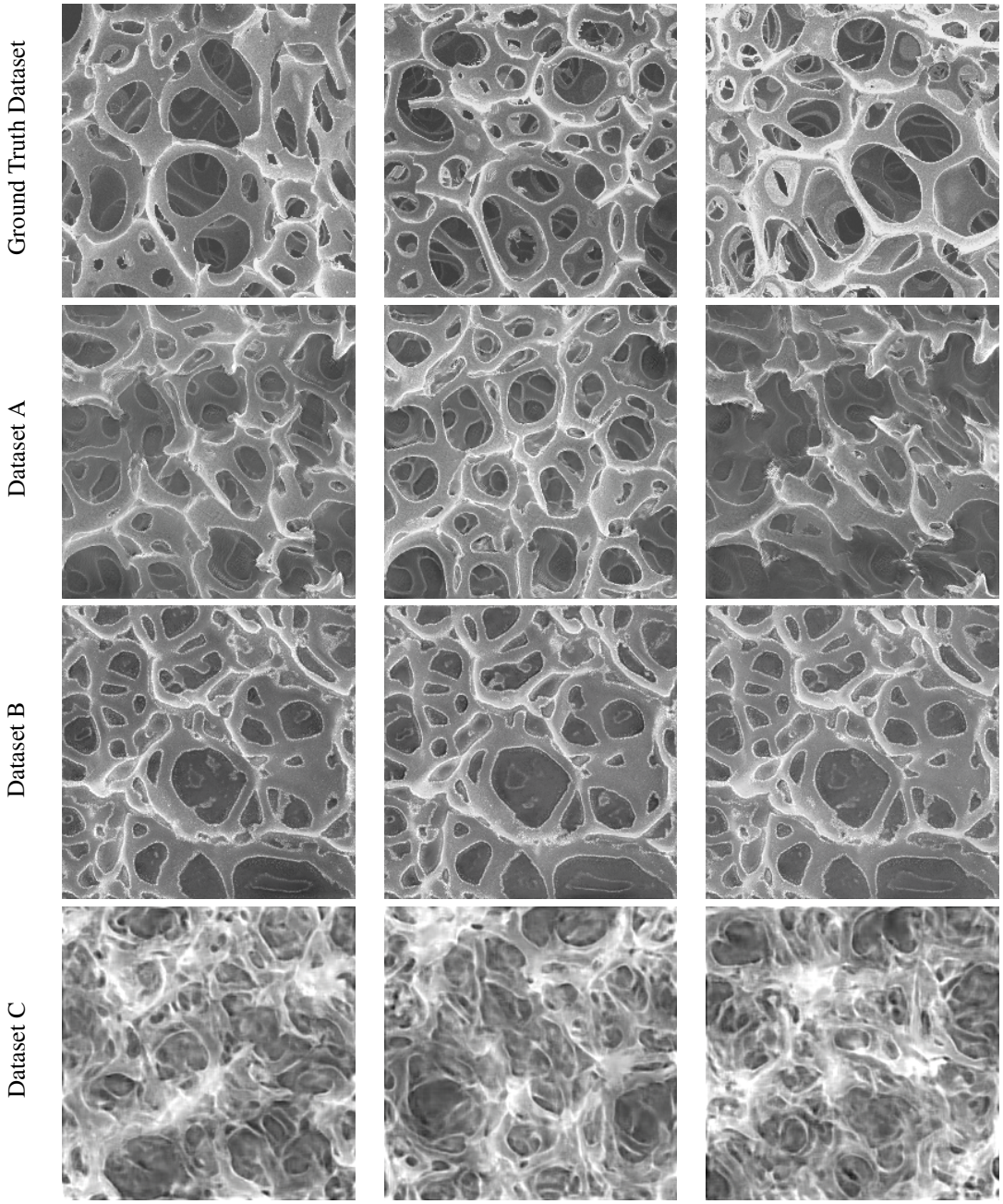


Figure 2: Specific image analysis on each of the 4 datasets we are working with. Note the variation in detail from each of the datasets. For a viewer outside the material science domain, both Dataset A and Dataset B seem to show very realistic images relative to the Ground Truth Dataset. On the other hand, it is clear the images in Dataset C are not of high quality and are synthetically generated.

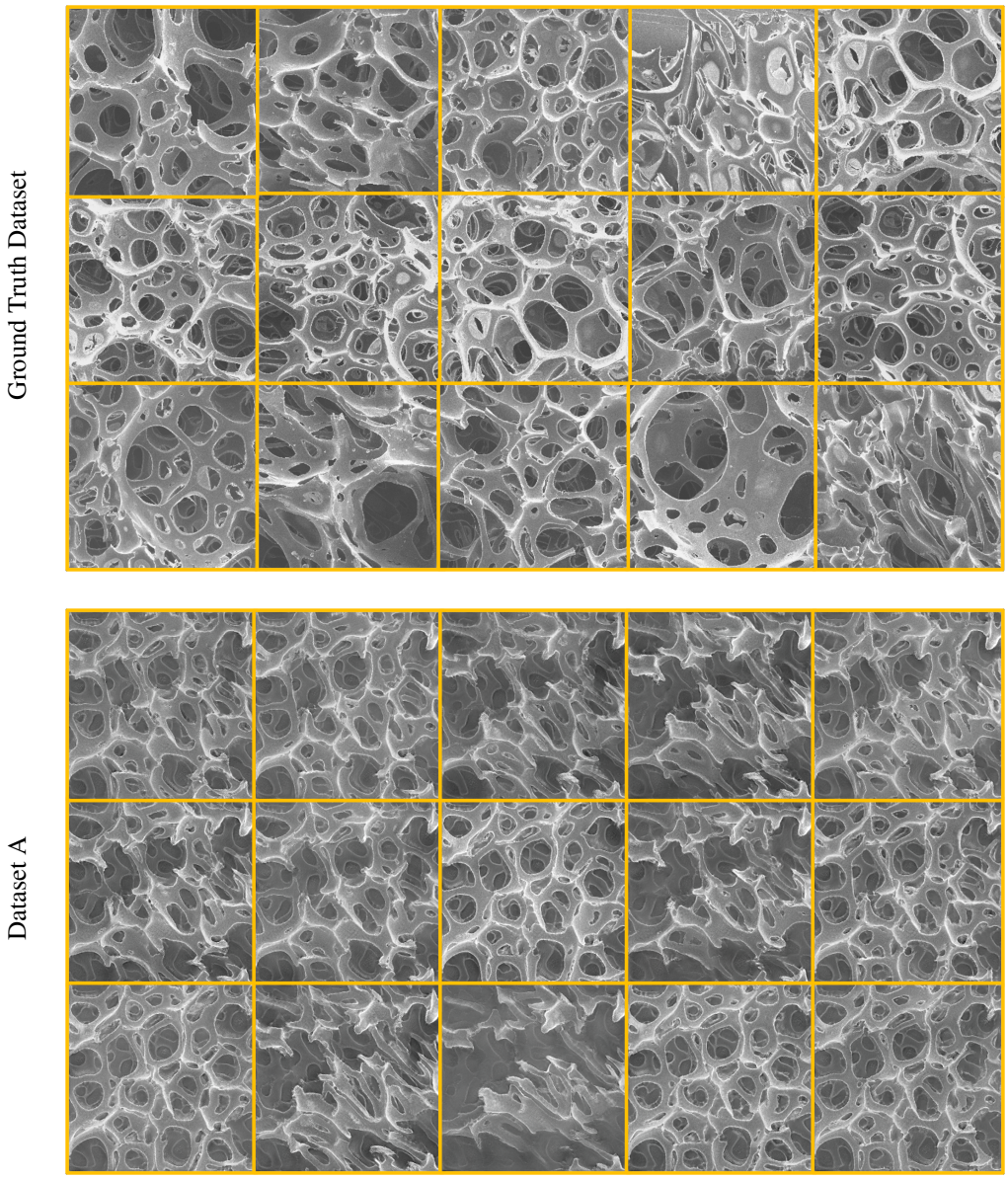


Figure 3: A random set 15 images were selected from each of the datasets and shown here side by side to highlight image diversity. As expected, the Ground Truth Dataset is extremely diverse as it comes from 7 different combinations of additives. Even repeated SEM images of the same experiment from different angles show different features in the Ground Truth Dataset. On the other hand, we see the model used for generating Dataset B is over fit and repeating the same image over with very minor changes. Dataset A provides greater diversity; however, we still notice multiple transitions between the same few phases. Contrastingly, Dataset C is the most diverse with unique micro-structure configurations in each of the images.

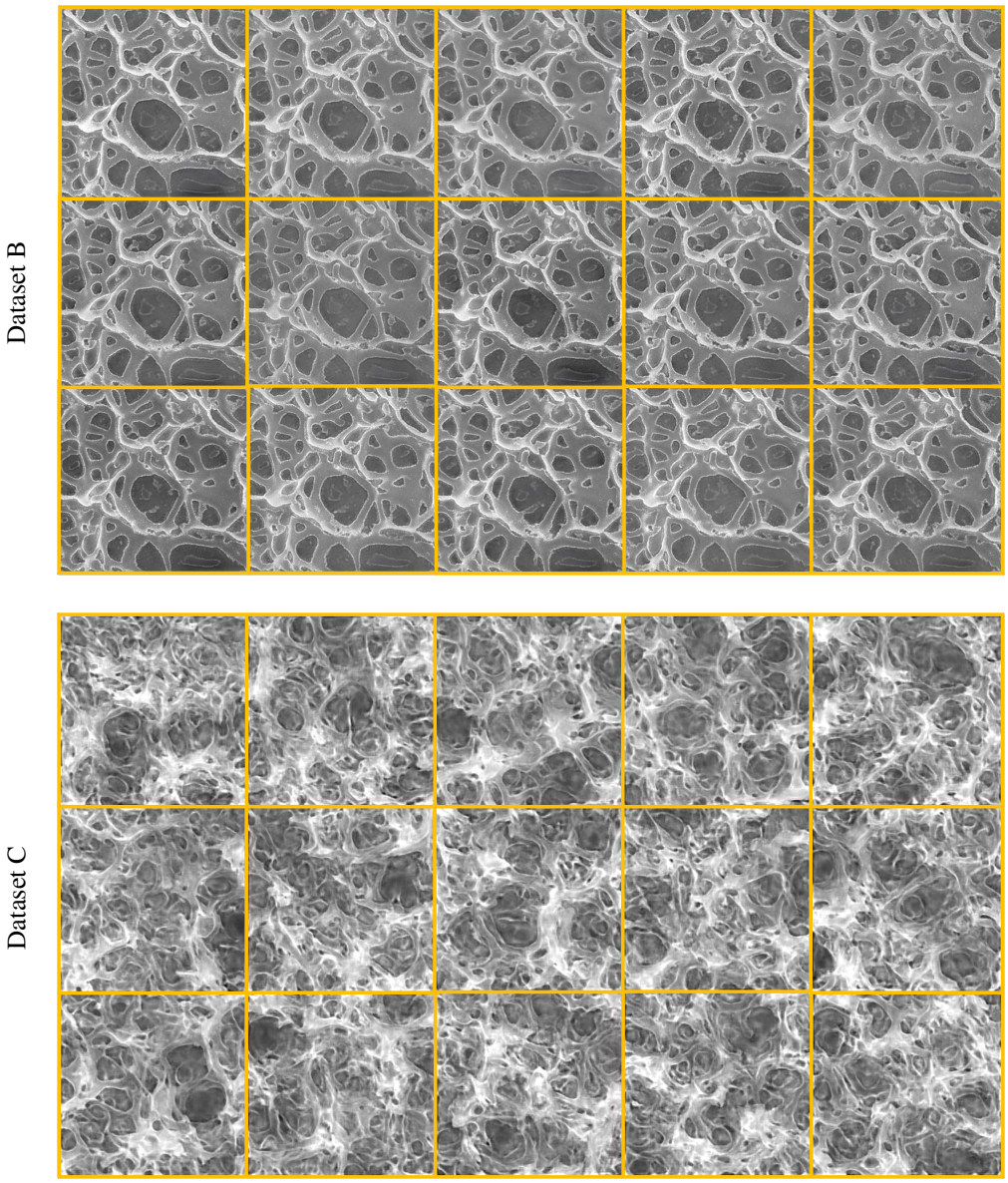
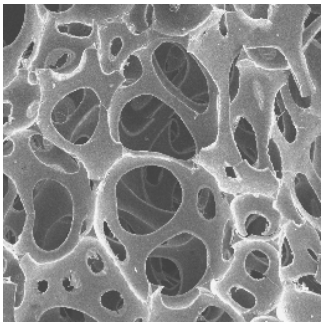


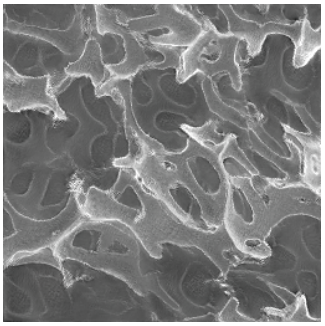
Figure 3: A random set 15 images were selected from each of the datasets and shown here side by side to highlight image diversity. As expected, the Ground Truth Dataset is extremely diverse as it comes from 7 different combinations of additives. Even repeated SEM images of the same experiment from different angles show different features in the Ground Truth Dataset. On the other hand, we see the model used for generating Dataset B is over fit and repeating the same image over with very minor changes. Dataset A provides greater diversity; however, we still notice multiple transitions between the same few phases. Contrastingly, Dataset C is the most diverse with unique micro-structure configurations in each of the images.

Ground Truth Dataset



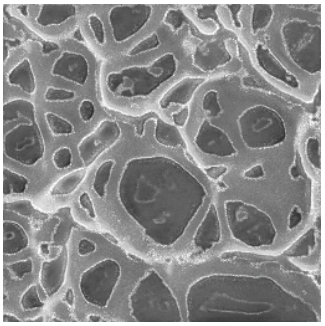
There are clear cell structures and thick walls outline each of the cells. Note, minimal number of the cells are broken with edges surrounding each intermediate cell hole. In addition, there is a clear evidence of depth in the micro-structure such that layers of the foam in the front are more reflective of lighting from the SEM.

Dataset A



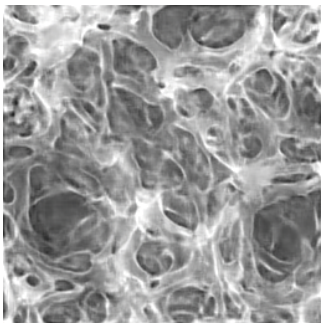
The images in this dataset lack complete cell structures. Several images may include complete cell walls; however, many of the images are representative of broken cells. This particular image lacks complete enclosed walls that represent the cell structure. The broken structure is representative of an image on the edge of a foam, a few of which were included in the training dataset. These are not experimentally valid structures.

Dataset B



There is a clear set of cell structures with thick walls outlining each of the cells. However, there is no sense of depth in this image. Beyond the initial layer of the micro-structure, we see an incomplete formation of cells in the background, thus not showing the dimensionality that is expected of a true micro-structure.

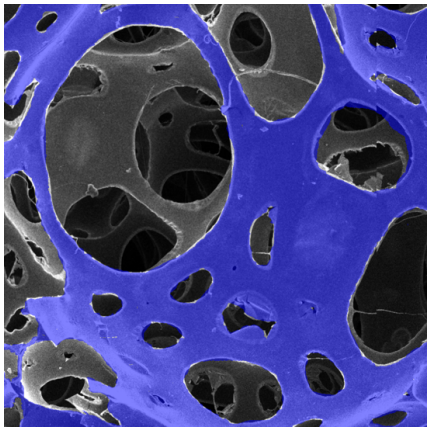
Dataset C



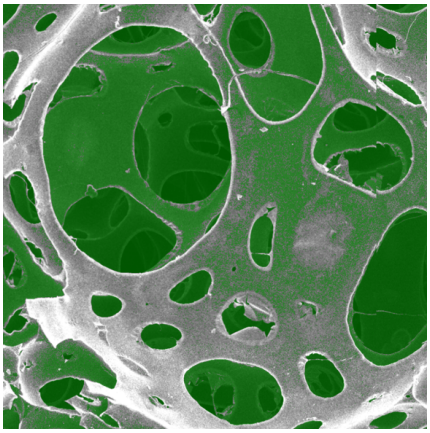
Despite low image quality, we see clear micro-structures and cell walls in the image. The image quality impacts the physically accurate scoring due to the inability get accurate measurements of cell wall thickness and cell sizes.

Figure 4: A descriptive analysis on the physically accurate scoring of each of the datasets. The key considerations which are visible to the human eye in an SEM image include cell structures, cell wall, and depth of cells beyond initial layer. We only consider visually interpretable conclusions currently; however, there is opportunity to leverage scoring metrics, as seen in Appendix D, to quantify such measurements.

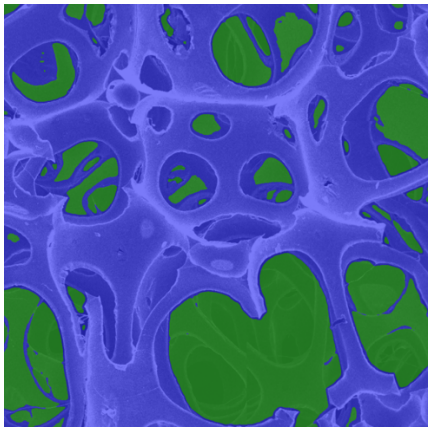
D Dragonfly Image Segmentation



(a) Manual segmentation using Dragonfly Deep Learning Tool



(b) Otsu thresholding of internal structure



(c) Segmentation map prediction of pores and cell structure

Figure 5: Dragonfly, a deep learning based image processing software, allows us to perform automated labeling on micro-structure images with high precision given very limited data. Beyond image segmentation, this tool allows us to quantify parameters such as number of cells, cell wall width and size of cells. Using such values, we can statistically score various additive types as well as synthetic images features.