Harnessing Synthetic Datasets: The Role of Shape Bias in Deep Neural Network Generalization

Elior Benarous
ETH Zürich
ebenarous@ethz.ch

Sotiris Anagnostidis
ETH Zürich

Luca Biggio
ETH Zürich

Thomas Hofmann
ETH Zürich

Abstract

Recent advancements in deep learning have been primarily driven by the use of large models trained on increasingly vast datasets. While neural scaling laws have emerged to predict network performance given a specific level of computational resources, the growing demand for expansive datasets raises concerns. To address this, a new research direction has emerged, focusing on the creation of synthetic data as a substitute. In this study, we investigate how neural networks exhibit shape bias during training on synthetic datasets, serving as an indicator of the synthetic data quality. Specifically, our findings indicate three key points: (1) Shape bias varies across network architectures and types of supervision, casting doubt on its reliability as a predictor for generalization and its ability to explain differences in model recognition compared to human capabilities. (2) Relying solely on shape bias to estimate generalization is unreliable, as it is entangled with diversity and naturalism. (3) We propose a novel interpretation of shape bias as a tool for estimating the diversity of samples within a dataset. Our research aims to clarify the implications of using synthetic data and its associated shape bias in deep learning, addressing concerns regarding generalization and dataset quality.

1 Introduction

The success of deep learning models hinges on their ability to extract meaningful patterns from limited data, emphasizing the importance of comprehending the guiding principles of their learning process. In machine learning, inductive biases are foundational, representing the prior assumptions algorithms make about data structure. While neural networks are theoretically universal approximators [22], practical constraints, such as finite training data and capacity, necessitate inductive bias to narrow the hypothesis space when encountering new instances [19].

Inspired by advancements in neural network architecture, Vision Transformers (ViTs) have demonstrated superior performance in visual tasks compared to Convolutional Neural Networks (CNNs) [14], provided sufficient computational resources. CNNs are designed around convolution and pooling operations, involving filters that capture local features, with inherent biases like translation equivariance and locality. Conversely, ViTs employ a self-attention mechanism, dividing input images into patches, linearly projecting them for embeddings, and subsequently passing these embeddings, along with positional encodings, through a Transformer encoder. This mechanism enables ViTs to capture global dependencies among patches, facilitating a more holistic understanding of image context and dependencies. This underlines the remarkable capabilities of ViTs in visual data.

In the pursuit of elucidating the exceptional performance of vision algorithms, a compelling avenue has emerged: comparing the inherent predilections of these algorithms with the proclivities of the human visual system. This exploration has yielded the concept of the "shape bias" as a pragmatic yardstick for quantifying disparities between machine learning models and human perceptual processes [17]. Notably, human cognition predominantly relies on shape-based recognition, constituting

NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

approximately ~96% of decision-making, while texture plays a minor role. Conversely, CNNs place substantial reliance on textural cues, with shape contributing to only about ~25% of their decision-making process. In the realm of Vision Transformers, recent work by Dehghani et al. [13] has spotlighted that with extensive training data – surpassing 4 billion images in this instance – ViTs not only attain state-of-the-art performance in out-of-distribution zero-shot classification and dense output tasks but also exhibit a remarkable shape bias, amounting to approximately ~87%. This achievement positions ViTs as the models closest to mirroring human perceptual biases. In light of these insights, we posit an intriguing question: *could the shape bias exhibited by deep neural networks elucidate the gap between their recognition capabilities and the remarkable acumen of human vision?* To address this, we scrutinize the correlation between the shape bias and their generalization capacity.

In recent research, significant effort has been directed towards addressing challenges posed by limited training data. One effective approach is the use of data augmentations, as demonstrated in prior work [28]. This technique proves especially valuable when dealing with smaller datasets that exhibit limited sample diversity. However, it is important to acknowledge the limitations of pixel-space augmentations. While they do help, they still maintain natural image statistics, which may lead models to rely on superficial cues [24]. Alternatively, a promising strategy involves leveraging synthetic datasets for pre-training purposes. By employing a trained generator, researchers can generate datasets of virtually unlimited size. This approach mitigates issues related to data bias and under-representation, offering particular advantages in scenarios with limited data availability, such as certain medical applications. Moreover, generative models offer a cost-effective alternative to the labor-intensive processes of data collection and annotation. They also sidestep potential privacy and legal concerns, as exemplified by the restrictions placed on the widely used ImageNet dataset due to privacy, ethics, and labeling issues [7, 35].

Recent studies (e.g., Baradad et al. [3] and Kataoka et al. [26]) have recently shown that utilizing unrealistic-looking synthetic samples for pre-training neural networks can significantly reduce the need for extensive datasets. In fact, some of these models outperform traditional ImageNet pre-trained models. According to Baradad et al. [3], the key factor for effective training data is diversity and naturalistic characteristics, rather than strict realism. In other words, synthetic samples need not replicate real-world images precisely; they should instead capture essential structural properties of real data. Nevertheless, there is still no consensus on the specific properties that make a synthetic dataset most effective in harnessing its potential benefits.

This paper explores the role of shape bias in assessing a model's generalization potential on a synthetic dataset. We propose that dataset diversity is a critical factor in the dataset's quality, as it influences the model's shape bias across training epochs. Our primary focus is to monitor the evolution of shape bias throughout training. Our key finding (1) challenges the conventional use of the shape bias metric as a reliable predictor for generalization, particularly when considering various architectural types and levels of supervision. Furthermore, (2) we reveal the intricate relationships between shape bias, dataset diversity, and naturalism in the context of generalization. As a novel contribution, (3) we demonstrate the potential of shape bias as a reliable tool for measuring the diversity of training samples, presenting an alternative to methods like Fréchet Inception Distance (FID) [21].

2 Related Work

Inductive bias. Inductive biases, whether explicit or implicit, are crucial constraints imposed by a model on the learning process, shaping its knowledge acquisition. These biases are paramount in determining what the model learns and how it learns.

Notably, specific biases emerge from the model's architecture [18]. Convolutional Neural Networks, known for their translation invariance, introduce a strong locality bias [12], even when initialized randomly. This bias enables CNNs to efficiently locate salient objects within images [9]. Furthermore, the prevalent adoption of square pooling geometries amplifies a bias in favor of natural images [11], where adjacent pixels exhibit higher correlation. Adjusting these architectural choices can enhance CNNs' adaptability across various data types [11]. Nevertheless, the locality bias inherent to CNNs hampers their ability to learn patterns extending over larger regions of the data. Conversely, Vision Transformers leverage a self-attention mechanism that empowers them to capture both short and long-range dependencies between tokens, irrespective of their spatial separation. This stands in contrast to the limitations of CNNs, as discussed earlier.

The selection of training data introduces biases. Finite datasets exhibit inherent biases [33], prompting models to develop efficient strategies that excel within the training distribution (in-distribution *i.d.*) but falter in out-of-distribution (*o.o.d*) scenarios. Notably, CNNs trained on ImageNet have demonstrated a propensity for optimal performance with constrained receptive field sizes [6] and resistance to global shape distortions while remaining sensitive to local shapes [2, 1]. This suggests that ImageNet-trained CNNs tend to emphasize textures over shapes [17]. Remarkably, CNNs can rapidly acquire a shape bias with minimal training data [15], leading to improved robustness against image distortions and higher *o.o.d* accuracy [20]. Geirhos et al. [17] and Li et al. [30] employ style transfer techniques [16] to augment their training data with images that challenge the balance between shape and texture, ultimately yielding improved accuracy and adversarial robustness Li et al. [30]. This work underscores the complementary relationship between shape and texture biases in models.

Learning from synthetic data. In recent developments, deep generative models, notably Generative Adversarial Networks (GANs), have emerged as valuable sources of synthetic data for improving classification [23, 5, 32] and semantic segmentation [36, 34, 29]. These models operate by learning from authentic image datasets and are adept at generating convincing synthetic samples. In contrast, Kataoka et al. [26] have directed their focus towards fractals, intricate geometric patterns characterized by self-similarity across different scales, which can be produced through purely procedural methods. Such patterns are prevalent in the natural world, exemplified in structures like leaves, snowflakes, and seashells. Furthermore, Baradad et al. [3] have undertaken an analysis of contrastive representation training using unrealistic data derived from statistical image models, procedural graphics models [26], and samples drawn from generative models initialized at random. In our study, we place particular emphasis on the use of such unrealistic data samples, deliberately sidestepping potential biases and contamination issues frequently encountered in natural datasets.

3 Method

To comprehensively investigate the impact of shape bias, we evaluate both the ResNet architecture, representing CNNs, and the standard Vision Transformer architecture. We conduct experiments with a ResNet-18 and a ResNet-50, as well as ViT-pX models, dividing the input into a grid of 8x8 and 16x16 patches, referred to as ViT-p8 and ViT-p16 respectively. Our customized ViTs all contain 7 encoder layers with 8 heads, and the hidden dimension is set to 512.

Training details. We train our models on 6 synthetic datasets. First is the Fractals dataset, which we generate following the process provided by Kataoka et al. [26] (see Appendix A for specifics). The other 5 synthetic datasets are generated from a GAN. Four are obtained through an untrained StyleGANv2 [25] whose parameters are sampled from different prior distributions. These prior distributions are sufficient to encode image properties without requiring any training on real data. A randomly initialized model is used to obtain StyleGAN - Random 4. StyleGAN - High-freq. 5 has supplementary noise maps and wavelets filters applied to increase the presence of noise. StyleGAN - Sparse 6 is intended to mimic the sparsity of natural images by adding a random bias to the convolutions and multiplying the noise maps to randomly sampled Laplacian noise. StyleGAN - Oriented 7 further aims to approach the properties of real image by integrating oriented structures through tied wavelets. The last dataset, Shaders-21k StyleGAN 8, is obtained by sampling from a StyleGANv2, which was trained on the procedurally generated dataset Shaders-21k. The latter, itself, is created by applying 21 thousand generative procedures to shaders. See Appendix B to visualize samples from each class and [25, 3, 4] for further details on the creation of these synthetic datasets. All datasets contain 95 thousand instances, sampled uniformly across generative parameters where applicable. Each image is obtained in size 64x64, except for those in Shaders-21k StyleGAN, which we resize from 224x224 to 64x64. Lastly, we test downstream *o.o.d* performance on Tiny ImageNet.

For the synthetic datasets that do not contain labels, we train with self-supervision using the normalized temperature-scaled cross-entropy loss [10]. The latter maximizes the similarity between the representations of two augmented versions of a same image, while minimizing the similarity with all other samples in the batch. The two versions are extracted as random crops of the same image. We set the temperature to 0.5 and use the cosine similarity as a similarity measure.

Other training hyperparameters are kept constant across model architectures, datasets, and supervision types (see Appendix A for specifics). This helps ensure a fair comparison between the aforementioned

variables. Prior to evaluation, we freeze the pre-trained encoder, randomly reset the fully connected layers, and fine-tune the latter with supervision during two epochs on the downstream dataset.

Finally, the training samples undergo augmentation through different schemes, which vary between datasets. For Tiny ImageNet, we apply the *simCLR* augmentation scheme [10]. We transform the Fractals dataset with random horizontal flip, and random resized crop with a scale ranging between 8% and 100% of the original image size. Finally, we apply the same augmentation scheme as Baradad et al. [3] to the GAN-generated samples.

Evaluation details. The benchmark dataset introduced by Geirhos et al. [17] to estimate the shape bias of a model is referred to as Cue conflict. It is composed of 1200 images generated using style transfer between two ImageNet samples. Consequently, these exhibit conflicting shape and texture features. Each image is assigned two labels, indicating the shape and the transferred texture. For each sample, we evaluate whether a model accurately classified the shape or texture. To this end, we design an alternative method, distinct from that of Geirhos et al. [17]. Rather than performing linear classification, we evaluate the embeddings with K-nearest-neighbours (K-NN) classification. This addresses the need of employing a model with the same output classes as those of ImageNet, but also allows us to specifically analyze the encoder. Given the smaller size of the dataset, we fix the number of neighbours to $K = 5$. We then classify the embedding of each sample twice, first according to the shape label of the remaining 1199 embeddings, and then according to their texture label. In both cases, we use the cosine distance to determine the closest embeddings. Ultimately, the shape bias is computed as the ratio between the number of samples correctly classified by their shape and the total number of samples correctly classified by either their shape or texture. The bias hence ranges from 0 to 1, where 0 indicates a strong texture bias and 1 indicates a strong shape bias.

4 Experiments and Results

4.1 Limitations of Shape Bias in the face of Architectural and Supervision Variations

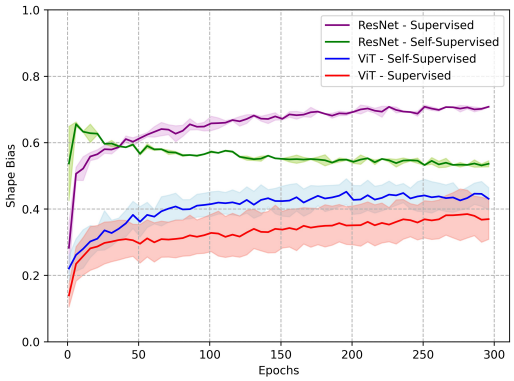


Figure 1: Variation of the K-NN Shape Bias. We display the results averaged between ViT models (ViT-p8 and ViT-p16) and ResNet models (ResNet-18 and ResNet-50). Darker lines correspond to the averages and the associated filled regions have for upper and lower bounds the values of each model.

We first investigate the evolution of the shape bias during training on a realistic dataset, Tiny ImageNet, where the concept of shapes is more tangible. Several observations emerge that shed light on the intricacies and nuances of this bias. These insights are pivotal in determining the validity of the shape bias metric as a reliable indicator of a dataset's capacity to teach models to generalize to real-world samples.

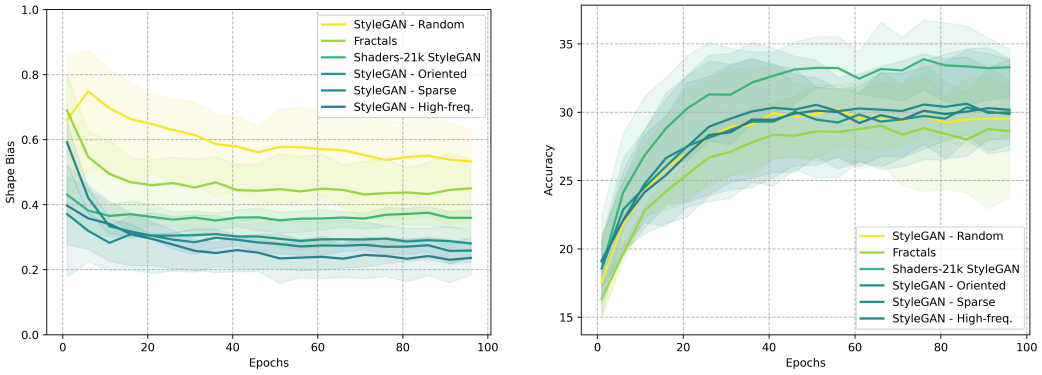
As depicted in Figure 1, each combination of architecture and supervision follows a specific trend of its own. All but the ResNets with self-supervision see their shape bias increase over the epochs. Also, training with supervision induces a higher preference for shape with the convolution-based models, but not with the attention-based ones.

ResNets display the larger K-NN shape bias across supervision types. However, this attempt to vaguely form groups according to architecture is limited by intra-class dissimilarities. For instance, we notice a much larger shaded region for the ViTs, irrespective of the supervision type. In both cases, the upper and lower bounds of the filled region correspond to the ViT-p8 and ViT-p16 respectively. The sole distinction between these models is the number of patches they extract. This highlights a compelling correlation between the number of tokens and the resultant shape bias. This seems intuitive, as more patches equate to smaller segments of the image, thereby emphasizing local over global information.

These observations reveal that the shape bias is not solely a product of the training dataset. It is induced quite distinctively by supervised and self-supervised learning, and is expressed differently by

CNNs and ViTs. This complexity challenges the feasibility of employing the shape bias as a robust and straightforward predictor for model generalization from synthetic to real-world datasets.

4.2 Does Synthetic Data Generalize well because of Shape Bias?



(a) (b)

Figure 2: Variation of the (a) K-NN Shape Bias and (b) Downstream accuracy on Tiny ImageNet's validation set over 100 epochs. In both plots, darker lines correspond to the average over the three architectures pre-trained on the same dataset. The associated shaded regions have for lower and upper bounds the values of the lowest and highest performing architectures respectively. For all datasets, the lower bound for the K-NN Shape Bias plot corresponds to the ViT-p8's performance. Colors follow the order of the average Shape Bias of each dataset after convergence (epoch 50 onward).

We now delve into the results obtained by training our models on the six synthetic datasets and testing them on Tiny ImageNet as our *o.o.d* dataset. For computational purposes, we focus on the ViT-p8 as our transformer architecture. We track the performance of the models throughout 100 epochs of training. We aim to determine whether there is a relation between the shape bias captured by a synthetic dataset and the generalization potential of a model pre-trained on that dataset.

An evident observation in Figure 2 is that the shape biases generally start at a higher value compared to those seen with Tiny ImageNet training. However, this value decreases over training epochs for all synthetic datasets. One might think that this is only caused by the choice of self-supervision. However this option is refuted by the evolution of the bias of the ViT-p8. Indeed, the K-NN shape bias of this architecture kept increasing when trained on Tiny ImageNet with self-supervision. However, it now also decreases from the first epoch onward. Hence, some inherent, possibly shared, property of these datasets must be the cause for that decrease in shape bias. As mentioned before, one plausible explanation, put forward by Baradad et al. [3], is the need for diversity.

When comparing the datasets between one another, we find that StyleGAN - Random induces the highest shape bias among the six datasets. Fractals and Shaders-21k StyleGAN take the second and third place respectively. But most importantly, the last three are in a particularly interesting order. Datasets that more closely resemble natural images achieved higher K-NN shape bias upon convergence. This suggests that our evaluation of the shape bias is relevant to understand the properties of synthetic datasets, especially in the aim of having them mimic real-world images. Nonetheless, one outlier remains, that is StyleGAN - Random. Although being the farthest statistically from real images, this dataset induces the strongest shape bias. Why is that? As discussed above, this metric is influenced by many parameters. Still, we can form the hypothesis that StyleGAN - Random samples are the ones that appear the smoothest visually, i.e. they contain the least texture. In that sense, we can interpret it being unable to induce a strong texture bias to any model because its samples contain none (or very few) high frequency signals. Consequently, even though a strong shape bias may be induced, this dataset is most likely lacking other crucial properties required in addition to shape bias for an overall strong generalization potential. Such properties include those mentioned by Baradad et al. [3]: diversity and statistical similarity of samples with the downstream dataset. When examining Figure 2b, we find that this is indeed the case. The order of the shape bias plot is lost and StyleGAN - Random does not stand out anymore on an *o.o.d* recognition task. Conversely, models pre-trained on Shaders-21k StyleGAN, supposedly the most diverse dataset, perform best.

Through our experiments on the six synthetic datasets, a key insight surfaced: while the K-NN evaluation of the shape bias offers valuable perspectives, it is not an all-encompassing metric for predicting generalization capability. Intricately interwoven with shape bias are the fundamental elements of diversity and naturalism. A case in point are models pre-trained on StyleGAN - Random, which exhibit a pronounced shape bias. However, this bias might be a manifestation of its samples' texture paucity, which potentially compromises its diversity and naturalism. As a result, despite its strong shape bias, this dataset doesn't surpass others in generalization.

4.3 Evaluating the Diversity of a Dataset through the Shape Bias

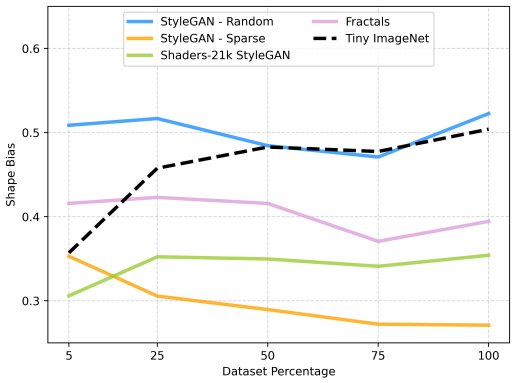


Figure 3: Average value of the K-NN Shape Bias after convergence (epoch 50 onward). Results are averaged between the architectures we experimented with.

Building on our earlier findings, we investigate the potential of shape bias as an estimator for dataset diversity. The commonly adopted FID [21] score computes a distance using a pre-trained model, inevitably infusing the bias inherent to its training dataset. In contrast, our proposed method leverages shape bias for diversity estimation, eliminating the need for an auxiliary model trained on a distinct dataset.

In accordance with Occam's razor, DNNs have the tendency to prioritize learning simple patterns of low frequency modes that generalize well [8], which Rahaman et al. [31] refer to as spectral bias. Hence, we make the hypothesis that for highly diverse datasets, an increase in the number of samples would result in an increased shape bias. This is because a model with more data would be less prone to overfitting and, therefore, less likely to rely on high-frequency textures over low-frequency shapes. For this study, we pre-train models using varying amounts of synthetic data, specifically: 5%, 25%, 50%, 75% and 100% of the total samples. We focus on the ViT-p8 and ResNet-18 architectures for computational purposes. Also, we center the discussion on StyleGAN - Random, due to its high K-NN shape bias in Cue conflict; Shaders-21k StyleGAN and Fractals, as both of which were visually very different from other datasets in our study; and StyleGAN - Sparse, because of its low shape bias but with some naturalistic elements by design.

The results displayed in Figure 3 align with our hypothesis. We observe an increase of the shape bias with the number of Tiny ImageNet samples, which we consider being diverse. In contrast, we notice very slim variations in the K-NN shape bias over the range of percentages for all synthetic datasets. Models trained on the Fractals and StyleGAN - Sparse datasets reach a maximum shape bias at 25% and 5% of the initial number of samples respectively. In a similar vein, the value at 25% for Shaders-21k StyleGAN and StyleGAN - Random are lower than the peak value, reached with the complete dataset, by approximately 0.5% and 1.1% respectively. In essence, the shape biases converge to consistent values, whether we use a reduced portion of the synthetic datasets or not. These results suggest that the models keep relying on similar cues, mainly texture-based, regardless of the number of samples, which implies a deficiency in shape diversity within the datasets.

5 Conclusion

In this paper, we investigate the extent to which the shape bias of a model pre-trained on a synthetic dataset can be used to predict its generalization to realistic samples. We first unveil the limitations of using such metric on any dataset due to the intricate dynamics with types of architecture and supervision. We then highlight the entangled impact of shape bias, diversity and naturalism of synthetic samples on generalization. Lastly, we suggest an alternative use of the shape bias as a proxy to estimate the diversity of samples. We hope that this study spurs further research in achieving a holistic understanding of properties of synthetic datasets and in elucidating strategies for crafting effective synthetic datasets.

References

[1] Antonio A. Abello, Roberto Hirata, and Zhangyang Wang. Dissecting the high-frequency bias in convolutional neural networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 863–871, 2021. doi: 10.1109/CVPRW53098.2021.00096.

[2] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14: e1006613, 12 2018. doi: 10.1371/journal.pcbi.1006613.

[3] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise, 2021. URL `https://arxiv.org/abs/2106.05963`.

[4] Manel Baradad, Chun-Fu Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning, 2022. URL `https://arxiv.org/abs/2211.16412`.

[5] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images, 2019. URL `https://arxiv.org/abs/1911.02888`.

[6] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *CoRR*, abs/1904.00760, 2019. URL `http://arxiv.org/abs/1904.00760`.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/buolamwini18a.html`.

[8] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *CoRR*, abs/1912.01198, 2019. URL `http://arxiv.org/abs/1912.01198`.

[9] Yun-Hao Cao and Jianxin Wu. A random CNN sees objects: One inductive bias of CNN and its applications. *CoRR*, abs/2106.09259, 2021. URL `https://arxiv.org/abs/2106.09259`.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL `https://arxiv.org/abs/2002.05709`.

[11] Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. *CoRR*, abs/1605.06743, 2016. URL `http://arxiv.org/abs/1605.06743`.

[12] Stéphane d'Ascoli, Levent Sagun, Joan Bruna, and Giulio Biroli. Finding the needle in the haystack with convolutions: on the benefits of architectural bias. *CoRR*, abs/1906.06766, 2019. URL `http://arxiv.org/abs/1906.06766`.

[13] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. URL `https://arxiv.org/abs/2302.05442`.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.

[15] Reuben Feinman and Brenden M. Lake. Learning inductive biases with simple neural networks. *CoRR*, abs/1802.02745, 2018. URL http://arxiv.org/abs/1802.02745.

[16] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *CoRR*, abs/1811.12231, 2018. URL http://arxiv.org/abs/1811.12231.

[18] Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning, 2022. URL https://arxiv.org/abs/2211.14699.

[19] David Haussler. Quantifying inductive bias: Ai learning algorithms and valiant's learning framework. *Artificial Intelligence*, 36(2):177–221, 1988. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(88)90002-1. URL https://www.sciencedirect.com/science/article/pii/0004370288900021.

[20] Katherine L. Hermann and Simon Kornblith. Exploring the origins and prevalence of texture bias in convolutional neural networks. *CoRR*, abs/1911.09071, 2019. URL http://arxiv.org/abs/1911.09071.

[21] 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, 2018. URL https://arxiv.org/abs/1706.08500.

[22] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

[23] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning, 2021. URL https://arxiv.org/abs/2106.05258.

[24] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, abs/1711.11561, 2017. URL http://arxiv.org/abs/1711.11561.

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019. URL https://arxiv.org/abs/1912.04958.

[26] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images, 2021. URL https://arxiv.org/abs/2101.08515.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[29] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization, 2021. URL https://arxiv.org/abs/2104.05833.

[30] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan L. Yuille, and Cihang Xie. Shape-texture debiased neural network training. *CoRR*, abs/2010.05981, 2020. URL https://arxiv.org/abs/2010.05981.

[31] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

[32] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models, 2019. URL https://arxiv.org/abs/1905.10887.

[33] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

[34] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation, 2021. URL https://arxiv.org/abs/2103.04379.

[35] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, jan 2020. doi: 10.1145/3351095.3375709. URL https://doi.org/10.1145%2F3351095.3375709.

[36] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort, 2021. URL https://arxiv.org/abs/2104.06490.

Appendix

A Further training details

Dataset generation. Regarding the creation of the Fractals dataset, we follow the process defined by Kataoka et al. [26]. We keep the intra-category augmentation parameters constant throughout. With such, we sample 125 sets of fractal generation parameters and create 1000 images per set of parameters. Accordingly, we obtain 125,000 samples.

Hyperparameters. We choose Adam [27] as our optimizer for both pre-training and finetuning. A summary of the hyperparameters can be found in Table 1. These are kept identical for both pre-training and finetuning.

Hyperparameter	**Value**
Adam β_1	0.9
Adam β_2	0.999
Learning Rate	10^{-4}
Weight Decay	None
Batch Size	64

Table 1: Training hyperparameters and associated values.

B Dataset Visualization

StyleGAN - Random

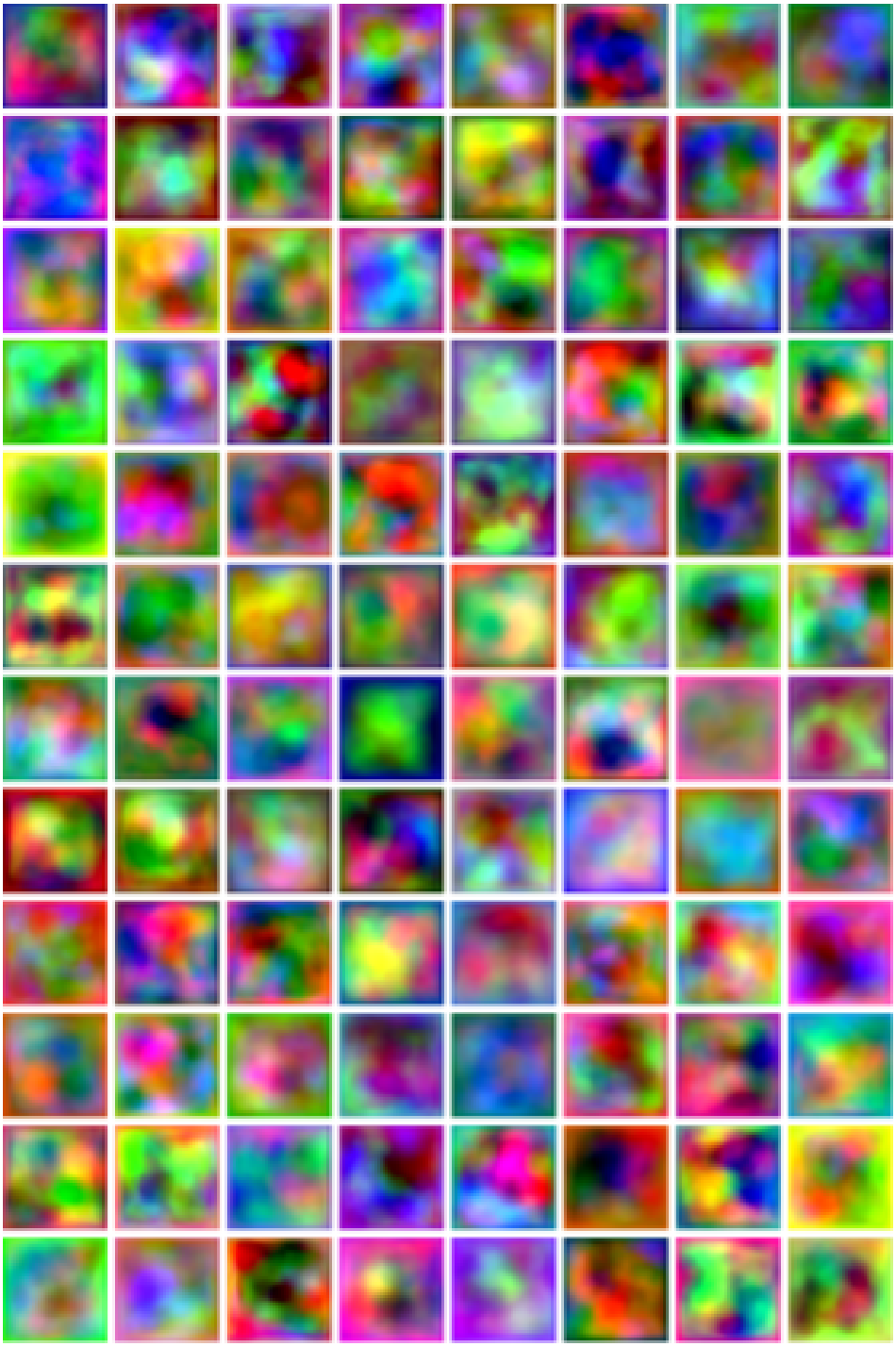


Figure 4: 96 random samples of the dataset StyleGAN - Random. Figure taken from [3]

StyleGAN - High freq.

Figure 5: 96 random samples of the dataset StyleGAN - High freq.. Figure taken from [3]

StyleGAN - Sparse

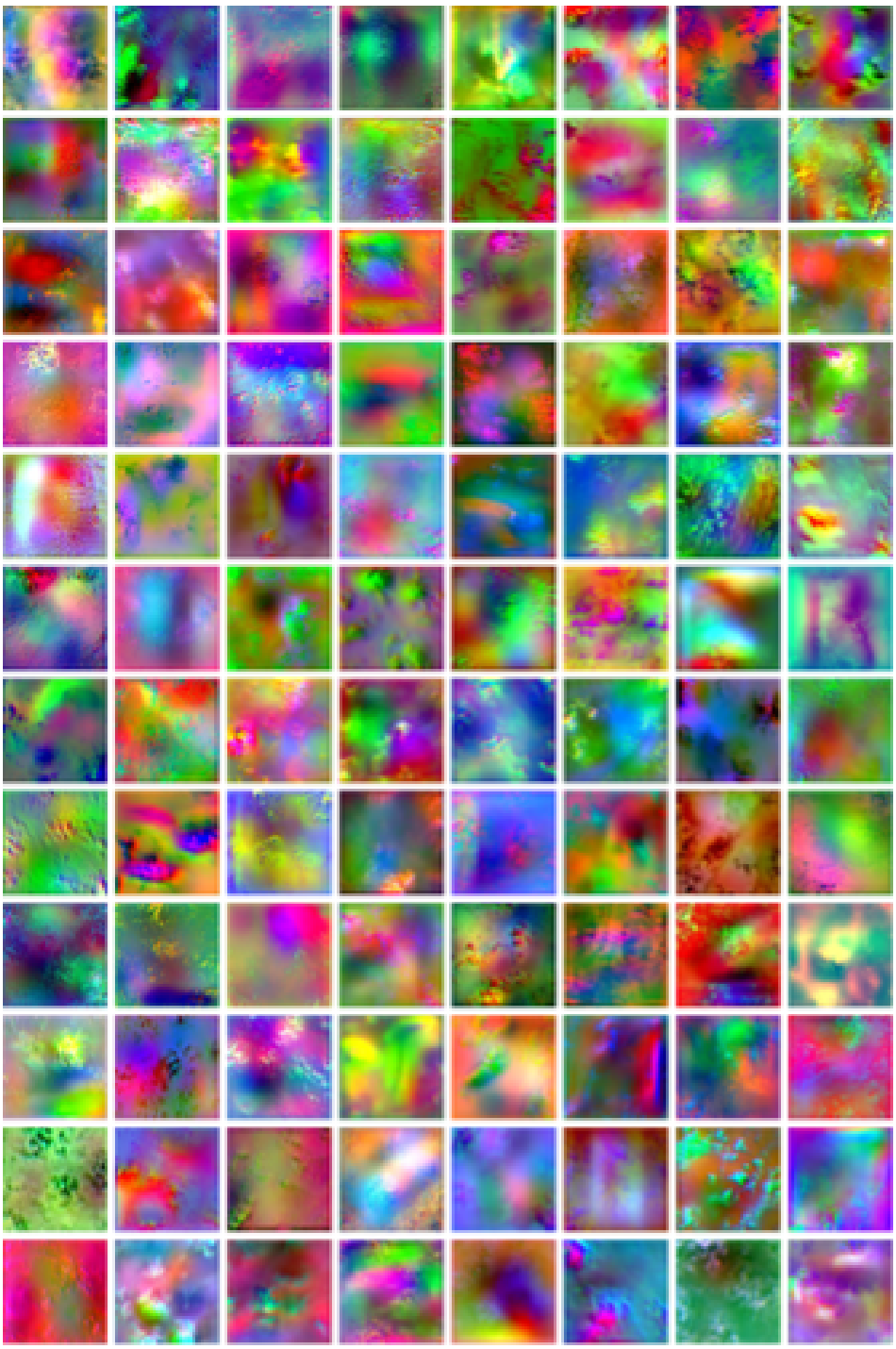


Figure 6: 96 random samples of the dataset StyleGAN - Sparse. Figure taken from [3]

StyleGAN - Oriented

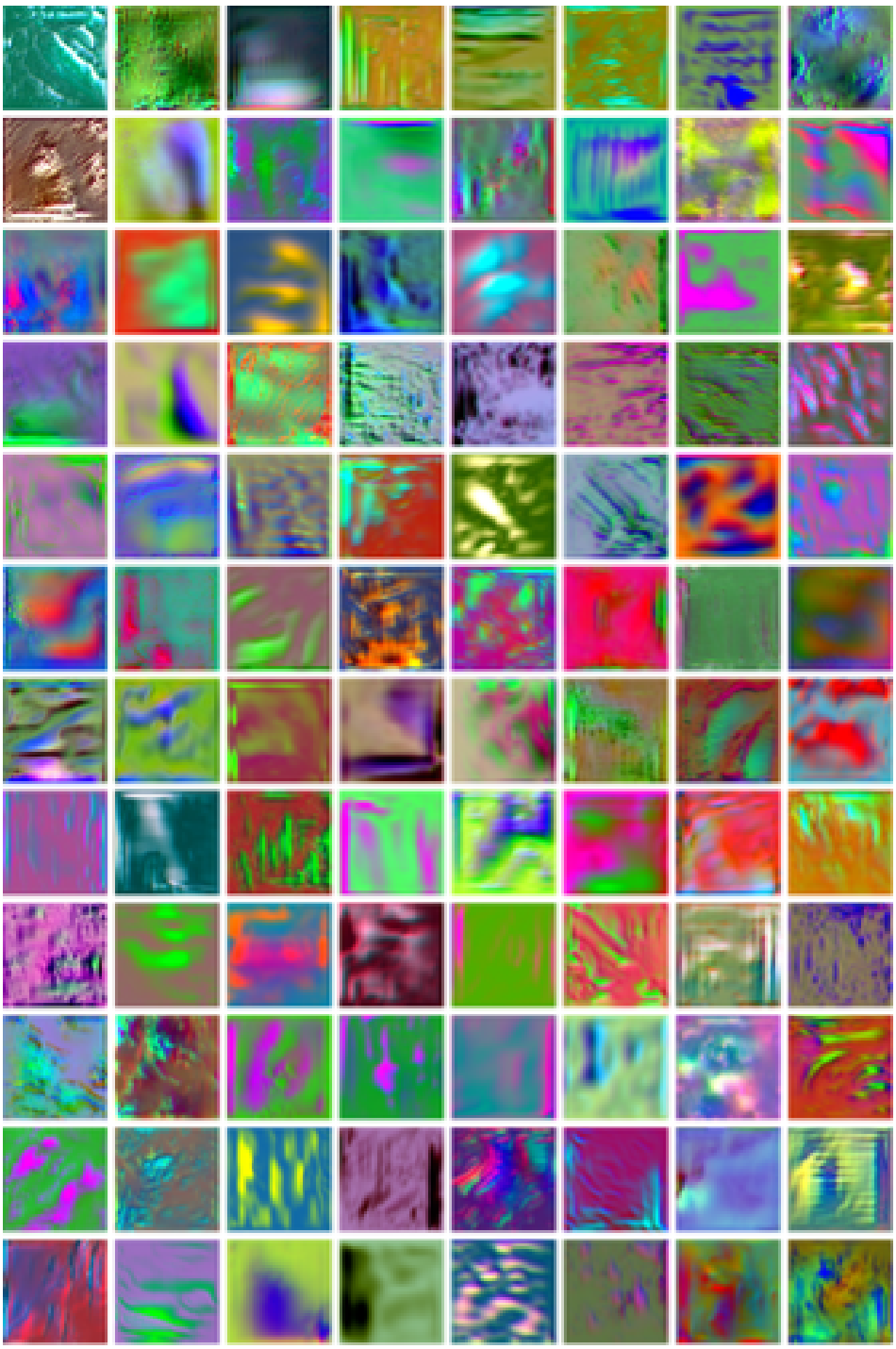


Figure 7: 96 random samples of the dataset StyleGAN - Oriented. Figure taken from [3]

S-21k StyleGAN

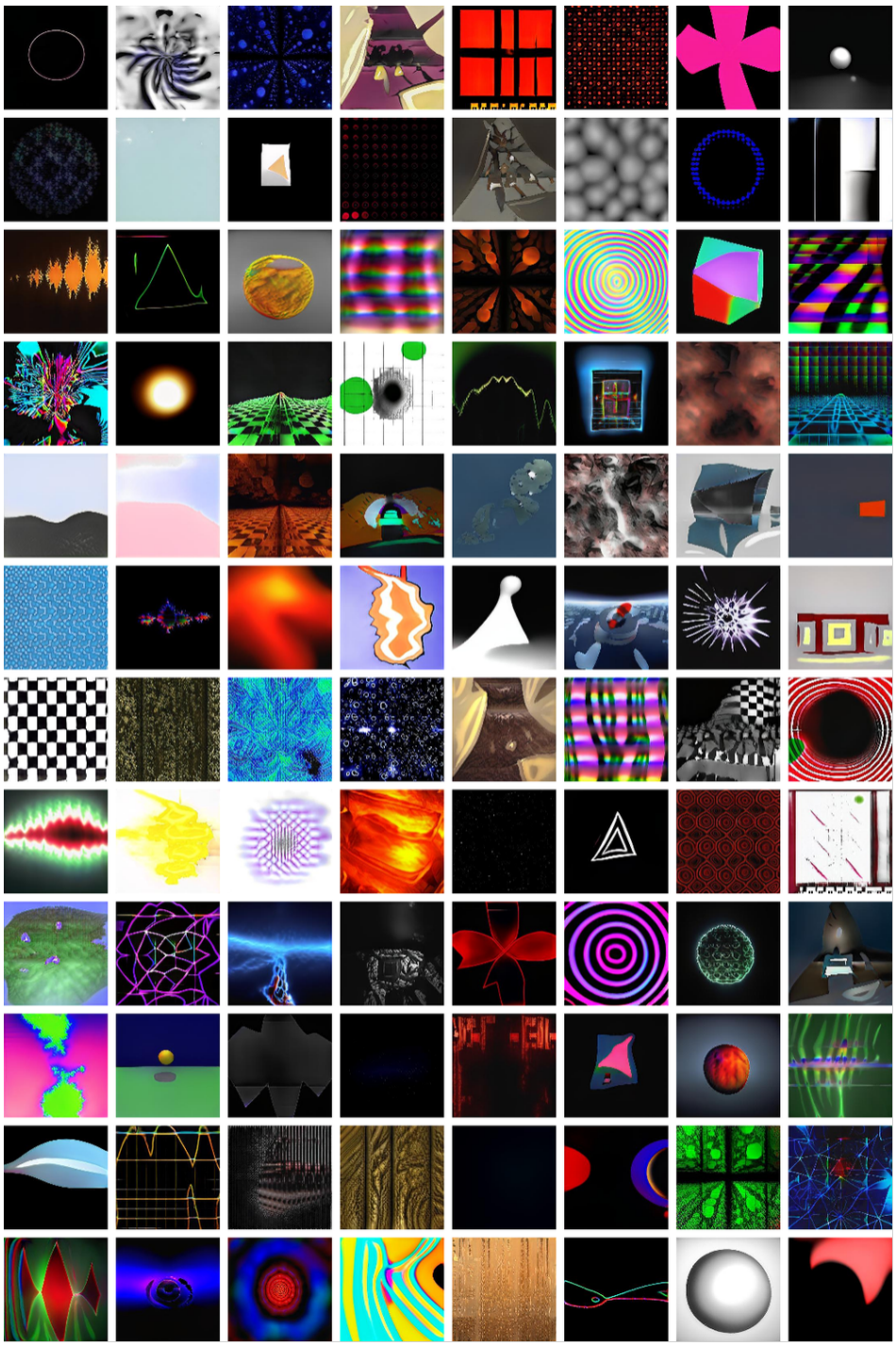


Figure 8: 96 random samples of the dataset S-21k StyleGAN. Figure taken from [4]

Fractals

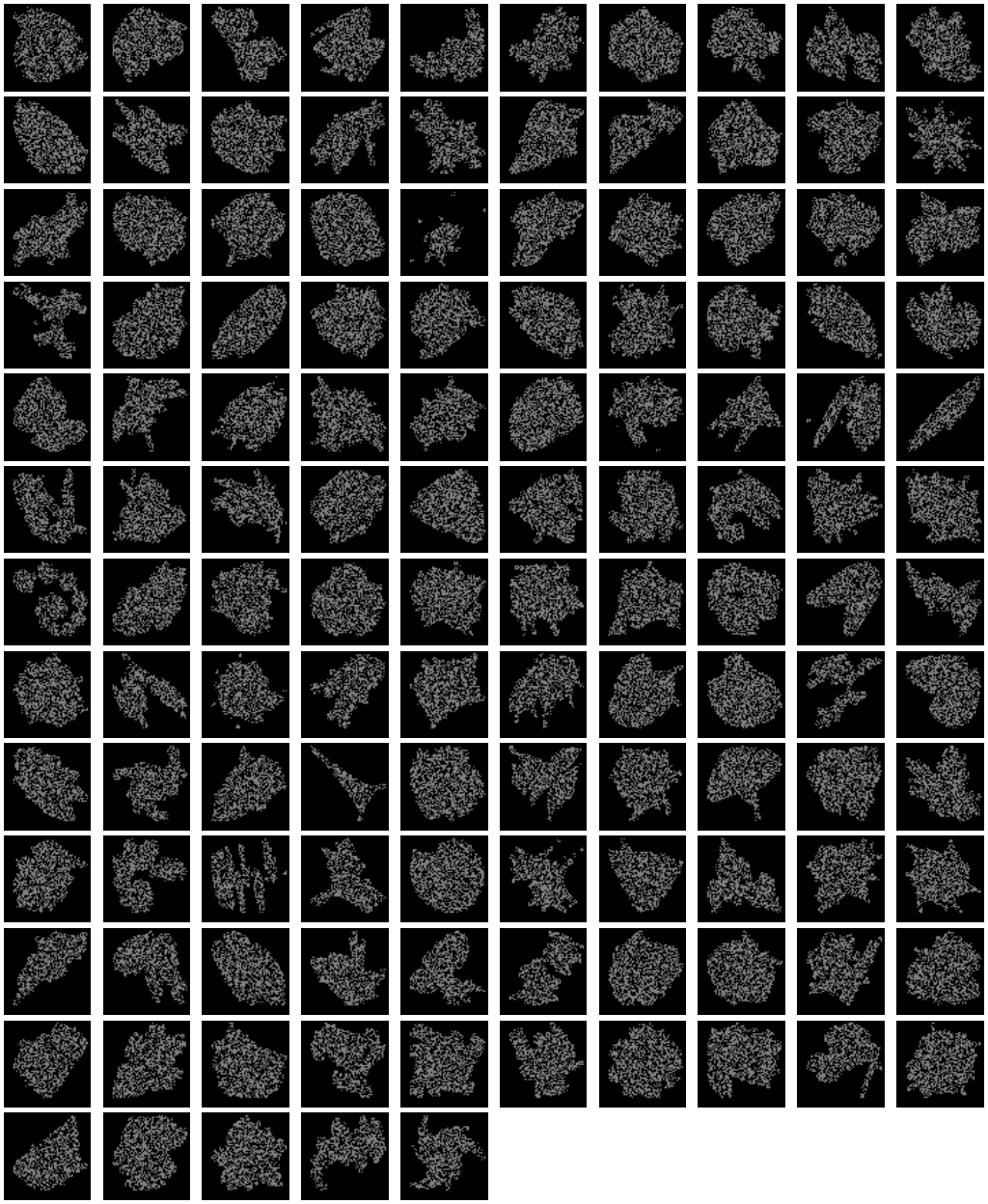


Figure 9: 125 random samples of the dataset Fractals