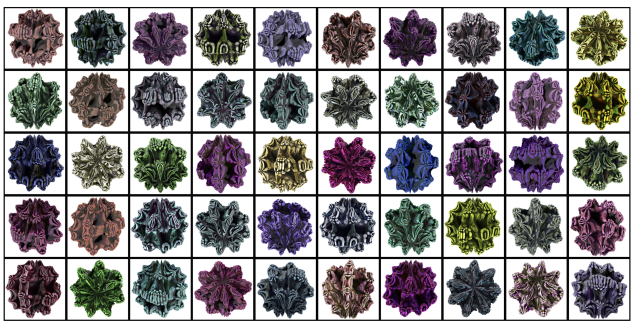
Pre-training Vision Models with Mandelbulb Variations

Supplementary Material

1. Data samples

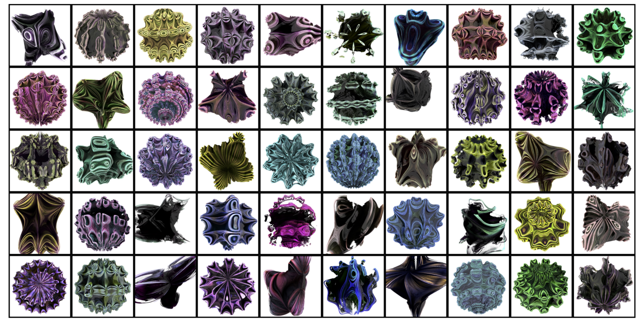
(a) MandelbulbVAR-1k classes



(b) MandelbulbVAR-1k instances

Figure 1. Example MandelbulbVAR-1k images. Best viewed in color and zoomed in. (a) Images of randomly selected 50 classes with one instance per class. (b) Images of randomly selected 50 instances in a randomly selected class.

Figs. 1 and 2 displays example images of our proposed datasets.



(a) MandelbulbVAR-Hybrid-21k classes

(b) MandelbulbVAR-Hybrid-21k instances

Figure 2. Example MandelbulbVAR-Hybrid-21k images. Best viewed in color and zoomed in. (a) Images of randomly selected 50 classes with one instance per class. (b) Images of the 50 instances in a randomly selected class.

2. Codes

We implement our original fractal modeling and rendering software based on OpenGL Shading Language (GLSL) [5] and we make it publicly available on our GitHub repository[1]. We include a permissive license that allows commercial use.

To pre-train CNNs and fine-tune them on supervised classification tasks, we use the codes released by the authors of [3][2]. To pre-train and fine-tune ViTs, we make use of the codes published by the authors of [7][3]. Regarding WideResNet-50 models for the anomaly detection task, we pre-train them using our own training codes. Once their pre-trained weights are saved, we run and evaluate PatchCore algorithms based on them by using the official codes released by the authors of PatchCore [6][4].

3. Training hyper-parameters for ViT pre-training and fine-tuning

Training hyper-parameters related to our experiments using ViTs are summarized in Tab. 1. They are almost the same as in the studies [7, 8].

4. CNN pre-training on MandelbulbVAR-Hybrid-21k

Tab. 2 shows the average classification accuracies of ResNet-50 models pre-trained on various datasets including MandelbulbVAR-Hybrid-21k. Tab. 3 includes the anomaly

[1] https://github.com/RistoranteRist/MandelbulbVariationsGenerator

[2] https://github.com/hirokatsukataoka16/FractalDB-Pretrained-ResNet-PyTorch

[3] https://github.com/masora1030/CVPR2023-FDSL-on-VisualAtom

[4] https://github.com/amazon-science/patchcore-inspection

Phase	Pre-training	Fine-tuning	
# Dataset classes	1k & 21k	1k	others
Epochs	300	300	1000
Batch size	1024	1024	768
Optimizer	AdamW	AdamW	SGD
LR	1.0e-3	1.0e-3	1.0e-2
Weight decay	0.05	0.05	1.0e-4
LR scheduler	Cosine	Cosine	Cosine
Warmup steps	5k	-	-
Warmup epochs	-	5	10
Resolution	224	224	224
Label smoothing	0.1	0.1	0.1
Drop path	0.1	0.1	0.1
Rand augment	9/0.5	9/0.5	9/0.5
Mixup	0.8	0.8	0.8
Cutmix	1.0	1.0	1.0
Erasing	0.25	0.25	0.25

Table 1. Hyper-parameters employed when training ViT-T and ViT-B models in our experiments.

Pre-training	Average accuracy
From scratch	74.6
FractalDB-1k	82.2
FractalDB-10k*	**83.5**
VisualAtom-1k	83.1
RCDB-1k	82.1
ExFractalDB-1k	82.4
MandelbulbVAR-1k	**84.5**
MandelbulbVAR-Hybrid-21k	**84.5**

Table 2. Average top-1 accuracy of ResNet-50 models over the validation sets of ImageNet-1k, CIFAR-10, CIFAR-100, Flowers and ImageNet-100. These models are either trained from scratch or fine-tuned after being pre-trained on different datasets. Best, second-best, and third-best scores are shown in underlined bold, bold, and underlined, respectively. The model with * is downloaded from the project page of the work [3].

detection performance recorded by PatchCore based on MandelbulbVAR-Hybrid-21k pre-trained WideResNet-50. On the one hand, MandelbulbVAR-Hybrid-21k pre-training outperforms existing FDSL regarding both anomaly detection and classification downstream tasks. For example, regarding anomaly detection, the average image-level AUROC recorded by PatchCore based on MandelbulbVAR-Hybrid-21k is better than the one recorded by the algorithm based on VisualAtom-1k (96.7% vs 92.7%). On the other hand, in any case, MandelbulbVAR-Hybrid-21k does not perform better than MandelbulbVAR-1k. For classification, they present the same average accuracy. For anomaly detection, MandelbulbVAR-1k pre-training is better than the one based on MandelbulbVAR-Hybrid-21k. For example, between these pre-training schemes, there is a gap in average image-level AUROC of 0.5% (97.2% vs 96.7%). We conclude that contrary to ViTs, CNNs do not benefit from the shape diversity of MandelbulbVAR-Hybrid-21k. As CNNs have lower shape bias than ViTs [9], this result is not surprising.

Pre-training	Img. AUROC	Pw. AUROC	PRO
Rand. init.	77.2	85.9	55.8
ImageNet-1k	**99.1**	**98.1**	**93.4**
ExFractalDB-1k	87.4	92.8	72.3
RCDB-1k	77.7	88.6	68.1
VisualAtom-1k	92.7	93.9	80.3
FractalDB-1k	90.6	90.7	70.8
MandelbulbVAR-1k	**97.2**	**96.8**	**89.6**
MandelbulbVAR-21k	96.7	**96.8**	89.3

Table 3. Anomaly detection performance (average image-level AUROC, pixel-wise AUROC and PRO in %) on MVTec AD [2]. PatchCore [6] with WideResNet-50 feature extractor is used. The memory bank subsampling rate is 10%. The pre-training column indicates the feature extractor has been either pre-trained on a dataset or randomly initialized. MandelbulbVAR-21k means MandelbulbVAR-Hybrid-21k. Best, second-best, and third-best scores are shown in underlined bold, bold, and underlined, respectively.

5. Impact of using colored images instead of grayscale ones

To measure the impact of using colored images instead of grayscale ones, we also generate MandelbulbVAR-1k-gray and MandelbulbVAR-Hybrid-21k-gray. To do so, we use the same generative parameters as MandelbulbVAR-1k and MandelbulbVAR-Hybrid-21k, respectively, except that all of the images in MandelbulbVAR-1k-gray and MandelbulbVAR-Hybrid-21k-gray are grayscale, similar to Fig. 3 in the main paper. According to Tab. 4, regarding CNN classification and anomaly detection, MandelbulbVAR-1k outperforms MandelbulbVAR-1k-gray by respective margins of 1.2% (84.5% *vs.* 83.3%) and 1.9% (97.2% *vs.* 95.3%). Regarding ViT classification, MandelbulbVAR-Hybrid-21k outperforms MandelbulbVAR-Hybrid-21k-gray by a gap of 0.1% (88.3% *vs.* 88.2%). These results quantify the gain in CNN and ViT pre-training performance brought by adding colors.

Regarding CNN pre-training, the positive performance gain brought by adding colors was also reported in past studies: [1] found that using fractal images generated with color and backgrounds leads to better pre-training. [3] found that adding colors to their 2D fractals made the pre-training better, compared to using grayscale images. Regarding ViT

Pre-training	Anomaly detection	CNN classification	ViT classification
MandelbulbVAR-1k	**97.2**	**84.5**	-
MandelbulbVAR-1k-gray	95.3	83.3	-
MandelbulbVAR-Hybrid-21k	-	-	**88.3**
MandelbulbVAR-Hybrid-21k-gray	-	-	88.2

Table 4. Comparison between colored and grayscale pre-training datasets. The anomaly detection column reports the average image-level AUROC recorded by PatchCore over MVTec AD. The CNN classification and ViT classification columns show the average accuracy recorded by ResNet50 and ViT-T respectively, over ImageNet-1k, CIFAR-10, CIFAR-100, Flowers and ImageNet-100. The best score is in bold. All metrics are in %.

pre-training, our result contradicts one of the discoveries reported in the work [4]: this study found that grayscale FractalDB images were better than the colored ones concerning ViT pre-training. However, in our study, adding colors benefits less ViT pre-training than CNN pre-training. The performance gain is smaller for ViT than CNN (0.1 points *vs.* 1.2 points regarding average classification accuracy). This result is not surprising, as ViTs have lower texture bias than CNNs [9].

6. Pre-training ResNet-50 for anomaly detection

pre-training	Img. AUROC	Pw. AUROC	PRO
Rand. init.	80.1	89.0	63.2
ImageNet-1k	<u>**99.0**</u>	<u>**98.1**</u>	<u>**93.1**</u>
ExFractalDB-1k	86.5	92.2	71.7
RCDB-1k	78.3	88.5	66.6
VisualAtom-1k	92.2	93.2	77.8
FractalDB-1k*	86.1	92.9	73.4
FractalDB-10k*	80.3	91.8	69.1
MandelbulbVAR-1k	**96.7**	**96.8**	**89.3**

Table 5. Anomaly detection performance (Average image-level AUROC, pixel-wise AUROC and PRO) on MVTec AD [2]. The PatchCore algorithm [6] with ResNet-50 feature extractor is used. The memory bank subsampling rate is 10%. The pre-training column indicates the feature extractor has been either pre-trained on a dataset or randomly initialized. Models with * are downloaded from the project page of the work [3]. Best, and second-best scores are shown in underlined bold, and bold, respectively.

Tab. 5 compares the anomaly detection performance on MVTec AD of PatchCore algorithms relying on ResNet-50 feature extractors. These networks are pre-trained on different datasets. We obtain the same conclusion as the main paper. First, in terms of each performance metric, PatchCore used along with the feature extractor pre-trained on MandelbulbVAR-1k performs the second best right after the algorithm based on the ImageNet-1k pre-training. The performance gaps between them are relatively low (2.3, 1.3 and 3.8 points in image-level AUROC, pixel-wise AUROC and PRO, respectively). Second, pre-training on our proposed dataset outperforms existing FDSL methods. Among them, the one based on VisualAtom-1k performs the best. But between the latter and ours, there are gaps of 4.5, 3.6 and 11.5 points in image-level AUROC, pixel-wise AUROC and PRO, respectively.

References

[1] Connor Anderson and Ryan Farrell. Improving fractal pre-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1300–1309, 2022. 2

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 2, 3

[3] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. *International Journal on Computer Vision (IJCV)*, 2022. 1, 2, 3

[4] Kodai Nakashima, Hirokatsu Kataoka, Asato Matsumoto, Kenji Iwata, Nakamasa Inoue, and Yutaka Satoh. Can vision transformers learn without natural images? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1990–1998, 2022. 3

[5] Randi J Rost, Bill Licea-Kane, Dan Ginsburg, John Kessenich, Barthold Lichtenbelt, Hugh Malan, and Mike Weiblen. *OpenGL shading language*. Pearson Education, 2009. 1

[6] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 1, 2, 3

[7] Sora Takashima, Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, and Rio Yokota. Visual atoms: Pre-training vision transformers with sinusoidal waves. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18579–18588, 2023. 1

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco

Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[9] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 2, 3