Pre-training Vision Models with Mandelbulb Variations

Benjamin Naoto Chiche, Yuto Horikawa, Ryo Fujita
Rist Inc.
830 Hongakujimae-cho, Kawaramachi Nishi-iru, Gojo-dori, Shimogyo-ku, Kyoto 600-8102 Japan
https://github.com/RistoranteRist/MandelbulbVariationsGenerator

Abstract

The use of models that have been pre-trained on natural image datasets like ImageNet may face some limitations. First, this use may be restricted due to copyright and license on the training images, and privacy laws. Second, these datasets and models may incorporate societal and ethical biases. Formula-driven supervised learning (FDSL) enables model pre-training to circumvent these issues. This consists of generating a synthetic image dataset based on mathematical formulae and pre-training the model on it.

In this work, we propose novel FDSL datasets based on Mandelbulb Variations. These datasets contain RGB images that are projections of colored objects deriving from the 3D Mandelbulb fractal. Pre-training ResNet-50 on one of our proposed datasets MandelbulbVAR-1k enables an average top-1 accuracy over target classification datasets that is at least 1% higher than pre-training on existing FDSL datasets. With regard to anomaly detection on MVTec AD, pre-training the WideResNet-50 backbone on MandelbulbVAR-1k enables PatchCore to achieve 97.2% average image-level AUROC. This is only 1.9% lower than pre-training on ImageNet-1k (99.1%) and 4.5% higher than pre-training on the second-best performing FDSL dataset i.e. *VisualAtom-1k (92.7%). Regarding Vision Transformer (ViT) pre-training, another dataset that we propose and coin MandelbulbVAR-Hybrid-21k enables ViT-Base to achieve 82.2% top-1 accuracy on ImageNet-1k, which is 0.4% higher than pre-training on ImageNet-21k (81.8%) and only 0.1% lower than pre-training on VisualAtom-1k (82.3%).*

1. Introduction

Formula-driven supervised learning (FDSL) to pre-train computer vision models such as Convolutional Neural Networks (CNNs) [12] or Vision Transformers (ViTs) [11, 17, 23, 24] has recently gained interest within the industry and the research community. This consists of the following steps. First, an artificial dataset of labeled images is generated based on mathematical formulae and rendering software. Second, a model undergoes the supervised learning over this labeled dataset. The trained model is then deployed to solve a target (synonym for downstream) task on a real-image dataset. In this transfer learning phase, the model is generally fine-tuned on the target dataset, if possible. FDSL presents remarkable advantages: first, its annotation is free of errors and cost since labels are automatically generated. Second, the synthesized dataset does not contain any societal biases that can be harmful from an ethical point of view. Third, it can replace other pre-training datasets that either are not publicly available (*e.g.* Instagram-3.5B [14] and JFT-300M/3B [29]) or present license, copyright or privacy issues. For example, as ImageNet [8] is only available for non-commercial use, ImageNet pre-trained models—that are largely available online today—are not really licensed for commercial use. Also, this dataset contains some images that either violate privacy protection laws or are not cleared in terms of copyright. Thus, using ImageNet pre-trained models can be legally not suitable in some countries.

Figure 1. Example MandelbulbVAR-1k images. 10 classes are randomly chosen and each image is an instance of one of them.

In this context, several FDSL datasets have been proposed. One of them is meant to pre-train CNNs [12], others are reportedly suitable to pre-train ViTs [11, 17, 24]. These datasets are classification datasets and are effective for downstream supervised classifications, that involve fine-tuning on target datasets. Yet, there is a small number of datasets that are effective in pre-training both ViTs and CNNs. Indeed, only FractalDB has been shown to be ef-

fective in both CNN and ViT pre-trainings, in two separate studies [12, 17]. Moreover, none of the existing FDSL studies verified whether their proposed datasets can be used in order to pre-train CNNs with a view to freezing them after pre-training and using them as off-the-shelf feature extractors. Indeed, such CNNs—if having been pre-trained on a large and diverse enough image dataset like ImageNet—have proven to be highly useful in *cold-start* anomaly detection [3, 6, 7, 21]. This task consists of having only a small number of normal instances as training data and detecting outliers in the test data. Such a task is frequently dealt with in industrial scenarios when it is easy to collect normal data but hard to collect abnormal ones. As previously stated, ImageNet pre-training of such feature extractors can be legally problematic. Therefore, an FDSL dataset that can effectively pre-train these feature extractors can be a great asset for the industry. To summarize, there is no FDSL dataset construction scheme that is effective in pre-training all of the following models at once: CNNs and ViTs for downstream supervised classifications and CNNs for downstream anomaly detection. Finally, existing FDSL datasets are images of two-dimensional (2D) mathematical objects and/or exclusively contain grayscale images. As target datasets often contain RGB images that are projections of real-life objects that are colored and three-dimensional (3D), we think that it is preferable to model 3D objects that are colored and project them onto RGB images. Moreover, in the 3D space, we can simulate external light sources and alter the color of the objects based on their light exposure. This process, called shading in computer graphics, would enable realistic depth perception and variation of darkness level, which would contribute to the pre-training.

Given these observations, our study proposes new FDSL datasets based on Mandelbulb Variations [19]. These datasets contain RGB images that are projections of 3D fractals augmented with colors and shading. Fig. 1 visualizes some instances of one of our proposed datasets, **MandelbulbVAR-1k**. The rest of this article is structured in the following way. First, we review related studies. Then, we explain the mathematical theory behind our proposed datasets. Finally, we empirically evaluate them and demonstrate their effectiveness.

2. Related work

2.1. FDSL

The work [12] introduced FDSL to pre-train CNNs for the first time. It proposed FractalDB, a dataset of either colored or grayscale images of 2D fractals generated based on an iterated function system. After this, the same dataset was used in [17] to pre-train ViTs. This study used FDSL to pre-train ViTs for the first time. The authors of this paper intuited that FDSL enables effective pre-training of a ViT thanks to its self-attention mechanism that enables it to focus on the formula-based patterns without considering background areas. They additionally showed that this mechanism makes the models focus on the contours *i.e*. outlines, rather than the textures. This discovery was in accordance with one of the conclusions from [25] stating that compared to CNNs, ViTs are more biased towards shape than texture. Motivated by these observations, [11] proposed ExFractalDB and Radial Contour DataBase (RCDB). The former is formed by grayscale images that are projections of 3D fractals. The latter consisted of grayscale images of 2D contours and enabled a better ViT-Base pre-training than ImageNet-21k when fine-tuned on ImageNet-1k. The work [24] represents an improvement in this direction by proposing VisualAtom, a dataset of grayscale images of 2D contours with a larger design space. This dataset showed even better pre-training performance than RCDB when considering ViTs.

Inspired by some of the previously mentioned studies, the following works proposed FDSL datasets for specific downstream tasks, other than classification: [23] proposed a dataset of 2D contours that is suitable for semantic segmentation ViT pre-training; [27] proposed a point cloud fractal database for 3D object detection. Besides, authors of [1] improved the CNN pre-training of [12] by proposing to generate the 2D fractal images on-the-fly at training time.

2.2. Anomaly detection

Studies such as [3, 6, 7, 21] showed that using ImageNet pre-trained and frozen CNNs to extract features from images was an effective approach in *cold-start* anomaly detection. This task consists in having only a small number of normal examples as training data and detecting abnormal ones in the test data. The paper [3] was the first study that showed that kNN combined with ImageNet pre-trained feature extractor was effective in anomaly detection. The study [6] proposed SPADE, an algorithm that uses memory banks containing several ImageNet-feature hierarchies to do kNN-based anomaly segmentation and image-level anomaly detection. Authors of [7] proposed PaDiM, which estimates statistics of patch-level ImageNet-features (mean and covariance) and considers patch-level Mahalanobis distances. Finally, [21] proposed PatchCore, which uses a memory bank of neighborhood-aware patch-level ImageNet features. Its coreset subsampling mechanism enables reduced inference cost while keeping high performance. As stated in Sec. 1, cold-start anomaly detection is frequently dealt with in an industrial context, but ImageNet pre-training may face some limitations. Having a dataset other than ImageNet that is licensed and free of privacy/copyright issues, labeling costs/errors and ethical biases—like an FDSL one—would be greatly beneficial to industrial anomaly detection applications.

To conclude this section, we observe that most of the existing FDSL datasets were either tailored to pre-train CNNs or ViTs. Moreover, their capabilities to pre-train CNNs to use them as off-the-shelf feature extractors for cold-start anomaly detection have not been evaluated. Finally, as target data are often colored images that are projections of 3D real-life objects, we think that differently from existing works, we should model 3D mathematical objects, color them and project them onto RGB images via virtual cameras. On top of this, shading would benefit the pre-training, enabling realistic depth perception and variation of darkness level. Considering these points, our contributions are the following.

- We propose novel FDSL datasets that are free of copyright/privacy issues, ethical biases and labeling costs/errors. They contain RGB images that are projections of Mandelbulb Variations [19], augmented with colorful, light and shaded areas. We explain the mathematical theory behind these 3D fractals. We implement our original fractal modeling and rendering software based on OpenGL Shading Language (GLSL) [20] and we make it publicly available on our GitHub repository, with a permissive license that allows commercial use.
- We evaluate the pre-training performance of our proposed datasets. For comparison purposes, we also evaluate existing FDSL datasets. We demonstrate the following. First, one of our proposed datasets **MandelbulbVAR-1k** is on average the best when pre-training CNNs for the downstream classification and anomaly detection tasks. Our dataset approaches the most ImageNet in pre-training performance. Second, regarding ViT pre-training, compared to existing FDSL datasets, another dataset that we propose and name **MandelbulbVAR-Hybrid-21k** performs the second-best when considering ImageNet-1k fine-tuning and average performance over CIFAR-10/100, ImageNet-100 and Flowers.

3. Dataset generation method

3.1. Motivation

Instead of relying on FDSL, simulating 3D scenes of real-life objects can also produce datasets that are free of copyright/privacy issues, ethical biases and labeling costs/errors. However, increasing the diversity (*i.e.* the number of classes) of such datasets requires coming up with a lot (thousand) of different 3D models, which is costly. In contrast, with FDSL we can increase the diversity by simply changing the parameters used in the generative formula. This is one of the motivations behind our FDSL process. Moreover, as target datasets often contain RGB images of real-life objects with colors, light and shaded areas, we seek to model a parametrized 3D mathematical object and augment it with colors and shading. Therefore, we propose new datasets that derive from the 3D Mandelbulb fractal [26]. This fractal extends the 2D Mandelbrot set—a typical example of a 2D fractal[15, 16]—to the 3D space. As fractals are objects in which the smaller portions are similar to those on a larger scale, they present infinite levels of smaller detail. Displaying it is only limited by computer capability. Therefore, their rendering can result in images with an important amount of information. If used for pre-training, they can hopefully make the model learn features that are useful when dealing with target tasks.

3.2. Mandelbulb Variations

Let $n \in \mathbb{N}$ and the following function $g_n : \mathbb{R}^3 \to \mathbb{R}^3$:

$$g_n : \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto r^n \begin{pmatrix} \sin(n\theta)\cos(n\varphi) \\ \sin(n\theta)\sin(n\varphi) \\ \cos(n\theta) \end{pmatrix} \tag{1}$$

where

$$\begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \theta = \arccos(z/r) \\ \varphi = \operatorname{atan2}(y, x) \end{cases} \tag{2}$$

A 3D Mandelbulb $\mathcal{M}_n$ is defined as the set of data points $\boldsymbol{c} \in \mathbb{R}^3$ for which the sequence $(\boldsymbol{v}_k)$ defined by $\boldsymbol{v}_{k+1} = g_n(\boldsymbol{v}_k) + \boldsymbol{c}$ does not diverge *i.e.* $\sup(||\boldsymbol{v}_k||) < +\infty$ when starting at $\boldsymbol{v}_0 = \mathbf{0}$ [26].

This definition of 3D Mandelbulb has only a parameter n. For our FDSL approach, we want a 3D model with more parameters, to increase the diversity of the generated dataset. Therefore, we rely on Mandelbulb Variation $\mathcal{V}_{(n,b)}$. Its definition [19] derives from the one of 3D Mandelbulb presented above. It relies on a new function $f_{(n,b)} : \mathbb{R}^3 \to \mathbb{R}^3$ parametrized by $b = (b_1, ..., b_9) \in \{0, 1\}^9$. b is equivalent to a boolean vector. We use the following definition:

$$\mathcal{V}_{(n,b)} = \left\{ \boldsymbol{c} \in \mathbb{R}^3 \;\middle|\; \begin{array}{c} \sup_k(\|\boldsymbol{v}_k\|) < +\infty \\ \boldsymbol{v}_{k+1} = f_{(n,b)}(\boldsymbol{v}_k) + \boldsymbol{c} \\ \boldsymbol{v}_0 = \mathbf{0} \end{array} \right\} \tag{3}$$

$$f_{(n,b)} : \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto r^n \begin{pmatrix} s_1 h_1(n\omega_1) \\ s_2 h_2(n\omega_2) \\ s_3 h_3(n\omega_3) \end{pmatrix} \tag{4}$$

where

$$\begin{cases} s_i = 1 - 2b_i \\ \omega_i = b_{i+3}\theta + (1 - b_{i+3})\varphi \\ h_i = b_{i+6}\sin + (1 - b_{i+6})\cos \\ r, \theta, \varphi \text{ are defined by Eq. (2).} \end{cases} \tag{5}$$

The vector b being of length 9, one would think the maximum number of different Mandelbulb Variations is $2^9 = 512$ for a given n. However, some vectors b result in

sets that either are not bounded or have inconvenient shapes. Therefore, we do not use these vectors. Then we apply additional selection and retain 61 vectors *i.e.* variations. Finally, by letting n take the 17 values in $[\![2, 18]\!]$ we obtain $61 \times 17 = 1037$ variations. We associate each of them with a different class label, resulting in 1037 classes.

We implement our original Mandelbulb Variation modeling and rendering software based on OpenGL Shading Language (GLSL) [20]. We make it publicly available on our GitHub repository. Using this software, we first simulate a Mandelbulb Variation model $\mathcal{V}_{(n,b)}$ by entering the input parameters n and b. We color this variation by randomly generating a coloring pattern and using an orbit trap algorithm [16], and an external light source is simulated to determine the surface brightness via shading. Then, we simulate a virtual camera with a random location that points toward the variation. Through the rendering process, the fractal is projected onto an RGB image of size 512×512 via the camera. Given a Mandelbulb Variation model $\mathcal{V}_{(n,b)}$, we repeat the above coloring, shading and rendering process 1000 times, which results in 1000 images of the variation having different colors and viewing angles. All of these images are labeled with the same class label corresponding to the variation. By repeating this process for all of the 1037 Mandelbulb Variations, we obtain a dataset of 1037 classes and 1000 images per class. This dataset has therefore around 1M images in total. We coin it **MandelbulbVAR-1k**. Fig. 2 shows 5 instances belonging to a same class of this dataset. Using an NVIDIA GeForce RTX 3090 GPU, it takes 38 hours (about 1.5 days) to generate this dataset.

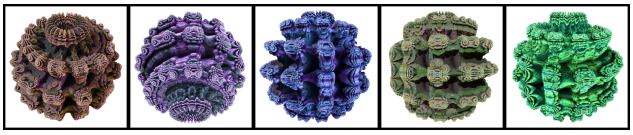


Figure 2. 5 instances of a class of MandelbulbVAR-1k. Better seen in color and zoomed in.

3.3. Hybrid Mandelbulb Variations

To further increase the fractal shape diversity, we model Hybrid Mandelbulb Variations. A Hybrid Mandelbulb Variation $\mathcal{V}_{(n,b),(n',b')}$ is generated by combining two Mandelbulb Variations $\mathcal{V}_{(n,b)}$ and $\mathcal{V}_{(n',b')}$ [16]:

$$\mathcal{V}_{(n,b),(n',b')} = \left\{ \boldsymbol{c} \in \mathbb{R}^3 \;\middle|\; \begin{array}{c} \sup_k(\|\boldsymbol{v}_k\|) < +\infty \\ \boldsymbol{u}_{k+1} = f_{(n,b)}(\boldsymbol{v}_k) + \boldsymbol{c} \\ \boldsymbol{v}_{k+1} = f_{(n',b')}(\boldsymbol{u}_k) + \boldsymbol{c} \\ \boldsymbol{v}_0 = \boldsymbol{0} \end{array} \right\} \quad (6)$$

The hybrid variation presents a shape combining properties of the two input variations. The general shape of the fractal is more similar to the one of $\mathcal{V}_{(n,b)}$ (because $f_{(n,b)}$ is first applied to compute $\boldsymbol{v}_1$). The presence of the second function $f_{(n',b')}$ in the subsequent computations of $(\boldsymbol{v}_k)$ introduces deformations. Shapes of these deformations are somehow similar to the shape of $\mathcal{V}_{(n',b')}$. Fig. 3 illustrates the formation of a Hybrid Mandelbulb Variation.

Figure 3. Illustration of a Hybrid Mandelbulb Variation. (Left) an instance of the Mandelbulb Variation $\mathcal{V}_{(8,u(302))}$. (Center) an instance of $\mathcal{V}_{(7,u(128))}$. (Right) an instance of the hybrid variation $\mathcal{V}_{(8,u(302)),(7,u(128))}$ combining the two variations. These images present the same viewing angle. We show uncolored variations for illustrative purposes. The bijection $u : [\![0, 511]\!] \to \{0, 1\}^9, u^{-1} : (b_1, \ldots, b_9) \mapsto \sum_i b_i 2^{i-1}$ converts the input integer into a vector containing its 9-bit binary representation. The general shape of the hybrid fractal is more similar to the one of $\mathcal{V}_{(8,u(302))}$ while the hybrid fractal presents deformations. Shapes of these deformations are somehow similar to the shape of $\mathcal{V}_{(7,u(128))}$.

From the 1037 Mandelbulb Variations that were used in order to generate MandelbulbVAR-1k, we combine around 20k pairs of them to define 20k hybrid variations. The 1037 Mandelbulb Variations and the 20k hybrid ones are added together to give a set of 21k variations. These variation models undergo the previously mentioned coloring, shading and rendering processes within our software, which generates a new dataset having 21k classes and 50 images per class. This dataset, again, has therefore around 1M images in total. We coin it **MandelbulbVAR-Hybrid-21k**. Using an NVIDIA GeForce RTX 3090 GPU, it takes 23 hours (about 1 day) to generate this dataset. This duration is similar to the ones needed to generate the existing VisualAtom-1k [24], RCDB-1k and ExFractalDB-1k [11].

4. Experiments

4.1. Datasets, models and implementations

We split our experiments into two parts. First, we evaluate the performance of MandelbulbVAR-1k regarding CNN pre-training when target tasks are supervised classifications and anomaly detection. Second, we evaluate our proposed datasets in terms of ViT pre-training when target tasks are supervised classifications. Each of these parts involves comparisons to existing pre-training datasets—including FDSL ones—and the cases where there is no pre-training. Therefore, we download existing FDSL datasets from the

project pages of the studies [11, 12, 24]. Unless otherwise specified, we also pre-train models on them and evaluate them on target tasks.

In the first part, when target tasks are supervised classifications, we pre-train ResNet-50 backbones [10]. When the target task is anomaly detection, we pre-train WideResNet-50 [28] backbones. In the second part, we pre-train ViT-Tiny (ViT-T) and ViT-Base (ViT-B) [9] backbones. Downstream classification tasks are evaluated on the following datasets: ImageNet-1k (IN1k) [8], CIFAR-10/100 (C10/100) [13], Flowers [18] and ImageNet-100 (IN100) [22]. The last dataset is a subset of ImageNet-1k with 100 classes. The anomaly detection is evaluated on the MVTec AD dataset [4]. The codes we use are detailed in the supplementary material.

4.2. Training procedures

To pre-train CNNs and fine-tune them on supervised classification tasks, we use the momentum stochastic gradient descent (SGD) with a momentum value of 0.9 and a basic batch size of 256. The initial learning rate equals respectively 0.1 and 0.01 at the pre-training and fine-tuning phases. The learning rate is divided by 10 when the training epoch reaches 100 and then again at epoch 200. Both pre-training and fine-tuning are done for 300 epochs. Training images are randomly cropped with the size 224×224. Training hyper-parameters we use to pre-train and fine-tune ViTs are provided in the supplementary material.

When dealing with the downstream anomaly detection task, WideResNet-50 feature extractors are pre-trained based on the following hyperparameters. We use the momentum SGD with a value of 0.9 and an overall batch size of 512. The learning rate is initially set to 0.1 and the weight decay equals 1e-4. The learning rate is divided by 10 when the training epoch reaches 300 and the training is performed up to 600 epochs. Training images are randomly cropped and resized to 224×224.

4.3. Evaluations

Regarding comparisons to existing FDSL datasets, unless otherwise specified, we mainly compare models pre-trained on our proposed datasets against models that are pre-trained using similar numbers of images *i.e*. similar products of number of epochs and total numbers of images in the datasets. Thus, unless otherwise indicated, we mainly compare our FDSL datasets against RCDB-1k, ExFractalDB-1k, FractalDB-1k and VisualAtom-1k. These datasets, like our proposed ones, contain around 1M images each.

Regarding performance metrics, for supervised classification tasks, we employ the top-1 accuracy. For the anomaly detection task, the image-level area under the receiver-operator curve (AUROC) measures the image-level anomaly detection performance. The pixel-wise AUROC and PRO metric measure the segmentation performance. The second metric weights ground-truth regions of different sizes equally, in contrast to the first one [5, 21].

For anomaly detection, we also make qualitative evaluations, by first visualizing anomaly scores outputted by PatchCore algorithms in the form of segmentation images. Second, we also visualize convolutional filters learned on MandelbulbVAR-1k and compare them to the ones learned on ImageNet-1k.

5. Results

5.1. CNN pre-training

For downstream supervised classification tasks. Tab. 1 shows accuracies of ResNet-50 on the validation sets of various datasets. When compared to the models trained from scratch or pre-trained on existing FDSL datasets, on 4 out of 5 target datasets (IN1k, C10, Flowers and IN100) the network pre-trained on MandelbulbVAR-1k performs the best. Furthermore, the performance on C10 and Flowers enabled by our dataset is the closest to the one achieved by ImageNet-1k pre-training. On the remaining C100, MandelbulbVAR-1k is the second-best FDSL dataset. On average, when compared to the models pre-trained on existing FDSL datasets or trained from scratch, the one pre-trained on our dataset performs the best. The gap in average accuracy between ours and the second-best FDSL dataset *i.e*. FractalDB-10k is 1.0% (84.5% *vs*. 83.5%). Since ResNet-50 pre-trained on FractalDB-10k has seen around 3 times more pre-training images than the one pre-trained on our dataset, this result confirms the high pre-training performance of our dataset. Indeed, the former model has been pre-trained for 90 epochs on a dataset that contains around 10 times more images than MandelbulbVAR-1k [12].

For downstream anomaly detection. Tab. 2 compares anomaly detection performance of PatchCore algorithms relying on different WideResNet-50 feature extractors. These networks are either pre-trained on different datasets or initialized with random weights. We observe that in terms of each performance metric, PatchCore used along with WideResNet-50 pre-trained on MandelbulbVAR-1k performs the second best right after the algorithm based on ImageNet-1k pre-training. The gaps between them are only 1.9, 1.3 and 3.8 points in average image-level AUROC (97.2% *vs*. 99.1%), pixel-wise AUROC (96.8% *vs*. 98.1%) and PRO (89.6% *vs*. 93.4%), respectively. Also, 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, the gaps are 4.5, 2.9 and 9.3 points in average image-level AUROC (92.7% *vs*. 97.2%), pixel-wise AUROC (93.9% *vs*. 96.8%) and PRO (80.3% *vs*. 89.6%), respectively.

Fig. 4 shows some segmentation images produced by the

Pre-training	IN1k	C10	C100	Flowers	IN100	Average
From scratch	71.8	88.9	62.1	78.0	72.1	74.6
FractalDB-1k	72.4	93.0	74.3	91.8	79.7	82.2
FractalDB-10k*	72.9	**93.9**	**77.1**	92.7	81.1	**83.5**
VisualAtom-1k	**73.2**	93.5	74.4	93.2	**81.4**	83.1
RCDB-1k	73.0	92.4	73.6	90.3	81.3	82.1
ExFractalDB-1k	72.5	93.0	73.7	93.4	79.2	82.4
MandelbulbVAR-1k (ours)	**73.4**	**93.9**	76.2	**96.6**	**82.3**	**84.5**
ImageNet-1k	-	**97.1**	**84.4**	**98.3**	-	-

Table 1. Top-1 accuracies of ResNet-50 models on the validation sets of various datasets. These models are either trained from scratch or fine-tuned after being pre-trained on different datasets. Bold and underlined values indicate the best scores, and bold values show the second-best scores. The model with * is downloaded from the project page of the work [12]. For the ImageNet-1k pre-trained model, we use the official weight that is available on PyTorch.

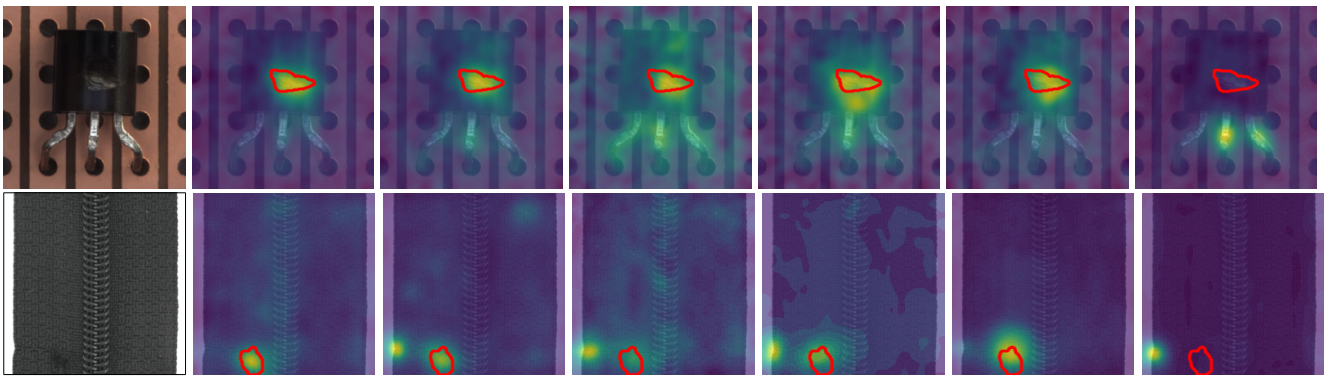
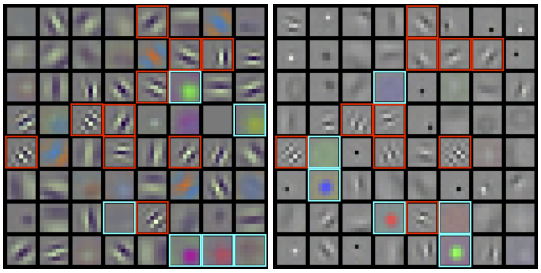


Figure 4. Segmentation images on MVtec AD produced by PatchCore algorithms, merged with the input images. These algorithms use WideResNet-50 feature extractors pre-trained on various datasets (indicated below the second row). First row: *transistor* object and *damaged case* anomaly class. Second row: *zipper* object and *fabric interior* anomaly class. The red boundaries denote the contours of the anomaly regions of the actual GT mask. Best viewed in color and zoomed in.

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**

Table 2. Anomaly detection performance (average image-level AUROC, pixel-wise AUROC and PRO in %) on MVTec AD [4]. PatchCore [21] 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. Best, and second-best scores are shown in underlined bold, and bold, respectively.



(a) ImageNet-1k (b) MandelbulbVAR-1k

Figure 5. Filters of the first convolutional layer of the WideResNet-50 backbones that give the results reported in Tab. 2. These backbones are pre-trained on (a) ImageNet-1k (b) MandelbulbVAR-1k. Each set of filters is normalized by the minimum and maximum values over all of its filters. Some of the Gabor-like and colored Gaussian-like filters are respectively framed in red and cyan. Best viewed in color and zoomed in.

PatchCore algorithms based on the WideResNet-50 backbones. These images visualize the anomaly score at each pixel with the viridis colormap *i.e*. pixels with low scores tend to be colored in blue and the higher the score is, the more the pixel is colored in yellow. The red boundaries denote the anomaly contours of the GT mask. Regarding the input image of *transistor* object, the segmentation image related to our proposed dataset seems to be the most in accordance with the GT mask, for the following reasons. First, the abnormal region is accurately covered with high anomaly scores. Second, it contains fewer false positives *i.e*. green or yellow regions outside of the red boundary than the segmentation images related to other datasets. When the input image is the one related to *zipper* object, PatchCore based on our dataset accurately covers the abnormal region with high anomaly scores. Moreover, contrary to the algorithm based on other datasets, it does not produce false positives at the lower-left border of the cloth *i.e*. it produces relatively low anomaly scores at this part of the image.

Discussion. A drawback of our proposed and existing FDSL datasets is that they do not contain any semantic content. As a result, they are potentially not good for semantic-related representation learning. On the one hand, we think that Mandelbulb Variation images are good for learning low-level statistics *i.e*. the first layer filters. This is shown by Fig. 5, which shows that the set of filters learned on our dataset is close to the one based on ImageNet-1k, presenting similar patterns of filters (Gabor-like and colored Gaussian-like filters) and similar gray-colored backgrounds. On the other hand, the high anomaly detection performance also shows that Mandelbulb Variation images are also good for learning mid-level feature representations. Indeed, these features (the outputs of the 2nd and 3rd blocks of WideResNet-50) are used in PatchCore [21].

As CNNs are more biased toward texture than shape [2, 25], the high CNN pre-training performance of MandelbulbVAR-1k suggests that this dataset is good for learning texture. This is indeed in accordance with the high anomaly detection performance, because a significant proportion of anomalies in MVTec AD occurs at the texture level.

5.2. ViT pre-training

Results on ImageNet-1K. Tab. 3 shows accuracies of ViT-T/B models on the validation set of ImageNet-1K. We first observe that regarding ViT-T, pre-trainings on MandelbulbVAR-1k and MandelbulbVAR-Hybrid-21k enable the same 73.8% top-1 accuracy. Regarding ViT-B, pre-training on MandelbulbVAR-Hybrid-21k enables a better performance than MandelbulbVAR-1k (82.2% *vs*. 82.1%; 0.1 points higher performance). These datasets contain similar overall numbers of images but differ in shape diversity thus in the number of classes. This result may suggest that the shape diversity of MandelbulbVAR-Hybrid-21k benefits the ViT pre-training.

Pre-training	ViT-T	ViT-B
Scratch	72.6	79.8
ImageNet-21k	**74.1**	81.8
ExFractalDB-1k	73.7	80.4
RCDB-1k	73.1	**82.3**
VisualAtom-1k	**74.2**	**82.3**
MandelbulbVAR-1k (ours)	73.8	82.1
MandelbulbVAR-Hybrid-21k (ours)	73.8	**82.2**

Table 3. Accuracy of ViT-T/B on the validation set of ImageNet-1k. The pre-training column indicates whether the models are trained from scratch or fine-tuned after being pre-trained on various datasets. Best, second-best, and third-best scores are shown in underlined bold, bold, and underlined, respectively. Results reported on non-gray rows are taken from [24].

As expected, ViT-T and ViT-B pre-trained on our proposed datasets outperform the ones that are trained from scratch. These performance gaps are 1.2 points for ViT-T and at least 2.3 points for ViT-B. More remarkably, ViT-B models pre-trained on MandelbulbVAR-1k (82.1%) and MandelbulbVAR-Hybrid-21k (82.2%) outperform the same model pre-trained on ImageNet-21k (81.8%) by 0.3 and 0.4 point gaps, respectively. This shows the high ViT-B pre-training performance of our datasets.

Compared to existing FDSL datasets, our datasets are better than ExFractalDB-1k for both ViT-T and ViT-B (by respective margins of 0.1 and at least 1.7 points). ViT-T pre-trained on our dataset performs better than the one pre-trained on RCDB-1k, by a 0.7 point gap. Regarding the pre-training of ViT-B, MandelbulbVAR-Hybrid-21k enables slightly worse performance than the top-performing RCDB-1k and VisualAtom-1k (only 0.1 point difference between 82.2% and 82.3%).

Results on CIFAR-10/100, Flowers and IN100. Tab. 4 shows accuracies of ViT-T/B on the validation sets of CIFAR-10/100, IN100 and Flowers. By first comparing the pre-training performance of MandelbulbVAR-1k and MandelbulbVAR-Hybrid-21k, the previously formulated suggestion is confirmed: the shape diversity of MandelbulbVAR-Hybrid-21k enables better pre-training performance regarding ViTs, even with a similar number of images in the dataset. When ViT-T is pre-trained, for all target datasets MandelbulbVAR-Hybrid-21k is better than MandelbulbVAR-1k. When ViT-B is pre-trained, the same tendency is observed except for IN100. The difference in average performance between ViT-T models pre-trained on these datasets is 0.7 points (91.9% *vs*. 91.2%). For ViT-B, this difference is 0.3 points (92.0% *vs*. 91.7%).

Again, the ViT-T models pre-trained on our datasets

Model	Pre-training	C10	C100	Flowers	IN100	Average
ViT-T	Scratch	78.3	57.7	77.1	75.2	72.1
	ImageNet-1k	**98.0**	**85.5**	**99.4**	**88.5**	**92.9**
	FractalDB-1k	96.8	81.6	98.3	86.4	90.8
	ExFractalDB-1k	97.2	81.8	**98.9**	87.4	91.3
	RCDB-1k	97.0	82.2	**98.9**	87.5	91.4
	VisualAtom-1k	**97.6**	**84.9**	**98.9**	**87.8**	**92.3**
	MandelbulbVAR-1k (ours)	97.2	81.7	98.7	87.1	91.2
	MandelbulbVAR-Hybrid-21k (ours)	97.3	83.8	**98.9**	87.5	91.9
ViT-B	RCDB-21k	96.8	82.9	**99.0**	**88.1**	91.7
	VisualAtom-21k	97.7	**86.7**	**99.0**	**88.6**	**93.0**
	MandelbulbVAR-1k (ours)	**97.8**	82.5	98.8	87.6	91.7
	MandelbulbVAR-Hybrid-21k (ours)	**98.2**	**83.6**	**98.9**	87.4	**92.0**

Table 4. Accuracy of ViT-T/B on CIFAR-10/100, Flowers and IN100. The pre-training column indicates whether the models are trained from scratch or fine-tuned after being pre-trained on various datasets. Best, second-best, and third-best scores are shown in underlined bold, bold, and underlined, respectively. Results reported on non-gray rows are taken from [24], except for IN100. For this dataset, models are fine-tuned by ourselves since we are not sure that we are using the same subset of ImageNet-1k as [24].

outperform the ones trained from scratch. Compared to existing FDSL datasets, regarding both ViT-T and ViT-B, MandelbulbVAR-Hybrid-21k enables the second-best average performance after VisualAtom-1k and VisualAtom-21k, respectively. Related performance gaps are 0.4% for ViT-T (91.9% *vs*. 92.3%) and 1.0% for ViT-B (92.0% *vs*. 93.0%). The latter gap is partly explained by the following fact: ViT-B pre-trained on VisualAtom-21k has seen around 6 times more pre-training images than the one pre-trained on our dataset. Indeed, the former model has been pre-trained for 90 epochs on a dataset that contains around 20 times more images than MandelbulbVAR-Hybrid-21k [11, 24]. Furthermore, ViT-T pre-trained on MandelbulbVAR-Hybrid-21k presents an average performance that is 0.5 points higher than the third-best average performance enabled by RCDB-1k (91.9% *vs*. 91.4%). ViT-B pre-trained on MandelbulbVAR-Hybrid-21k shows an average performance that is 0.3 points above the third-best average performance enabled by RCDB-21k (92.0% *vs*. 91.7%). This shows the power of our dataset because again, ViT-B pre-trained on RCDB-21k has seen around 6 times more pre-training images than the one pre-trained on our dataset, for the same reason as described previously.

Discussion. In contrast with CNN-pretraining, regarding ViT-pretraining, MandelbulbVAR-1k does not outperform, on average, the datasets composed of images of contours, namely RCDB and VisualAtom. Making ViTs learn contours *i.e*. shapes is indeed a good strategy when pre-training. As ViTs have higher shape bias than CNNs [25], this result is not surprising. However, Hybrid Mandelbulb Variations make MandelbulbVAR-Hybrid-21k contain more diverse fractal shapes than MandelbulbVAR-1k. This is why MandelbulbVAR-Hybrid-21k performs better than MandelbulbVAR-1k when pre-training ViTs. At the end of the day, MandelbulbVAR-Hybrid-21k performs better than RCDB and worse than VisualAtom when considering the average ViT-T/B pre-training performance over CIFAR-10/100, Flowers and IN100.

6. Conclusion

We proposed new FDSL datasets containing colored images that are projections of Mandelbulb Variations. These are 3D fractals augmented with colors and shading. We made their rendering codes publicly available, with a license that allows commercial use. Contrary to many existing natural image datasets, their annotation is free of error and cost, they do not contain any societal and ethical biases, and they are free of privacy, copyright and license issues. Regarding CNN pre-training, one of our proposed datasets MandelbulbVAR-1k outperformed existing FDSL datasets and approached the most ImageNet-1k when target tasks were supervised classifications (at least 1% higher in average accuracy than existing FDSL datasets) and anomaly detection (for instance only 1.9% lower image-level AUROC than ImageNet pre-training). Regarding ViT pre-training, compared to existing FDSL datasets, another dataset that we proposed and named MandelbulbVAR-Hybrid-21k performed the second-best when considering both ImageNet-1k fine-tuning and average performance over CIFAR-10/100, Flowers and IN100. Notably, ViT-B pre-trained on our dataset recorded 82.2% accuracy on ImageNet-1k, which was 0.4% higher than pre-training on ImageNet-21k and only 0.1% lower than pre-training on the best-performing existing FDSL dataset.

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] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol*, 14 (12):e1006613, 2018. 7

[3] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 2

[4] 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. 5, 6

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 5

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5

[9] 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. In *International Conference on Learning Representations*, 2021. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima, Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, and Rio Yokota. Replacing labeled real-image datasets with auto-generated contours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21232–21241, 2022. 1, 2, 4, 5, 8

[12] 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, 5, 6

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[14] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 1

[15] Benoit B Mandelbrot and Benoit B Mandelbrot. *The fractal geometry of nature*. WH freeman New York, 1982. 3

[16] Krzysztof Marczak, Graeme McLarekin, Jennen Sebastian, Sink Stephen, and Pancoast Robert. Mandelbulber user manual version 2.24.0.0, 2020. 3, 4

[17] 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. 1, 2

[18] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[19] Jason Rampe. New mandelbulb variations. https://softologyblog.wordpress.com/2011/07/21/new-mandelbulb-variations/, 2011. Accessed: 2022-11-01. 2, 3

[20] Randi J Rost, Bill Licea-Kane, Dan Ginsburg, John Kessenich, Barthold Lichtenbelt, Hugh Malan, and Mike Weiblen. *OpenGL shading language*. Pearson Education, 2009. 3, 4

[21] 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. 2, 5, 6, 7

[22] Ambesh Shekhar. Imagenet100. https://www.kaggle.com/datasets/ambityga/imagenet100, 2021. Accessed: 2023-09-01. 5

[23] Risa Shinoda, Ryo Hayamizu, Kodai Nakashima, Nakamasa Inoue, Rio Yokota, and Hirokatsu Kataoka. Segrcdb: Semantic segmentation via formula-driven supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20054–20063, 2023. 1, 2

[24] 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, 2, 4, 5, 7, 8

[25] 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, 7, 8

[26] Daniel White. The unravelling of the real 3d mandelbrot fractal. https://www.skytopia.com/project/fractal/2mandelbulb.html, 2009. Accessed: 2022-11-01. 3

[27] Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, and Tetsuya Ogata. Point cloud pre-training with natural 3d structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21283–21293, 2022. 2

[28] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5

[29] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 1