Can Vision Transformers Learn without Natural Images?

Kodai Nakashima[1,2]*, Hirokatsu Kataoka[1]*, Asato Matsumoto[1,2],
Kenji Iwata[1], Nakamasa Inoue[3] and Yutaka Satoh[1,2]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]University of Tsukuba, Japan
[3]Tokyo Institute of Technology, Japan
{nakashima.kodai, hirokatsu.kataoka, matsumoto-a, kenji.iwata, yu.satou}@aist.go.jp, inoue@c.titech.ac.jp

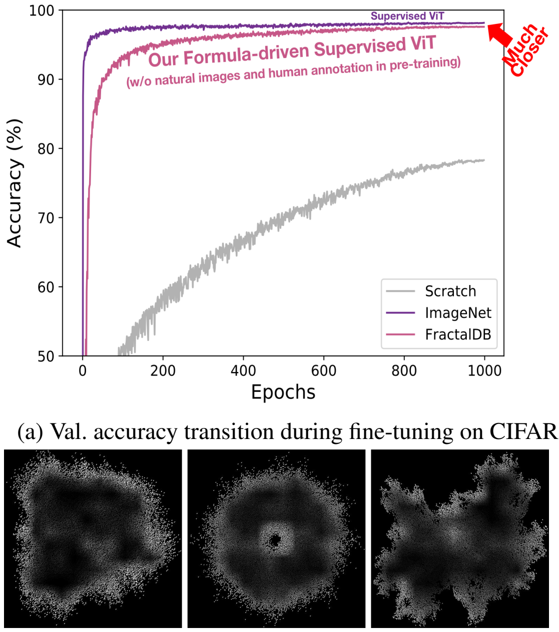
Abstract

Is it possible to complete Vision Transformer (ViT) pre-training without natural images and human-annotated labels? This question has become increasingly relevant in recent months because while current ViT pre-training tends to rely heavily on a large number of natural images and human-annotated labels, the recent use of natural images has resulted in problems related to privacy violation, inadequate fairness protection, and the need for labor-intensive annotations. In this paper, we experimentally verify that the results of formula-driven supervised learning (FDSL) framework are comparable with, and can even partially outperform, sophisticated self-supervised learning (SSL) methods like SimCLRv2 and MoCov2 without using any natural images in the pre-training phase. We also consider ways to reorganize FractalDB generation based on our tentative conclusion that there is room for configuration improvements in the iterated function system (IFS) parameter settings of such databases. Moreover, we show that while ViTs pre-trained without natural images produce visualizations that are somewhat different from ImageNet pre-trained ViTs, they can still interpret natural image datasets to a large extent. Finally, in experiments using the CIFAR-10 dataset, we show that our model achieved a performance rate of 97.8, which is comparable to the rate of 97.4 achieved with SimCLRv2 and 98.0 achieved with ImageNet.

1 Introduction

In contemporary visual recognition, transformer architectures (Vaswani et al. 2017) are gradually replacing the usage of convolutional neural networks (CNNs), which has long been dominated in the field of computer vision (CV).

Transformer architectures, which are based on self-attention mechanisms, were initially employed in natural language processing (NLP) tasks such as machine translation and semantic analysis. Recently, however, we have witnessed the development of epoch-making methods that use transformer modules such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) and Generative pre-trained Transformers (GPT)-{1, 2, 3} (Radford et al. 2018a,b; Brown and et al. 2020). The trend is now gradually shifting from NLP to CV tasks, where one of the most active topics is the use of Vision Transformers (ViTs) for image classification (Dosovitskiy et al. 2021), primarily because ViTs can effectively process and recognize an image based on Transformers with minimum modifications. Additionally, even though the reimplementation process is reasonably straightforward, it has been shown that ViTs often perform at least as well as state-of-the-art transfer learning. However, it is also true that ViT tends to require a large amount of data in the pre-training phase. For example, Dosovitskiy *et al.* (Dosovitskiy et al. 2021) reported



(a) Val. accuracy transition during fine-tuning on CIFAR-10.

(b) Attention maps in fractal images. The brighter areas are those receiving more attention.

Figure 1: The impact of FDSL ViT pre-training on a FractalDB. These results show that (a) the pre-trained model facilitates fine-tuning accuracy improvements to a level that is much closer to an ImageNet pre-trained ViT, and (b) the FractalDB pre-trained ViT recognizes fractal images while identifying complex contours. We believe that this property contributes to the pre-training effect even though fractal images do not have background areas.

*These authors contributed equally.

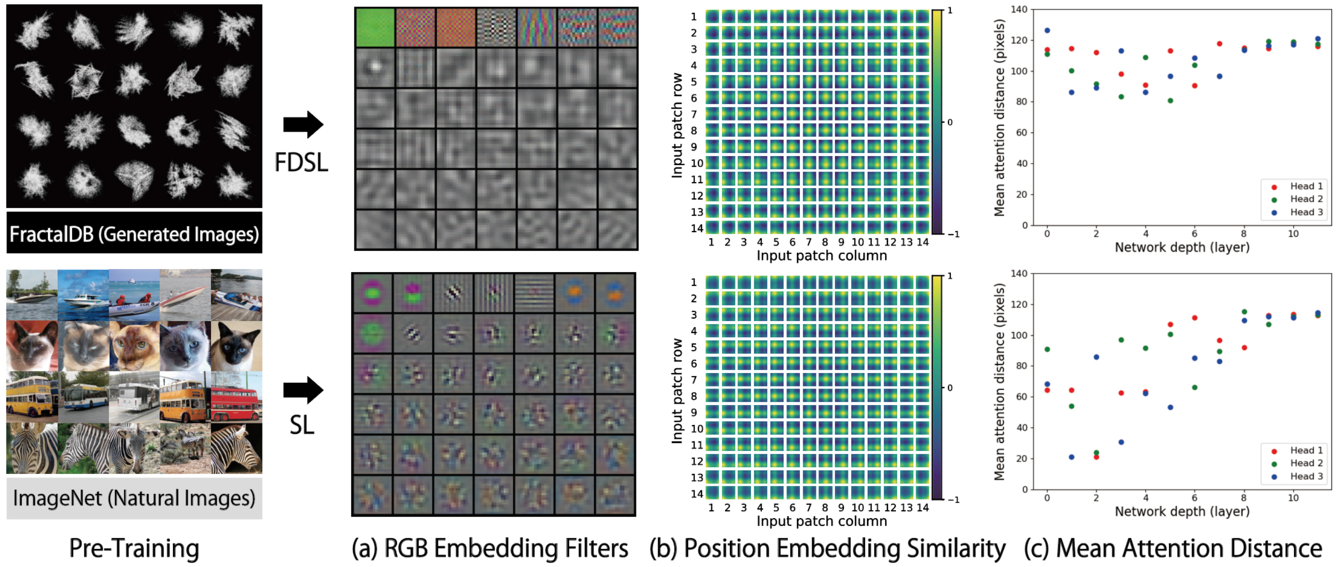


Figure 2: Following (Dosovitskiy et al. 2021), we list the (a) RGB embedding filters, (b) position embedding similarity, and (c) mean attention distance in the frameworks of formula-driven supervised learning (FDSL) with FractalDB and supervised learning (SL) with human-annotated ImageNet. When compared to SL, we found that the FDSL pre-trained ViT enables the acquisition of slightly different filters, the same position embedding, and a wider receptive field. Of particular note, in the mean attention comparison with SL, we found that FDSL enables us to acquire a wider receptive field from the early layers.

that unless a ViT architecture is pre-trained with more than 100 million images, its accuracy is inferior to a CNN. This pre-training problem can be somewhat alleviated by Data-efficient image Transformer (DeiT) (Touvron et al. 2021).

On the other hand, the use of large-scale image datasets may be problematic from the perspective of privacy preservation, annotation labor, and artificial intelligence (AI) ethics. In fact, the use of representative datasets that include natural images taken by camera is currently restricted to academic or educational usage. This problem cannot be resolved by using self-supervised learning (SSL) for automatic natural image labeling because training on natural image datasets will still raise privacy-violation and fairness-protection concerns. In relation to this, other large-scale datasets (e.g., human-related images in ImageNet (Yang et al. 2020) and 80M Tiny Images (Torralba, Fergus, and Freeman 2008)[1]) are now being deleted due to AI ethic issues. To date, other huge datasets such as JFT-300M (Sun et al. 2017) and Instagram-3.5B (Mahajan et al. 2018) are not publicly available. Since these dataset-related problems significantly limit research opportunities in this field, the research community must carefully consider the use of large-scale datasets in terms of availability and reliability while working to overcome them. However, if a method could be developed to train a ViT without using natural images, we believe that the result would have an impact similar to previously proposed models such as BERT, GPT, and ViT. This is because, even though a transformer is used to compose the core modules in each of these models, the resulting insights obtained to date have led to significant research community advancements.

In this context, the formula-driven supervised learning (FDSL) concept, which involves automatically generating image patterns and their labels based on a mathematical formula that includes rendering functions and real-world rules, was first proposed in late 2020 (Kataoka et al. 2020). In that paper, Kataoka *et al.* showed why fractal geometry (Mandelbrot 1983; Barnsley 1988) was the best way to construct a dataset in the FDSL framework, while the present paper examines the question of whether or not ViTs can be pre-trained with FractalDB in FDSL. We contend that the ViTs can successfully be pre-trained with the FDSL framework because the self-attention mechanism enables background effects between fractal and natural images to be eliminated and because the mechanism can understand wholly fractal shapes consisting of iteratively recursive patterns. Herein, we also report on our rethink of the fractal rendering function in relation to an iterated function system (IFS) (Barnsley 1988). In our careful 'sanity check', we reveal that the process of reconsidering the IFS relationship has shown there is still room for improving the pre-training effect and enhancing natural image classifications. Figure 1, 2 shows both the impact of FractalDB pre-trained ViT and the characteristics of its training properties.

The contributions of this paper are as follows: We clarify that the pre-training of ViT can be completed without any natural images and human annotations. Our reconsideration of the fractal rendering function show that the performance of a FractalDB-10k pre-trained ViT is similar to SL approaches (see Table 4) and slightly surpasses the self-supervised ImageNet with a SimCLRv2 pre-trained ViT (see

[1] https://groups.csail.mit.edu/vision/TinyImages/

'Average' in Table 5). More specifically, in experiments using the CIFAR-10 dataset, we show that our model achieved a performance rate of 97.8, which is comparable to the rate of 97.4 achieved with SimCLRv2 and the 98.0 rate achieved with ImageNet. Additionally, we found that FractalDB pre-training allows us to acquire slightly different filters, similar position embeddings, and a wider range of receptive fields (see Figure 2(c)). Furthermore, the FractalDB pre-trained model appears to pay more attention to the contour areas (see Figure 1b), which helps to promote effective pre-training that facilitates natural image understanding.

2 Related Work

2.1 Network Architectures for Image Recognition

As mentioned previously, CNNs are currently popular in the visual recognition fields, and several well-defined structures have emerged through a large number of trials over the last decade (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015; Szegedy et al. 2015; He et al. 2016; Xie et al. 2017; Huang et al. 2017; Tan and Le 2019). Very recently, at the end of 2020, the architecture began shifting to Transformers (Vaswani et al. 2017) originating from NLP. Since this mechanism has enabled the construction of revolutionary models (e.g., BERT (Devlin et al. 2019), GPT-$\{1, 2, 3\}$ (Radford et al. 2018a,b; Brown and et al. 2020)), the CV community is now in the process of replacing the defacto-standard CNN-based architectures with Transformer.

Of these, one of the most insightful ViT architectures is (Dosovitskiy et al. 2021), which has been found to perform comparably to state-of-the-art alternative approaches, even though it is a basic architecture in terms of image input. Separately, experiments have shown that the JFT-300M/ImageNet-21k pre-trained ViT performed well in terms of accuracy but is hampered by the fact that ViTs require over ten-million-order labeled images in representation learning. Furthermore, even though the issue of learning with large-scale datasets has been somewhat alleviated with the introduction of DeiTs (Laplan et al. 2020), the pre-training problem still exists in image classification tasks.

2.2 Image Dataset and Training Framework

It has been said that the deep learning era started with the ILSVRC (Russakovsky et al. 2015), and it is undeniable that transfer learning with large-scale image datasets has contributed to accelerating visual training (He, Girshick, and Dollár 2019). Initially, the ImageNet (Deng et al. 2009) pre-trained model were widely used for diverse tasks not limited to image classification. However, even in million-scale datasets, there are concerns regarding issues such as AI ethics and copyrights, e.g., fairness protection, privacy violations, and offensive labeling. Due to these sensitive issues, as mentioned above, human-related labels in ImageNet (Yang et al. 2020) and 80M Tiny Images (Torralba, Fergus, and Freeman 2008) are in the process of being deleted, and it has become necessary for community members to consider terms-of-use in large-scale image datasets in order to create pre-trained models responsibly.

Here, it should be noted that efforts to reduce the image labeling labor required by human annotators in SSL methods have progressed significantly in recent years. The first such methods created pseudo labels based on semantic concepts (Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016; Noroozi et al. 2018; Noroozi, Pirsiavash, and Favaro 2017; Gidaris, Singh, and Komodakis 2018) and trained feature representations through image reconstruction (Zhang, Isola, and Efros 2016). However, in terms of performance rates, current SSL methods are closer to SL with human annotations (e.g., MoCo (He et al. 2020; Chen et al. 2020c), SimCLR (Chen et al. 2020a,b), SwAV (Caron et al. 2020)).

In this context, in addition to alleviating the annotation labor issue, FDSL (Kataoka et al. 2020) was proposed to overcome the problems of AI ethics and copyrights. The FDSL framework is similar to SSL, but FDSL methods do not require any natural images taken by a camera. Instead, the framework simultaneously and automatically generates image patterns and paired labels for pre-training image representations. With this point in mind, we investigate whether a formula-driven image dataset could sufficiently optimize a ViT in the pre-training phase. At the same time, we compare a pre-trained ViT created through the FDSL framework with supervised and self-supervised pre-training. It is our contention that if the supervised/self-supervised pre-training can be replaced by FDSL, it could be possible to pre-train ViTs in the future without using any natural images.

3 FDSL for ViTs

At the beginning of the section, we describe a brief review of FractalDB (Kataoka et al. 2020), followed by how to apply this auto-generated pre-training dataset to ViT (Section 3.1). Then we consider how to re-organize FractalDB in order to improve from the original one (Section 3.2).

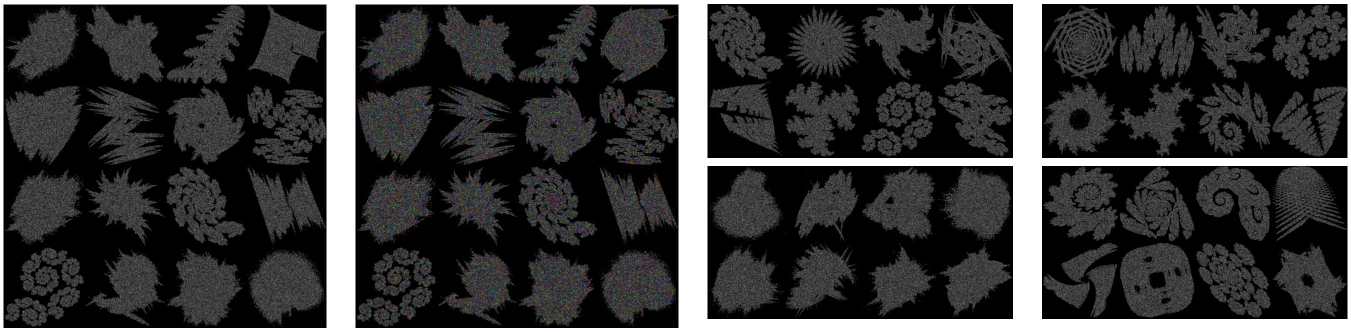
3.1 FractalDB

The most successful approach in FDSL relies on FractalDB, a dataset consists of fractal images generated with the IFS. On the 2D Euclidian space $\mathbb{R}^2$, IFS is defined by

$$\text{IFS} = \{\mathbb{R}^2; w_1, w_2, \cdots, w_N; p_1, p_2, \cdots, p_N\}, \quad (1)$$

where $w_i : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is an affine transformation, p_i is a probability, and N is the number of transformations. With IFS, a fractal $S = \{x_t\}_{t=1}^{\infty}$ is constructed by the random iteration algorithm, which repeats the following two steps: (1) select an affine transformation w^* from $\{w_1, w_2, \cdots, w_N\}$ under pre-defined probabilities $p_i = p(w^* = w_i)$, and (2) produce the next point by $x_{t+1} = w^*(x_t)$. Here, the initial point $x_1 = (0, 0)$ corresponds to the center of an image, and t is incremented by 1 at each iteration.

The original FractalDB (Kataoka et al. 2020) consisted of 1,000 or 10,000 different fractal categories with 1,000 instances per category. In their experiments, the FractalDB pre-trained ResNet-50 partially outperformed the same model pre-trained with human-annotated datasets such as ImageNet and Places, but they did not show whether or not ViTs could learn without natural images because their success rates rely on convolutional architectures. Nevertheless, even though a FractalDB image does not contain a



(a) FractalDB (b) Random colorization (c) Number of transformations (d) Range of parameters

Figure 3: Reconsidering item of FractalDB. (a) Original FractalDB in (Kataoka et al. 2020). (b) Randomly colorized FractalDB on the foreground fractal areas, since natural images are not grayscale. (c) FractalDB by considering #transformations (N). Although the conventional FractalDB consists of combined $N = \{2, 3, 4, 5, 6, 7\}$ transformations (see also (a) in the figure), the first row shows the images with $N = 2$, the second row illustrates the images with $N = 6$. The lower N tends to be a 'vivid' shape in a fractal image.Here, the range of the transformation parameters is fixed at ± 1.0. (d) In IFS, the transformation parameters basically take values in the range ± 1.0. Since changing the range to ± 0.8 (first row) or ± 1.5 (second row) produces different shapes even though the other parameters are fixed.

background area, we believe that the self-attention mechanism can effectively focus on the fractal patterns while following complex contours and ignoring background areas.

3.2 Reconsidering FractalDB

We introduce several modifications to FractalDB that were not considered in (Kataoka et al. 2020). Specifically: (i) grayscale vs. color, (ii) the number of transformations (N in equation (1)), (iii) the range of transformation parameters, and (iv) training epochs.

Grayscale vs. color. Conventional FractalDB images are drawn by moving dots or patches in grayscale. However, the natural images used for common pre-training include not only grayscale, but also images with various color combinations. In the color generation process, color points or patches were drawn randomly at each iteration time (see Figure 3b). Conducting pre-training with a dataset consisting of colored fractal images makes it possible for the model to acquires feature representations related to color.

Number of transformations. FractalDB consisted of $N = \{2, 3, 4, 5, 6, 7\}$ transformations. However, in Figure 3c, fractal shapes can change drastically due to the N parameter. According to this figure, smaller values tend to render 'vivid' shapes while larger values tend to generate 'vague' shapes. Therefore, to determine what shapes are effective in a pre-trained ViT, we explore the FractalDB configuration with a fixed $N = \{2, 4, 6, 8, 10\}$.

Range of transformation parameters. The parameters in IFS take values in the range ± 1.0. We found that the transformation parameters also make a significant difference in fractal images, as can be seen in Figure 3d. In this study, four different parameters ±{0.8, 1.0, 1.2, 1.5} were set.

Training epoch. Recent SSL methods have begun to consider longer training epochs. For example, SimCLR tried to lengthen training epochs up to 1k [epoch] (Chen et al. 2020a). Therefore, although the first FDSL study (Kataoka et al. 2020) was conducted with just 90 [epochs], we conducted further tests to determine. We also investigated longer training epochs in light of the more recent SSL methods. We evaluate up to 300 epochs to determine if they provide additional improvements to FractalDB pre-training.

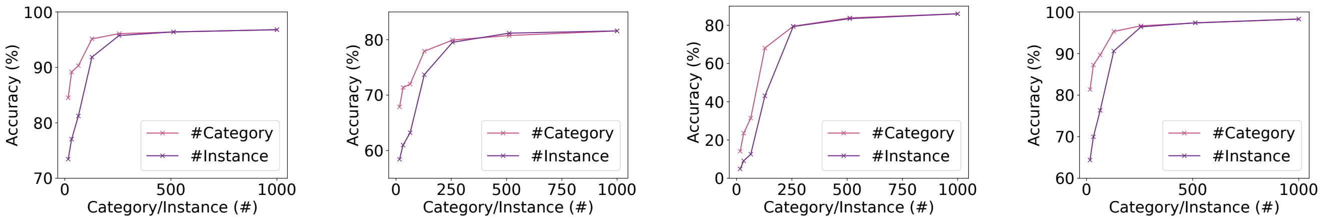
4 Experiments

This section reports on experiments conducted to verify the effectiveness of FractalDB pre-trained ViTs from various aspects. We begin by attempting to identify a better FractalDB configuration for ViT. Then, we evaluate the best configuration obtained on several image datasets. Specifically, by following the procedure outlined in (Touvron et al. 2021), we evaluate the CIFAR-10/100 (C10/C100), Stanford Cars (Cars), and Flowers-102 (Flowers) datasets. We also quantitatively compare the FractalDB pre-trained ViT with the pre-training performed with representative large-scale image datasets (e.g., ImageNet-1k and Places-365) and architectures (e.g., ResNet-50). For simplicity, this study uses the ViT tiny model and DeiT (Touvron et al. 2021) settings to confirm the properties of the FractalDB pre-trained model (hereafter, this is called ViT/ViT-Ti).

4.1 Exploration Study

Following the reconsideration discussed in Section 3.2, we explored an effective FractalDB configuration for ViT using the process described in (Kataoka et al. 2020). We began by carrying out experiments related to the FDSL-family (FractalDB, BezierCurveDB, and PerlinNoiseDB; see Table 1), architectures (ViT, gMLP, and CNN; see Table 2), and #category/#instance (see Figure 4).

Comparison with other FDSL methods (see Table 1). In addition to FractalDB, Kataoka et al. (Kataoka et al. 2020) proposed datasets based on Perlin noise (PerlinNoiseDB) and Bezier curves (BezierCurveDB). Accordingly, we conducted pre-training and fine-tuning experiments on the ViT



(a) C10 (b) C100 (c) Cars (d) Flowers

Figure 4: Effects of #category and #instance. The other parameter is fixed at 1,000, e.g., #Category is fixed at 1,000 while #Instance varies among {16, 32, 64, 128, 256, 512, 1,000}.

	C10	C100	Cars	Flowers
Scratch	78.3	57.7	11.6	77.1
PerlinNoiseDB	94.5	77.8	62.3	96.1
BezierCurveDB	96.7	80.3	82.8	**98.5**
FractalDB-1k	**97.1**	**82.6**	**87.1**	98.3

Table 1: Comparisons of ViT pre-training among FractalDB in (Kataoka et al. 2020), other FDSL with BezierCurvesDB, and PerlinNoiseDB.

Arch.	Params (M)	C10	C100	Cars	Flowers
ResNet-50	25	96.1	80.0	82.5	98.2
gMLP-Ti/16	6	95.4	77.4	78.7	94.2
ViT-Ti/16	5	**97.1**	**82.6**	**87.1**	**98.3**

Table 2: Comparison among ViT, gMLP, and ResNet with conventional FractalDB-1k pre-training.

with those FDSL methods as well.

From Table 1, we can confirm the existence of higher accuracy levels compared to scratch training for all of the FDSL methods. The improvements are up to {+18.8, +24.9, +75.5, +21.2} higher accuracy with FractalDB-1k on the {C10, C100, Cars, Flowers} datasets, respectively. Note that the configuration used here is based on the original and standard FractalDB-1k, which contains 1,000 [categories] × 1,000 [instances]. In FDSL methods, the FractalDB pre-trained ViT outperforms the other pre-trained models. The accuracy levels from BezierCurveDB pre-trained ViT are {+0.4, +2.3, +4.3, -0.2} on the {C10, C100, Cars, Flowers} datasets, respectively. Although the results showed that the BezierCurveDB pre-trained model recorded a better accuracy level on Flowers, we decided to use the FractalDB pre-trained model in consideration of the other performance rates.

ViT vs. gMLP/ResNet (see Table 2). We additionally verify ResNet-50, gMLP-Ti/16 (Liu et al. 2021), and ViT-Ti/16. We assigned the same data augmentation methods as DeiT to other architecture since the investigation was comprehensively executed. Practically, the FractalDB-1k pre-trained ResNet and gMLP were improved by using the data augmentation methods. Especially, the accuracy with ResNet-50 on C10 dataset was increased from 94.1 (reported by (Kataoka et al. 2020)) to 96.1 (+2.0 pt). The performance rates are listed in Table 2. For all fine-tuning datasets, the accuracy of ViT is higher than others. Therefore the self-attention module based architecture is more compatible with FractalDB than CNN or MLP based architecture.

#category/#instance (see Figures 4a, 4b, 4c, 4d). Figure 4 shows the effects of category and instance increases on FractalDB pre-training. We set category and instance as variables, fixing the former at 1000 and varying the others to values of {16, 32, 64, 128, 256, 512, 1000}. From the experimental results, we determined that larger category and instance values tend to lead to higher accuracy levels on a fine-tuning dataset. Particularly in the case of FractalDB pre-training, we found that the category increase is more effective for transfer learning on an image dataset. This result is intuitive because the task is easier for datasets with fewer categories and more instances. Hereafter, we set the category at 1,000 × 1,000 [instance] as the basic FractalDB setting.

Grayscale vs. color (see Table 3a). The table shows that pre-training on FractalDB works better with grayscale images than colored images. In particular, it can be seen that in C100 and Cars datasets, the improvement gaps are +1.0 pt and +1.1 improved over the pre-training performed with colored fractal images. These results confirm that colored representation is not required in ViT architecture.

Number of transformations (see Table 3b). Reducing the #transformation tends to improve the pre-training effect. This is particularly true when N is 2, which results in the highest accuracy for all the fine-tuning datasets. This indicates that it is important to input images with complex shapes for ViT pre-training.

Range of transformation parameters (see Table 3c). According to the results shown in the table, the original parameter [-1.0, 1.0] provided the best level of accuracy. However, it is necessary to carefully consider the range since the performance gap may be up to three points, for example, on Cars, between 88.4 in [-1.0, 1.0] and 85.4 in [-1.2, 1.2].

Training epochs. In FractalDB pre-training, as in the SSL, a longer training epoch tends to achieve better performance rates. The accuracy levels obtained in 300 epoch pre-training recorded the best scores in three out of the four different

	C10	C100	Cars	Flowers
Grayscale	**97.1**	**82.6**	**87.1**	**98.3**
Color	96.8	81.6	86.0	**98.3**

(a) Grayscale vs. color. The configuration of IFS follows the conventional FractalDB.

N	C10	C100	Cars	Flowers
2	**97.4**	**84.3**	**88.4**	96.5
4	97.3	82.4	87.7	98.0
6	97.1	82.7	86.3	**98.7**
8	97.2	82.3	88.1	97.4
10	97.1	82.0	86.1	98.3

(b) Number of transformations. The parameter range is fixed at ± 1.0.

Range	C10	C100	Cars	Flowers
±0.8	96.9	82.3	88.1	96.7
±1.0	**97.4**	**84.3**	**88.4**	96.5
±1.2	97.1	81.9	85.4	**98.4**
±1.5	97.1	82.5	87.4	**98.4**

(c) Range of transformation parameters. N is fixed at 2.

Table 3: Effects of colorization and IFS parameters.

PT	PT Img	PT Type	C10	C100	Cars	Flowers	VOC12	P30	IN100
Scratch	–	–	78.3	57.7	11.6	77.1	64.8	75.7	73.2
Places-30	Natural	Supervision	95.2	78.5	69.4	96.7	77.6	–	86.5
Places-365	Natural	Supervision	97.6	83.9	**89.2**	**99.3**	84.6	–	**89.4**
ImageNet-100	Natural	Supervision	94.7	77.8	67.4	97.2	78.8	78.1	–
ImageNet-1k	Natural	Supervision	**98.0**	**85.5**	**89.9**	**99.4**	**88.7**	**80.0**	–
FractalDB-1k	Formula	Formula-supervision	97.4	**84.3**	88.4	96.5	81.4	79.3	**88.2**
FractalDB-10k	Formula	Formula-supervision	**97.8**	83.1	89.1	98.8	82.6	**80.8**	88.0

Table 4: Comparison of pre-training ViT on various datasets. The pre-trained image (PT img) types shown are {natural image (Natural) and formula-driven image dataset (Formula)}; while the pre-training types (PT Types) shown are {SL (supervision) and FDSL (formula-supervision)}. The Underlined bold and bold scores show the best and second-best values.

datasets. By comparing the pre-trained model at 100 epochs, we can see that, in terms of accuracy, the 300 epoch model increased {+0.7, +0.5, +4.0, +1.8} points on the {C10, C100, Cars, Flowers} datasets, respectively.

4.2 Comparisons

Based on the results of Section 4.1, we construct a new FractalDB-{1k, 10k} and experiment with the new FractalDB-{1k, 10k} in this subsection.

In addition to performing training from scratch with additional fine-tuning datasets. We compared the performance of FractalDB pre-trained ViT with {ImageNet-1k, ImageNet-100, Places-365, Places-30} pre-trained ViT. The ImageNet-100 (IN100) and Places-30 (P30) categories were randomly selected from ImageNet-1k and Places-365 presented in (Kataoka et al. 2020). We also evaluate the models on IN100, P30, and Pascal VOC 2012 (VOC12). Additionally, we also evaluated SSL with {Jigsaw, Rotation, MoCov2, SimCLRv2} on ImageNet-1k. The results obtained show the effectiveness of the FDSL in the compared properties of human supervision with natural images (Table 4) and self-supervision with natural images (Table 5).

Larger categories. Table 4 indicates that a larger FractalDB pre-training enhances the transformer in image classification. In fact, the accuracy levels are improved to {97.4, 97.8} by FractalDB-{1k, 10k} pre-training on C10.

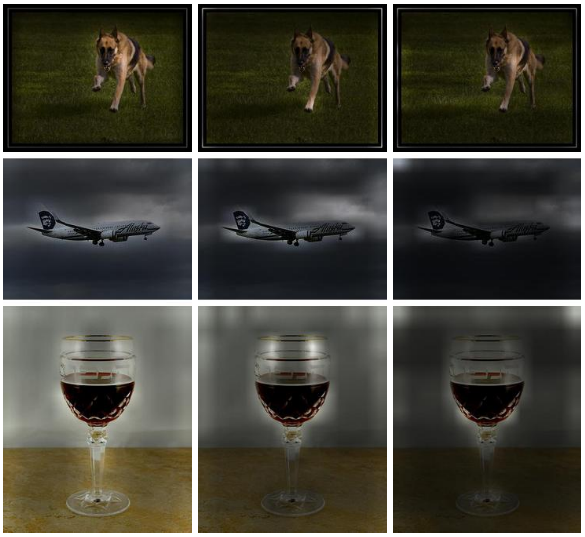
FDSL vs. SL (see Table 4). Next, natural image datasets and the FractalDB in the pre-training phase were compared. Table 4 describes the pre-training (PT), pre-training images (PT img), and their performance levels in terms of accuracy. Initially, the FractalDB-1k/10k pre-trained ViTs outperformed the pre-trained models on 100k-order labeled datasets (ImageNet-100 and Places-30). Furthermore, even though the FractalDB-10k pre-trained ViT did not exceed the performance with million-order labeled datasets (ImageNet-1k and Places-365), the scores obtained were similar to the ImageNet-1k pre-trained model.

FDSL vs. SSL (see Table 5). Through comparisons with SimCLRv2, we clarified that the FractalDB-10k pre-trained ViT performs slightly higher (FractalDB-10k 88.7 vs. SimCLRv2 88.5), in terms of average accuracy, on the representative datasets. Additionally, we found that FractalDB-10k pre-training outperformed SimCLRv2 pre-training on C10 (97.8 vs. 97.4), Cars (89.1 vs. 84.9), and P30 (80.8 vs. 80.0). However, the accuracy obtained was lower on C100 (83.5 vs. 84.1), Flowers (98.8 vs. 98.9), and VOC12 (82.6 vs. 86.2). In addition to SimCLRv2, we implemented Jigsaw, Rotation, and MoCov2 in order to compare their results with those of FractalDB-10k. The results show that while the FractalDB-10k pre-trained model tends to higher accuracy levels than other SSL methods.

Visualization (see Figures 1b, 2 and 5). As for ViT, the filters of the first linear embedding, similarity of positional embedding, and mean attention distance can be visualized by following the process described in (Dosovitskiy et al. 2021). Figure 2(a) shows filters trained with ImageNet-1k and FractalDB-1k. Although ViT pre-trained on both ImageNet-1K and FractalDB-1K acquire similar filters, the FractalDB-1k pre-trained ViT filters tend to spread over wide areas, while the ImageNet-1k pre-trained ViT filters seem to concentrate on the center areas. Figure 2(b) shows the cosine similarity of positional embedding corresponding

Method	Use Natural Images?	C10	C100	Cars	Flowers	VOC12	P30	Average
Jigsaw	YES	96.4	82.3	55.7	98.2	82.1	**80.6**	82.5
Rotation	YES	95.8	81.2	70.0	96.8	81.1	79.8	84.1
MoCov2	YES	96.9	**83.2**	78.0	98.5	**85.3**	**80.8**	87.1
SimCLRv2	YES	**97.4**	**84.1**	**84.9**	**98.9**	**86.2**	80.0	**88.5**
FractalDB-10k	NO	**97.8**	83.1	**89.1**	**98.8**	82.6	**80.8**	**88.7**

Table 5: Detailed results with FDSL vs. SSL. The additional column labeled 'Use Natural Images?' indicates whether or not natural images were used in the pre-training phase. 'Average' is the average accuracy of all datasets in the table. Note that ImageNet-100 has been eliminated from this table because the listed SSL methods are trained by images on ImageNet-1k. The Underlined bold and bold scores show the best and second-best values.



(a) ImageNet-1k (b) FractalDB-1k (c) FractalDB-10k

Figure 5: Attention maps.

to the input patch at each row and column. From the visualized figures, it can be seen that the FractalDB-1k pre-trained ViT acquired position embeddings at each row and column that are similar to ImageNet-1k pre-trained ViT.

Figure 2(c) shows mean attention distance in the same manner as the original ViT (Dosovitskiy et al. 2021). In comparison to the ImageNet-1k pre-training, the FractalDB-1k pre-trained ViT appears to look at widespread areas in an image.

Figure 5 shows attention maps with different pre-training. Here, it can be seen that, similar to ImageNet pre-training (Figure 5a), the FractalDB-1k pre-trained ViT focuses on the object areas (Figure 5b). Additionally, we can see that the FractalDB-10k pre-trained ViT looks at more specific areas (Figure 5c) compared to FractalDB-1k pre-training. Figure 1b shows attention maps in fractal images. From these figures, it can be seen that the FractalDB pre-training seems to perform recognition by observing contour lines. In relation to Figure 2(c), we believe that the ability to recognize complex and distant contour lines enabled the extraction of features from a widespread area.

5 Conclusion and Discussion

This study demonstrated that the FDSL framework makes it possible to train ViTs without the use of natural images or human-annotated labels. Additionally, we showed how our FractalDB pre-trained ViT achieved similar performance rates to the human-annotated ImageNet pre-trained model, and partially outperformed SimCLRv2 self-supervised ImageNet pre-trained model. Furthermore, based on the results in this study, the following findings are presented:

Feature representation with FractalDB pre-trained ViT. We see that the FractalDB pre-trained ViT acquired different feature representations in the first linear embeddings (Figure 2(a)) and similarly arranged position embeddings (Figure 2(b)) when compared to the ImageNet-1k pre-training. Moreover, Figure 1b shows that the ViT tends to pay attention to contour areas in the pre-training. We believe that this pre-trained model enabled feature acquisition in a wider range of areas than the ImageNet-1k pre-trained model (Figure 2(c)). We also identified the complex contour lines used to classify fractal categories in the pre-training phase.

The effect of FractalDB reconsideration. Extending the process outlined in (Kataoka et al. 2020), we noticed the potential for improvement and attempted to implement several novel configurations such as using colored FractalDB, increasing the number of transformations, using a wider range of transformation parameters, and applying longer training epochs. As shown in the experimental section, the highest scores occurred when the following parameter combination was used: grayscale fractal images, $N = 2$ for #transformation, range of ±1.0, and 300 epochs in the pre-training.

Is it possible to complete pre-training of ViT without natural images and human-annotated labels? In Table 5, we showed that the performance of FractalDB-10k was comparable in terms of accuracy to the SimCLRv2 pre-trained ViT. Although 10M images are used in FractalDB-10k, no natural images were used in the pre-training. Therefore, an FDSL-based pre-training dataset can be safely and effectively used to train a ViT in terms of AI ethics and image copyright, providing we can exceed the SL accuracy achieved via human annotation (Table 4). In relation to this point, we showed that the FDSL pre-training recorded similar accuracy levels to those achieved via SL, which means that it is possible that the ImageNet pre-trained model may be replaced by a model without any natural images in the near future.

Acknowledgement

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by JSPS KAKENHI Grant Number JP19H01134. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used. We want to thank Yoshihiro Fukuhara and Munetaka Minoguchi for their helpful comments in research discussions.

References

Barnsley, M. F. 1988. Fractals Everywhere. *Academic Press. New York.*

Brown, T. B.; and et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS).*

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems (NeurIPS).*

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML).*

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020b. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems (NeurIPS).*

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020c. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL).*

Doersch, C.; Gupta, A.; and Efros, A. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *IEEE International Conference on Computer Vision (ICCV).*

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representation (ICLR).*

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representation (ICLR).*

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

He, K.; Girshick, R.; and Dollár, P. 2019. Rethinking ImageNet Pre-training. In *IEEE International Conference on Computer Vision (ICCV).*

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

Huang, G.; Liu, Z.; Maaten, L. v. d.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

Kataoka, H.; Okayasu, K.; Matsumoto, A.; Yamagata, E.; Yamada, R.; Inoue, N.; Nakamura, A.; and Satoh, Y. 2020. Pre-training without Natural Images. In *Asian Conference on Computer Vision (ACCV).*

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS).*

Laplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Liu, H.; Z., D.; So, D. R.; and Le, Q. V. 2021. Pay Attention to MLPs. arXiv:2105.08050.

Mahajan, D.; Girshick, R.; Ranathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Maaten, L. v. d. 2018. Exploring the Limits of Weakly Supervised Pretraining. In *European Conference on Computer Vision (ECCV).*

Mandelbrot, B. 1983. The fractal geometry of nature. *American Journal of Physics*, 51(3).

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *European Conference on Computer Vision (ECCV).*

Noroozi, M.; Pirsiavash, H.; and Favaro, P. 2017. Representation Learning by Learning to Count. In *IEEE International Conference on Computer Vision (ICCV).*

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting Self-Supervised Learning via Knowledge Transfer. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018a. Improving language understanding by generative pre-training. In *Technical Report, OpenAI.*

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018b. Language Models are Unsupervised Multitask Learners. In *International Conference on Machine Learning (ICML).*

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3).

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *International Conference on Computer Vision (ICCV)*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*.

Torralba, A.; Fergus, R.; and Freeman, W. T. 2008. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Conference on Fairness, Accountability and Transparency (FAT)*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful Image Colorization. In *European Conference on Computer Vision (ECCV)*.