Scaling Backwards: Minimal Synthetic Pre-training?

Ryo Nakamura[1,*], Ryu Tadokoro[2,*], Ryosuke Yamada[1], Yuki M. Asano[3], Iro Laina[4], Christian Rupprecht[4], Nakamasa Inoue[5], Rio Yokota[5], and Hirokatsu Kataoka[1]

[1] National Institute of Advanced Industrial Science and Technology (AIST)
[2] Tohoku University
[3] University of Amsterdam, now at University of Technology Nuremberg
[4] University of Oxford
[5] Tokyo Institute of Technology

Abstract. Pre-training and transfer learning are an important building block of current computer vision systems. While pre-training is usually performed on large real-world image datasets, in this paper we ask whether this is truly necessary. To this end, we search for a minimal, purely synthetic pre-training dataset that allows us to achieve performance similar to the 1 million images of ImageNet-1k. We construct such a dataset from a single fractal with perturbations. With this, we contribute three main findings. (i) We show that pre-training is effective even with minimal synthetic images, with performance on par with large-scale pre-training datasets like ImageNet-1k for full fine-tuning. (ii) We investigate the single parameter with which we construct artificial categories for our dataset. We find that while the shape differences can be indistinguishable to humans, they are crucial for obtaining strong performances. (iii) Finally, we investigate the minimal requirements for successful pre-training. Surprisingly, we find that a substantial reduction of synthetic images from 1k to 1 can even lead to an *increase* in pre-training performance, a motivation to further investigate "scaling backwards". Finally, we extend our method from synthetic images to real images to see if a single real image can show similar pre-training effect through shape augmentation. We find that the use of grayscale images and affine transformations allows even real images to "scale backwards". The code is available at `https://github.com/SUPER-TADORY/1p-frac`.

Keywords: Synthetic pre-training · Limited data · Vision transformers

1 Introduction

In image recognition, pre-training allows discovering fundamental visual representations for downstream task applications. Pre-training enhances the performance of visual tasks and enables the use of small-scale task-specific datasets.

[*] These authors contributed equally

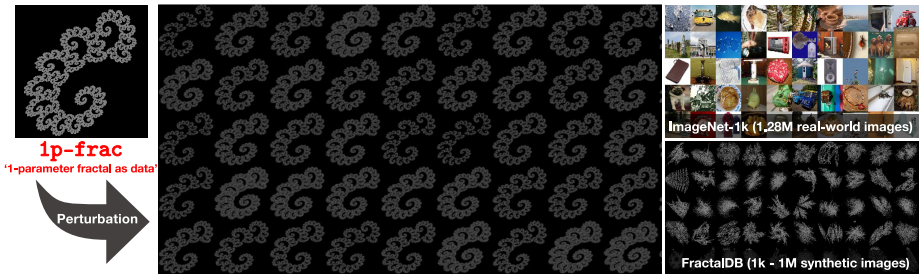

Fig. 1: Comparison of ImageNet-1k, FractalDB and `1p-frac` (ours). `1p-frac` consists of only a single fractal for pre-training. With `1p-frac`, neural networks learn to classify perturbations applied to the fractal. In our study "single" means a very narrow distribution over parameters that leads to images that are roughly equivalent from a human visual perspective. While the shape differences of perturbed images can be indistinguishable to humans, models pre-trained on `1p-frac` achieve comparable performance with those pre-trained on ImageNet-1k or FractalDB.

Recently, pre-training has been used as a key technology to construct foundation models trained on massive datasets with over hundreds of millions of images. In some cases, a foundation model enables adaptation of zero-shot recognition without the need for additional data.

Pre-training is often interpreted as discovering universal structures in large-scale datasets that later facilitate adaptation to down-stream tasks. In this paper, we challenge this interpretation by providing a *minimal* pre-training dataset, generated from a single fractal, that achieves similar downstream performance. At the heart of this investigation is the question of whether pre-training might simply be a better weight initialization rather than the discovery of useful visual concepts. If true, performing expensive pre-training with hundreds of millions of images might not be necessary. This additionally frees pre-training from licensing or ethical issues.

Since the rise of deep neural networks, the ImageNet dataset [16] has been one of the most commonly used pre-training datasets. Originally, pre-training has been conducted through supervised learning (SL) with human-provided teacher labels. However, it has become clear that pre-training can also be achieved without human-provided labels, through self-supervised learning (SSL) [7, 9–12, 17, 21–23, 37, 38, 56].

In this context, Asano *et al.* [3] successfully acquired visual representations while dramatically reducing the number of required images. They concluded that SSL can yield sufficient image representations even with a single training example, but only for earlier layers of a recognition model. However, it is unclear how these findings translate to modern architectures and representation learning methods. Based on this, vision transformers (ViT) [18] have been pre-trained with only 2,040 real images [52] through an instance discrimination learning signal.

More recently, it has been shown that basic visual representations can be acquired without using real images and human-provided labels. The trend of

artificially generated labeled images is on the rise for synthetic pre-training [5, 28, 30, 46]. Formula-driven supervised learning (FDSL) generates images from generative formulas, and labels from their parameters [30]. Under the FDSL framework, one can adapt the synthetic pre-training image dataset by altering formulas [24, 27, 29, 44]. While a million-order image dataset was constructed in FractalDB [30], our findings suggest that synthetic pre-training can be reduced to significantly fewer fractal images.

Motivated by these findings, we believe it is possible to find the key essence of pre-training for image recognition. ViT training can be done with as few as 1,000 artificially generated images [35]. Here, we believe that equivalent performance can be achieved with even fewer images. This consideration is undoubtedly important as we approach a minimal synthetic pre-training dataset in image recognition, which goes against the trend of foundation models toward increasing the dataset scale.

In the present paper, we thus introduce a *minimal* synthetic dataset, namely 1-parameter Fractal as Data (`1p-frac`), which consists of a single fractal as shown in Figure 1, as well as a loss function for pre-training with it. Our contributions regarding minimal synthetic pre-training are as follows:

Ordinal minimalism: We introduce the locally perturbed cross entropy (LPCE) loss for pre-training with a single fractal. It utilizes perturbed fractal images for training, where neural networks learn to classify small perturbations. In experiments, we demonstrate that pre-training can be performed even with a single fractal. The pre-training effect of `1p-frac` is comparable to that of a million-scale labeled image dataset.

Distributional minimalism: We introduce the locally integrated empirical (LIEP) distribution p_{Δ} that has a controllable perturbation degree Δ to investigate the minimal support of the probability density distribution of synthesized images. We observed positive pre-training effects even with a small Δ producing shape differences that cannot be distinguished by humans. We also show that if Δ is too small the visual pre-training collapses. From these observations, we establish the general bounds for generating good pre-training images from a mathematical formula.

Instance minimalism: Based on the experimental results, the synthetic images should not simply contain complex shapes. The use of recursive image patterns, similar to objects in nature, should be applied in visual pre-training. Experiments with augmented categories from a real image have shown that good pre-training effects can be achieved by performing "affine transformations on edge-emphasized objects in grayscale". These operations are found to be almost synonymous with the configuration of the proposed `1p-frac`.

Table 1: Scaling backwards in synthetic pre-training (Accuracies on CIFAR-100, Real: ImageNet, Synth: Fractal images).

Type\#Img	1	1k	1M
Real	N/A	76.9	85.5
Synth	84.2	84.0	81.6

In summary, we significantly reduce the size of the pre-training dataset from 1M images (fractal database (FractalDB) [30]) or 1k images (one-instance fractal

database (OFDB) [35]) down to 1 and show that this even improves the pre-training effect (Table 1), which motivates "scaling backwards".

2 Related Work

The present study focuses primarily on finding the minimal pre-training dataset with artificially generated images.

Pre-training in image recognition. Pre-training in image recognition has demonstrated significant success with vast amounts of labeled images. Pre-training helps neural networks to acquire fundamental visual representations which are essential for improving the performance of diverse downstream tasks in image recognition. In particular, researchers have used larger-scale image datasets for supervised learning (SL), starting from million-scale datasets (e.g., ImageNet [16], Places [59]). More recently, the largest datasets used are reaching billion-scale (e.g., IG-3.5B [34], JFT-300M/3B/4B [15, 42, 53]).

Pre-training can be also achieved without any human intervention. One of the most effective approaches is self-supervised learning (SSL [7, 9–12, 17, 21–23, 37, 38, 56]), which has been developed in order to alleviate the manual annotations on large-amount of images. SSL has been studied as an expected alternative to supervised pre-training and has sometimes been shown to surpass supervised pre-training.

Despite the development of supervised and self-supervised pre-training, the number of real images used in visual pre-training continues to increase. Exploratory research questions such as "what is a minimal synthetic pre-training dataset?" have received less attention. However, they could give us more insights on what is learned during pre-training and how it could be improved.

Pre-training with limited data. In order to address ethical issues such as privacy and fairness associated with real images, a pre-training with a very limited number of real images and synthetic pre-training without any real images have been on the rise. In this context, synthetic pre-training with copy-and-paste learning [19, 39], domain randomization from 3D graphics [43, 46], learning from primitive patterns [4, 5, 41], and dataset distillation [8, 50, 57] have been recognized as promising approaches to acquiring visual representations. More recently, a sophisticated approach, formula-driven supervised learning (FDSL) [28, 30], enabled automatic construction of a labeled image dataset without any real images or human annotations. FDSL can generate images and their corresponding labels using a mathematical formula. Along these lines, One-instance FractalDB (OFDB) has successfully pre-trained a ViT using only 1,000 synthetic images generated from a mathematical formula, without using real images [35]. Moreover, several studies have been reported in the context of extremely limited data [3, 49, 52]. In particular, Asano *et al.* have shown that learning visual representations is possible using a single image for self-supervised learning. Venkataramanan *et al.* achieved a comparable method to ImageNet pre-training with only 10 walking videos and their cropped images.

By analyzing the acquisition of visual representations through the use of minimal synthetic pre-training datasets, we believe that the FDSL framework will be key to discovering the essence of pre-training mechanisms in image recognition.

3 Scaling Backwards with a Single Fractal

This section presents the proposed methodology to explore the minimal requirements for successful pre-training by utilizing purely synthetic images of fractals. Specifically, we introduce the 1 parameter Fractal as Data (`1p-frac`), which consists of only a single fractal, with a method to pre-train neural networks on it. Our key idea is to introduce *the locally integrated empirical (LIEP) distribution* p_{Δ} over perturbed fractal images, enabling pre-training even with a single fractal image. Since the LIEP distribution is designed so that it converges to the empirical distribution $p_{\text{data}}(x) = \delta(x - I)$ of a single image I when the perturbation degree $\Delta \in \mathbb{R}_{\geq 0}$ goes to zero, we can narrow down the support of the distribution by decreasing Δ as shown in Figure 2a. Below, we begin with a preliminary for defining the empirical distributions of two fractal databases, namely FractalDB [30] and OFDB [35]. We then introduce `1p-frac` with the LIEP distribution.

3.1 Preliminary

FractalDB [30]. Kataoka *et al.* have introduced a method for effectively pre-training neural networks with FractalDB, a set of fractal images generated by the iterated function systems (IFSs). Specifically, FractalDB $\mathcal{F}$ consists of one million synthesized images: $\mathcal{F} = \{(\Omega_c, \{I_i^c\}_{i=0}^{M-1})\}_{c=0}^{C-1}$, where Ω_c is an IFS, I_i^c is a fractal image generated by Ω_c, $C = 1{,}000$ is the number of fractal categories, and $M = 1{,}000$ is the number of images per category. Each IFS characterizes each fractal category c and is defined as follows:

$$\Omega_c = \{\mathbb{R}^2; w_1, w_2, \ldots, w_{N_c}; p_1, p_2, \ldots, p_{N_c}\}, \tag{1}$$

where $w_j : \mathbb{R}^2 \to \mathbb{R}^2$ is a 2D affine transformation given by

$$w_j(\boldsymbol{v}) = \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \boldsymbol{v} + \begin{bmatrix} e_j \\ f_j \end{bmatrix} \quad (\boldsymbol{v} \in \mathbb{R}^2), \tag{2}$$

and p_j is a probability mass distribution. Each fractal image I_i^c renders a fractal $F = \{\boldsymbol{v}_t\}_{t=0}^{T} \subset \mathbb{R}^2$ into a 2D image, where points $\boldsymbol{v}_t$ are determined by a recurrence relation that applies affine transformations as $\boldsymbol{v}_{t+1} = w_{\sigma_t}(\boldsymbol{v}_t)$ for $t = 0, 1, 2, \cdots, T-1$. Here, the initial point is set as $\boldsymbol{v}_0 = (0, 0)^\top$, and the index σ_t is sampled at each t independently from the probability mass distribution $p(\sigma_t = j) = p_j$. Pre-training with FractalDB utilizes the cross-entropy loss function:

$$\mathcal{L} = -\mathbb{E}_{x,y \sim p_{\text{data}}}[\log p_\theta(y|x)], \tag{3}$$

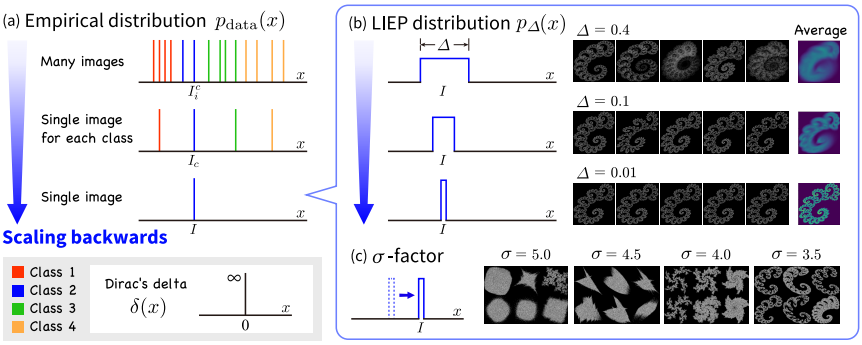

Fig. 2: Scaling backwards from many images to a single synthetic image. (a) Empirical distribution p_{data}. Colors indicate classes. With a single image, the distribution is given by a single Dirac's delta function. (b) LIEP distribution p_{Δ}. The support of the distribution narrows as the degree of perturbation Δ decreases. (c) σ-factor for investigating fractal shapes. A small σ produces complex fractals.

where p_θ is the category distribution predicted by a neural network and θ is a set of learnable parameters. The joint empirical distribution p_{data} is defined over the dataset as follows:

$$p_{\text{data}}(x, y; \mathcal{F}) = \frac{1}{MC} \sum_{i=0}^{M-1} \sum_{c=0}^{C-1} \delta(x - I_i^c)\delta(y - c) \tag{4}$$

where δ is the Dirac's delta function. Models pre-trained on this dataset perform comparable to those pre-trained on real-world image datasets, such as ImageNet-1k and Places365, in some downstream tasks.

OFDB [35]. This dataset consists of 1,000 fractal images. Specifically, OFDB $\mathcal{F}_{\text{OF}}$ involves only one representative image per category, *i.e.*, $\mathcal{F}_{\text{OF}} = \{\Omega_c, I_c\}_{c=0}^{C-1}$. Therefore, the joint empirical distribution reduces to

$$p_{\text{data}}(x, y; \mathcal{F}_{\text{OF}}) = \frac{1}{C} \sum_{c=0}^{C-1} \delta(x - I_c)\delta(y - c). \tag{5}$$

Models pre-trained on this dataset perform comparable to or even better than those pre-trained on FractalDB. This work has shown that there exists a small but essential set of images for visual pre-training. However, reducing the number of fractals C to less than 1,000 degrades the performance.

3.2 Pre-training with a Single Fractal

Scaling backwards. To further facilitate the analysis of what images are minimally required for successful visual pre-training, we introduce `1p-frac`, which ultimately reduces the number of IFSs and images to one each as $\mathcal{F}_{\text{OP}} = (\Omega, I)$. With this dataset, the empirical distribution is given by

$$p_{\text{data}}(x, y; \mathcal{F}_{\text{OP}}) = \delta(x - I)\delta(y). \tag{6}$$

However, we notice that with this distribution, training neural networks using cross-entropy loss is not straightforward because $p_\theta(y = 0|x) \equiv 1$ $(\forall x)$ gives a trivial solution to the loss minimization problem. To address this, we introduce locally perturbed cross entropy (LPCE) loss $\mathcal{L}_\Delta$, a variant of the cross entropy loss defined with the LIEP distribution.

Definition 1. *Let $I_{\boldsymbol{\epsilon}} \in \mathcal{X}$ be a perturbed image, where $\mathcal{X}$ is a set of images, $\boldsymbol{\epsilon} \in \mathbb{R}^d$ is a small perturbation with $d \in \mathbb{N}_{>0}$ and $I_{\mathbf{0}} = I$ is the original image. We define the LIEP distribution by*

$$p_\Delta(x, y) = \frac{1}{|\mathcal{R}_\Delta|} \int_{\mathcal{R}_\Delta} \delta(x - I_{\boldsymbol{\epsilon}})\delta(y - \boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \tag{7}$$

where $\mathcal{R}_\Delta \subset \mathbb{R}^d$ is a compact set containing the origin and $|\mathcal{R}_\Delta|$ is its volume with order $O(|\Delta|^d)$.

Definition 2. *We define the LPCE loss by*

$$\mathcal{L}_\Delta = -\mathbb{E}_{x,y\sim p_\Delta} \left[\log p_\theta(y|x)\right], \tag{8}$$

where p_Δ is the LIEP distribution.

If $\mathcal{R}_\Delta$ is a small hypercube or hypersphere, it is straightforward to see that p_Δ approaches to the empirical distribution of Eq. (6) when Δ goes to zero. Therefore, this loss allows us to analyze visual pre-training effects by narrowing the support of the distribution around a single image.

With `1p-frac`, we apply perturbation to the affine transformations. As such, the perturbation $\boldsymbol{\epsilon}$ is in $\mathbb{R}^{6*j}$. A perturbed image $I_{\boldsymbol{\epsilon}}$ is obtained by the noisy affine transformations:

$$w_j\left(\boldsymbol{v}; \boldsymbol{\epsilon}_j\right) = \left(\begin{bmatrix} a_j & b_j & e_j \\ c_j & d_j & f_j \end{bmatrix} + \boldsymbol{\epsilon}_j \right) \begin{bmatrix} \boldsymbol{v} \\ 1 \end{bmatrix} \tag{9}$$

where $\boldsymbol{\epsilon}_j \in \mathcal{R}_\Delta = [-\Delta/2, \Delta/2]^{6*j} \subset \mathbb{R}^{6*j}$ is a hypercube with a side length Δ, and $|\mathcal{R}_\Delta| = \Delta^{6*j}$. Note that the numerical integration of Eq. (7) is used in practice, which is an approximation obtained by uniformly sampling L points in $\mathcal{R}_\Delta$, where we set $L = 1{,}000$ by default.

Visualization. Figure 2b shows examples of perturbed images to compute the LPCE loss. While most of shape differences are indistinguishable to humans, a neural network learns to distinguish the perturbations applied to the single image by minimizing the LPCE loss.

Complexity of Ω. We use the σ-factor proposed by Anderson *et al.* [1] to evaluate the complexity of IFSs. As shown in Figure 2c, small values of σ produce complex fractal shapes.

4 Experiments

Experimental setup. What is minimal visual pre-training, and what properties of images lend themselves to reducing their count? In order to clarify these questions, we conduct experiments following these steps:

- **Exploration study (Tables 2-4):** We verify whether a single fractal is sufficient for pre-training. Moreover, we compare the extent to which the shapes used in pre-training can be simplified, using object contours, Gaussian, and uniform distributions.
- **Hyperparameter study (Tables 5-7):** We investigate the effects of three hyperparameters Δ, σ and L.
- **Scaling study (Table 8):** We compare `1p-frac` with other large-scale datasets in terms of fine-tuning accuracy on ImageNet-1k.
- **Analysis and discussion (Tables 9-12):** We discuss data augmentation, computational cost for synthesizing images, and pre-training using a single real-world image.
- **Applications (Tables 13 and 14):** We conduct comparisons in various fine-tuning datasets.

Implementation details. In order to verify the effects of pre-training, we measure the accuracy of downstream tasks through fine-tuning. We use ViT [18] for all experiments. Specifically, we employ ViT-Tiny (ViT-T) in the the exploration and hyperparameter studies, and employ ViT-Base (ViT-B) for the scaling study. The pre-training parameters and data augmentation methods for pre-training and fine-tuning follow conventional methods in OFDB. In OFDB, data augmentation is based on DeiT. For exploration studies, CIFAR-100 (C100) [32] and/or ImageNet-100 (IN100) [40] are used as fine-tuning datasets. In the comparisons, we assign seven representative datasets used in OFDB paper, namely C10/C100, Cars [31], Flowers [36], Pascal VOC 2012 (VOC12) [20], Places-30 (P30) [59], IN100. See supplementary material for more detailed experimental settings.

4.1 Exploration Study

Pre-training with `1p-frac`. Table 2 compares `1p-frac` with FracdalDB (1M images, 1k categories) and OFDB (1k images, 1k categories). As can be seen, the pre-training effect of `1p-frac` is comparable to that of FractalDB and OFDB. The number of fractal categories is significantly reduced from $C = 1{,}000$ to only 1. This shows the effectiveness and efficiency of `1p-frac` and the LPCE loss.
Comparison with SOTA FDSL datasets. Table 3 compares `1p-frac` with two SOTA FDSL datasets: (1) Radial Contour Database (RCDB) [28], which consists of 1 million images of synthetic polygons, and (2) Visual Atoms (VA) [45], which consists of 1 million images of parameterized wave functions. The "1K" counterparts in the table reduce the number of images per category to one. We also applied the LPCE loss to these two datasets by applying perturbations to the radios of a polygon or wave. As can be seen, `1p-frac` performs the best when a single image is used for pre-training and outperforms the 1K FDSL datasets.
Pre-training with a noisy image. To validate the necessity of a complex shape when pre-training with a single image, we applied the LPCE loss to two noisy images: an image of Gaussian noise and an image of uniform noise. Since these two noise images are generated by determining the parameters of the parametric distributions, Gaussian and uniform, the LPCE loss can be computed by

Table 2: Comparison of `1p-frac` with FractalDB [30] and OFDB [35]. $^{\diamond}$ indicates the use of LPCE loss. We report top-1 accuracies (%).

Method	#Img	C100	IN100
Scratch	-	64.2	74.9
FractalDB	1M	81.6	88.5
OFDB	1k	84.0	88.6
`1p-frac`$^{\diamond}$	**1**	**84.2**	**89.0**

Table 3: Comparison with SOTA FDSL datasets. $^{\diamond}$ indicates the use of LPCE loss.

Method	#Img	C100	IN100
RCDB	1M	81.6	88.5
RCDB	1K	80.4	87.5
RCDB$^{\diamond}$	**1**	82.5	87.9
VA	1M	84.9	90.3
VA	1K	82.1	88.5
VA$^{\diamond}$	**1**	82.6	88.0
`1p-frac`$^{\diamond}$	**1**	84.2	89.0

Table 4: Comparison with pre-training with a single noise image.

Method	C100	IN100
Gaussian$^{\diamond}$	1.1	5.7
Uniform$^{\diamond}$	2.0	71.1
`1p-frac`$^{\diamond}$	**84.2**	**89.0**

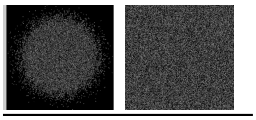

Table 5: Effects of perturbation degree Δ ($\sigma = 3.5$).

Δ	C100	IN100
0.001	1.2	1.9
0.01	19.9	61.8
0.05	83.0	88.2
0.1	**84.2**	**89.0**
0.2	83.4	88.5
1.0	82.6	88.1

Table 6: Effects of σ. ($\Delta = 0.1$).

σ	C100	IN100
6.0	81.3	86.3
5.0	82.2	87.1
4.5	81.9	86.9
3.5	**84.2**	**89.0**
4.0	82.8	87.9

Table 7: Effects of number of sampling points L for numerical integration.

L	C100	IN100
16	78.7	85.3
64	82.3	88.0
512	82.4	88.0
1000	**84.2**	**89.0**

giving small perturbations to the parameters. Table 4 shows these images and fine-tuning results. For example, here the ϵ is added to the mean vector and covariance matrix of the Gaussian distribution. The results revealed that noise images fail in visual pre-training, performing worse than training from scratch. This suggests that acquiring fundamental visual representations through pre-training requires an image of a certain structured object, such as a fractal.

4.2 Hyperparameter Study

Perturbation degree. Table 5 shows the results obtained by different perturbation degrees Δ for the LIEP distribution. The results indicate that setting the value to 0.1 yields the highest performance. Interestingly, the performance significantly improved when Δ was increased from 0.01 to 0.05. This suggests that the support of the empirical distribution must have a certain size to obtain positive pre-training effects.

σ-factor. Table 6 investigates the effects of the σ-factor, which measures the complexity of fractal shapes. As can be seen, the most complex shape yields the highest performance. It is worth noting that, even with a fractal of $\sigma = 6.0$, which looks like Gaussian noise, we observed positive pre-training effects. This suggests that the shapes obtained from IFS are crucial for pre-training.

Numerical integration. Tables 7 reduces the number of sampling points L of the numerical integration when approximately computing loss with Eq. (7).

Table 8: Comparison of pre-training datasets on ImageNet-1k fine-tuning. Fine-tuning accuracies calculated by using ViT-Base (B) are listed in the table. 21k indicates the number of categories in each pre-training dataset.

Pre-training	#Img	Type	ViT-B
Scratch	–	–	79.8
ImageNet-21k	14M	SL	81.8
FractalDB-21k	21M	FDSL	81.8
ExFractalDB-21k	21M	FDSL	82.7
RCDB-21k	21M	FDSL	82.4
VA-21k	21M	FDSL	82.7
OFDB-21k	21k	FDSL	82.2
3D-OFDB-21k	21k	FDSL	82.7
`1p-frac` (ours)	**1**	FDSL	82.1

Table 9: Ablation study on data augmentation methods. The DeiT augmentation [47] is used as a baseline.

Method	C100
Baseline	84.2
w/o Random Aug. [14]	83.4
w/o Random Crop	**80.1**
w/o Rand Aspect	83.6
w/o Rand Erasing [58]	84.0
w/o Mixup [55]	83.6
w/o Cutmix [51]	83.3
w/o Mixup/Cutmix	**80.4**
w/o Flipping	84.3
w/o Color Jittering	84.2

We see that larger numbers result in high performance. Interestingly, even with $L = 16$, the pre-training effect was positive (better than training from scratch).

4.3 Scaling Study

ImageNet-1k fine-tuning (Table 8). While this work discusses minimal requirements for visual pre-training, increasing the approximation precision of the numerical integration could further improve the performance and benefit training of large models. In Table 8, we applied the LPCE loss with $L = 21{,}000$ to the ViT-B model and compare fine-tuning accuracy on ImageNet-1k. Surprisingly, and despite using a single fractal as data, our ViT-B pre-trained on `1p-frac` outperforms pre-training on ImageNet-21k. This shows the strength of our approach in particular because we simply keep the same fine-tuning protocols as used for the original Imagenet-21k to ImageNet-1k transfer proposed in [18], putting us at a potential disadvantage.

4.4 Analysis and Discussion

Relationship with data augmentation (Table 9). We consider that the variation in image patterns also depends on data augmentation. While traditional methods primarily rely on the data augmentation techniques described in the DeiT paper [47], we investigated which data augmentation methods contribute significantly by excluding one or two techniques from the base data augmentation methods and observing the changes in fine-tuning accuracy on C100 dataset. The experimental results in Table 9 reveal that excluding Random Cropping (RandomCrop) significantly deteriorates accuracy, suggesting that omitting parts of the image region and causing partial loss facilitates the acquisition of the ability to focus on any part without relying on grasping the overall shape, greatly reducing the pre-training effect. Furthermore, when both Mixup and Cutmix are

Table 10: Dataset construction time (hours). FractalDB (1M images), OFDB (1k images) and `1p-frac` (1k perturbations for one image) are compared. The processing time is shown separately for fractal category search (Search), image rendering (Render), and total time (Total).

Dataset	Search	Render	Total
FractalDB	2.37	16.86	19.23
OFDB	2.37	0.41	2.78
`1p-frac`	0.0022	0.036	0.0382

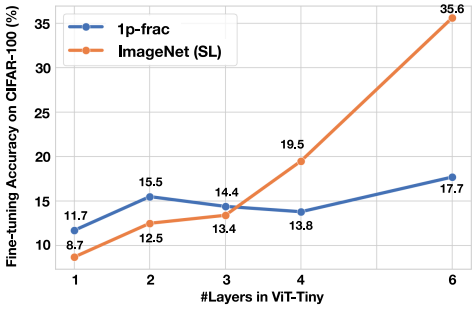

Fig. 3: Linear probing following the experiments from [3]. We verified the performance with {1, 2, 3, 4, 6} layers in the ViT-Tiny model on C100 dataset. Here, 2 and 4 show the use of up to 1–2 and 1–4 layers.

excluded, a degradation in accuracy is observed, indicating that mixing images and categories enables better feature recognition.

Dataset construction times (Table 10). We calculated the processing times in fractal category search and image rendering. By comparing to the conventional approaches, our `1p-frac` recorded much faster category search (0.0022 hours nearly equals to 8 seconds) and image rendering (0.036 hours nearly equals to 129 seconds). `1p-frac` requires only one category with a parameter set. This very simple procedure leads to efficiently construct a minimum requirement for pre-training in image recognition.

Linear probing (Figure 3). To compare supervised ImageNet-1k pre-training with our `1p-frac` pre-training in the context of early layer visual representations, we executed a linear probing experiment on the C100 dataset. In the ViT-T model with {1, 2, 3} layers, the `1p-frac` pre-trained model outperformed the model pre-trained on ImageNet-1k. This demonstrates the superior quality of early, low-level representations learned from our dataset compared to ImageNet. In effect, layers up to layer 3 in ViT-T can be sufficiently pre-trained and frozen using our `1p-frac`.

Pre-training with a single real image and shape augmentation. It was found that pre-training effects comparable to a dataset of one million real-world images can be achieved with `1p-frac`, but is it possible to perform pre-training similarly with a real image? In this section, we verify whether visual features can be acquired through pre-training using a single real image (each image in Figure 5) with the LPCE loss using image representations similar to FDSL (Canny [6] images) and geometric transformations for perturbation ({affine, elastic, polynomial} transformations in Figure 4). The results of pre-

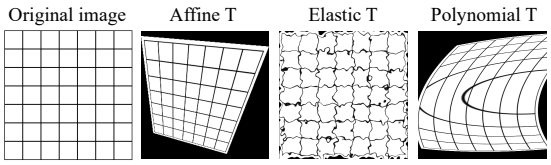

Fig. 4: Shape augmentation with three geometry transformations from a single real image.

Fig. 5: The five different images used in the pre-training with a single real image and shape augmentation.

Table 11: Comparison of pre-training effect between RGB and Canny images when shape augmentation is applied. The shape augmentation uses affine transformations.

Image	Mean Accuracy
RGB image	81.6
Canny image	**82.2**

Table 12: Comparison of pre-training effects of affine, elastic, and polynomial image transformations. Real Img type with Canny applied to the images.

Transformation	Mean Accuracy
Affine trans.	**81.9**
Elastic trans.	37.0
Polynomial trans.	81.3

training effects with $L = 1,000$ are shown in Tables 11 and 12. Note that the results in these tables represent the average score on C100 over the five real images in Figure 5. See the supplementary material for results for each of image. According to the results, we find that contour-emphasized Canny images omitting color information and affine transformation perturbation yield better pre-training, as measured by the fine-tuning accuracies. The operations are similar to the classification of labeled fractal images via IFS.

4.5 Applications

Fine-tuning on commonly used datasets (Table 13). We compare fine-tuning accuracies in three types of dataset configurations. For supervised learning (SL) we compare pretraining on ImageNet-1k [16] and Places-365 [59], for self-supervised learning (SSL) we compare DINO [7] and masked auto-encoder (MAE) [22] pre-trained on ImageNet and PASS, and models trained on the fractal datasets FDSL on FractalDB-1k and RCDB-1k. Following the setting of one-instance per category [35], we also evaluate using random patch augmentation ("w/ Aug."). We compare our `1p-frac` with 1-image SSL [3] with DINO and MAE (we use Real Img #1 of Figure 5). The results show that our `1p-frac` pre-training with one single image demonstrated high fine-tuning accuracies (87.5% and 88.2% with and without random patch augmentation) which are close to the performance of self-supervised ImageNet-1k pre-training with MAE that utilizes 1.28M images and yields 88.5%. Surprisingly, despite using only one fractal image with perturbations, our `1p-frac` without random patch augmentation has demonstrated fine-tuning accuracies that are equal to or even higher than that of conventional OFDB-1k/FractalDB-1k, which uses 1,000 parameter sets with fractal categories and 1,000/1,000,000 images: `1p-frac` obtains mean performances of 87.5% vs OFDB-1k 87.2% vs FractalDB-1k 87.1%.

Fine-tuning on specialized and structured datasets in VTAB [54] (Table 14). We also verify performance on a couple of specialized and structured

Table 13: Comparison among pre-training methods in fine-tuning accuracies. Best values at each dataset scale are in bold. ViT-T is used for all experiments. #OrgImg indicates number of collected images in real-image datasets or number of original images used in synthetic datasets. Type shows supervised types within supervised learning using cross-entropy loss (SL), self-supervised learning using DINO [7] or MAE [22] (SSL:D/SSL:M), and formula-supervised learning (FDSL).

Pre-training	#OrgImg	Type	C10	C100	Cars	Flowers	VOC12	P30	IN100	Mean
Scratch	–	–	80.3	62.1	13.7	71.4	56.2	76.2	74.8	62.1
Places-365 [59]	1.80M	SL	97.6	83.9	89.2	99.3	84.6	–	89.4	–
ImageNet-1k [16]	1.28M	SL	**98.0**	85.5	**89.9**	**99.4**	**88.7**	80.0	–	–
ImageNet-1k [16]	1.28M	SSL:D	97.7	82.4	88.0	98.5	74.7	78.4	89.0	86.9
ImageNet-1k [16]	1.28M	SSL:M	97.4	**85.8**	86.6	96.3	83.3	**80.2**	**90.0**	**88.5**
PASS [2]	1.43M	SSL:D	97.5	84.0	86.4	98.6	82.9	79.0	82.9	87.8
FractalDB-1k [30]	1.00M	FDSL	96.8	81.6	86.0	98.3	80.6	78.4	88.3	87.1
RCDB-1k [28]	1.00M	FDSL	97.0	82.2	86.5	98.9	80.9	79.7	88.5	87.6
ImageNet-1k [35]	1,000	SL	94.3	76.9	57.3	94.8	73.8	78.2	84.3	79.9
ImageNet-1k [35]	1,000	SSL:D	94.9	78.0	71.2	94.6	75.5	78.6	84.9	82.5
OFDB-1k [35]	1,000	FDSL	96.9	84.0	84.5	97.1	79.9	79.9	88.0	87.2
3D-OFDB-1k [35]	1,000	FDSL	97.1	83.8	85.5	**98.4**	80.8	80.0	89.1	87.8
OFDB-1k w/ Aug. [35]	1,000	FDSL	**97.2**	**85.3**	**87.6**	98.3	**81.4**	**80.4**	**89.5**	**88.5**
3D-OFDB-1k w/ Aug. [35]	1,000	FDSL	97.0	84.7	85.6	98.3	81.2	79.8	88.9	87.9
1-image SSL [3]	**1**	SSL:D	95.7	79.5	73.1	93.8	69.0	80.4	88.7	82.8
1-image SSL [3]	**1**	SSL:M	**97.2**	**84.7**	82.1	94.6	78.5	80.3	**89.0**	86.6
`1p-frac` (ours)	**1**	FDSL	96.9	84.2	84.5	97.4	80.5	**80.6**	**89.0**	87.5
`1p-frac` w/ Aug. (ours)	**1**	FDSL	96.5	**84.7**	**87.0**	**98.1**	**80.9**	80.5	88.9	**88.2**

datasets in VTAB (visual task adaptation benchmark; Retinopathy [26], Resisc45 [13], Camelyon [48], CLEVR-Count [25], and sNORB-Azim [33]). The experiment is more suitable to evaluate and compare the pre-training methods since the image domains and their labels are distinct from both real and fractal images on ImageNet and `1p-frac`. According to Table 14, our `1p-frac` pre-trained model performed relatively similar fine-tuning accuracies on all listed datasets in the table. By comparing to self-supervised ImageNet-1k with MAE, our `1p-frac` surpassed the fine-tuning accuracies on three (Retinopathy, Resisc45, and sNORB-Azim) out of the five datasets.

5 Conclusion and Discussion

This paper examined what a minimal dataset for synthetic pre-training might look like. We proposed the 1-parameter Fractal as Data (`1p-frac`) dataset that succeeds in pre-training even with a single fractal by utilizing the locally perturbed cross entropy (LPCE) loss. The following findings were presented:

What is a minimal synthetic pre-training dataset? In this paper, we applied IFS to generate labeled images. In this context, preparing a single set of parameters $a - f$ in IFS is the least informative. It recorded fine-tuning accuracy nearly equivalent to conventional methods such as FractalDB and OFDB (Tables 1 and 2), succeeding in pre-training with minimal image pattern utiliza-

Table 14: Performance comparisons on five fine-tuning datasets: Retinopathy (Retino) [26], Resisc45 [13], Camelyon [48], CLEVR-Count (CLEVR-C) [25], and sNORB-Azim (sNORB-A) [33]. Best and second-best values are in bold and underlined, respectively.

	#OrgImg	Type	Retino	Resisc45	Camelyon	CLEVR-C	sNORB-A
Scratch	–	–	66.9	89.2	74.4	46.4	11.5
ImageNet-1k [16]	1.28M	SL	**79.1**	**97.0**	<u>82.7</u>	<u>89.3</u>	**29.2**
ImageNet-1k [16]	1.28M	SSL:M	76.7	95.6	**83.9**	**89.5**	26.3
`1p-frac` (ours)	**1**	FDSL	<u>77.9</u>	<u>95.8</u>	<u>82.7</u>	86.2	<u>28.9</u>

tion. Moreover, it is not that any perturbation in image space is acceptable, as Gaussian/uniform distributions failed in pre-training (Table 4). It clarified the importance of geometric transformations with certain regularities like fractal geometry or contour shapes (Table 3).

What properties of images lead to image reduction? In `1p-frac`, we investigated shape changes and pre-training effects by adopting the LPCE loss and the σ-factor that contribute to fractal shape variation. It was revealed that pre-training efficacy varies with shape variation (Tables 5, 6). For real images, it is not just about including more edge components, but about including affine transformations, which closely resembles the classification of labeled fractal images through IFS (Tables 11 and 12). By extensively ablating data augmentation showed, we demonstrated the essential role of random cropping and MixUp.

Pre-training effects with the proposed method. Supervised pre-training on ImageNet was more accurate for datasets where object labels are assigned to real images (Table 13), but in VTAB's Specialized/Structured datasets (Table 14), performances were relatively close. On the other hand, our `1p-frac` pre-trained ViT performed a similar mean accuracy by comparing to MAE self-supervised ImageNet pre-training (MAE 88.5 vs. our `1p-frac` w/ Aug. 88.2). Furthermore, increasing the number of sampling points L showed performance improvement even with 21k categories and demonstrated good performance in ImageNet-1k fine-tuning (Table 8), indicating the potential for building large models with limited data resources.

Broader impact. Our proposed dataset `1p-frac` does not suffer from licensing or ethical issues such as biases and has a large potential to serve as a clean pretraining dataset. While not investigated in this paper, this approach also opens the possibility for significant gains in training speed by keeping the data on the GPU and applying transformations there, removing the GPU-CPU transfer bottleneck.

Limitation. In this paper, we have explored achieving a minimal synthetic pre-training in image recognition based on fractal geometry with IFS (a single fractal, and their perturbations). Despite this, we believe it is also necessary to find a similarly minimal pre-training dataset containing real images – for some applications and increased interpretability. This might allow for quick and clear calibration of pre-training, similar to how calibration images are used in photography. We did not investigate this direction in this paper but leave this for future work.

Acknowledgements

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

1. Anderson, C., Farrell, R.: Improving fractal pre-training. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1300–1309 (2022)
2. Asano, Y., Rupprecht, C., Zisserman, A., Vedaldi, A.: Pass: An imagenet replacement for self-Supervised pretraining without humans. In: Proceedings of the 2021 Neural Information Processing Systems Track on Datasets and Benchmarks (2021)
3. Asano, Y.M., Rupprecht, C., Vedaldi, A.: A critical analysis of self-supervision, or what we can learn from a single image. In: Proceedings of the 2020 International Conference on Learning Representations. (ICLR) (2020)
4. Baradad, M., Chen, C.F., Wulff, J., Wang, T., Feris, R., Torralba, A., Isola, P.: Procedural image programs for representation learning. In: Proceedings of the 2022 Advances in Neural Information Processing Systems (NeurIPS) (2022)
5. Baradad, M., Wulff, J., Wang, T., Isola, P., Torralba, A.: Learning to See by Looking at Noise. In: Proceedings of the 2021 Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 2556–2569 (2021)
6. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **8**(6), 679–698 (1986)
7. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
8. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10708–10717 (2022)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 2020 International Conference on Machine Learning (ICML). vol. 119, pp. 1597–1607 (2020)
10. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15750–15758 (2021)
12. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9640–9649 (2021)
13. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
14. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical data augmentation with no separate search. arXiv preprint abs/2104.07918 (2019)

15. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme Ruiz, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., Steenkiste, S.V., Elsayed, G.F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M., Gritsenko, A.A., Birodkar, V., Vasconcelos, C.N., Tay, Y., Mensink, T., Kolesnikov, A., Pavetic, F., Tran, D., Kipf, T., Lucic, M., Zhai, X., Keysers, D., Harmsen, J.J., Houlsby, N.: Scaling vision transformers to 22 billion parameters. In: Proceedings of the 2023 International Conference on Machine Learning. vol. 202, pp. 7480–7512 (2023)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 248–255 (2009)
17. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the 2015 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1422–1430 (2015)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the 2021 International Conference on Learning Representation (ICLR) (2021)
19. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1310–1319 (2017)
20. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision (IJCV) **111**(1), 98–136 (2015)
21. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representation (ICLR) (2018)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020)
24. Inoue, N., Yamagata, E., Kataoka, H.: Initialization using perlin noise for training networks with a limited amount of data. In: Proceedings of the 2020 International Conference on Pattern Recognition (ICPR). pp. 1023–1028 (2020)
25. Johnson, J., Hariharan, B., Maaten, L.v.d., Li, F.F., Zitnick, L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1988–1997 (2017)
26. Kaggle, Eyepacs: Kaggle diabetic retinopathy detection (2015)
27. Kataoka, H., Hara, K., Hayashi, R., Yamagata, E., Inoue, N.: Spatiotemporal initialization for 3D CNNs with generated motion patterns. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 737–746 (2022)
28. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E.J., Inoue, N., Yokota, R.: Replacing labeled real-image

datasets with auto-generated contours. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21232–21241 (2022)
29. Kataoka, H., Matsumoto, A., Yamada, R., Satoh, Y., Yamagata, E., Inoue, N.: Formula-driven supervised learning with recursive tiling patterns. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4098–4105 (2021)
30. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. International Journal of Computer Vision (IJCV) **130** (2022)
31. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the 2013 International IEEE Workshop on 3D Representation and Recognition (3DRR-13). pp. 554–561 (2013)
32. Krizhevsky, A., Hinton, G.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, Department of Computer Science, University of Toronto (2009)
33. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. II–104 Vol.2 (2004)
34. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Maaten, L.v.d.: Exploring the Limits of Weakly Supervised Pretraining. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). pp. 181–196 (2018)
35. Nakamura, R., Kataoka, H., Takashima, S., Noriega, E.J.M., Yokota, R., Inoue, N.: Pre-training vision transformers with very limited synthesized images. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20303–20312. IEEE (2023)
36. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the 2008 Indian Conference on Computer Vision, Graphics and Image Processing. pp. 722–729 (2008)
37. Noroozi, M., Favaro, P.: Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). pp. 69–84 (2016)
38. Noroozi, M., Vinjimoor, A., Favaro, P., Pirsiavash, H.: Boosting Self-Supervised Learning via Knowledge Transfer. In: IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9359–9367 (2018)
39. Remez, T., Huang, J., Brown, M.: Learning to Segment via Cut-and-Paste. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). pp. 39–54 (2018)
40. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3) (2015)
41. Sharma, P., Rott Shaham, T., Baradad, M., Fu, S., Rodriguez-Munoz, A., Duggal, S., Isola, P., Torralba, A.: A vision check-up for language models. In: arXiv preprint arXiv:2401.01862 (2024)
42. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 843–852 (2017)

43. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3D orientation learning for 6D object detection from RGB images. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). pp. 712–729 (2018)
44. Tadokoro, R., Yamada, R., Nakashima, K., Nakamura, R., Kataoka, H.: Primitive Geometry Segment Pre-training for 3D Medical Image Segmentation. In: Proceedings of the 2023 British Machine Vision Conference (BMVC) (2023)
45. Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., Yokota, R.: Visual atoms: Pre-training vision transformers with sinusoidal waves. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18579–18588 (2023)
46. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 23–30 (2017)
47. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 2021 International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357 (2021)
48. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., , Welling, M.: Rotation equivariant cnns for digital pathology. In: Proceedings of the 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2018)
49. Venkataramanan, S., Rizve, M.N., Carreira, J., Asano, Y.M., Avrithis, Y.: Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. arXiv preprint arXiv:2310.08584 (2023)
50. Wang, T., Zhu, J., Torralba, A., Efros, A.A.: Dataset Distillation (2018)
51. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031 (2019)
52. Yun-Hao Cao, H.Y., Wu, J.: Training vision transformers with only 2040 images. In: Proceedings of the 2022 European Conference on Computer Vision (ECCV). pp. 220–237 (2022)
53. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1204–1213 (2022)
54. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., Houlsby, N.: A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. arXiv preprint arXiv:1910:04867 (2019)
55. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. arXiv preprint abs/1710.09412 (2017)
56. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 649–666 (2016)
57. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: Proceedings of the 2021 International Conference on Learning Representation (ICLR) (2021)
58. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint abs/1708.04896 (2017)

59. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(6), 1452–1464 (2017)