Pre-training Vision Transformers with Very Limited Synthesized Images –Supplementary Material–

1. Dataset details in our paper used

We introduce sample sets of ImageNet$^{\diamondsuit}$, ImageNet$^{\diamondsuit}$-gray, ImageNet$^{\diamondsuit}$-binary, ImageNet$^{\diamondsuit}$-cany, 2D-OFDB, and 3D-OFDB that we could not show in the paper. You can find more details about FractalDB [5], RCDB [4], and ExFractaLDB [4], which are not presented in this Supplemental Material, in the references in the paper.

2D-OFDB-1k (Figure 1a). We call the dataset 2D-OFD-1k, which consists of samples representing the categories of FractalDB-1k. The full dataset for 2D-OFDB-1k is shown in Figure 1a. The total number of images is 1000, the same as the number of categories in FractalDB-1k. We also share the full 2D-OFDB-1k dataset and its training code in the supplementary material. If you are interested, please try to execute the commands in the shell script.

3D-OFDB-1k(Figure 1b). We call the dataset 3D-OFD-1k, which consists of 3Dmodels representing the categories of ExFractalDB-1k. In training 3D-OFDB-1k, we use images taken from arbitrary viewpoints from these 3Dmodels for pre-training in image recognition. The set of samples taken is Figure 1b). In the pre-training, the viewpoint is changed for each epoch. This paper shows the sample set of rotated yaw angles in Figure 1c.

ImageNet$^{\diamondsuit}$(Figure 2a). We call the dataset ImageNet$^{\diamondsuit}$ from ImageNet-1k, using random instances representing the categories. ImageNet-1k is the 1000-category dataset, so the number of data in ImageNet$^{\diamondsuit}$ is 1000, which is the same as the number of categories in ImageNet-1k. The specific sample set is shown in Figure 2a. In the experiment, we carefully paid attention to the sampling bias of ImageNet$^{\diamondsuit}$, and the results of five times averaging are used as the experimental values.

ImageNet$^{\diamondsuit}$-gray(Figure 2b). We call ImageNet$^{\diamondsuit}$-gray after converting ImageNet$^{\diamondsuit}$ into a gray scale [3] image. If you want to get more specific on the process, please refer to grayscale processing [3].

Table 1: Relationship between batch size and accuracy on 2D/3D-OFDB-1k.

Pre-training	64	128	256	512
2D-OFDB-1k	80.4	82.3	**84.0**	83.7
3D-OFDB-1k	77.1	81.0	**83.8**	82.8

Table 2: Comparison of ViT, gMLP, and ResNet with 2D-OFDB-1k and 3D-OFDB-1k pre-training.

PT Dataset	Arch.	C10	C100	Cars	Flowers
2D-OFDB-1k	ResNet	95.6	79.3	76.5	92.4
2D-OFDB-1k	gMLP	95.3	79.1	84.2	**97.1**
2D-OFDB-1k	ViT	**96.9**	**84.0**	**84.5**	**97.1**
3D-OFDB-1k	ResNet	95.2	78.9	76.1	95.3
3D-OFDB-1k	gMLP	95.3	78.8	83.3	97.3
3D-OFDB-1k	ViT	**97.1**	**83.8**	**85.5**	**98.4**

ImageNet$^{\diamondsuit}$-gray is the dataset of 1000 images as well as ImageNet$^{\diamondsuit}$. A specific set is shown in Figure 2b. In the experiment, the five averages are given in the table, as in ImageNet$^{\diamondsuit}$.

ImageNet$^{\diamondsuit}$-binarry (Figure 2c). We call ImageNet$^{\diamondsuit}$-gray the image converted from ImageNet$^{\diamondsuit}$ to the image applying Otsu Method [7]. If you want to learn more about the specific process, please refer to the Otsu method's paper [7]. Specific sets are shown in Figure 2c. In the experiment, the five averages are listed in the table as in ImageNet$^{\diamondsuit}$. .

ImageNet$^{\diamondsuit}$-canny (Figure 2d). We call the image converted from ImageNet$^{\diamondsuit}$ to an image applying the canny edge method [1] ImageNet$^{\diamondsuit}$-canny. You can see the specific process description at the canny edge process. The parameters required for canny edge are the same as those for canny edge provided by open–cv. ImageNet$^{\diamondsuit}$-canny is a dataset of 1000 images, similar to ImageNet$^{\diamondsuit}$. The specific sample set is shown in Figure 2d. In the experiment, as with ImageNet$^{\diamondsuit}$, the five averages are shown in the table.

Table 3: Hyper-parameters of pre-training and fine-tuning in our experiments. Basically, they are same as the configuration used by Kataoka *et al.* [4]. *: When pre-training the 21k dataset for ViT-T, LR is 1.0e-3 for Imagenet-21k$^{\diamond}$ only. Also, when pre-training the ViT-B 21k dataset, LR is 2.5e-4 for 2D-OFDB21k and 1.0e-3 for 2D-OFDB21k.

Training Step	Pre-training			Fine-tuning	
Model	ViT-T		ViT-B	ViT-T/B	
Dataset Category	1k	21k	21k	1k	Others
Image per Category	One/Full	One/Full	One/Full	Full	Full
Epochs	80000/300	90	90	300	1000
Batch Size	256/1024	1024/8192	1024/8192	1024	768
Optimizer	AdamW	AdamW	AdamW	AdamW	SGD
LR	1.0e-3	5.0e-4*/8.0e-3	5.0e-4*/1.0e-3	1.0e-3	1.0e-2
Weight Decay	0.05	0.05	0.05	0.05	1.0e-4
LR Scheduler	Cosine	Cosine	Cosine	Cosine	Cosine
Warmup Steps	15.238k/5k	15.238k/5k	15.238k/5k	5 (epochs)	10 (epochs)
Resolution	224	224	224	224/384	224
Label Smoothing	0.1	0.1	0.1	0.1	0.1
Drop Path	0.1	0.1	0.1	0.1	0.1
Rand Augment	(9,0.5)	(9,0.5)	(9,0.5)	(9,0.5)	(9,0.5)
Mixup	0.8	0.8	0.8	0.8	0.8
Cutmix	1.0	1.0	1.0	1.0	1.0
Erasing	0.25	0.25	0.25	0.25	0.25

2. Hyper-parameters in our experiments

We show the hyperparameters used in each experiment in Table3. These hyper-parameters are based on the configuration used by Kataoka *et al.* [4]. More fundamentally, they are based on the paper proposing DeiT [10].

In the case of One image per category, we adjust the epochs of the training so that the iterations are aligned. Also, Warmup Steps are adjusted so that the number of times Warmup runs between One/Full are aligned. LR is also adjusted according to the data set in the case of One. Specific LR values are listed in the caption of Table3. For some 21k one instance data sets, the default LR did not lower the training error correctly, and it was difficult to compare legitimately, so we adjusted the LR to lower the error stably. For experiments on datasets with a reduced number of data other than One, Epoch is adjusted so that the number of training updates are aligned, and the other parameters are the values for One in Table3.

In addition, we almost used Github repository[1] published by Kataoka *et al.* for each experiments.

[1]Github repository of Kataoka *et al.* [4] : https://github.com/masora1030/CVPR2022-Pretrained-ViT-PyTorch

3. Additional Experiments

We explore the hyper-parameters in relation to a key parameter in OFDBs. In this subsection, we basically use the data augmentation methods and hyper-parameters used in the paper on the DeiT, unless we mention the changed parameters from those of the DeiT. We use fine-tuning accuracy on CIFAR-100 (C100) as an evaluation measure.

Batch size. Table 1 shows the relationship between the performance rate in OFDB pre-training and batch size. Since the used dataset contains 1,000 images in total, we evaluated batch sizes of {64, 128, 256, and 512} in this experiment. We see that 256 is the best batch size for both 2D-OFDB and 3D-OFDB.

Performance on gMLP and ResNet. Table 2 shows the experimental results for gMLP with a 16×16 patch [6] and ResNet-50 [2] with 2D/3D-OFDB-1k. We employ gMLP-Tiny with a 16×16 patch and ResNet-50. The results show that ViT is more accurate than gMLP or ResNet in 2D/3D-OFDB-1k pre-training.

Category selection with data pruning. We have employed 'accidentally' found fractal categories in FDSL pre-training. However, a one-instance setting in FDSL does not required to augment image instances, that is, it makes easier to evaluate image categories on the FDSL dataset. Therefore, we test whether category selection can improve the pre-training effects of

OFDB with the data pruning method [9]. We used the EL2N score as the metric for data pruning. The EL2N score [8] is calculated by several training steps of the model, but in the case of OFDB, since the dataset size is small, we calculated the EL2N score after 1,000 epochs of training.

We analyzeed tendency of selected categories from 21k to 1k category dataset. results of category selection with data pruning showsed the relationship between easy sample usage with data pruning and fine-tuning accuracy. The results describes the balanced dataset (50:50 with easy:hard samples) recorded the best accuracy (84.35) by comparing to other settings. On the contrary, a dataset consists of hard samples (10:90 with easy:hard samples) is lower fine-tuning accuracy (83.38) than the OFDB-1k pre-trained ViT-T. The 50:50 with easy:hard samples is shown in Figure 3. Also the 10:90 with easy:hard samples is shown in Figure 3a

References

[1] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[3] Christopher Kanan and Garrison Cottrell. Color-to-Grayscale: Does the Method Matter in Image Recognition? *PloS one*, 7:e29740, 01 2012. 1

[4] Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima, Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, and Rio Yokota. Replacing Labeled Real-Image Datasets With Auto-Generated Contours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21232–21241, June 2022. 1, 2

[5] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without Natural Images. volume 130, 2022. 1

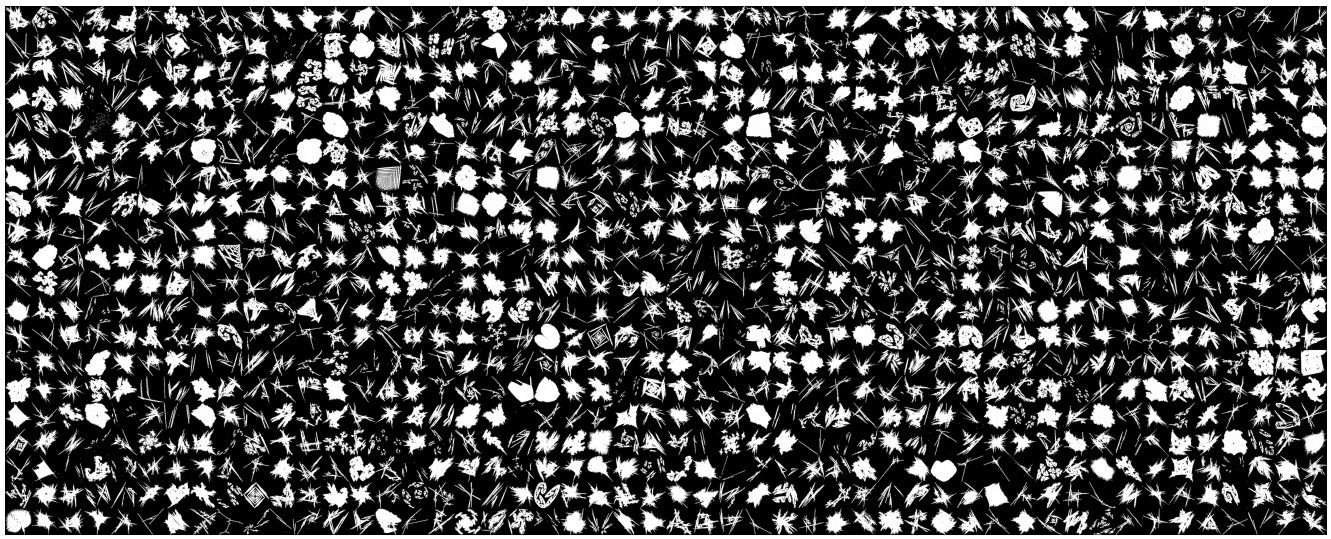
[6] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay Attention to MLPs. *CoRR*, abs/2105.08050, 2021. 2

[7] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 1

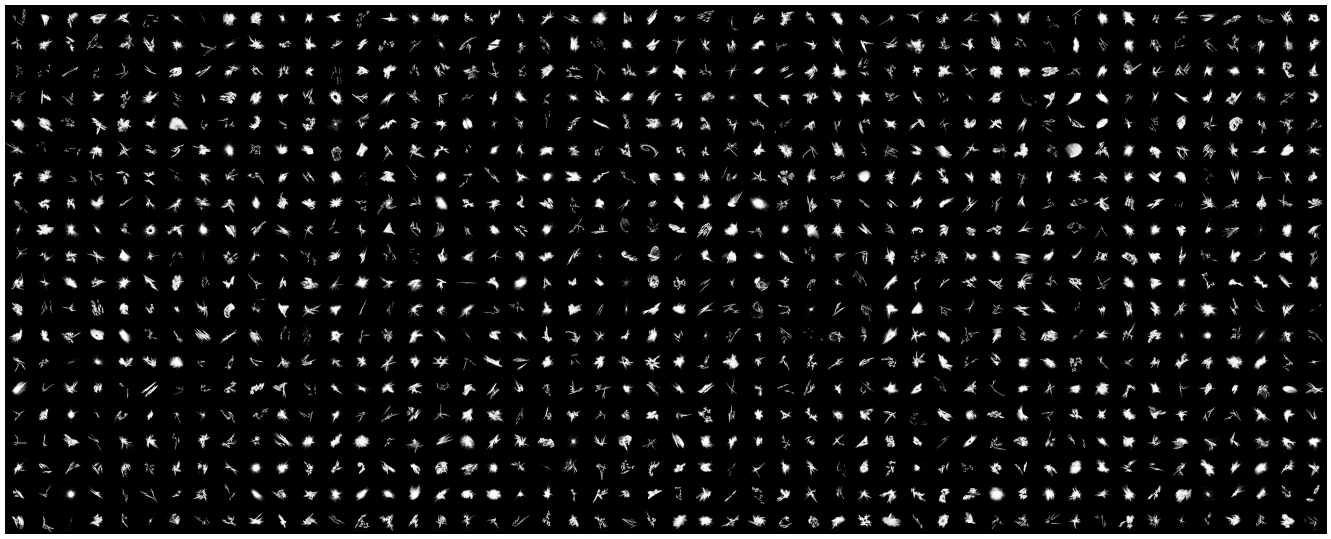
[8] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 3

[9] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

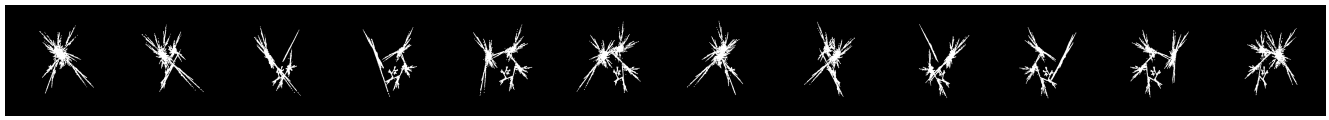
[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 2



(a) Full set of 2D-OFDB-1k.



(b) Sample set of 3D-OFDB-1k.

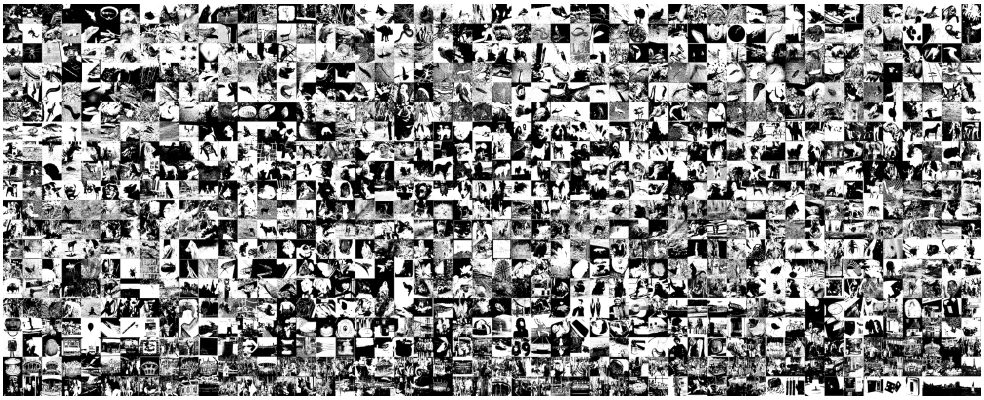


(c) Sample 3D-OFDB data augmentation of yaw rotation.

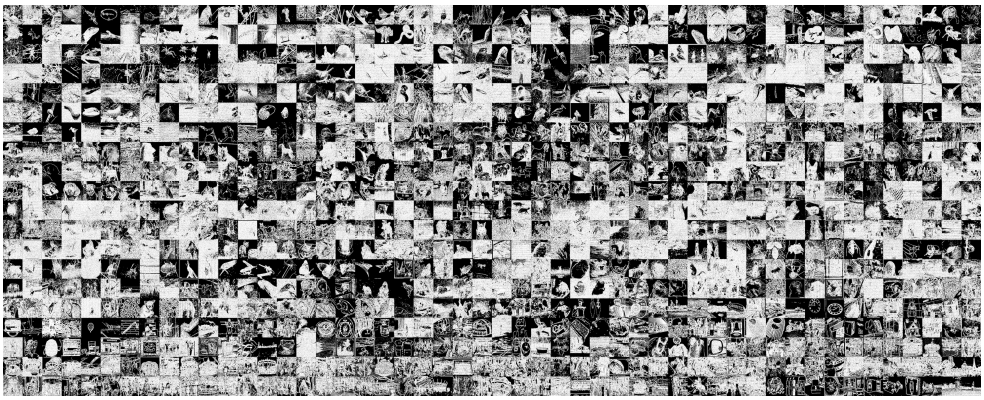
Figure 1: 2D-OFDB-1k and 3D-OFDB-1k

(a) Full set of ImageNet$^{\diamond}$.

(b) Full set of ImageNet$^{\diamond}$–gray.

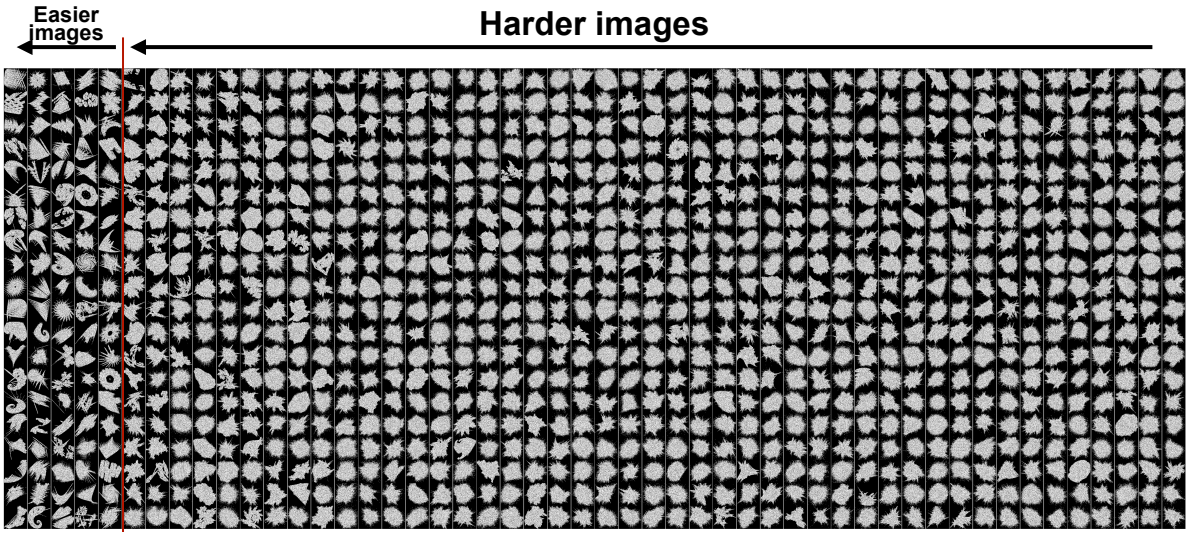


(c) Full set of ImageNet$^{\diamond}$–binary.

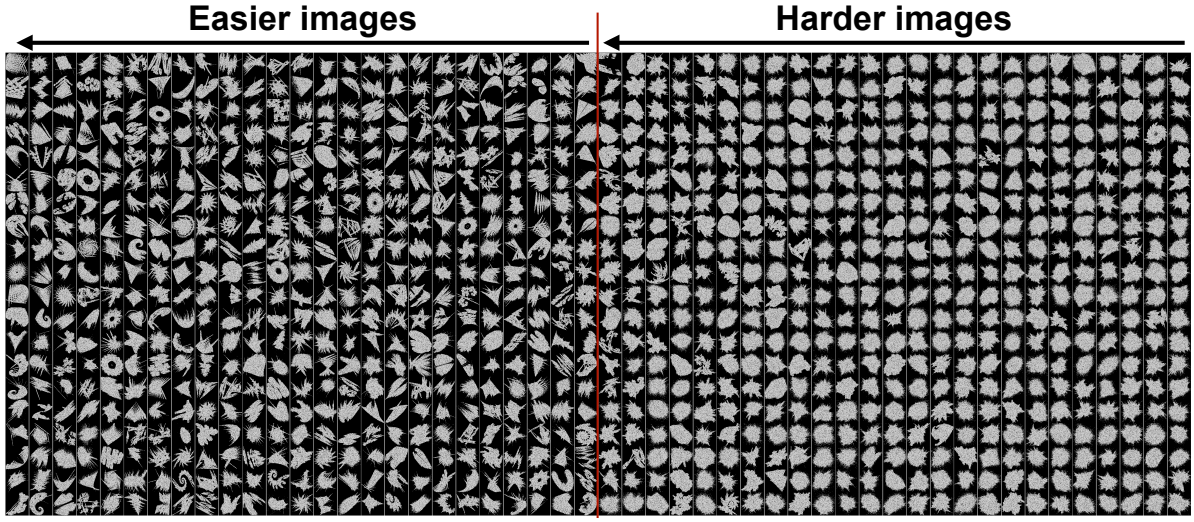


(d) Full set of ImageNet$^{\diamond}$–canny.

Figure 2: Full set of ImageNet$^{\diamond}$-canny.



(a) 10:90 with easy:hard samples



(b) 50:50 with easy:hard samples

Figure 3: Dataset constructed by Data pruning