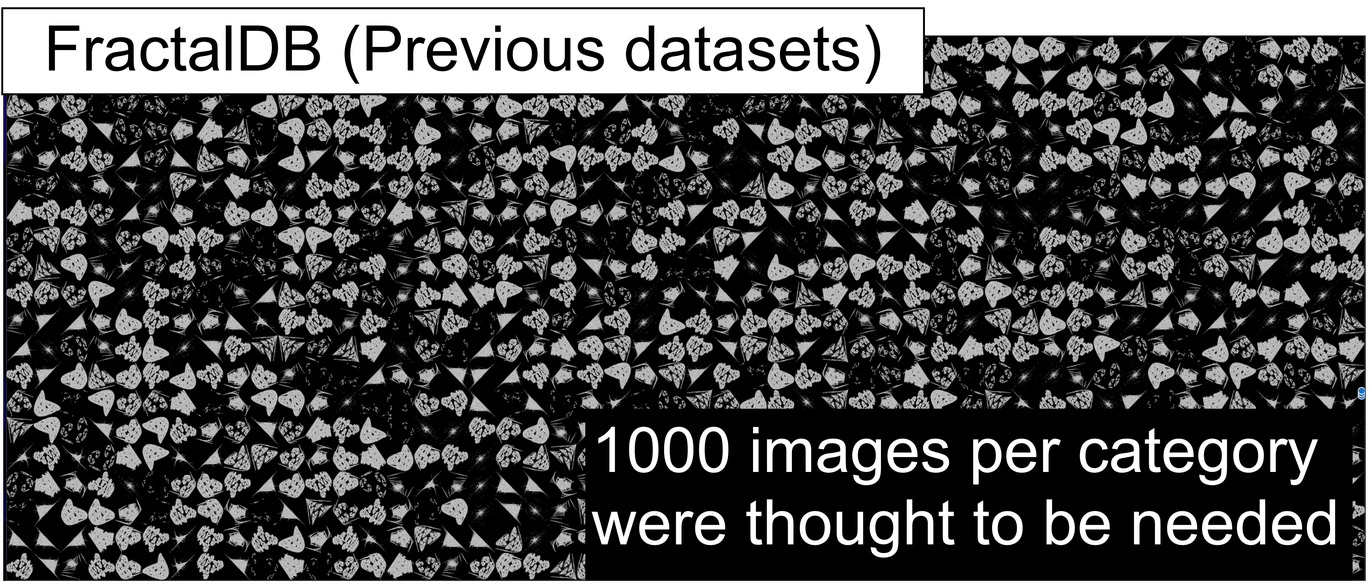
Pre-training Vision Transformers with Very Limited Synthesiszed Images

Ryo Nakamura[1,2], Hirokastu Kataoka[1], Sora Takashima[1,3], Edgar Josafat Martinez Noriega[1,3], Rio Yokota[1,3], Nakamasa Inoue[1,3]
[1]National Institute of Advanced Industrial Science and Technology(AIST), [2]Fukuoka University, [3]Tokyo Institute of Technology

Vision Transformer (ViT) can be pre-trained with 1k synthesized images!

Number of Images required for ViT pre-training：
JFT (300M), ImageNet-1k/21k (1.28M / 14M), FractalDB-1k/21k (1M/21M). OFDB-1k/21k (1k/21k)

FractalDB (Previous datasets)



Contribution

Same level of accuracy

GPU hours 78% faster

High performance with limited data pre-training (For ViT)

2D-OFDB

Proposal methods

Why can pre-train with 1 image per category? Update

FractalDB generates 1000 images from a single fractal image ↓
Minibatch data augmentation to perform essentially the same conversion

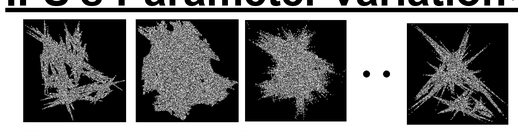
Pattern augmentation

Texture augmentation

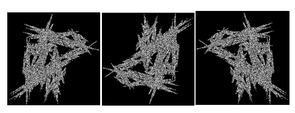
Experiment：Verification performance of Data augmentation

DeiT	✓	✓	✓	✓	✓	✓
IFS		✓				
Rotation			✓			
Rand. Pat.				✓		✓
Rand. Text.					✓	✓
2D-OFDB	84.0	81.6	84.1	**85.3**	84.8	84.3
3D-OFDB	83.8	-	-	84.7	**85.1**	**85.1**

IFS's Parameter variation×25

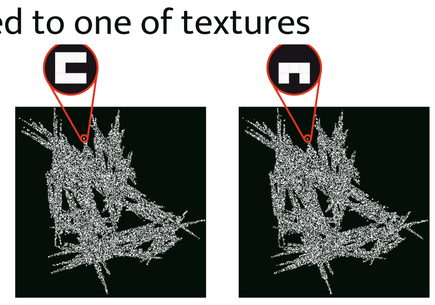


Rotation ×4



Patch pattern×10

Possible to generate a variety Of Fractals with similar shapes



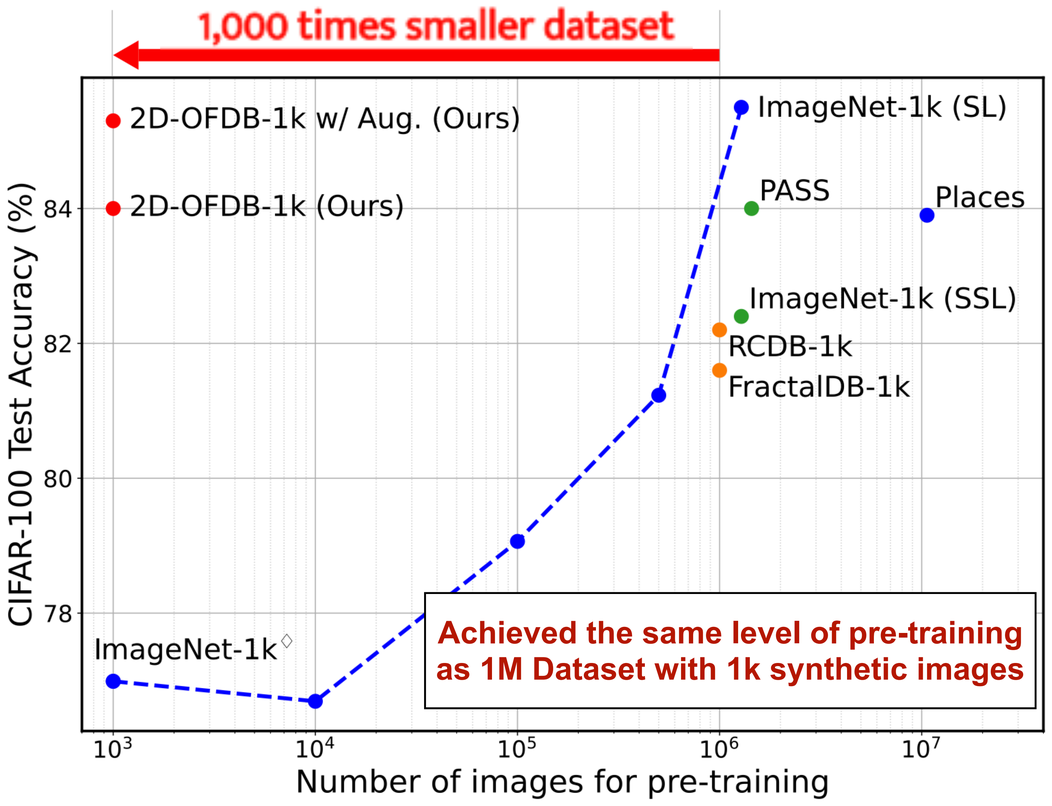
The same accuracy can be achieved with a single image, and the patch augmentation can be further improved.

Experiment

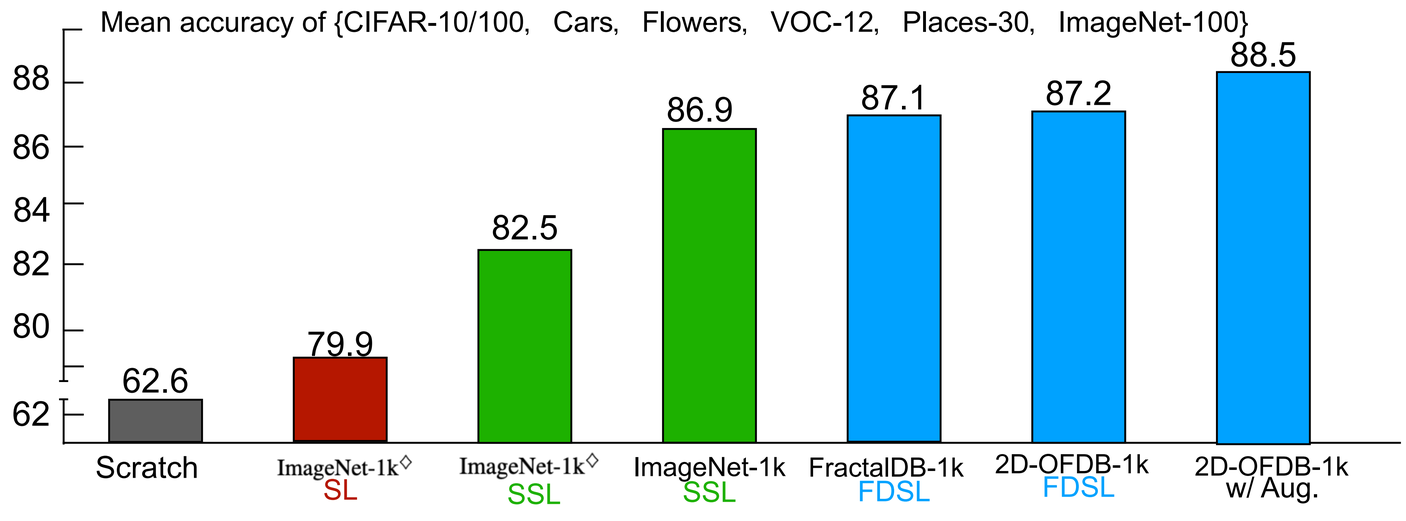
Experimental setting

Verify pre-training effectiveness of datasets：Transition training using pre-trained network weights as initial values

Evaluated by test performance on the dataset for transfer learning, data augmentation use DeiT setting.



Comparison of pre-training effect with 1k category dataset



Accuracy close to 1,000,000 images despite 1,000 images of data

Comparison of ViT prior learning with SoTA on limited data

Pre-training	#Img	Flowers	Pets	DTD	Indoor-67	CUB	Aircraft	Cars	Average
Scratch	–	76.4	67.2	44.2	58.7	54.4	23.0	78.6	57.5
SimCLR [10]	2,040 – 8,144	90.1	82.8	62.3	66.6	**68.5**	74.4	89.3	76.3
IDMM [42]	2,040 – 8,144	92.4	83.2	66.9	68.5	**69.8**	73.4	87.8	77.4
IDMM-ImageNet [42]	2,040	90.5	82.4	66.8	**68.8**	66.8	91.8	87.6	79.2
2D-OFDB-1k (ours)	1,000	**93.7**	**84.6**	**67.5**	66.1	67.7	**95.0**	**91.0**	**80.8**

Higher accuracy with smaller amount of data than IDMM with synthetic images

Comparison of pre-training effectiveness with 21k category data sets

Pre-training	#Img	Type	ViT-T	ViT-B	GPU hours	Batch	#Iterations
Scratch	–	–	72.6	79.8	–	–	–
ImageNet-21k	14M	SL	**74.1**	81.8	3,657	8,192	300k
FractalDB-21k	21M	FDSL	73.0	81.8	5,120	8,192	300k
ImageNet-21k◇	21k	SL	71.0	81.1	1,132	1,024	300k
2D-OFDB-21k	21k	FDSL	**73.8**	82.2	1,088	1,024	300k

Scaling Data Improves AccuracyViT-B exceeds ImageNet-21k Pre-training time reduced by 78

Enables ViT pre-training for anyone with limited data and computational resources