REPLACING LABELED REAL-IMAGE DATASETS WITH AUTO-GENERATED CONTOURS

Hirokatsu Kataoka[*], Ryo Hayamizu[*], Ryosuke Yamada[*], Kodai Nakashima[*], Sora Takashima[*,**], Xinyu Zhang[*,**], Edgar Josafat Martinez-Noriega[*,**], Nakamasa Inoue[*,**], Rio Yokota[*,**]
National Institute of Advanced Industrial Science and Technology (AIST) *, Tokyo Institute of Technology**

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

We show that the performance of FDSL can match that of ImageNet-21k without the use of real images, human-/self-supervision during the pre-training of ViTs.

FORMULA-DRIVEN SUPERVISED LEARNING (FDSL)

- Real images come with privacy/copyright issues and societal biases
- Do we actually need real images to pre-train vision transformers?

To enhance the performance of FDSL, we test the following hypotheses 1 & 2

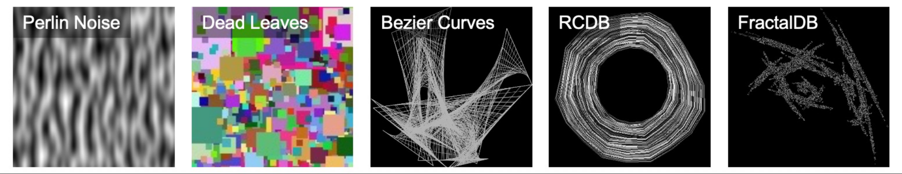
METHOD

- Classes are defined by parameters used to generate the images
- We generate 1k instances per class through transformations
- We use the same parameters as the DeiT paper
- Each of the large-scale runs were performed only once

HYPOTHESIS 1: OBJECT CONTOURS ARE WHAT MATTER IN FDSL DATASETS

- In our preliminary study we found that attention was focused on the outer contours of the fractals
- We created a new dataset that consists only of contours – Radial Contour DataBase (RCDB)
- Despite the lack of any texture, RCDB performed close to FractalDB and outperformed ImageNet-21k

Pre-training	C10	C100	Cars	Flowers
Scratch	78.3	57.7	11.6	77.1
Perlin Noise [21]	95.0	78.4	70.6	96.1
Dead Leaves [3]	95.9	79.6	72.8	96.9
Bezier Curves [21]	96.7	80.3	82.8	98.5
RCDB	**96.8**	**81.6**	84.2	**98.7**
FractalDB [27]	**96.8**	**81.6**	**86.0**	98.3

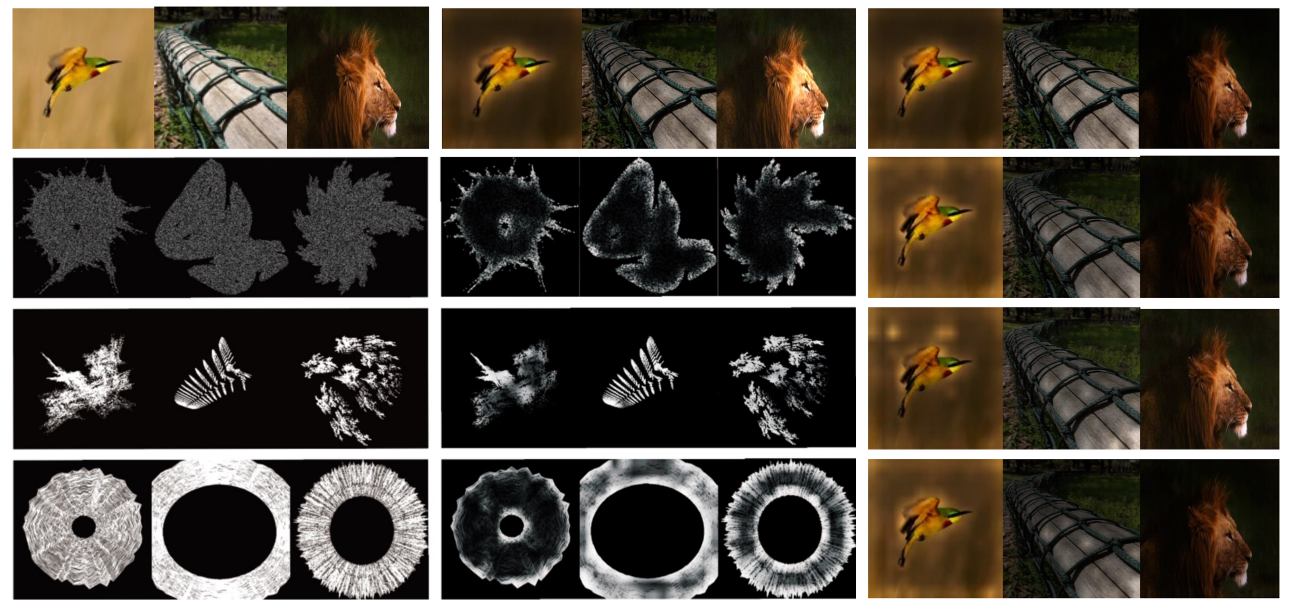


HYPOTHESIS 2: INCREASED NUMBER OF PARAMETERS IN FDSL PRE-TRAINING

- We tested various synthetic datasets with varying complexity of images
- For RCDB, we changed the number of polygons, radius, line width, resizing factor, and Perlin noise
- Complex images increases the difficulty of the pre-training task and leads to better downstream performance

Pre-training	C10	C100	Cars	Flowers
BC	96.9 (0.2)	81.4 (1.1)	85.9 (3.1)	97.9 (-0.6)
RCDB	97.0 (0.2)	**82.2** (0.6)	86.5 (2.4)	**98.9** (0.2)
ExFractalDB	**97.2** (0.4)	81.8 (0.2)	**87.0** (1.0)	**98.9** (0.6)

COMPARISON: IMAGENET-1K / MSCOCO



Pre-training	Fine-tuning @ ImageNet-1k Top-1 w/ ViT-Base
IN-21k ImageNet	81.8
FDB-21k FractalDB	81.8
EFDB-21k ExFractalDB	82.7 (Ours1)
RCDB-21k RadialContourDB	82.4 (Ours2)

Pre-training	COCO Det AP_{50} / AP / AP_{75}	COCO Inst Seg AP_{50} / AP / AP_{75}
Scratch	63.7 / 42.2 / 46.1	60.7 / 38.5 / 41.3
ImageNet-1k	69.2 / 48.2 / 53.0	66.6 / 43.1 / 46.5
ImageNet-21k	**70.7 / 48.8 / 53.2**	**67.7 / 43.6 / 47.0**
ExFractalDB-1k	69.1 / **48.0** / **52.8**	66.3 / **42.8** / 45.9
ExFractalDB-21k	**69.2** / **48.0** / 52.6	**66.4** / **42.8** / **46.1**
RCDB-1k	68.3 / 47.4 / 51.9	65.7 / 42.2 / 45.5
RCDB-21k	67.7 / 46.6 / 51.2	64.8 / 41.6 / 44.7

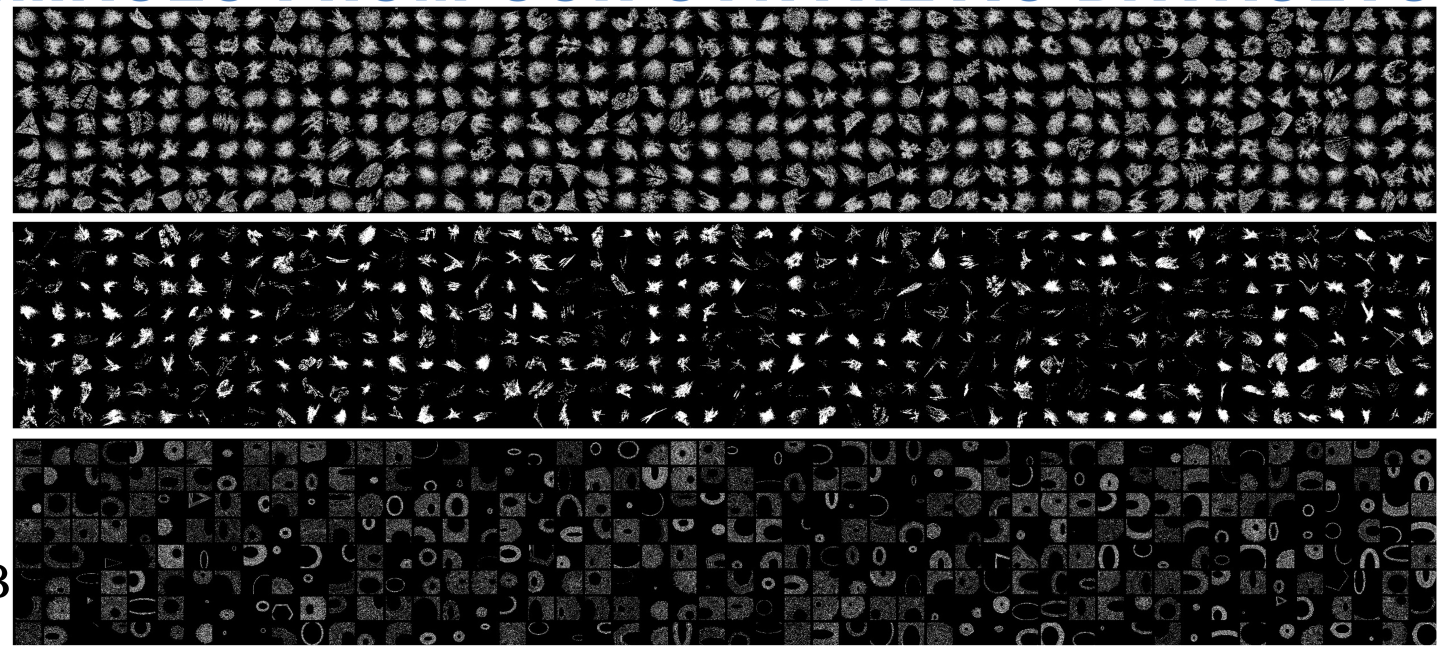
Swin Transformer backbone, Mask R-CNN head, 60 epochs fine-tuning

SAMPLES IMAGES FROM OUR SYNTHETIC DATASETS

FDB-21k FractalDB

EFDB-21k ExFractalDB

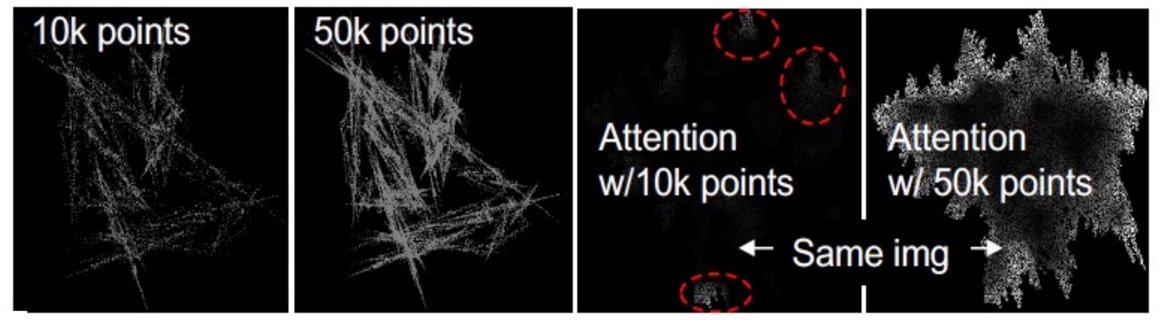
RCDB-21k RadialContourDB

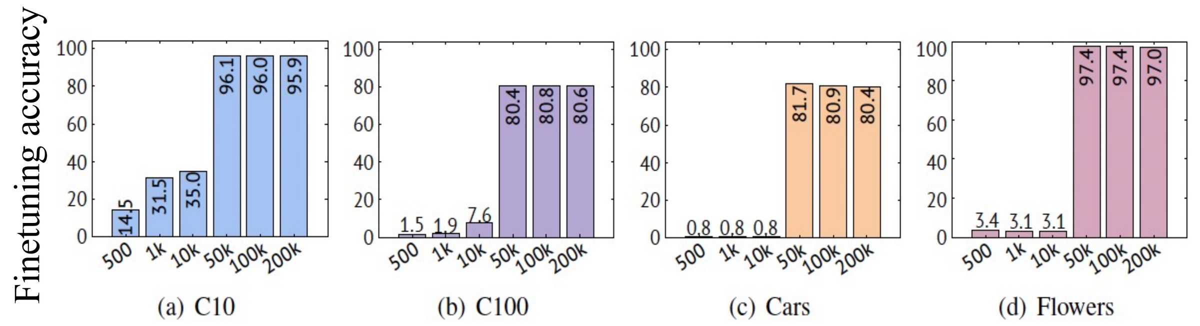


FAILURE MODES OF FDSL

Results, image examples, and attention maps in point-rendered FractalDB and RCDB with corruption

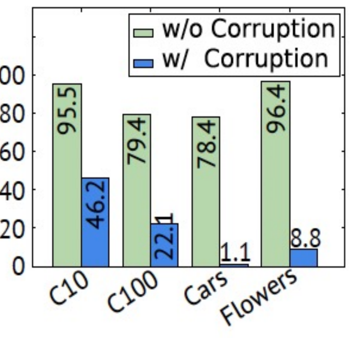
Failure modes in FractalDB





Although the fractal images with 50k+ points successfully trained the visual representations, the fractal images with 10k- points failed

Failure modes in RCDB



- Example of RCDB with broken contours
- Attention maps with the pre-trained models on RCDB w/ and w/o broken contours, respectively

CONCLUSION

- We provided empirical evidence that support our two hypotheses
- Our proposed method can surpass the accuracy of a ViT pre-trained on ImageNet-21k
- We performed ablation studies to identify failure modes of FDSL