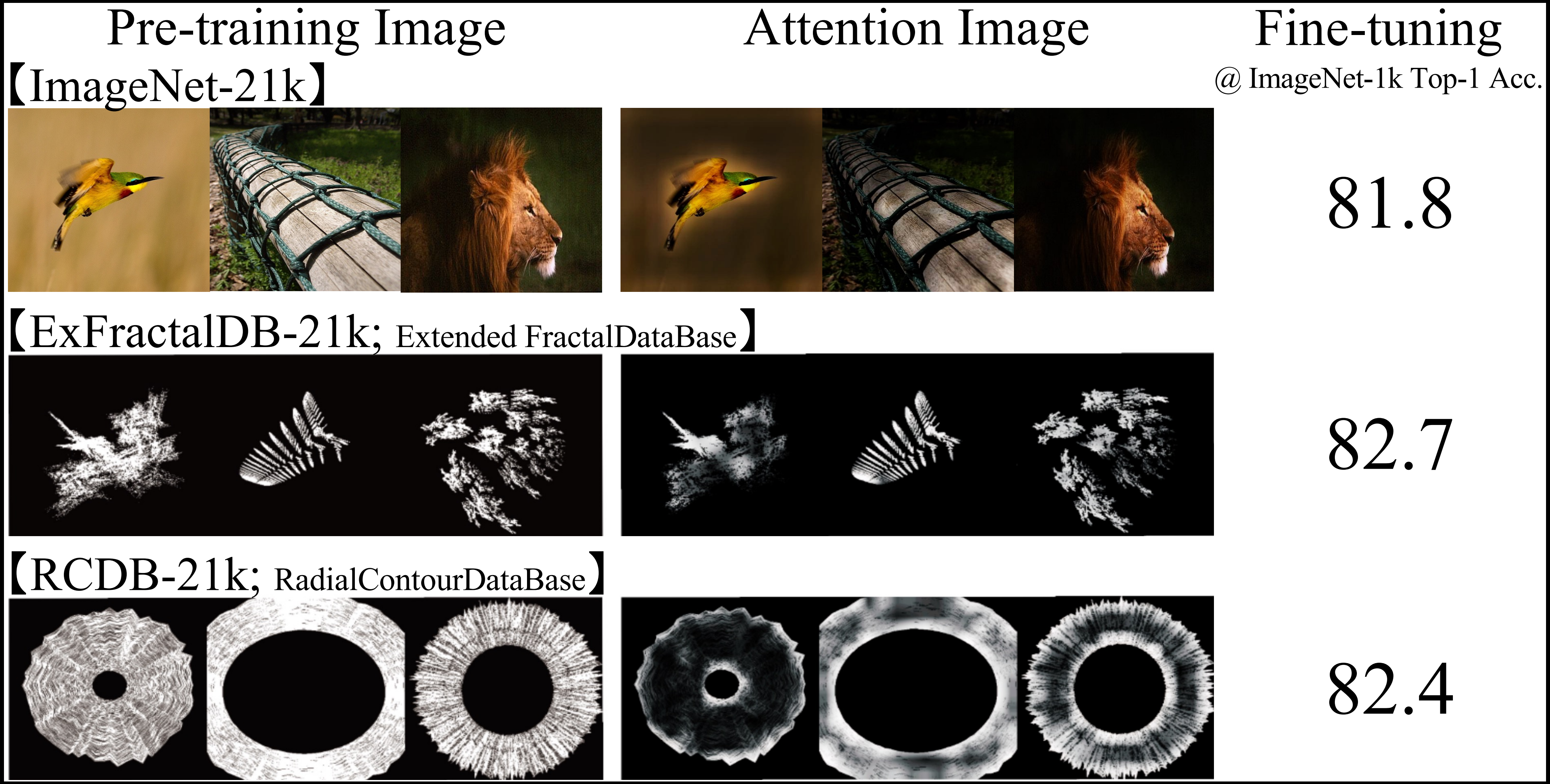
Replacing Labeled Real-image Datasets with Auto-generated Contours

CVPR 2022

Hirokatsu Kataoka[*], Ryo Hayamizu[*], Ryosuke Yamada[*], Kodai Nakashima[*], Sora Takashima[*,**], Xinyu Zhang[*,**], Edgar Josafat MARTINEZ-NORIEGA[*,**], Nakamasa Inoue[*,**], Rio Yokota[*,**]

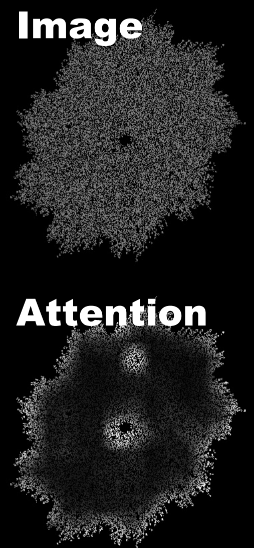
* National Institute of Advanced Industrial Science and Technology (AIST)
**Tokyo Institute of Technology



The performance of FDSL(formula-driven supervised learning) can match or even exceed that of ImageNet-21k without the use of real images, human-and self-supervision during the pre-training of ViT (vision transformers).

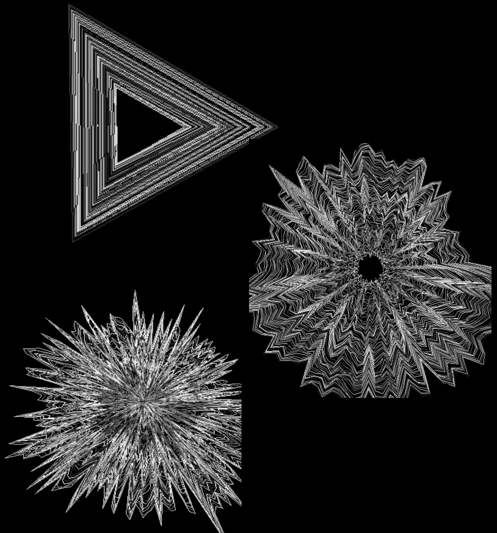
Hypothesis 1:
Object contours are what matter in FDSL datasets

@FractalDB
[Kataoka+, ACCV20]



ViT activated on contours

@RadialContourDB
(RCDB)

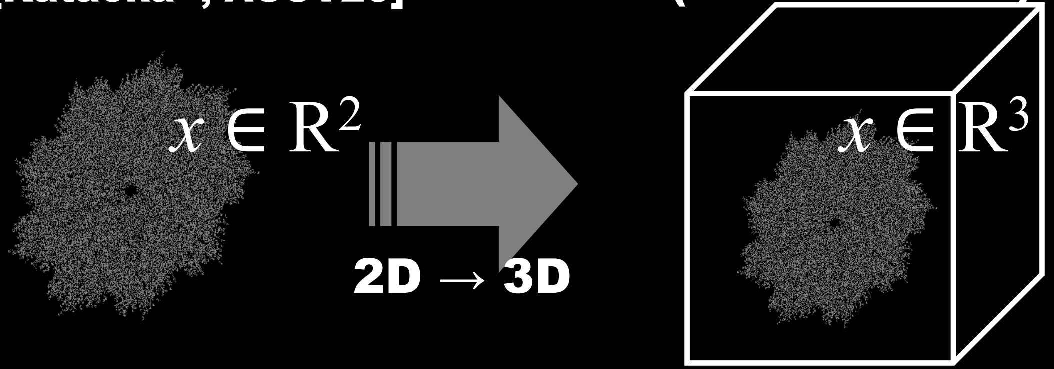


Mainly consists of contours from the activation

Hypothesis 2:
Task difficulty matters in FDSL pre-training

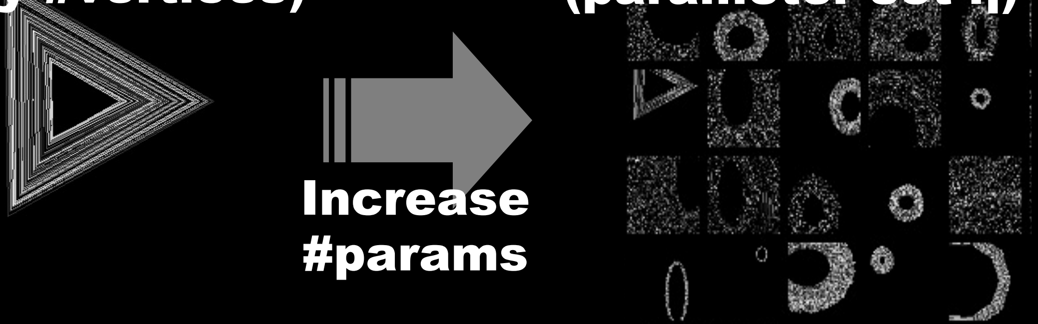
@FractalDB
[Kataoka+, ACCV20]

@Extended FractalDB
(ExFractalDB)



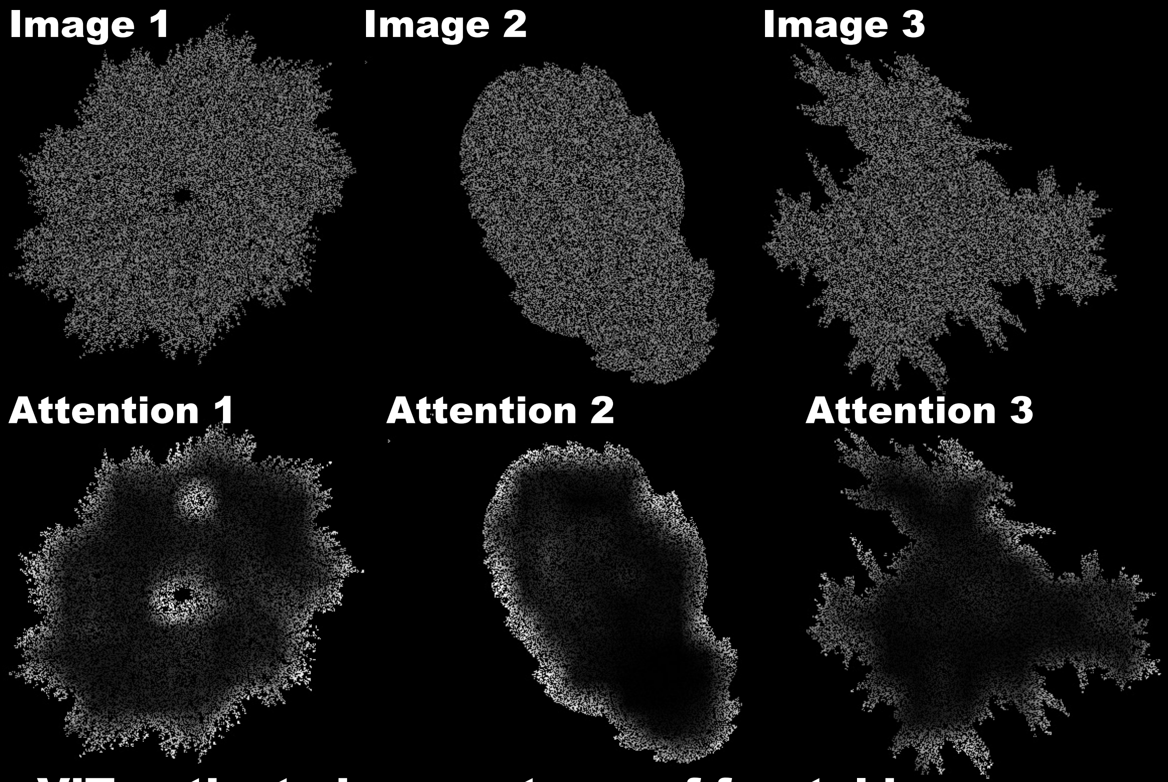
@RadialContourDB
(only #vertices)

@RadialContourDB
(parameter set η)



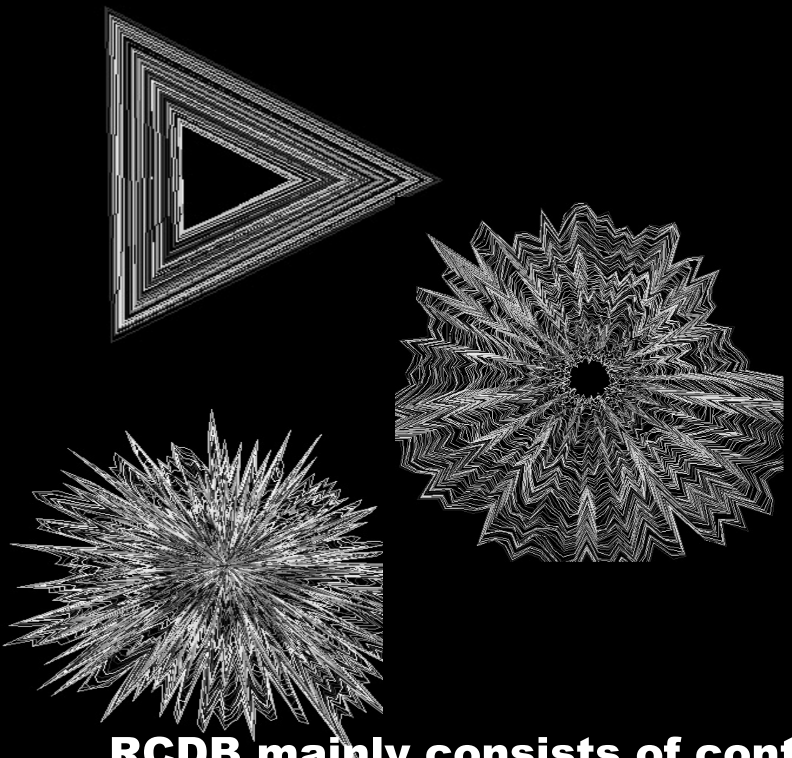
Hypothesis 1:
Object contours are what matter in FDSL datasets

@FractalDB [Kataoka+, ACCV20]



ViT activated on contours of fractal images

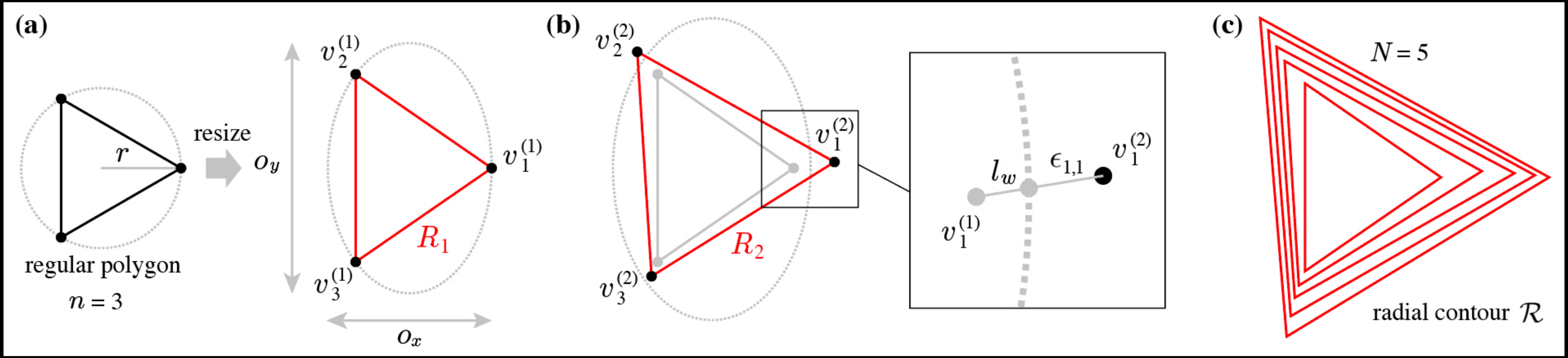
@RadialContourDB (RCDB)



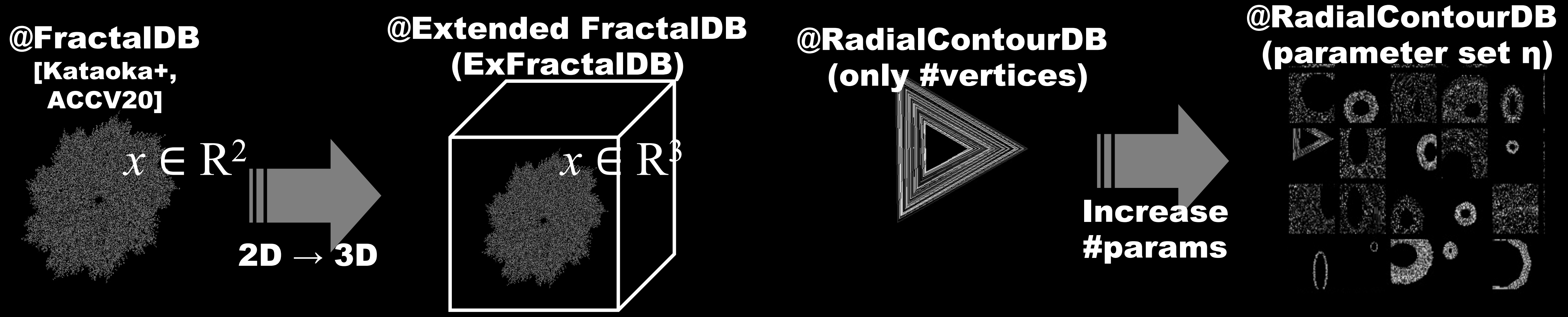
RCDB mainly consists of contours

Hypothesis 1:
Object contours are what matter in FDSL datasets

Procedure for generating radial contours



Hypothesis 2:
Increased number of parameters in FDSL pre-training



Verification of Hypotheses

Table 3. Comparison of FDSL methods. Hereafter, the best values are in bold.

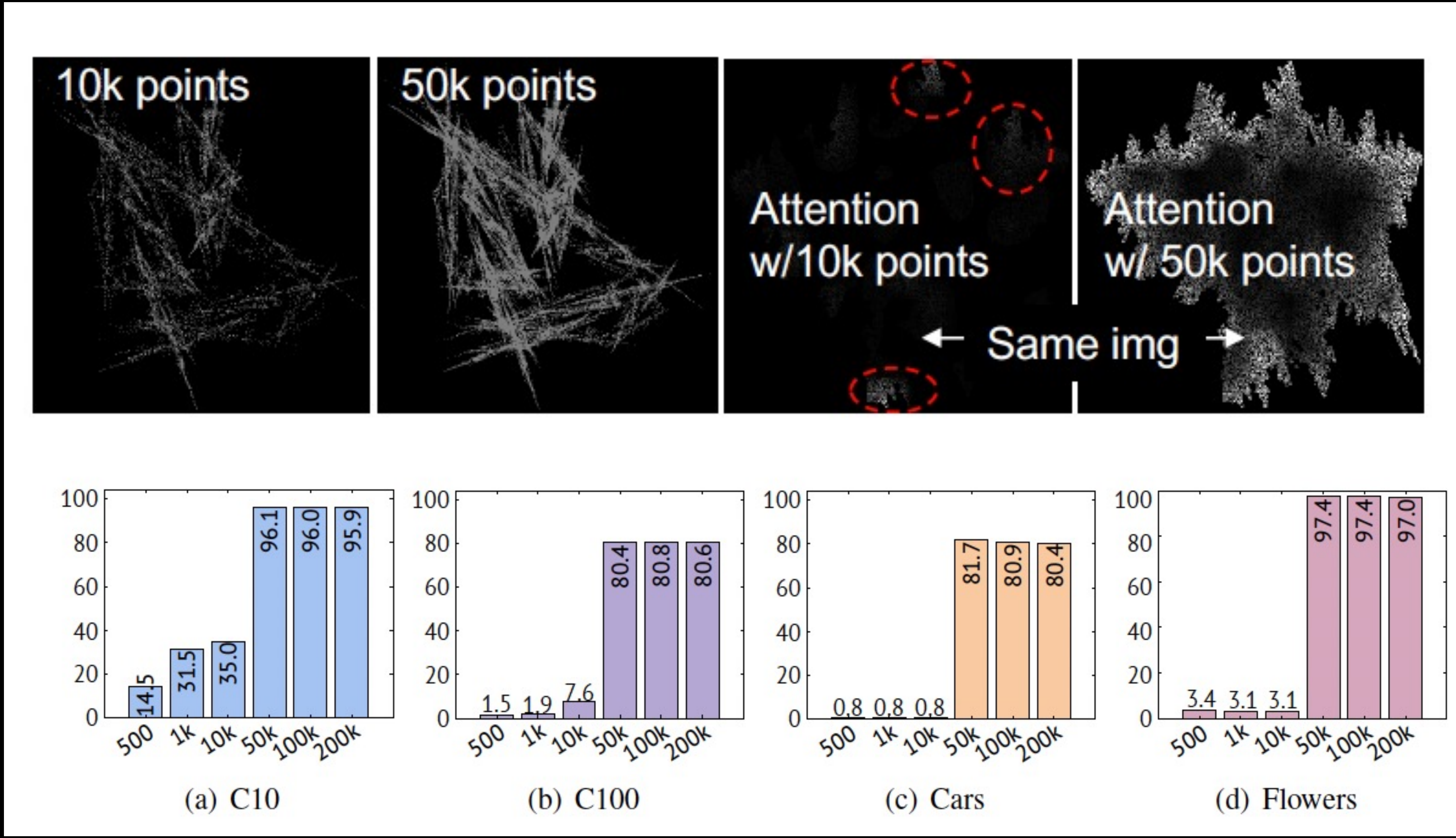
Pre-training	C10	C100	Cars	Flowers
Scratch	78.3	57.7	11.6	77.1
Perlin Noise [21]	95.0	78.4	70.6	96.1
Dead Leaves [3]	95.9	79.6	72.8	96.9
Bezier Curves [21]	96.7	80.3	82.8	98.5
RCDB	**96.8**	**81.6**	84.2	**98.7**
FractalDB [27]	**96.8**	**81.6**	**86.0**	98.3

Regarding Hypothesis 1, we confirm that image representation using object contours tends to yield higher scores: RCDB and FractalDB give the highest accuracy.

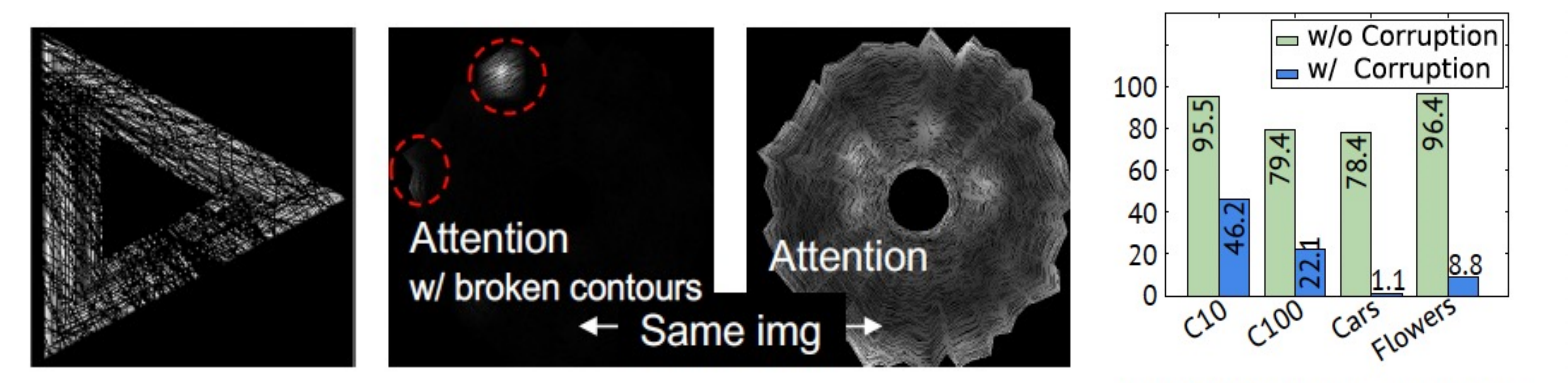
Table 5. Effect of task difficulty by using multiple parameters in FDSL methods. BC stands for Bezier curves. Values in parentheses indicate the difference from the case with fewer parameters.

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)

Regarding Hypothesis 2, we confirm that more difficult tasks improved the accuracy of RCDB and FractalDB (here, ExFractalDB).



In point-rendered FractalDB, although the fractal images with 50k points trained the visual representations, the fractal images with 10k points failed.



At the same time, the RCDB with broken contours failed to acquire a visual representation. The attention and accuracy were also broken from the visualization and result.

Comparison

Pre-training	Img	Type	ViT-Ti	ViT-B
Scratch	–	–	72.6	79.8
ImageNet-21k	Real	SL	**74.1**	81.8
FractalDB-21k	Synth	FDSL	73.0	81.8
FractalDB-50k	Synth	FDSL	73.4	82.1
ExFractalDB-21k	Synth	FDSL	73.6	**82.7**
ExFractalDB-50k	Synth	FDSL	**73.7**	82.5
RCDB-21k	Synth	FDSL	73.1	82.4
RCDB-50k	Synth	FDSL	73.4	**82.6**

- **Pre-training with ExFractalDB-21k (82.7), RCDB-21k (82.4) outperformed that with ImageNet-21k (81.8).**
- **We can match the accuracy of pre-training on ImageNet-21k with synthetic datasets in FDSL.**

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-Base backbone, Mask R-CNN head, 60 epochs fine-tuning

In COCO detection and segmentation, our pre-trained model achieved scores similar to those for the model pre-trained with ImageNet-1k.

Pre-training	Img	Type	C10	C100	Cars	Flowers	VOC12	P30	IN100	Average
Scratch	–	–	78.3	57.7	11.6	77.1	64.8	75.7	73.2	62.6
Places-365	Real	SL	97.6	83.9	89.2	99.3	84.6	–	89.4	–
ImageNet-1k	Real	SL	**98.0**	**85.5**	**89.9**	**99.4**	**88.7**	**80.0**	–	–
ImageNet-1k	Real	SSL (D)	**97.7**	82.4	**88.0**	98.5	74.7	78.4	**89.0**	86.9
PASS	Real	SSL (D)	97.5	**84.0**	86.4	**98.6**	**82.9**	**79.0**	82.9	**87.8**
FractalDB-1k [27]	Synth	FDSL	96.8	81.6	86.0	98.3	80.6	78.4	88.3	87.1
RCDB-1k	Synth	FDSL	97.0	82.2	86.5	98.9	80.9	**79.7**	88.5	87.6
ExFractalDB-1k	Synth	FDSL	97.2	81.8	87.0	**98.9**	80.6	78.0	88.1	87.4
ExFractalDB-1k*	Synth	FDSL	**97.5**	**82.6**	**90.3**	**99.6**	**81.4**	79.4	**89.2**	**88.6**

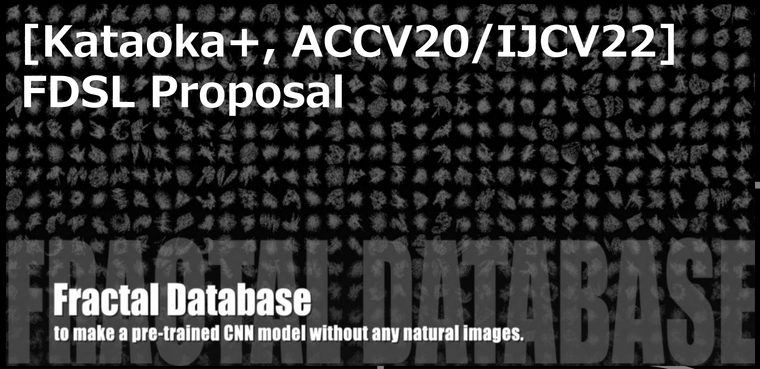
* Rate calculated for 1.4M images, which is the same number of images in PASS dataset..

- **In comparison to SSL, ExFractalDB-1k with 1.4k instances achieved a higher average accuracy (88.6) than that of the self-supervised PASS dataset (87.8).**
- **PASS and FDSL both attempt to improve ethics in datasets.**

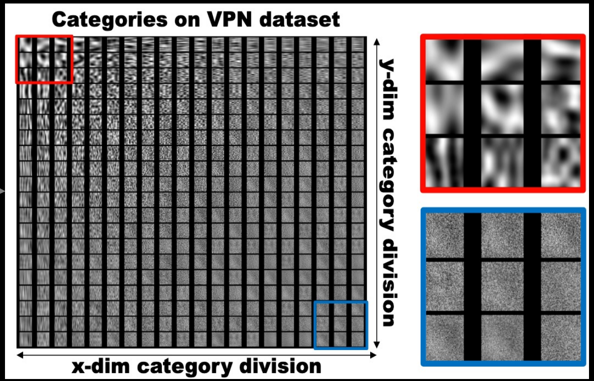
In this paper, one of our major findings is that we can surpass the accuracy of a ViT pre-trained on ImageNet-21k using our FDSL datasets.

We believe that further improvements in contour shapes and a more complex classification task are possible, which leaves open the possibility to scale up the pre-training on synthetic datasets to one day outperform huge-scale datasets (e.g. JFT-300M/3B, IG-3.5B).

[Kataoka+, ACCV20/IJCV22]
FDSL Proposal



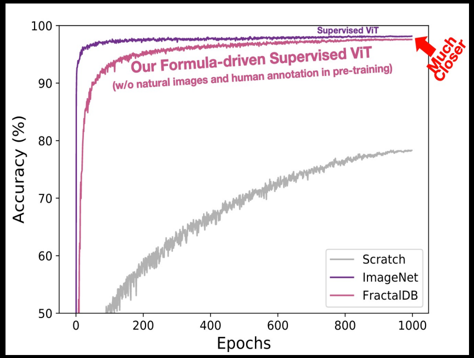
Spatiotemporal Domain



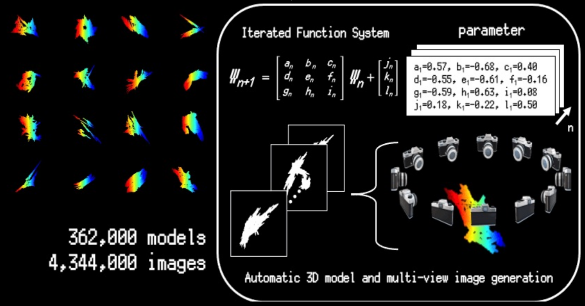
Video Perlin Noise
[Kataoka+, WACV22]

Vision Transformers

3D Domain



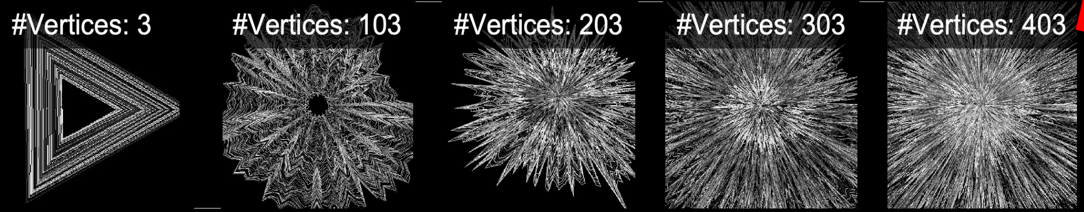
FractalDB Pre-trained ViT
[Nakashima+, AAAI22]



Multi-viewpoint / Point Cloud
[Yamada+, IROS22/CVPR22]

What's Next??

Enhanced by Hypotheses



Replacing Labeled Real-image Datasets (This work)
[Kataoka+, CVPR22]