Pre-training without Natural Images

Hirokatsu Kataoka

AIST

http://www.hirokatsukataoka.net/

Hirokatsu Kataoka

Chief Senior Researcher, Computer Vision Research Team, AIST

Profile :

- Ph.D. in Engineering at Keio University (Mar 2014)
- Chief Senior Researcher, AIST（Apr 2023 - Present）
- PI, cvpaper.challenge (May 2015 – Present; Research community with 1,000+ collaborators)
- Adjunct Researcher, LY Corp. (Oct 2023 - Present)
- Researcher, TICO-AIST Advanced Logistics Lab.（Oct 2016 - Present）
- Researcher, Tokyo Denki University（Apr 2016 - Present）
- Mentor, Tatsujin Program（Nov 2020 - Present）
- Editor, Computer Vision Frontier（Dec 2021 - Present）

片（かた）岡（おか）裕（ひろ）雄（かつ）

Recently Selected Projects (within 2 years) :

“Pre-training Vision Transformers with Very Limited Synthesized Images (ICCV23)”

“SegRCDB: Semantic Segmentation via Formula-Driven Supervised Learning (ICCV23)”

“Visual Atoms: Pre-training Vision Transformers with Sinusoidal Waves (CVPR23)”

”Replacing Labeled Real-Image Datasets with Auto-Generated Contours (CVPR22)”

“Point Cloud Pre-training with Natural 3D Structures (CVPR22)”

“Pre-training without Natural Images (IJCV22)”

“Can Vision Transformers Learn without Natural Images? (AAAI22)”

http:// hirokatsukataoka.net/

@HirokatuKataoka hirokatsukataoka16

Hirokatsu Kataoka

Chief Senior Researcher, Computer Vision Research Team, AIST

1 Pre-training without Natural Images

Representation learning from a natural law

- ACCV 2020 Best Paper Honorable Mention Award
- Accepted to IJCV’22 CVPR’22 ’23, AAAI’22, ICCV’23, BMVC’23 Oral
- MIT Technology Review (Feb. 4th, 2021)
- AIST Best Paper 2022

片（かた）岡（おか）裕（ひろ）雄（かつ）

2 Spatiotemporal 3D ResNet

Strong baseline for 3D convolution in video understanding

- Accepted to CVPR’18（1.9k+ citations; Top 0.5% in 8k+ 5-year CVPR papers)
- AIST Best Paper 2019
- GitHub 3.0k Stars (Top-1 in video recognition at the time of published)

http:// hirokatsukataoka.net/

Pre-training without Natural Images

ACCV 2020 Best Paper Honorable Mention Award
International Journal of Computer Vision (IJCV), 2022
AAAI 2022

Hirokatsu Kataoka

AIST

http://www.hirokatsukataoka.net/

What has the DNNs brought?

Benefits

– Solving various AI tasks, e.g., vision, language, audio, are widely recognized

Challenges in DNN research

– Annotation labor, privacy-preserving on the Internet photos

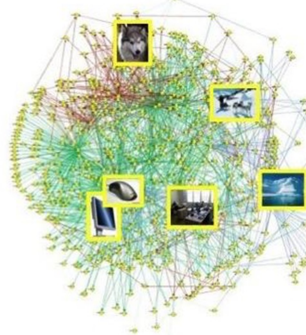
– Ethical issues have occurred

【Large amount of annotation】

http://image-net.org/explore?wnid=n01503061

Takes 2 years, around 50k participants
14M images across 21k categories

【Privacy-preserving】



IMAGENET

http://www.image-net.org/

Privacy is a concern, limiting the use of these image to academic/educational purposes

【Offensive labels】

- 80M Tiny Images had offensive labels
- The dataset was suspended from public access due to the difficulty of labeling and resolution

Issues of annotation & privacy pose significant challenges for AI applications

SL/SSL on huge-scale datasets

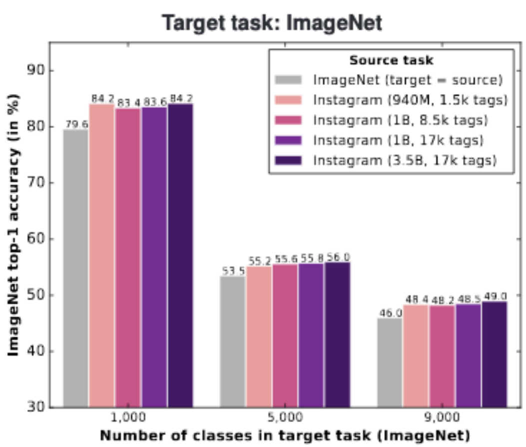
JFT-300M (Google, 2017/2021) / IG-3.5B (Meta, 2018)

300M images / 375M labels

3.5B images / 3.5B weak labels

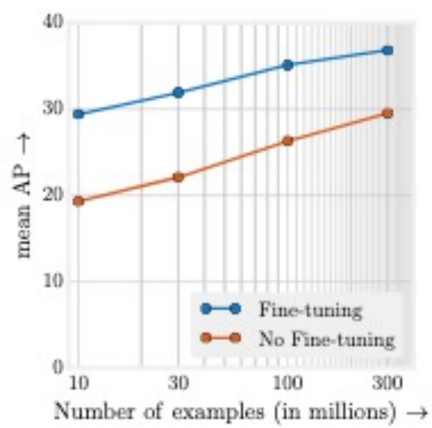
These datasets are x100 larger than ImageNet, improve image representation and recognition performance

-> large-scale datasets benefits both CNN and ViT in pre-training



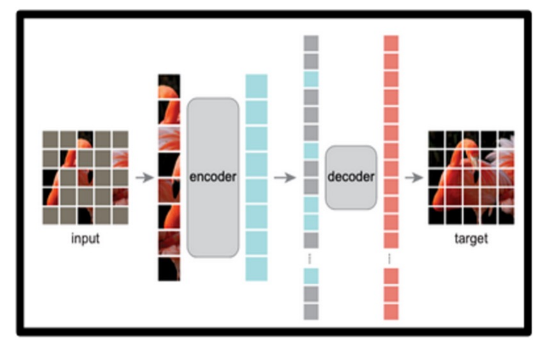
Meta (IG-3.5B), ECCV 2018

https://arxiv.org/pdf/1805.00932.pdf



Google (JFT-300M), ICCV 2017

http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Revisiting_Unreasonable_Effectiveness_ICCV_2017_paper.pdf



MAE, CVPR 2022

https://arxiv.org/abs/2111.06377

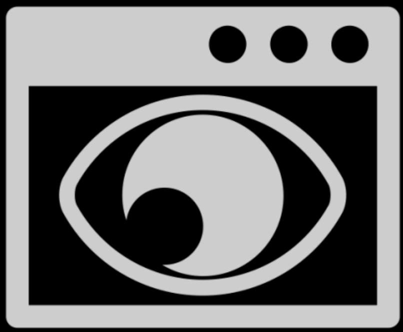
Ethical problems can occur as long as we use real images

To overcome the problems, it is better to automatically create datasets without any natural images

Annotation

FATE

Fairness, Accountability, Transparency and Ethics



Privacy

Can we pre-train DNN without any natural images?

Formula-driven Supervised Learning (FDSL)

- Generate image patterns and their labels
- Using mathematical formulas and/or functions

Observed fractal geometry on ImageNet dataset

We hypothesize DNN could learn natural principles from ImageNet?

Directly render and train Fractals

Our goal is to find a way to pre-train without any real images and human labels

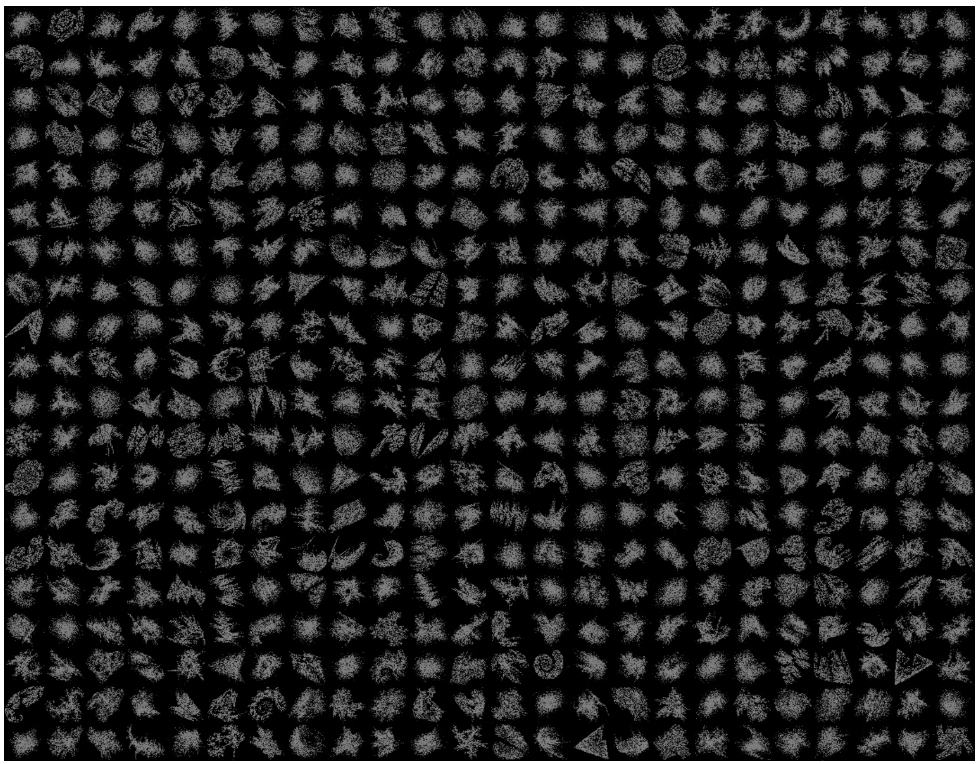
Proposed method: FractalDB Pre-trained CNN

FractalDB

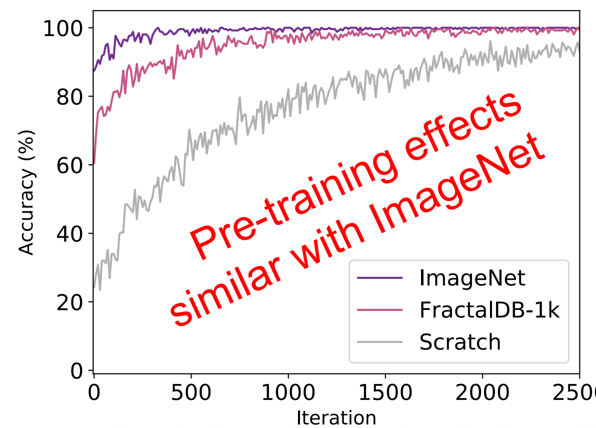
1) to make a pre-trained CNN without any natural images

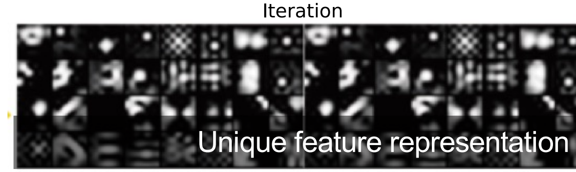
2) for a concept of Formula-driven Supervised Learning

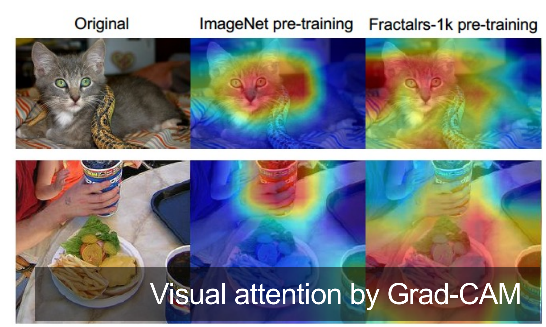
Ability to effectively train models based on natural laws

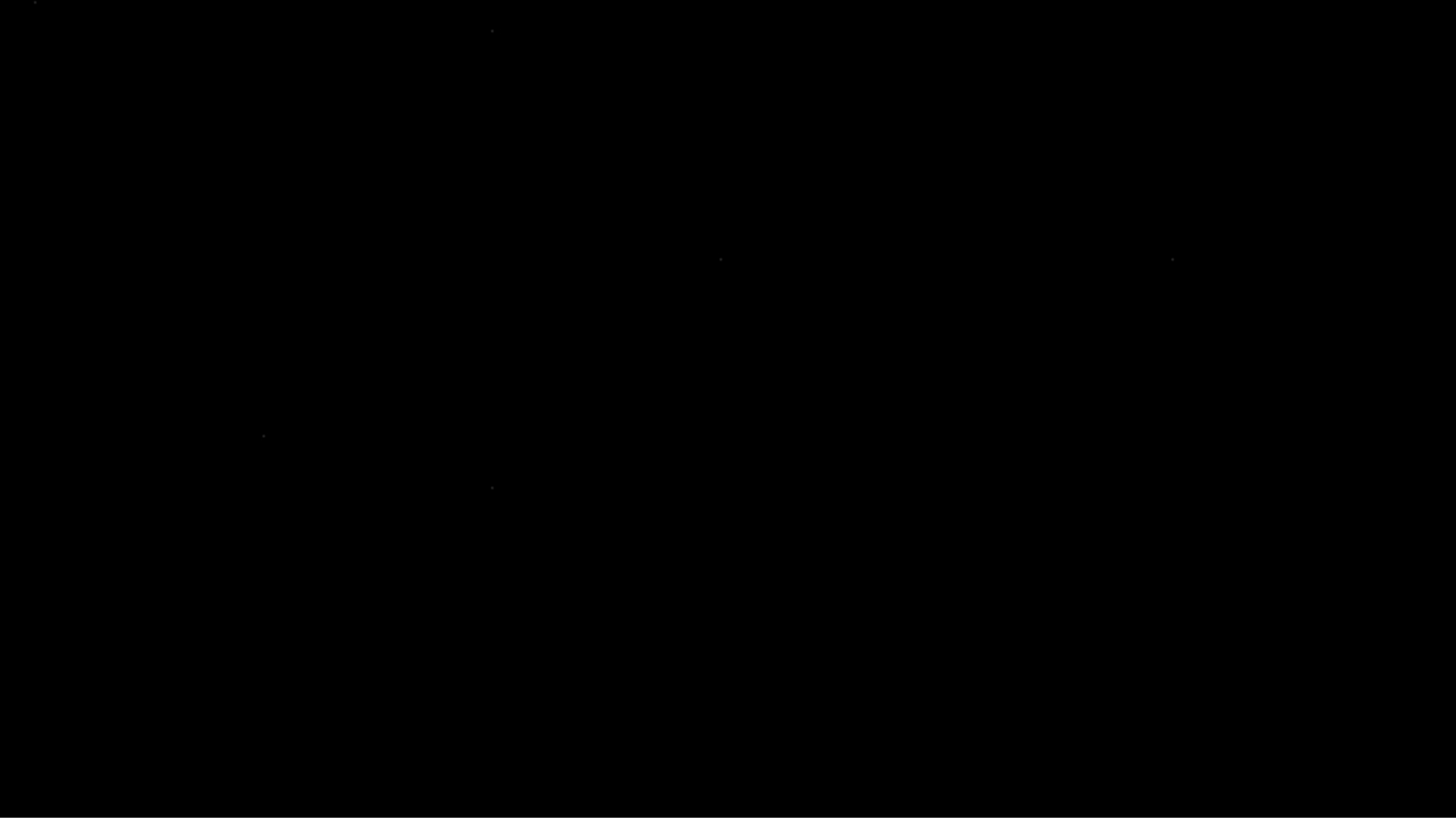


Pre-train









$$\mathrm{IFS} = \{\mathcal{X}; w_1, w_2, \cdots, w_N; p_1, p_2, \cdots, p_N\}$$ # Transformation probability

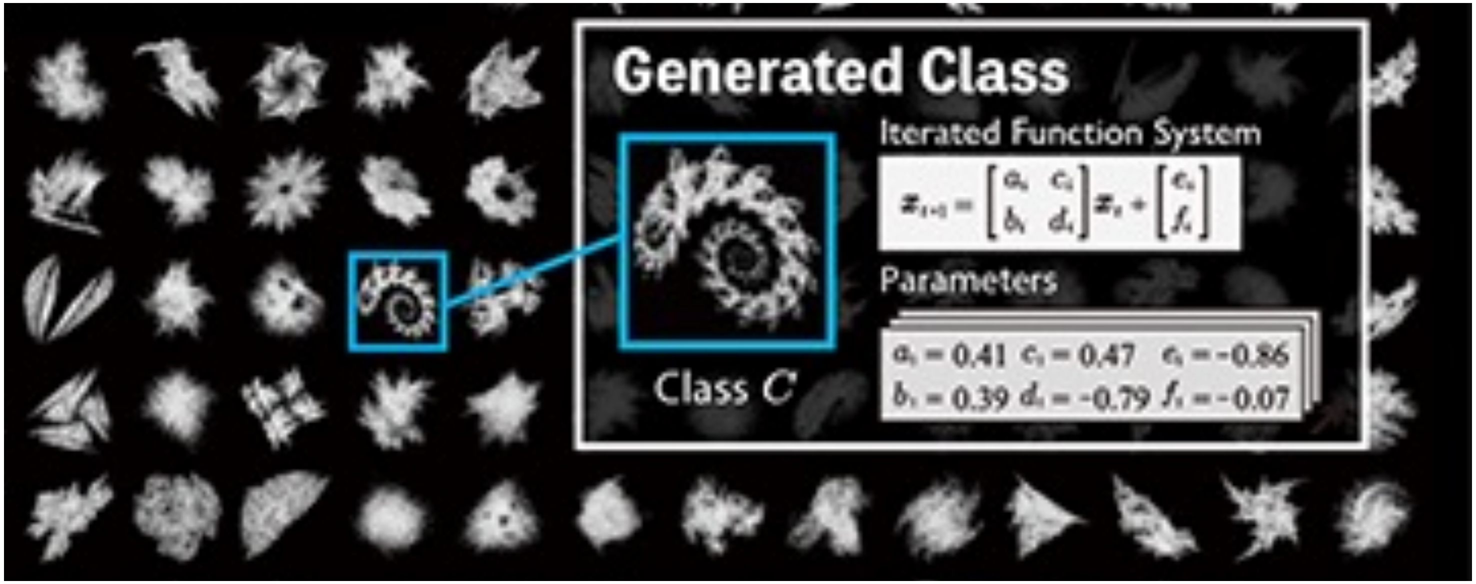
$$w_i(\boldsymbol{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$ # Affine transformation

Iterativesly renders a large number of dots or patches in an image

Search for fractal categories

Randomly select parameters to render

1. Fractal image rendering with randomized params $a \sim f$, w w/ IFS
2. If the filling rate ($> r$), the fractal category is added to DB
3. Repeated up to defined #category (C)
   - Different parameters make a different fractal category

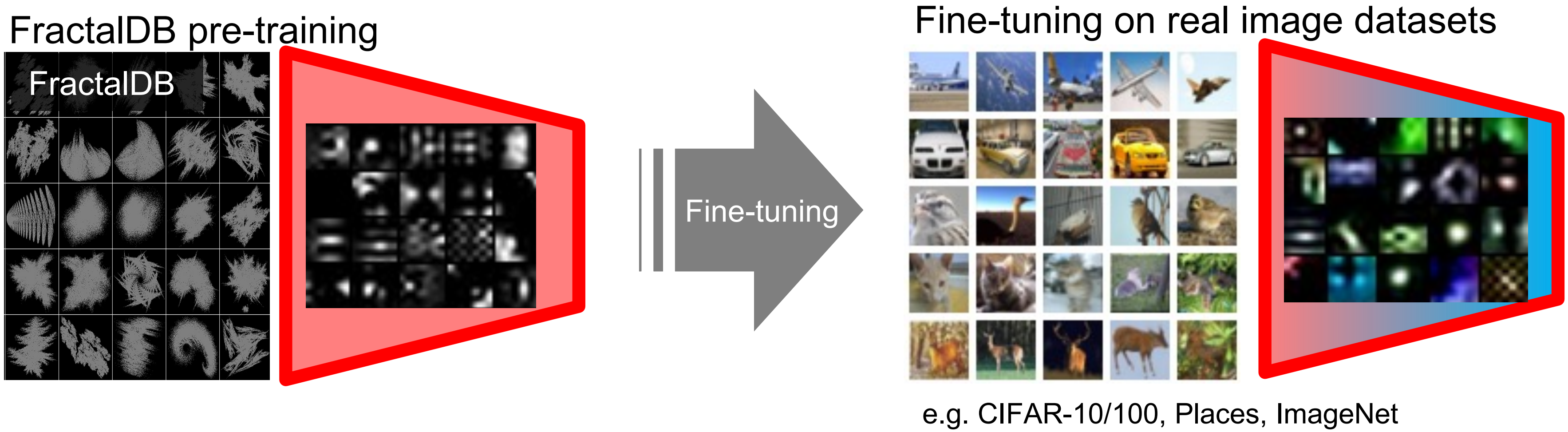


Fractal categories on FractalDB

Experimental setting

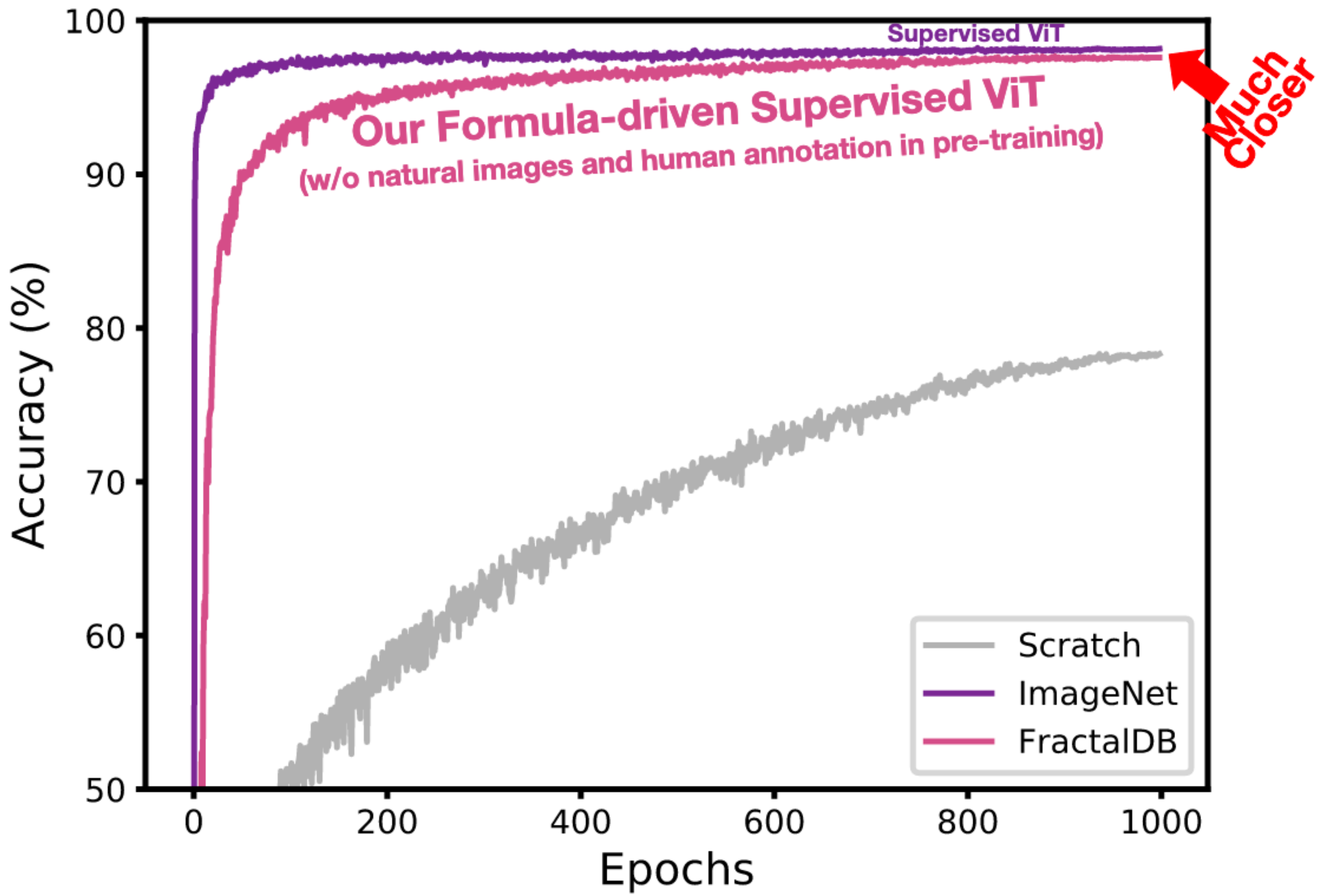
Pre-training, fine-tuning, and model

- Pre-training done without using any real images
- Fine-tuning in a traditional manner
- Vision Transformer model
  - No architecture difference from the original vision transformer
  - We assign data augmentation proposed in DeiT without distillation



FractalDB pre-trained Vision Transformer

– We succeeded a ViT pre-training without real images



Results (1/2)

vs. Supervised Learning

PT	PT Img	PT Type	C10	C100	Cars	Flowers	VOC12	P30	IN100
Scratch	–	–	78.3	57.7	11.6	77.1	64.8	75.7	73.2
Places-30	Natural	Supervision	95.2	78.5	69.4	96.7	77.6	–	86.5
Places-365	Natural	Supervision	**97.6**	**83.9**	**89.2**	**99.3**	84.6	–	<u>**89.4**</u>
ImageNet-100	Natural	Supervision	94.7	77.8	67.4	97.2	78.8	78.1	–
ImageNet-1k	Natural	Supervision	<u>**98.0**</u>	<u>**85.5**</u>	<u>**89.9**</u>	<u>**99.4**</u>	<u>**88.7**</u>	<u>**80.0**</u>	–
FractalDB-1k	Formula	Formula-supervision	96.8	81.6	86.0	98.3	84.5	78.0	87.3
FractalDB-10k	Formula	Formula-supervision	**97.6**	83.5	87.7	98.8	**86.9**	**78.5**	**88.1**

<u>**Underlined bold**</u>: best score, **Bold**: second best score

FractalDB pre-trained model showed significantly improved performance compared to training from scratch

Results (1/2)

vs. Supervised Learning

PT	PT Img	PT Type	C10	C100	Cars	Flowers	VOC12	P30	IN100
Scratch	–	–	78.3	57.7	11.6	77.1	64.8	75.7	73.2
Places-30	Natural	Supervision	95.2	78.5	69.4	96.7	77.6	–	86.5
Places-365	Natural	Supervision	**97.6**	**83.9**	**89.2**	**99.3**	84.6	–	<u>**89.4**</u>
ImageNet-100	Natural	Supervision	94.7	77.8	67.4	97.2	78.8	78.1	–
ImageNet-1k	Natural	Supervision	<u>**98.0**</u>	<u>**85.5**</u>	<u>**89.9**</u>	<u>**99.4**</u>	<u>**88.7**</u>	<u>**80.0**</u>	–
FractalDB-1k	Formula	Formula-supervision	96.8	81.6	86.0	98.3	84.5	78.0	87.3
FractalDB-10k	Formula	Formula-supervision	**97.6**	83.5	87.7	98.8	**86.9**	**78.5**	**88.1**

<u>**Underlined bold**</u>: best score, **Bold**: second best score

Though our method was not able to beat the ImageNet pre-trained model,

the FractalDB pre-trained model partially surpassed the Places

Results (2/2)

vs. Self-supervised Learning

Method	Use Natural Images?	C10	C100	Cars	Flowers	VOC12	P30	Average
Jigsaw	YES	96.4	82.3	55.7	98.2	82.1	**80.6**	82.5
Rotation	YES	95.8	81.2	70.0	96.8	81.1	79.8	84.1
MoCov2	YES	96.9	83.2	78.0	98.5	85.3	<u>**80.8**</u>	87.1
SimCLRv2	YES	**97.4**	<u>**84.1**</u>	**84.9**	<u>**98.9**</u>	**86.2**	80.0	**88.5**
FractalDB-10k	NO	<u>**97.6**</u>	**83.5**	<u>**87.7**</u>	**98.8**	<u>**86.9**</u>	78.5	<u>**88.8**</u>

<u>**Underlined bold**</u>: best score, **Bold**: second best score

The proposed method recorded higher scores compared to SSL methods such as MoCoV2, rotation, and jigsaw puzzle

Results (2/2)

vs. Self-supervised Learning

Method	Use Natural Images?	C10	C100	Cars	Flowers	VOC12	P30	Average
Jigsaw	YES	96.4	82.3	55.7	98.2	82.1	**80.6**	82.5
Rotation	YES	95.8	81.2	70.0	96.8	81.1	79.8	84.1
MoCov2	YES	96.9	83.2	78.0	98.5	85.3	<u>**80.8**</u>	87.1
SimCLRv2	YES	**97.4**	<u>**84.1**</u>	**84.9**	<u>**98.9**</u>	**86.2**	80.0	**88.5**
FractalDB-10k	NO	<u>**97.6**</u>	**83.5**	<u>**87.7**</u>	**98.8**	<u>**86.9**</u>	78.5	<u>**88.8**</u>

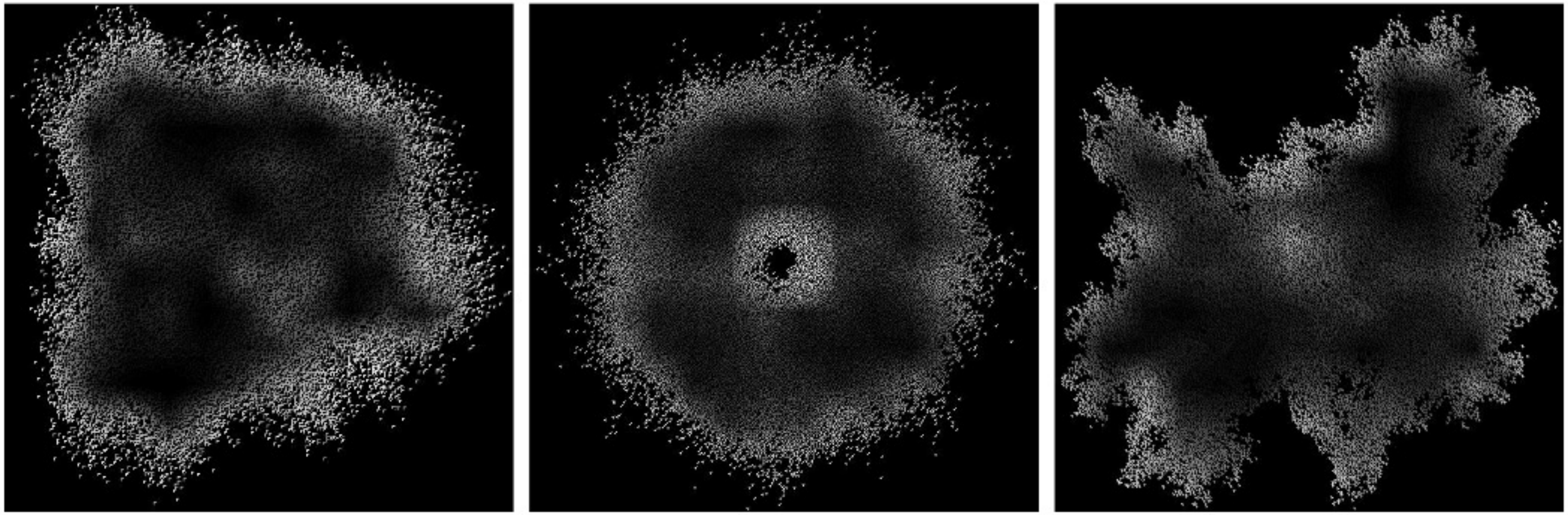
<u>**Underlined bold**</u>: best score, **Bold**: second best score

FractalDB-10k pre-trained ViT recorded a slightly higher in average accuracy on various benchmarks (88.8 vs. 88.5)

Visualization of attention maps

FractalDB pre-trained model focuses on contours

– The figures show attention on fractal images



(d) Attention maps in fractal images with FractalDB-1k pre-trained DeiT. The brighter areas show more attentive areas.

Can vision transformers learn without natural images?

→ Answer is “Yes”. The FractalDB pre-training achieved comparable performance to ImageNet-1k pre-training

Replacing Labeled Real-image Datasets with Auto-generated Contours

CVPR 2022, CVPR 2023
ExFractalDB/RCDB, VisualAtoms

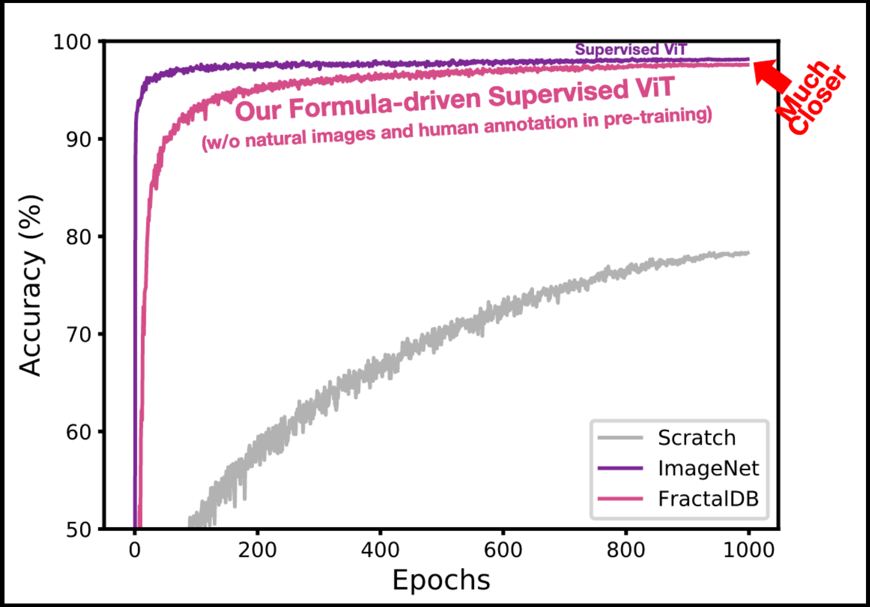
Hirokatsu Kataoka[*], Ryo Hayamizu[*], Ryosuke Yamada[*], Kodai Nakashima[*], Sora Takashima[*,**], Xinyu Zhang[*,**], Edgar Josafat MARTINEZ-NORIEGA[*,**], Nakamasa Inoue[*,**], Rio Yokota[*,**]

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

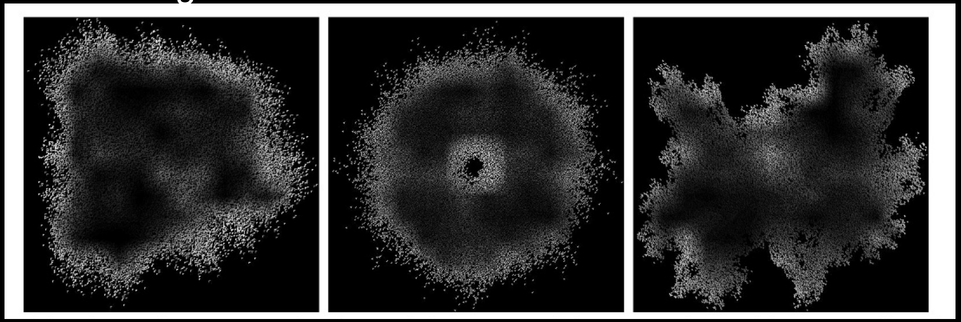
Can Vision Transformers Learn without Natural Images? (AAAI22)

Successfully trained a FractalDB pre-trained ViT

- Reducing the use of real images 14M to 0
- Exploring the reason behind the success



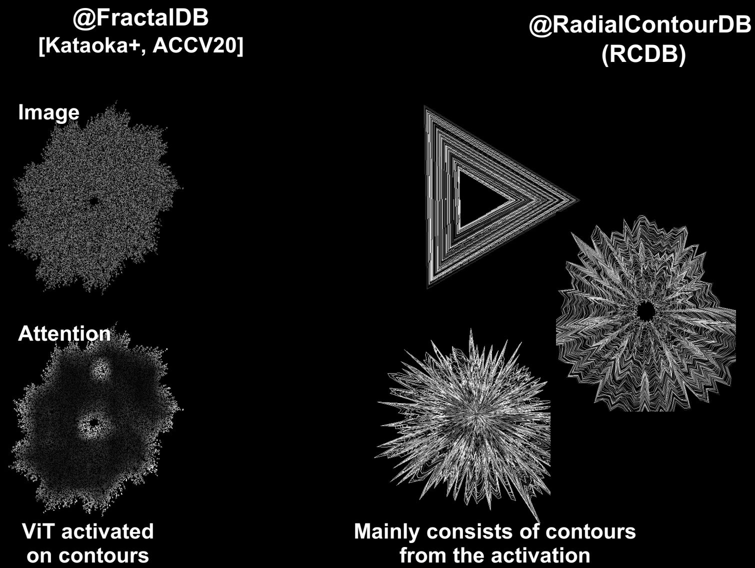
Visualizing self-attention in ViT



→ The fact describes that it focuses on object contours, rather than use of fractals

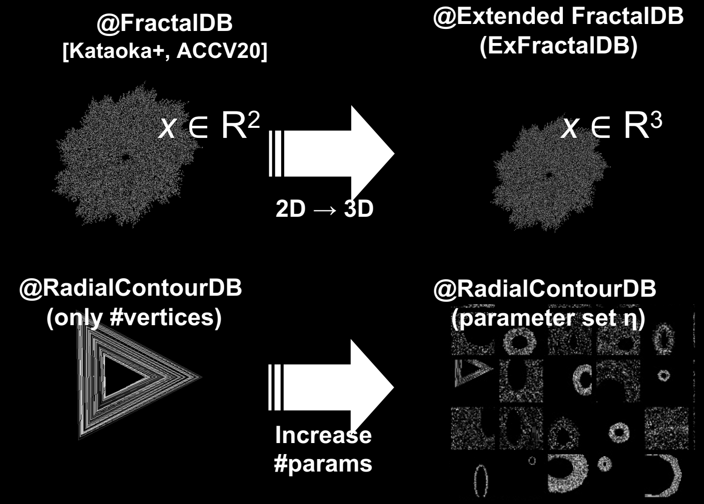
Two hypotheses regarding FDSL pre-training

Hypothesis 1:
Object contours are what matter



As the extreme case of contour classification, we implemented RCDB mainly consists of contours in an image

Hypothesis 2:
Task difficulty matters



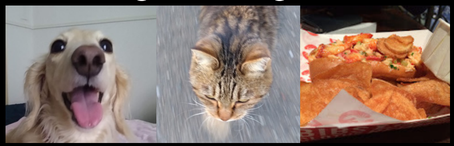
Our finding showed that #parameters are linked to task difficulty

Evaluation on classification, detection, and segmentation

ImageNet-1k / MS COCO dataset

Image Classification / Object Detection, Instance Segmentation

Real images: ImageNet-21k



Accuracy on ImageNet-1k

81.8%

3D fractal images: ExFractalDB-21k

82.7%

Contour images: RCDB-21k

82.4%

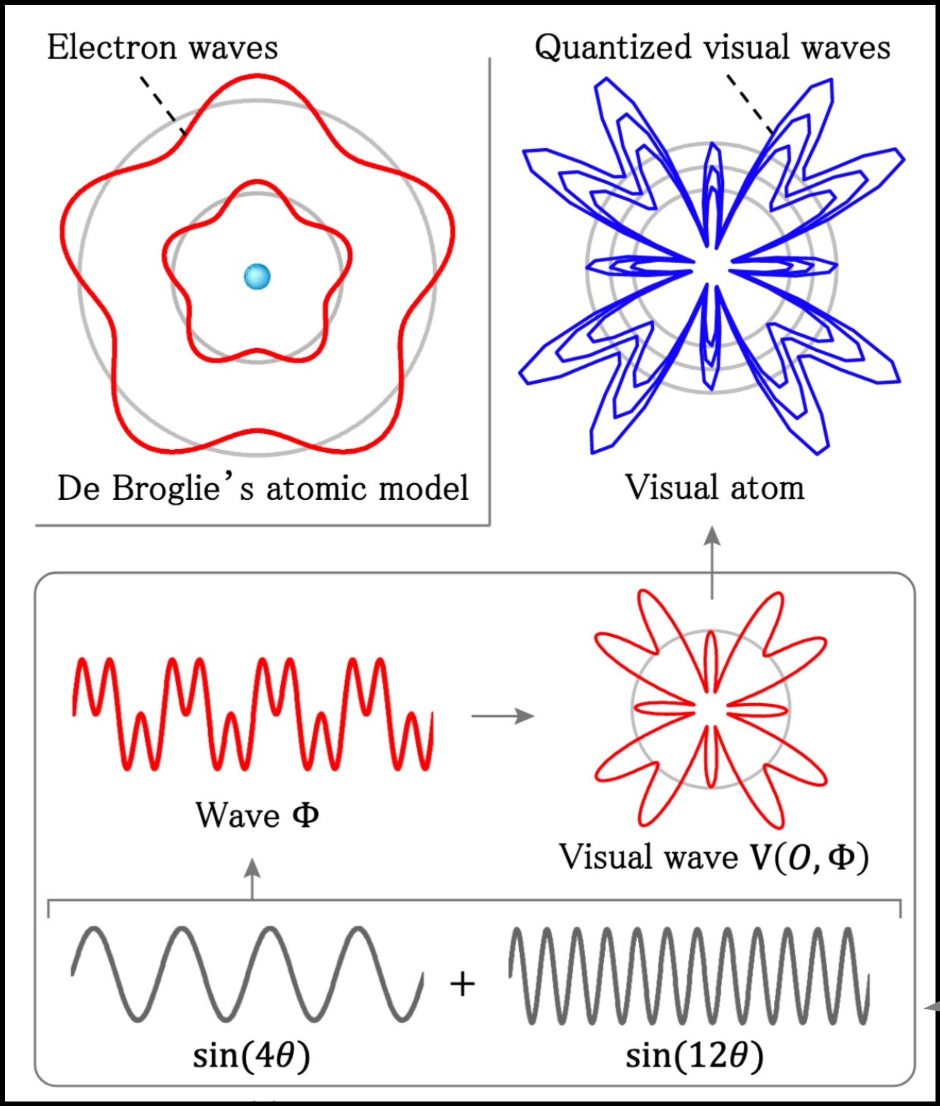
Pre-training	COCO Det AP_{50} / AP / AP_{75}	COCO Inst Seg AP_{50} / AP / AP_{75}
Scratch	63.7 / 42.2 / 46.1	60.7 / 38.5 / 41.3
ImageNet-1k	69.2 / 48.2 / 53.0	66.6 / 43.1 / 46.5
ImageNet-21k	**70.7 / 48.8 / 53.2**	**67.7 / 43.6 / 47.0**
ExFractalDB-1k	69.1 / **48.0 / 52.8**	66.3 / **42.8** / 45.9
ExFractalDB-21k	**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

Exceeded ImageNet-21k pre-training
Radial contours also surpassed the accuracy with ImageNet pre-training in addition to Fractal pre-training

Our pre-trained models perform good fine-tuning results on COCO with a pre-training from only contour classification

How contours important in pre-training?

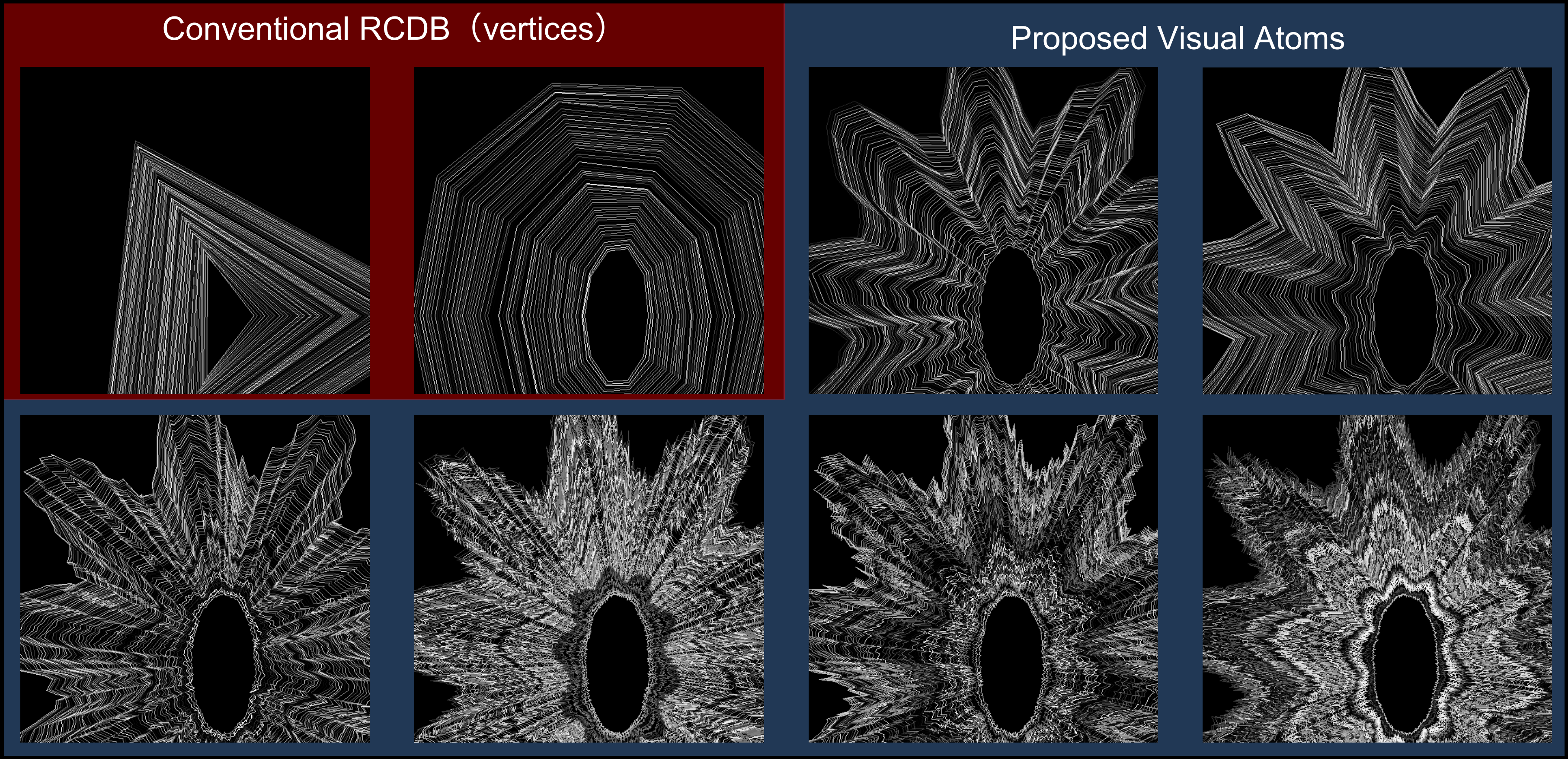
Throughout many experiments, the diversity of contours



Combined two different sine curves (sinusoidal waves)

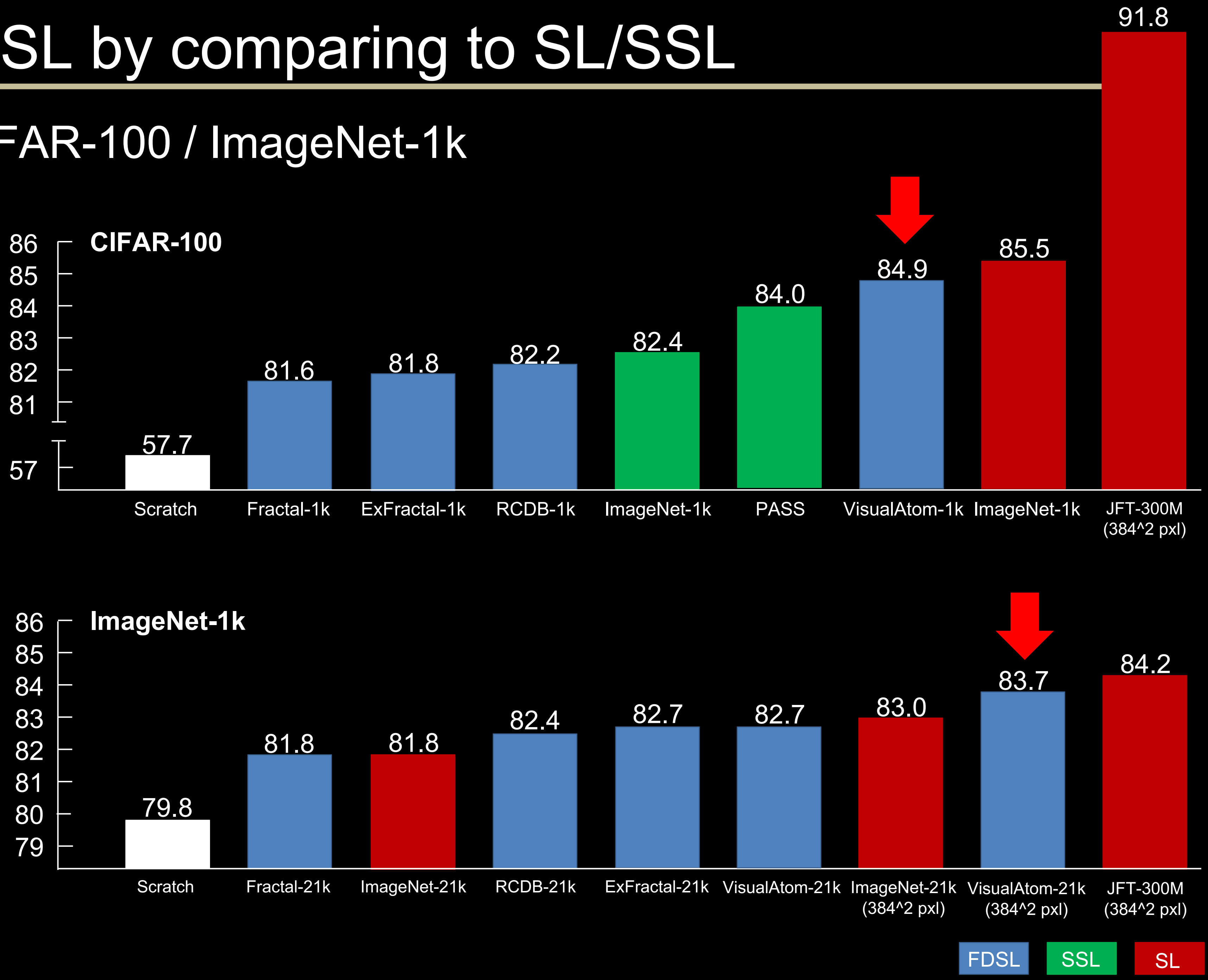
RCDB (vertices) vs. VisualAtom (2 sine curves)

Vertices vs. 2 sine curves



FDSL by comparing to SL/SSL

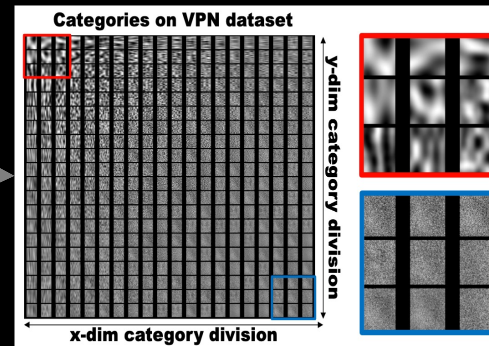
CIFAR-100 / ImageNet-1k



【FDSL Family】

FDSL [Kataoka, ACCV20/IJCV22]
ViT + FDSL [Nakashima, AAAI22]
OFDB [Nakamura, ICCV23]

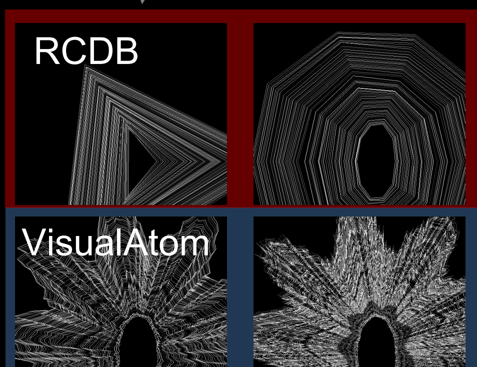
Spatiotemporal Domain



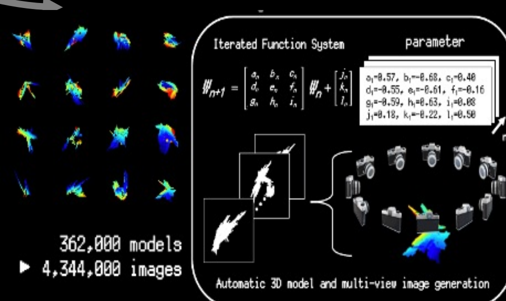
Video Perlin Noise [Kataoka, WACV22]

Object Contours

3D Domain



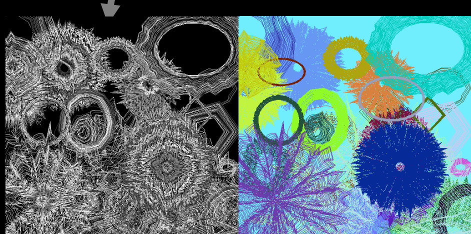
ExFractal / RCDB [Kataoka, CVPR22]
VisualAtom [Takashima, CVPR23]



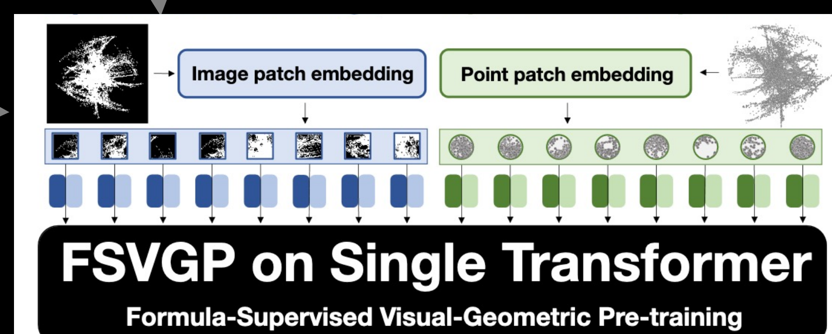
Multi-viewpoint [Yamada, IROS22]
Point Cloud [Yamada, CVPR22]

Merging 2D images and 3D point clouds

Segmentation labels



SegRCDB [Shinoda, ICCV23]

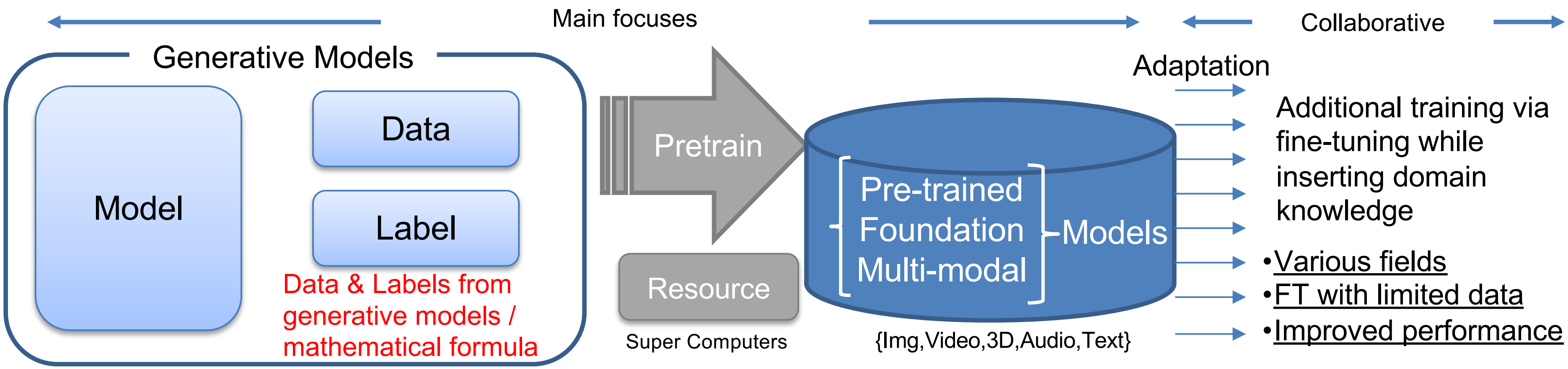


Visual-Geometric FractalDB (On-going work)

Future direction

Conducting pre-training with generative models

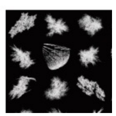
(Foundation, Multi-modal)



Modality

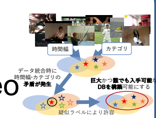
Images

Images/labels are generated from diffusion models or formulas



Videos

Videos/labels are generated with video generative models



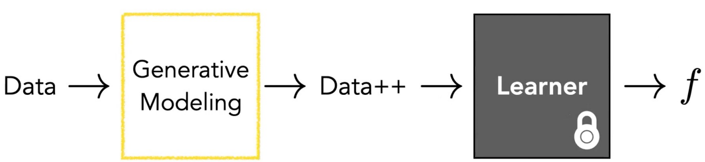
Audio

1D signal/labels are generated. The 1D signal is like a noise generation

Texts

Language models are constructed from a word probability / language models

The concept relates to…



Phillip Isola (MIT)
https://www.youtube.com/watch?v=YuRAeQsTSo8

Three general approaches to employ generative models.

1. To solve the task directly
2. As priors
3. To generate training data

Christian Rupprecht (Univ. of Oxford)
https://www.youtube.com/watch?v=HUyP2C2rYto

Our goal is to improve FDSL to potentially replace the pre-trained model done with real images and human annotations, addressing concerns around ethical and annotation issues

Thank you.