Pre-training without Natural Images

Hirokatsu Kataoka

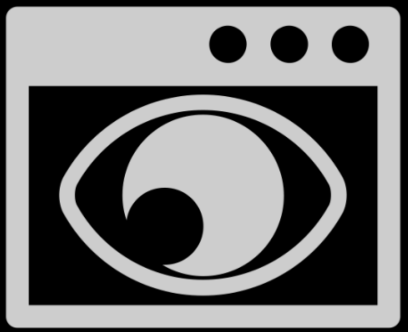
AIST

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

Annotation

FATE

Fairness, Accountability, Transparency and Ethics



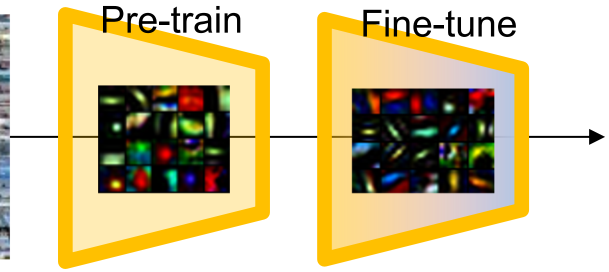
Privacy

Recent vision-driven learning

Supervised Learning

e.g. ImageNet, Places, Open Images

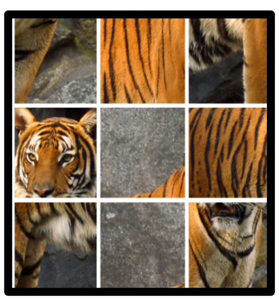
gluon-cv.mxnet.io



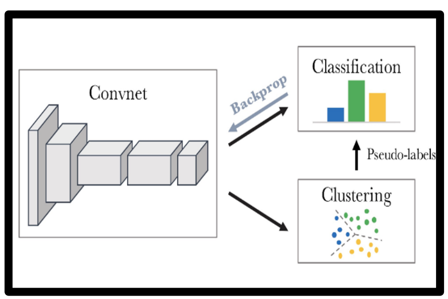
ImageNet + ResNet-50
76% @ImageNet val.

[He et al. CVPR16]

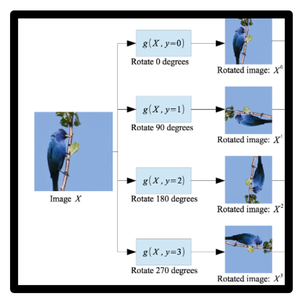
Self-supervised Learning (SSL)



Jigsaw Puzzle
[Noroozi al. ECCV16]



DeepCluster
[Caronet al. ECCV18]



Rotation Classify
[Gidaris et al. ICLR18]

SimCLR + ResNet-50
69%@ImageNet val.

[Chen et al. ICML20]

Existing the problems of image downloading and privacy-violations.

Can we pre-train CNN without any natural images?

Formula-driven Supervised Learning (FDSL)

- Automatically make image patterns and their labels
- With any mathematical formulas and functions

Fractal geometry from ImageNet dataset

CNN trains a natural principle from ImageNet dataset?

Directly render and train Fractals

To replace a human-annotated dataset in context of pre-training without any natural images and human labels

Proposed method: FractalDB

FractalDB

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

2) for a concept of Formula-driven Supervised Learning

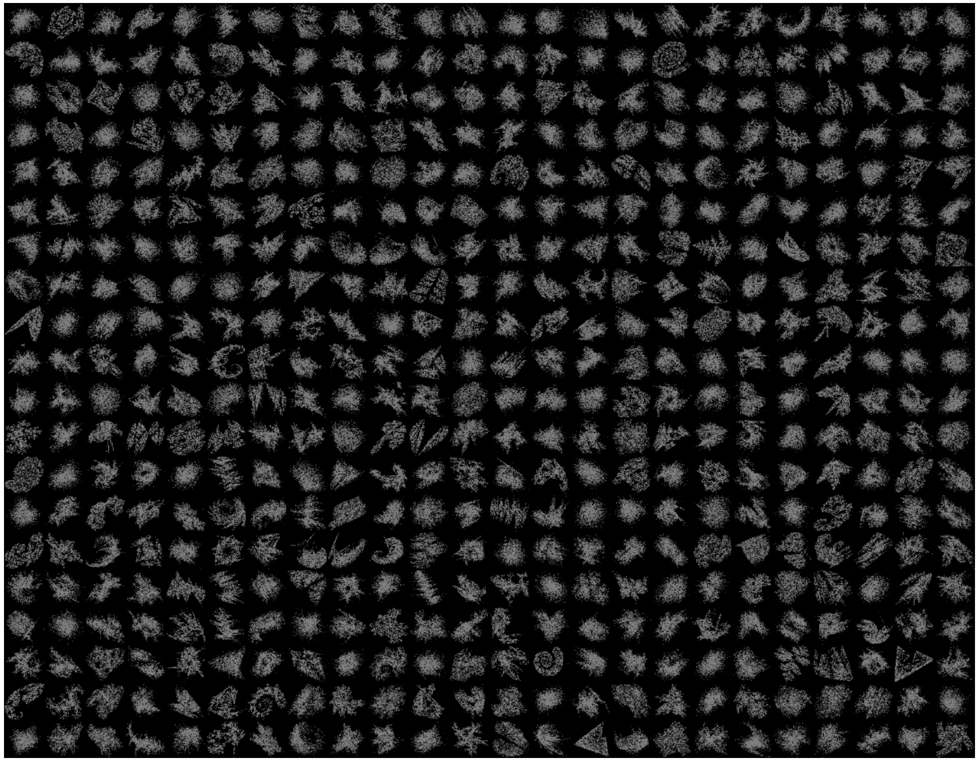
Proposed method: FractalDB

FractalDB

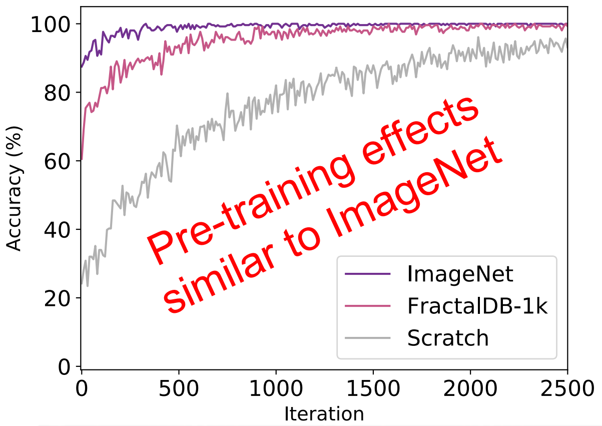
1) to make a pre-trained CNN without any natural images

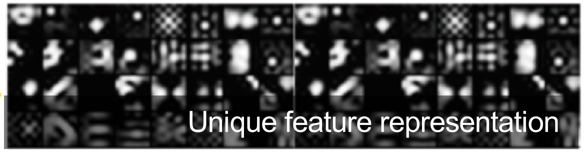
2) for a concept of Formula-driven Supervised Learning

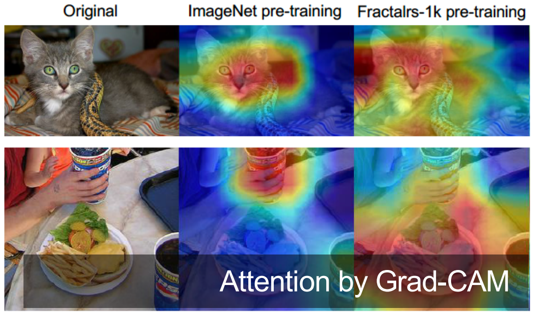
Surprising results which are similar to the effects of a supervised dataset



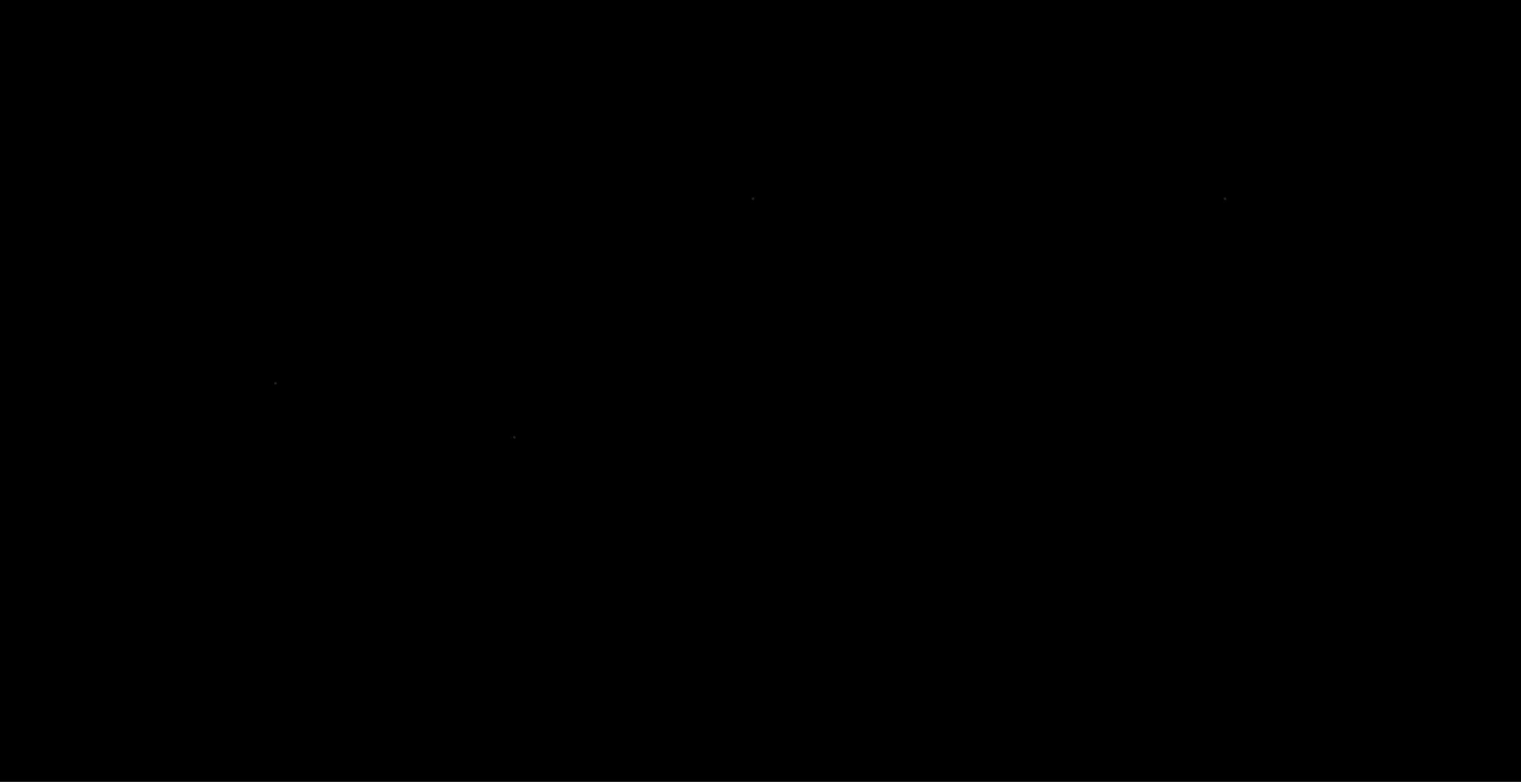
Pre-train







Fractal image rendering with Iterated Function System (IFS)



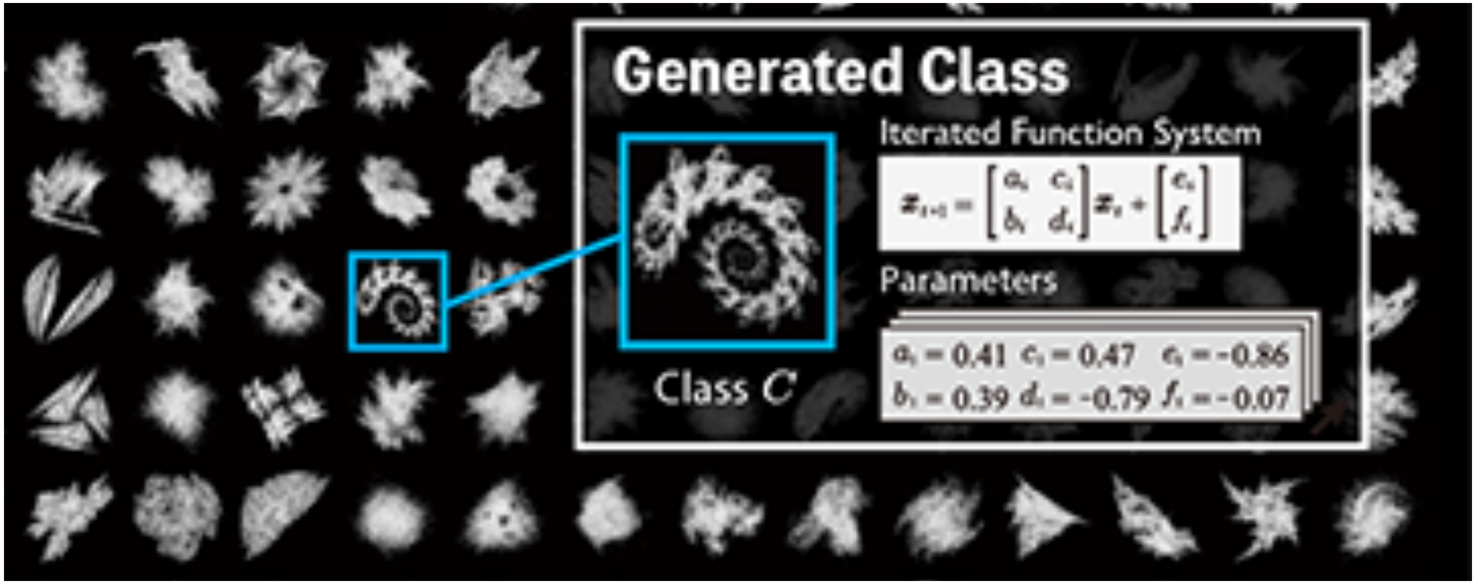
$$\mathrm{IFS} = \{\mathcal{X}; w_1, w_2, \cdots, w_N; p_1, p_2, \cdots, p_N\}$$ # Transformation probability

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

Definition of fractal category

Randomly searched image category

1. Image rendering with randomized $a \sim f$, w through IFS
2. Add category c if filling rate (> r) in the image
3. Iterate up to defined #category (C)
   - Parameter separation makes a different category

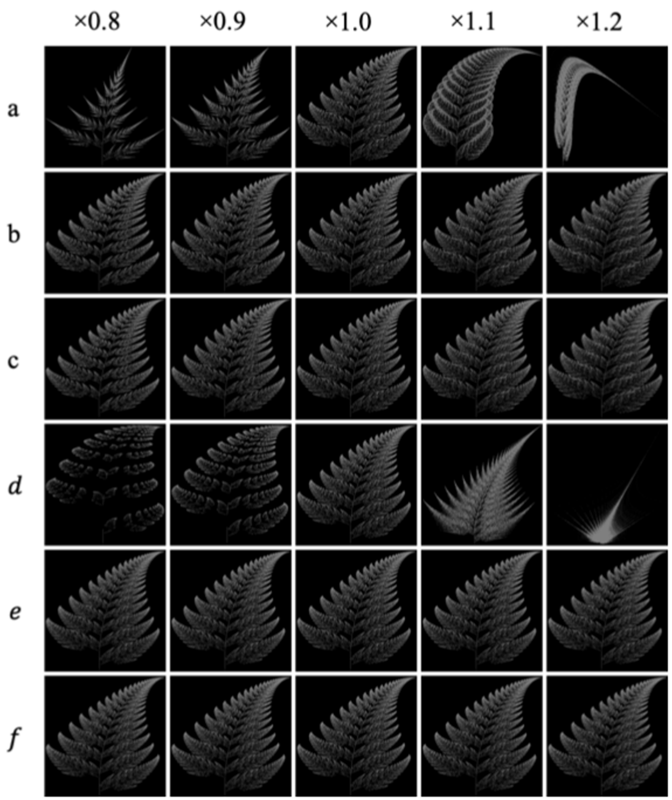


Fractal categories in FractalDB

Instance augmentation

Three different augmentation methods

1. Fluctuation of parameter set (x25)
2. Image rotation (x4)
3. Patch pattern (x10)



Parameter set (x25)

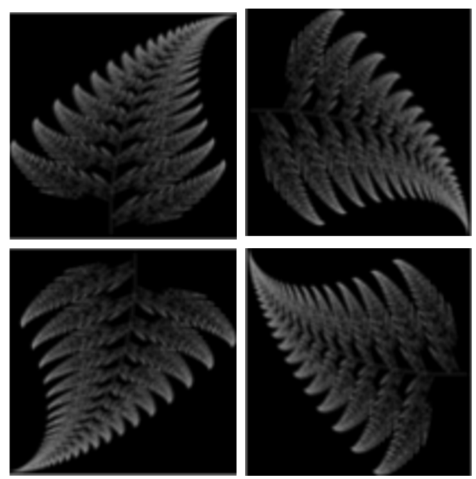
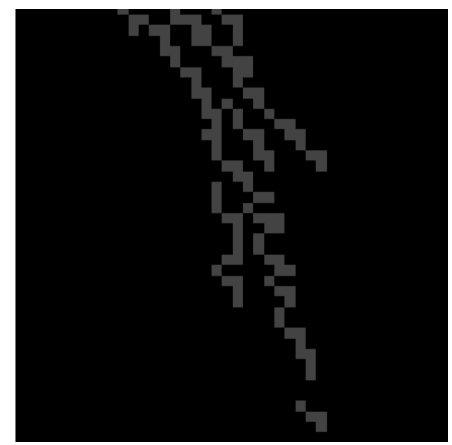


Image rotation (x4)



Patch pattern (x10)

Select ten randomly generated 3x3 patch patterns out of 511 (2^9-1)

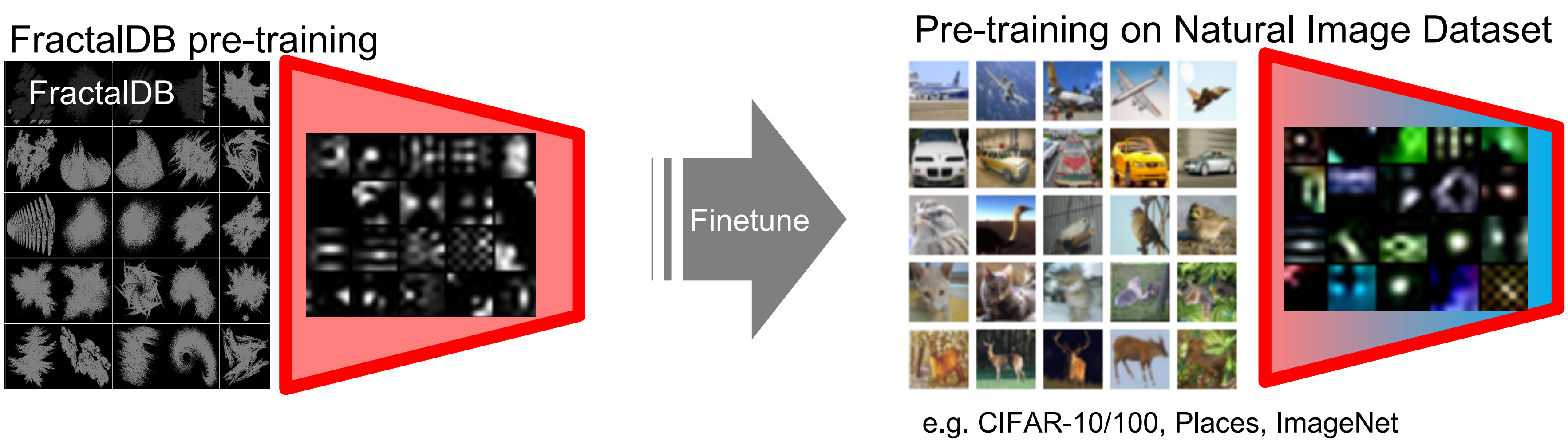
Up to x1000 instances per category

Experimental setting

Pre-training & Fine-tuning

– Pre-training without any natural images

– Fine-tuning in an ordinal way

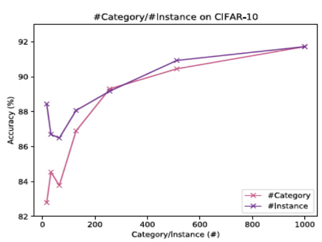


Parameters on FractalDB

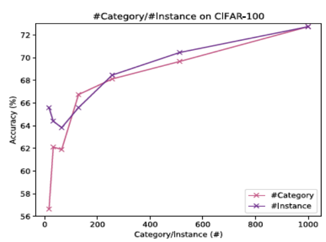
After the burden of exploration study,

– #Category, #instance, and patch-rendering are the most effective parameters in pre-training

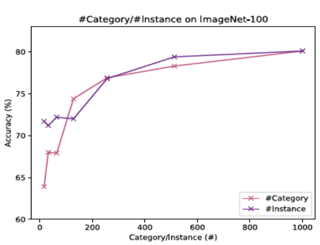
– A more difficult pre-train is slightly better in weights



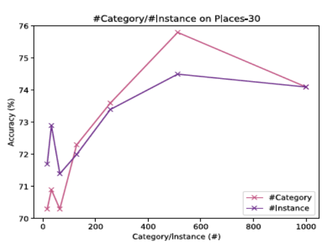
(a) CIFAR10



(b) CIFAR100



(c) ImageNet100



(d) Places30

Table 1. Patch vs. point.

	C10	C100	IN100	P30
Point	87.4	66.1	73.9	73.0
Patch (random)	**92.1**	**72.0**	**78.9**	**73.2**
Patch (fix)	**92.9**	**73.6**	**80.0**	**75.0**

Table 2. Filling rate.

	C10	C100	IN100	P30
.05	91.8	**72.4**	80.2	74.6
.10	**92.0**	72.3	**80.5**	**75.5**
.15	91.7	71.6	80.2	74.3
.20	91.3	70.8	78.8	74.7
.25	91.1	63.2	72.4	74.1

Table 3. Weights.

	C10	C100	IN100	P30
.1	92.1	72.0	78.9	73.2
.2	92.4	72.7	79.2	73.9
.3	92.4	72.6	79.2	74.3
.4	**92.7**	**73.1**	**79.6**	**74.9**
.5	91.8	72.1	78.9	73.5

Table 4. #Dot.

	C10	C100	IN100	P30
100k	**91.3**	70.8	78.8	74.7
200k	90.9	**71.0**	79.2	**74.8**
400k	90.4	70.3	**80.0**	74.5

Table 5. Image size.

	C10	C100	IN100	P30
256	**92.9**	**73.6**	80.0	75.0
362	92.2	73.2	**80.5**	**75.1**
512	90.9	71.0	79.2	73.0
724	90.8	71.0	79.2	73.0
1024	89.6	68.6	77.5	71.9

Please refer to our main paper.

Results (1/5)

Method	Pre-train Img	Type	C10	C100	IN1k	P365	VOC12	OG
Scratch	–	–	87.6	62.7	**<u>76.1</u>**	49.9	58.9	1.1
DC-10k	Natural	Self-supervision	89.9	66.9	66.2	**<u>51.5</u>**	67.5	15.2
Places-30	Natural	Supervision	90.1	67.8	69.1	–	69.5	6.4
Places-365	Natural	Supervision	**94.2**	76.9	71.4	–	**78.6**	10.5
ImageNet-100	Natural	Supervision	91.3	70.6	–	49.7	72.0	12.3
ImageNet-1k	Natural	Supervision	**<u>96.8</u>**	**<u>84.6</u>**	–	50.3	**<u>85.8</u>**	17.5
FractalDB-1k	Formula	Formula-supervision	93.4	75.7	70.3	49.5	58.9	**20.9**
FractalDB-10k	Formula	Formula-supervision	94.1	**77.3**	**71.5**	**50.8**	73.6	**<u>29.2</u>**

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

Results (1/5)

Method	Pre-train Img	Type	C10	C100	IN1k	P365	VOC12	OG
Scratch	–	–	87.6	62.7	<u>**76.1**</u>	49.9	58.9	1.1
DC-10k	Natural	Self-supervision	89.9	66.9	66.2	<u>**51.5**</u>	67.5	15.2
Places-30	Natural	Supervision	90.1	67.8	69.1	–	69.5	6.4
Places-365	Natural	Supervision	**94.2**	76.9	71.4	–	**78.6**	10.5
ImageNet-100	Natural	Supervision	91.3	70.6	–	49.7	72.0	12.3
ImageNet-1k	Natural	Supervision	<u>**96.8**</u>	<u>**84.6**</u>	–	50.3	<u>**85.8**</u>	17.5
FractalDB-1k	Formula	Formula-supervision	93.4	75.7	70.3	49.5	58.9	**20.9**
FractalDB-10k	Formula	Formula-supervision	94.1	**77.3**	**71.5**	**50.8**	73.6	<u>**29.2**</u>

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

FractalDB pre-trained model achieved much higher rates than training from scratch

Results (1/5)

Method	Pre-train Img	Type	C10	C100	IN1k	P365	VOC12	OG
Scratch	–	–	87.6	62.7	**76.1**	49.9	58.9	1.1
DC-10k	Natural	Self-supervision	89.9	66.9	66.2	**51.5**	67.5	15.2
Places-30	Natural	Supervision	90.1	67.8	69.1	–	69.5	6.4
Places-365	Natural	Supervision	**94.2**	76.9	71.4	–	**78.6**	10.5
ImageNet-100	Natural	Supervision	91.3	70.6	–	49.7	72.0	12.3
ImageNet-1k	Natural	Supervision	**96.8**	**84.6**	–	50.3	**85.8**	17.5
FractalDB-1k	Formula	Formula-supervision	93.4	75.7	70.3	49.5	58.9	**20.9**
FractalDB-10k	Formula	Formula-supervision	94.1	**77.3**	**71.5**	**50.8**	73.6	**29.2**

Underlined bold: best score, **Bold**: second best score

In the most cases, our method is better than the DeepCluster with 10k categories

Results (1/5)

Method	Pre-train Img	Type	C10	C100	IN1k	P365	VOC12	OG
Scratch	–	–	87.6	62.7	<u>**76.1**</u>	49.9	58.9	1.1
DC-10k	Natural	Self-supervision	89.9	66.9	66.2	<u>**51.5**</u>	67.5	15.2
Places-30	Natural	Supervision	90.1	67.8	69.1	–	69.5	6.4
Places-365	Natural	Supervision	**94.2**	76.9	71.4	–	**78.6**	10.5
ImageNet-100	Natural	Supervision	91.3	70.6	–	49.7	72.0	12.3
ImageNet-1k	Natural	Supervision	<u>**96.8**</u>	<u>**84.6**</u>	–	50.3	<u>**85.8**</u>	17.5
FractalDB-1k	Formula	Formula-supervision	93.4	75.7	70.3	49.5	58.9	**20.9**
FractalDB-10k	Formula	Formula-supervision	94.1	**77.3**	**71.5**	**50.8**	73.6	<u>**29.2**</u>

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

The FractalDB pre-trained model is still better than 100k-order supervised datasets

Results (1/5)

Method	Pre-train Img	Type	C10	C100	IN1k	P365	VOC12	OG
Scratch	–	–	87.6	62.7	<u>**76.1**</u>	49.9	58.9	1.1
DC-10k	Natural	Self-supervision	89.9	66.9	66.2	<u>**51.5**</u>	67.5	15.2
Places-30	Natural	Supervision	90.1	67.8	69.1	–	69.5	6.4
Places-365	Natural	Supervision	**94.2**	76.9	71.4	–	**78.6**	10.5
ImageNet-100	Natural	Supervision	91.3	70.6	–	49.7	72.0	12.3
ImageNet-1k	Natural	Supervision	<u>**96.8**</u>	<u>**84.6**</u>	–	50.3	<u>**85.8**</u>	17.5
FractalDB-1k	Formula	Formula-supervision	93.4	75.7	70.3	49.5	58.9	**20.9**
FractalDB-10k	Formula	Formula-supervision	94.1	**77.3**	**71.5**	**50.8**	73.6	<u>**29.2**</u>

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

Our method partially surpasses the ImageNet/Places pre-trained models

Results (2/5)

Mtd	PT Img	C10	C100	IN1k	P365	VOC12	OG
DC-10k	Natural	89.9	66.9	66.2	51.2	67.5	15.2
DC-10k	Formula	83.1	57.0	65.3	**53.4**	60.4	15.3
F1k	Formula	93.4	75.7	70.3	49.5	58.9	20.9
F10k	Formula	**94.1**	**77.3**	**71.5**	50.8	**73.6**	**29.2**

Bold: best score

DC-10k with fractal images cannot effectively pre-train to recognize natural images

This shows our method assigns an appropriate image pattern and the category

Results (3/5)

Freezing layer(s)	C10	C100	IN100	P30
Fine-tuning	93.4	75.7	82.7	75.9
Conv1	92.3	72.2	77.9	74.3
Conv1–2	92.0	72.0	77.5	72.9
Conv1–3	89.3	68.0	71.0	68.5
Conv1–4	82.7	56.2	55.0	58.3
Conv1–5	49.4	24.7	21.2	31.4

Full fine-tuning is the best

Moreover, earlier layers tend to be good feature representations

Results (4/5)

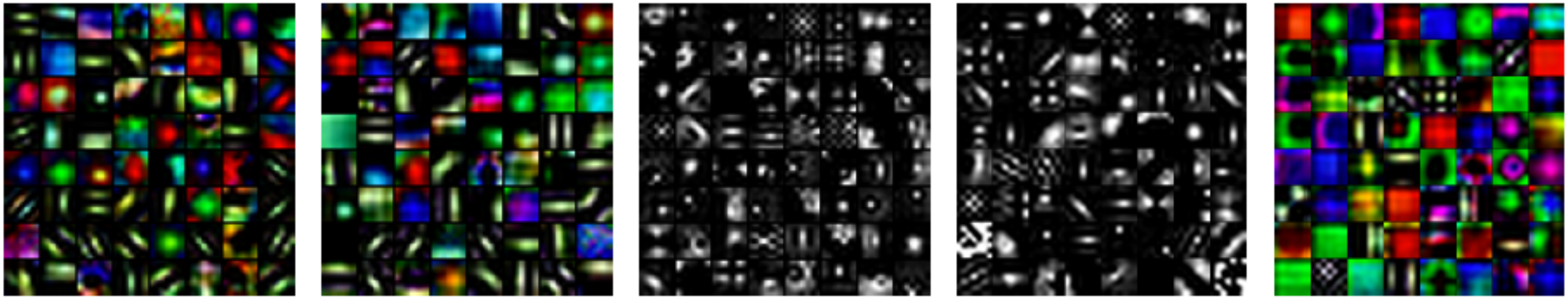
Pre-training	C10	C100	IN100	P30
Scratch	87.6	60.6	75.3	70.3
Bezier-144	87.6	62.5	72.7	73.5
Bezier-1024	89.7	68.1	73.0	73.6
Perlin-100	90.9	70.2	73.0	73.3
Perlin-1296	90.4	71.1	79.7	74.2
FractalDB-1k	**93.4**	**75.7**	**82.7**	**75.9**

We compare Formula-driven Supervised Learning with other principles

The FractalDB pre-trained model outperforms other methods

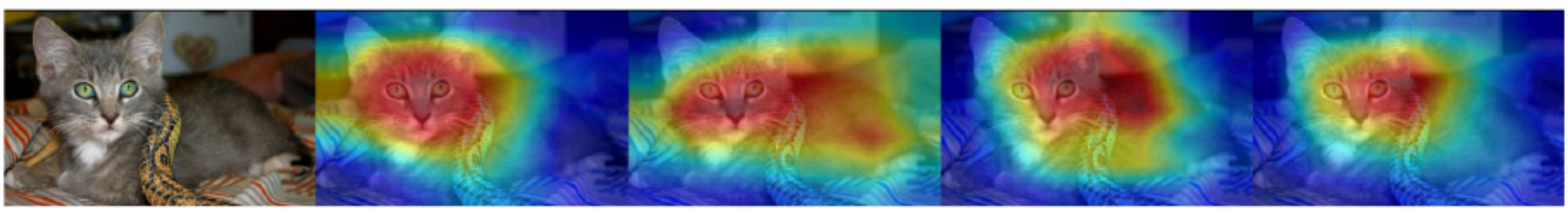
Results (5/5)

Visualization of Conv1



(a) ImageNet (b) Places365 (c) Fractal-1K (d) Fractal-10K (e) DC-10k

Original | ImageNet-1k →CIFAR-10 | Places365 →CIFAR-10 | FractalDB-1k →CIFAR-10 | FractalDB-10k →CIFAR-10



FractalDB pre-trained model acquires different representations yet look at a similar area

Thereafter…

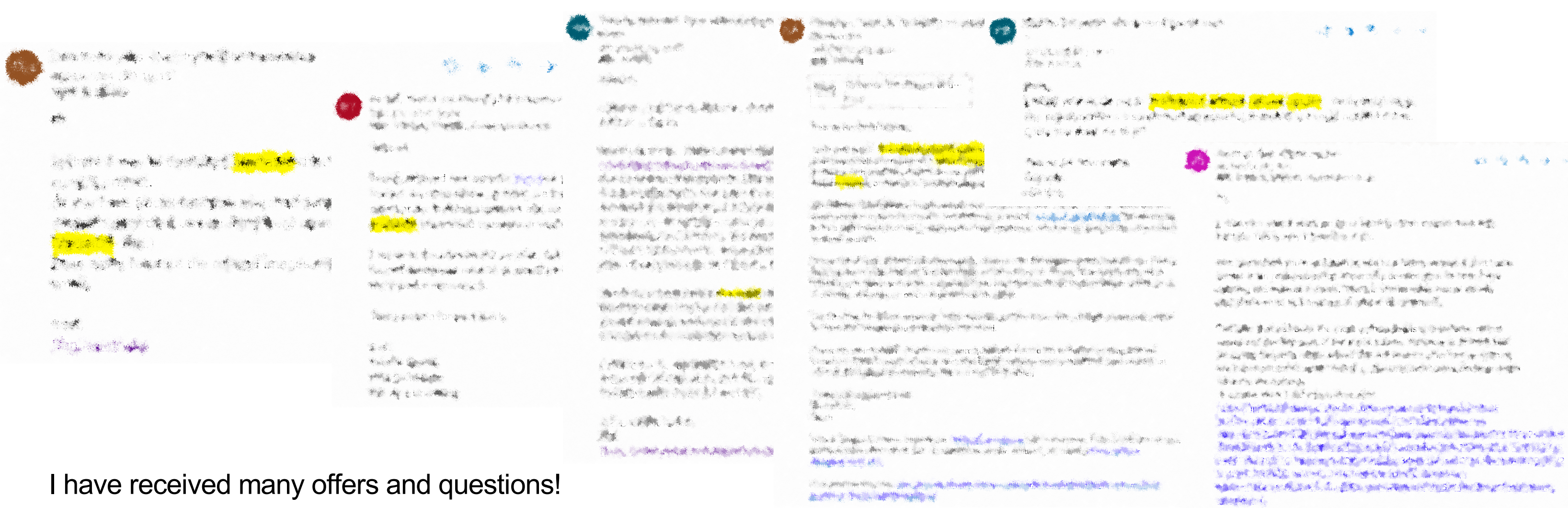
Best Paper
Honorable Mention

Pre-training without Natural Images

Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto,
Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue,
Akio Nakamura, Yutaka Satoh

Microsoft

Thanks to ACCV committee, our paper was authorized as an awardee🎉🎉🎉

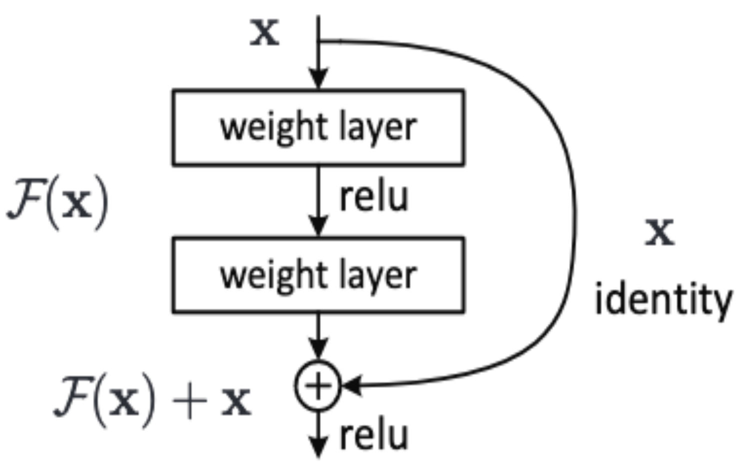


…

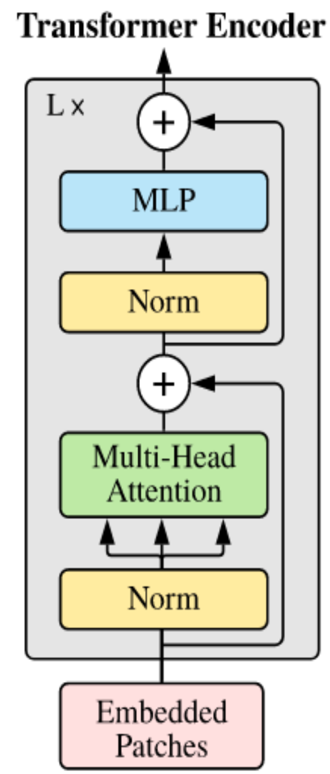
I have received many offers and questions!

Paradigm Shift in Computer Vision

‘Convolution’ to ‘Self-attention’



[He al. CVPR16]



[Vaswani al. NIPS17]
Figure from [Dosovitskiy al. ICLR21]

Can Vision Transformers Learn without Natural Images?

Hirokatsu Kataoka

AIST

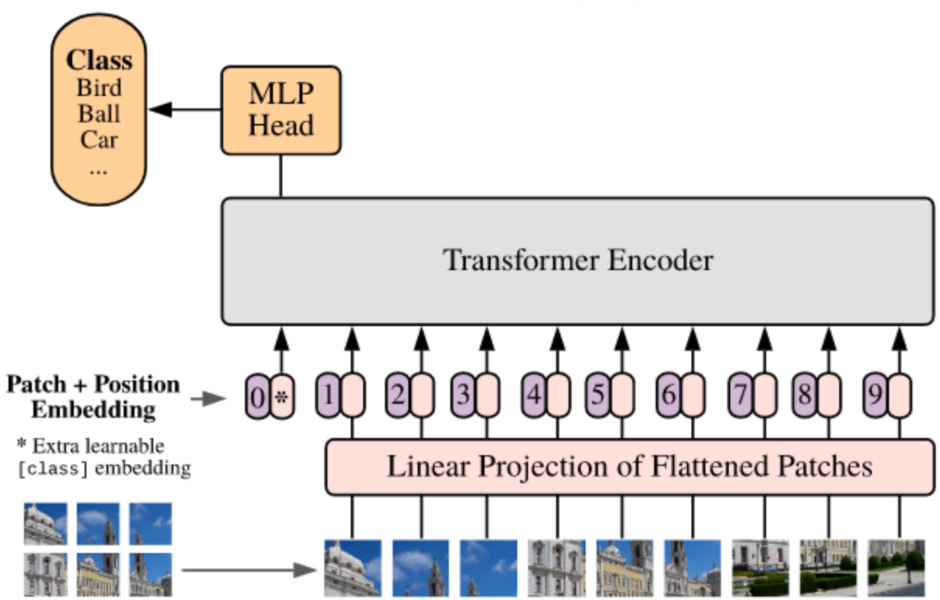
Vision Transformer (ViT), so far

One more shift in Transformer

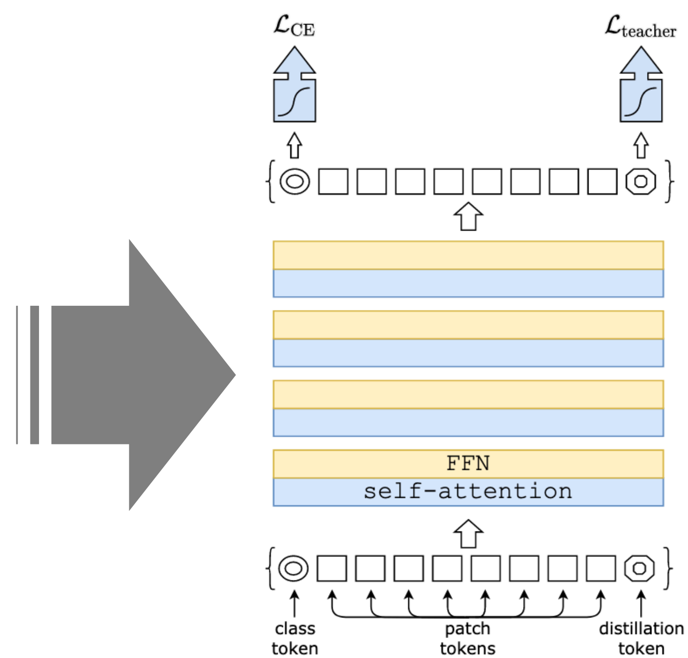
– ViT to DeiT (Data-efficient image Transformer)

– JFT-300M to ImageNet-1k in pre-training

Don't we require any natural images in ViT/DeiT?



[Dosovitskiy al. ICLR21]



[Touvron al. arXiv20]

Settings of Architecture and Dataset

Architecture

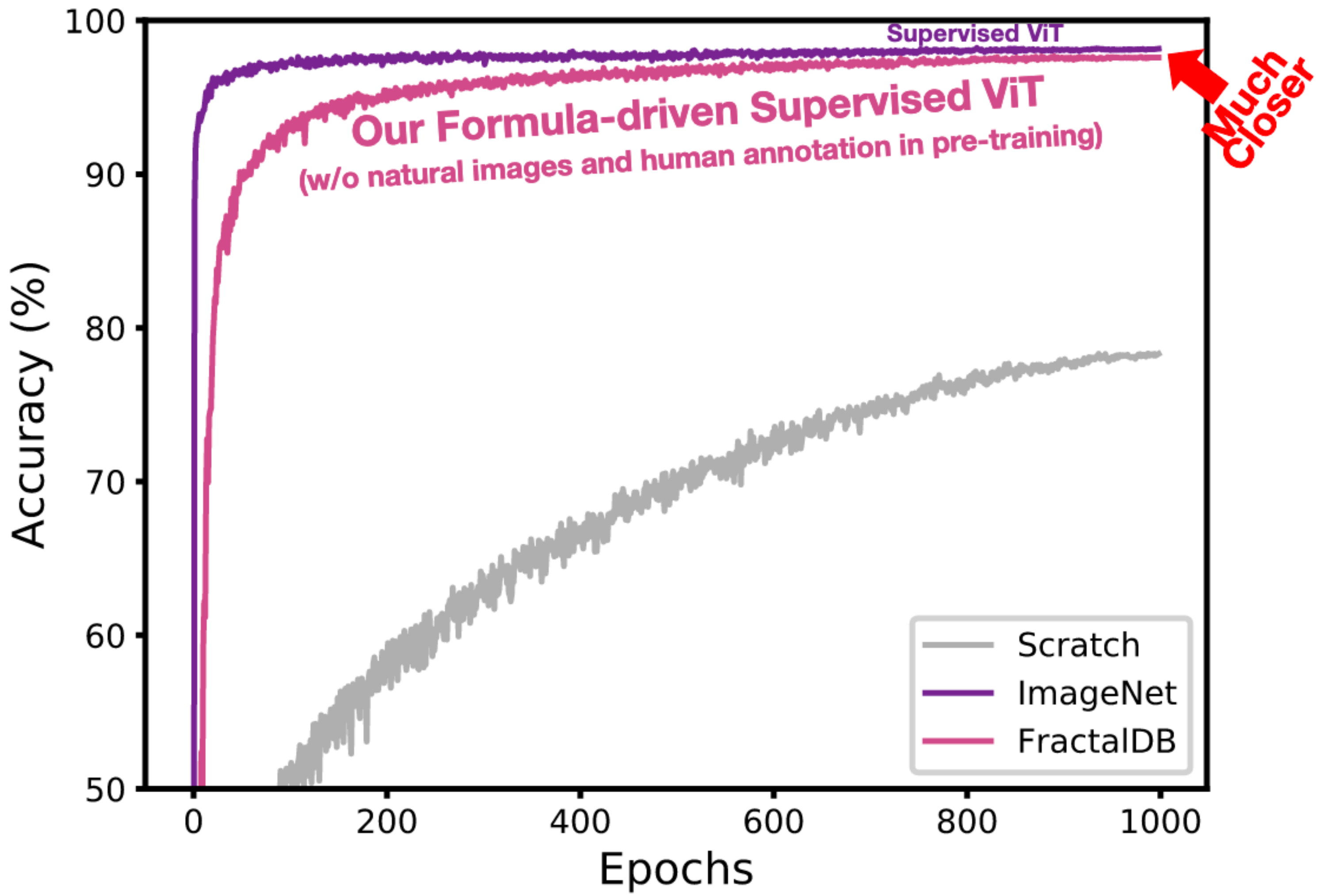
- DeiT
  - No big difference from DeiT on natural image datasets

Dataset

- FractalDB
  - Grayscale is better than colored FractalDB
    - ResNet: colored FractalDB is slightly better
    - DeiT: grayscale FractalDB is better
  - Longer training is better
    - 300 epochs in DeiT, instead of 90/200 epochs in ResNet

FractalDB pre-trained DeiT

– We succeeded a DeiT training without natural 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 achieved much higher rates than 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 cannot beat the ImageNet pre-trained model,

the FractalDB pre-trained model partially surpasses the Places pre-trained models

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 accuracies than SSL methods

with MoCoV2, Rotation, and Jigsaw

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>

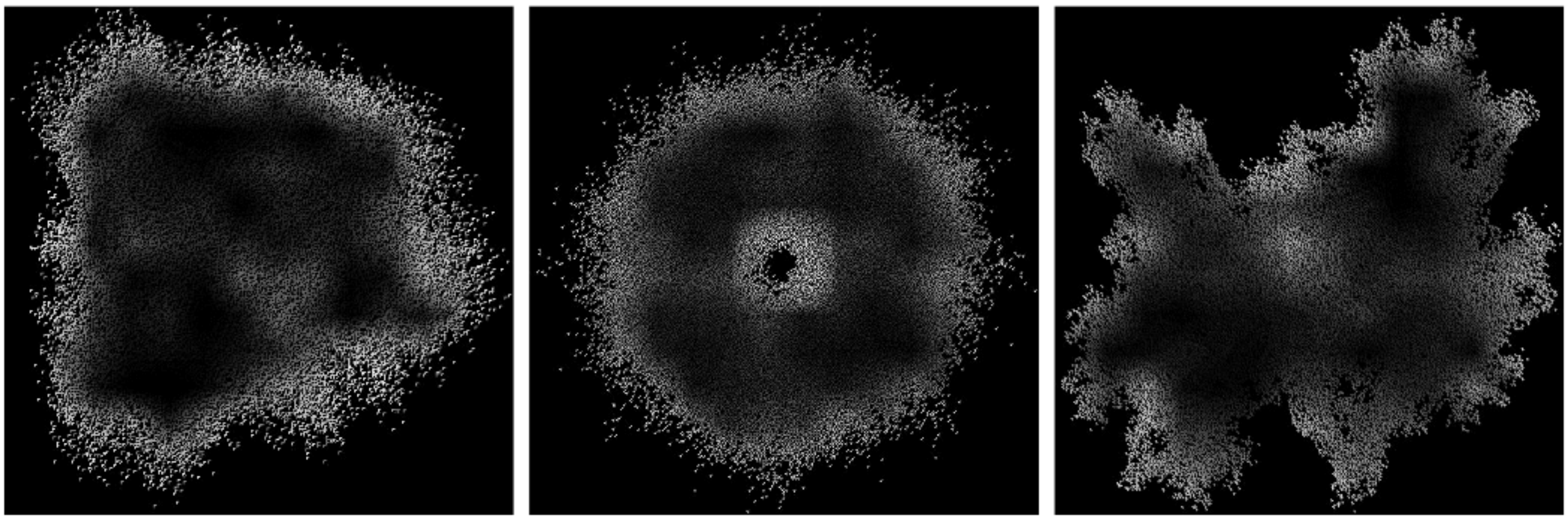
<u>**Underlined bold**</u>: best score, **Bold**: second best score

The FractalDB-10k pre-trained DeiT performs slightly higher in average accuracy on representative datasets (88.8 vs. 88.5)

Visualization of attention maps

FractalDB pre-trained model focused on contours

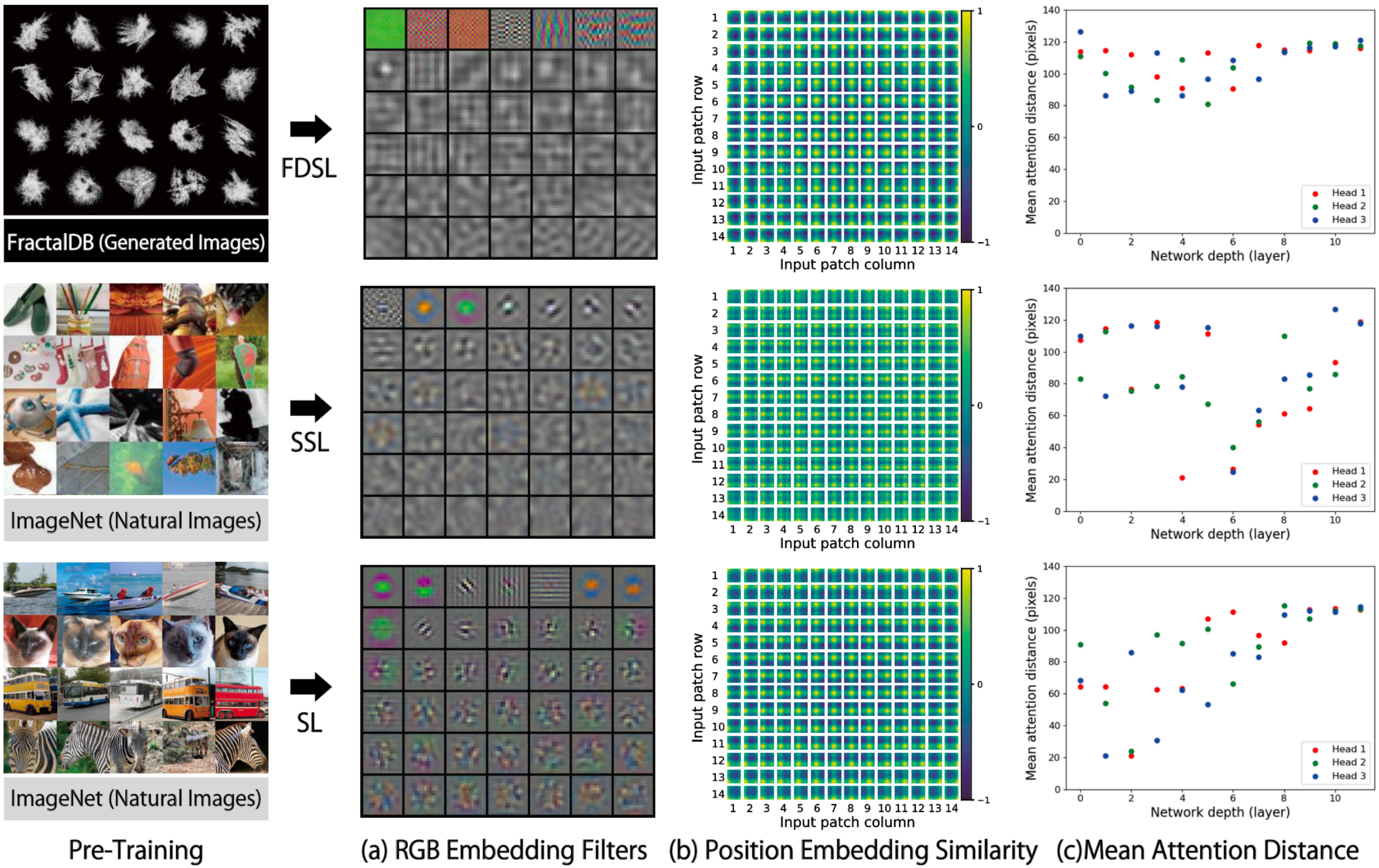
– The figures show attention on fractal images



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

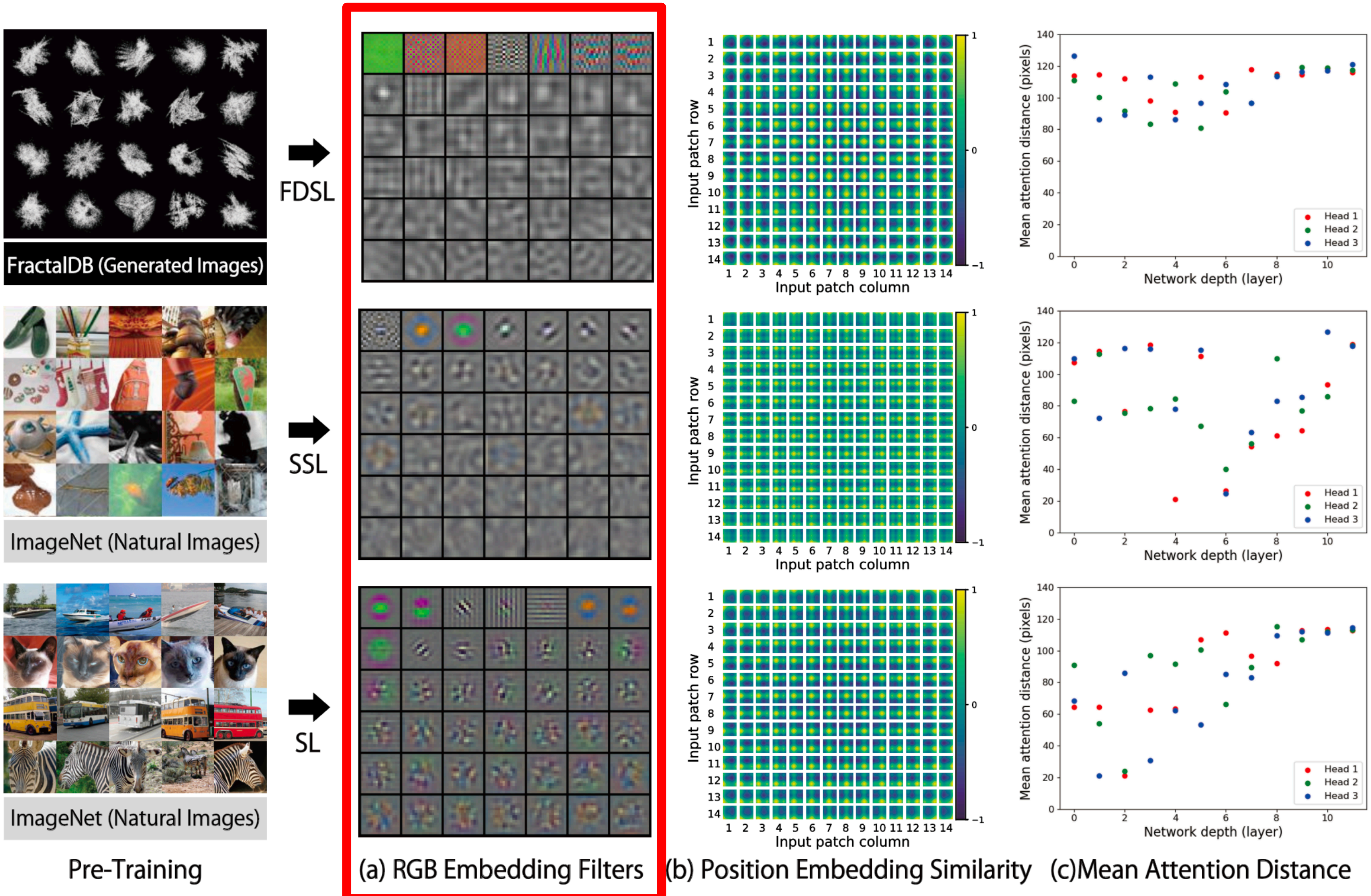
Visualization

Characteristics of FDSL, SSL, and SL



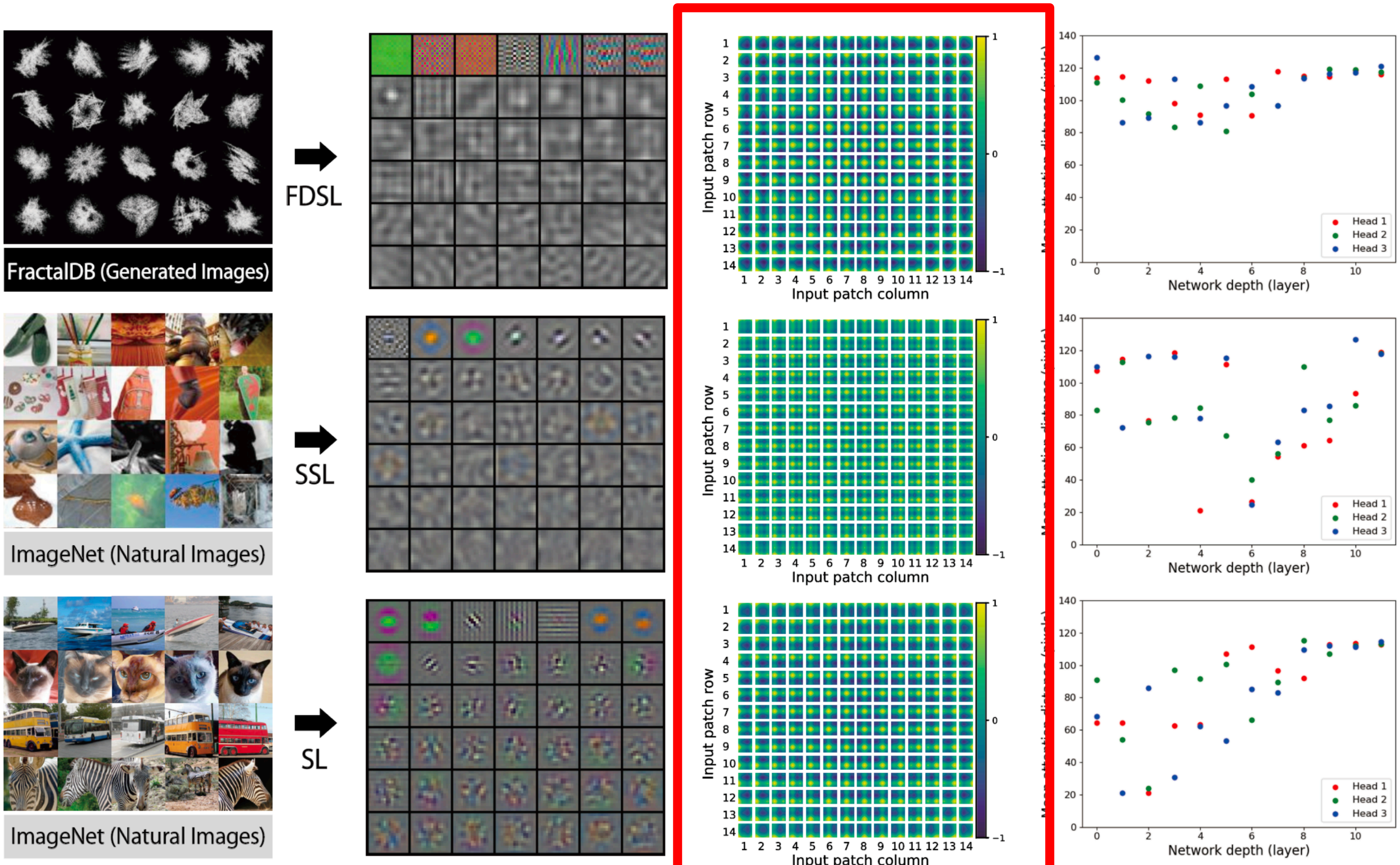
Visualization of embedding filters

Visual representation in the initial filter



Visualization of position embedding similarity

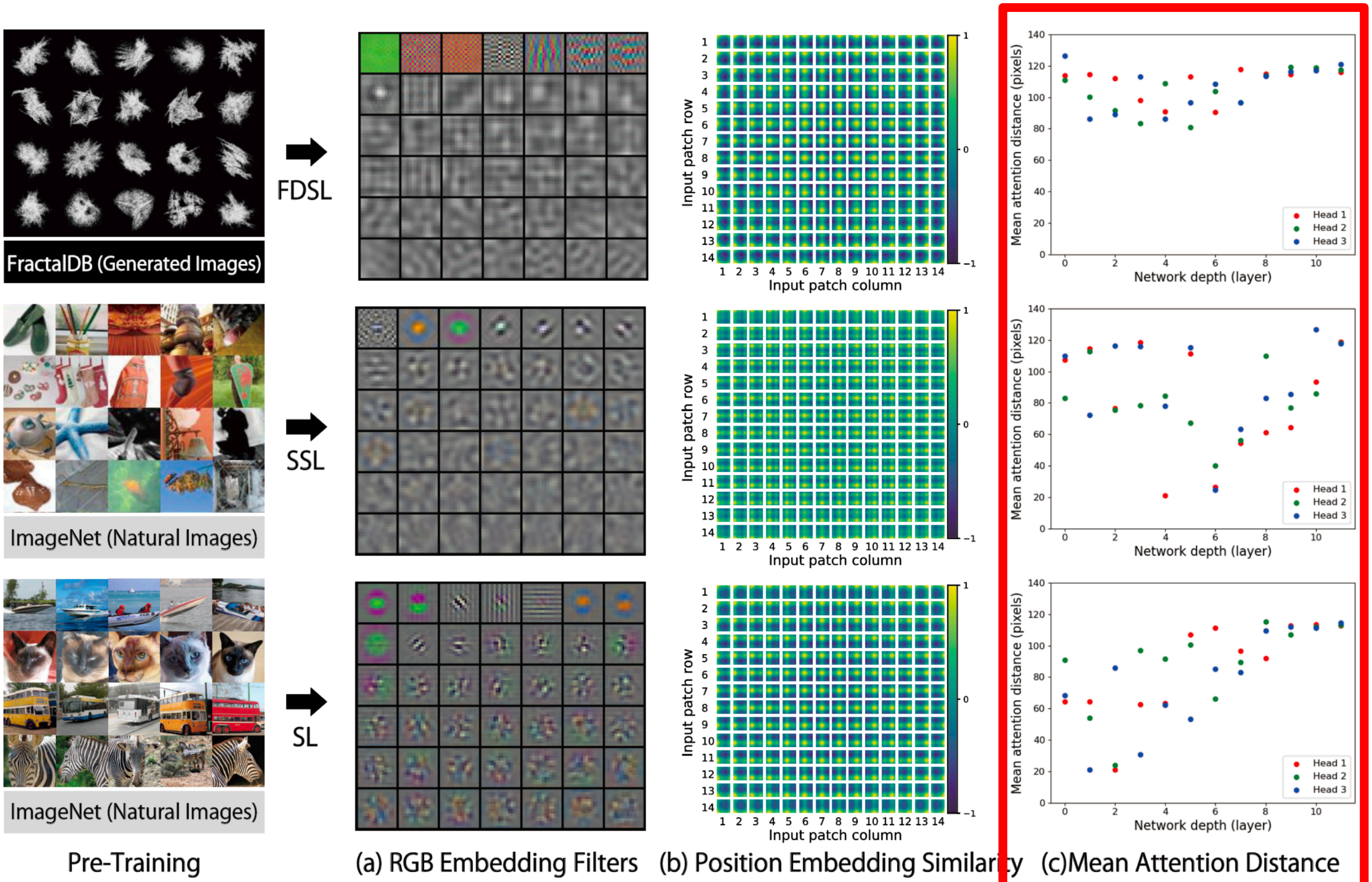
Cosine similarity of positional embedding



Pre-Training (a) RGB Embedding Filters (b) Position Embedding Similarity (c)Mean Attention Distance

Visualization of mean attention distance

FDSL tends to look at wide-spread areas



Can vision transformers learn without natural images?

→ Probably “Yes”. The FractalDB pre-training achieved to nearly perform the ImageNet-1k pre-training.

Future direction (1/3)

Towards a better pre-trained dataset

- FractalDB pre-trained model partially outperformed ImageNet-1k/Places-365 pre-trained models
- 80M Tiny Images/ ImageNet (human-related categories) withdrew public access
- We got a good feature representation without natural images

Future direction (2/3)

Different image representation from human annotated datasets

- FractalDB pre-trained model acquire a unique feature
- Steerable pre-training may be available
- Flexible dataset construction: Object detection, semantic segmentation…

Future direction (3/3)

Are fractals a good rendering formula?

- We are looking for better image patterns and their categories
- There is scope to improve the image representation and use a better rendering engine
- Any mathematical formulas, natural laws, and rendering functions can be employed to create image patterns and their image labels in the automatically created dataset

For the research community

@MIT A. Torralba Lab

Learning to See by Looking at Noise

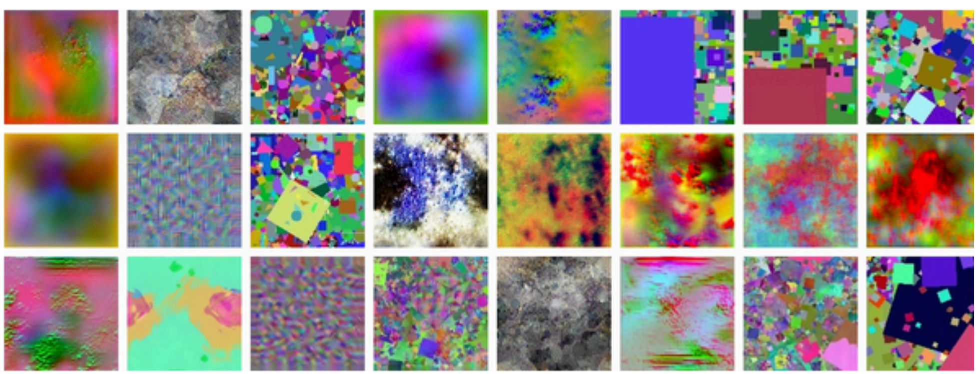
Manel Baradad*
MIT CSAIL

Jonas Wulff*
MIT CSAIL

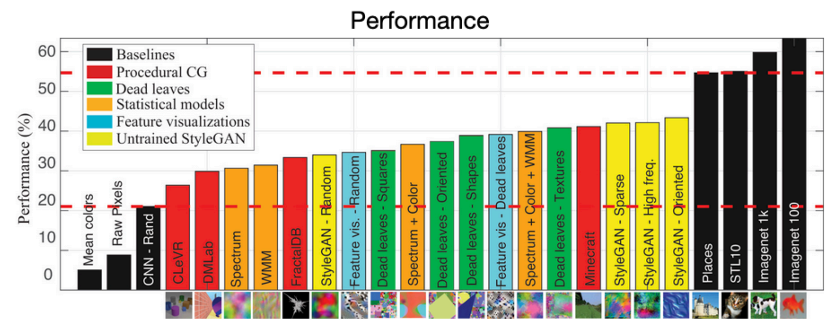
Tongzhou Wang
MIT CSAIL

Phillip Isola
MIT CSAIL

Antonio Torralba
MIT CSAIL



[Paper] [Code] [Datasets]



Top-1 accuracy for the different models proposed and baselines for Imagenet-100. The horizontal axis shows generative models sorted by performance. The two dashed lines represent approximated upper and lower bounds in performance that one can expect from a system trained from samples of a generic generative image model.

https://mbaradad.github.io/learning_with_noise/

For classification on ImageNet itself, the current state-of-the-art in self-supervised learning is, of course, much higher (81.0% [68]) than our results. Yet, only a few years ago self-supervised methods reported a similar accuracy to what we report here. We therefore believe it is an open and worthwhile challenge to improve learning from noise over the next 4 years as much as self-supervised learning improved over the last 4 years.

If we could improve the FDSL, ImageNet pre-trained model may be replaced so as to protect fairness, preserve privacy, and decrease annotation labor.

Thank you for watching.