数式から自動学習するAI

片岡 裕雄

産業技術総合研究所　人工知能研究センター

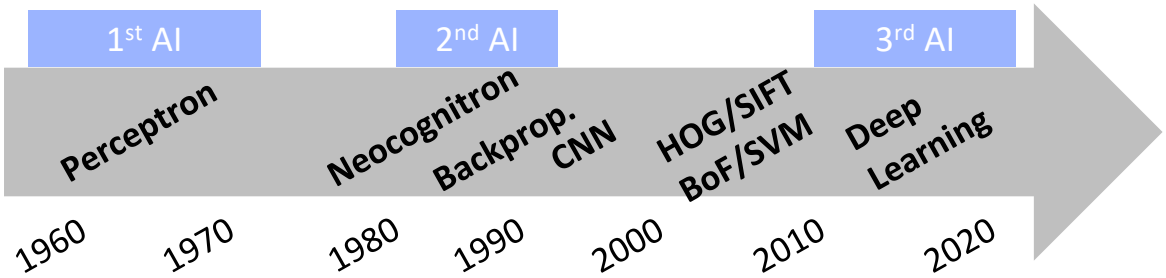
http://www.hirokatsukataoka.net/

深層学習は何をもたらした？

深層学習 隆盛の光と影

– メリット
  - もはや説明不要

– デメリット
  - 膨大なアノテーション/画像DLによる個人情報保護が必要

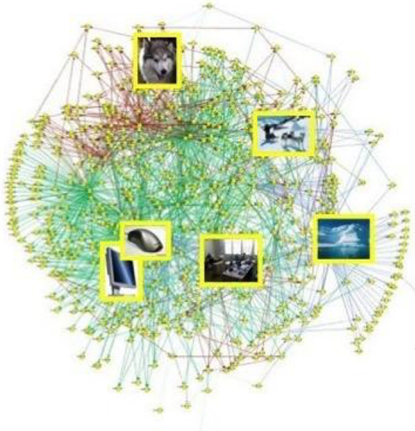


【膨大なアノテーション】

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

AMTにより5万人弱が参加，約2年を要した
数億画像DL，1400万枚収録，2.2万カテゴリ

【プライバシーの保護】



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

実はプライバシなど**権利が不透明**
現在でも学術・教育目的のみ

アノテーション問題/法令遵守の障壁は非常に大きい

近年発覚した問題：AI倫理問題

大規模データセット関連にて発生

– デメリット

- 公平性及び透明性などのAI倫理問題

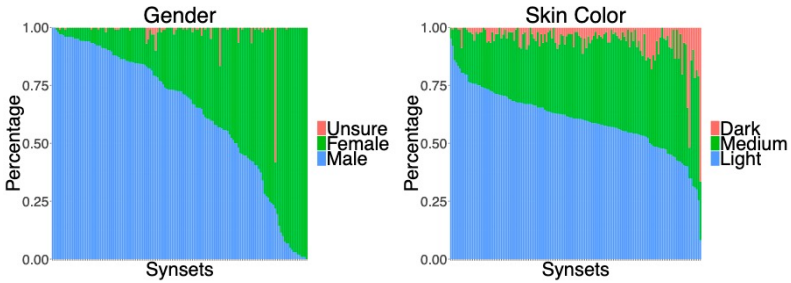
【攻撃的ラベル問題】

攻撃的ラベルを含むとして80M Tiny Imagesが公開停止に追い込まれた

https://groups.csail.mit.edu/vision/TinyImages/

【不公平性問題】

カテゴリごとに性別・人種のラベル数に偏りが生じていた

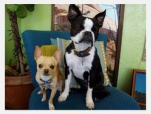


https://arxiv.org/pdf/1912.07726.pdf

【不透明性問題】

ILSVRC2012 の“疑わしい画像”の調査


ジャガー?


チワワ?


インゲン豆?

→ 不適切なアノテーション等が存在

専門家30人弱で3日間のワークショップを開催
“疑わしい画像”を収集し，議論をもとに分類

ImageNetの28.4万/128万画像を目視で再ラベル付
約6%にあたる17,419枚が疑わしい画像と判明

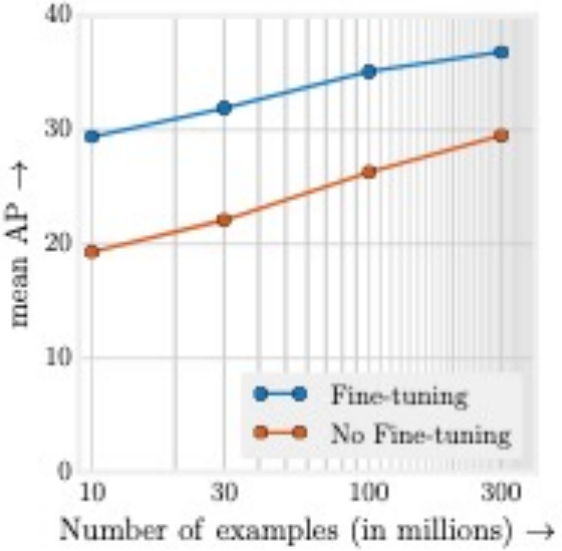
実用化に向けてAI倫理問題は無視できない課題

億単位の超大規模データセット

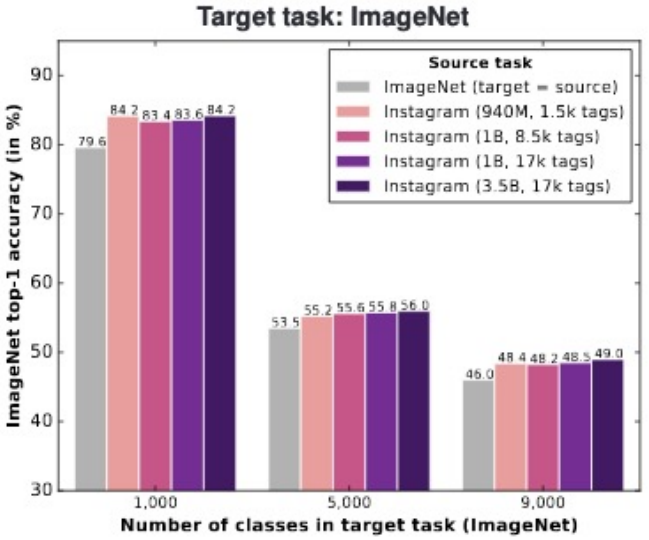
JFT-300M (Google, 2017) / IG-3.5B (Facebook, 2018)

約３億画像　　約３５億画像

ImageNetの数百倍のデータセットは認識性能の向上に寄与するか？
-> YES: 対数レベルで比例して性能は向上（10倍ごとに+数%; 左下図）
-> 35億枚の学習画像を用いた場合，モデルの変更なしで当時の最高精度達成（右下図）



Google, ICCV 2017
http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Revisiting_Unreasonable_Effectiveness_ICCV_2017_paper.pdf



Facebook, ECCV 2018
https://arxiv.org/pdf/1805.00932.pdf

「超大規模データは正義」だが現在も未公開データセット

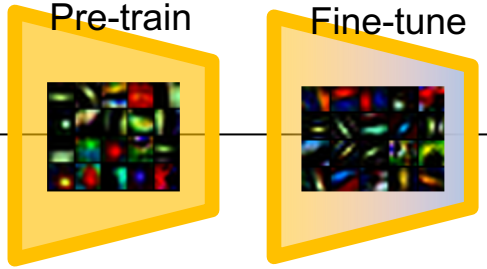
近年の学習戦略

教師あり学習（Supervised Learning）

人間のラベル付により学習の成功を確約，非常に強い特徴表現を獲得

e.g. ImageNet, Places, Open Images

gluon-cv.mxnet.io

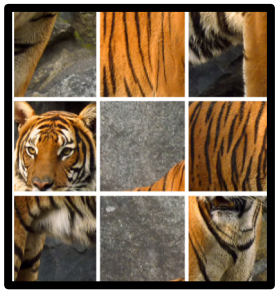


ImageNet + ResNet-50
76% @ImageNet val.

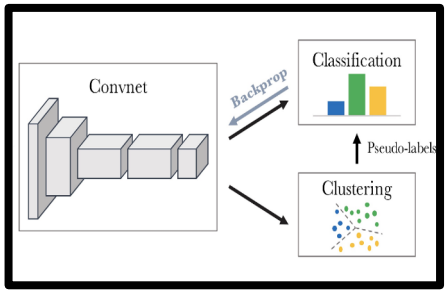
[He et al. CVPR16]

自己教師あり学習（Self-supervised Learning）

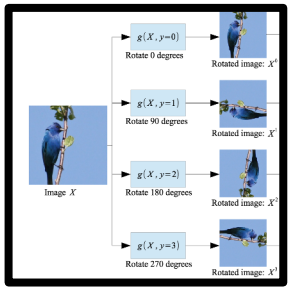
一定の法則により画像にラベル付，基礎的な視覚特徴獲得を自動化



Jigsaw Puzzle
[Noroozi al. ECCV16]



DeepCluster
[Caronet al. ECCV18]



Rotation Classify
[Gidaris et al. ICLR18]

SimCLR + ResNet-50
69%@ImageNet val.

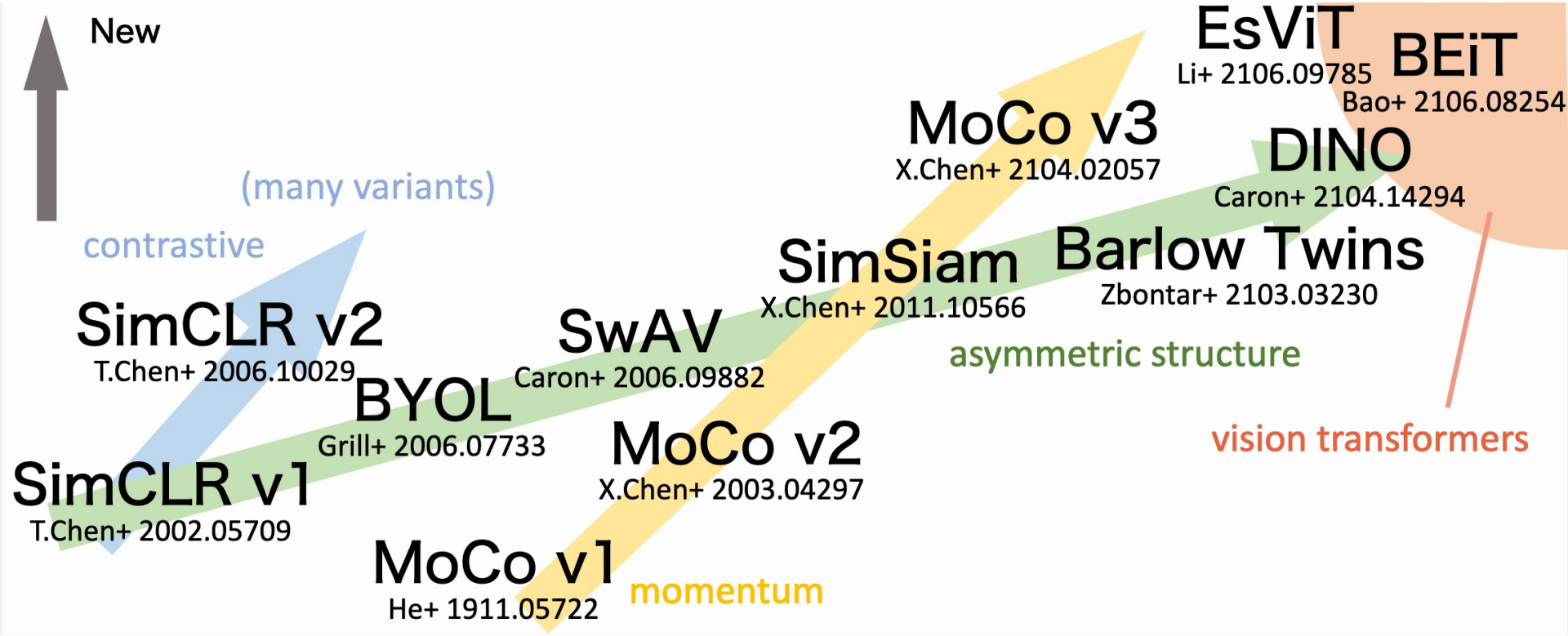
[Chen et al. ICML20]

自己教師あり学習は現在も進展中

自己教師あり学習の2021年現在

手法改善と精度向上

最近の手法（BEiTやSimSiam等）は一部で人間の教師を置き換える性能まで到達



MIRU2021チュートリアル「限られたデータからの深層学習」より引用

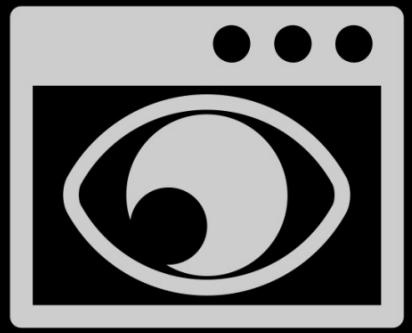
但し，AI倫理問題は実画像を使う限り起こり得る

３つの問題を解決する壁はあまりにも大きいのでは？

Annotation

FATE

Fairness, Accountability, Transparency and Ethics



Privacy

実画像を用いずに視覚特徴の学習はできるのか？

実画像に含まれる自然法則

– ImageNetを眺めていると法則が見えてくる

Fractal geometry from ImageNet dataset

深層学習では視覚特徴をある種の
自然法則から学んでいるのでは？

自然法則から画像に直接投影・学習することを着想

Pre-training without Natural Images

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

片岡 裕雄

産業技術総合研究所　人工知能研究センター

http://www.hirokatsukataoka.net/

提案の枠組み

数式ドリブン教師あり学習

Formula-driven Supervised Learning (FDSL)

– 生成規則（本発表：フラクタル幾何）から画像と教師ラベルを同時生成

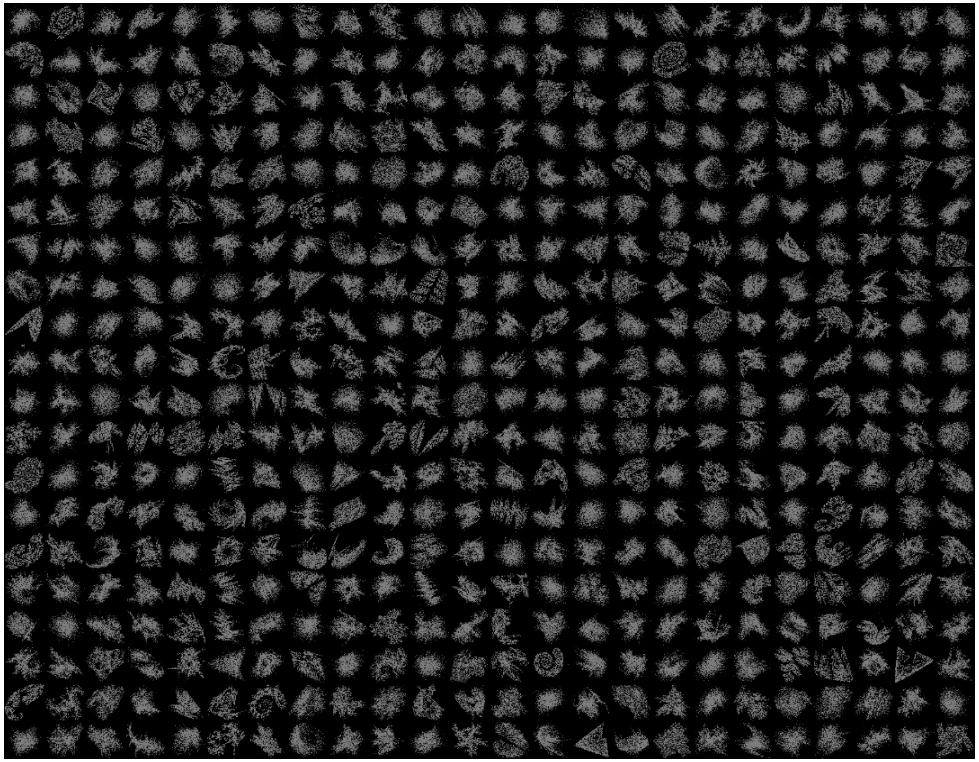
人間によるラベル付・実画像を一切用いず基礎的な視覚特徴を獲得

提案法FractalDBの効果

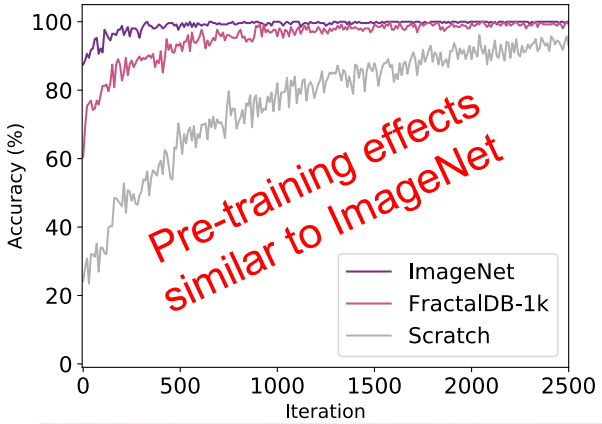
FractalDB

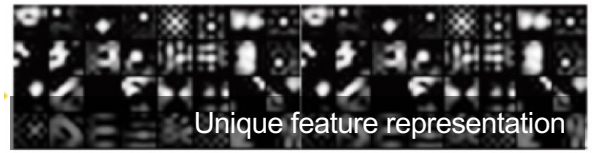
– 教師あり学習に近い事前学習効果を享受

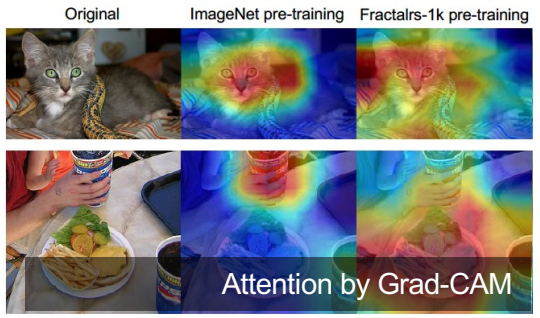
収束速度/特徴表現/アテンションを可視化



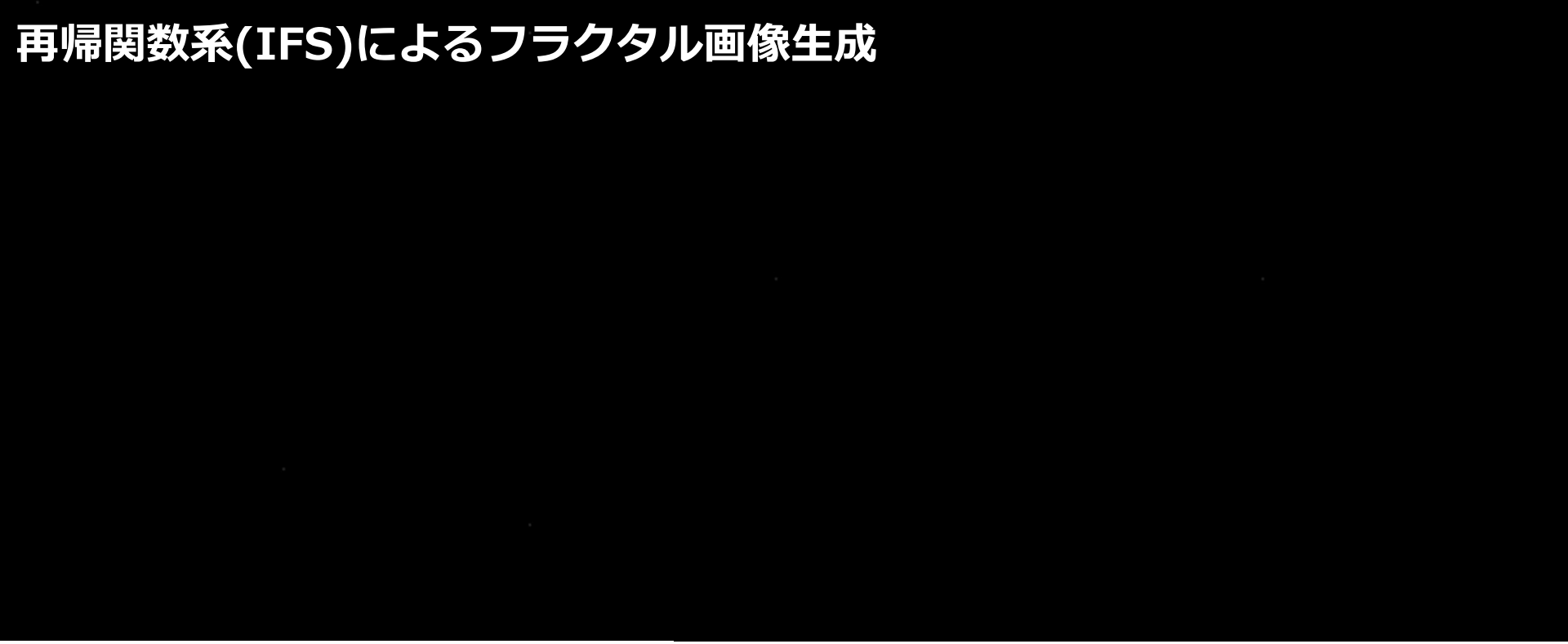
Pre-train







再帰関数系(IFS)によるフラクタル画像生成



生成規則(式)

ラベルの生成 (ランダムサンプリング)

$$\Theta = \{(\theta_i, p_i)\}_{i=1}^{N}$$

データの生成 (IFS)

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

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

$$p_i = p(w^* = w_i) \qquad \boldsymbol{x}_{t+1} = w^*(\boldsymbol{x}_t)$$

カテゴリ探索

フラクタル画像生成

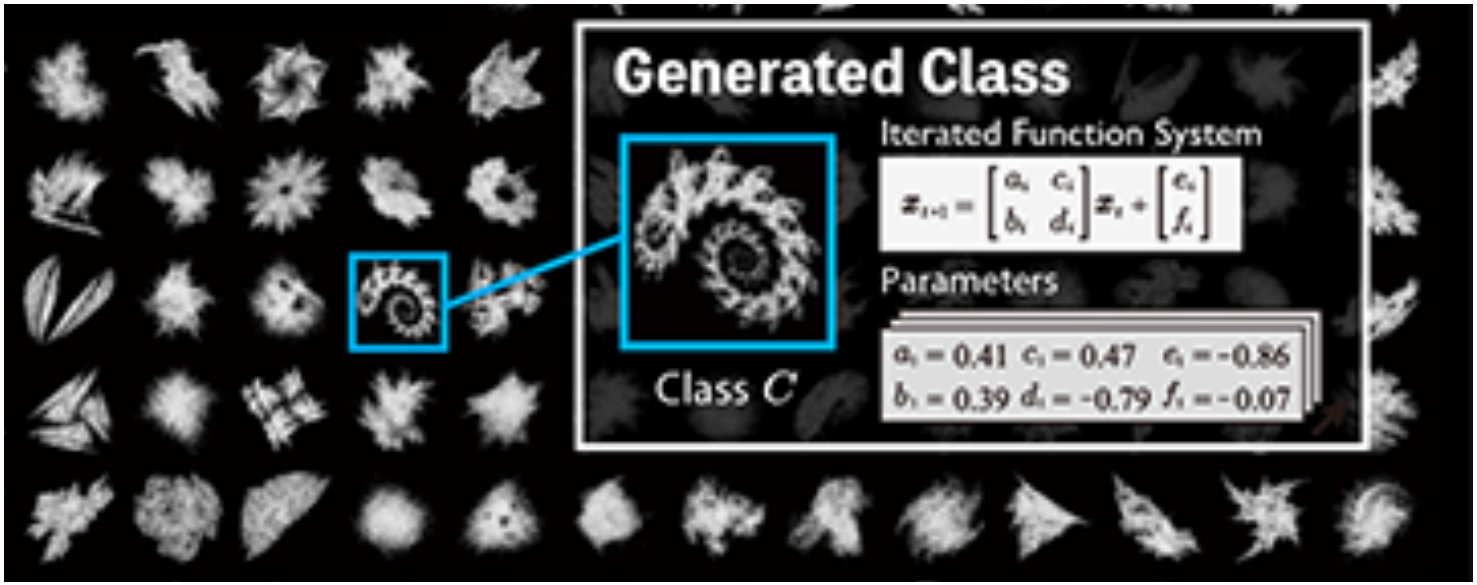
インスタンス拡張

について説明

カテゴリ探索

ランダム生成・簡単なチェックにより判断

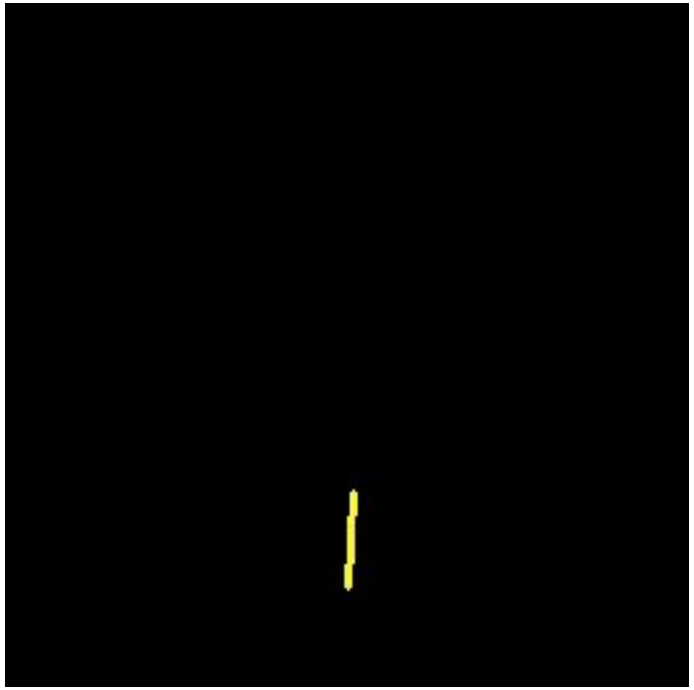
1. ランダムパラメータセット $\Theta = \{(\theta_i, p_i)\}_{i=1}^{N}$ に従い画像生成
2. フラクタル画像の占有率 r が閾値(> 0.2)以上ならカテゴリ c を登録
3. 設定カテゴリ数 C になるまで繰り返し
   - パラメータセット自体がフラクタルカテゴリとなる



Fractal categories in FractalDB

フラクタル画像生成

初期値設定と再帰的レンダリング



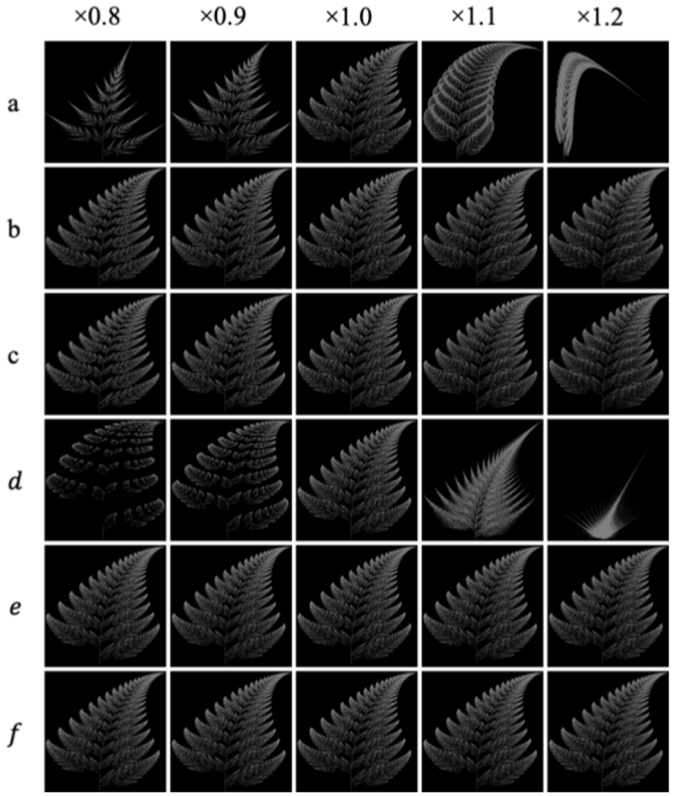
Fractal image with IFS

1. 座標 x_0 をランダムに選択，$t = 0$とする
2. 座標 x_t に点をプロット
3. 次の座標 $x_{t+1} = w^*(x_t)$ を計算
4. $t < T$ なら $t = t + 1$ として$2 \sim 3$を繰り返す

画像インスタンス拡張

3種の画像インスタンス拡張手法

1. パラメータセットの変動 (x25)
2. 画像回転 (x4)
3. 3x3のパッチパターン (x10)



Parameter set (x25)

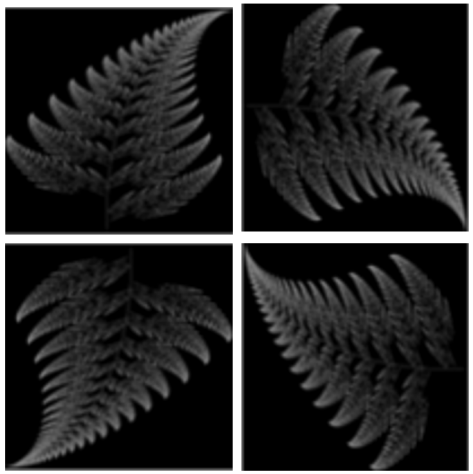
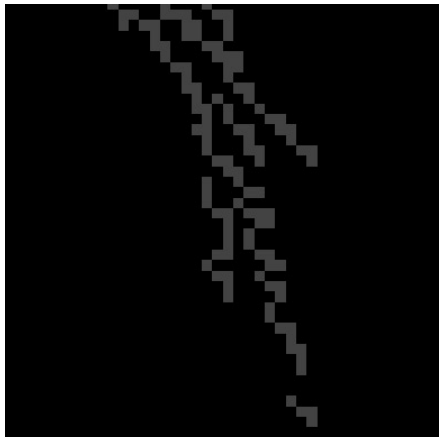


Image rotation (x4)



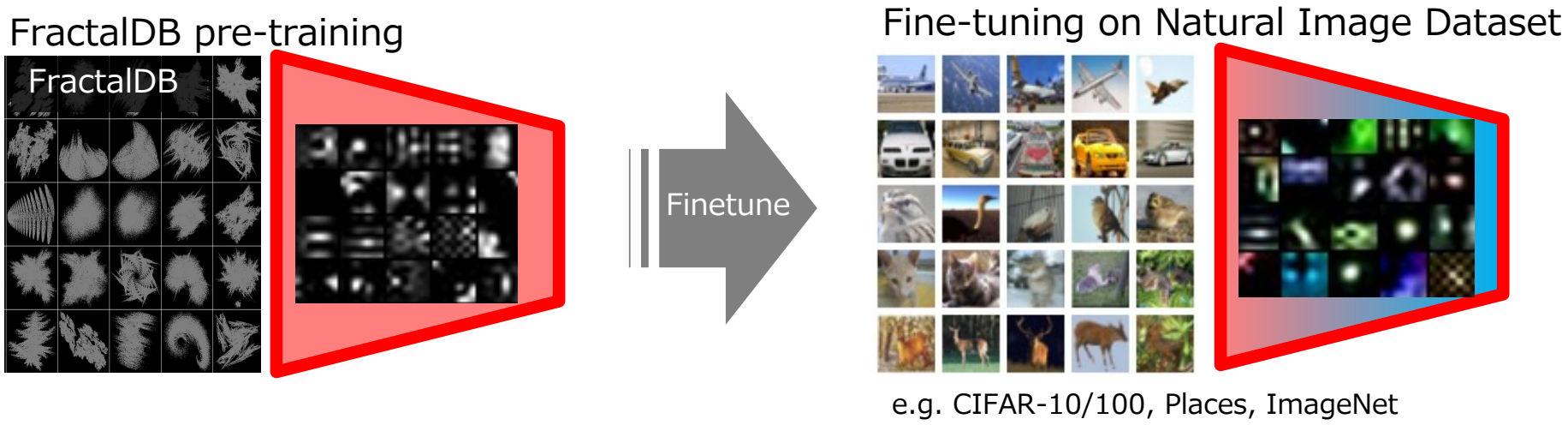
Patch pattern (x10)

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

学習の方式

事前学習Pre-training & 追加学習Fine-tuning

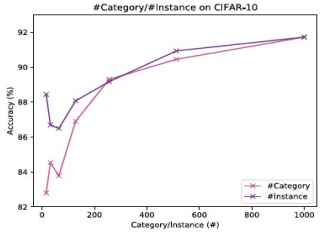
- 基本は実画像による手順と同様
- FractalDBはフラクタルカテゴリを推定するタスク



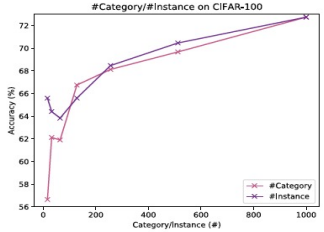
FractalDBのカテゴリ

膨大なパラメータ調整の結果,

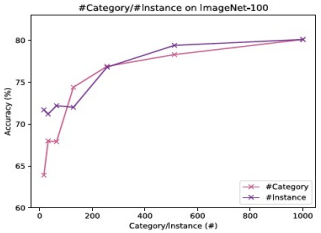
- #Category, #Instance, Patch Rendering が有効と判断
- パラメータセットの変動重み w により形状変化を調整し精度改善



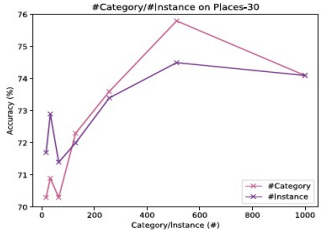
(a) CIFAR10



(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

詳細は論文をご参照ください

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

- 実験ではResNet-50を適用
- パラメータはImageNet事前学習とほぼ同様

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

- Scratchと比較すると精度は良い傾向
- ImageNet/Placesのような大規模データは単体で学習可能

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

- ほとんどの場合，DeepClusterよりも高精度

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

- 全ての設定において10万画像規模の事前学習よりも高精度

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

- この時点では100万画像規模の教師あり学習には及んでない
- 一部，ImageNet/Places事前学習を超える精度

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

- FractalDBと自己教師では視覚特徴を獲得しきれない傾向

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

- 1, 2層あたりが良好な視覚特徴を獲得
- 3~5層を固定して追加学習すると精度低下著しい

Results (4/5)

We compare Formula-driven Supervised Learning with other principles

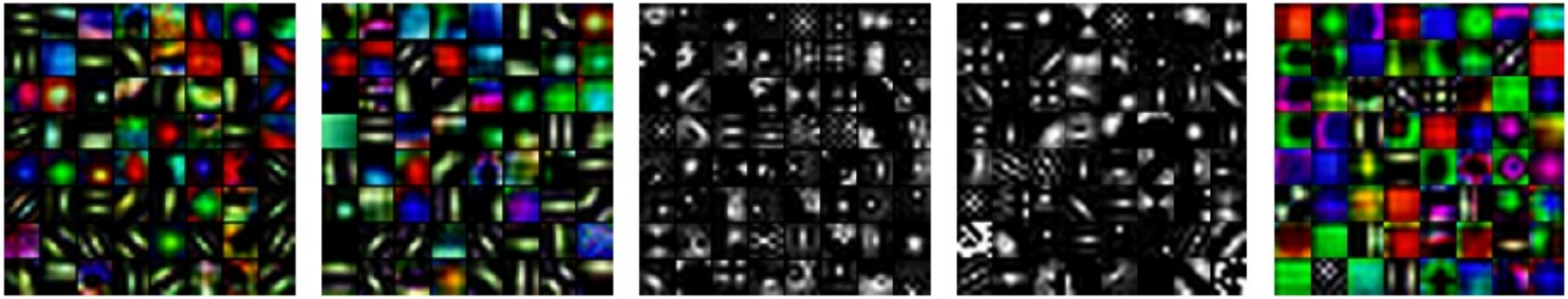
The FractalDB pre-trained model outperforms other methods

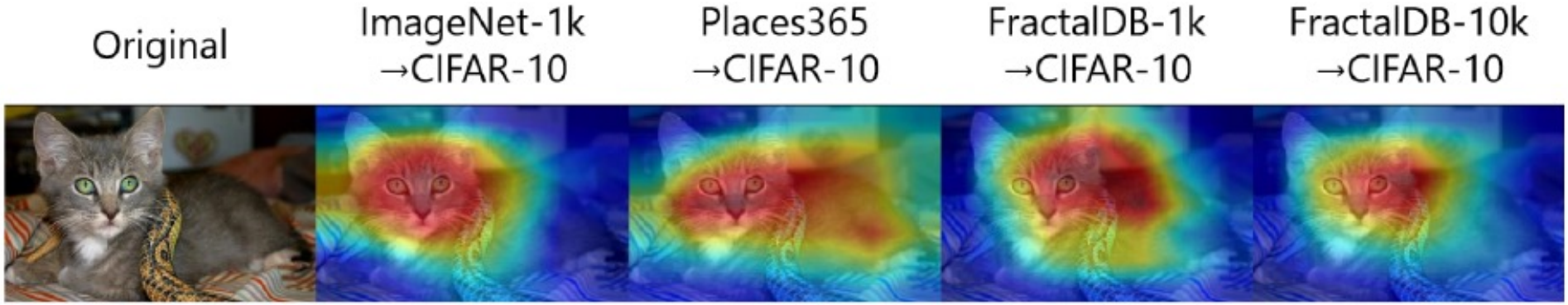
Pre-training	C10	C100	IN100	P30
Scratch	87.6	60.6	75.3	70.3
Bezier-144	87.6	62.5	72.7	73.5
Bezier-1024	89.7	68.1	73.0	73.6
Perlin-100	90.9	70.2	73.0	73.3
Perlin-1296	90.4	71.1	79.7	74.2
FractalDB-1k	**93.4**	**75.7**	**82.7**	**75.9**

- 他のFDSL方式（ベジェ曲線/パーリンノイズ）と比較
- フラクタル画像が最も高精度

Results (5/5)

Visualization of Conv1





- 実画像DBとは異なる視覚特徴を獲得
- 画像の反応領域は対象物体に配置

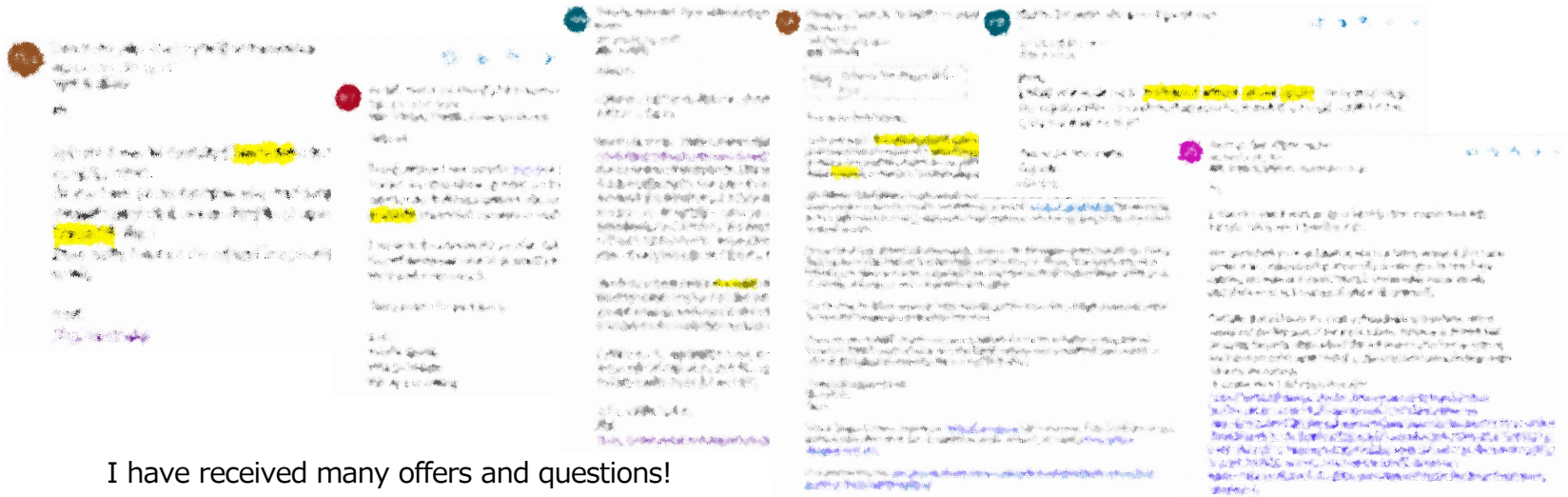
提案後，，，

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

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

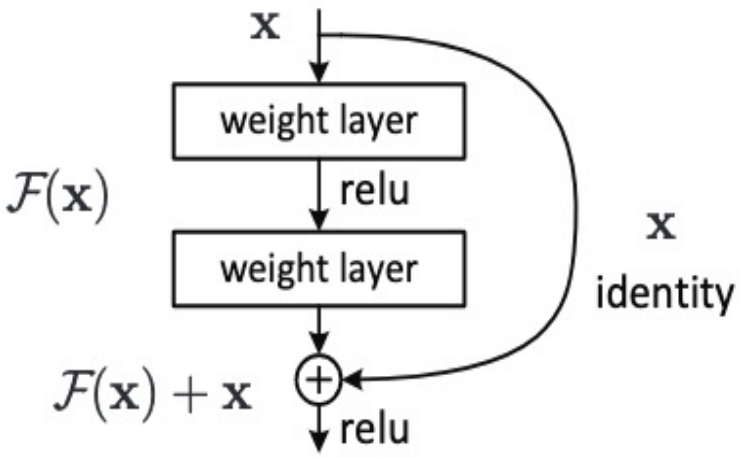


...

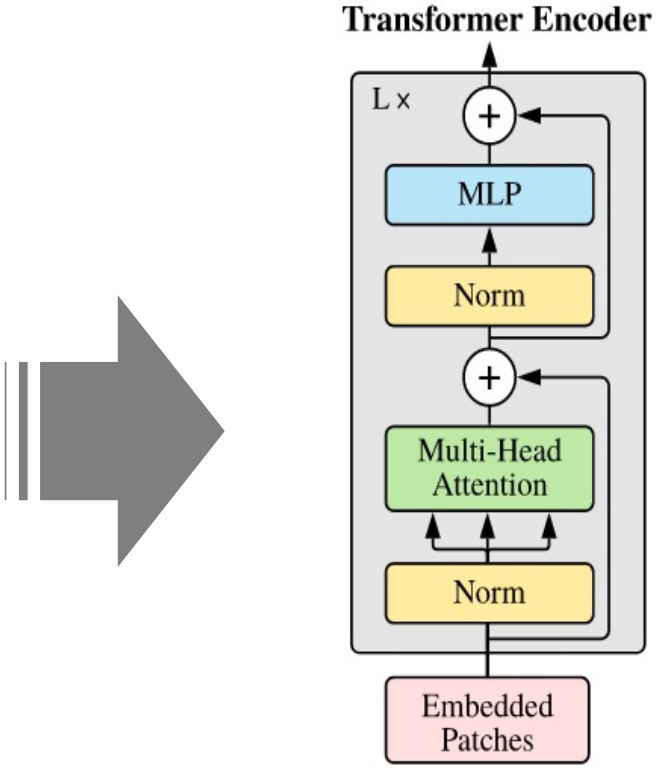
I have received many offers and questions!

Computer Vision分野のパラダイムシフト

‘Convolution’ から ‘Self-attention’ へ



[He al. CVPR16]



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

Can Vision Transformers Learn without Natural Images?

AAAI 2022

片岡 裕雄

産業技術総合研究所　人工知能研究センター

http://www.hirokatsukataoka.net/

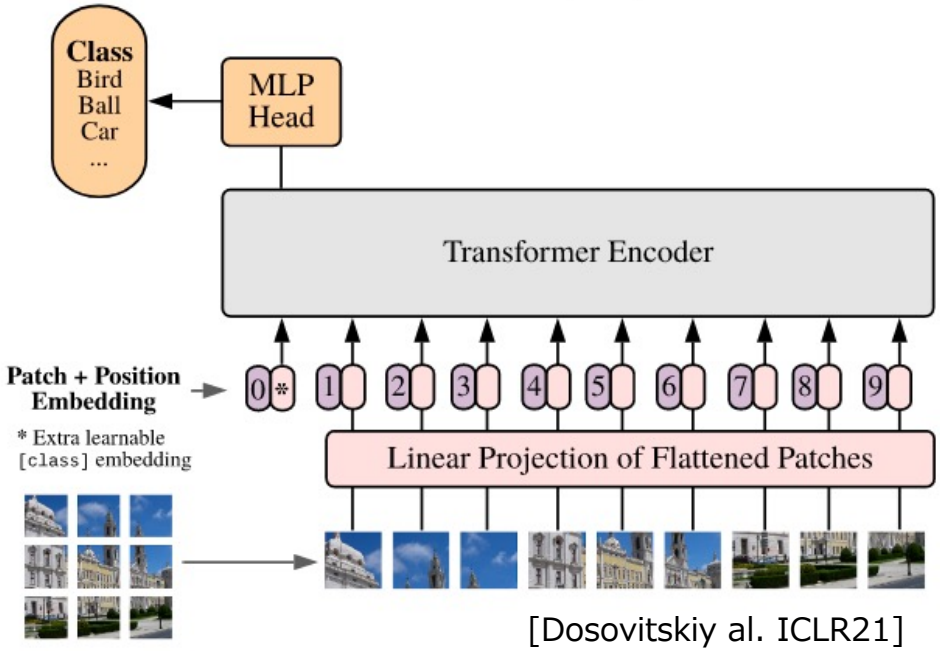
Vision Transformer (ViT)のこれまで

もうひとつのシフト（データ利用）

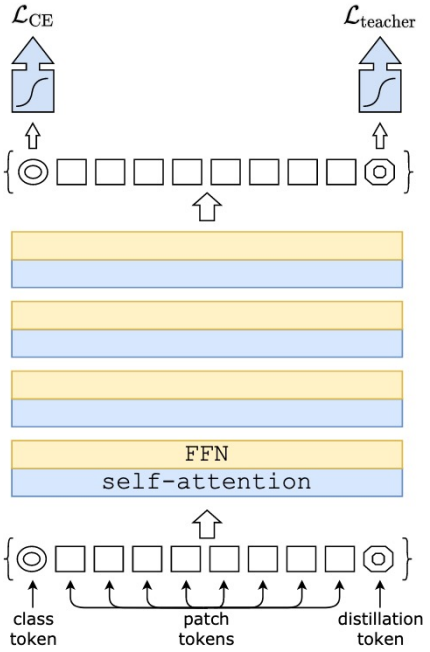
– ‘ViT’ から ‘DeiT’ (Data-efficient image Transformer)

– 事前学習の画像数を 3億 から 100万 へ削減

ViTの事前学習は実画像ではなくても良いのではないか？



[Dosovitskiy al. ICLR21]



[Touvron al. arXiv20]

ViT + FractalDBで学習

アーキテクチャ

- ViT
  - 但し，教師の与え方はDeiTを利用

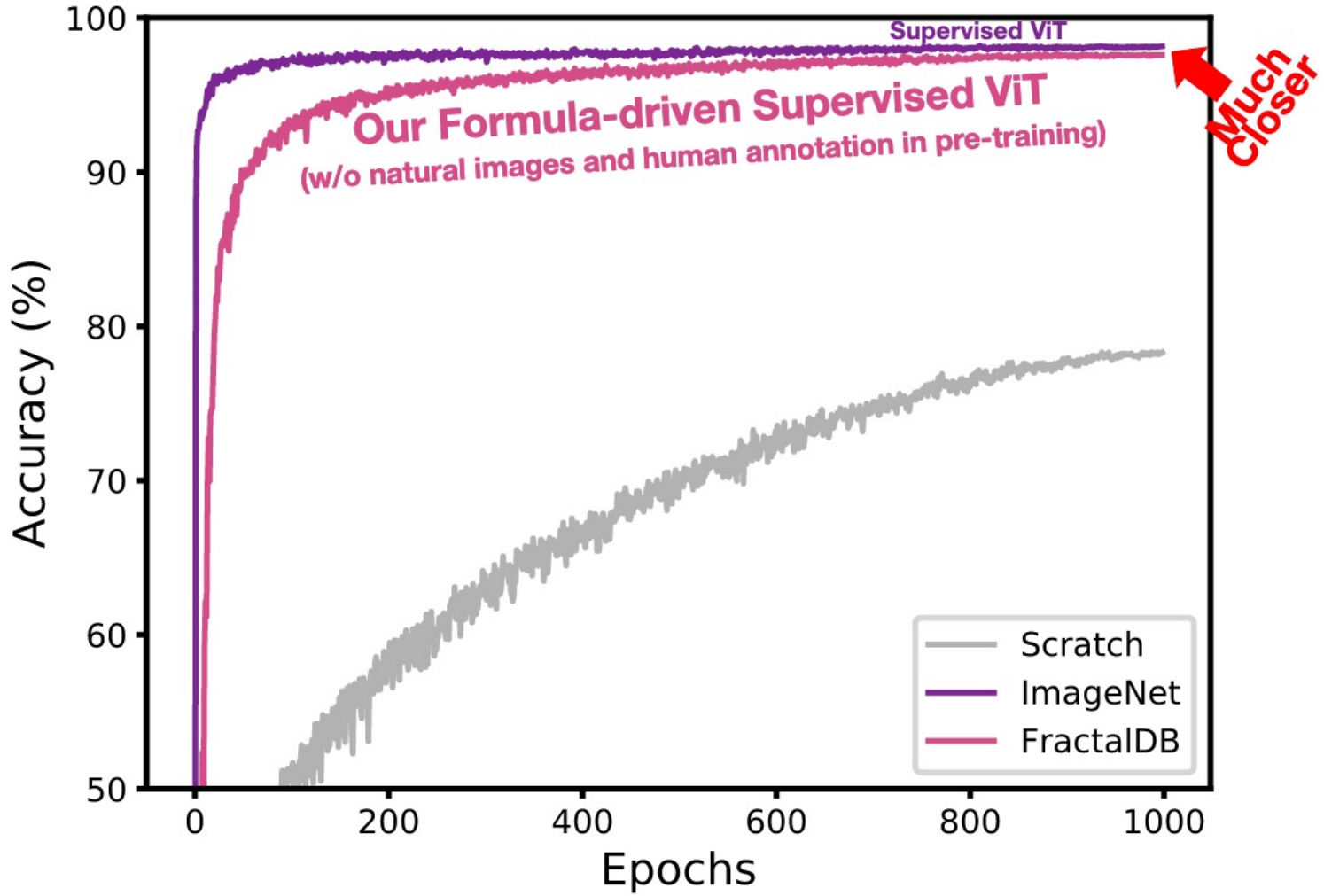
データセット

- FractalDB
  - カラーよりもグレースケールの方が高精度
    - ResNet: カラーが高精度
    - DeiT: グレースケールが高精度
  - 長い学習時間
    - ViTでは300エポック学習

FractalDB pre-trained Vision Transformer

– Scratchよりも圧倒的に高精度

– ImageNet事前学習により近い事前学習効果



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	–	**89.4**
ImageNet-100	Natural	Supervision	94.7	77.8	67.4	97.2	78.8	78.1	–
ImageNet-1k	Natural	Supervision	**98.0**	**85.5**	**89.9**	**99.4**	**88.7**	**80.0**	–
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**

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

- Places-365を一部凌駕，ImageNetに近い精度で認識

Results (2/2)

The proposed method recorded higher accuracies than SSL methods with MoCoV2, Rotation, and Jigsaw

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	**80.8**	87.1
SimCLRv2	YES	**97.4**	**84.1**	**84.9**	**98.9**	**86.2**	80.0	**88.5**
FractalDB-10k	NO	**97.6**	**83.5**	**87.7**	**98.8**	**86.9**	78.5	**88.8**

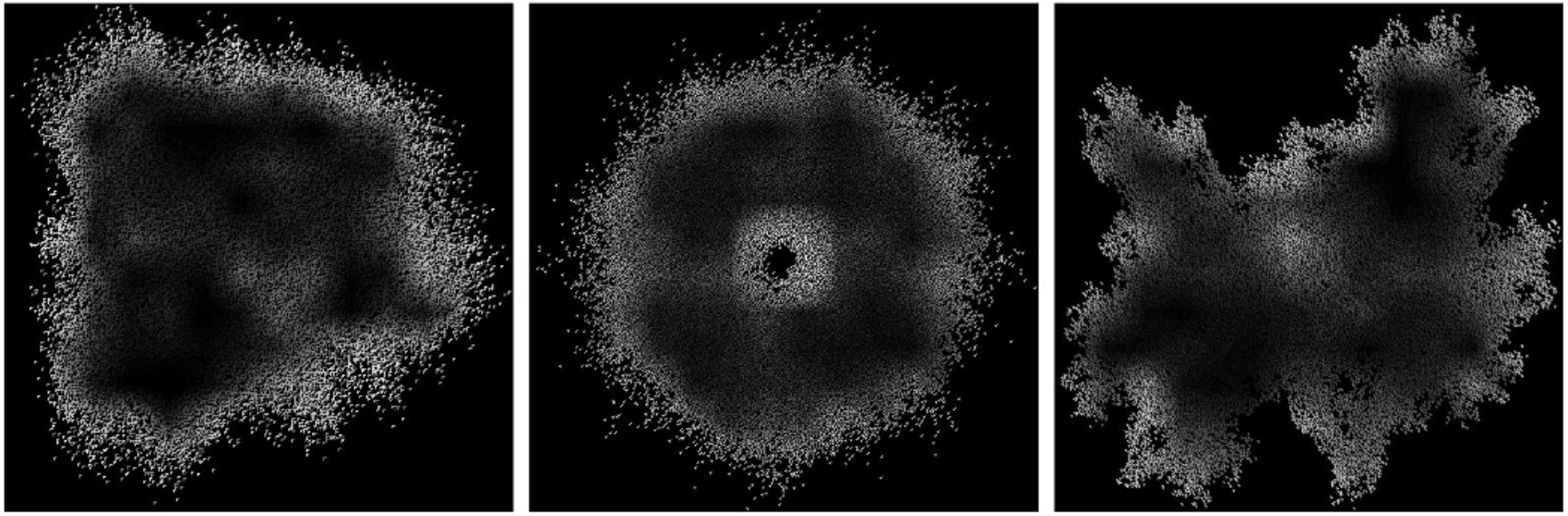
Underlined bold: best score, **Bold**: second best score

- 平均値ではImageNet + SimCLRv2を超える精度を達成
- MoCov2, Rotation, Jigsawによる自己教師よりも高精度

アテンションマップの可視化

FractalDB事前学習は輪郭にフォーカスする傾向？

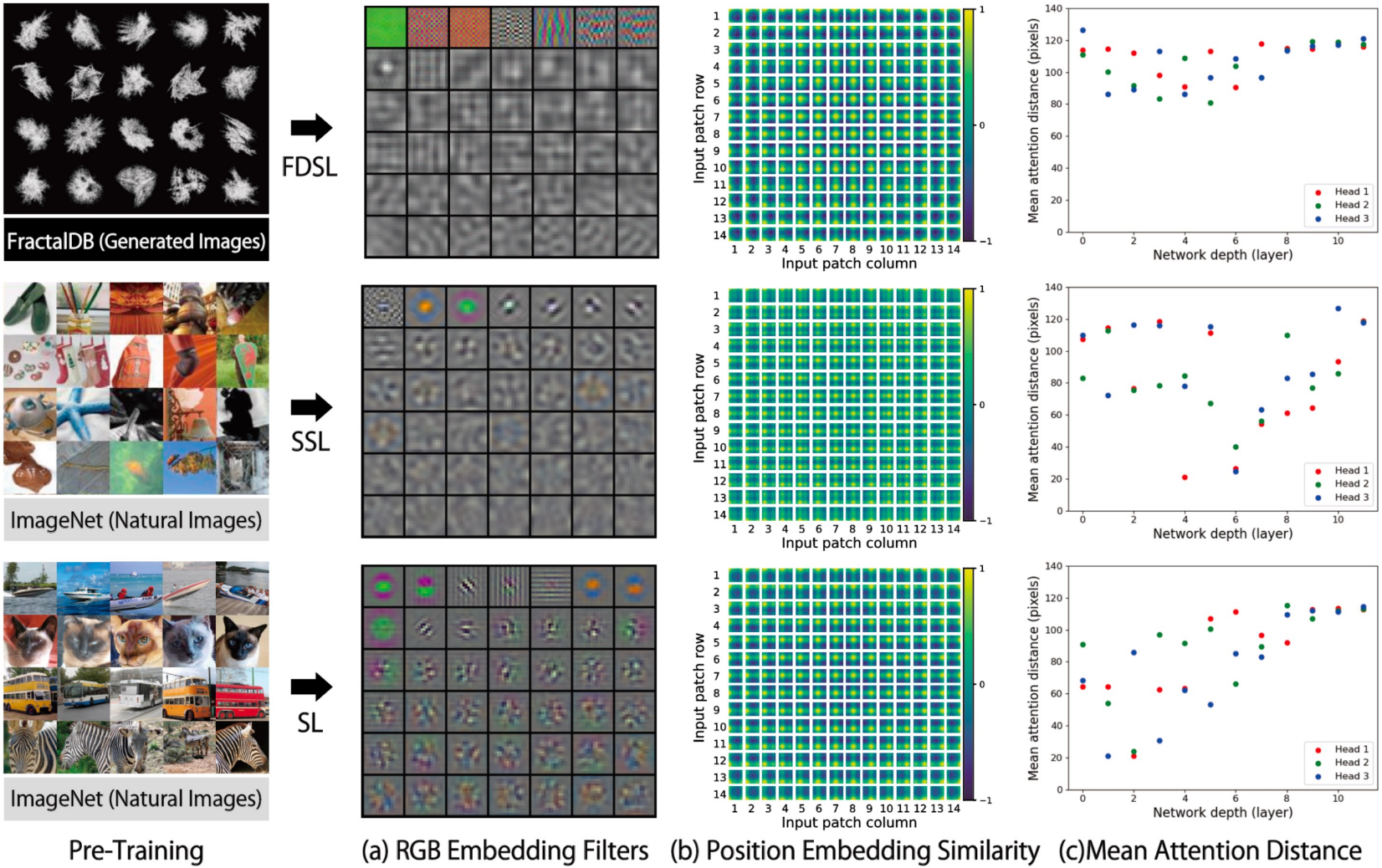
– 画像はFractalDB事前学習モデルのself-attentionを可視化



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

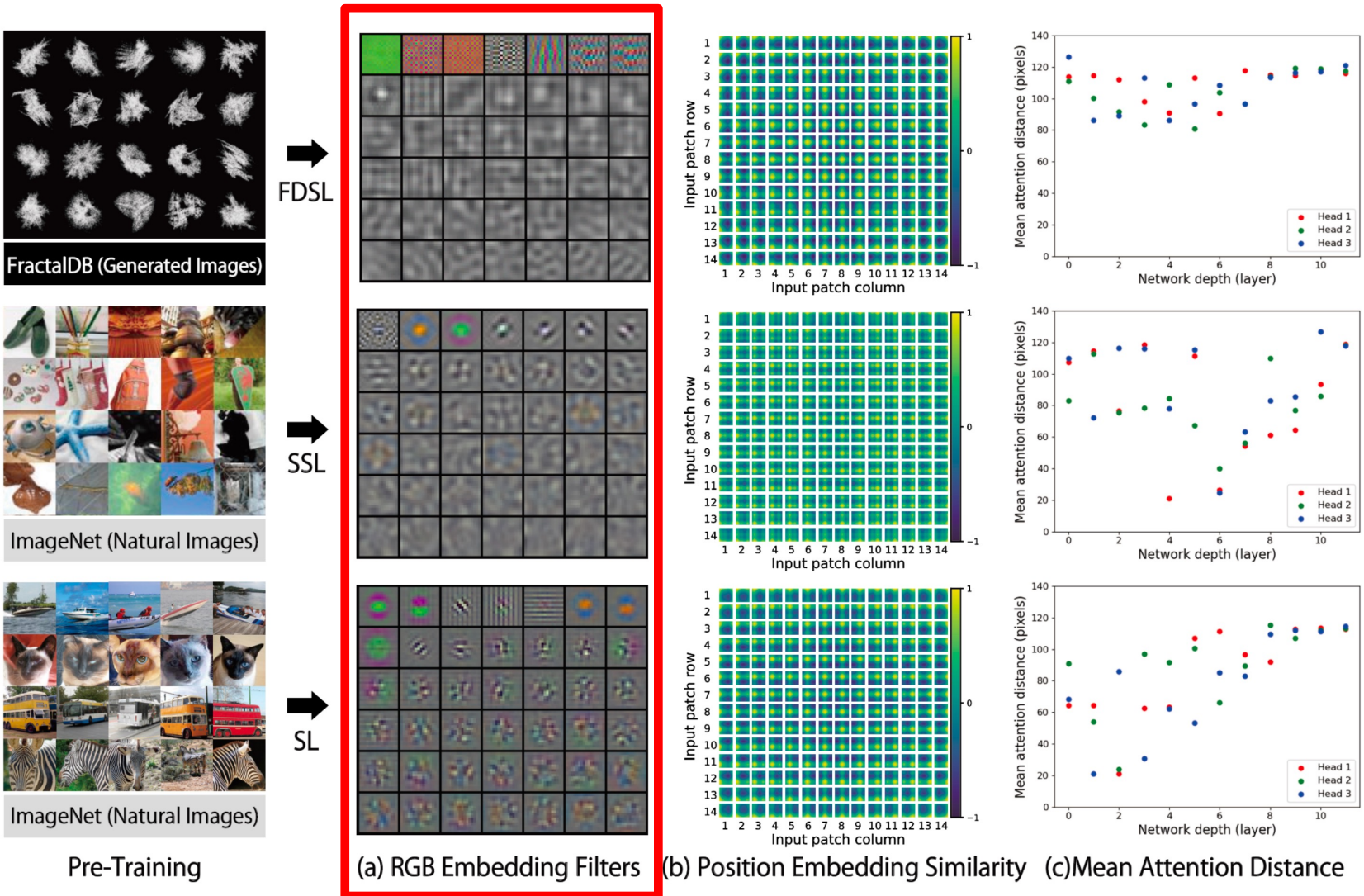
可視化

教師ありSL，自己教師SSL，数式教師FDSL比較



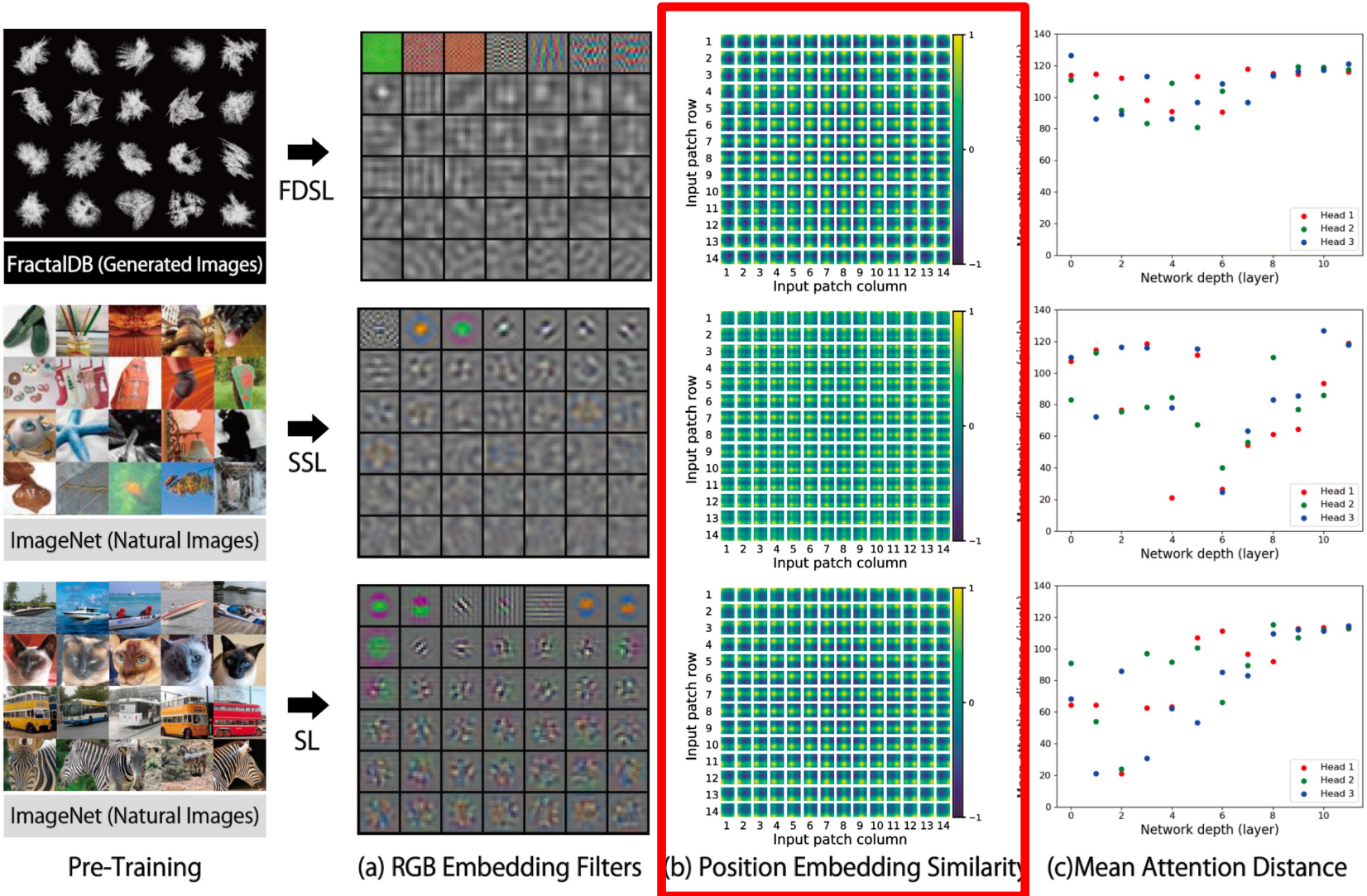
可視化：Embedding Filters

初期フィルタの視覚特徴



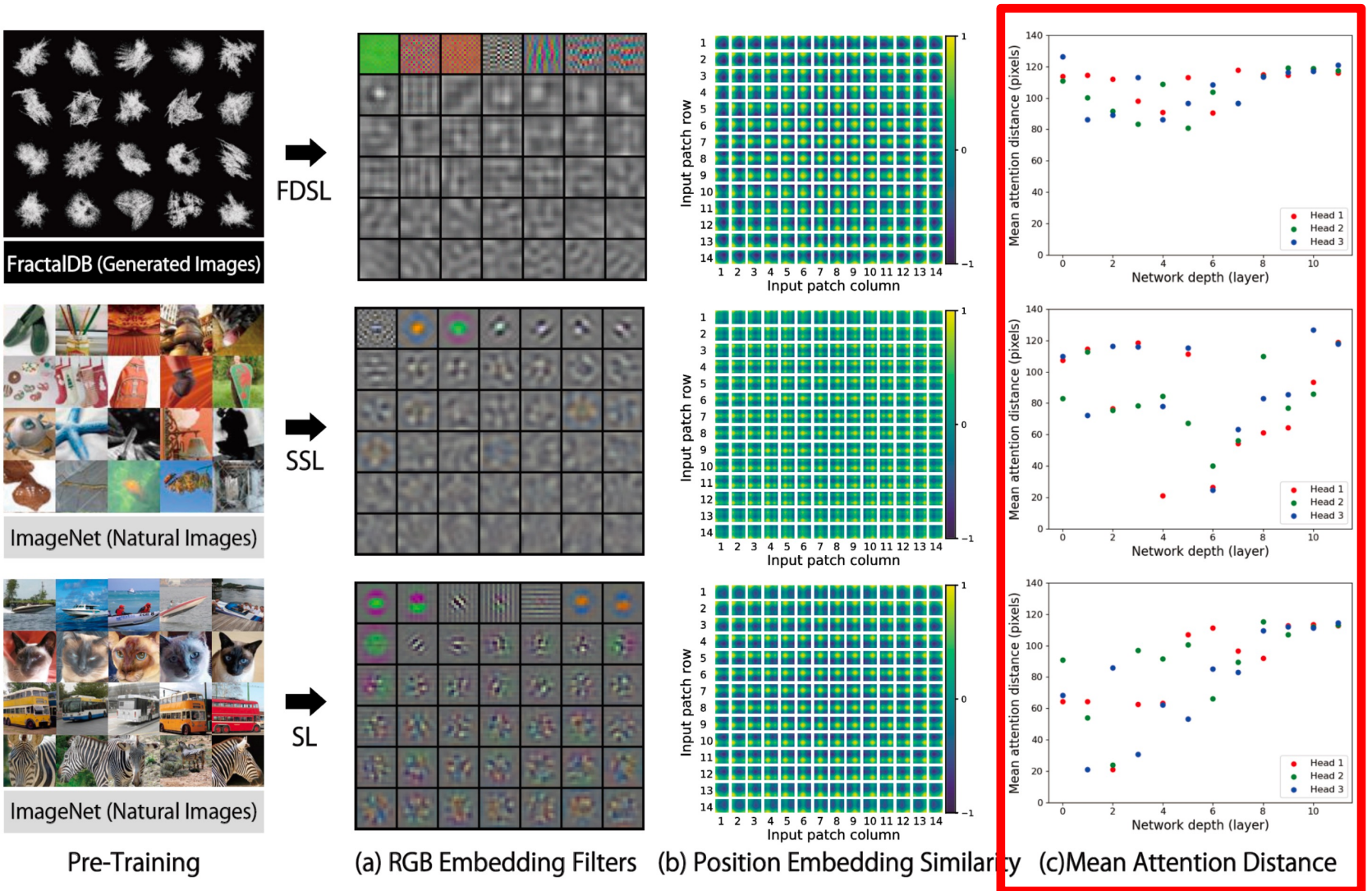
可視化：Position Embedding Similarity

位置による可視化



可視化：Mean Attention Distance

FDSLは初期層から離れた位置から特徴抽出



ViTは実画像を用いず学習できるのか？

→ 恐らく“Yes”：

- 収束・視覚特徴・自己注視の点で高い事前学習効果を発揮
- 今後も精度は向上し続けると予想

Replacing Labeled Real-image Datasets with Auto-generated Contours

CVPR 2022

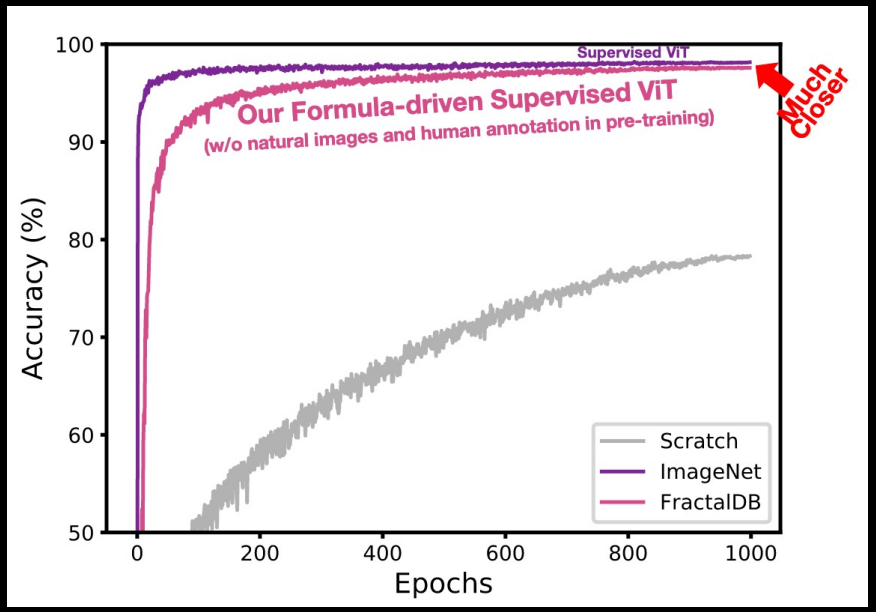
Hirokatsu Kataoka[*], Ryo Hayamizu[*], Ryosuke Yamada[*], Kodai Nakashima[*], Sora Takashima[*,], Xinyu Zhang[*,**], Edgar Josafat MARTINEZ-NORIEGA[*,**], Nakamasa Inoue[*,**], Rio Yokota[*,**]**

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

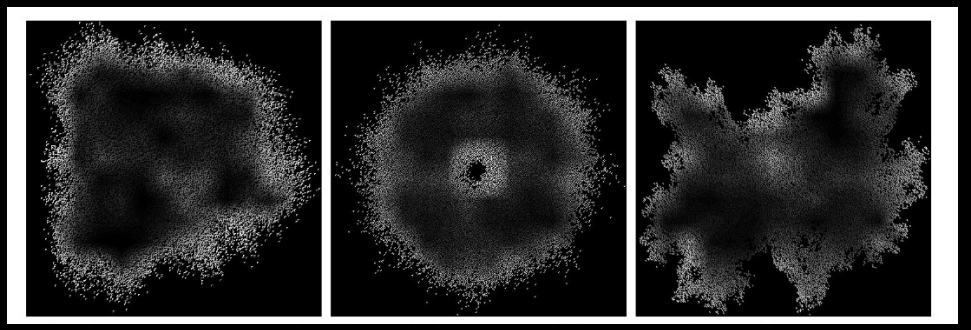
Can Vision Transformers Learn without Natural Images? (AAAI22)

FractalDBでVision Transformer（ViT）の事前学習に成功

- 実画像枚数を従来の14,000,000から実質0に



ViTの自己注視可視化結果

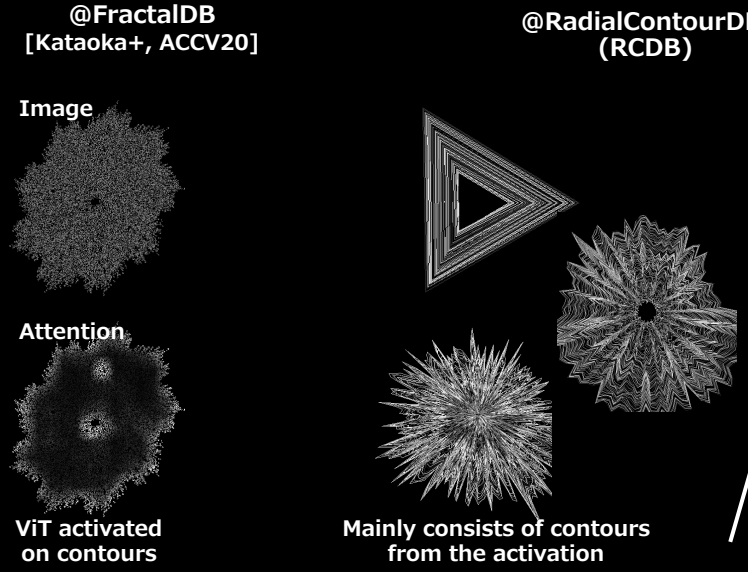


→実はフラクタルが重要なのではなく，
輪郭が複雑であれば良いのでは？

自動生成輪郭でラベル付実画像データセットを置き換える

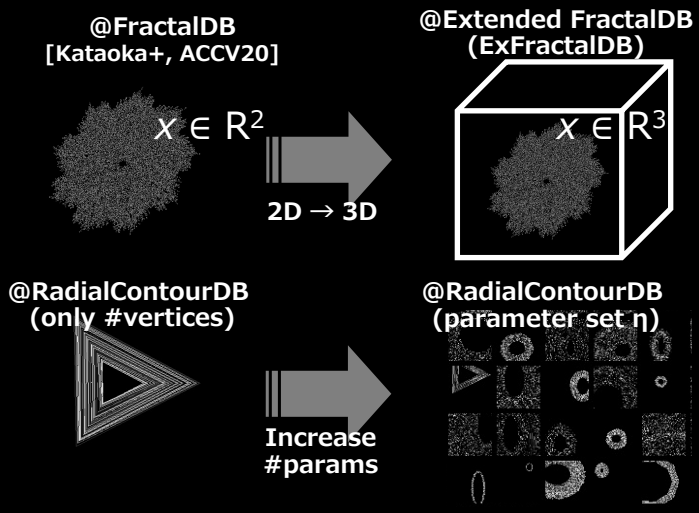
- ふたつの仮説を検証

Hypothesis 1:
Object contours are what matter in FDSL datasets
FDSLの画像表現では物体輪郭こそが重要



輪郭形状の「極端な例」として，画像の主要成分が輪郭である放射輪郭（Radial Contour）を実装

Hypothesis 2:
Task difficulty matters in FDSL pre-training
FDSL事前学習のタスク難易度が精度向上に寄与

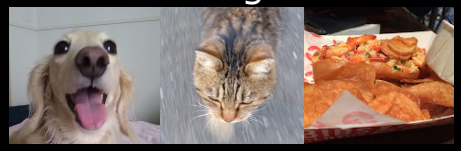


事前学習タスクの難易度と直結する成分は「数式のパラメータ数」

一般物体認識・物体検出・領域分割タスクにおける評価

ImageNet-1k / MS COCOデータセット

実画像: ImageNet-21k



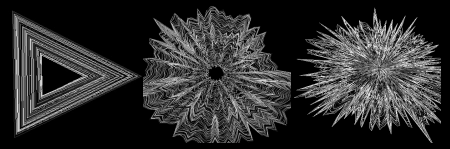
一般物体認識の性能
ImageNet における精度

81.8%

3Dフラクタル幾何画像:
ExFractalDB-21k

82.7%

放射輪郭画像: 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

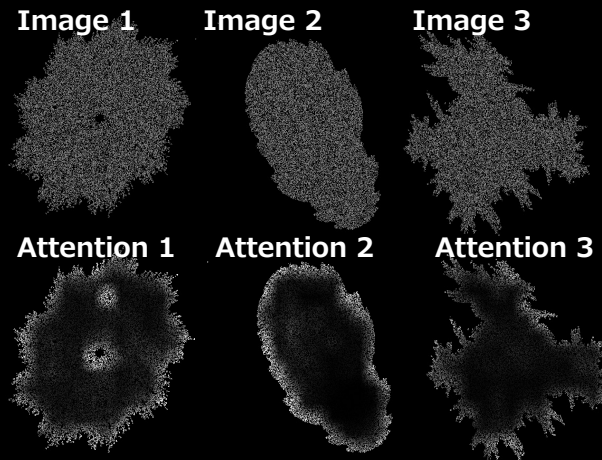
ImageNet-21kを超える事前学習効果
Fractalでラベル付実画像データセットを超えるのは凄いが，放射輪郭画像のみでも置き換え可能！

輪郭識別のみの事前学習ながら，物体検出・インスタンスセグメンテーションはImageNet-1kに近い事前学習効果を発揮

仮説１：FDSLの画像表現では物体輪郭こそが重要

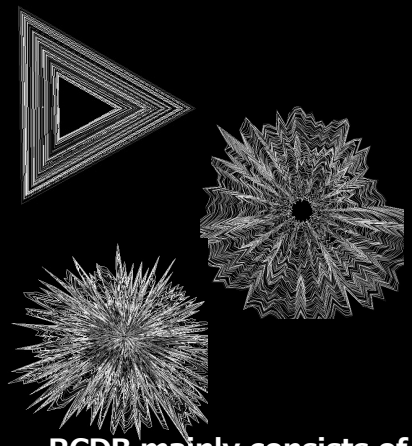
Hypothesis 1:
Object contours are what matter in FDSL datasets

@FractalDB [Kataoka+, ACCV20]



ViT activated on contours of fractal images

@RadialContourDB (RCDB)



RCDB mainly consists of contours

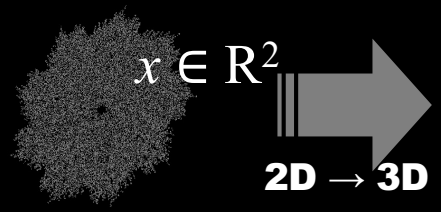
Pre-training	C10	C100	Cars	Flowers
Scratch	78.3	57.7	11.6	77.1
Perlin Noise [21]	95.0	78.4	70.6	96.1
Dead Leaves [3]	95.9	79.6	72.8	96.9
Bezier Curves [21]	96.7	80.3	82.8	98.5
RCDB	**96.8**	**81.6**	84.2	**98.7**
FractalDB [27]	**96.8**	**81.6**	**86.0**	98.3

ハイパラ探索含め，高度なことをせずともRCDBはFractalDBに近い精度に到達

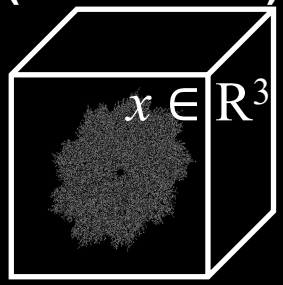
仮説２：FDSL事前学習のタスク難易度が精度向上に寄与

Hypothesis 2:
Task difficulty matters in FDSL pre-training

@FractalDB
[Kataoka+, ACCV20]

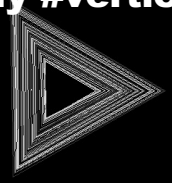


@Extended FractalDB
(ExFractalDB)



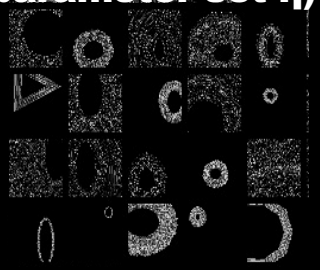
3D Fractalを描画，ランダム視点から2D画像に投影

@RadialContourDB
(only #vertices)



Increase #params

@RadialContourDB
(parameter set η)



頂点数がメインだが，半径長・輪郭数・滑らかさなど
パラメータセットを調整しつつカテゴリ定義

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)

事前学習データセットを生成する数式のパラメータ数増加が追加学習の精度に貢献

Point Cloud Pre-training with Natural 3D Structures

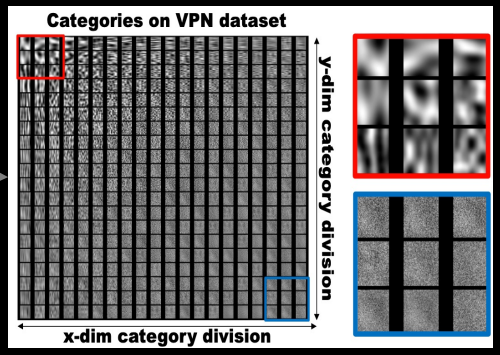
CVPR 2022

Ryosuke Yamada*, Hirokatsu Kataoka*, Naoya Chiba, Yukiyasu Domae*, Testuya Ogata*, ****

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

【FDSL Family】

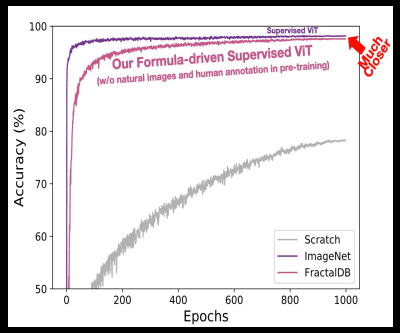
Spatiotemporal Domain



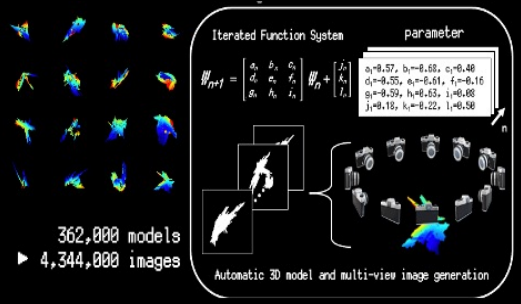
Video Perlin Noise
[Kataoka+, WACV22]

3D Domain

Vision Transformers



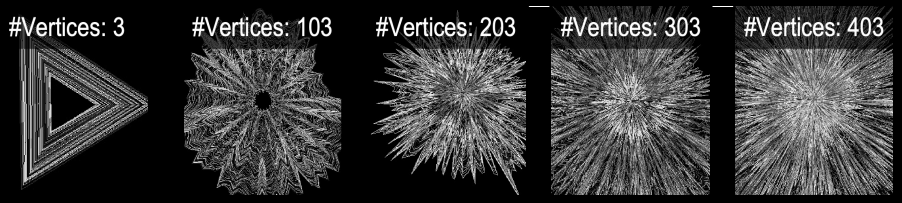
FractalDB Pre-trained ViT
[Nakashima+, AAAI22]



3Dドメインにも適用できないか？

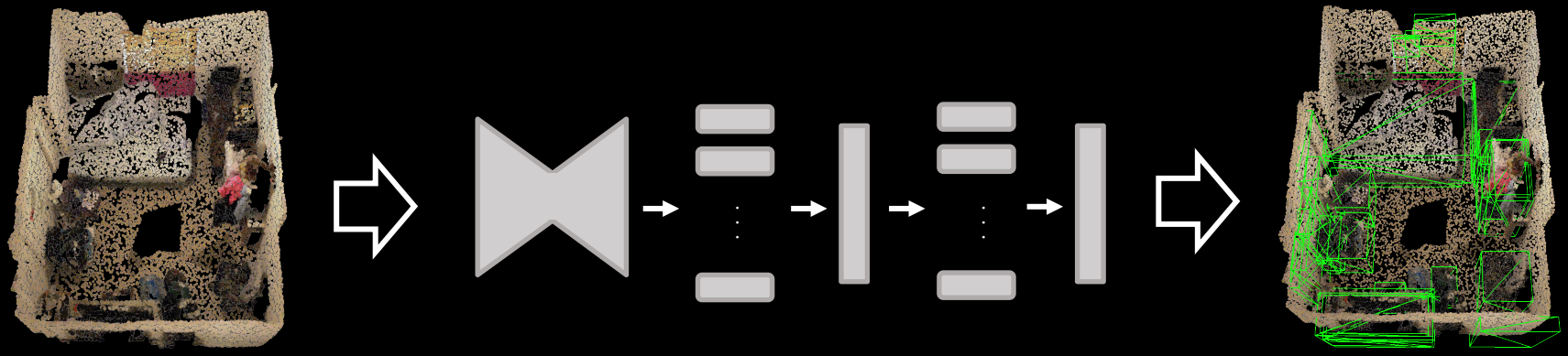
Multi-viewpoint / Point Cloud
[Yamada+, IROS22/CVPR22]

Enhanced by Hypotheses



Replacing Labeled Real-image Datasets
[Kataoka+, CVPR22]

3Dドメインでは決定版の事前学習データセットが不在
データセット構築のコストが高すぎる



数式ドリブン3D点群事前学習

実世界の生成規則を学習することで
従来の3Dデータセットより
汎用的特徴が獲得できるのでは？

Point Cloud Fractal Database: 3Dフラクタルモデル生成

3Dフラクタルモデルをいかに作るか？ → 変換行列を3Dに拡張するのみ

$3D\ IFS = \{(w_j, p_j)\}_{j=1}^{N}$ w_j: Affine Transformation, p_j: Selection probability

1. 3D-IFS parameters setting

$$w_1 = \begin{bmatrix} 0.57 & -0.68 & 0{,}40 \\ -0.55 & -0.61 & -0.16 \\ -0.59 & 0.63 & 0.08 \end{bmatrix} + \begin{bmatrix} 0.18 \\ -0.22 \\ 0.50 \end{bmatrix}$$

N

2. Affine transformation

$$\boldsymbol{x}_i = w_j\, \boldsymbol{x}_{i-1}$$

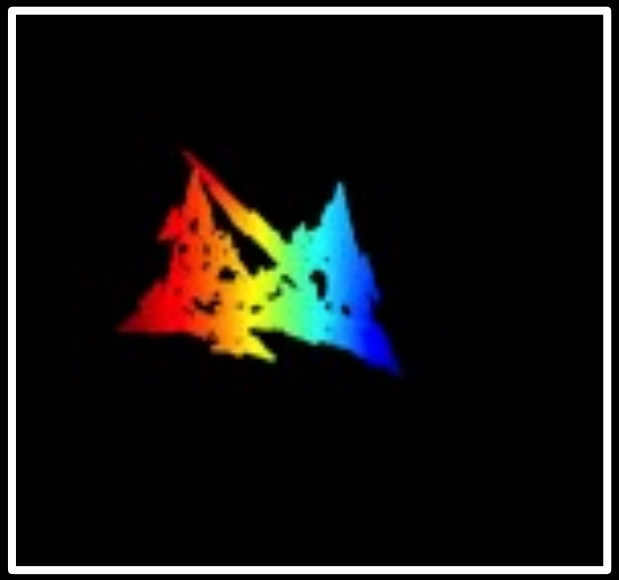
$(i = 1, 2, 3, .., n)$

$\boldsymbol{x} = [x, y, z]^T$

3D fractal model: $P = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

3. Variance check & Fractal category definition

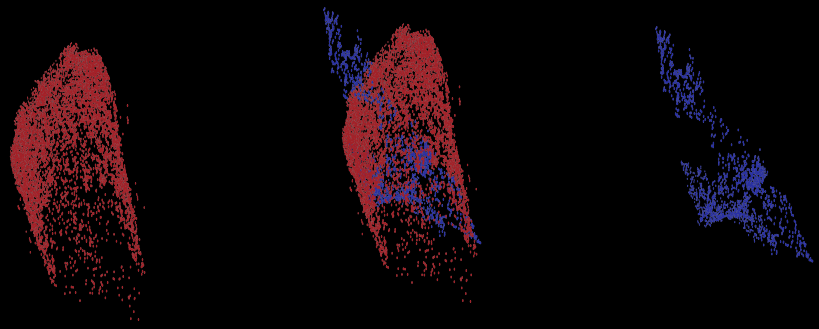
$$min(Var[x], Var[y], Var[z]) = \mathbf{0.17\ \ldots} > 0.15$$



Point Cloud Fractal Database: インスタンス拡張方法

インスタンス拡張 / 3Dシーン構築

インスタンスはMix

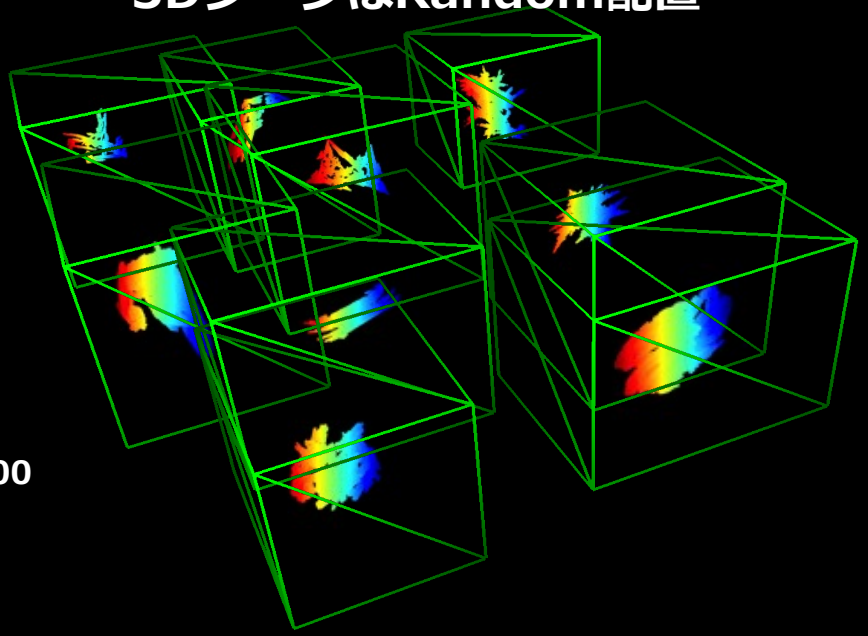


Main Category
Point number: 3,200

FractalNoiseMix
Point number: 4,000

Noise Category
Point number: 800

3DシーンはRandom配置



実験結果: 3D物体検出精度比較

ScanNetV2 / SUN RGB-Dによる比較

Pre-training	Backbone	Parameter	Input	ScanNetV2		SUN RGB-D	
				mAP@0.25	mAP@0.50	mAP@0.25	mAP@0.50
Scratch	PointNet++	0.95M	Geo + Height	57.9	32.1	57.4	32.8
Scratch	SR-UNet	38.2M	Geo	57.0	35.8	56.1	34.2
RandomRooms [51]	PointNet++	0.95M	Geo + Height	61.3	36.2	59.2	35.4
PointContrast [67]	SR-UNet	38.2M	Geo	59.2	38.0	57.5	34.8
CSC [26]	SR-UNet	38.2M	Geo	-	**39.3**	-	**36.4**
PC-FractalDB	PointNet++	0.95M	Geo + Height	**61.9**	38.3	**59.4**	33.9
PC-FractalDB	PointNet++ ×2	38.2M	Geo + Height	**63.4**	**39.9**	**60.2**	35.2
PC-FractalDB	SR-UNet	38.2M	Geo	59.4	37.0	57.1	**35.9**

Underlined bold: best score　**Baseline**　**Ours**

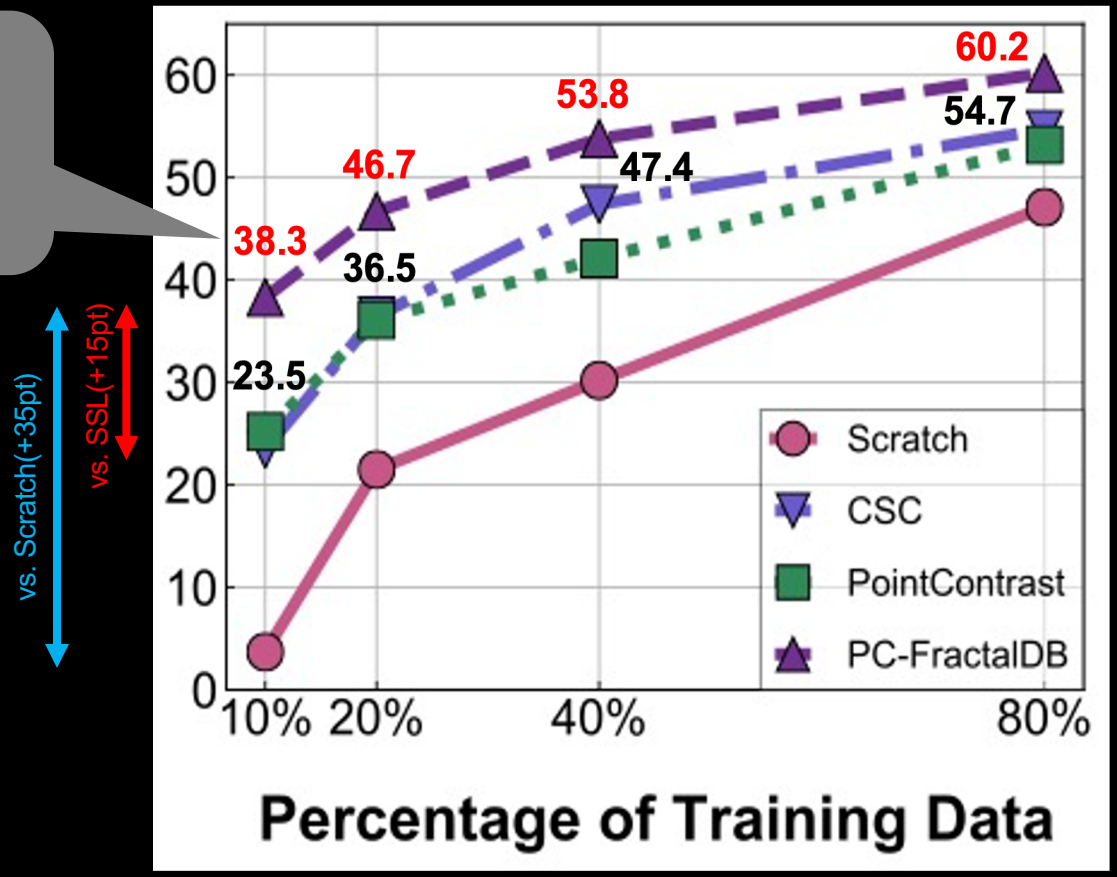
PC-FractalDB 61.9 vs 59.2 (PointContrast; ECCV 2020)
vs 61.3 (RandomRoom; ICCV 2021)

ScanNetV2 / mAP @ 0.25 により計測

実験結果: 少量データ&アノテーションにおける実験

限られたデータの学習においても高精度

データ量10%使用時：
SSL比較で約+15%
Scratch比較約+35%



少量の学習データに対する実験結果（mAP@0.25）

Future direction (1/3)

より良い事前学習データセットへ

- FractalDB事前学習モデルは部分的にImageNet/Places事前学習モデルの精度を超えた
- 80M Tiny Images / ImageNet（人間関連ラベル）が公開停止
- 自然画像なしで良好な視覚特徴を獲得できたことは有意義

Future direction (2/3)

教師あり学習とは異なる視覚特徴を獲得

- FractalDB事前学習モデルはユニークな視覚特徴
- データに合わせて操作可能な事前学習という可能性
- 柔軟なデータ構成：物体検出・領域分割など

Future direction (3/3)

フラクタル幾何(のみ)が良い特徴表現か？

- より良い画像パターン/教師の生成エンジンを模索中
- 画像と教師のペアさえ用意できれば任意の関数で良いので，今後の研究によってはフラクタル幾何よりも良い法則が見つかる可能性大

世界の動向

@MIT A. Torralba Lab

Learning to See by Looking at Noise

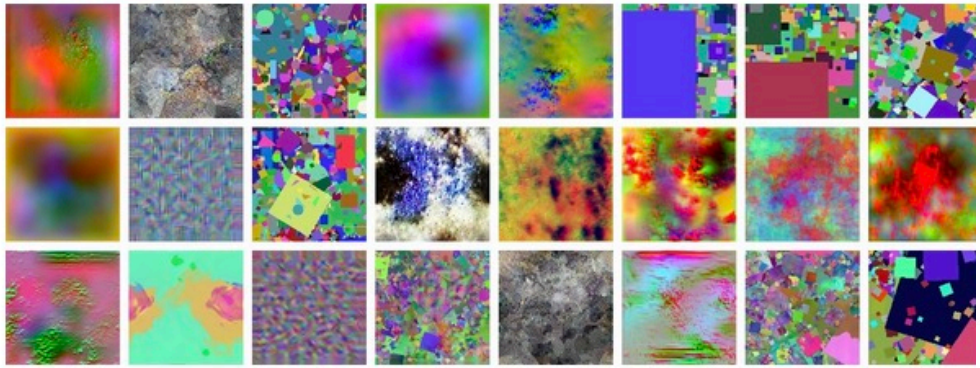
Manel Baradad*
MIT CSAIL

Jonas Wulff*
MIT CSAIL

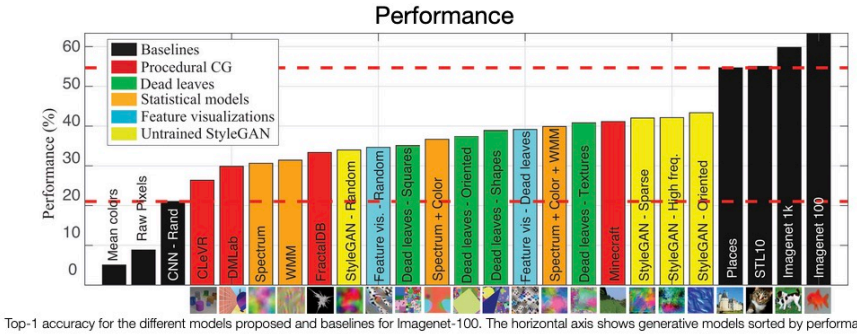
Tongzhou Wang
MIT CSAIL

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.

自然画像を用いない事前学習は分野全体で解決したい

Hirokatsu Kataoka | 片岡裕雄
@HirokatuKataoka

我々の研究により「画像認識AIの事前学習（Pre-training）データを人間が集める時代は終わった」と言えるような世界にしていきたいですね！さらに、AI開発を加速させるべく、今回我々からは商用利用を制限する権利を設けておりません。

https://twitter.com/HirokatuKataoka/status/1536284511696490498