少量教師データ下の深層学習について

実践データ駆動科学オンラインセミナー
第10回「データ科学・AIにおける数理の威力」

岡谷 貴之
東北大学 情報科学研究科
(理化学研究所AIPセンター)

自己紹介

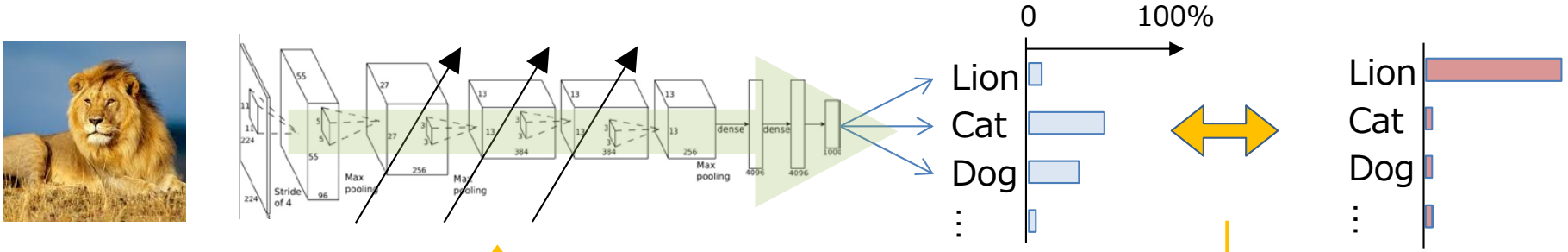
- 略歴
  - 1999年東京大学計数工学専攻博士課程修了
  - 1999年～東北大学大学院情報科学研究科，2013年から現職
  - 2016年～理化学研究所AIPセンター兼務
- 研究対象
  - コンピュータビジョン＋深層学習
- 研究の方向性
  - 基礎：新規技術＋学生教育
  - 応用：企業との連携（共同研究＋コンサルティング）

深層学習～教師あり学習

- 深層ニューラルネットワークを大量のデータを用いて訓練

数十層からなる数千万オーダの
パラメータを持つネットワーク

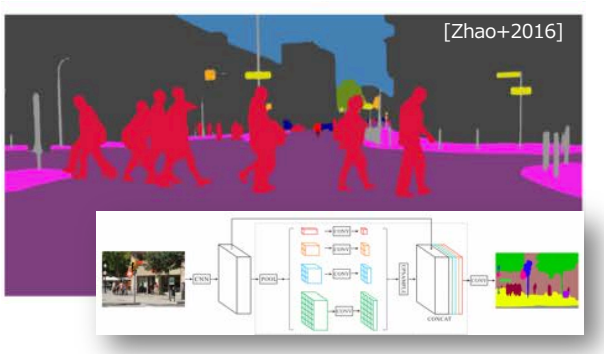
数万～数百万
の学習データ



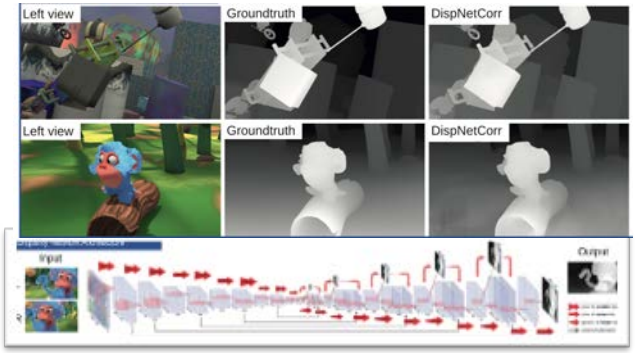
深層学習～教師あり学習

- 多数の問題に同じアプローチ → 大成功

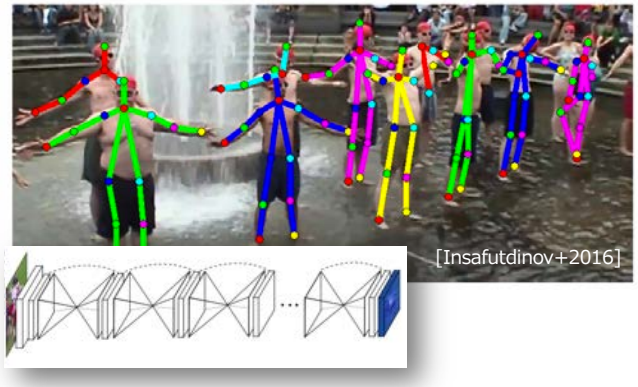
画素レベルの認識



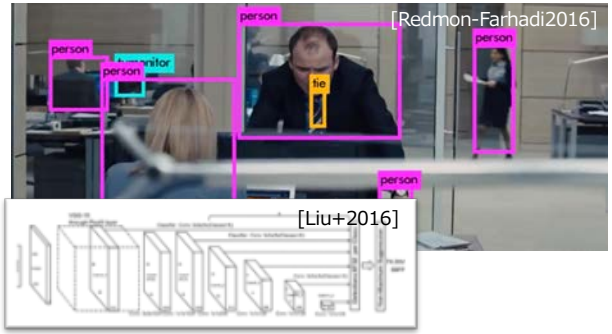
視差（ステレオ画像） [Mayer+2017]



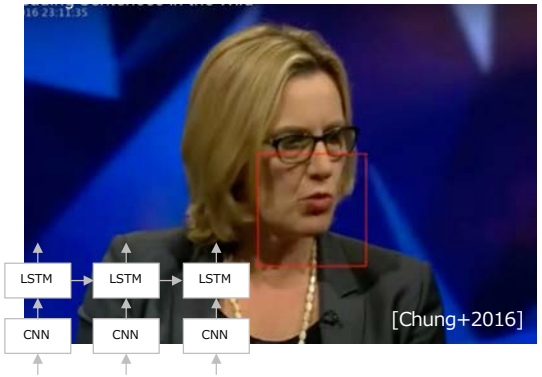
人体ポーズ



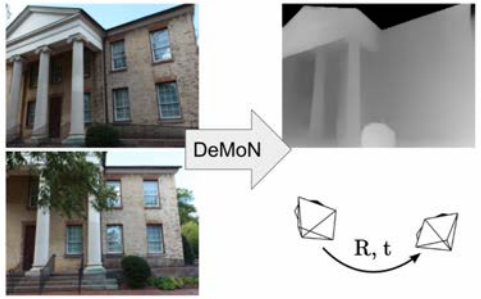
物体検出



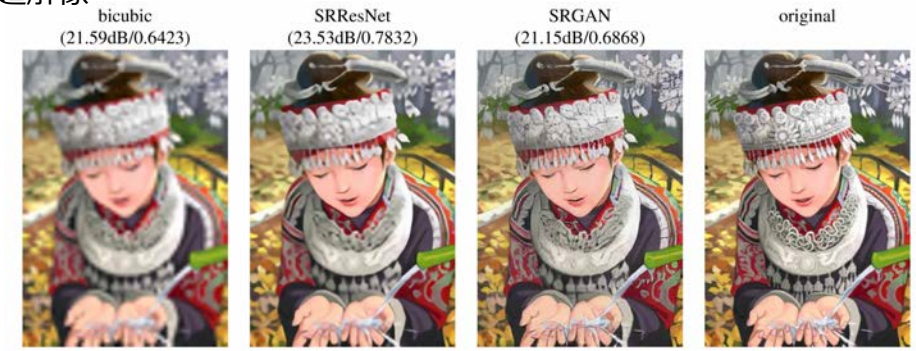
読唇

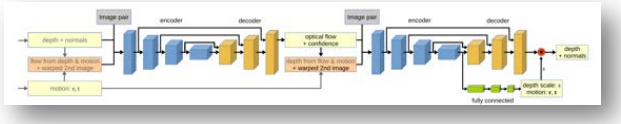


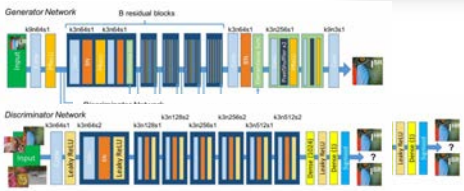
カメラ姿勢 [Ummenhofer+2017]



超解像





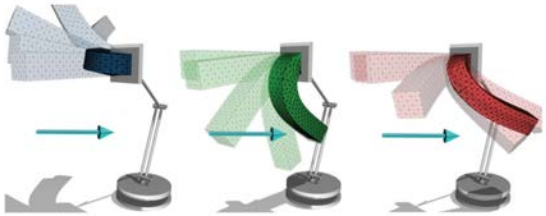


深層学習はAIの枠を超えた道具に

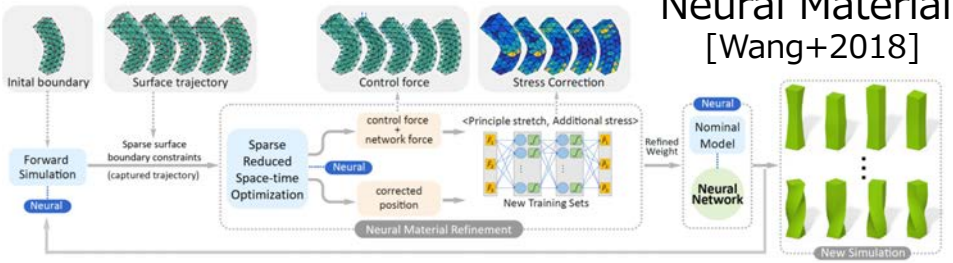
- あらゆる工学・サイエンスに浸透開始 → 解けなかった問題を解く
- 計算時間短縮の目的でも利用が始まる → 順伝播計算１回で解ける

構造体の変形計算

DeepWarp
[Luo+2018]

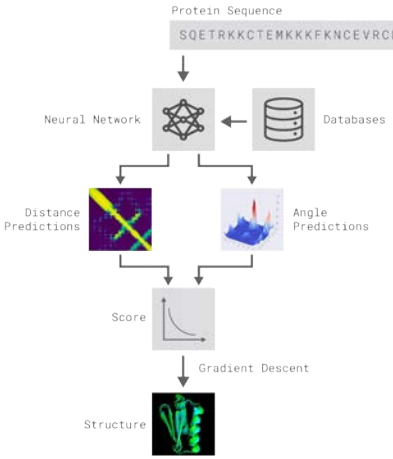


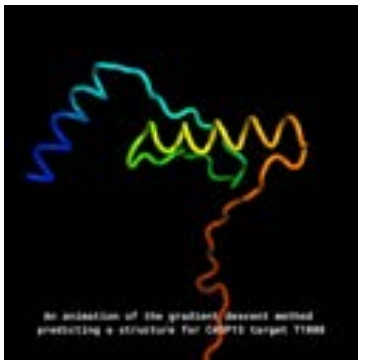
Neural Material
[Wang+2018]



タンパク質構造予測

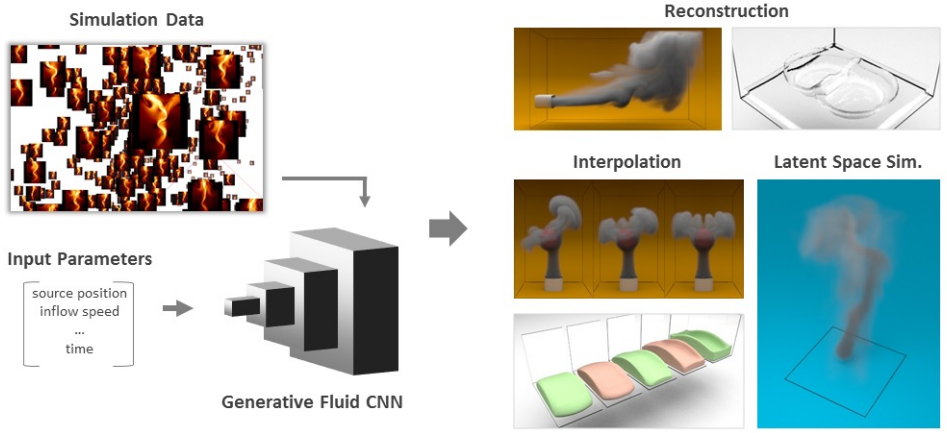
AlphaFold [DeepMind2018]





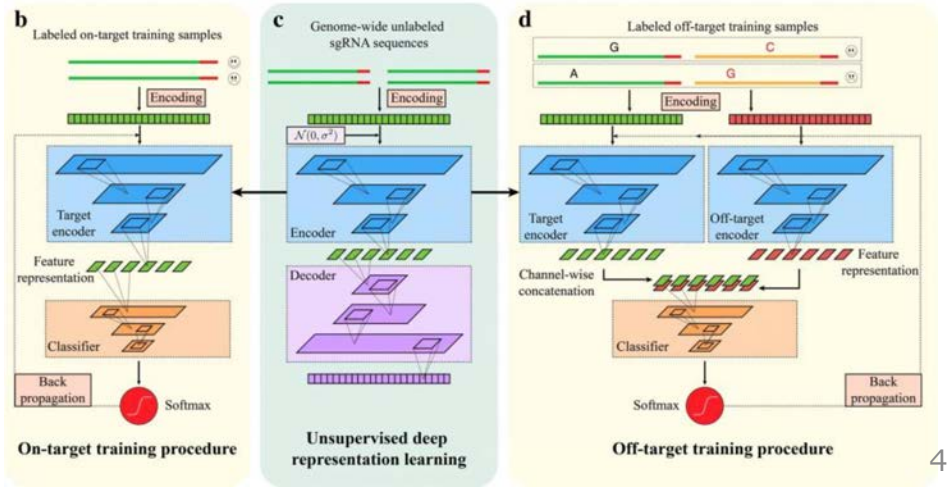
流体シミュレーション

Deep Fluids [Kim+2018]



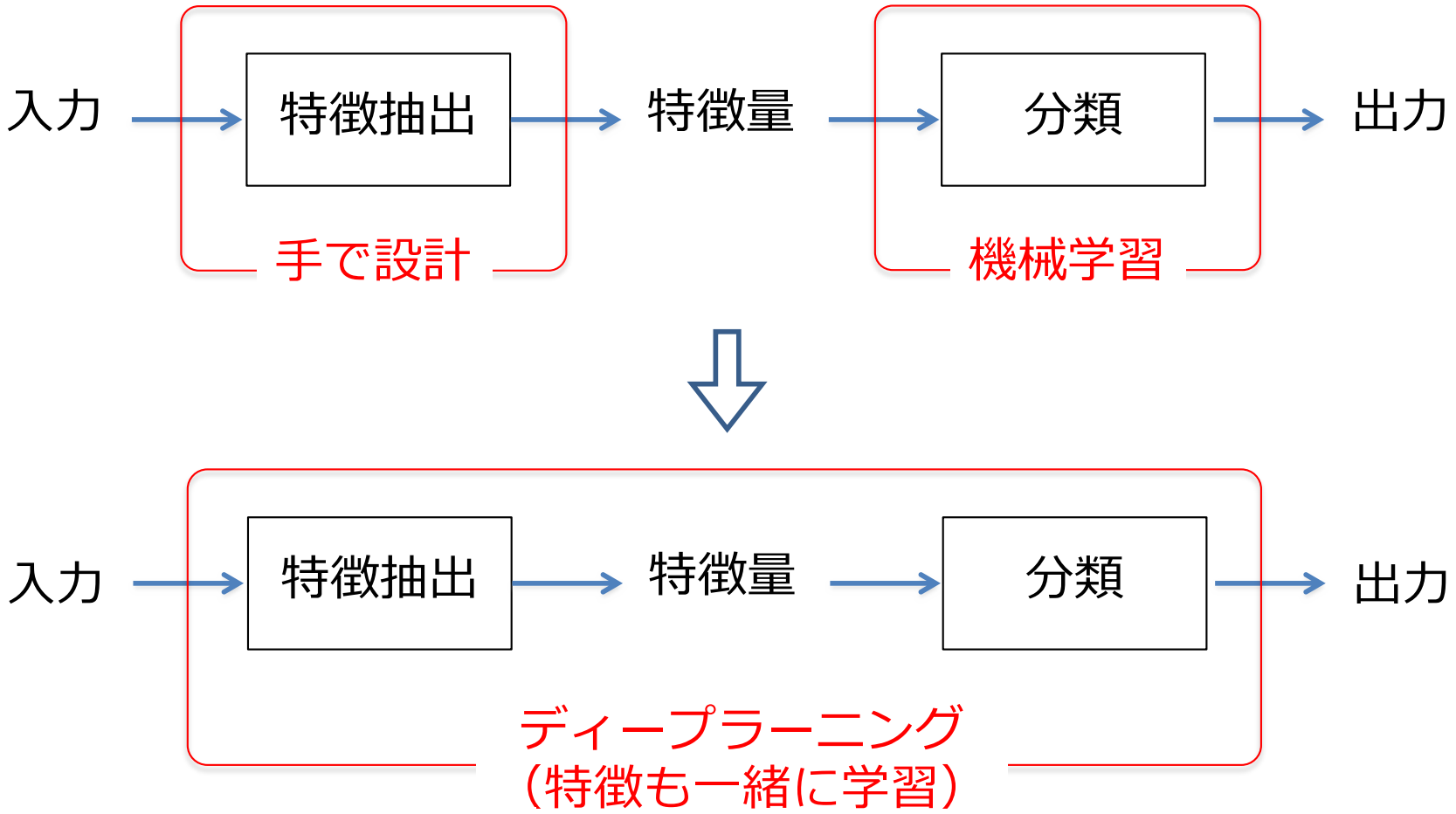
遺伝子操作応用

DeepCRISPR [Chuai+2018]



深層学習は何が違うか

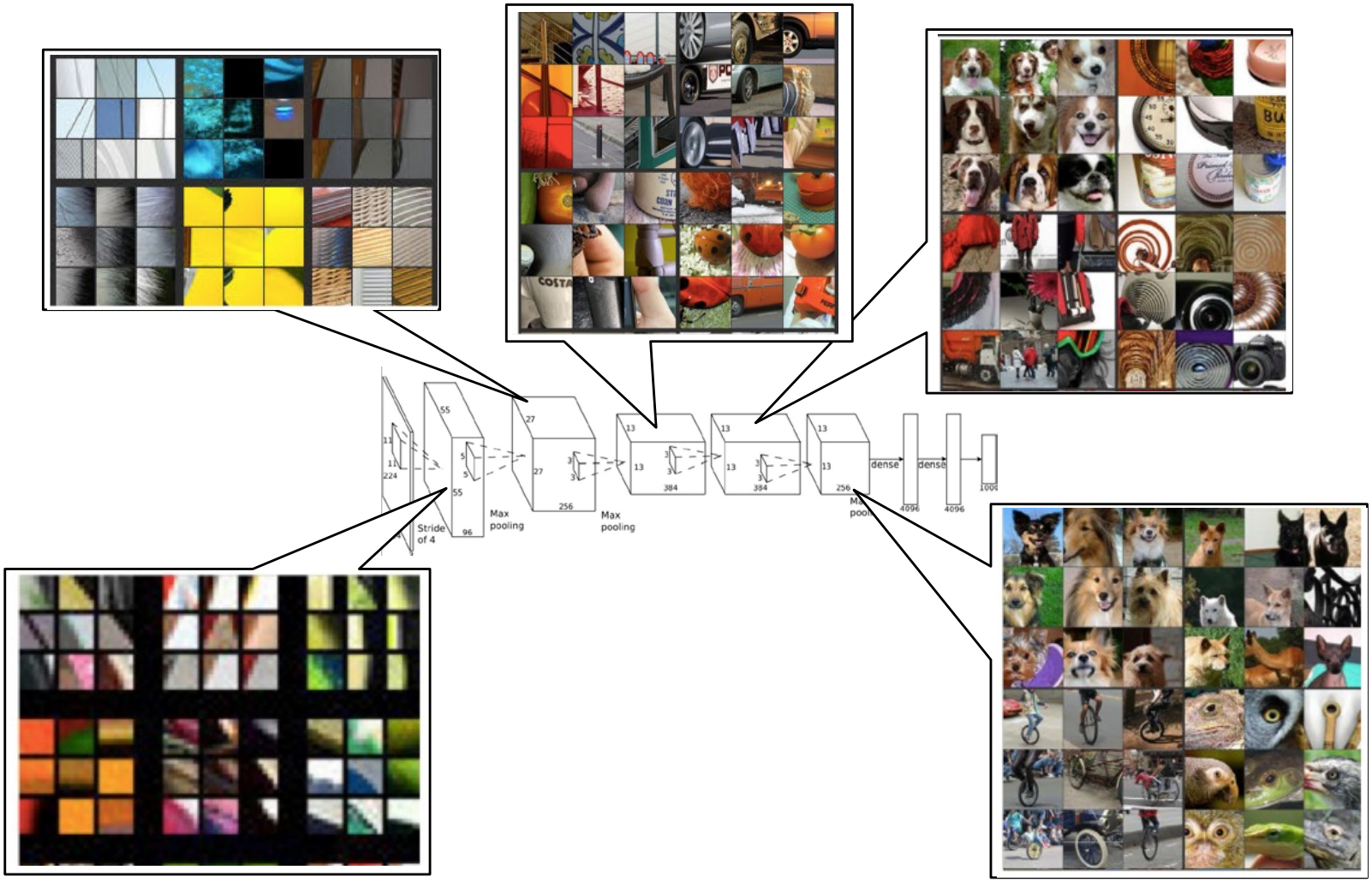
- 手で設計していた特徴抽出を学習できる



学習で獲得する特徴表現

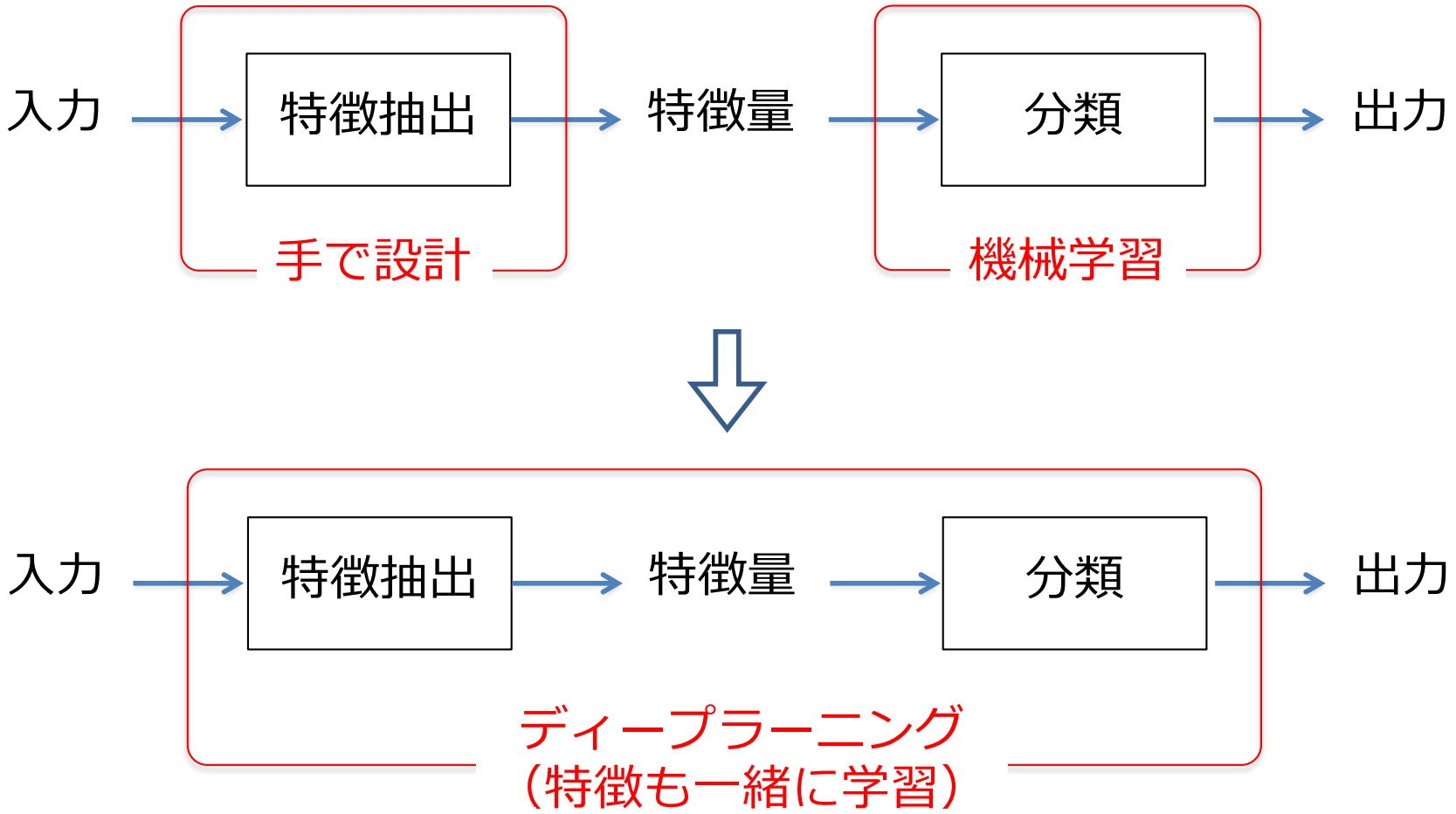
Zeiler-Fergus, Visualizing and understanding convolutional networks, arxiv2014

- 各層各ユニットを最も活性化する入力画像



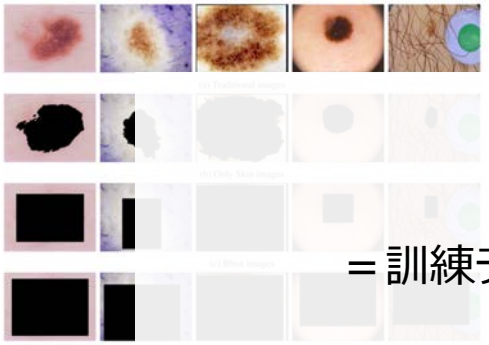
深層学習は何が違うか

- 手で設計していた特徴抽出を学習できる



このアプローチの限界＝訓練データへの依存

不明な画像特徴



[Bissoto+2019]

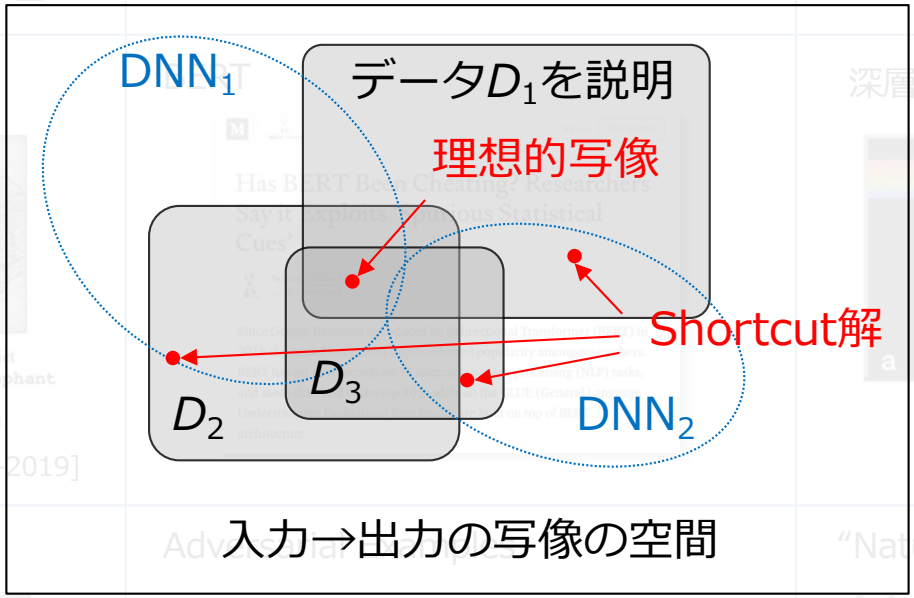
ドメインシフト

見慣れない姿勢の物体

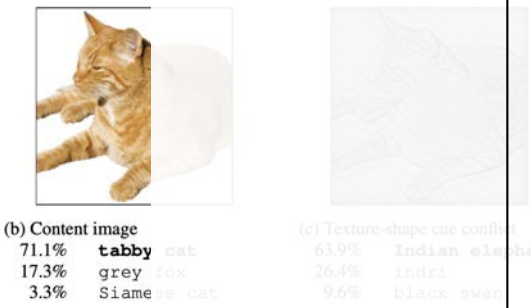
[Alcorn+2018]

ショートカット学習

＝訓練データ固有の「偽りの統計的性質」をてっとり早く学習



形状よりテクスチャ優位



[Geirhos+2019]

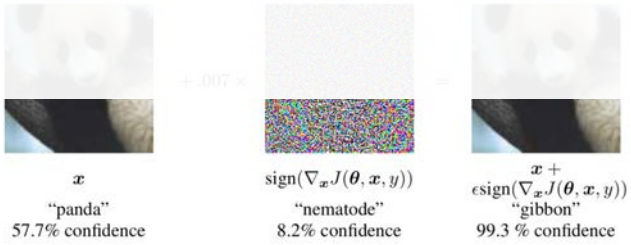
深層強化学習×ビデオゲーム

[Kansky+2017]

Fooling samples

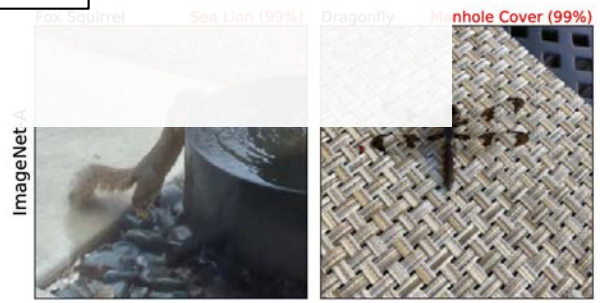
[Nguyen+2014]

Adversarial Examples



[Goodfellow+2015]

"Natural Adversarial Examples"



[Hendrycks+2020]

今あるソリューション = データの大規模化

IMAGENET http://image-net.org/index
- 一般物のクラス認識
- 21841クラス・14,197,122枚
- スタンフォード大・プリンストン大他

CITYSCAPES https://www.cityscapes-dataset.com/examples/
- 車載カメラの画素単位の物体クラス
- 30クラス・50都市・５,000枚/20,000枚
- ダイムラー・ダルムシュタット工科大他

COCO http://cocodataset.org
- 物体クラスとその画像領域
- 80物体+91物体以外クラス・330,000枚

CelebA http://mmlab.ie.cuhk.edu
- 顔画像
- 40属性・10,177人物・2
- 香港中文大学

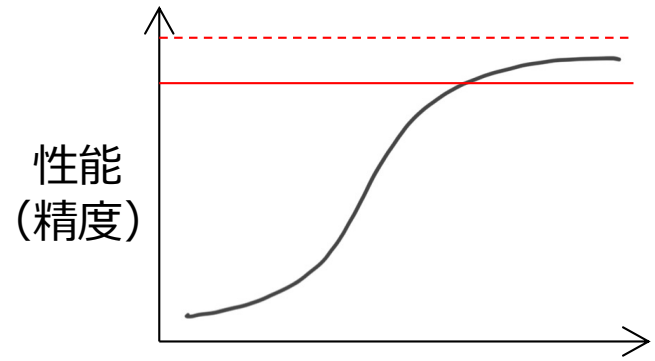
MPII Huma
- 人体ポーズ
- 前身関節位置
- Max Planck I

- 画像中のシーンに対する質問と答え
- 画像当たり平均5.4質問・10回答・265,016枚
- Virginia Tech., Georgia Tech.

Q: Where is the cat?
Q: Is the cat safe?
Q: What color is the car?

Q: What toy does the elephant have in his cage?
Q: Is this animal in an enclosure?
Q: Where is the woman?

性能-コストの「S-カーブ」



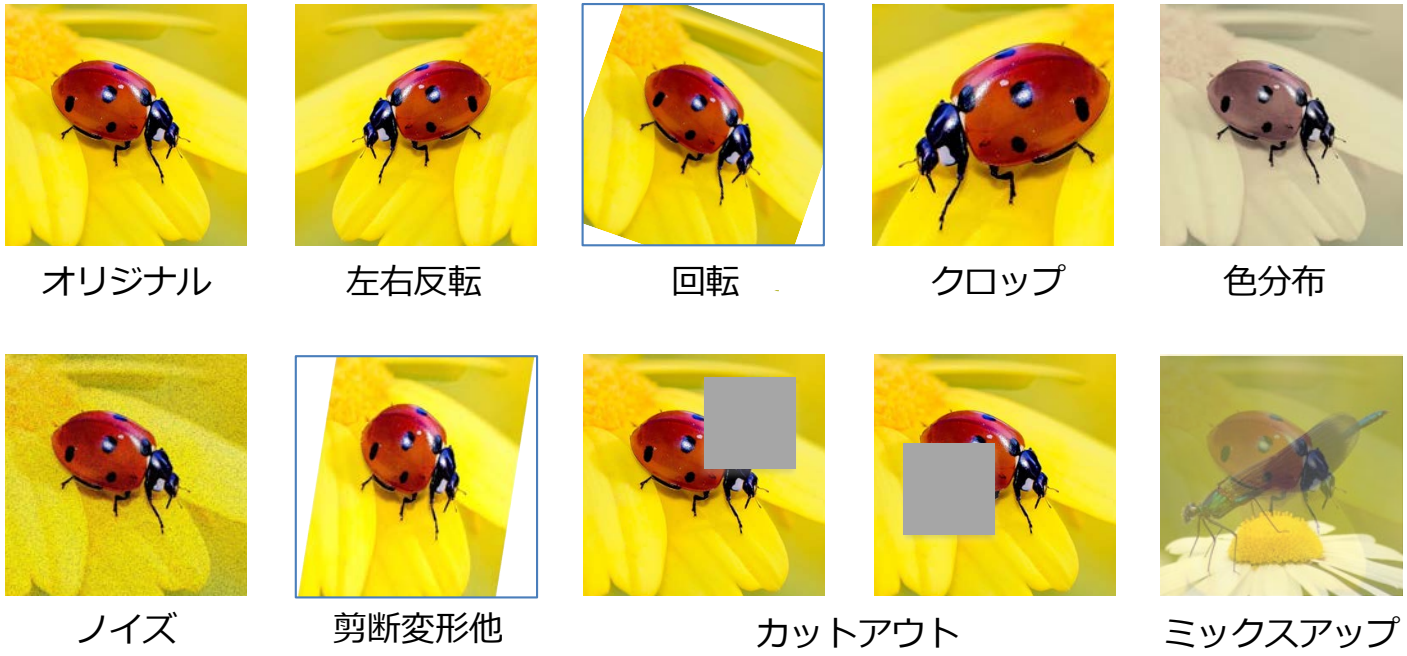
データが少ない場合の対策

- 定番の方法～研究段階の方法まで
  - データ拡張
  - 転移学習
  - ドメイン適応（domain adaptation）
  - 少数事例学習（few-shot learning）
  - 半教師学習（semi-supervised），弱教師学習（weakly-supervised）

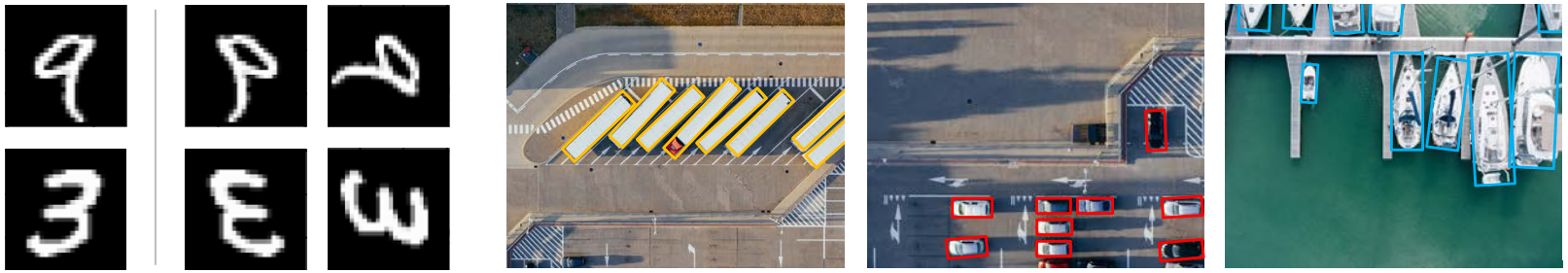
- 今後の方向
  - 自己教師学習を基礎とする巨大なDNNモデル

データ拡張

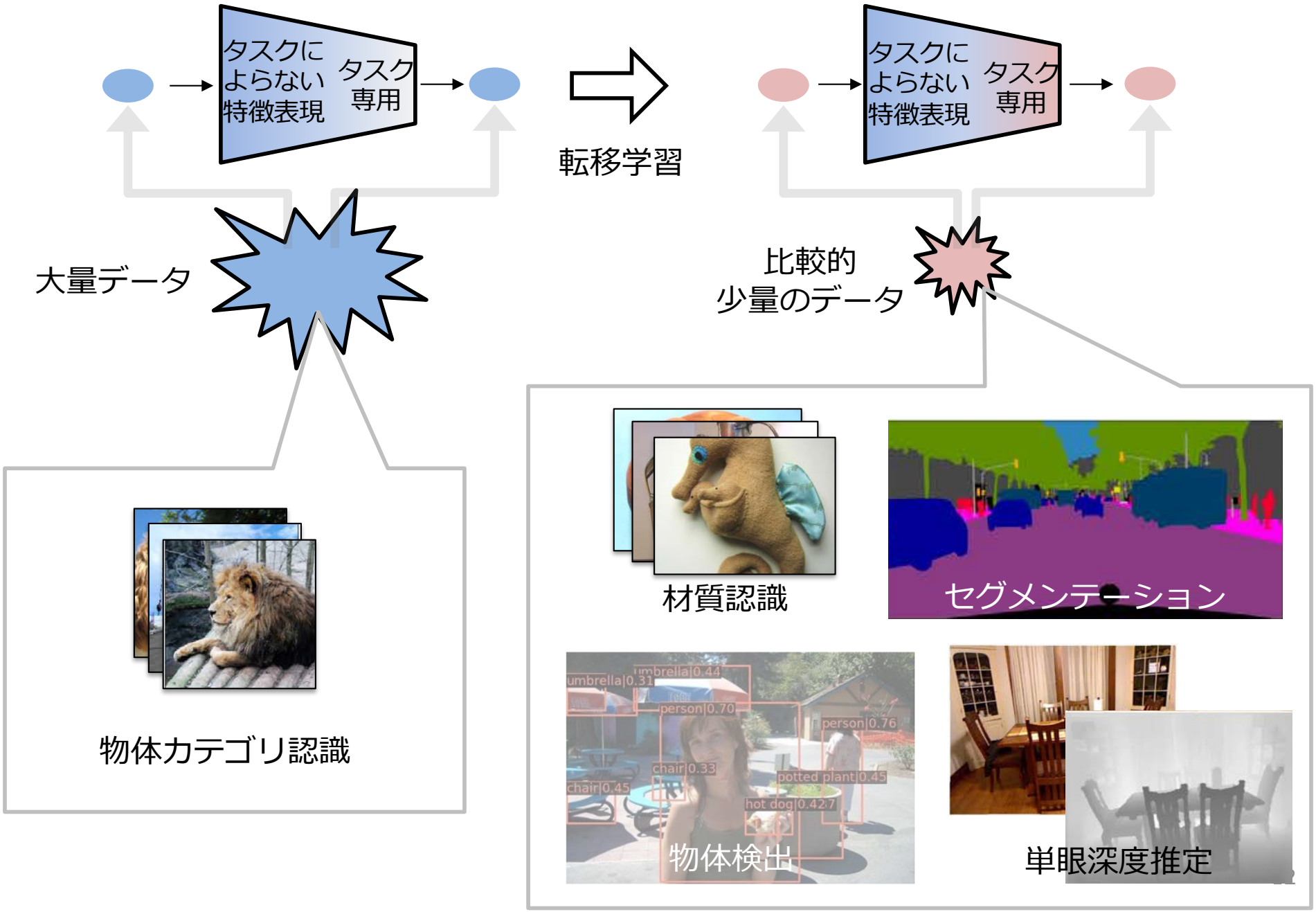
- コンテンツを変えない変換を適用，サンプルを水増し



- タスク・データの性質に応じて変換を考える



転移学習


タスクに
よらない
特徴表現
タスク
専用
転移学習
タスクに
よらない
特徴表現
タスク
専用
大量データ
比較的
少量のデータ
物体カテゴリ認識
材質認識
セグメンテーション
umbrella 0.31
umbrella 0.44
person 0.70
person 0.76
chair 0.33
chair 0.45
potted plant 0.45
hot dog 0.427
物体検出
単眼深度推定

データが少ない場合の対策

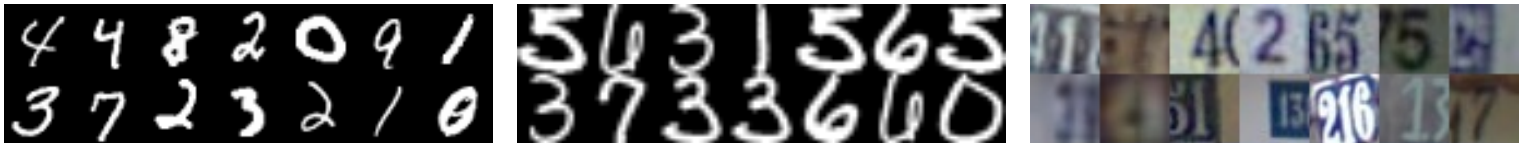
- 定番の方法～研究段階の方法まで
  - データ拡張
  - 転移学習
  - ドメイン適応（domain adaptation）
  - 少数事例学習（few-shot learning）
  - 半教師学習（semi-supervised），弱教師学習（weakly-supervised）

- 今後の方向
  - 自己教師学習を基礎とする巨大なDNNモデル

ドメインの違うデータ

MNIST　　USPS　　SVHN

数字



車載カメラ映像（セグメンテーション）

Cityscapes
（実写）

GTA5
（ゲームCG）

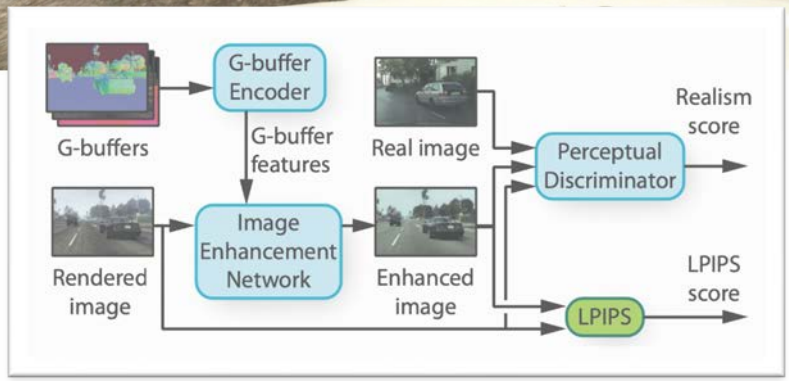
商品画像

合成画像の実写化

Richter+, Enhancing photorealism enhancement, arXiv2021

Rendered images from GTA V

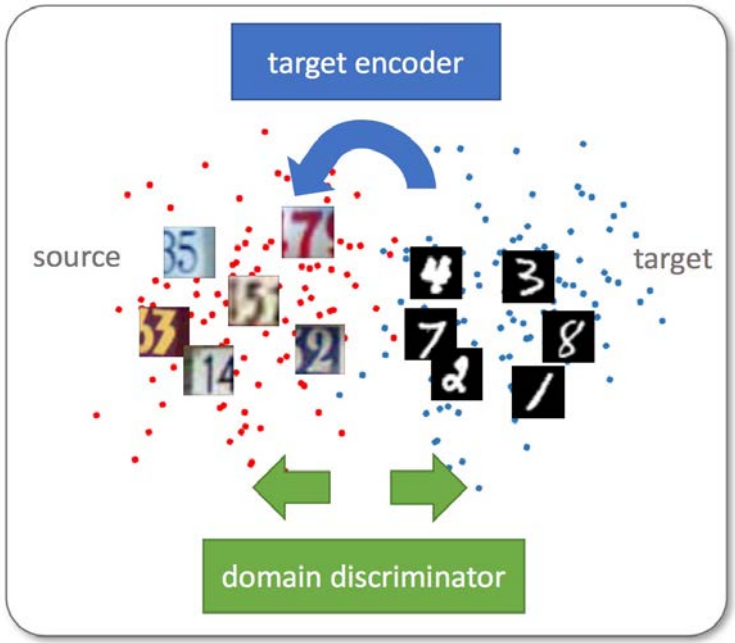
Enhancement by our method (trained to mimic Cityscapes)

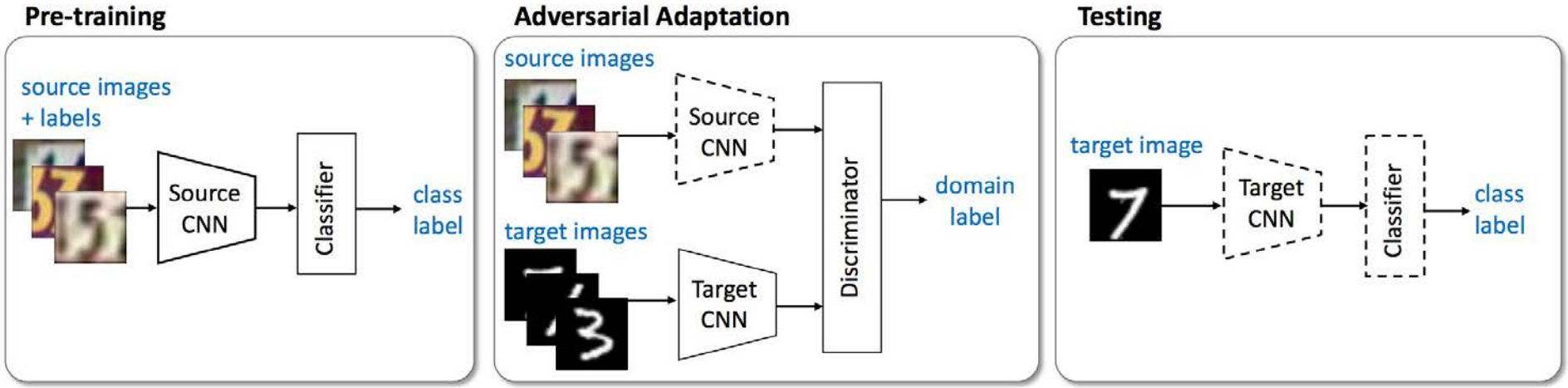


無教師ドメイン適応(UDA)

Tzeng+, Adversarial Discriminative Domain Adaptation, CVPR2017

- ソースドメインからターゲットドメインへ適応
  - ソースは正解ラベル付・ターゲットはラベルなし(画像のみ)
- 定番：敵対的学習
  - ソースとターゲットの特徴が同じように見えるような特徴空間を得る
  - ソースの分類器をターゲットの特徴空間に適用し，ターゲットの画像を分類

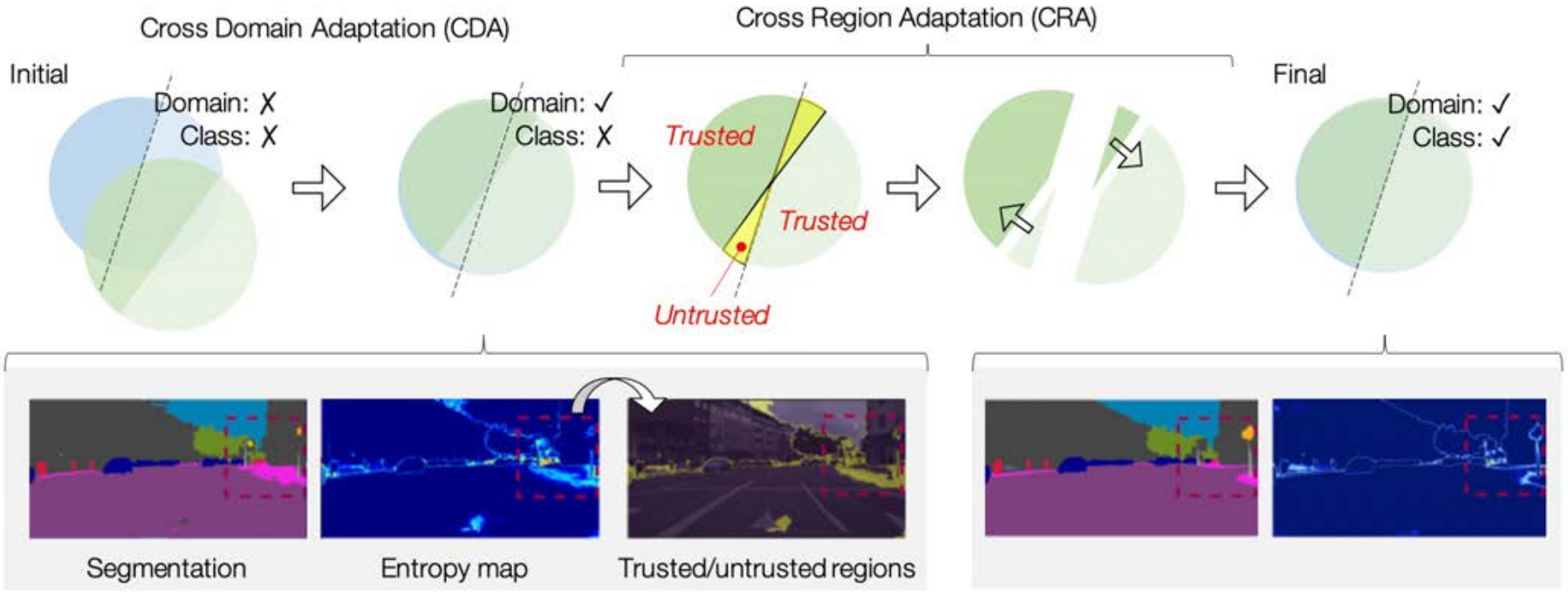




セグメンテーションでのUDA最新手法

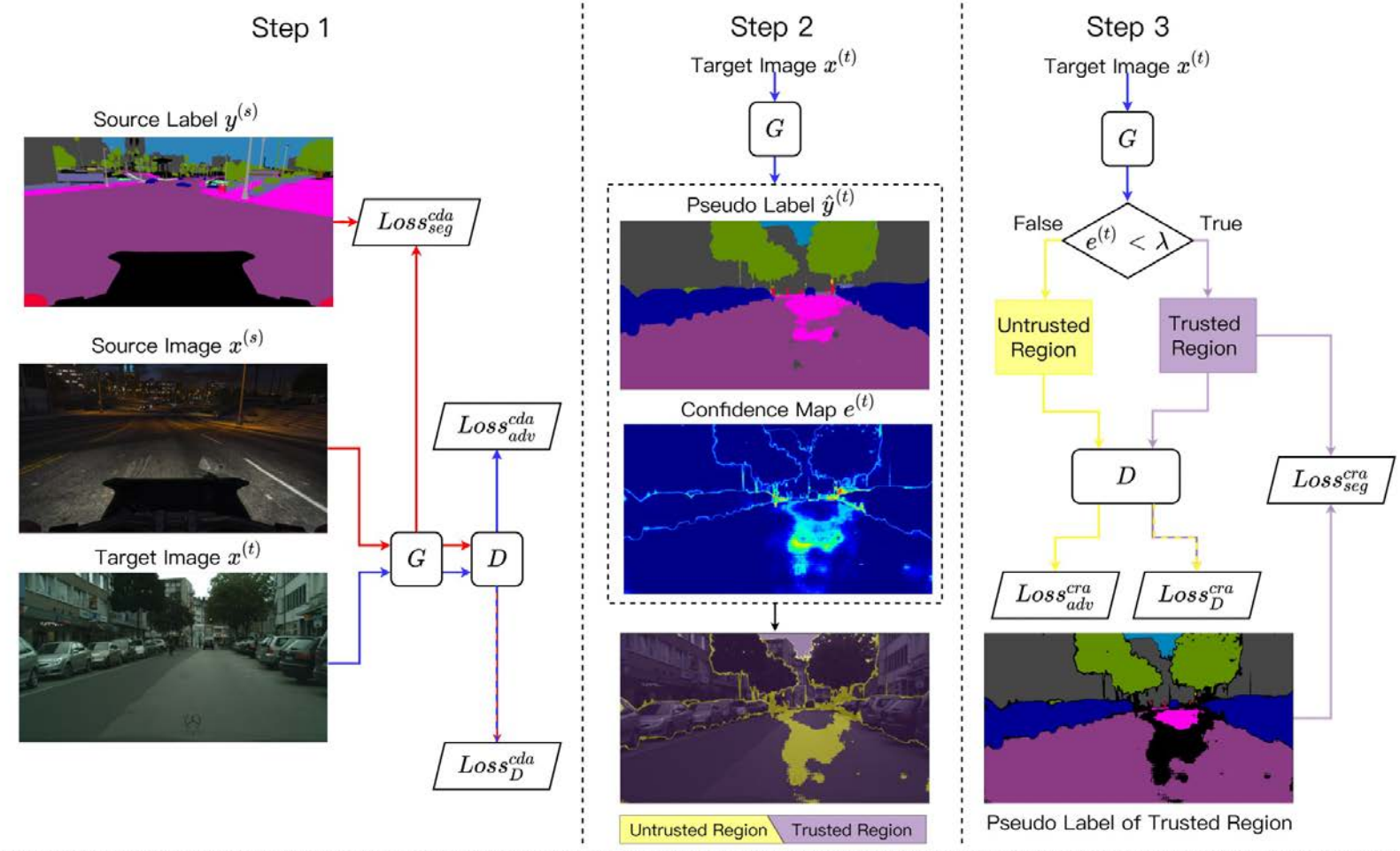
Wang, Liu, Suganuma, Okatani, Cross-Region Domain Adaptation for Class-level Alignment, arXiv 2021

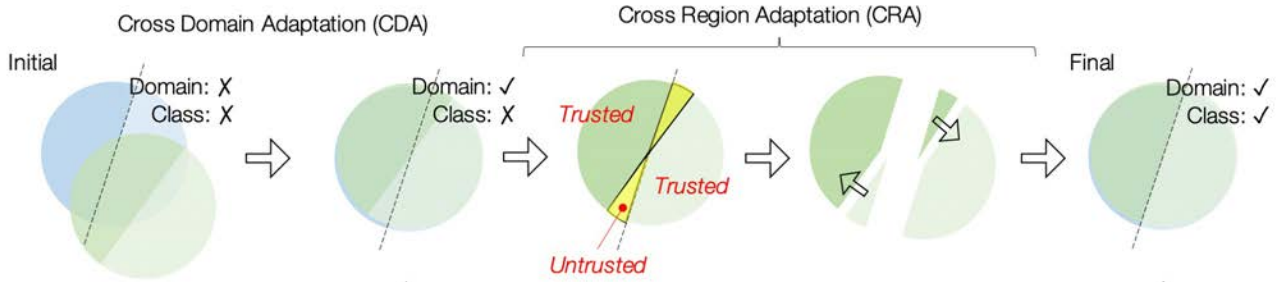
- CG画像(GTA) → 実写画像(Cityscapes)



セグメンテーションでのUDA最新手法

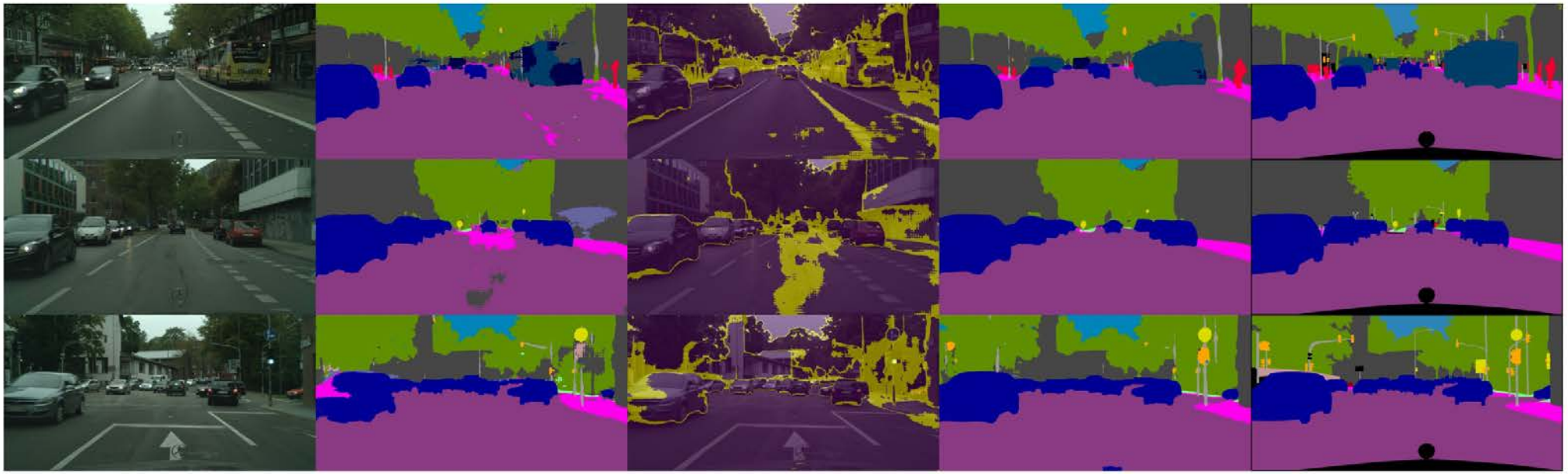
Wang, Liu, Suganuma, Okatani, Cross-Region Domain Adaptation for Class-level Alignment, arXiv 2021





セグメンテーションでのUDA最新手法

Wang, Liu, Suganuma, Okatani, Cross-Region Domain Adaptation for Class-level Alignment, arXiv 2021



画像　　従来手法 (FADA)　　不確実性　　提案手法　　GT

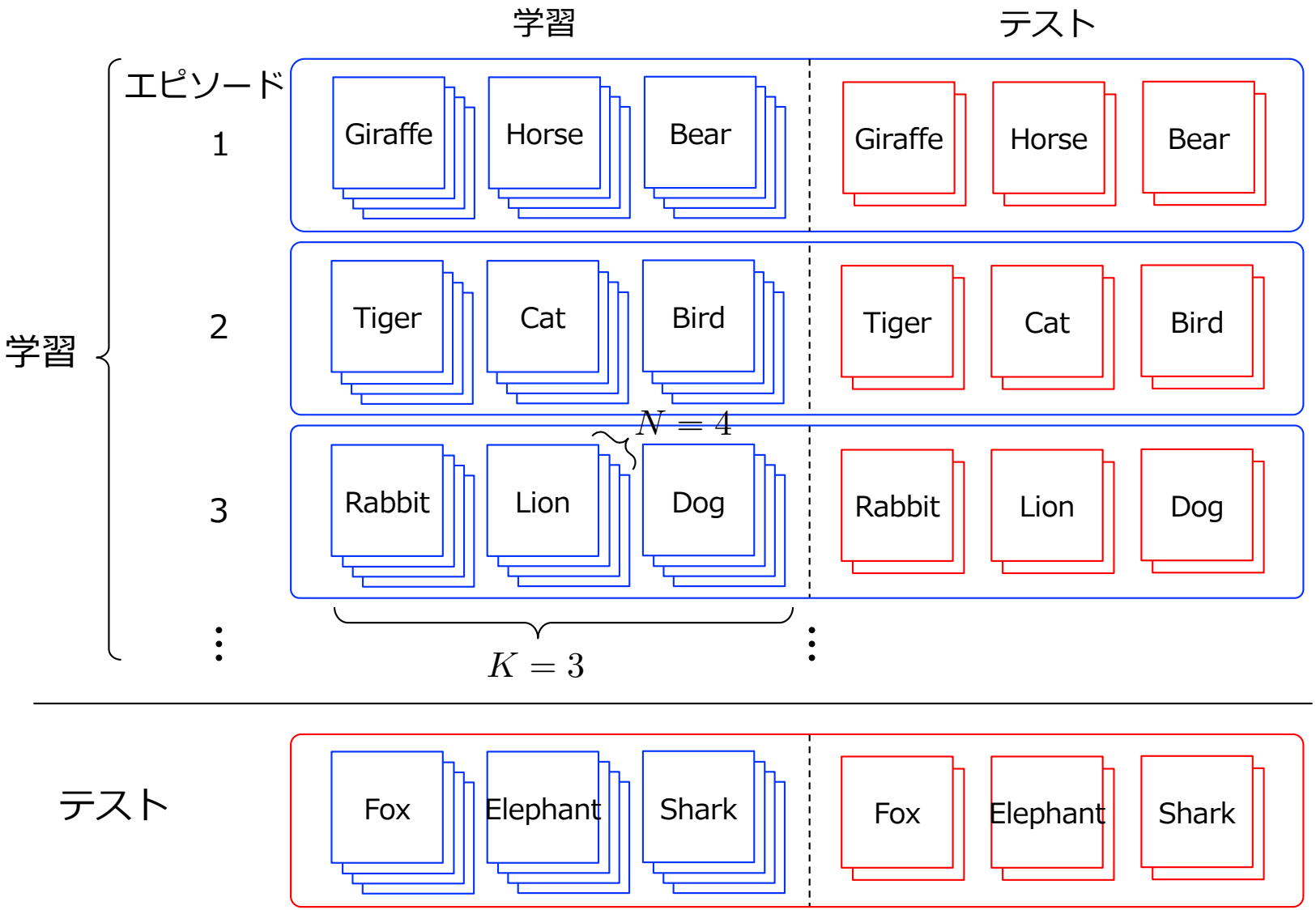
Method	Base CDA	+CRA	Δ
AdaptSegNet (Tsai et al. 2018)	42.4	43.4	+1.0
ADVENT (Vu et al. 2019a)	43.8	46.7	+2.9
FADA (Wang et al. 2020)	50.1	52.2	+2.1
IAST (Mei et al. 2020)	52.2	54.1	+1.9
ProDA (Zhang et al. 2021)	57.5	58.6	+1.1

データが少ない場合の対策

- 定番の方法～研究段階の方法まで
  - データ拡張
  - 転移学習
  - ドメイン適応（domain adaptation）
  - 少数事例学習（few-shot learning）
  - 半教師学習（semi-supervised），弱教師学習（weakly-supervised）

- 今後の方向
  - 自己教師学習を基礎とする巨大なDNNモデル

少数事例（few-shot）学習



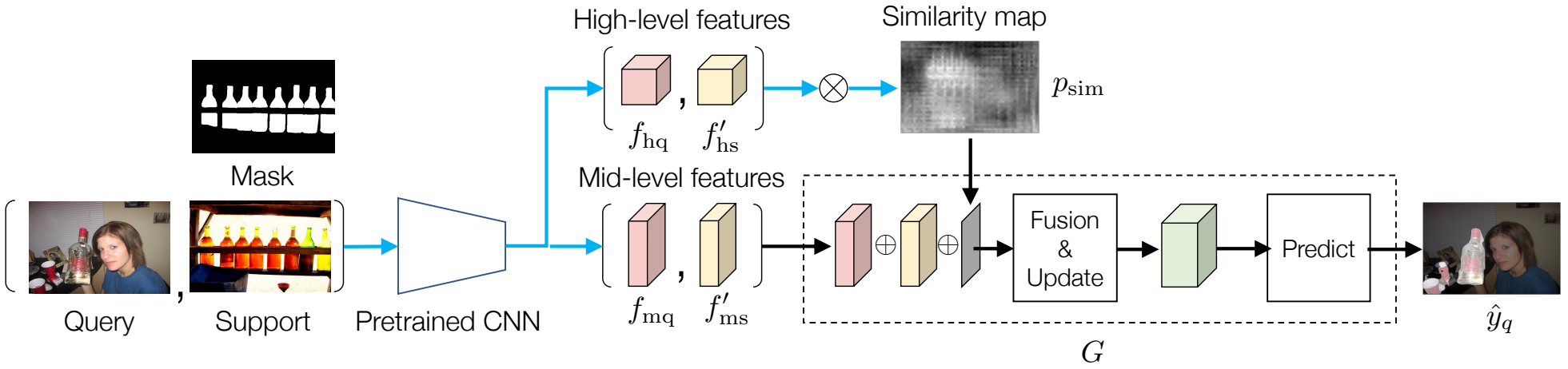
"K-way N-shots”

少数事例からのセグメンテーション

Wang, Suganuma, Okatani, Improved Few-shot Segmentation by Redefinition of the Roles of Multi-level CNN Features, arXiv 2021

- 最少で１事例のみで対象物を指定し，セグメンテーションを実行

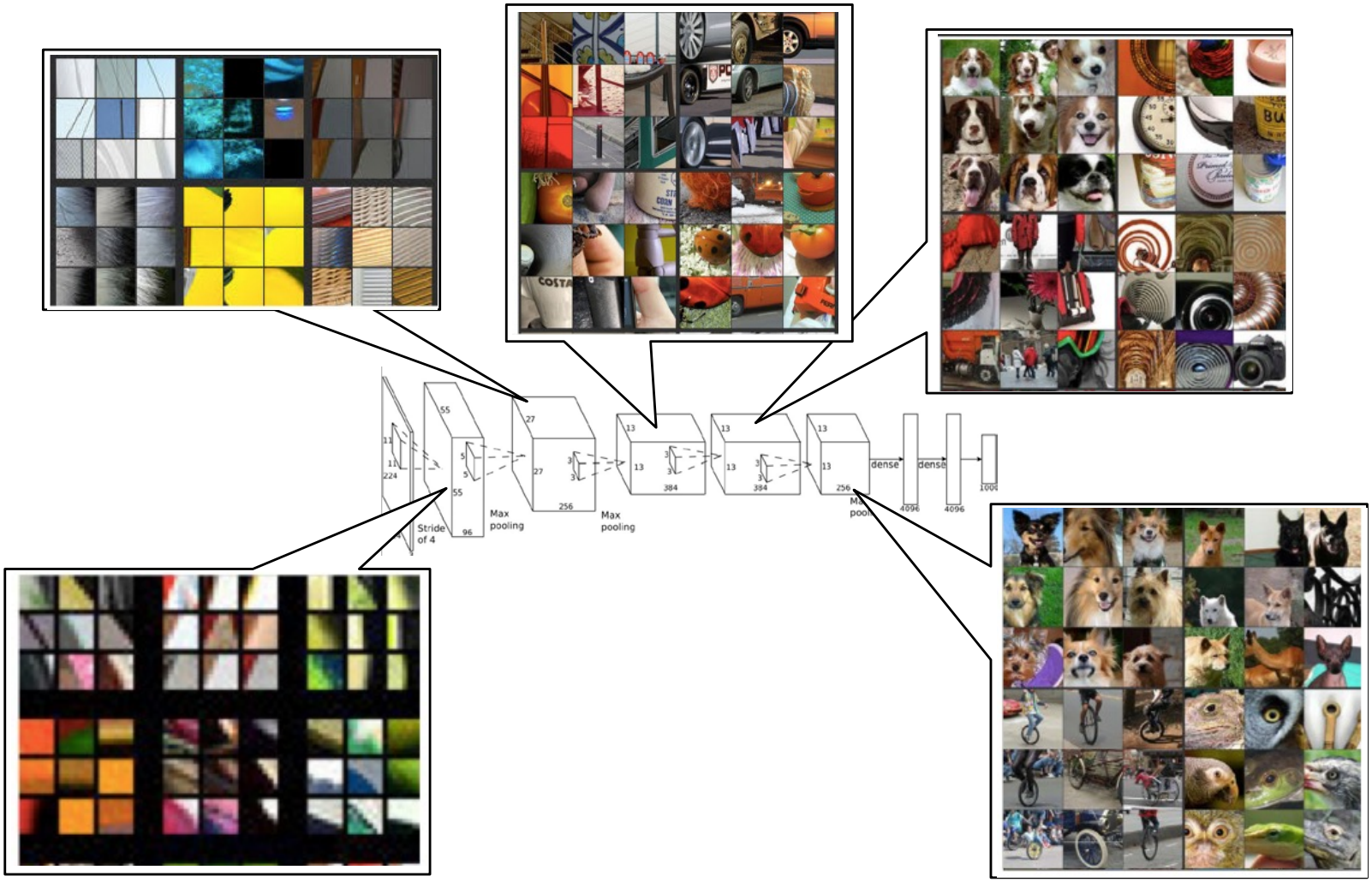
例：「ボトル」



学習で獲得する特徴表現

Zeiler-Fergus, Visualizing and understanding convolutional networks, arxiv2014

- 各層各ユニットを最も活性化する入力画像



Wang, Suganuma, Okatani, Improved Few-shot Segmentation by Redefinition of the Roles of Multi-level CNN Features, arXiv 2021

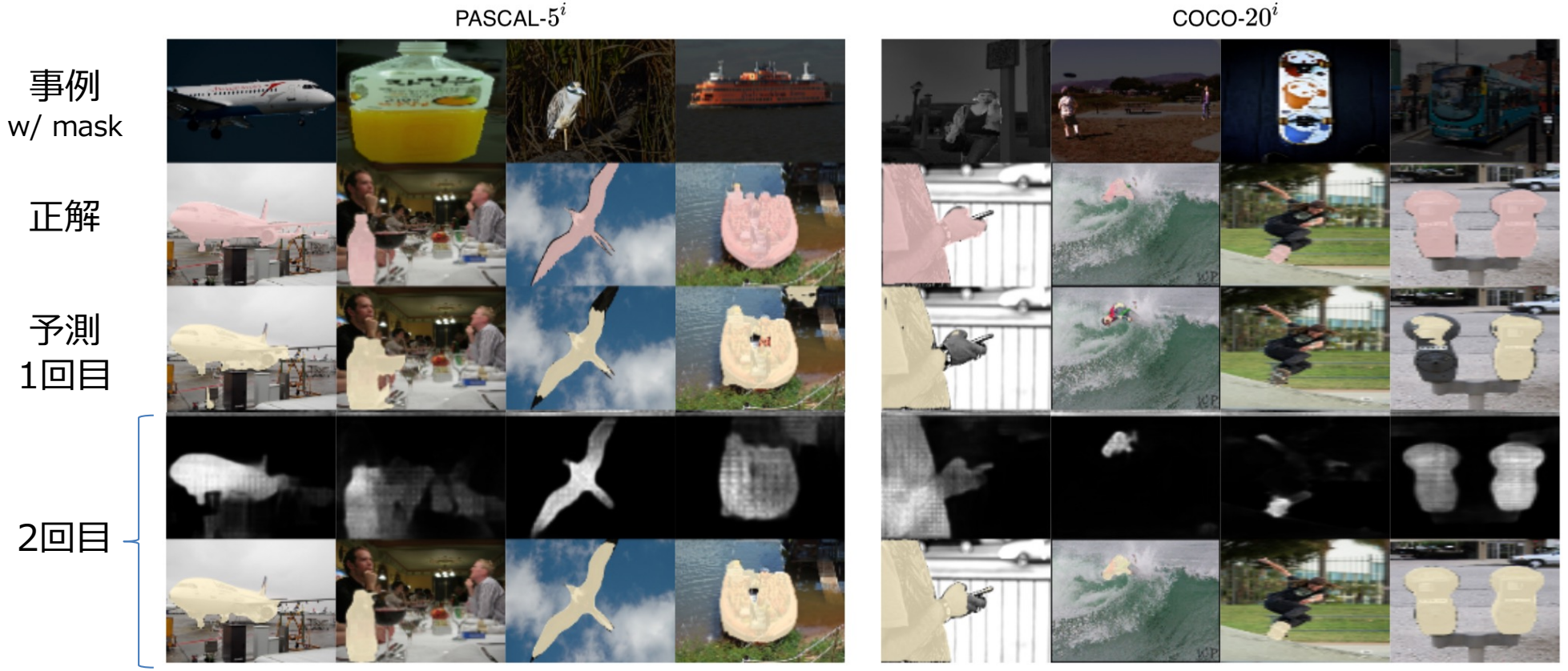


Table 2: Comparison with state-of-the-art methods under 1-shot and 5-shot settings on COCO-20^i. mIoU values for each test split and the averaged mIoU values (termed as mean) for four test splits are shown.

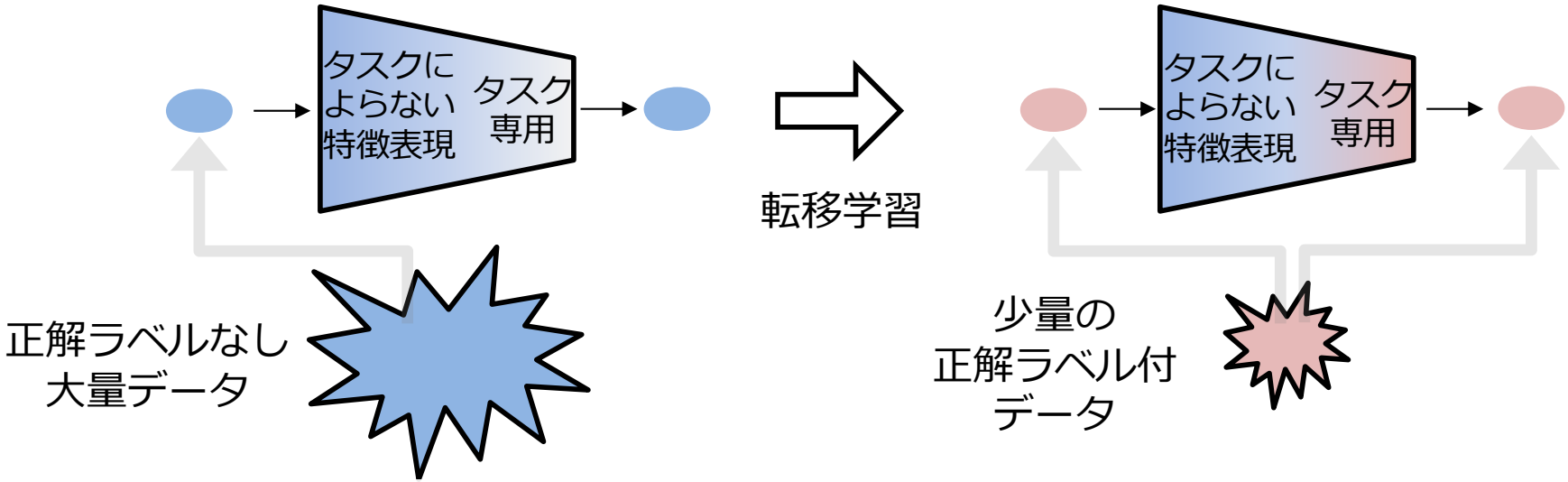
Methods	BB.	1-shot					5-shot				
		S0	S1	S2	S3	Mean	S0	S1	S2	S3	mean
FWB (Nguyen and Todorovic 2019)	VGG	18.4	16.7	19.6	25.4	20.0	20.9	19.2	21.9	28.4	22.6
PANet (Wang et al. 2019)	VGG	-	-	-	-	20.9	-	-	-	-	29.7
PFENet (Tian et al. 2020)	VGG	33.4	36.0	34.1	32.8	34.1	39.2	47.1	41.5	40.4	42.1
Ours	VGG	**34.6**	**36.6**	**35.9**	**35.0**	**35.5**	**40.3**	**48.0**	**44.0**	**43.0**	**43.8**
RePRI (Boudiaf et al. 2021)	RES	31.2	38.1	33.3	33.0	34.0	38.5	46.2	40.0	43.6	42.1
ASGNet (Li et al. 2021)	RES	34.9	36.9	34.3	32.1	34.6	41.0	48.3	40.1	40.5	42.5
PFENet (Tian et al. 2020)	RES	35.7	41.4	38.9	35.4	37.9	38.6	47.7	45.2	40.3	43.0
Ours	RES	**37.5**	**41.4**	**40.0**	**38.1**	**39.3**	**42.2**	**49.9**	**47.3**	**46.3**	**46.4**

データが少ない場合の対策

- 定番の方法～研究段階の方法まで
  - データ拡張
  - 転移学習
  - ドメイン適応（domain adaptation）
  - 少数事例学習（few-shot learning）
  - 半教師学習（semi-supervised），弱教師学習（weakly-supervised）

- 今後の方向
  - 自己教師学習を基礎とする巨大なDNNモデル

自己教師学習



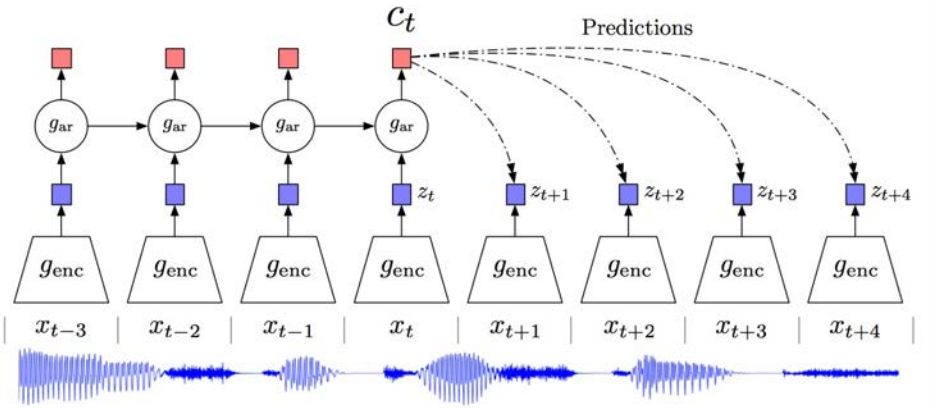
転移後，自己学習（self-training）を実行することでさらなる性能向上

SimCLR2[Chen+2020]

自己教師学習

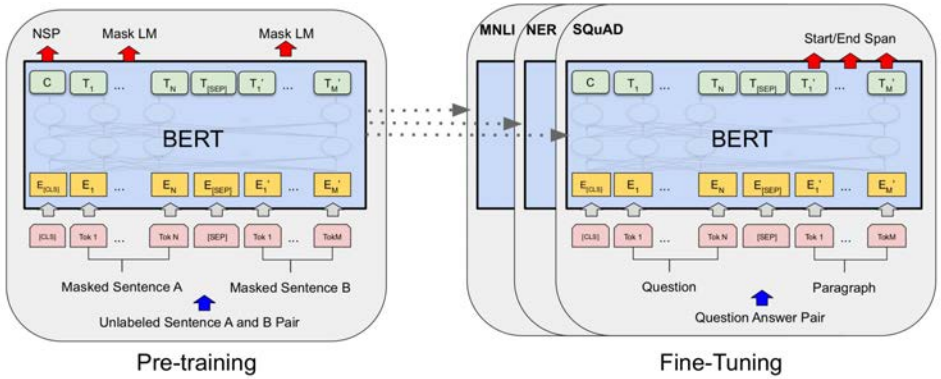
- 教師ラベルが「ただ」で手に入るようなタスク（=pretext task）で事前学習 → 少数訓練データで学習

音声信号：将来の信号を予測



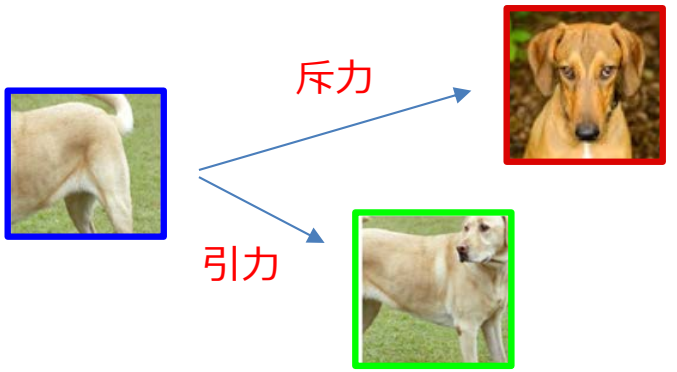
Contrastive Predictive Coding [van den Oord-Li-Vinyals arXiv18]

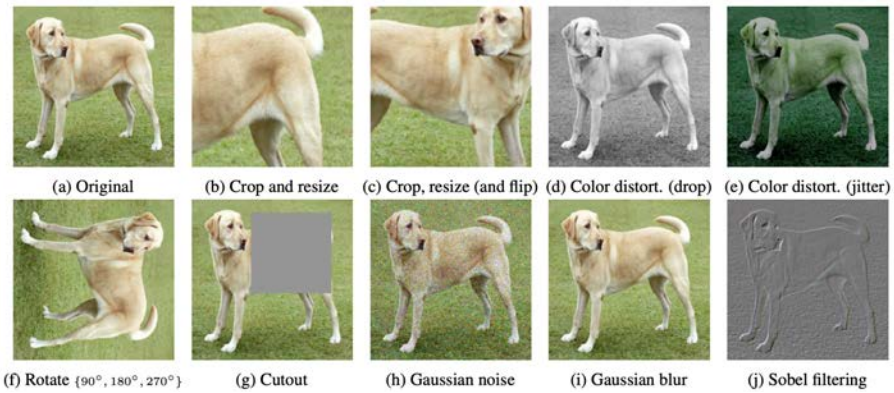
言語：隠した単語を予測



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Devlin+18]

画像：内容を変えない変換を施した画像を同一視（対照表現学習）



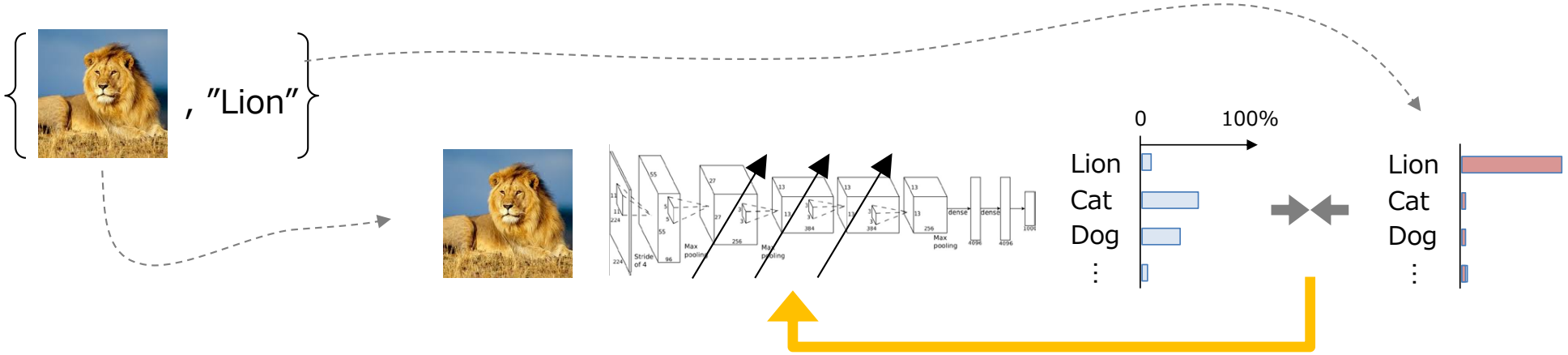


SimCLR [Chen+20]

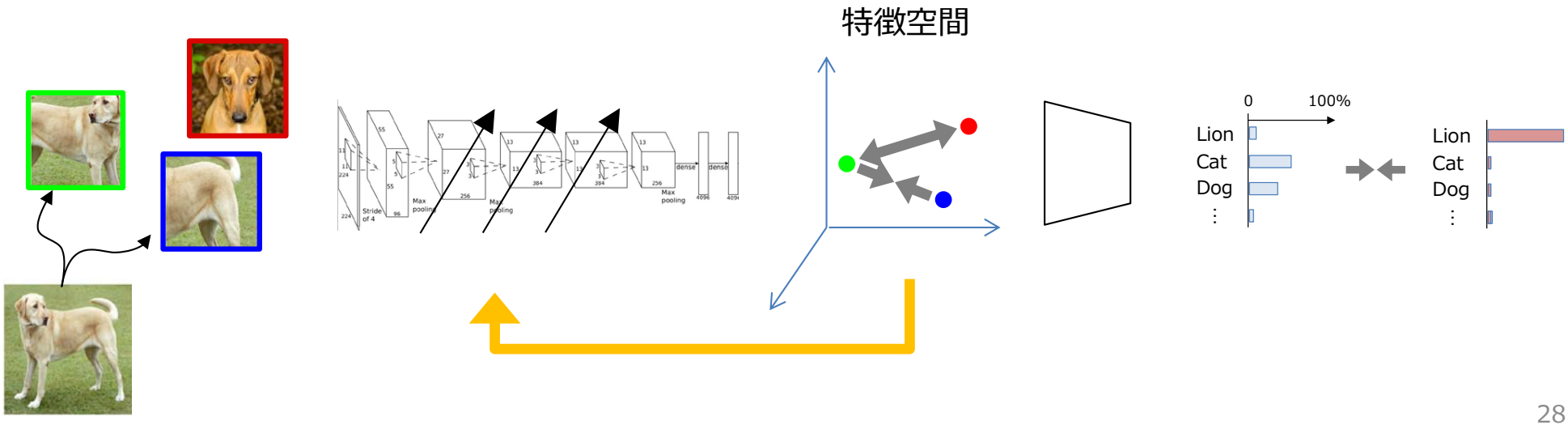
自己教師学習：対照学習

(*contrastive learning*)

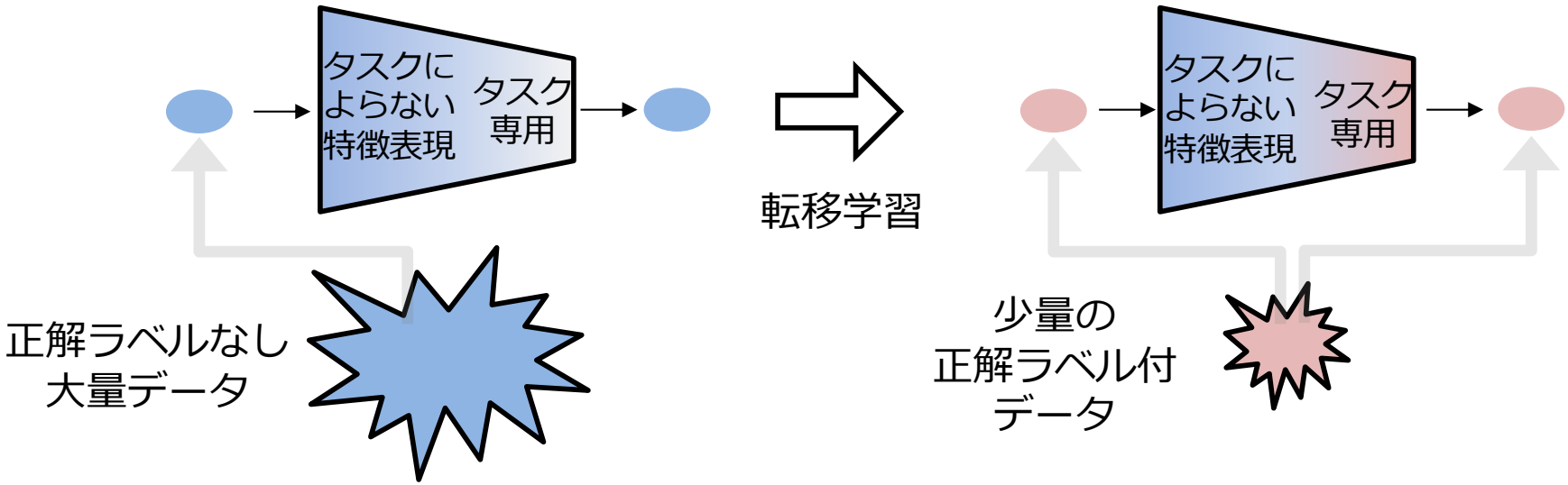
教師あり学習：1対1の入出力関係（画像→“Lion”）



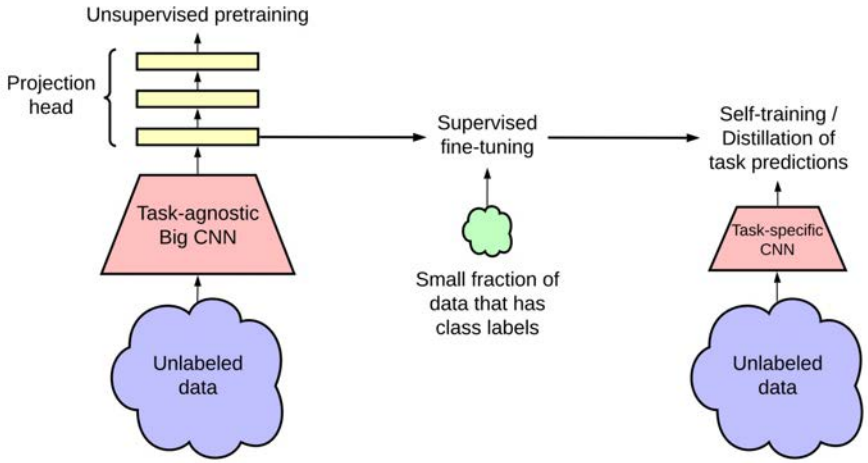
自己教師学習：特徴表現（画像の内容の同一性）

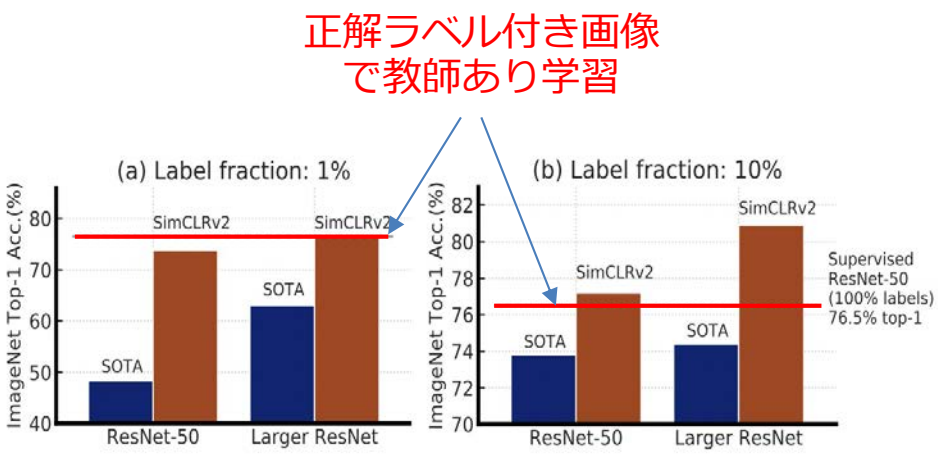


自己教師学習



転移後，自己学習（self-training）を実行することでさらなる性能向上





SimCLR2[Chen+2020]

学習データの量に対する見方

深層ニューラルネットワークは人に比肩する精度で物体認識が可能

ただしそれには正解ラベル付き画像100万枚の教師データを要する

深層ニューラルネットワークは人に比肩する精度で物体認識が可能

~~ただしそれには正解ラベル付き画像100万枚の教師データを要する~~

「自己教師学習」や「自己訓練」でラベル付き画像はそんなに要らなそう

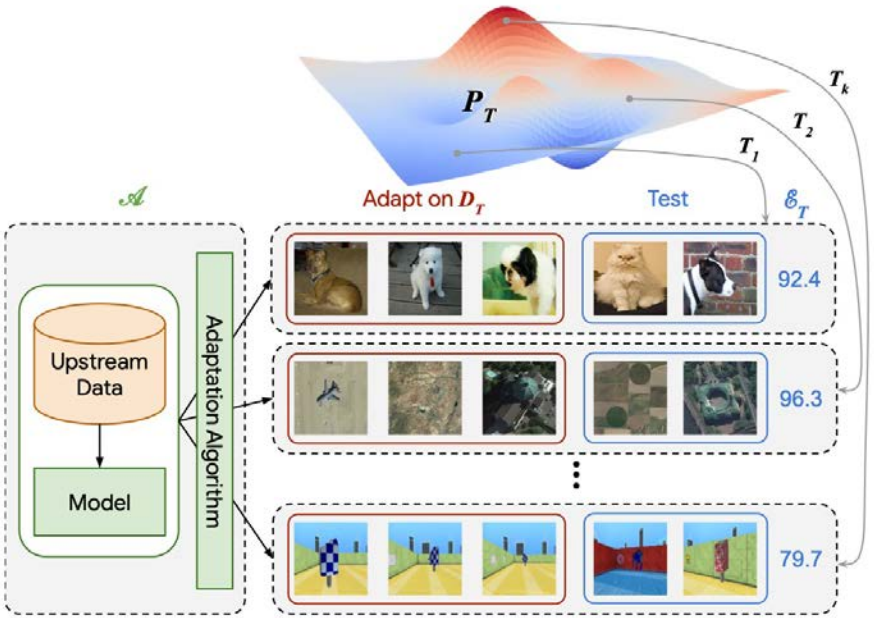
self-supervised learning　　self-training

事前学習　　半教師あり学習

トレンド：万能特徴表現の学習へ

- 完全教師あり学習の限界
  - 大量データ依存性・ショートカット学習
  - 定式化できない問題
- 現在の方向性：万能特徴表現の学習
  - あらゆるタスクに転移可能な特徴表現の獲得

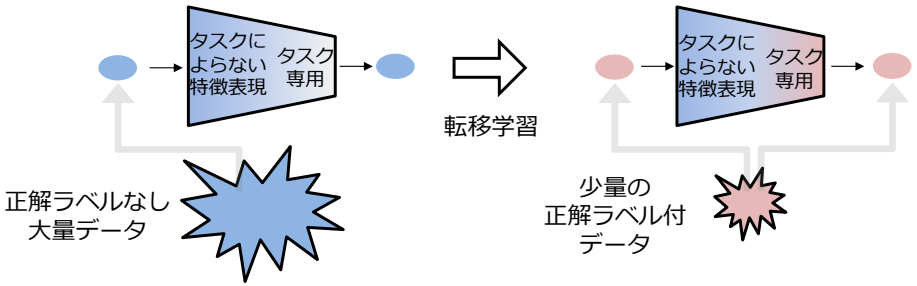
VTAB(Visual Task Adaptation Benchmark) [Zhai+2019]

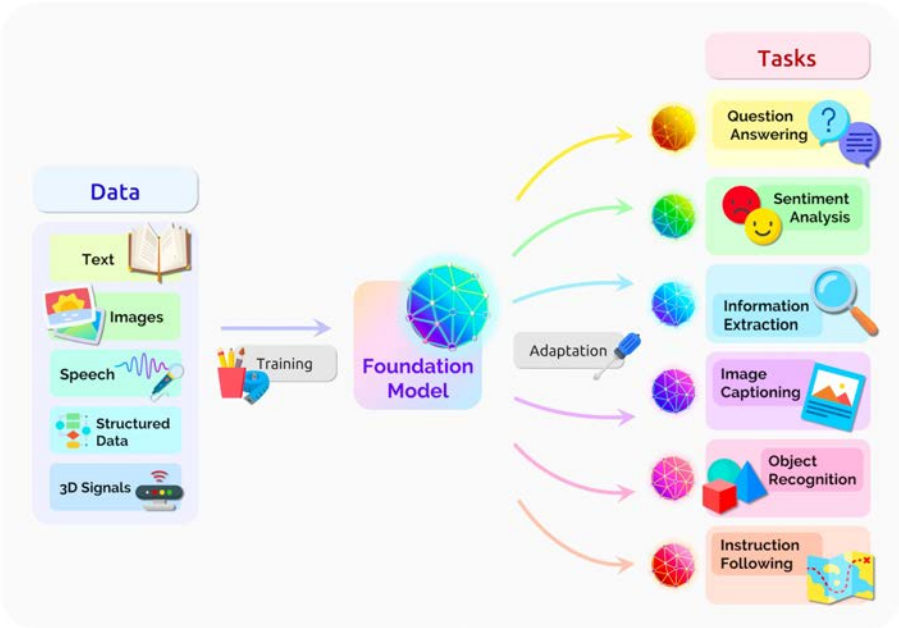


Category	Dataset	Train size	Classes	Reference
● Natural	Caltech101	3,060	102	(Li et al., 2006)
● Natural	CIFAR-100	50,000	100	(Krizhevsky, 2009)
● Natural	DTD	3,760	47	(Cimpoi et al., 2014)
● Natural	Flowers102	2,040	102	(Nilsback & Zisserman, 2008)
● Natural	Pets	3,680	37	(Parkhi et al., 2012)
● Natural	Sun397	87,003	397	(Xiao et al., 2010)
● Natural	SVHN	73,257	10	(Netzer et al., 2011)
● Specialized	EuroSAT	21,600	10	(Helber et al., 2019)
● Specialized	Resisc45	25,200	45	(Cheng et al., 2017)
● Specialized	Patch Camelyon	294,912	2	(Veeling et al., 2018)
● Specialized	Retinopathy	46,032	5	(Kaggle & EyePacs, 2015)
● Structured	Clevr/count	70,000	8	(Johnson et al., 2017)
● Structured	Clevr/distance	70,000	6	(Johnson et al., 2017)
● Structured	dSprites/location	663,552	16	(Matthey et al., 2017)
● Structured	dSprites/orientation	663,552	16	(Matthey et al., 2017)
● Structured	SmallNORB/azimuth	36,450	18	(LeCun et al., 2004)
● Structured	SmallNORB/elevation	36,450	9	(LeCun et al., 2004)
● Structured	DMLab	88,178	6	(Beattie et al., 2016)
● Structured	KITTI/distance	5,711	4	(Geiger et al., 2013)

まとめ～今後の姿

- 限られた訓練データで行う深層学習
  - ドメイン適応
  - 少量事例学習
- 超大規模な事前学習を行なったDNN = “Foundation Models”
  - 自然言語，画像，今後さらなるモダリティ？
  - コア技術＝自己教師学習
  - 膨大な計算量 → 限られた団体のみが作れる
  - 個別の問題への適応が主な仕事に





Bommasani+, On the Opportunities and Risks of Foundation Models, arXiv2022