Self-Supervised Learning in Computer Vision: Past, Present, Trends

Han Hu (胡瀚)
Visual Computing Group
Microsoft Research Asia (MSRA)

June 2nd, 2021 @BAAI

A Story about Cake (in Yann LeCun's Turing Talk)

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Credit by Yann LeCun

Why Self-Supervised Learning?

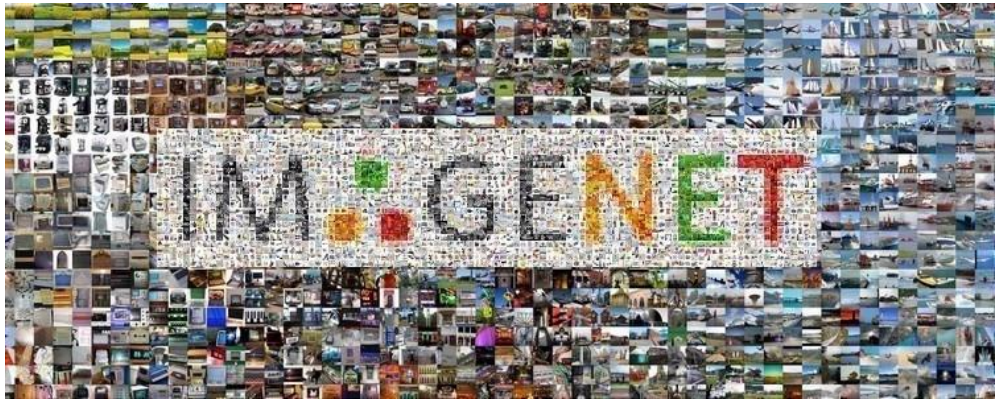
- Baby learns to see the world largely by observation

Photos courtesy of Emmanuel Dupoux

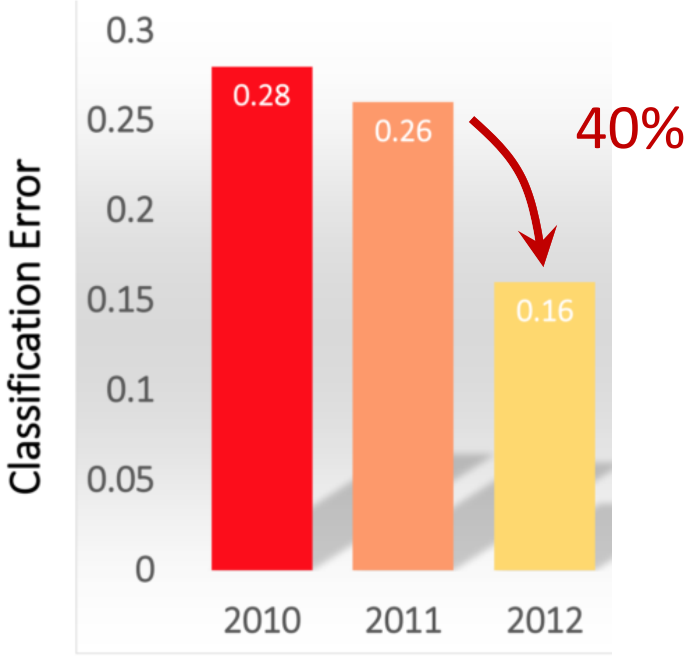
Credit by Yann LeCun

A Story about ImageNet

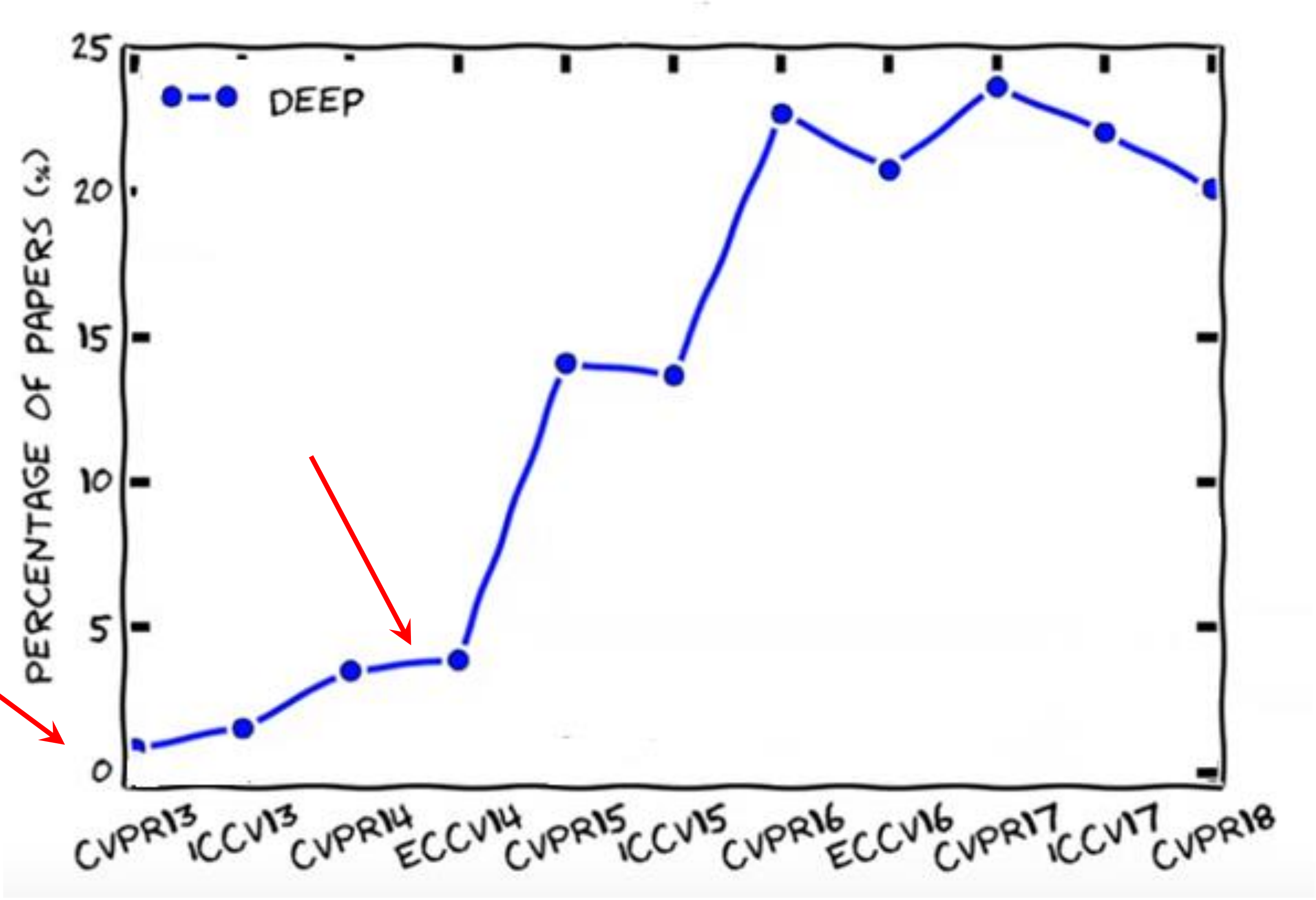
- AlexNet (NIPS'2012)



ImageNet Challenge

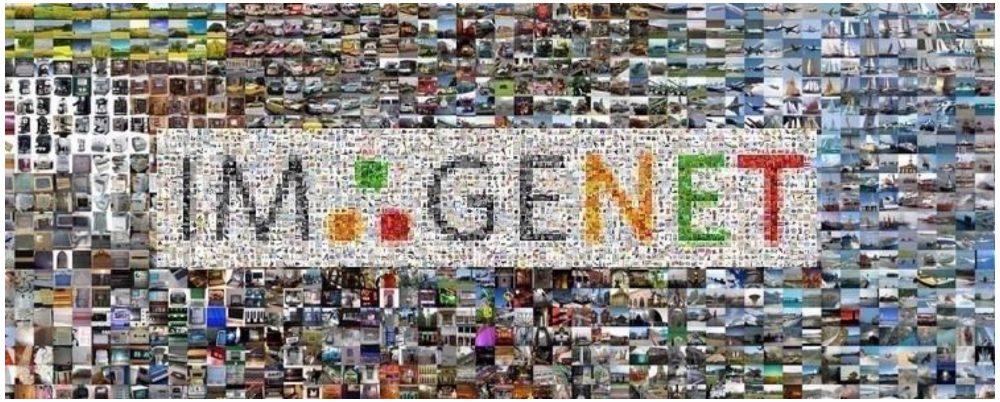


A Story about ImageNet



AlexNet

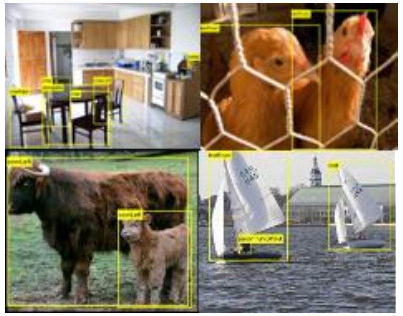
Supervised Pretraining + Finetuning (2014)



Pretraining on ImageNet Classification

Finetuning

Semantic Segmentation



Object Detection

Fine-grained Classification

Two Stories Meet Each Other

- Unsupervised Pretraining + Finetuning

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

2019.11

MoCo

FAIR

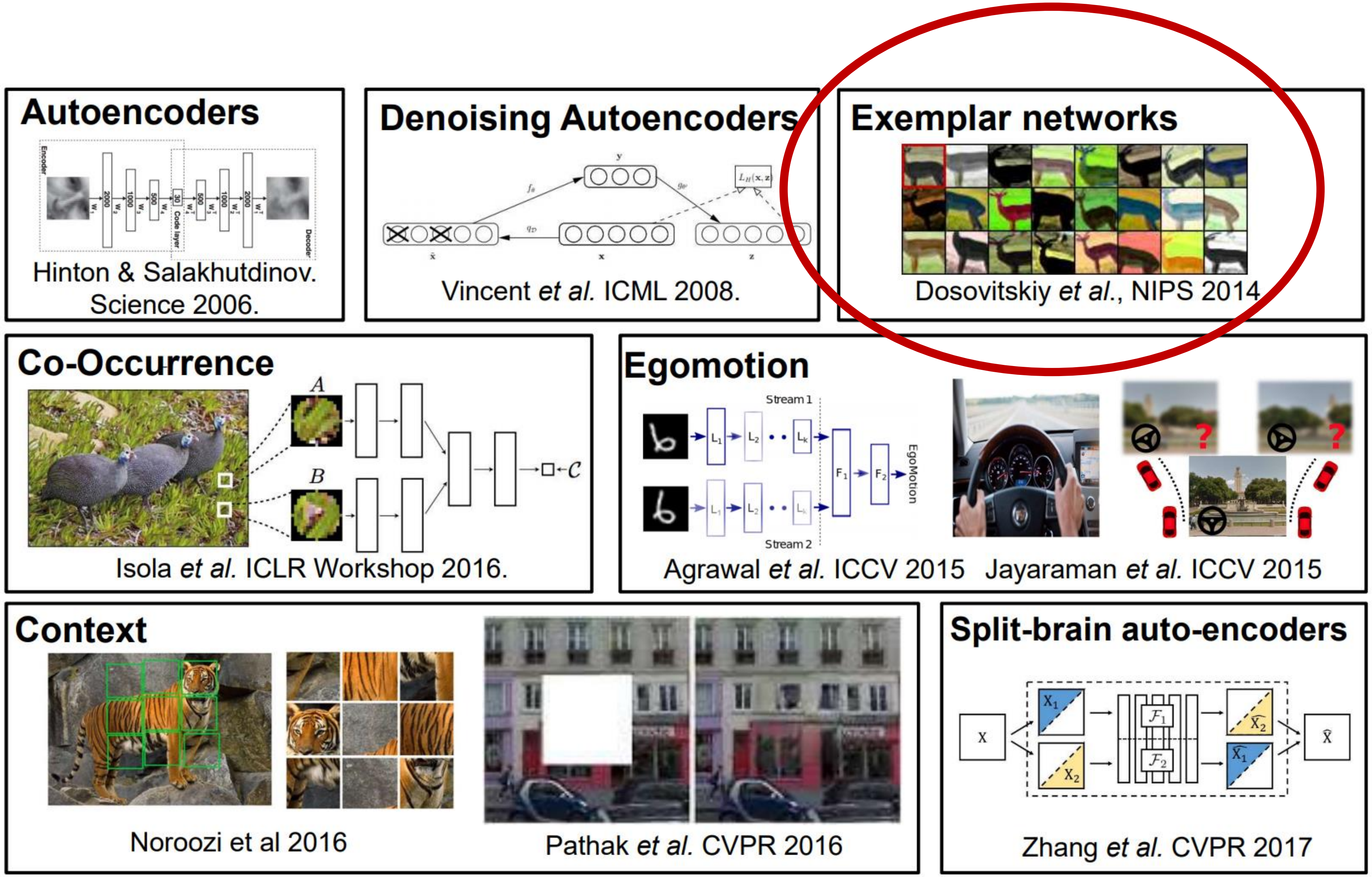
- For the first time, unsupervised pretraining outperform supervised pretraining on 7 down-stream tasks

The Self-Supervised Learning Era!

- Can utilize unlimited data
- Similar way as that of human baby learning

How Did We Get Here?

Credit mostly by Andrew Zisserman



How Did We Get Here?

Image #1 Image #2 Image #3

Pre-text task： Image discrimination

2014.6

Exemplar

Dosovitskiy et al, NIPS’2014

2018.5

Memory bank

Wu et al, CVPR’2018

2018.12

Deep metric transfer

MSRA

2019.11

MoCo

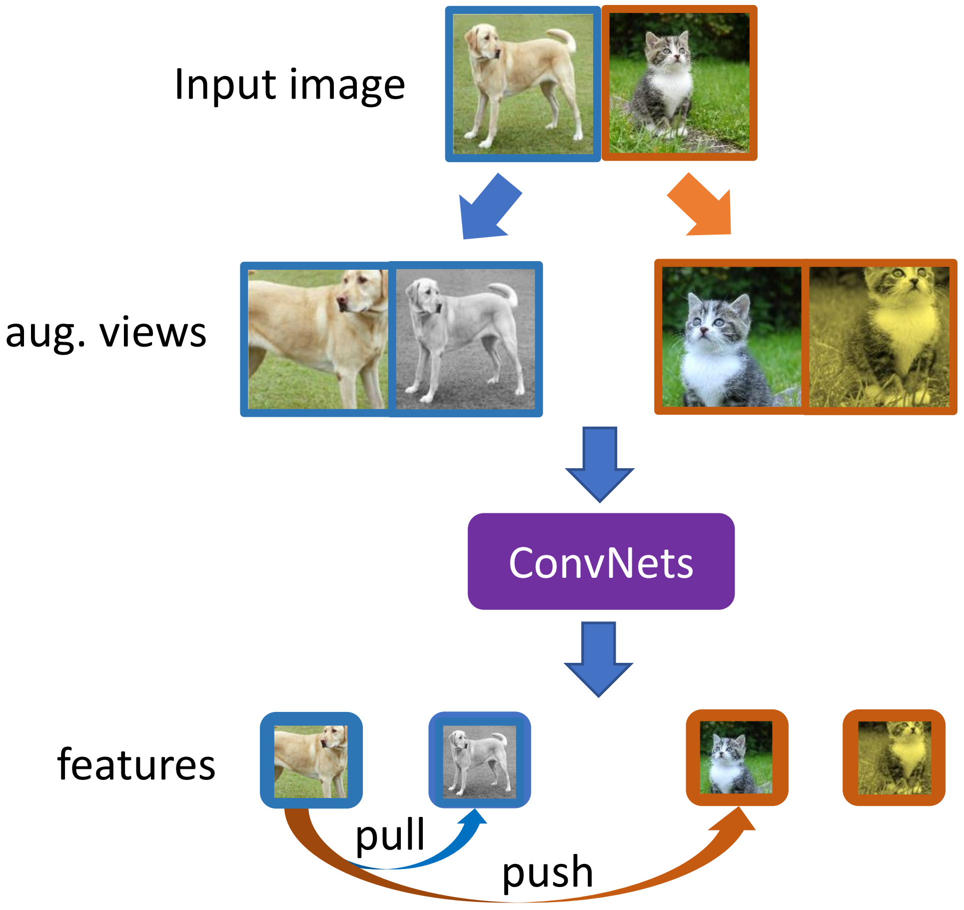
FAIR

- For the first time, unsupervised pretraining outperform supervised pretraining on 7 down-stream tasks

Contrastive Learning for Instance Discrimination

Pre-text task： Image discrimination

contrastive learning

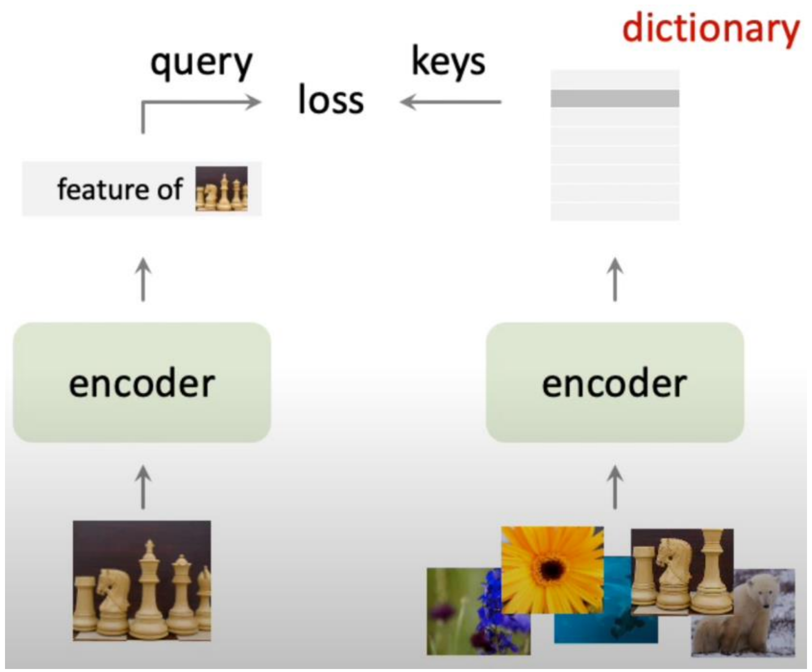


MoCo (CVPR'2020)

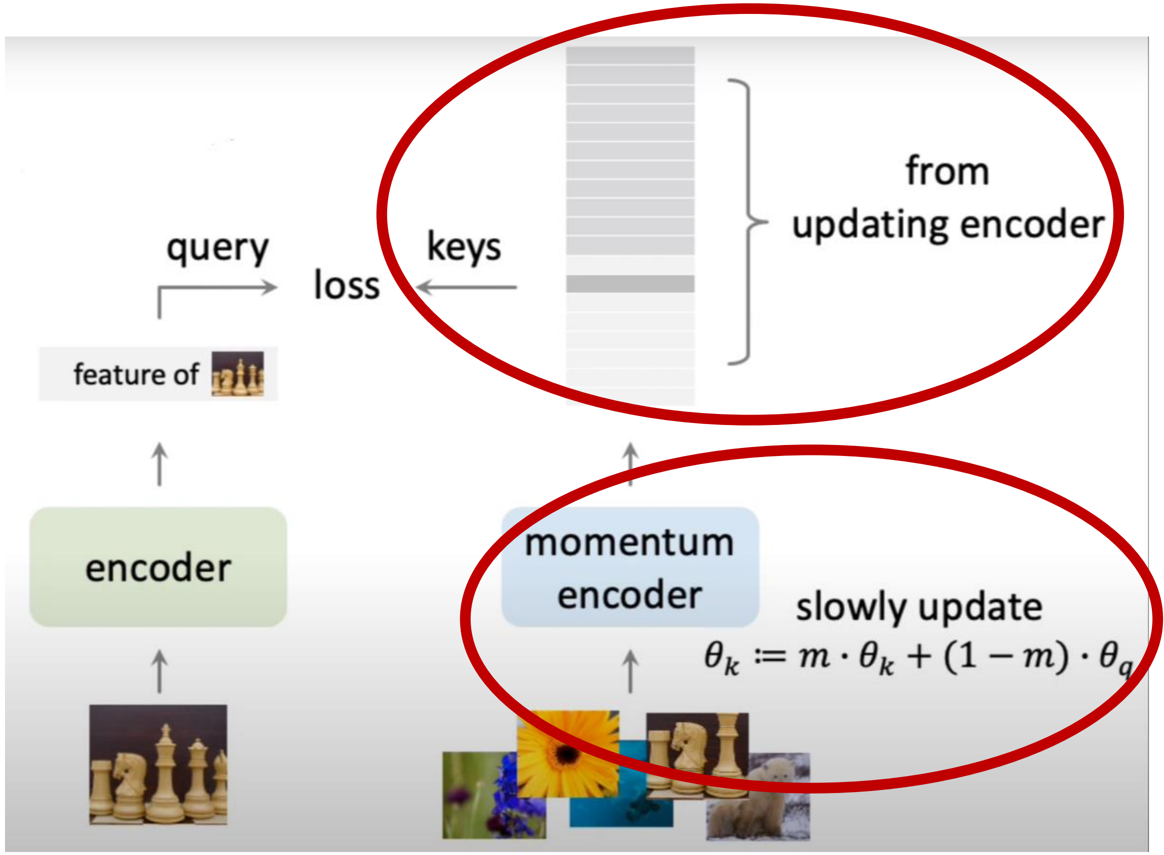
Credit by Kaiming He

- Large dictionary
- Consistent dictionary by momentum encoder

contrastive learning



MoCo

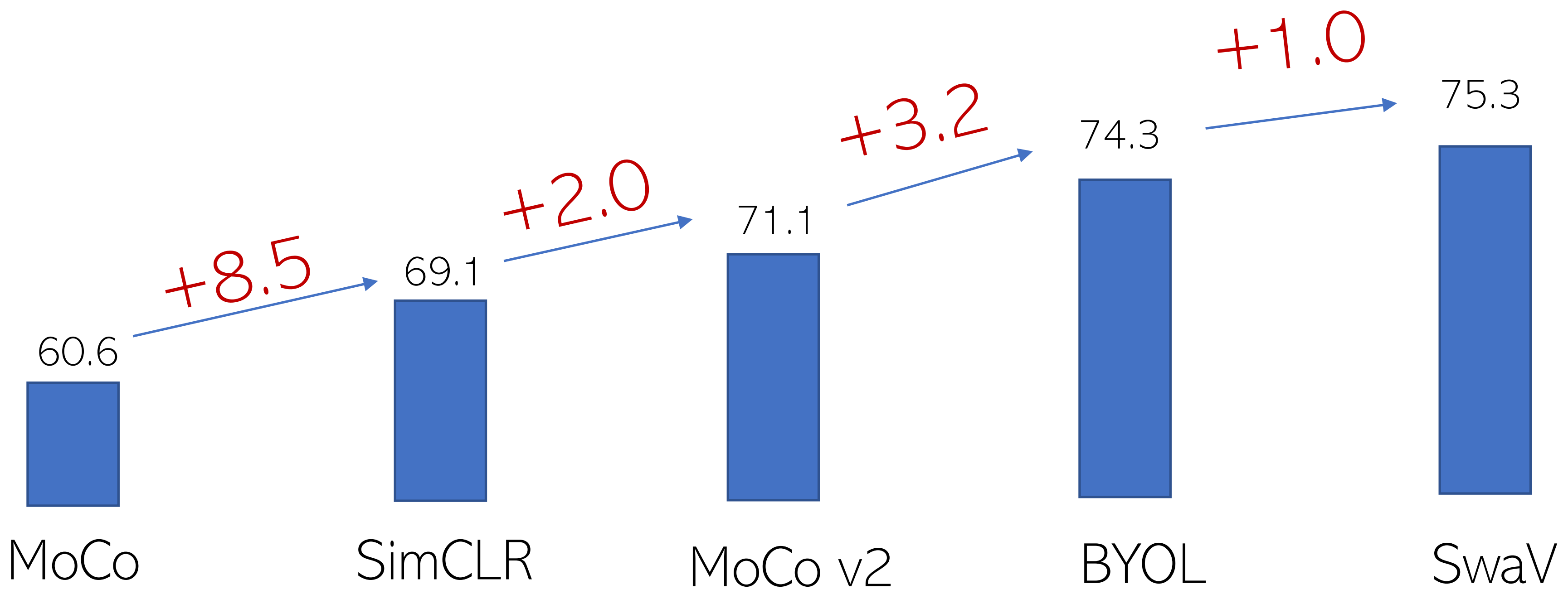


Post MoCo until NeurIPS’2020

2019.11-2020.7

Main Theme

- Improving ImageNet-1K linear evaluation (top-1 acc)



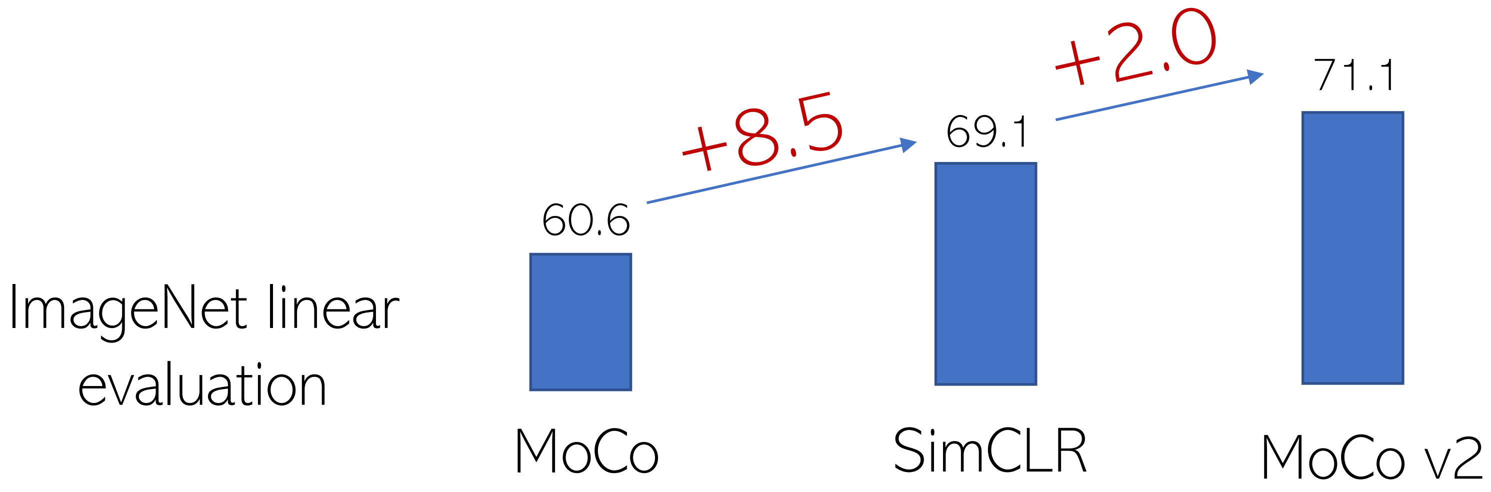
Totally absolute 14.7% improvements in 6 months!

Representative Works

- SimCLR (ICML'2020)
- SimCLR v2 (NeurIPS'2020)
- BYOL (NeurIPS'2020)
- SwaV (NeurIPS'2020)
- PIC (NeurIPS'2020)
- …

SimCLR (ICML'2020)

- Simpler: no momentum, no memory (dictionary)
- Sufficient distance between pretext tasks and downstream tasks
  - a linear projection layer -> a MLP layer
- Self-supervised learning benefit significantly from longer training



More Insights in SimCLR

- Self-supervised learning benefit more from larger models
- Self-supervised learning benefit significantly for semi-supervised learning

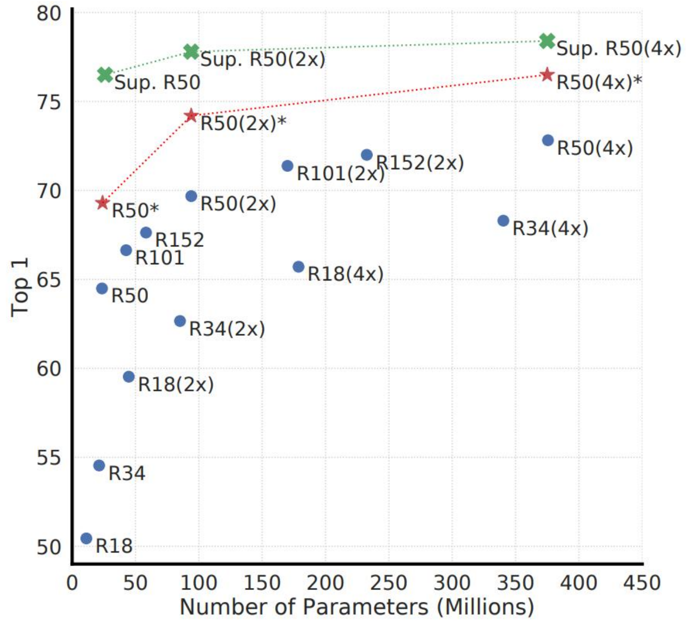


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs[7] (He et al., 2016).

Method	Architecture	Label fraction 1% (Top 5)	Label fraction 10% (Top 5)
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

+27.1

Table 7. ImageNet accuracy of models trained with few labels.

SimCLR v2 (NeurIPS'2020)

- "Big Self-Supervised Models are Strong Semi-Supervised Learners"

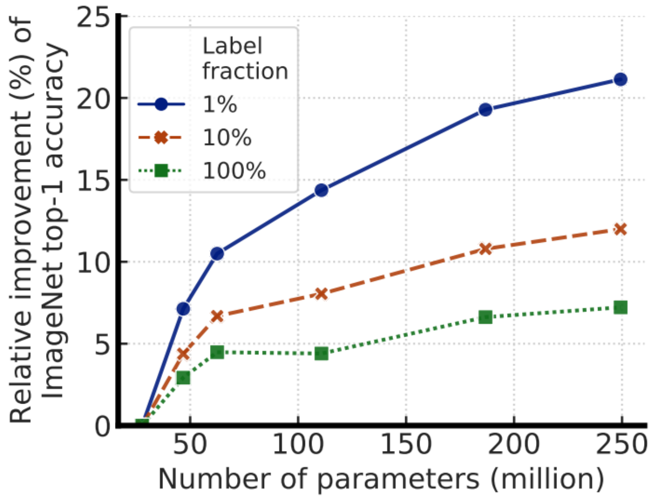
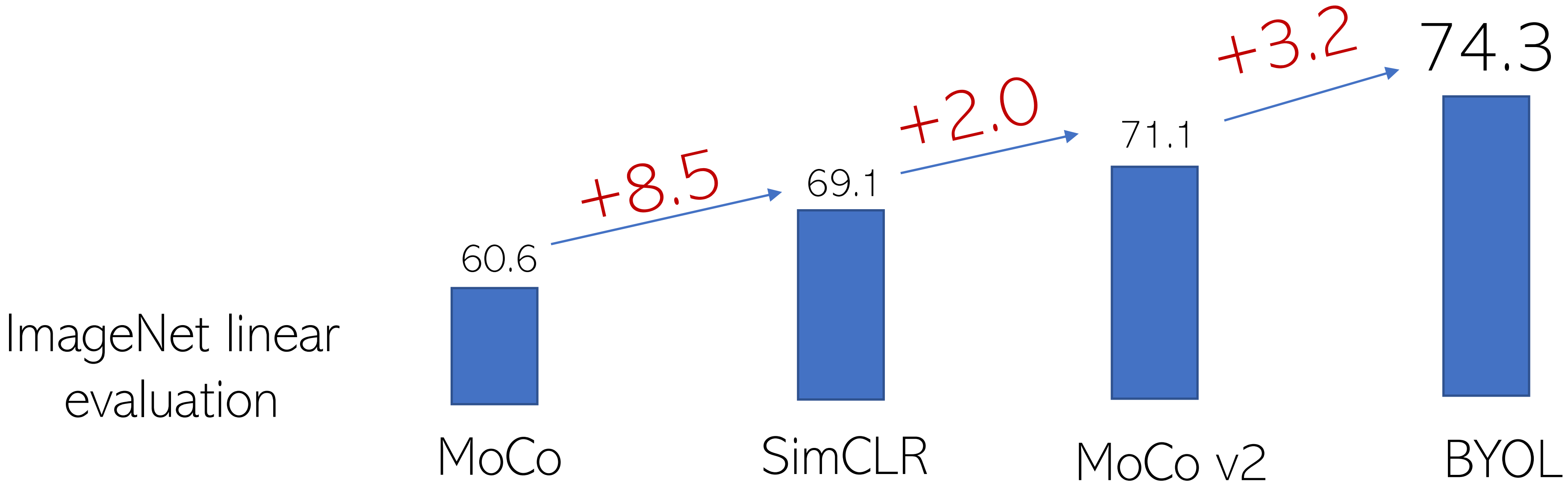


Figure 1: Bigger models yield larger gains when fine-tuning with fewer labeled examples.

Similar as that of GPT-3 in NLP!

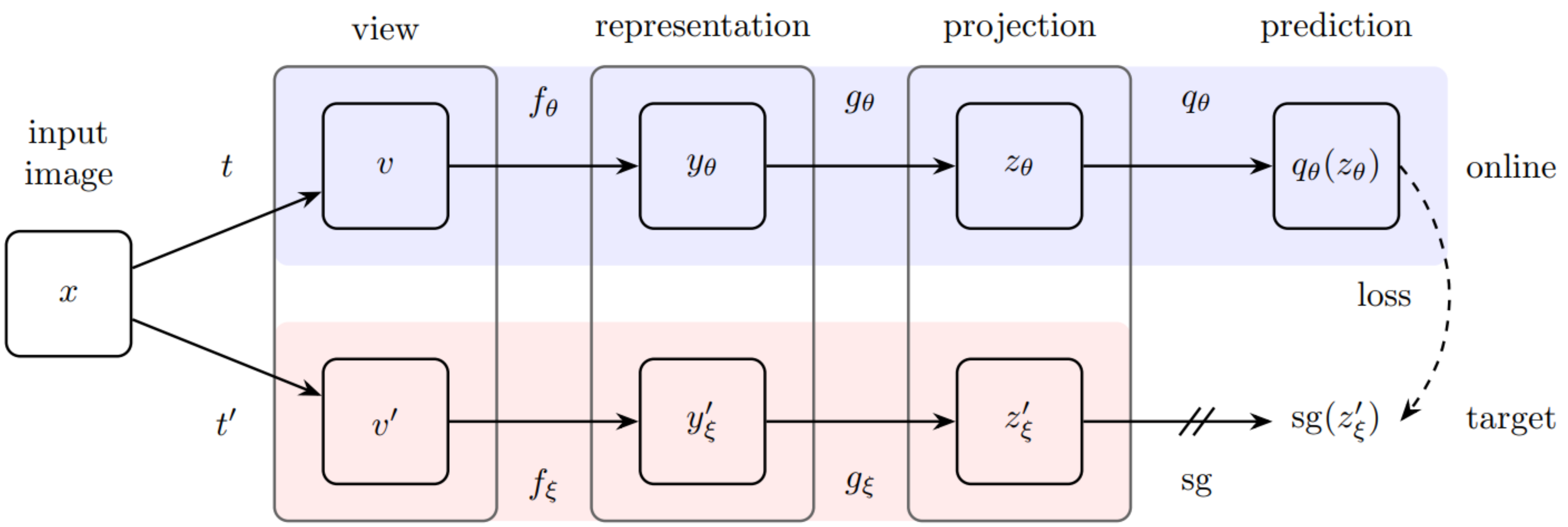
BYOL (NeurIPS'2020)

- Bootstrap Your Own Latent

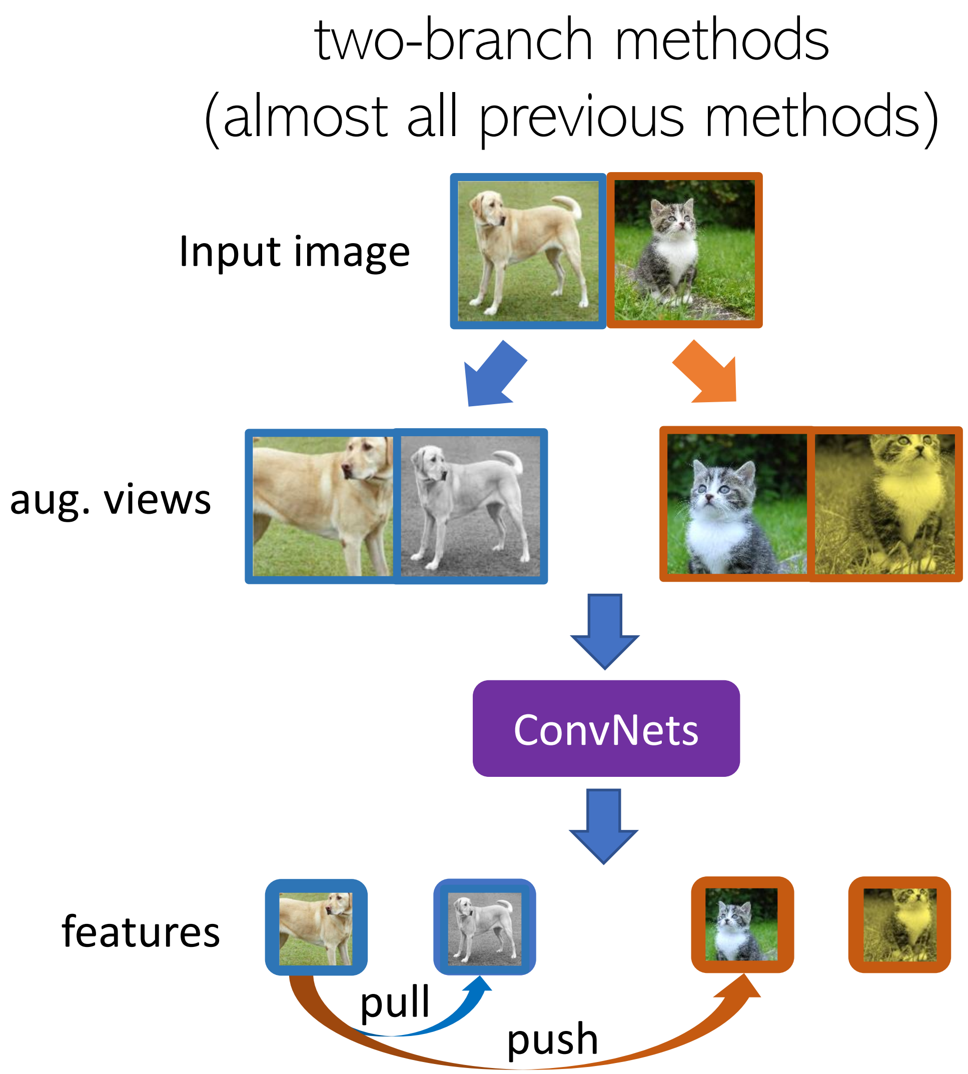


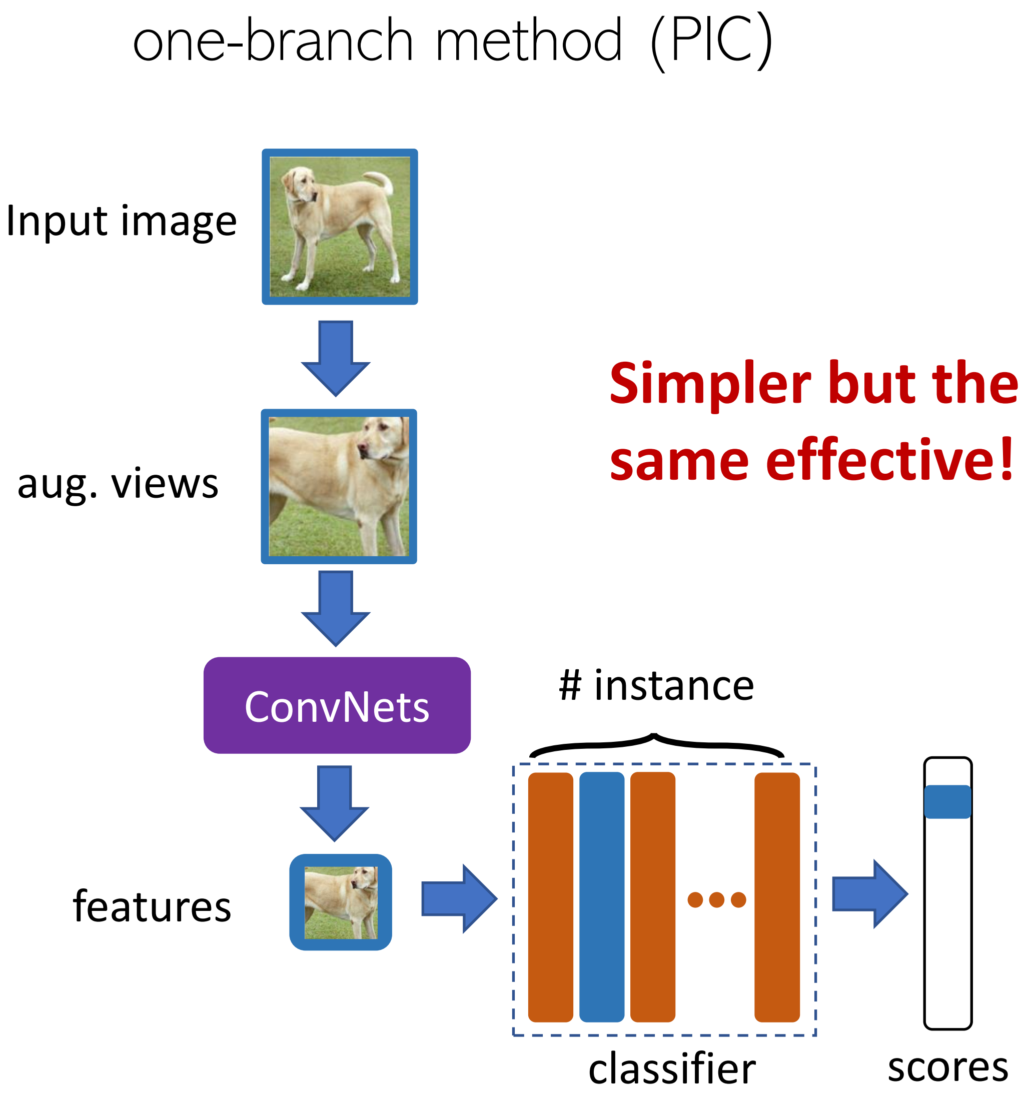
A Finding by BYOL

- MoCo: we need larger dictionary size (more negative pairs)
- BYOL: we do not need negative pairs anymore
  - an asymmetric design



PIC: a Single-Branch Method (NeurIPS'2020)





Simpler but the same effective!

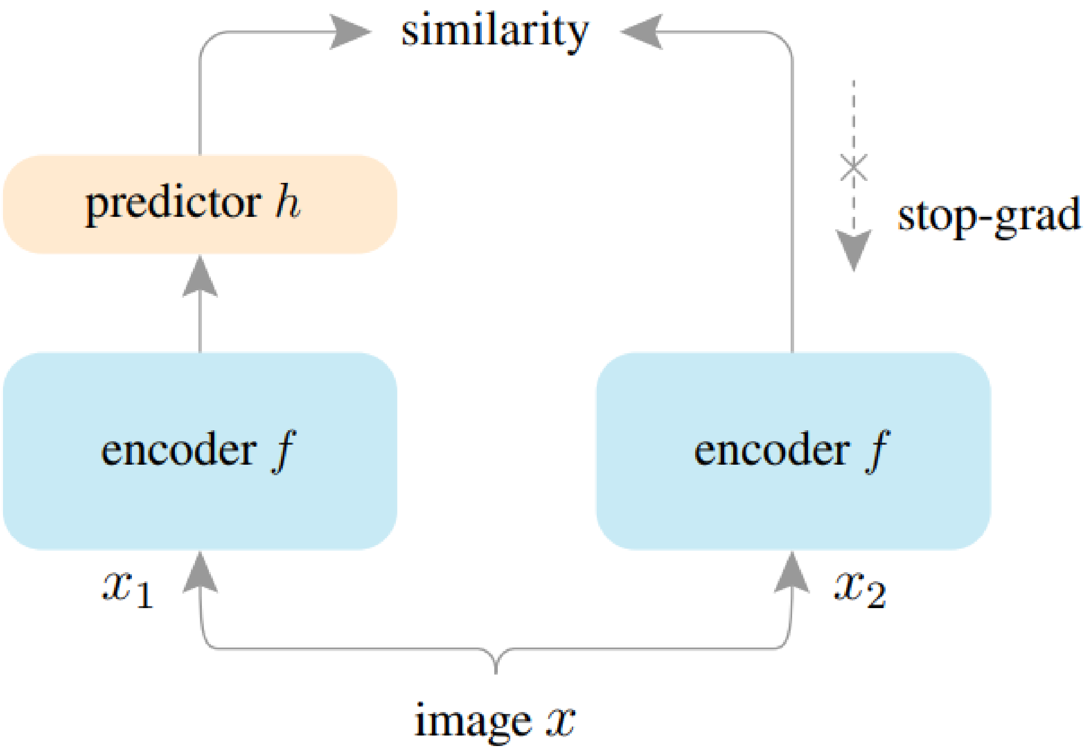
Post NeurIPS’2020

2020.8-present

Three Main Trends after NeurIPS'2020

- More study on why BYOL does not collapse
  - BYOL (Arxiv v3), SimSiam (CVPR'2021)
- Pre-training good features for down-stream tasks
  - Pixel-level pre-training
    - *PixPro*, DenseCL (CVPR'2021)
  - Object-level pre-training
    - SoCo (tech report)
- Self-supervised learning + Transformers
  - MoCo v3 (tech report), DINO (tech report)
  - SSL-Swin/MoBY (tech report)

SimSiam, BYOL (arxiv v3)



Another paper: understanding SSL dynamics without contrastive pairs (ICML'2021)

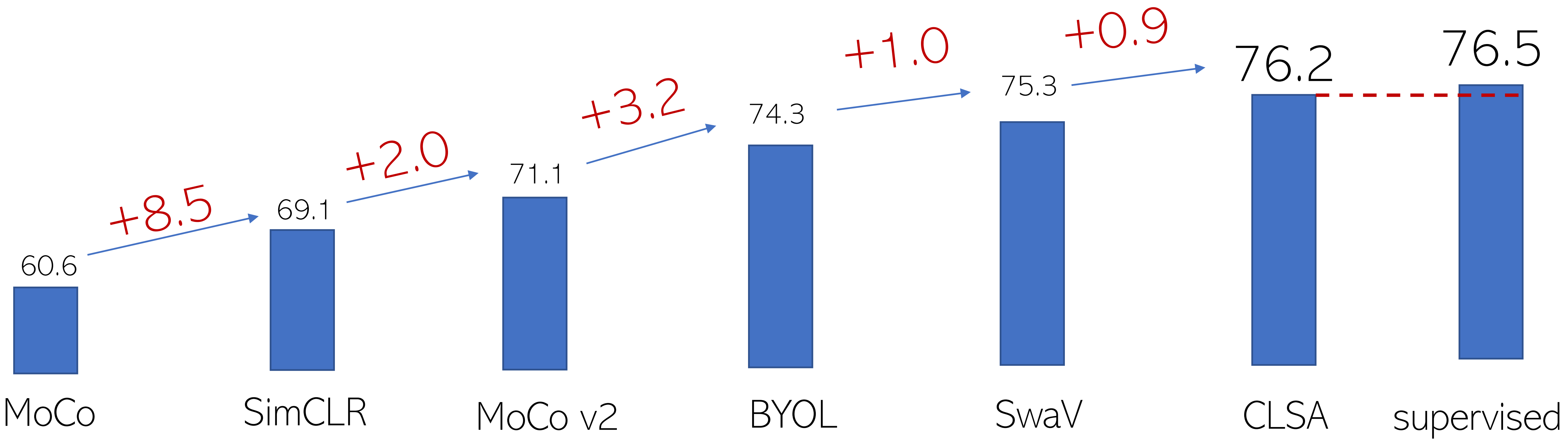
Trends after NeurIPS'2020

- ~~More study on BYOL why it does not collapse~~
  - ~~BYOL (Arxiv v3), SimSiam (CVPR'2021)~~
- Pre-training features which are good for down-stream tasks
  - Pixel-level pre-training
    - *PixPro*, DenseCL (CVPR'2021)
  - Object-level pre-training
    - SoCo (tech report)
- Self-supervised learning + Transformers
  - MoCo v3 (tech report), DINO (tech report)
  - SSL-Swin/MoBY (tech report)

Trends after NeurIPS'2020

- ~~More study on BYOL why it does not collapse~~
  - ~~BYOL (Arxiv v3), SimSiam (CVPR'2021)~~
- Pre-training features which are good for down-stream tasks
  - Pixel-level pre-training
    - *PixPro*, DenseCL (CVPR'2021)
  - Object-level pre-training
    - SoCo (tech report)
- Self-supervised learning + Transformers
  - MoCo v3 (tech report), DINO (tech report)
  - SSL-Swin/MoBY (tech report)

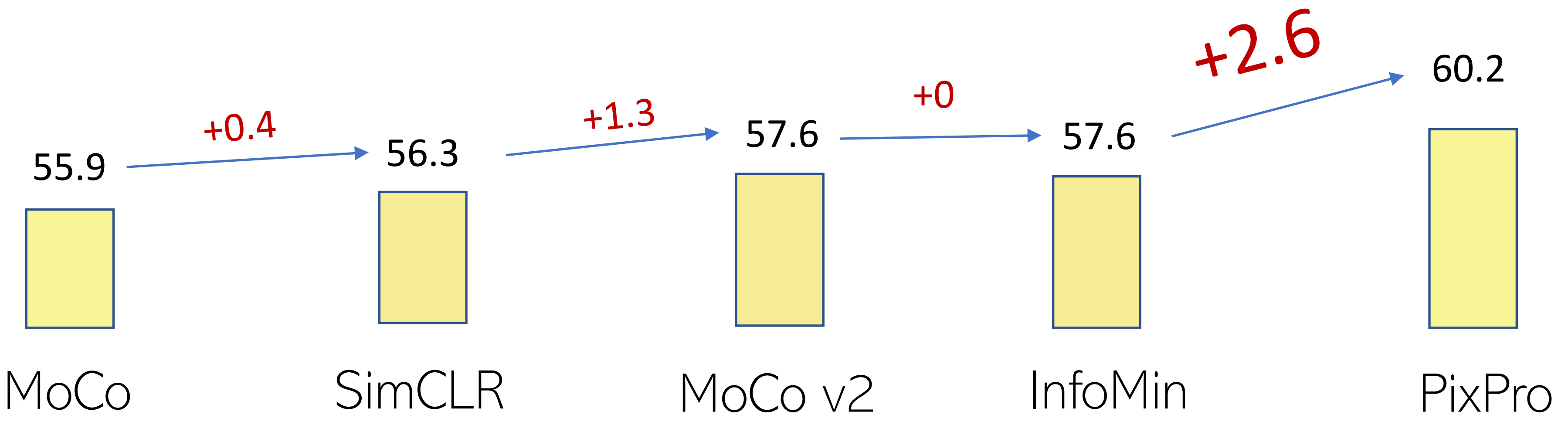
Improvements on ImageNet-1K linear evaluation



Totally 15.6% absolute improvements in 1 year!

Improvements on Pascal VOC object detection

- PixPro (CVPR'2021)



Totally 1.7% absolute improvements in 1 year!

Zhenda Xie et al. *Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning*. CVPR'2021

PixPro Results

- VOC detection (+2.6 mAP)
- COCO FPN detection (+0.8 mAP) COCO C4 (+1.0 mAP)
- Cityscape segmentation (+1.0 mIoU)

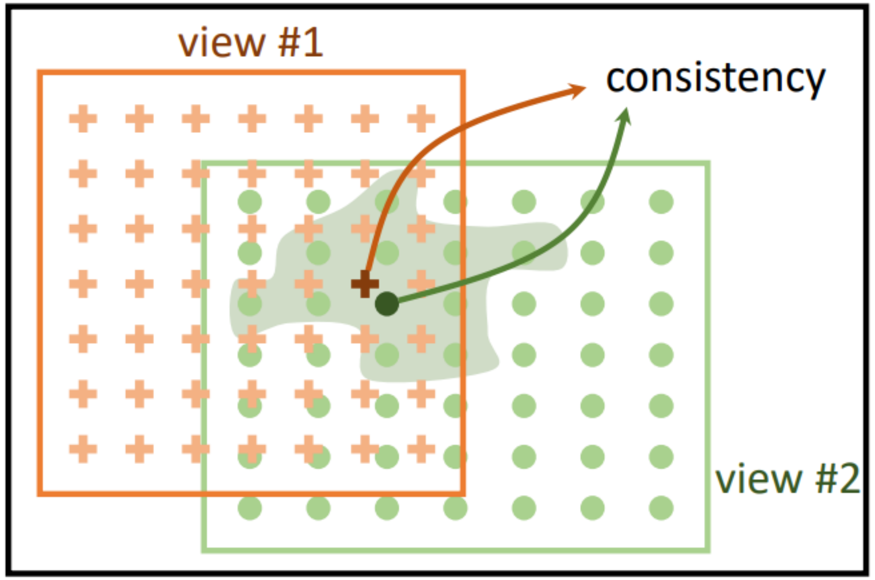
Method	#. Epoch	Pascal VOC (R50-C4)			COCO (R50-FPN)			COCO (R50-C4)			Cityscapes (R50)
		AP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mIoU
scratch	-	33.8	60.2	33.1	32.8	51.0	35.3	26.4	44.0	27.8	65.3
supervised	100	53.5	81.3	58.8	39.7	59.5	43.3	38.2	58.2	41.2	74.6
MoCo [18]	200	55.9	81.5	62.6	39.4	59.1	43.0	38.5	58.3	41.6	75.3
SimCLR [8]	1000	56.3	81.9	62.5	39.8	59.5	43.6	38.4	58.3	41.6	75.8
MoCo v2 [9]	800	57.6	82.7	64.4	40.4	60.1	44.3	39.5	59.0	42.6	76.2
InfoMin [30]	200	57.6	82.7	64.6	40.6	60.6	44.6	39.0	58.5	42.0	75.6
InfoMin [30]	800	57.5	82.5	64.0	40.4	60.4	44.3	38.8	58.2	41.7	75.6
PixPro (ours)	100	58.8	83.0	66.5	41.3	61.3	45.4	39.6	59.2	42.8	76.8
PixPro (ours)	400	**60.2**	**83.8**	**67.7**	**41.4**	**61.6**	**45.4**	**40.5**	**59.8**	**44.0**	**77.2**

+2.6 mAP　+0.8 mAP　+1.0 mAP　+1.0 mIoU

From Instance-Level to Pixel-Level Learning

Memory bank, MoCo,
SimCLR, BYOL, SwaV, PIC, …

Image #1 Image #2 Image #3

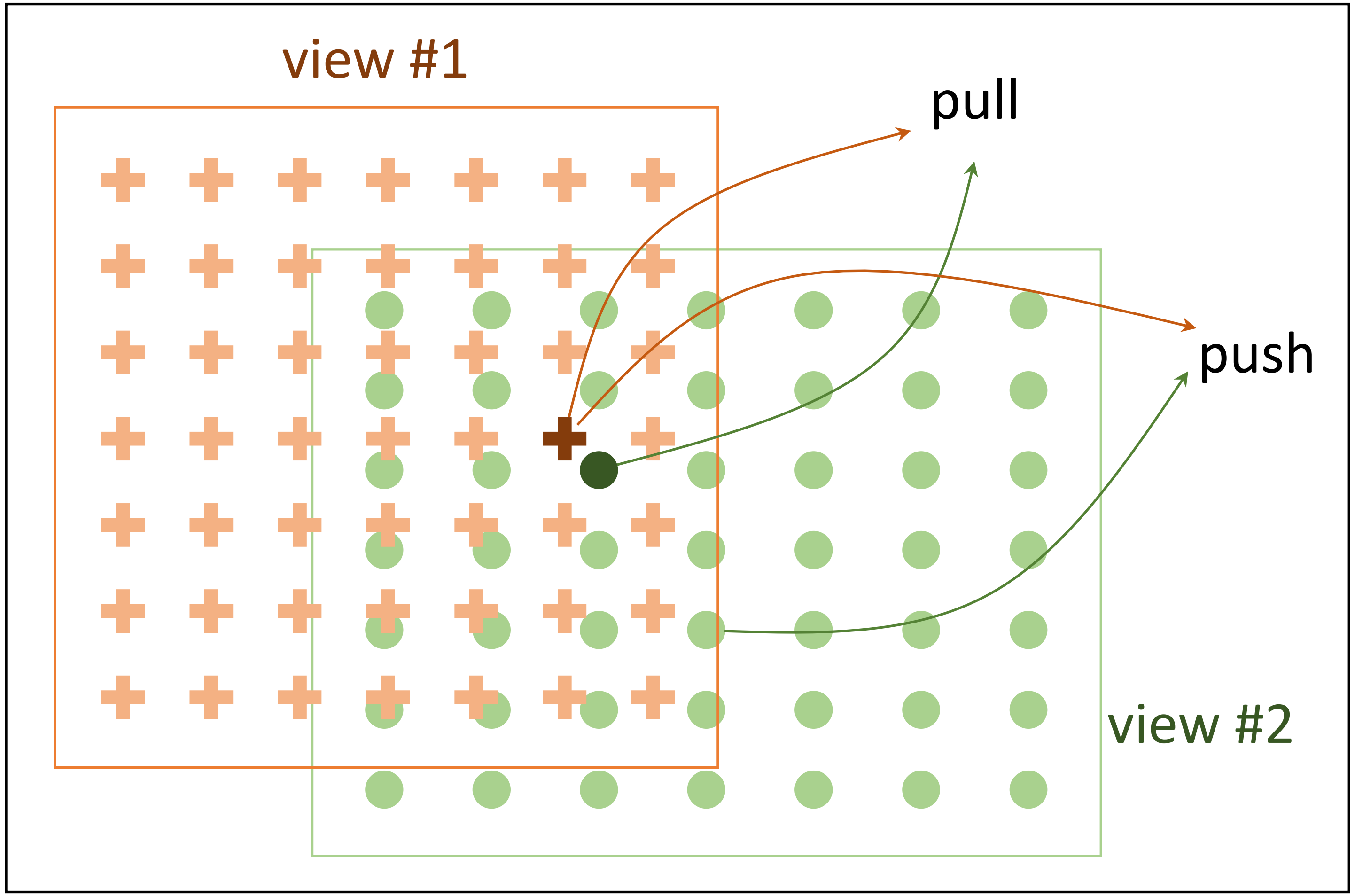


Previous pre-text tasks： instance discrimination

pixel-level pretext task

Pixel-Level Contrastive Learning

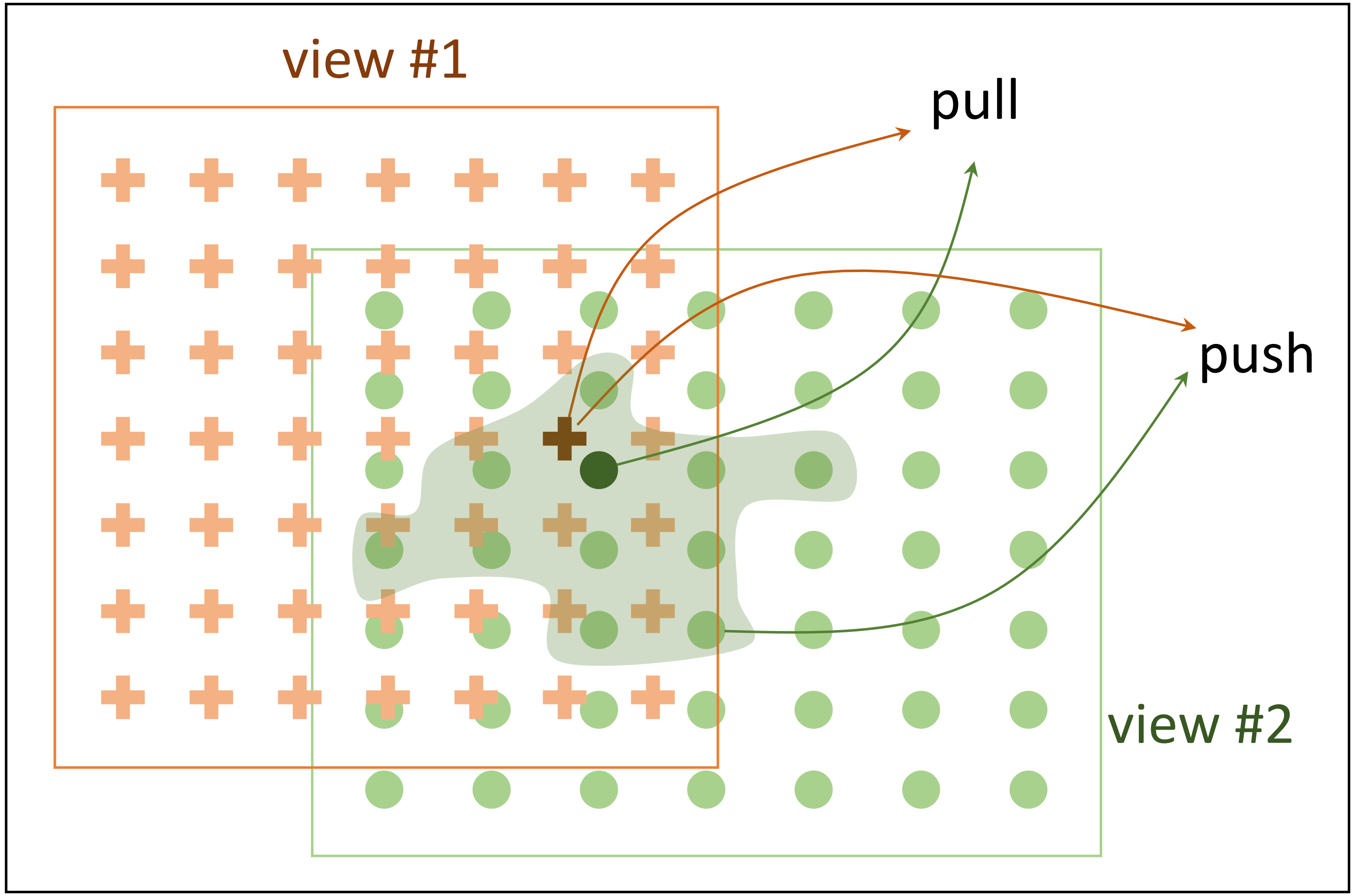
an image



pixel discrimination

Pixel-to-Propagation Consistency

an image



Pixel-to-propagation consistency

Pixel-to-Propagation Consistency

- **Pixel contrast:** spatial sensitivity
- **Propagation:** spatial smoothness

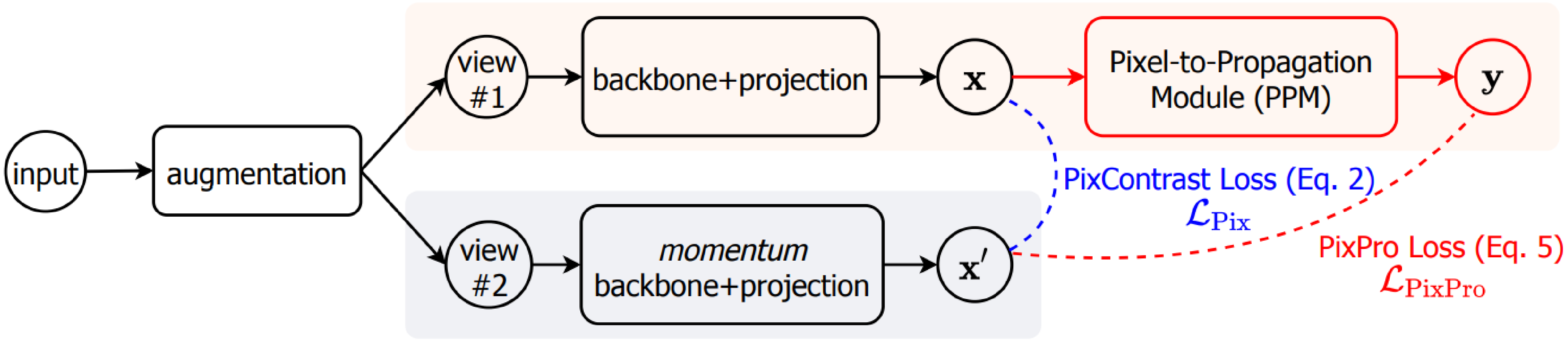
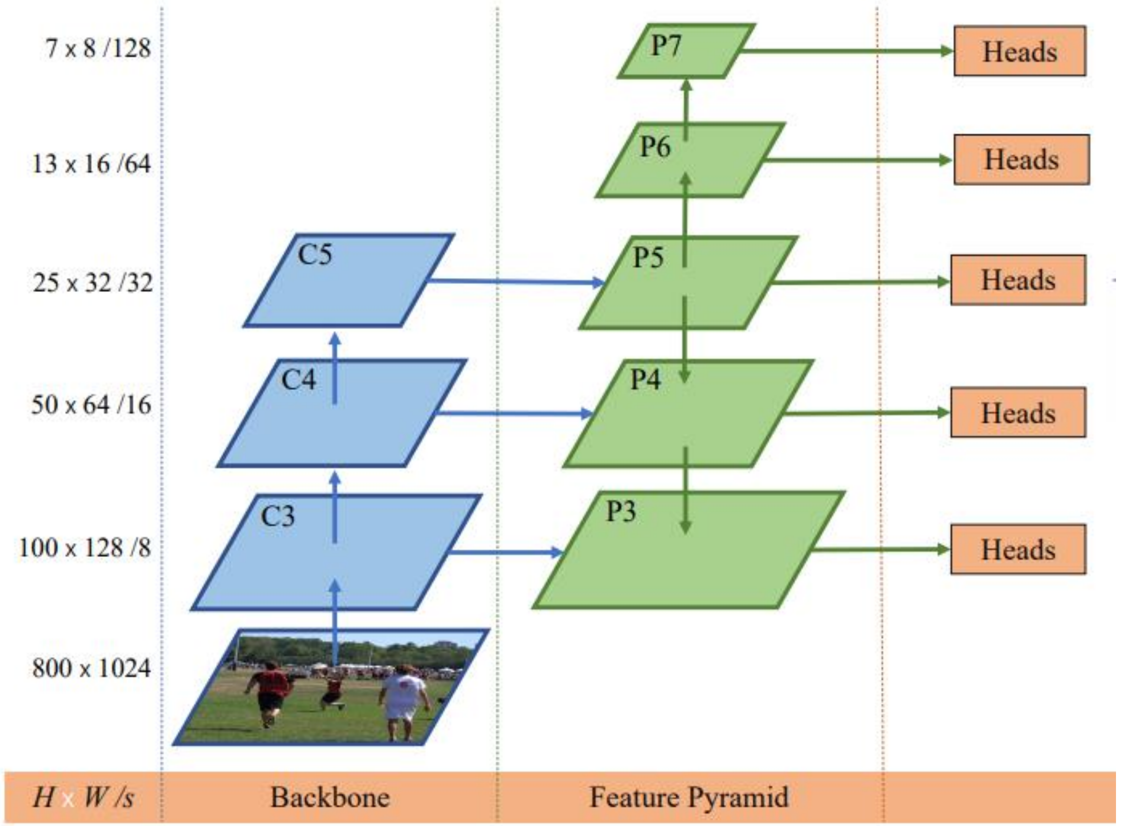


Figure 2. Architecture of the *PixContrast* and *PixPro* methods.

Aligning Pre-Training to Downstream Networks

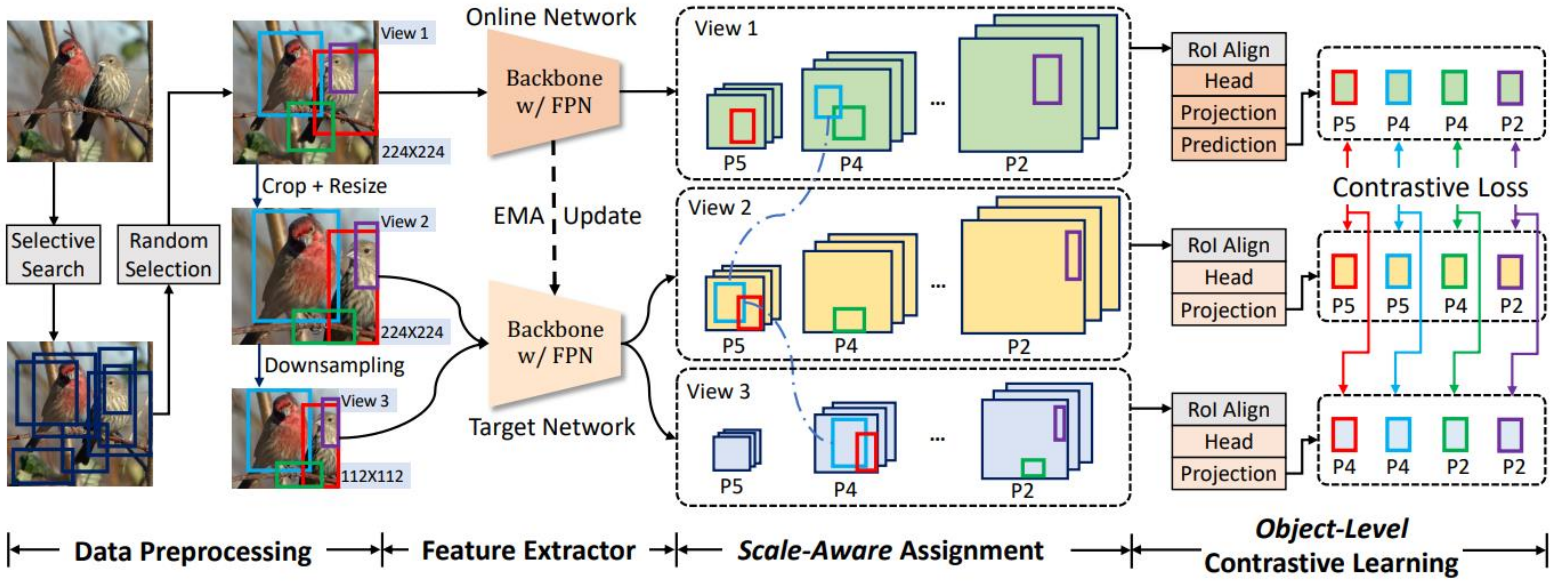
- Using the same architecture as in downstream tasks



An architecture in FCOS detector

Object-Level Pre-Training

- Aligning pretraining for object detection
  - SoCo (tech report, 2021)



Fangyun Wei et al. *Aligning Pretraining for Detection via Object-Level Contrastive Learning.* Tech Report 2021

Object-Level Pre-Training (SoCo)

- Results

Table 1: Comparison with state-of-the-art methods on **COCO** by using Mask R-CNN with **R50-FPN**.

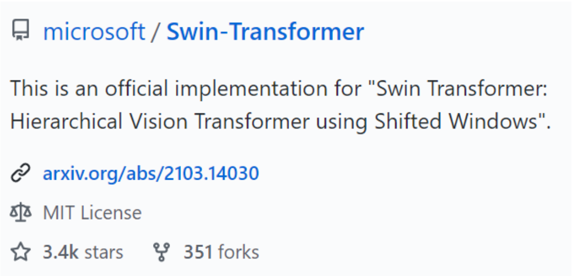
Methods	Epoch	1× Schedule AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	2× Schedule AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Scratch	-	31.0	49.5	33.2	28.5	46.8	30.4	38.4	57.5	42.0	34.7	54.8	37.2
Supervised	90	38.9	59.6	42.7	35.4	56.5	38.1	41.3	61.3	45.0	37.3	58.3	40.3
MoCo [4]	200	38.5	58.9	42.0	35.1	55.9	37.7	40.8	61.6	44.7	36.9	58.4	39.7
MoCo v2 [5]	200	40.4	60.2	44.2	36.4	57.2	38.9	41.7	61.6	45.6	37.6	58.7	40.5
InfoMin [6]	200	40.6	60.6	44.6	36.7	57.7	39.4	42.5	62.7	46.8	38.4	59.7	41.4
BYOL [3]	300	40.4	61.6	44.1	37.2	58.8	39.8	42.3	62.6	46.2	38.3	59.6	41.1
SwAV [7]	400	-	-	-	-	-	-	42.3	62.8	46.3	38.2	60.0	41.0
ReSim-FPN^T [41]	200	39.8	60.2	43.5	36.0	57.1	38.6	41.4	61.9	45.4	37.5	59.1	40.3
PixPro [10]	400	41.4	61.6	45.4	-	-	-	-	-	-	-	-	-
InsLoc [12]	400	42.0	62.3	45.8	37.6	59.0	40.5	43.3	63.6	47.3	38.8	60.9	41.7
DenseCL [11]	200	40.3	59.9	44.3	36.4	57.0	39.2	41.2	61.9	45.1	37.3	58.9	40.1
DetCon$_S$ [13]	1000	41.8	-	-	37.4	-	-	42.9	-	-	38.1	-	-
DetCon$_B$ [13]	1000	42.7	-	-	38.2	-	-	43.4	-	-	38.7	-	-
SoCo	100	42.3	62.5	46.5	37.6	59.1	40.5	43.2	63.3	47.3	38.8	60.6	41.9
SoCo	400	43.0	63.3	47.1	38.2	60.2	41.0	44.0	64.0	48.4	39.0	61.3	41.7
SoCo*	400	**43.2**	**63.5**	**47.4**	**38.4**	**60.2**	**41.4**	**44.3**	**64.6**	**48.9**	**39.6**	**61.8**	**42.5**

+1.8 mAP

Trends after NeurIPS’2020

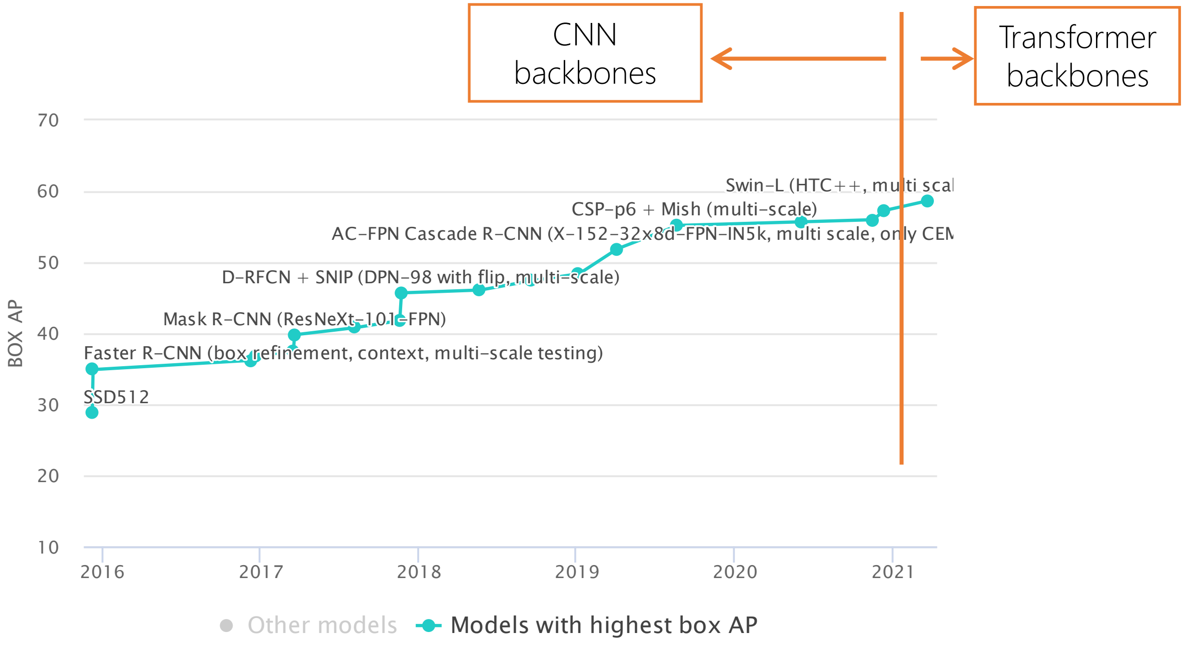
- ~~More study on BYOL why it does not collapse~~
  - ~~BYOL (Arxiv v3), SimSiam (CVPR’2021)~~
- Pre-training features which are good for down-stream tasks
  - Pixel-level pre-training
    - *PixPro*, DenseCL (CVPR’2021)
  - Object-level pre-training
    - SoCo (tech report)
- Self-supervised learning + Transformers
  - MoCo v3 (tech report), DINO (tech report)
  - SSL-Swin/MoBY (tech report)

SSL on Transformer?

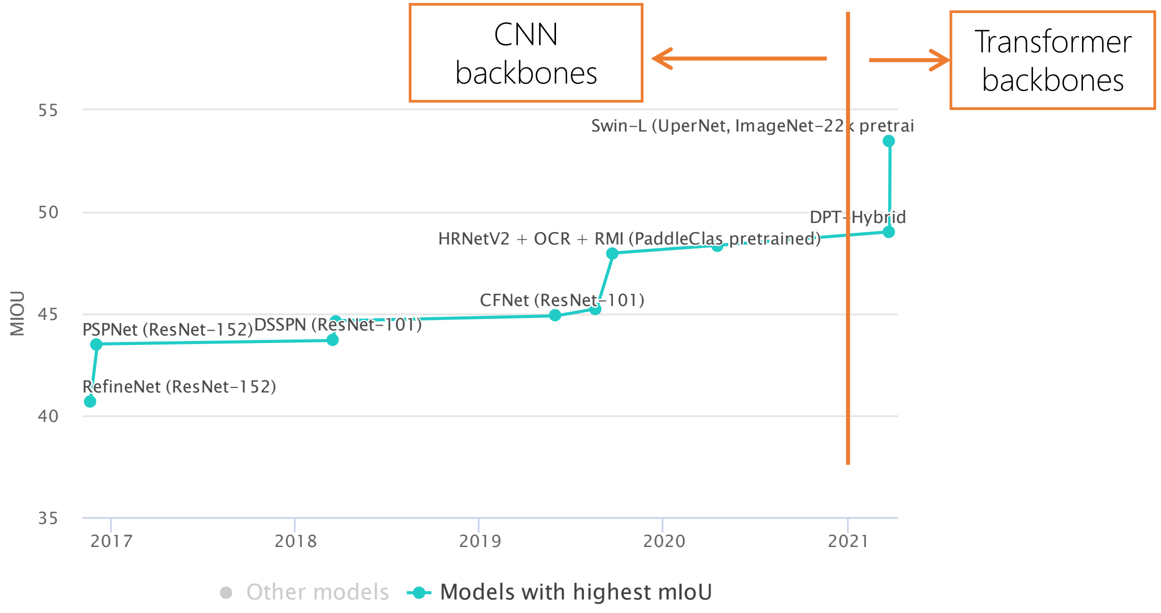


3400 stars

COCO object detection



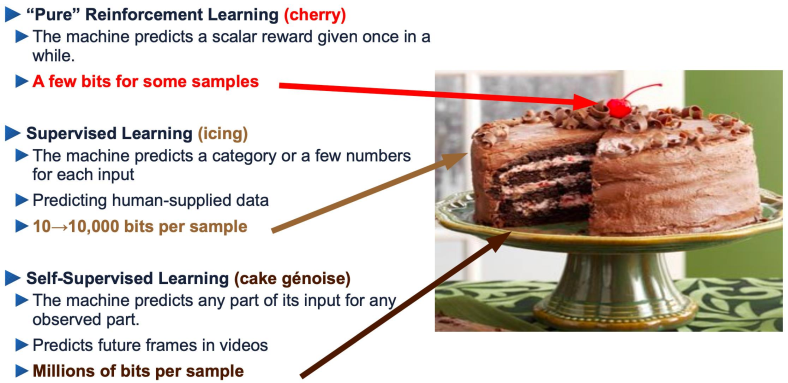
ADE20K semantic segmentation

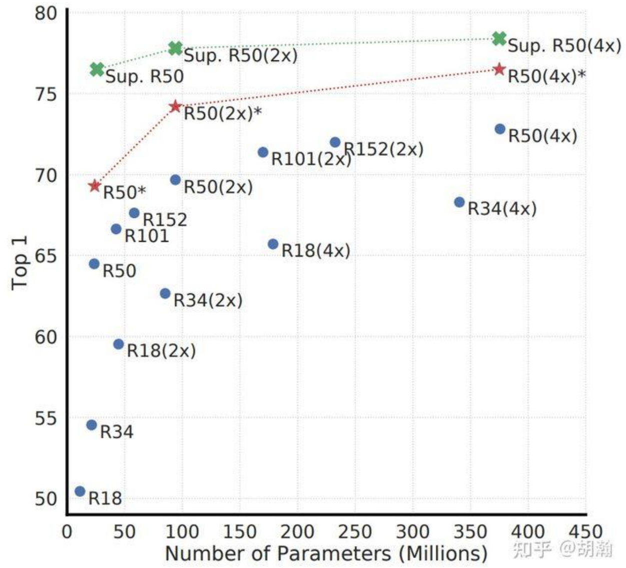


Evolving of state-of-the-art approaches for years

Self-supervised learning + Transformer

- "Golden combination"
  - SSL can better leverage the model capacity



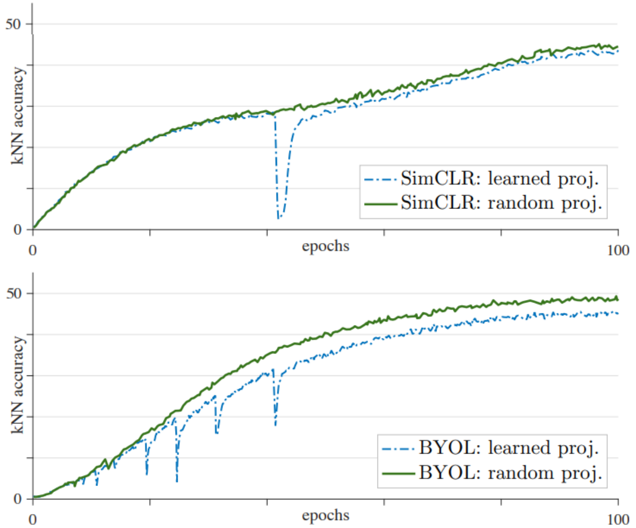


- Transformers has significantly stronger modeling power than CNN

https://www.zhihu.com/question/457507120

MoCo v3 (tech report, 2021/04)

- Transformer is difficult to be tamed for SSL
  - Fixed patch projection



	SimCLR	BYOL
learned patch proj.	69.3	69.7
random patch proj.	**70.1**	**71.0**

Figure 6. **Random *vs.* learned patch projection** (ViT-B/16, 100-epoch ImageNet, AdamW, batch 4096). **Top**: SimCLR: lr=2e-4, wd=0.1. **Bottom**: BYOL: lr=1e-4, wd=0.03.

DINO (tech report, 2021/05)

- Transformer is better at learn segmentation

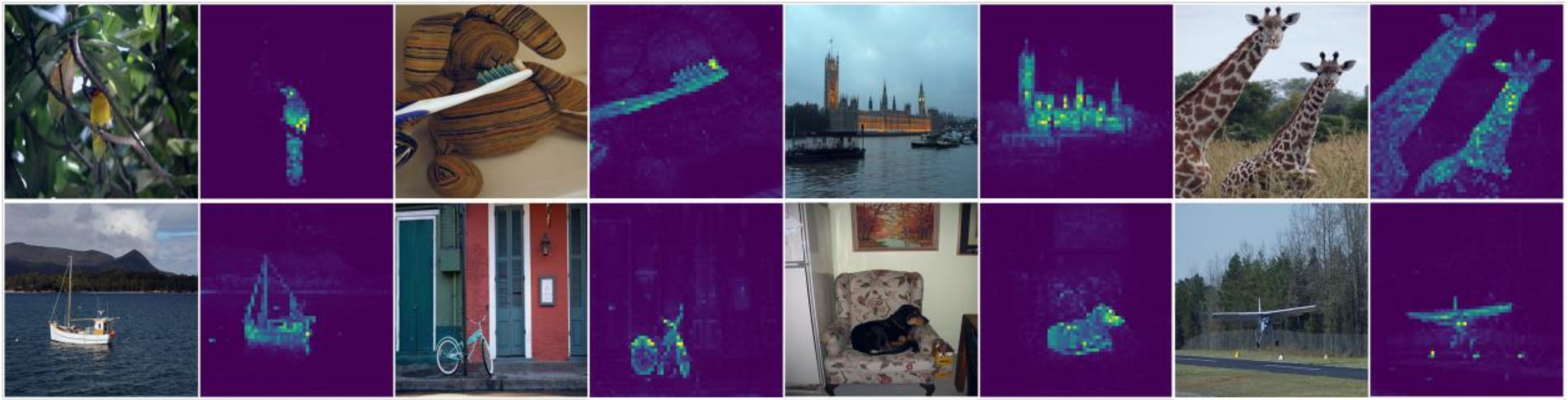
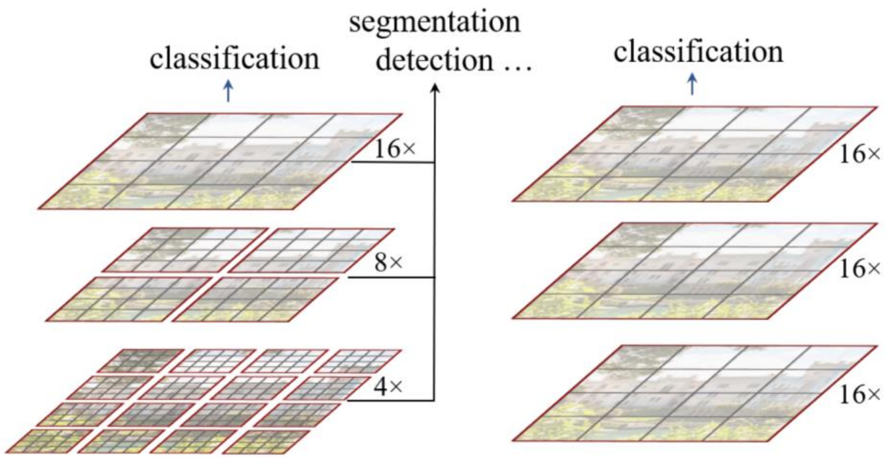


Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

SSL-Swin (MoBY)

- Provide baselines to evaluation transferring performance on down-stream tasks



(a) Swin Transformer (ours)

Used in MoBY

(b) ViT

Used in MoCo v3/DINO

- <u>No better than supervised approaches</u>

Method	Model	Schd.	box AP		
			mAP^{bbox}	AP_{50}^{bbox}	AP_{75}^{bbox}
Swin-T (mask R-CNN)	Sup.	1x	43.7	66.6	47.7
	MoBY	1x	43.6	66.2	47.7
	Sup.	3x	46.0	68.1	50.3
	MoBY	3x	46.0	67.8	50.6
Swin-T (Cascade mask R-CNN)	Sup.	1x	48.1	67.1	52.2
	MoBY	1x	48.1	67.1	52.1
	Sup.	3x	50.4	69.2	54.7
	MoBY	3x	50.2	68.8	54.7

COCO object detection

Method	Model	Schd.	mIoU
Swin-T (UPerNet)	Sup.	160K	44.51
	MoBY	160K	44.06
	Sup.†	160K	45.81
	MoBY†	160K	45.58

ADE20K semantic segmentation

SSL-Swin (MoBY)

- Higher accuracy than DINO/MoCo v3, with much fewer additional tricks

Method	Arch.	Epochs	Params (M)	FLOPs (G)	img/s	Top-1 acc (%)
Sup.	DeiT-S	300	22	4.6	940.4	79.8
Sup.	Swin-T	300	29	4.5	755.2	81.3
MoCo v3	DeiT-S	300	22	4.6	940.4	72.5
DINO	DeiT-S	300	22	4.6	940.4	72.5
DINO†	DeiT-S	300	22	4.6	940.4	75.9
MoBY	DeiT-S	300	22	4.6	940.4	72.8
MoBY	Swin-T	100	29	4.5	755.2	70.9
MoBY	Swin-T	300	29	4.5	755.2	**75.0**

+0.3 mAP vs. MoCo v3/DINO

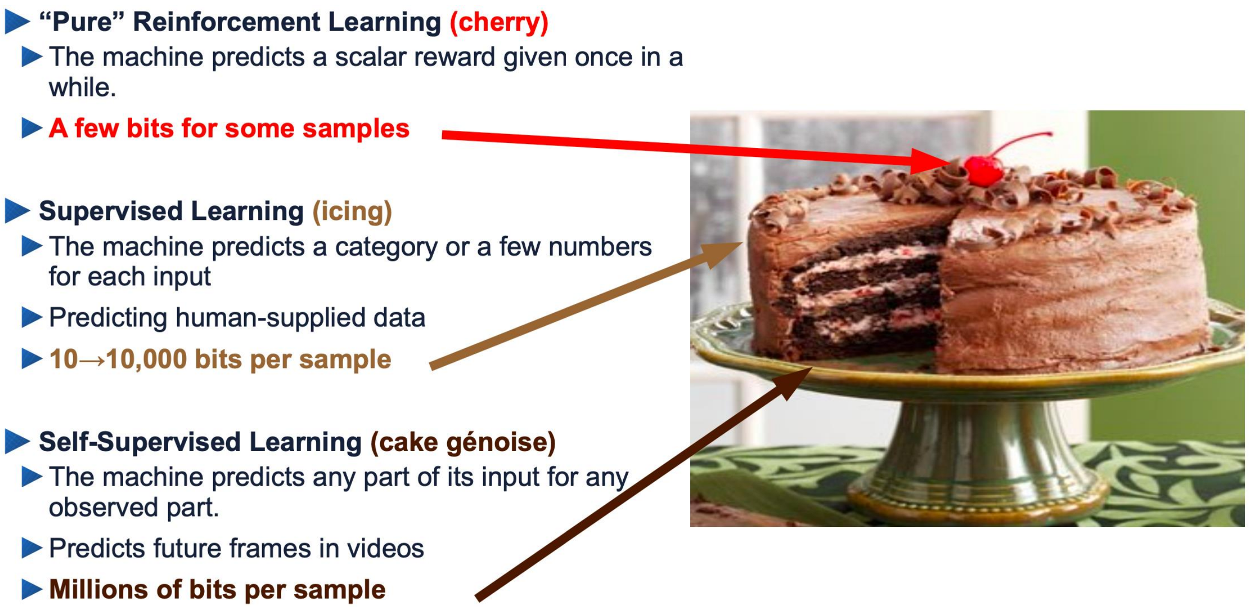
+2.2 mAP vs. DeiT

Table 1: Comparison of different SSL methods and different Transformer architectures on ImageNet-1K linear evaluation. † denotes DINO with a multi-crop scheme in training.

https://github.com/SwinTransformer/Transformer-SSL

Take-Home Message

- Enjoy the “cake”
- Two directions:
  - Aligning pre-training to down-stream tasks
  - SSL + Swin Transformers



Reference

[1] Yann LeCun. Self-Supervised Learning. AAAI 2020 Turing Talk https://drive.google.com/file/d/1r-mDL4IX_hzZLDBKp8_e8VZqD7fOzBkF/view?usp=sharing

[2] Kaiming He, et al. Momentum contrast for unsupervised visual representation learning. CVPR, 2020

[3] Andrew Zisserman. Self-Supervised Learning. 2018 https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf

[4] Alexey Dosovitskiy, et al. Discriminative unsupervised feature learning with convolutional neural networks. NIPS, 2014

[5] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. CVPR, 2018

[6] Bin Liu, Zhirong Wu, Han Hu and Stephen Lin. Deep Metric Transfer for Label Propagation with Limited Annotated Data. CVPRW, 2019

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. ICML, 2020

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. NeurIPS, 2020

[9] Jean-Bastien Grill, et al. Bootstrap your own latent: A new approach to self-supervised Learning. NeurIPS, 2020

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS, 2020

[11] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric Instance Classification for Unsupervised Visual Feature Learning. NeurIPS, 2020

[12] Xinlei Chen, Kaiming He. Exploring Simple Siamese Representation Learning. CVPR, 2021.

[13] Yuandong Tian, Xinlei Chen, Surya Ganguli. Understanding self-supervised Learning Dynamics without Contrastive Pairs. ICML, 2021

[14] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, Han Hu. Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning. CVPR, 2021

[14] Xinlong Wang, et al. Dense Contrastive Learning for Self-Supervised Visual Pre-Training. CVPR, 2021

[15] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, Stephen Lin. Aligning Pretraining for Detection via Object-Level Contrastive Learning. Tech report

[16] Alexey Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021

[17] Ze Liu, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Tech report, 2021

[18] Xinlei Chen, Saining Xie, Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. Tech report https://arxiv.org/abs/2104.02057

[19] Mathilde Caron et al. Emerging Properties in Self-Supervised Vision Transformers. Tech report https://arxiv.org/abs/2104.14294

[20] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, Han Hu. Self-Supervised Learning with Swin Transformers. Tech report https://arxiv.org/abs/2105.04553