Recent Progress on Self-Supervised Visual Representation Learning

Han Hu (胡瀚)
Principal Researcher
Visual Computing Group, Microsoft Research Asia (MSRA)
November 27^{th}, 2020

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 how to world works largely by observation

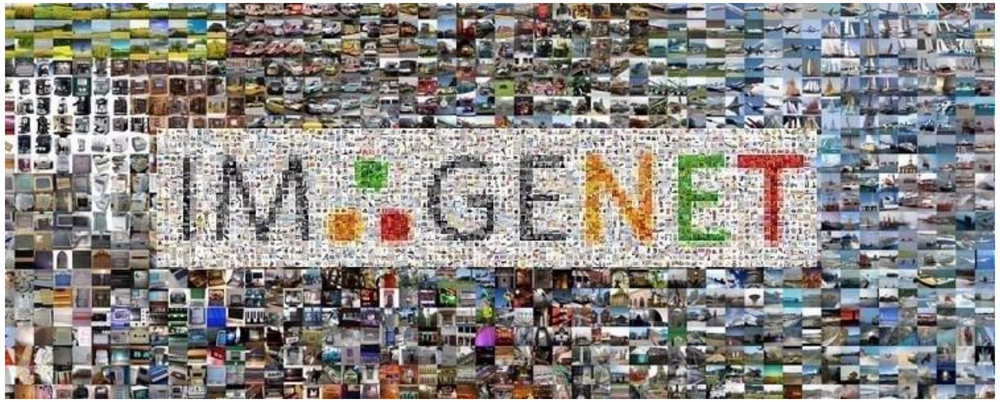
Photos courtesy of Emmanuel Dupoux

Credit by Yann LeCun

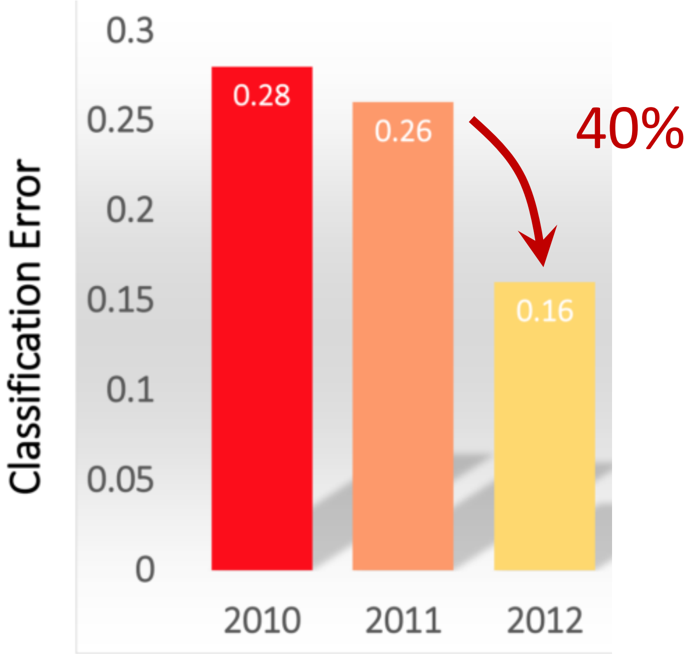
Linda Smith, Michael Gasser. The Development of Embodied Cognition: Six Lessons from Babies, 2005

A Story about ImageNet

- AlexNet (NIPS'2012)

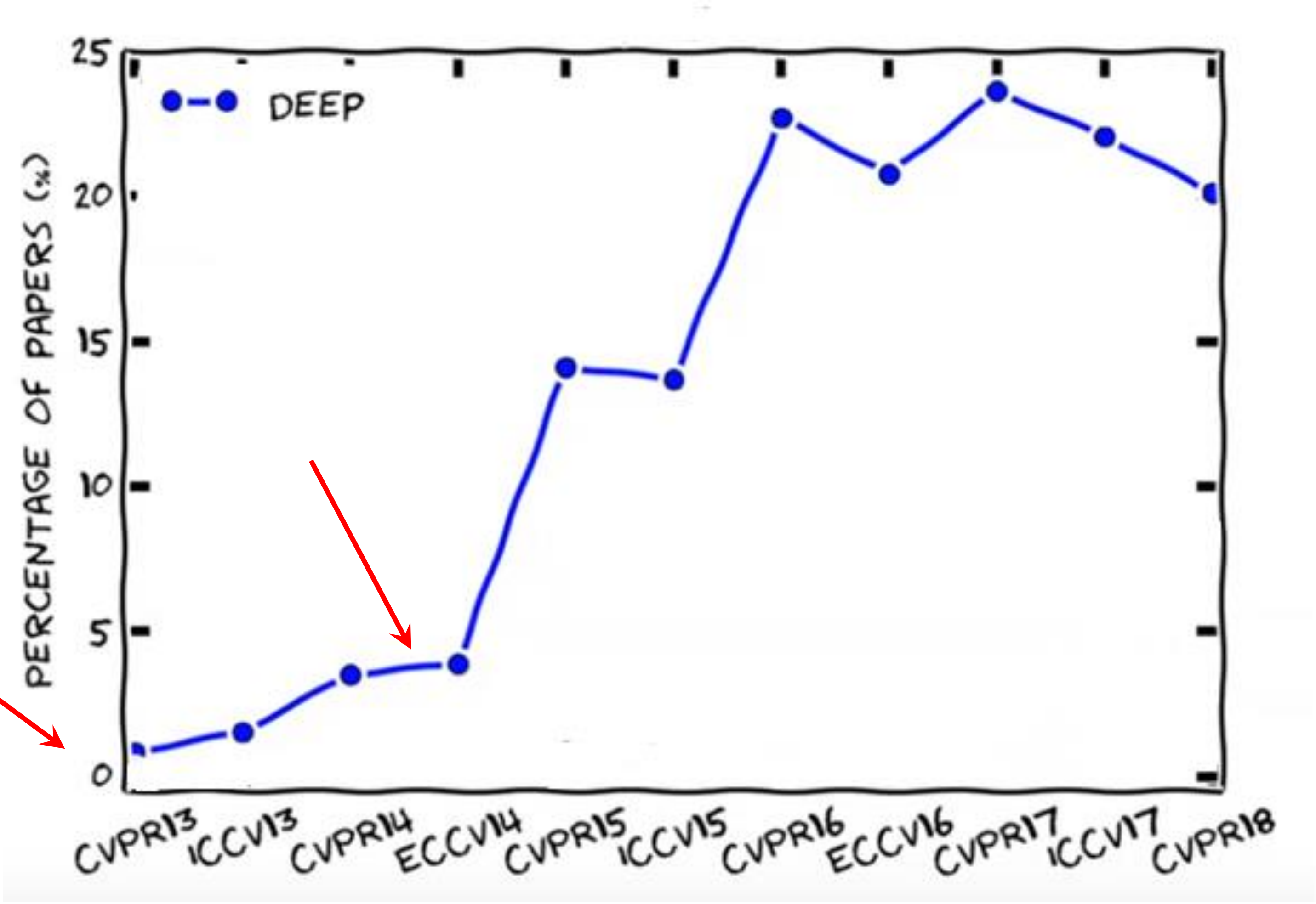


ImageNet Challenge



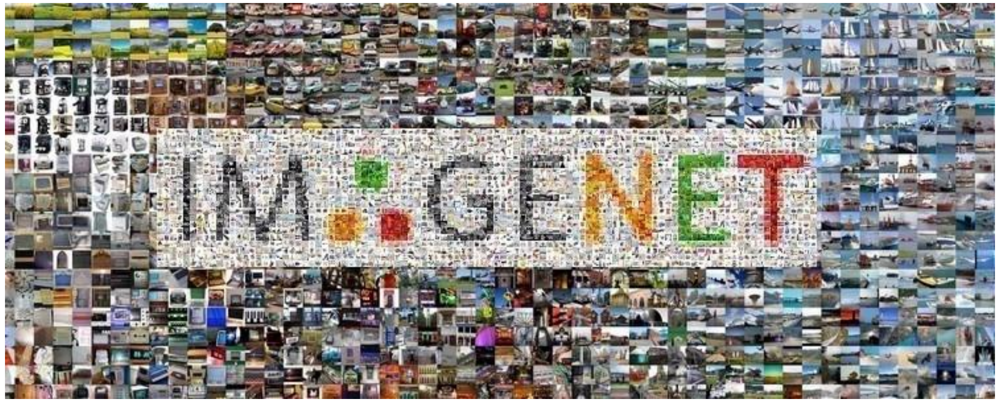
A Story about ImageNet

AlexNet



Supervised Pretraining + Finetuning (2014)

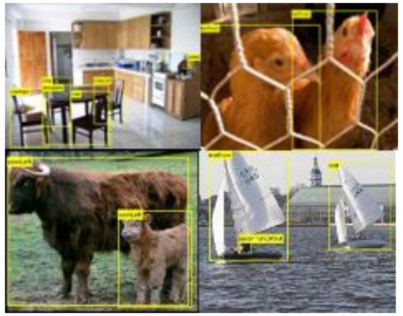
- A kind of transfer learning paradigm



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

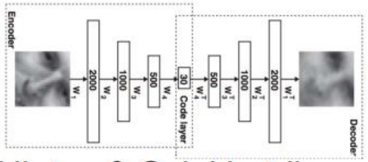
The Self-Supervised Learning Era!

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

How Did We Get Here?

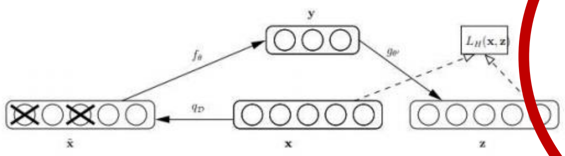
Credit mostly by Andrew Zisserman

Autoencoders



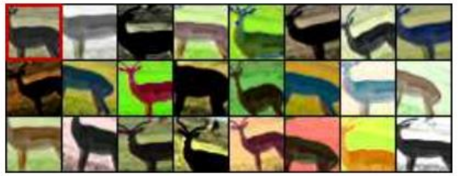
Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



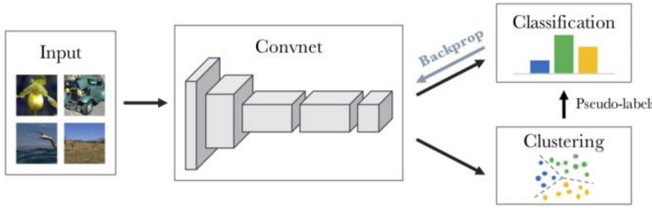
Vincent *et al.* ICML 2008.

Exemplar networks



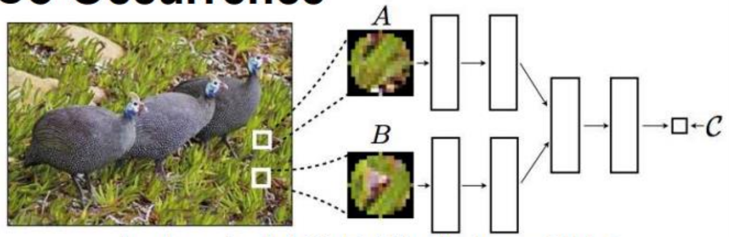
Dosovitskiy *et al.*, NIPS 2014

Deep Clustering



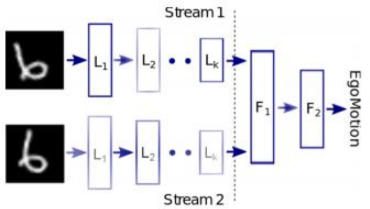
Caron et al, ECCV'2018

Co-Occurrence



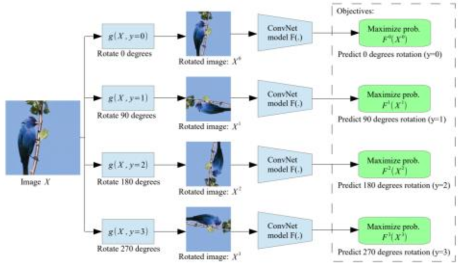
Isola *et al.* ICLR Workshop 2016.

Egomotion



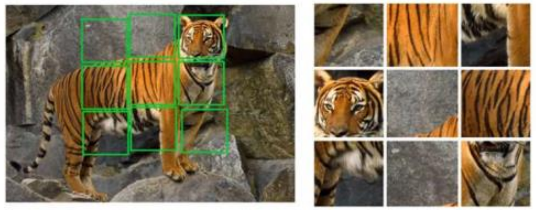
Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

Rotation Prediction



Gidaris et al, ICLR'2018

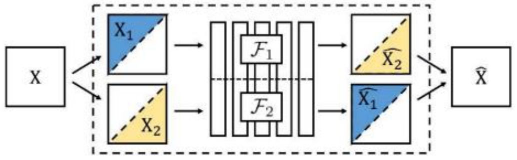
Context



Noroozi et al 2016

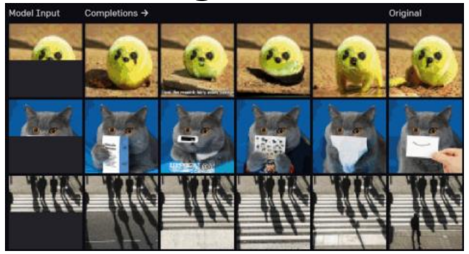
Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

Image GPT



Chen et al, ICML'2020

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

2019.11

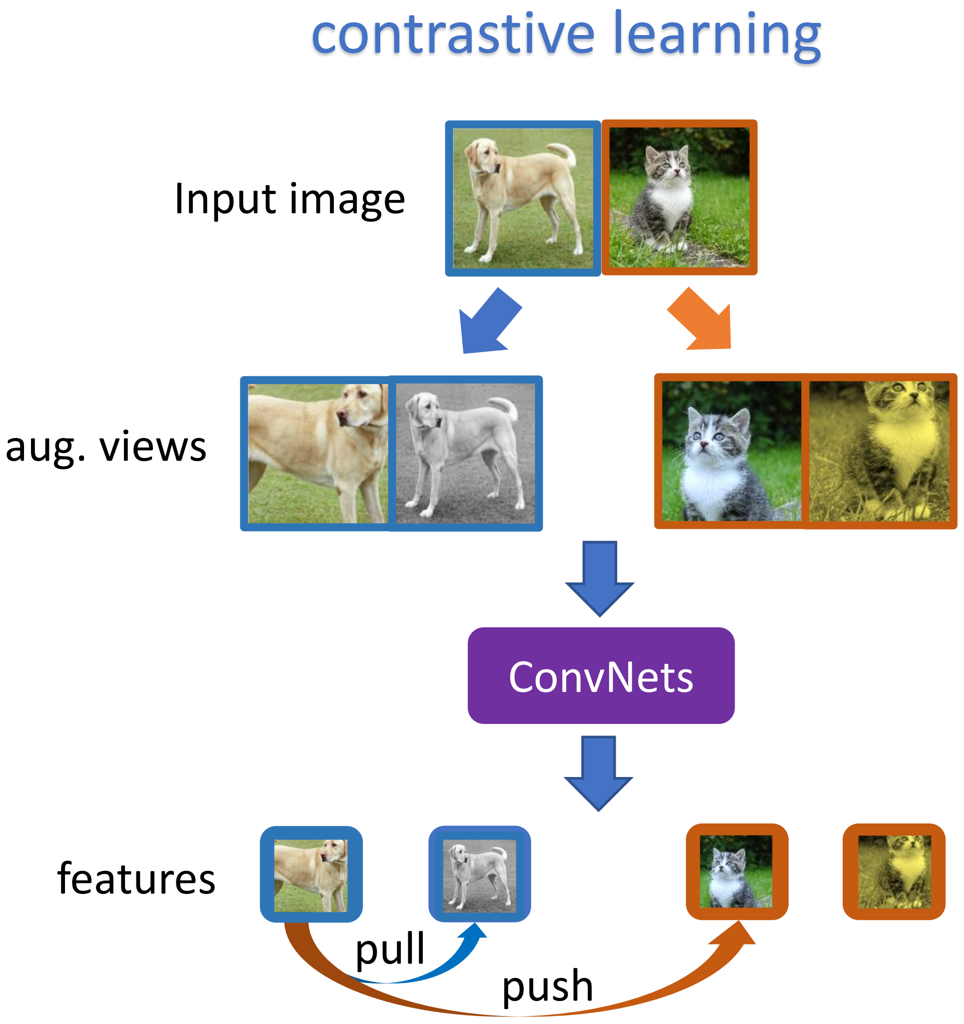
MoCo

FAIR

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

Contrastive Learning for Instance Discrimination

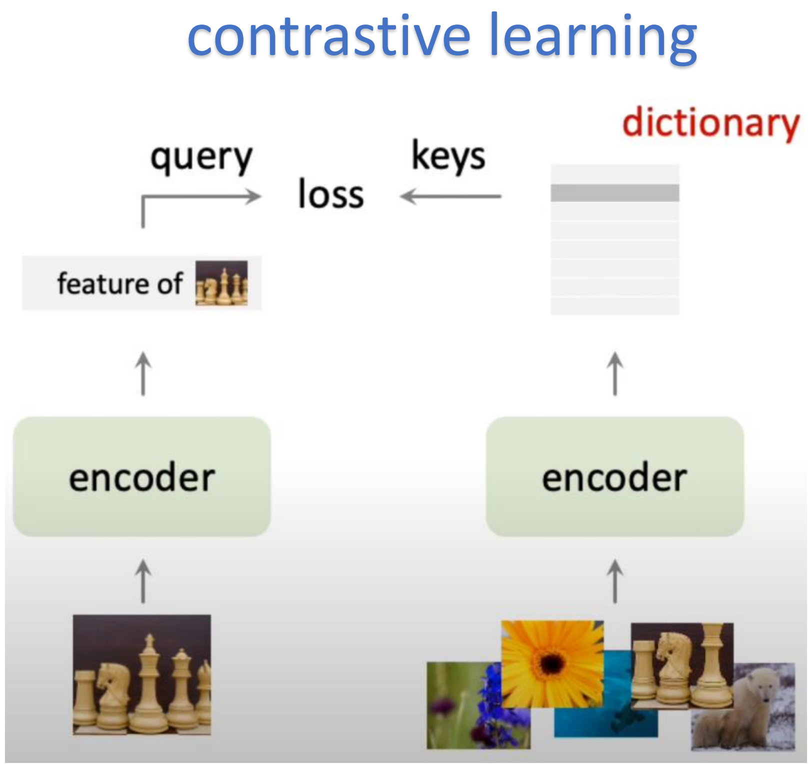
Pre-text task： Image discrimination

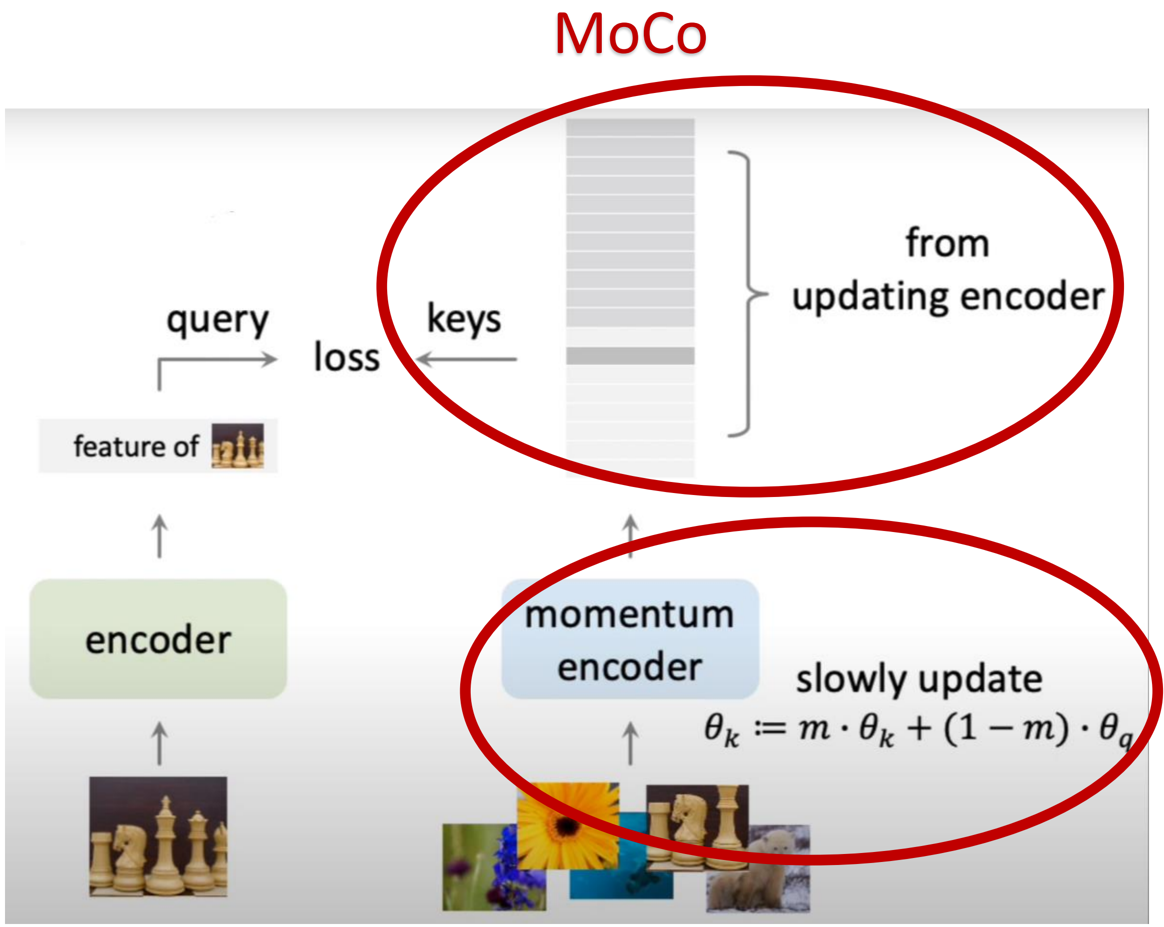


MoCo (CVPR'2020)

Credit by Kaiming He

- Large dictionary
- Consistent dictionary by momentum encoder

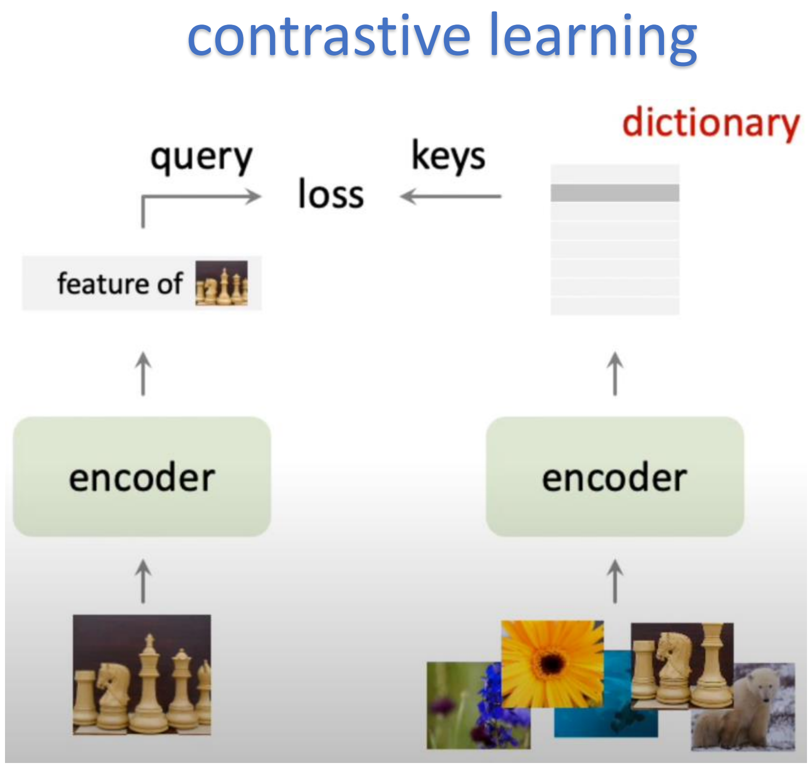


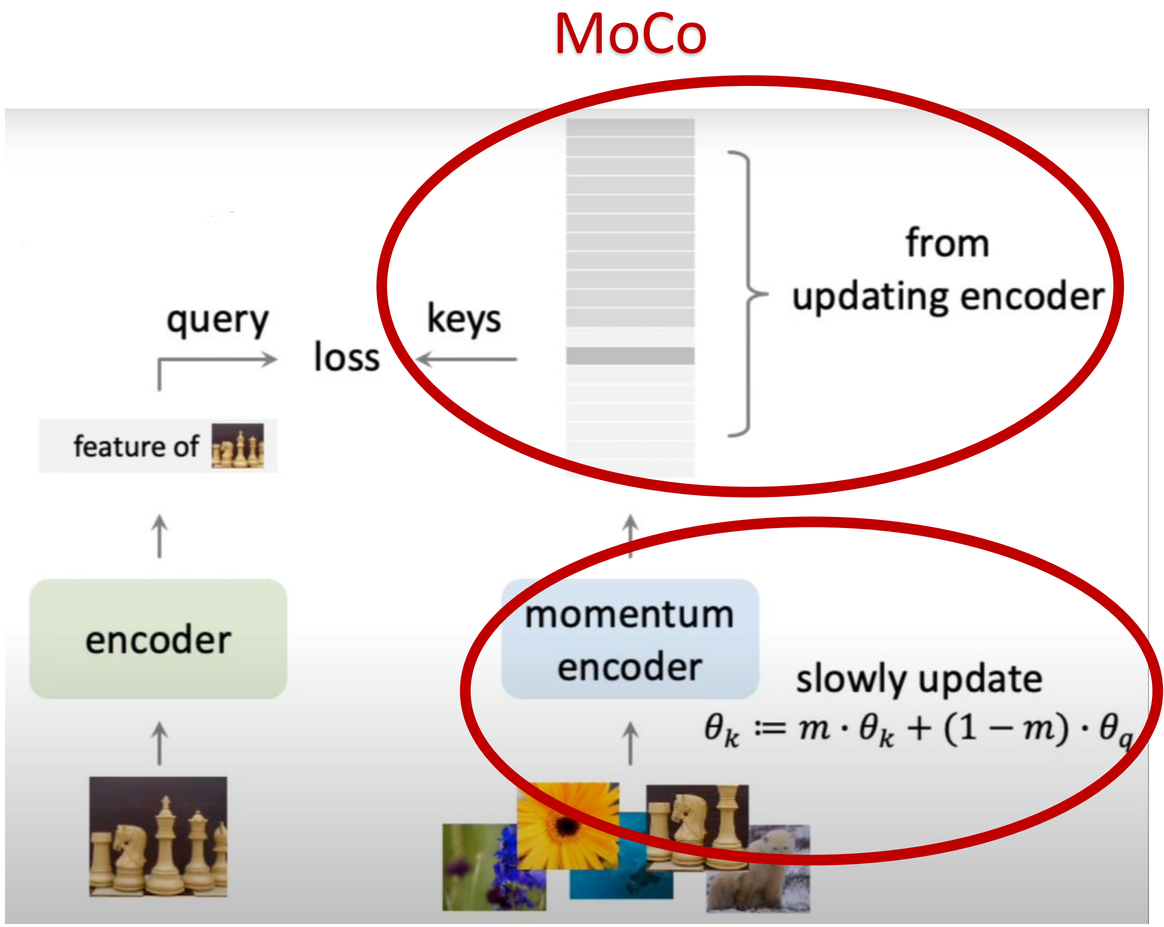


After MoCo

Credit by Kaiming He

- ~~Large dictionary~~
- ~~Consistent dictionary~~ by momentum encoder





MoCo Results

- Outperforms supervised methods on 7 down-stream tasks for the first time

pre-train	AP_{50}	AP	AP_{75}
random init.	60.2	33.8	33.1
super. IN-1M	81.3	53.5	58.8
MoCo IN-1M	81.5 (+0.2)	55.9 (+2.4)	62.6 (+3.8)
MoCo IG-1B	82.2 (+0.9)	57.2 (+3.7)	63.7 (+4.9)

(b) Faster R-CNN, R50-**C4**

Table 2. **Object detection fine-tuned on PASCAL VOC**

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	31.0	49.5	33.2	28.5	46.8	30.4
super. IN-1M	38.9	59.6	42.7	35.4	56.5	38.1
MoCo IN-1M	38.5 (−0.4)	58.9 (−0.7)	42.0 (−0.7)	35.1 (−0.3)	55.9 (−0.6)	37.7 (−0.4)
MoCo IG-1B	38.9 (0.0)	59.4 (−0.2)	42.3 (−0.4)	35.4 (0.0)	56.5 (0.0)	37.9 (−0.2)

(a) Mask R-CNN, R50-**FPN**, 1× schedule

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	26.4	44.0	27.8	29.3	46.9	30.8
super. IN-1M	38.2	58.2	41.2	33.3	54.7	35.2
MoCo IN-1M	38.5 (+0.3)	58.3 (+0.1)	41.6 (+0.4)	33.6 (+0.3)	54.8 (+0.1)	35.6 (+0.4)
MoCo IG-1B	39.1 (+0.9)	58.7 (+0.5)	42.2 (+1.0)	34.1 (+0.8)	55.4 (+0.7)	36.4 (+1.2)

(c) Mask R-CNN, R50-**C4**, 1× schedule

Table 5. **Object detection and instance segmentation fine-tuned on COCO**

pre-train	**COCO keypoint detection** AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
random init.	65.9	86.5	71.7
super. IN-1M	65.8	86.9	71.9
MoCo IN-1M	66.8 (+1.0)	87.4 (+0.5)	72.5 (+0.6)
MoCo IG-1B	66.9 (+1.1)	87.8 (+0.9)	73.0 (+1.1)

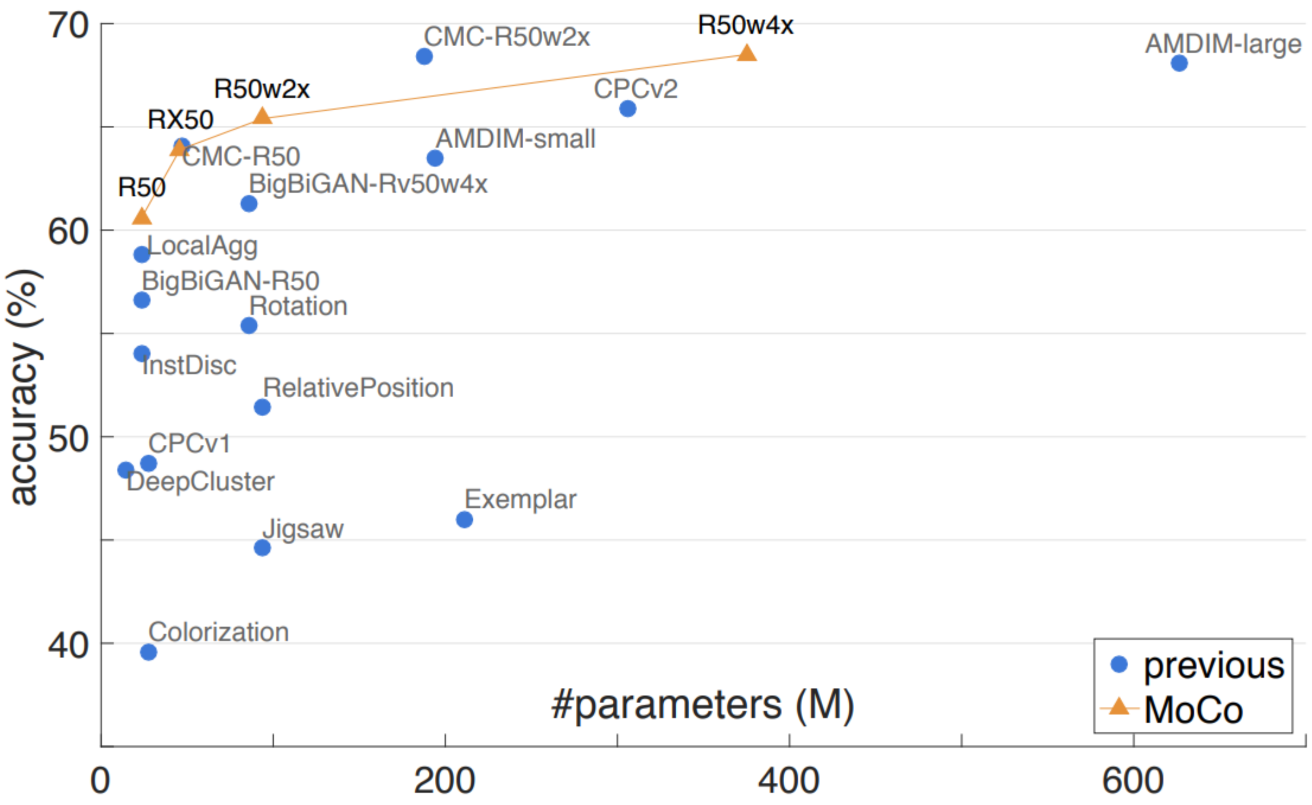
pre-train	**COCO dense pose estimation** AP^{dp}	AP^{dp}_{50}	AP^{dp}_{75}
random init.	39.4	78.5	35.1
super. IN-1M	48.3	85.6	50.6
MoCo IN-1M	50.1 (+1.8)	86.8 (+1.2)	53.9 (+3.3)
MoCo IG-1B	50.6 (+2.3)	87.0 (+1.4)	54.3 (+3.7)

pre-train	**LVIS v0.5 instance segmentation** AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	22.5	34.8	23.8
super. IN-1M†	24.4	37.8	25.8
MoCo IN-1M	24.1 (−0.3)	37.4 (−0.4)	25.5 (−0.3)
MoCo IG-1B	24.9 (+0.5)	38.2 (+0.4)	26.4 (+0.6)

pre-train	**Cityscapes instance seg.** AP^{mk}	AP^{mk}_{50}	**Semantic seg.** (mIoU) **Cityscapes**	**VOC**
random init.	25.4	51.1	65.3	39.5
super. IN-1M	32.9	59.6	74.6	74.4
MoCo IN-1M	32.3 (−0.6)	59.3 (−0.3)	75.3 (+0.7)	72.5 (−1.9)
MoCo IG-1B	32.9 (0.0)	60.3 (+0.7)	75.5 (+0.9)	73.6 (−0.8)

MoCo Results

- ImageNet-1K linear evaluation



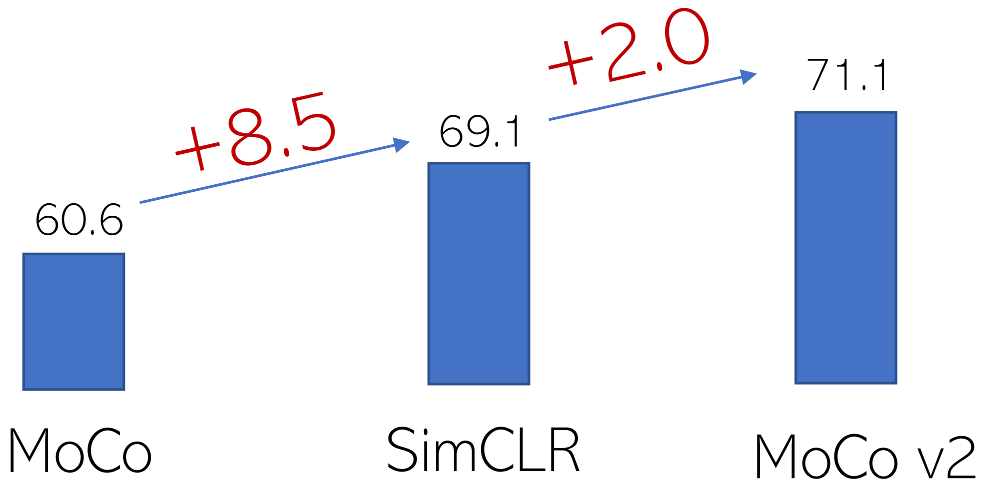
After MoCo

- SimCLR (ICML'2020)
- NeurIPS'2020 papers
- After 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
- Carefully tuning data augmentation methods

ImageNet linear evaluation



More Insights in SimCLR

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

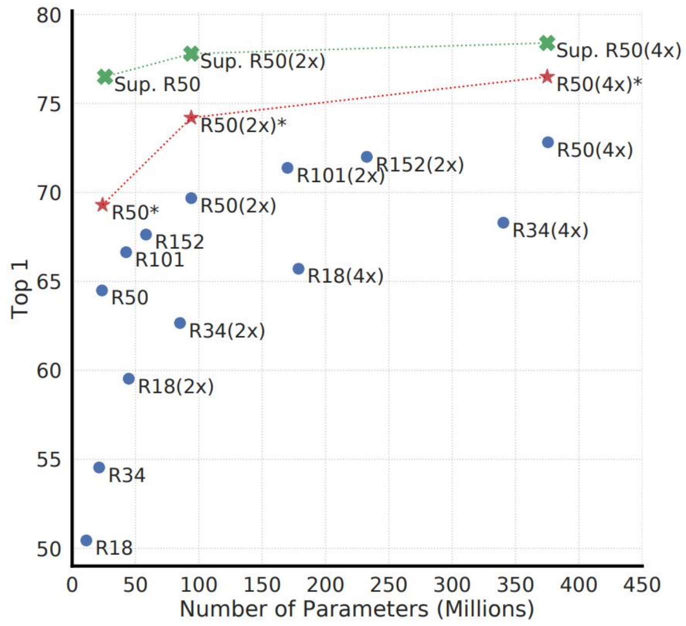


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs[7] (He et al., 2016).

Method	Architecture	Label fraction 1% (Top 5)	Label fraction 10% (Top 5)
Supervised baseline	ResNet-50	$\boxed{48.4}$	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	$\boxed{75.5}$	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

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

+27.1

Similar as that of GPT-3 in NLP!

SimCLR v2 (NeurIPS'2020)

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

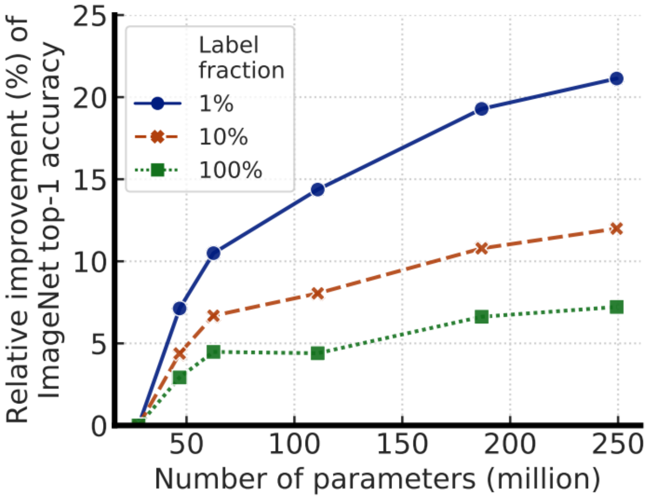


Figure 1: Bigger models yield larger gains when fine-tuning with fewer labeled examples.

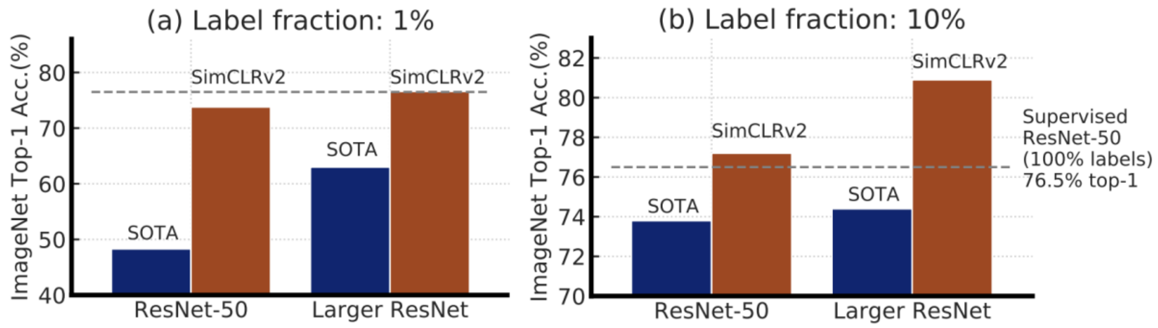


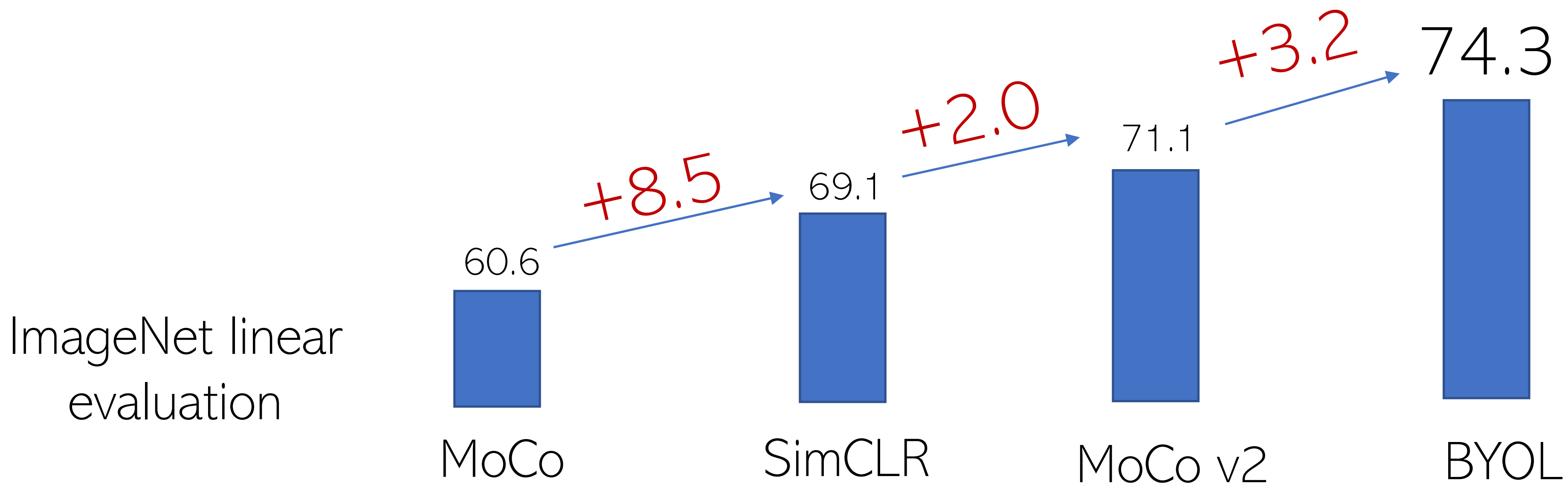
Figure 2: Top-1 accuracy of previous state-of-the-art (SOTA) methods [1, 2] and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50 trained with 100% of labels. Full comparisons in Table 3.

“Unsupervised” Papers on NeurIPS’2020

- 130 papers by a keyword “unsupervised” (totally about 1,900)
- Representative works
  - BOYL (DeepMind)
  - SwaV (Facebook AI Research)
  - InfoMin (MIT, Google Research)
  - ~~SimCLR v2 (Google Brain)~~
  - PIC (talk #4 by Zhenda Xie, MSRA)

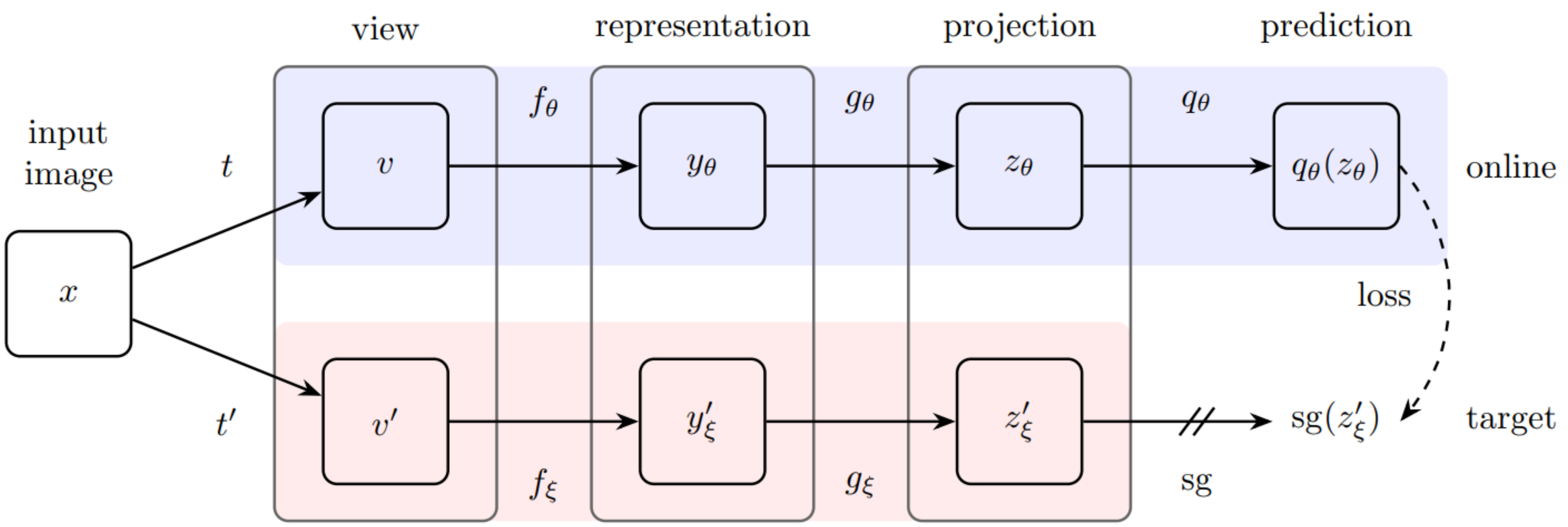
BYOL

- Bootstrap Your Own Latent



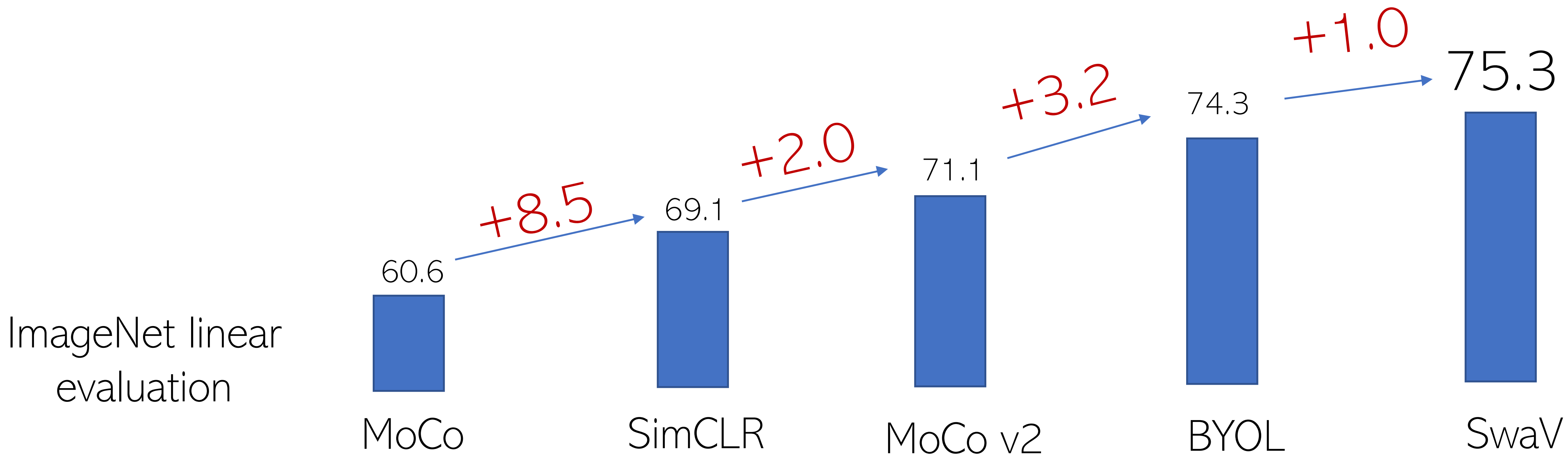
A Finding by BYOL

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



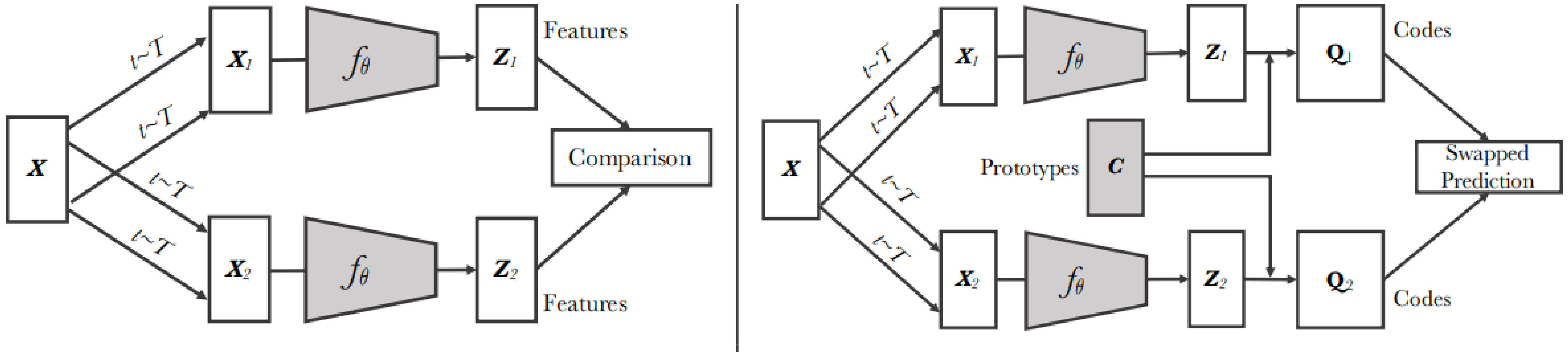
SwaV

- Contrasting Cluster Assignments



SwaV

- Deep clustering (ECCV'2018) + contrastive learning
- Additional small patches in view generation



Contrastive instance learning

Swapping Assignments between Views (Ours)

InfoMin: What Makes for Good Views for Contrastive Learning?

- Empirical study on augmentation methods
- Extensive/good results on Pascal VOC and COCO detection
  - Previous methods mostly focus on improving ImageNet linear evaluation accuracy

pre-train	AP_{50}	AP	AP_{75}	ImageNet Acc(%)
random init.*	60.2	33.8	33.1	-
supervised*	81.3	53.5	58.8	76.1
InstDis	80.9	55.2	61.2	59.5
PIRL	81.0	55.5	61.3	61.7
MoCo*	81.5	55.9	62.6	60.6
InfoMin Aug. (ours)	**82.7**	**57.6**	**64.6**	**70.1**

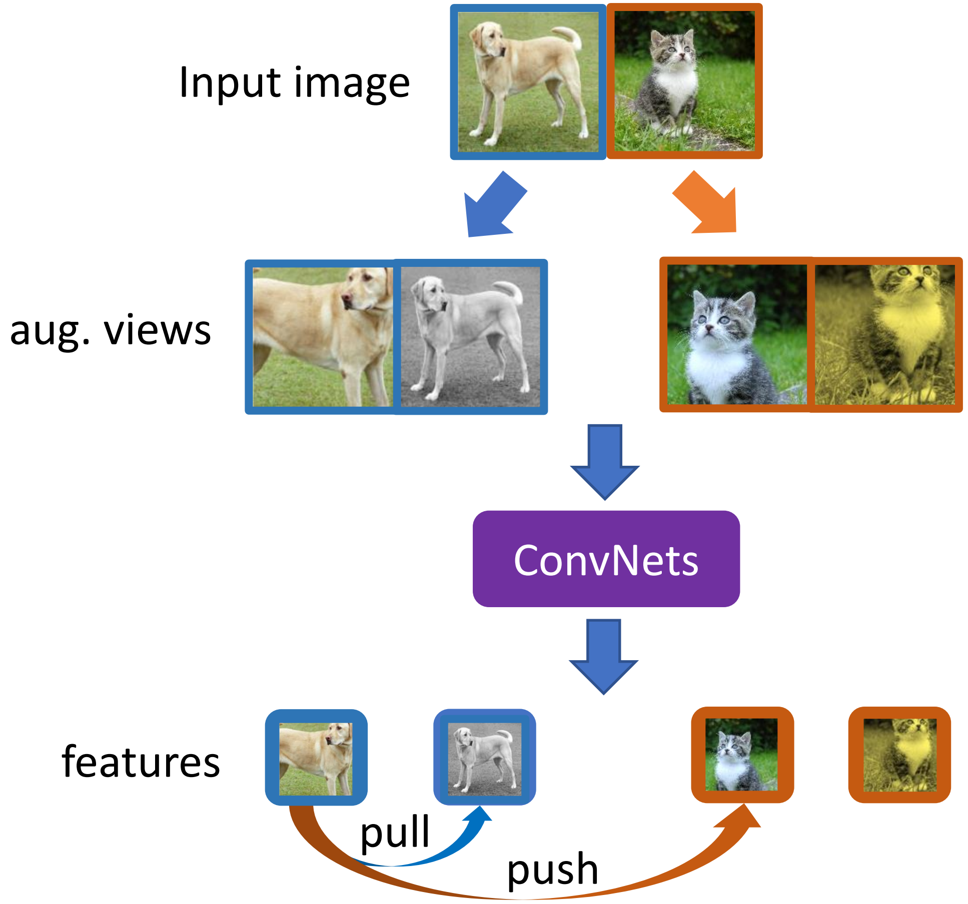
Pascal VOC object detection

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
random init*	26.4	44.0	27.8
supervised*	38.2	58.2	41.2
MoCo*	38.5($\uparrow$0.3)	58.3($\uparrow$0.1)	41.6($\uparrow$0.4)
InfoMin Aug.	39.0($\uparrow$0.8)	58.5($\uparrow$0.3)	42.0($\uparrow$0.8)

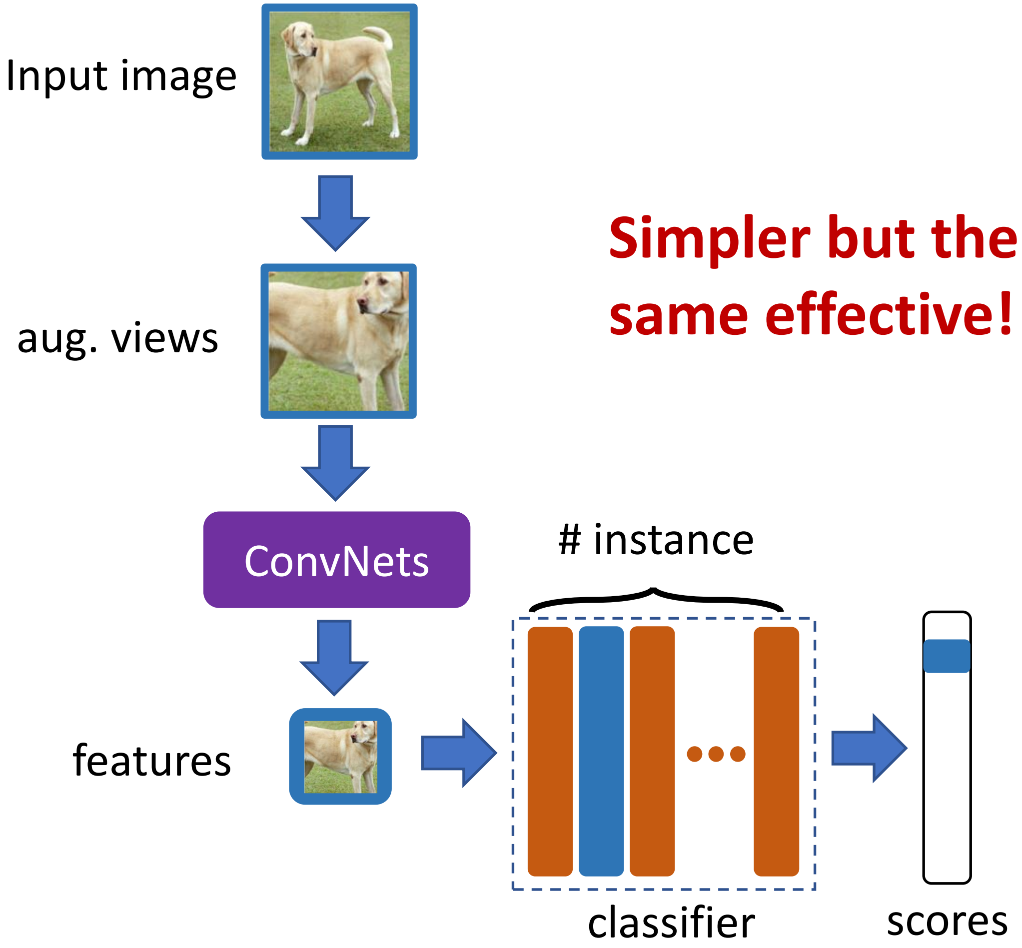
COCO object detection

PIC: a Single-Branch Method (Talk #4)

two-branch methods
(almost all previous methods)



one-branch method (PIC)



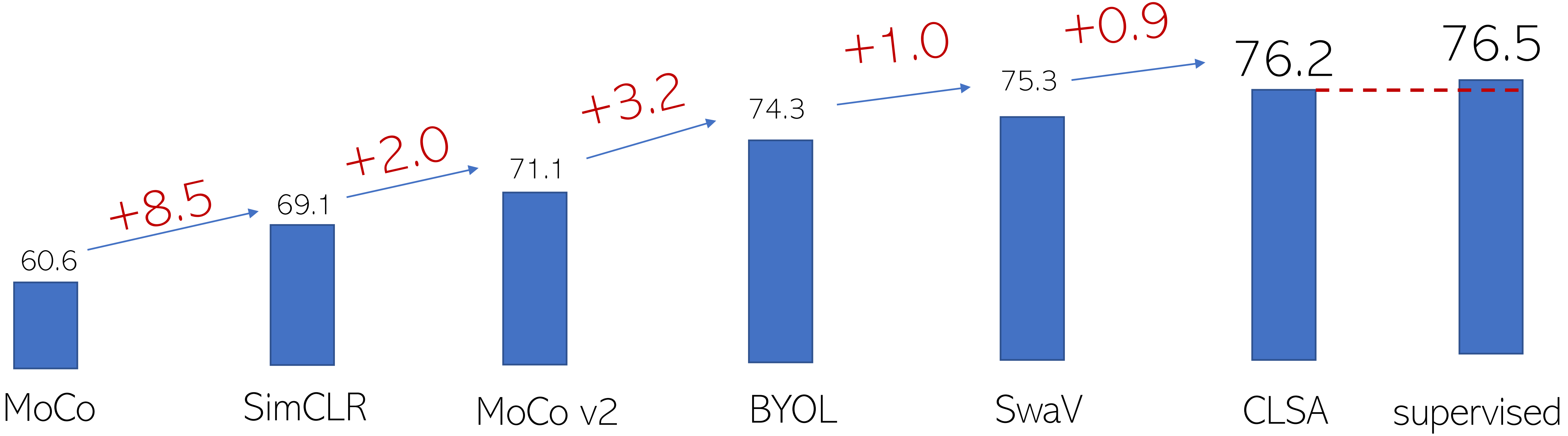
Simpler but the same effective!

Representative Works after NeurIPS'2020

- Higher ImageNet-1K linear evaluation accuracy
  - Contrastive learning with stronger augmentations (CLSA)
  - (ICLR'2021 submission) CLSA 76.2 vs supervised 76.5
- Better understanding
  - What makes instance discrimination good for transfer learning?
  - (ICLR'2021 submission) it is mainly the low-level features that effect!
- More study on BYOL why it does not collapse
  - BYOL (Arxiv v3)
  - Exploring Simple Siamese Representation Learning (CVPR'2021 submission)
- Pixel-level pretext tasks
  - *PixPro*, for more spatially fine-grained representation learning

Motivation of PixPro

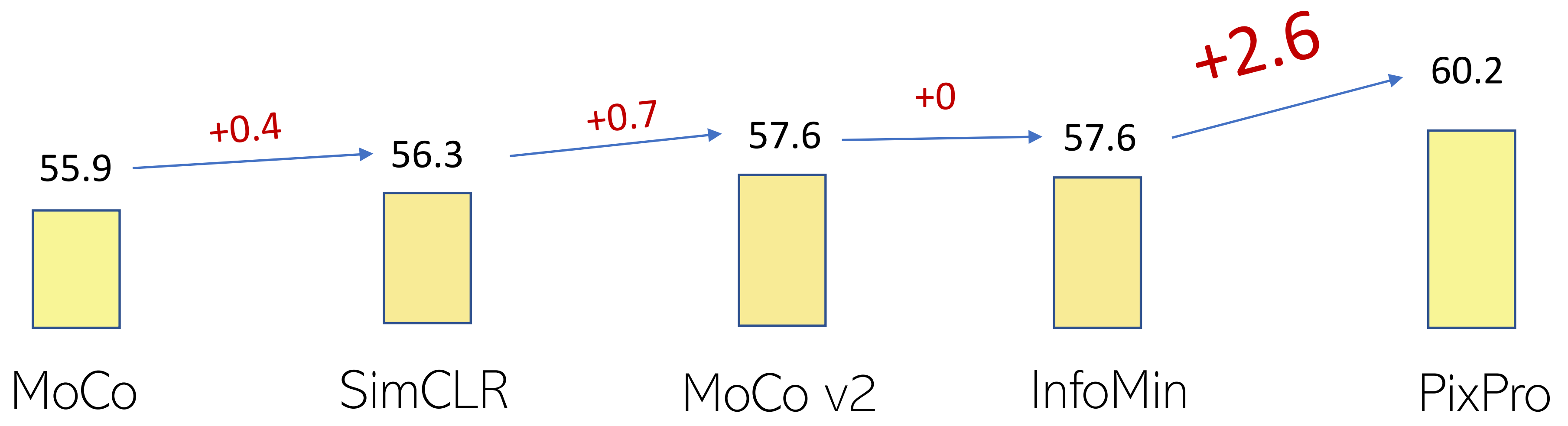
- Improvements on ImageNet-1K linear evaluation



Totally 15.6% absolute improvements in 1 year!

PixPro

- Improvements on Pascal VOC object detection (C4)
- Zhenda Xie et al. *Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning.* Tech Report 2020



Totally 1.1% absolute improvements in 1 year!

PixPro Results

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

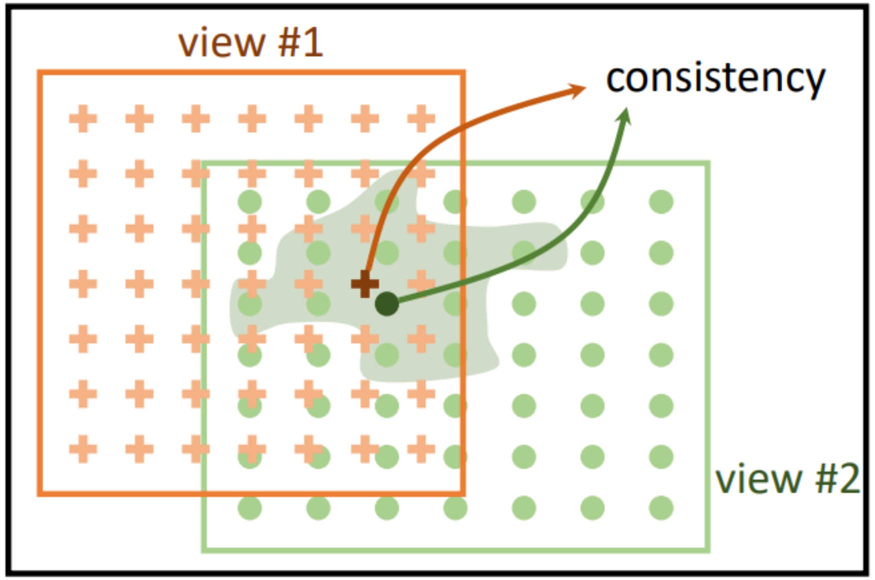
Method	#. Epoch	Pascal VOC (R50-C4)			COCO (R50-FPN)			COCO (R50-C4)			Cityscapes (R50)
		AP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mIoU
scratch	-	33.8	60.2	33.1	32.8	51.0	35.3	26.4	44.0	27.8	65.3
supervised	100	53.5	81.3	58.8	39.7	59.5	43.3	38.2	58.2	41.2	74.6
MoCo [18]	200	55.9	81.5	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

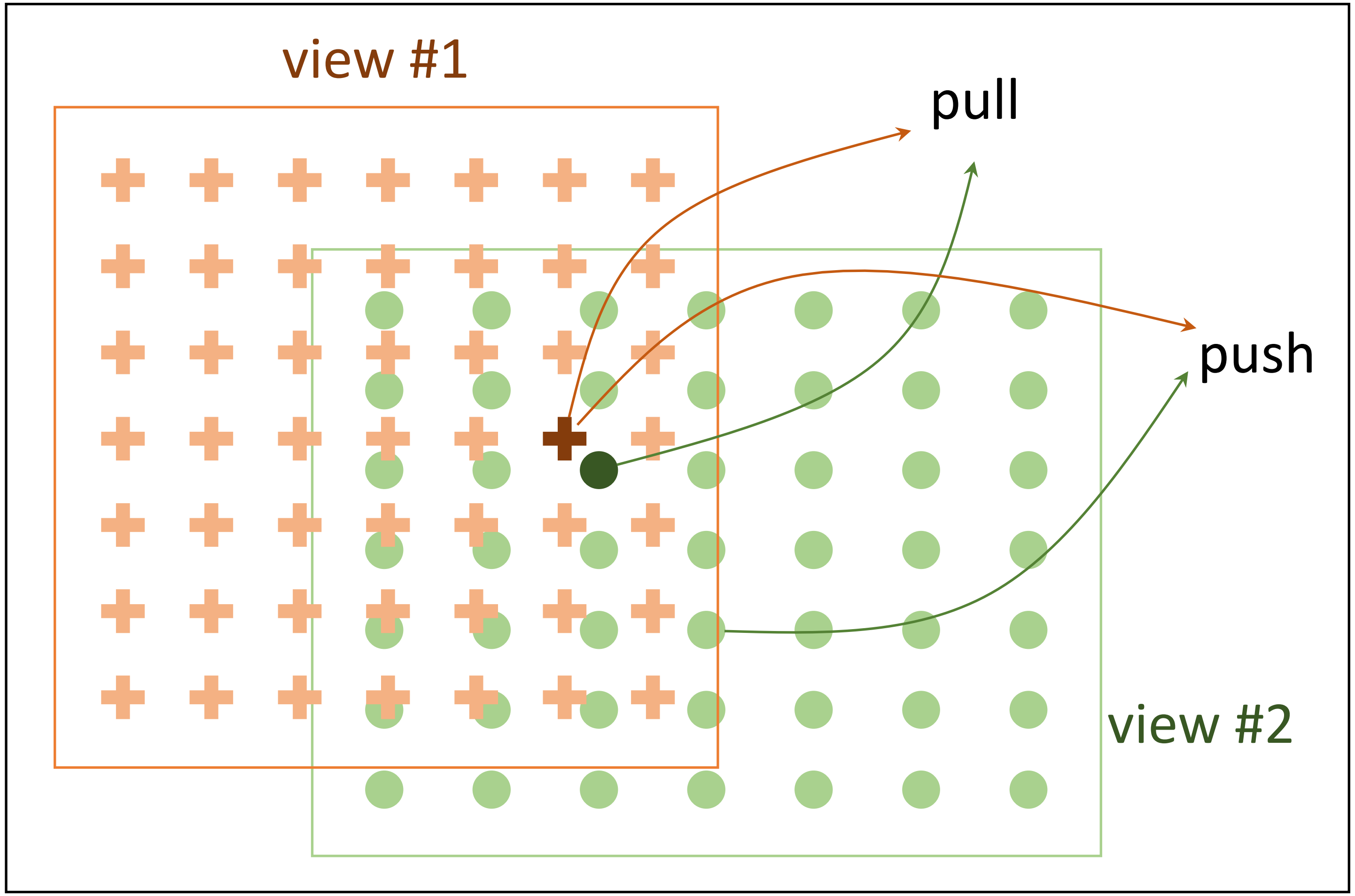
Previous pre-text tasks： instance discrimination



pixel-level pretext task

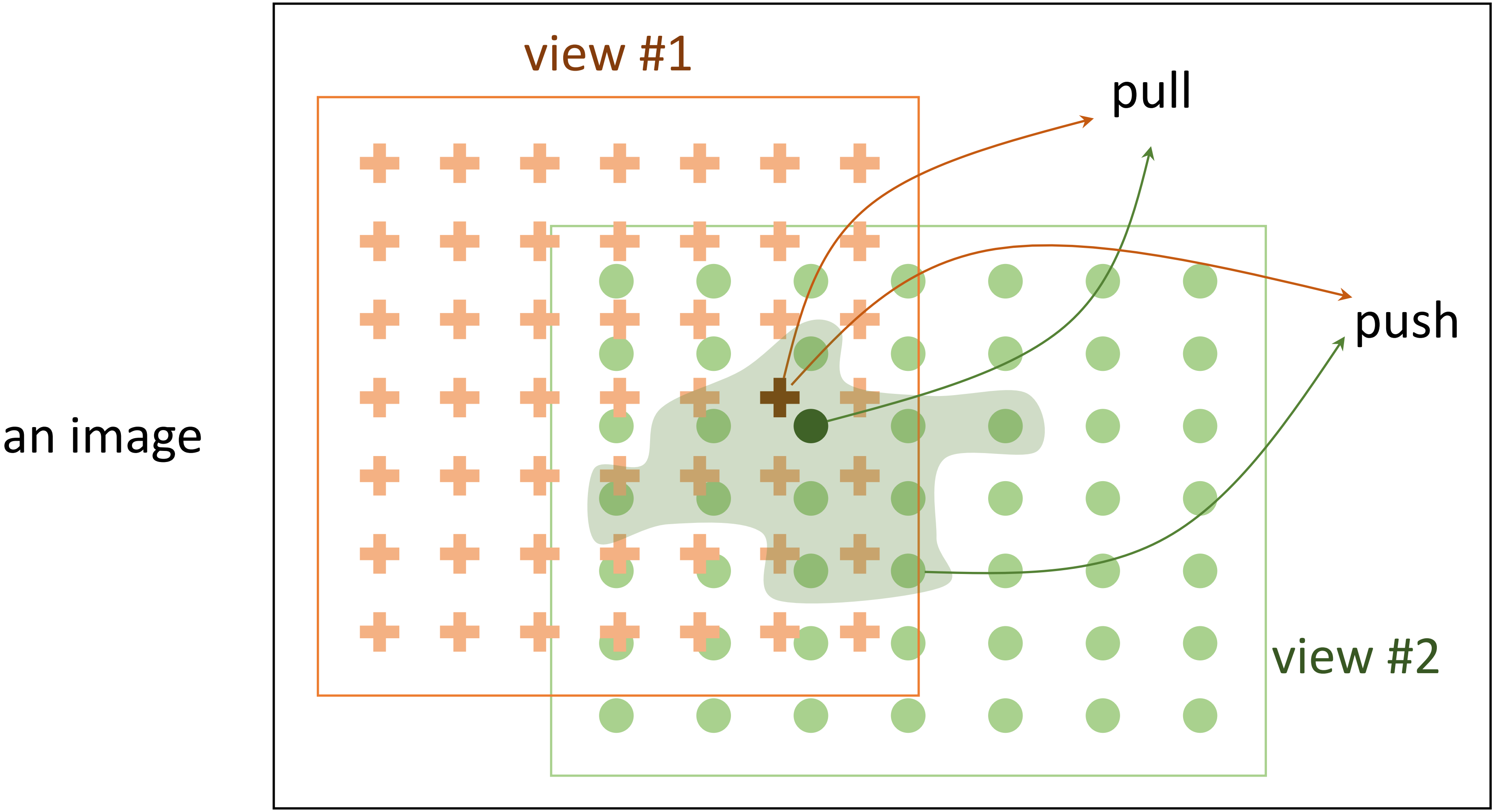
Pixel-Level Contrastive Learning

an image



pixel discrimination

Pixel-to-Propagation Consistency



Pixel-to-propagation consistency

Pixel-to-Propagation Consistency

- Pixel contrast: spatial sensitivity
- Propagation: spatial smoothness

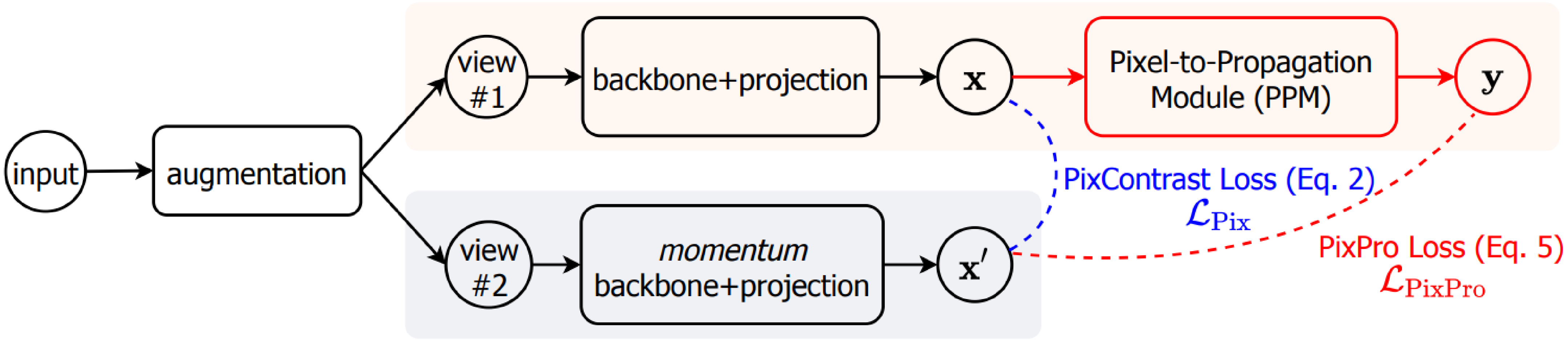
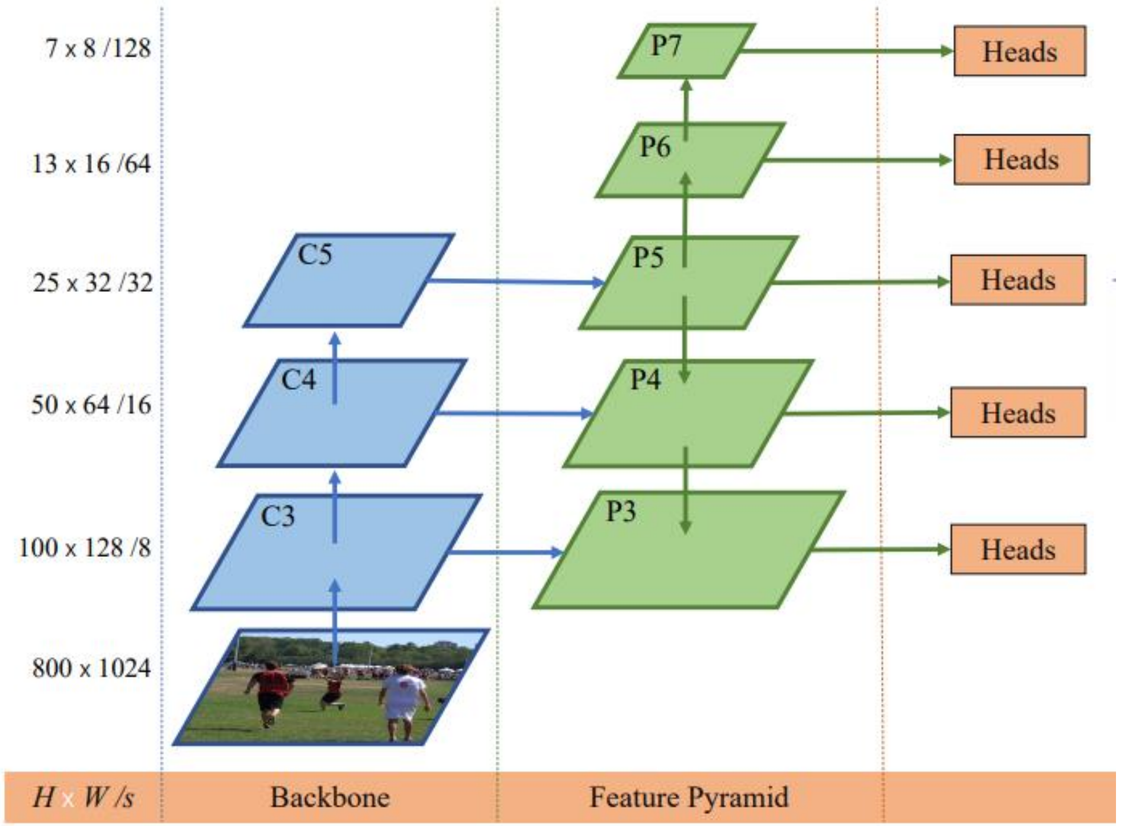


Figure 2. Architecture of the *PixContrast* and *PixPro* methods.

Aligning Pre-Training to Downstream Networks

- Using the same architecture as in downstream tasks

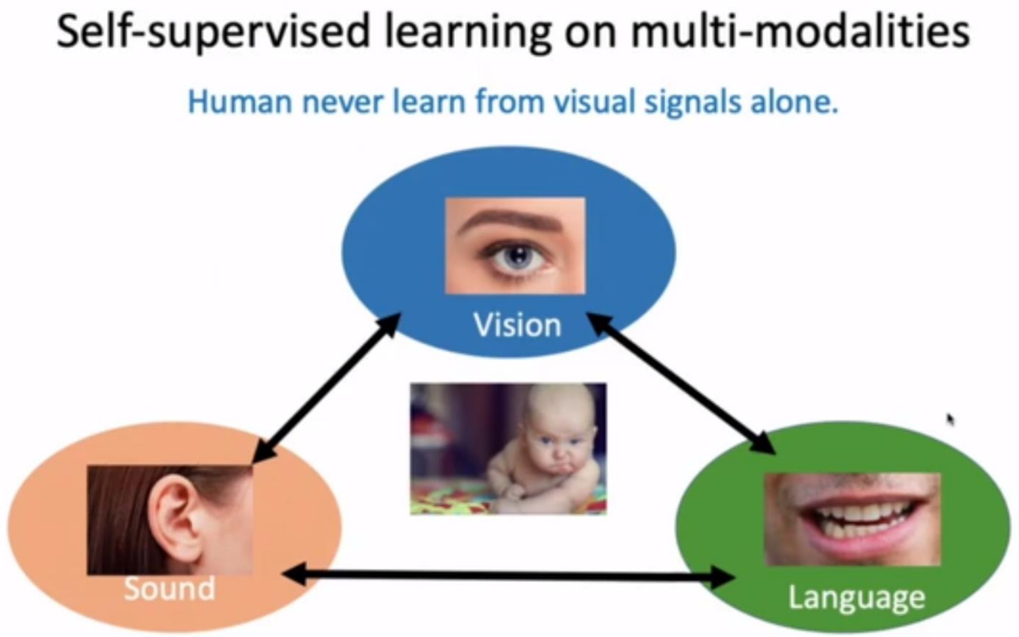


An architecture in FCOS detector

Beyond Image-based Unsupervised Pre-training

- Video based pre-training
    - Representative researchers
        - Andrew Zisserman, Weidi Xie, Xiaolong Wang, Alexei Efros et al

- Multi-modality pre-training



Take-Home Message

- Enjoy the "cake"

- **"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**