DeepMind

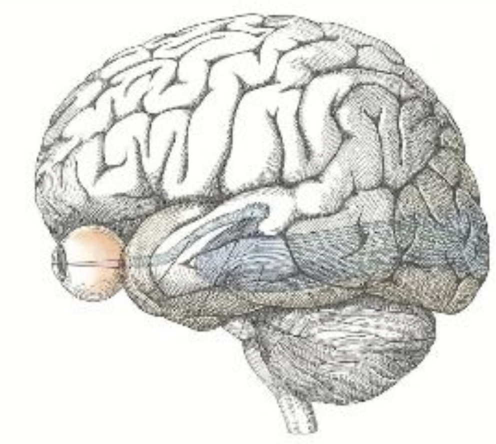
The virtuous cycle of object discovery and representation learning

Olivier Hénaff

ECCV 2022 Workshop on Self-Supervised Representation Learning in Computer Vision

Principles of biological and artificial intelligence

Efficient generalization



Neural representations

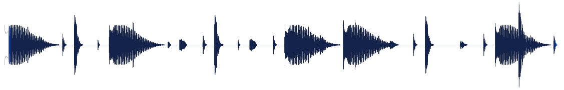
- Enable intelligent behavior

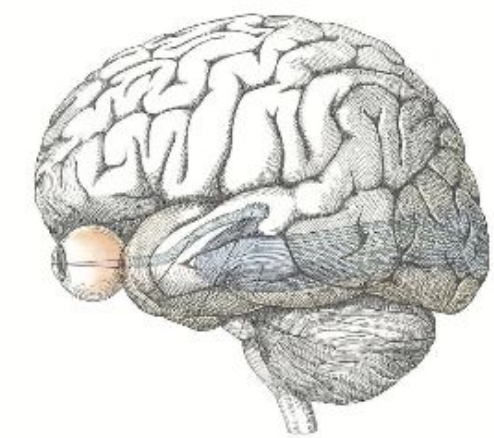
Principles of biological and artificial intelligence

Self-supervised pretraining

Efficient generalization

But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.





Neural representations

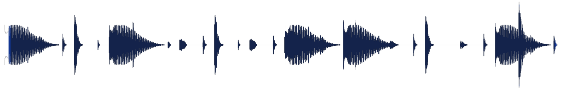
- Enable intelligent behavior
- Require minimal supervision
- Are generally applicable

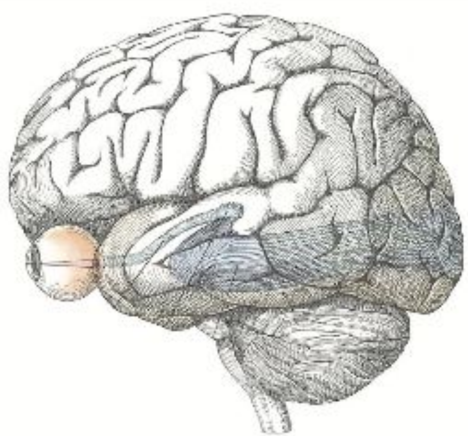
Principles of biological and artificial intelligence

Self-supervised pretraining

Efficient generalization

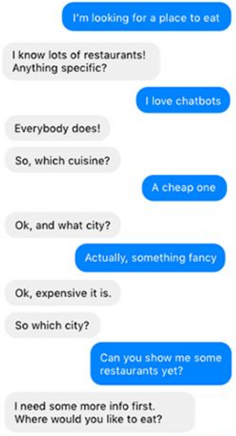
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.





Neural representations

- Enable intelligent behavior
- Require minimal supervision
- Are generally applicable

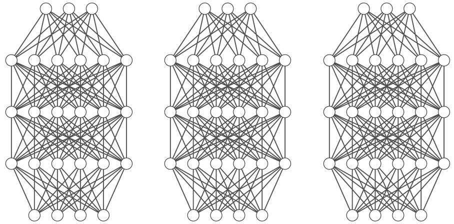


Principles of biological and artificial intelligence

Self-supervised pretraining

But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

“cat” “dog” “cycle”

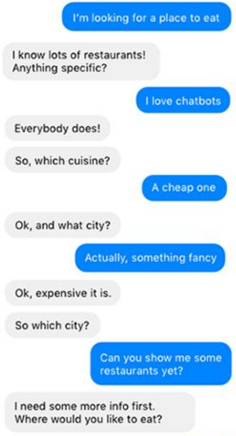


Supervised ConvNet

Neural representations

- Enable intelligent behavior
- Require minimal supervision
- Are generally applicable

Efficient generalization



Principles of biological and artificial intelligence

Self-supervised pretraining

But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

image colorization
(Zhang, 2016)

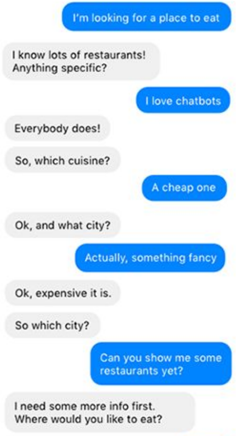
Neural representations

- Enable intelligent behavior

- Require minimal supervision

- Are generally applicable

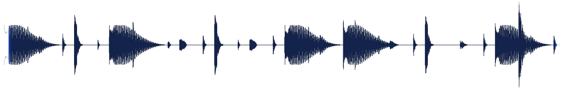
Efficient generalization

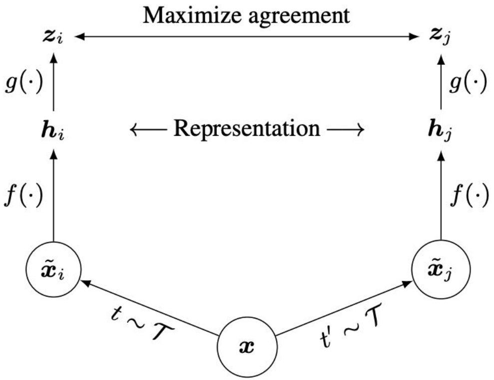


Principles of biological and artificial intelligence

Self-supervised pretraining

But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.



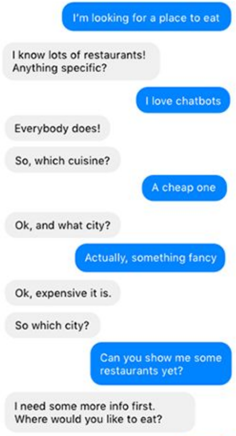


Contrastive learning (Chen, 2020)

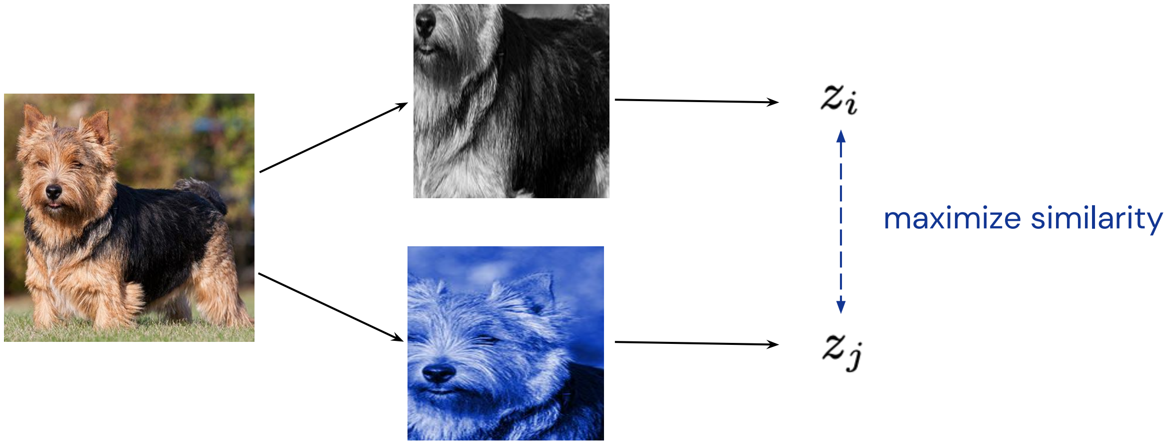
Neural representations

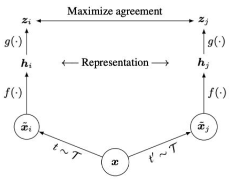
- Enable intelligent behavior
- Require minimal supervision
- Are generally applicable

Efficient generalization

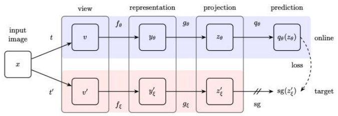


Is the current self-supervised paradigm too simple?

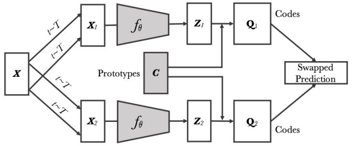




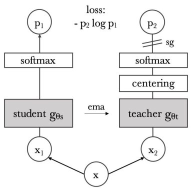
SimCLR



BYOL



SwAV



DINO

Is the current self-supervised paradigm too simple?

Real-world data is complex

- multiple objects in natural scenes
- multiple speakers in natural speech
- multiple scenes in natural videos



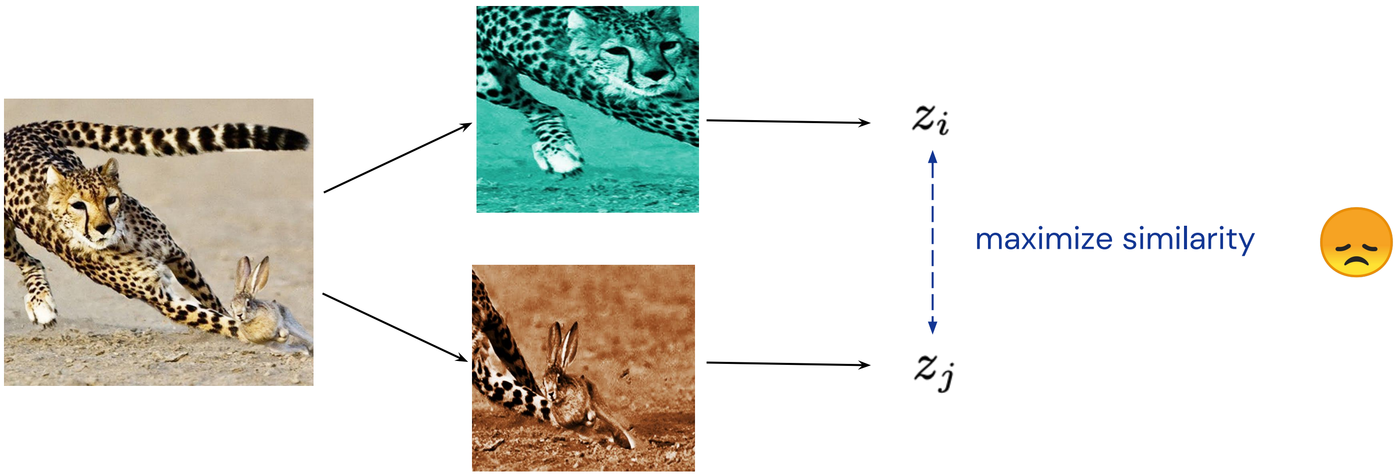



Is the current self-supervised paradigm too simple?

Real-world data is complex

- multiple objects in natural scenes
- multiple speakers in natural speech
- multiple scenes in natural videos

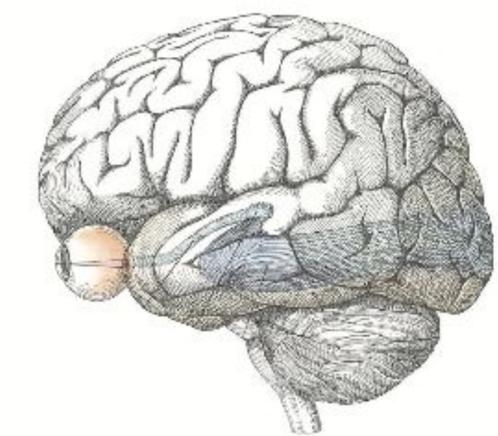
→ invariance across views dampens instance selectivity



Is the current self-supervised paradigm too simple?

Pretrain a ResNet-50 on ImageNet

Fine-tune for object detection, segmentation



Neural representations

- Enable intelligent behavior
- Require minimal supervision
- Are domain-agnostic

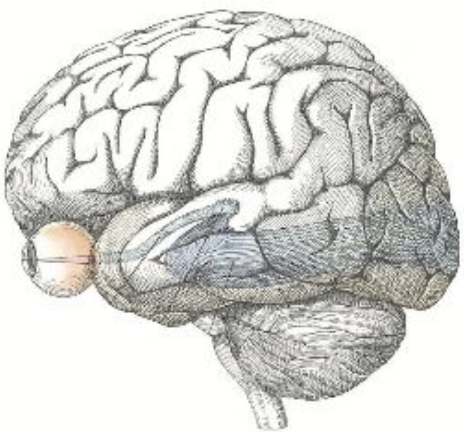
Fine-tune on

- Semantic segmentation (PASCAL or ADE20K)
- Object detection (COCO or LVIS)

Is the current self-supervised paradigm too simple?

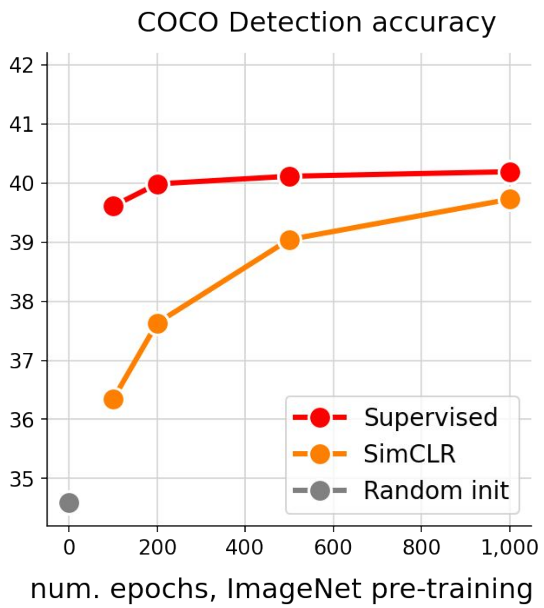
Pretrain a ResNet-50 on ImageNet

Fine-tune for object detection, segmentation



Neural representations

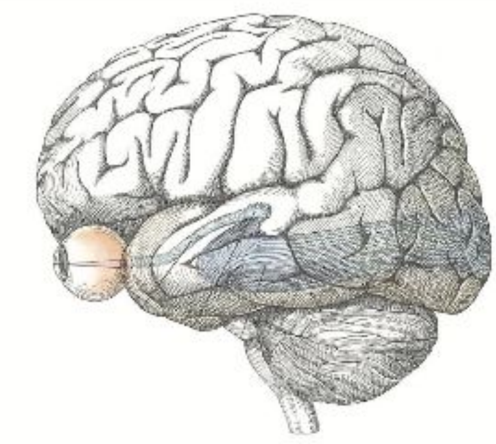
- Enable intelligent behavior
- Require minimal supervision
- Are domain-agnostic



Is the current self-supervised paradigm too simple?

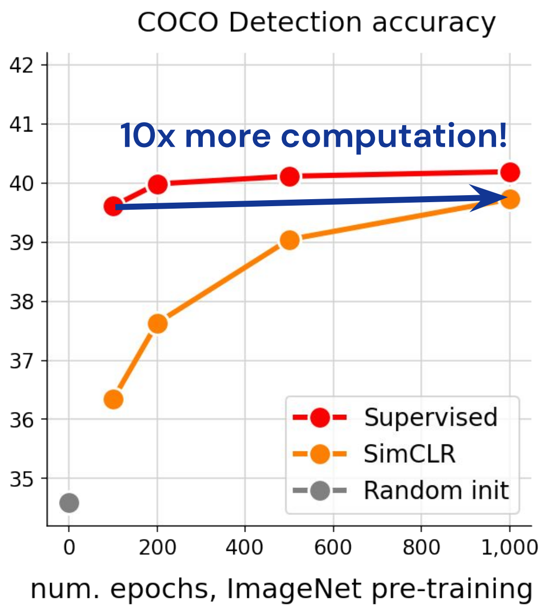
Pretrain a ResNet-50 on ImageNet

Fine-tune for object detection, segmentation



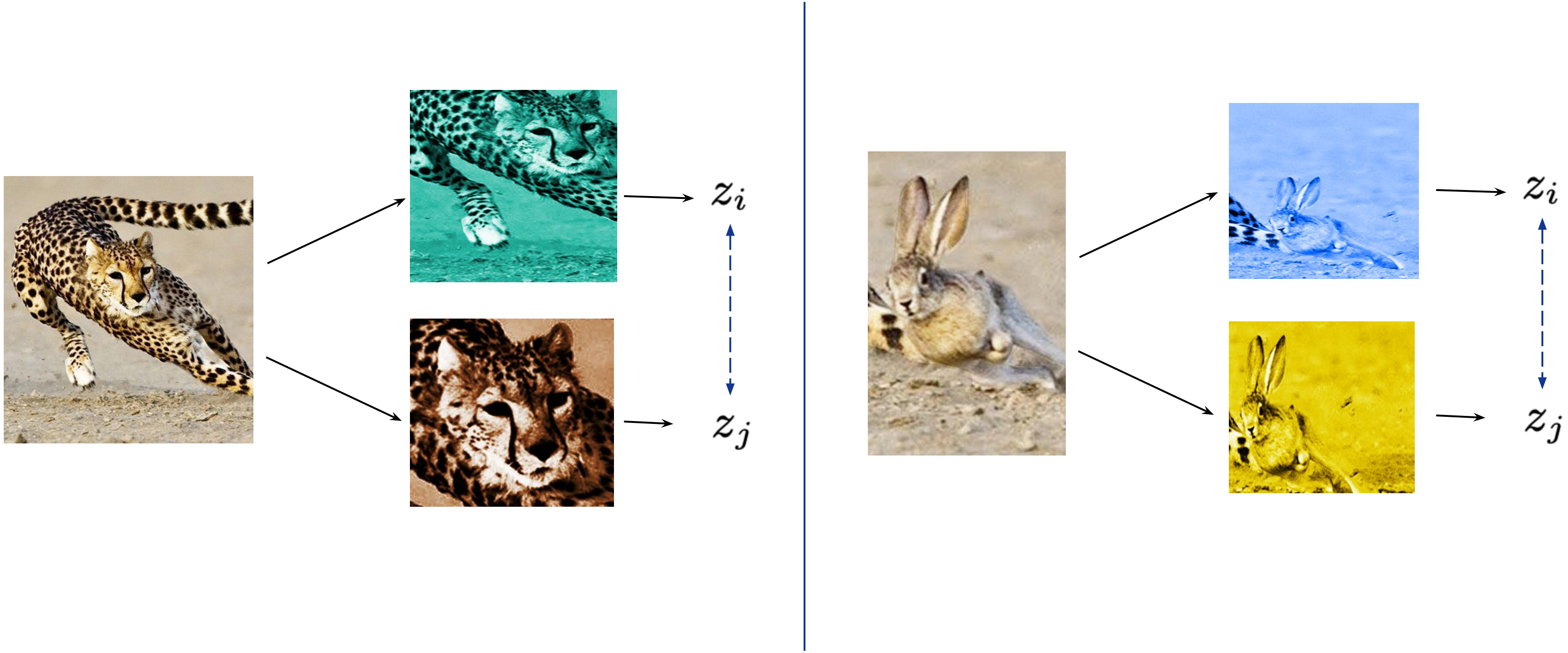
Neural representations

- Enable intelligent behavior
- Require minimal supervision
- Are domain-agnostic



Hypothesis for handling real-world data

- Break images into their constituent objects
- Apply contrastive learning to each **object** rather than each **image**



Outline

1. **Knowledge of objects accelerates and improves representation learning**

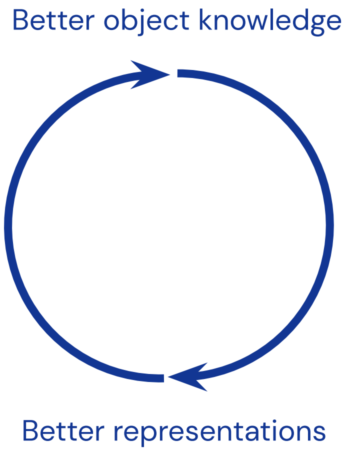
    → DetCon objective (ICCV, 2021)

2. **Knowledge of objects can be extracted from learned representations**

    → Odin framework (ECCV, 2022)

3. **Videos can be used to learn strong image representations**

    → VITO framework (arXiv, 2022)



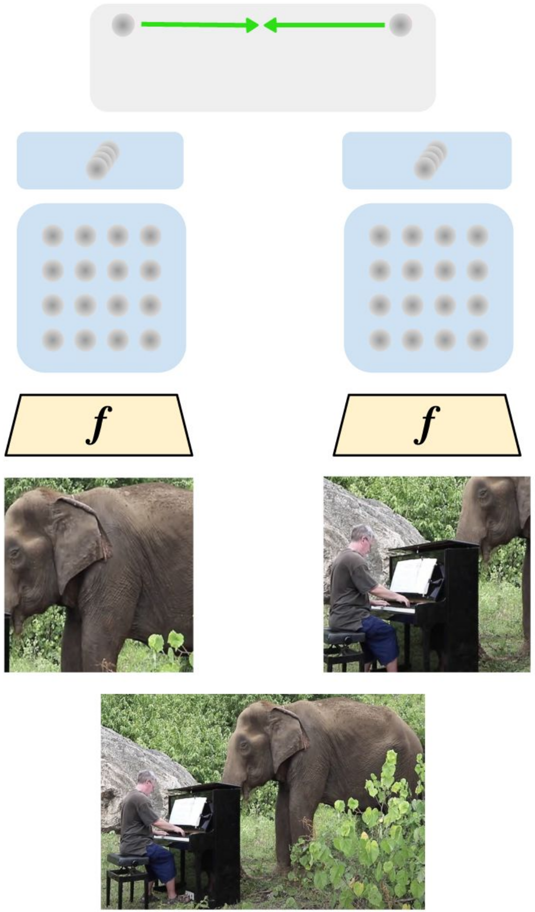
DeepMind

Efficient visual pretraining with contrastive detection

Olivier Hénaff, Skanda Koppula, Jean-Baptiste Alayrac,
Aaron van den Oord, Oriol Vinyals, João Carreira

ICCV 2021

Contrastive learning



Contrastive objective

Global pooling

Convolutional features

Encoder

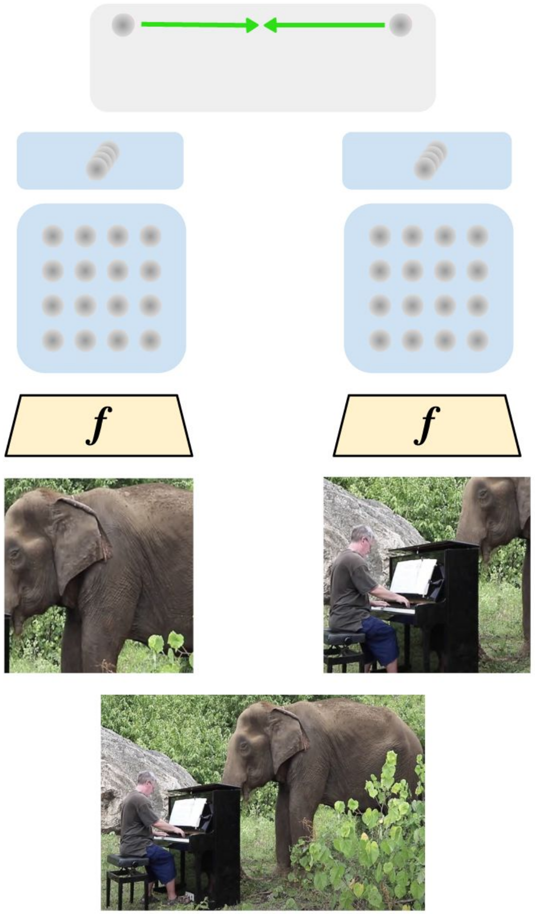
Augmented views

Training image

$$\mathcal{L} = -\log \frac{\exp(\boldsymbol{v} \cdot \boldsymbol{v}')}{\exp(\boldsymbol{v} \cdot \boldsymbol{v}') + \sum_{\boldsymbol{n}} \exp(\boldsymbol{v} \cdot \boldsymbol{v}_{\boldsymbol{n}})}$$

SimCLR, Chen et al. 2020

Contrastive learning



Contrastive objective

Global pooling

Convolutional features

Encoder

Augmented views

Training image

1 positive pair per image

$$\mathcal{L} = -\log \frac{\boxed{\exp(\boldsymbol{v} \cdot \boldsymbol{v}')}}{\exp(\boldsymbol{v} \cdot \boldsymbol{v}') + \boxed{\sum_n \exp(\boldsymbol{v} \cdot \boldsymbol{v}_n)}}$$

1 negative sample per image

- Each image contributes a single positive pair and negative sample
- Positive pairs can be semantically different

SimCLR, Chen et al. 2020

Contrastive detection

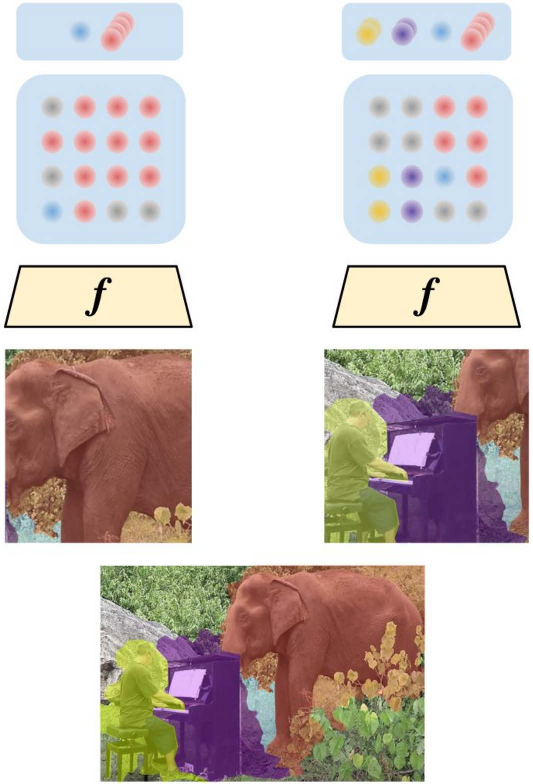
Training image and heuristic masks

Contrastive detection

Augmented views

Training image and heuristic masks

Contrastive detection



Masked pooling

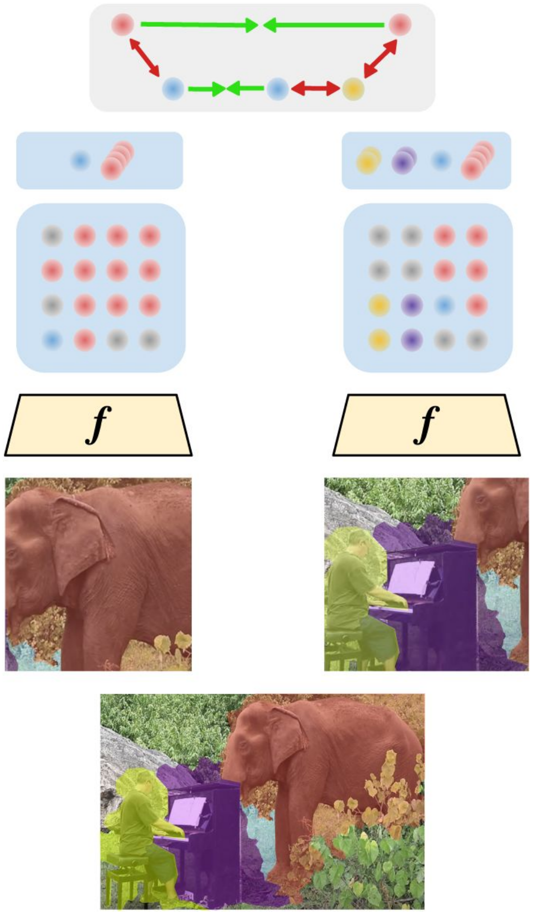
Convolutional features

Encoder

Augmented views

Training image and heuristic masks

Contrastive detection



DetCon objective

Masked pooling

Convolutional features

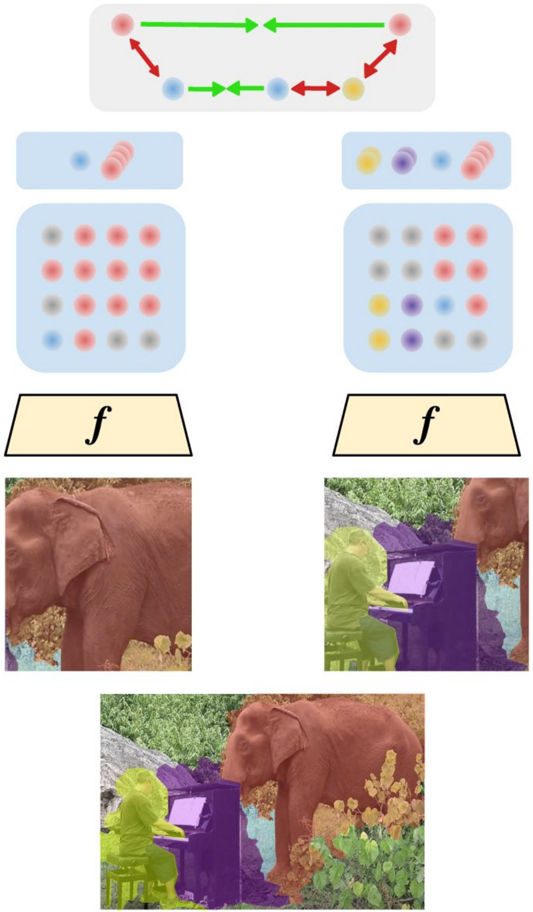
Encoder

Augmented views

Training image and heuristic masks

$$\mathcal{L} = -\sum_{m} \log \frac{\exp(\boldsymbol{v}_m \cdot \boldsymbol{v}'_m)}{\exp(\boldsymbol{v}_m \cdot \boldsymbol{v}'_m) + \sum_n \exp(\boldsymbol{v}_m \cdot \boldsymbol{v}_n)}$$

Contrastive detection



DetCon objective

Masked pooling

Convolutional features

Encoder

Augmented views

Training image and heuristic masks

M positive pairs per image

$$\mathcal{L} = -\sum_m \log \frac{\boxed{\exp(v_m \cdot v'_m)}}{\exp(v_m \cdot v'_m) + \boxed{\sum_n \exp(v_m \cdot v_n)}}$$

M negative samples per image

- Each image contributes multiple positive pairs and negative samples
- Positive pairs are spatially aligned

Unsupervised segmentation

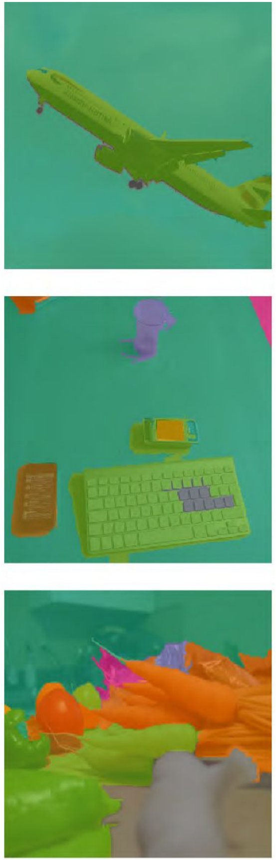
Spatial heuristic | Heuristic: FH | Heuristic: MCG

Felzenszwalb & Huttenlocher 2004

Arbeláez et al. 2014

Unsupervised segmentation

Heuristic: FH



Felzenszwalb & Huttenlocher 2004

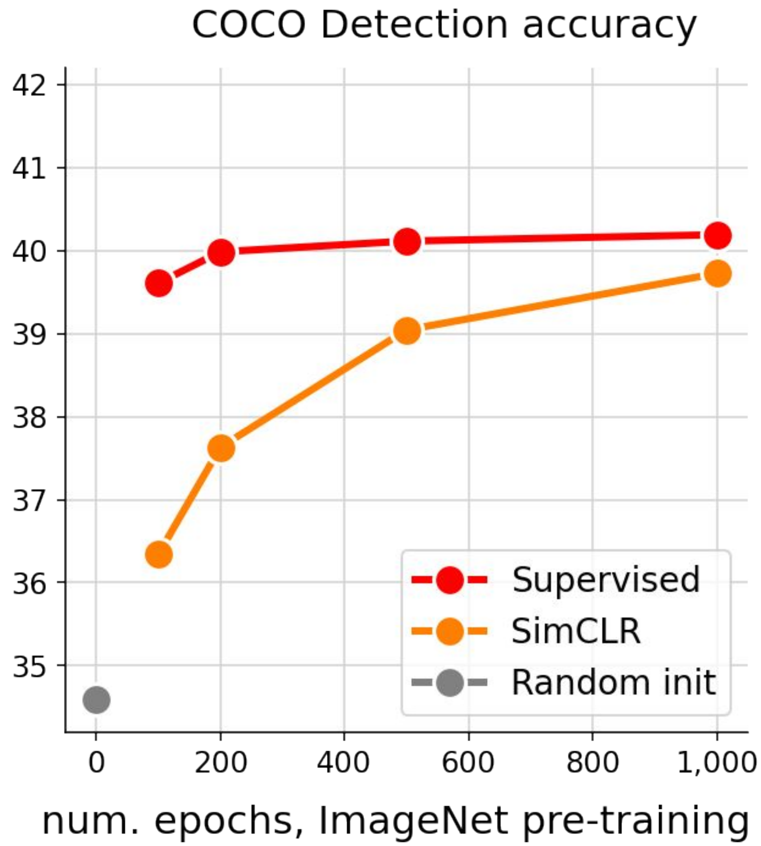
skimage.segmentation.felzenszwalb

Experiments

Pretrain ResNet-50 on ImageNet

Objective: Supervised, SimCLR, or DetCon

Transfer to COCO detection and instance segmentation using Mask-RCNN

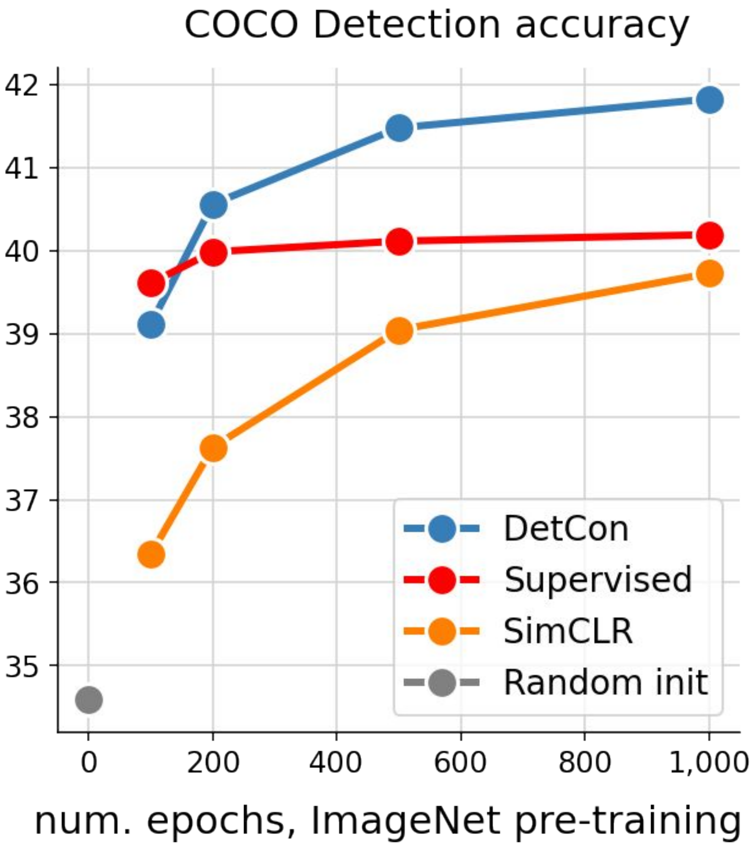


Experiments

Pretrain ResNet-50 on ImageNet

Objective: Supervised, SimCLR, or DetCon

Transfer to COCO detection and instance segmentation using Mask-RCNN

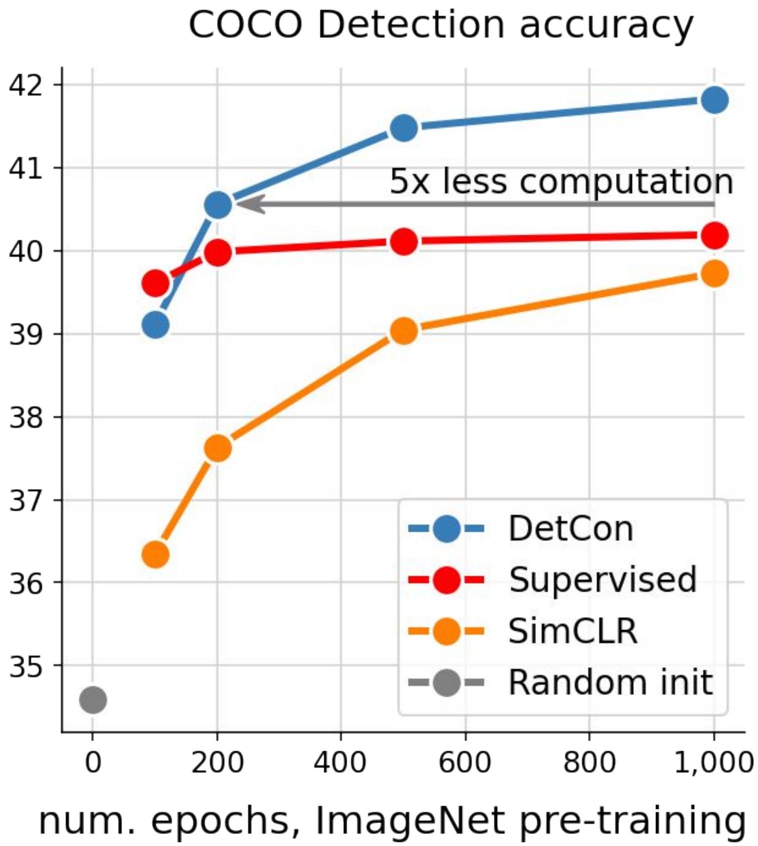


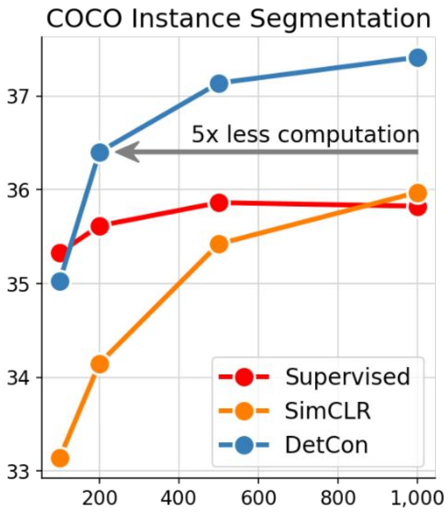
Experiments

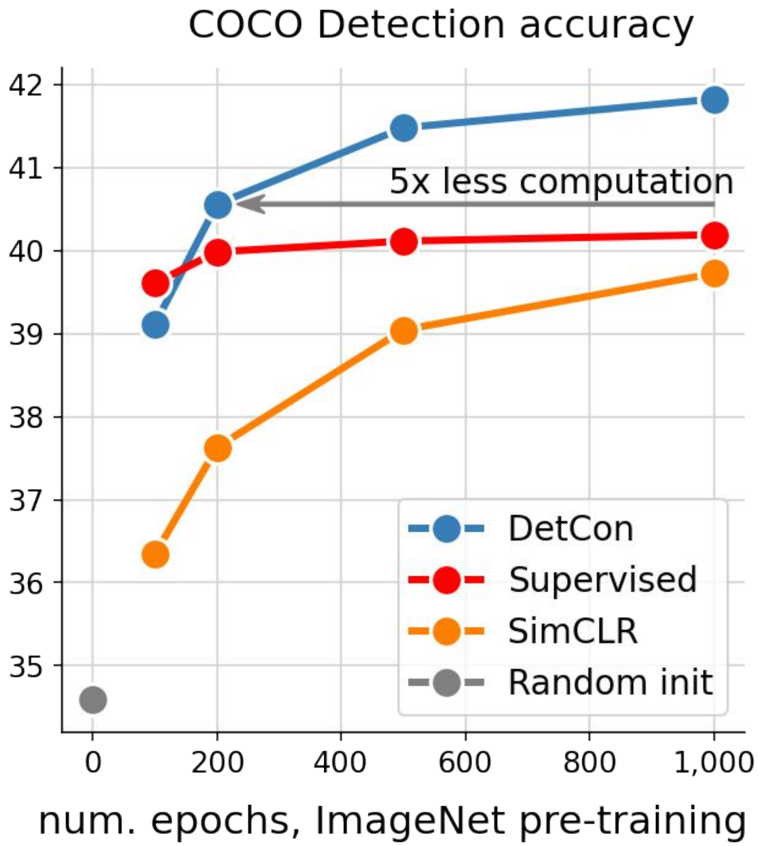
Pretrain ResNet-50 on ImageNet

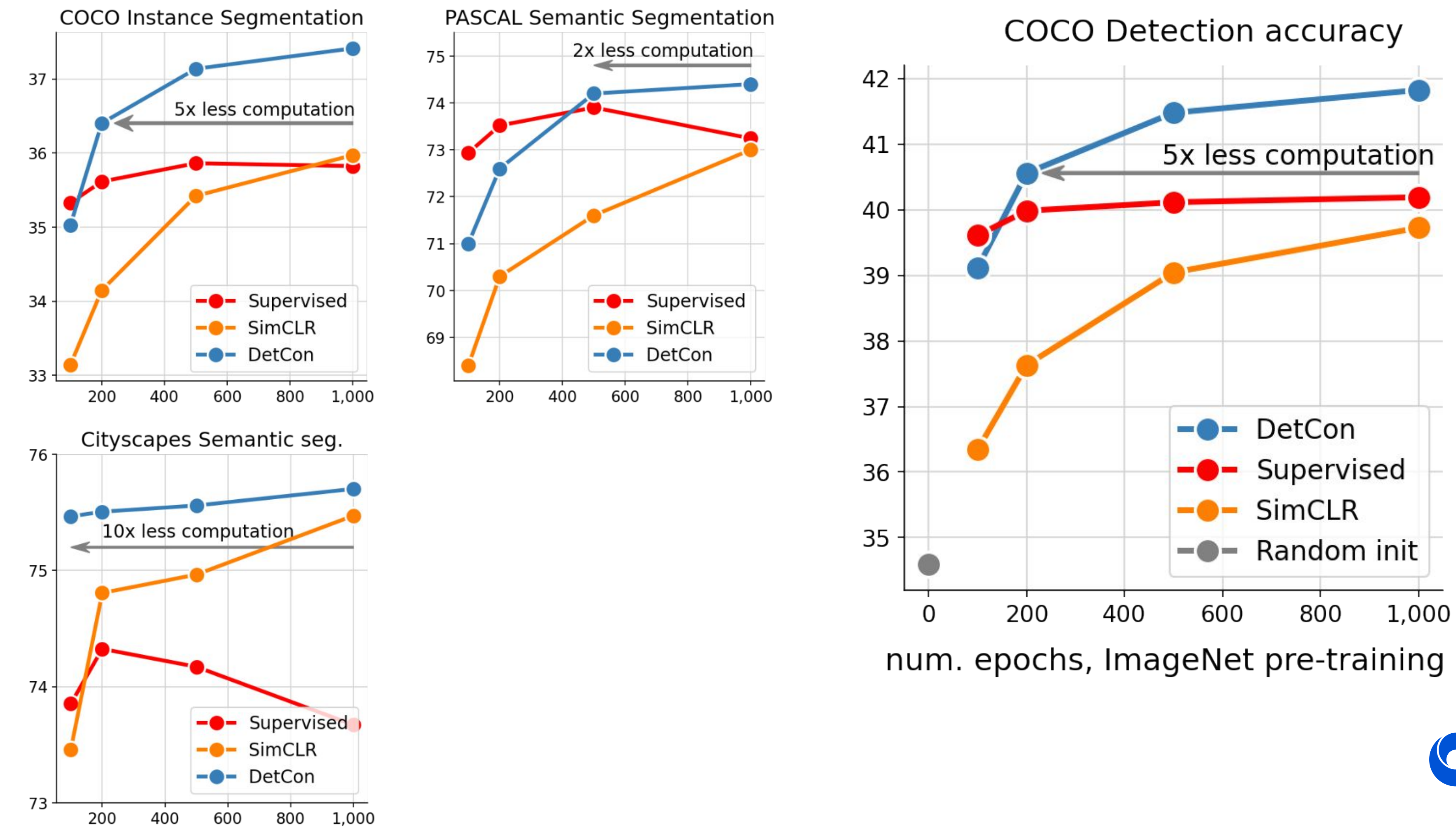
Objective: Supervised, SimCLR, or DetCon

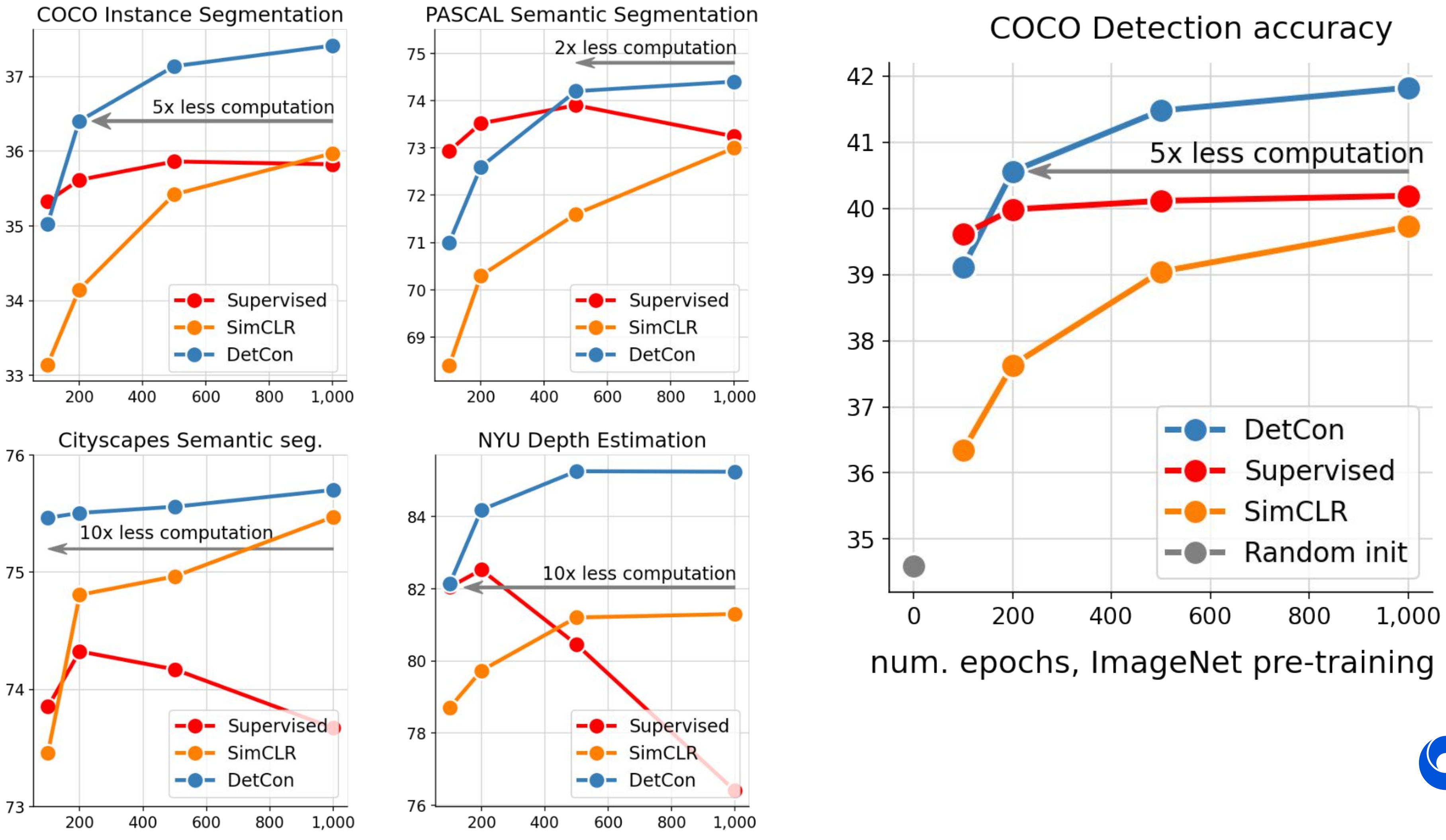
Transfer to COCO detection and instance segmentation using Mask-RCNN




COCO Instance Segmentation
5x less computation
Supervised
SimCLR
DetCon
33
34
35
36
37
200
400
600
800
1,000


COCO Detection accuracy
5x less computation
DetCon
Supervised
SimCLR
Random init
35
36
37
38
39
40
41
42
0
200
400
600
800
1,000
num. epochs, ImageNet pre-training


COCO Instance Segmentation
5x less computation
Supervised
SimCLR
DetCon
PASCAL Semantic Segmentation
2x less computation
Supervised
SimCLR
DetCon
Cityscapes Semantic seg.
10x less computation
Supervised
SimCLR
DetCon
COCO Detection accuracy
5x less computation
DetCon
Supervised
SimCLR
Random init
num. epochs, ImageNet pre-training



COCO Instance Segmentation
5x less computation
Supervised
SimCLR
DetCon
PASCAL Semantic Segmentation
2x less computation
Supervised
SimCLR
DetCon
Cityscapes Semantic seg.
10x less computation
Supervised
SimCLR
DetCon
NYU Depth Estimation
10x less computation
Supervised
SimCLR
DetCon
COCO Detection accuracy
5x less computation
DetCon
Supervised
SimCLR
Random init
num. epochs, ImageNet pre-training

Experiments

SimCLR $\xrightarrow{\text{2-10x gain in efficiency}}$ DetCon

Experiments

$$\begin{tikzcd}
\mathrm{SimCLR} \arrow[r, "\text{2-10x gain in efficiency}"] \arrow[d, "\begin{array}{l}\text{+ prediction heads}\\ \text{+ target networks}\\ \text{+ better augmentations}\end{array}"] & \mathrm{DetCon}_S \arrow[d] \\
\mathrm{BYOL} \arrow[r, "\text{gain in efficiency?}"'] & \mathrm{DetCon}_B
\end{tikzcd}$$

Experiments

$$\begin{tikzcd} \text{SimCLR} \arrow[r, "\text{2-10x gain in efficiency}"] \arrow[d, "\begin{array}{l}\text{+ prediction heads}\\ \text{+ target networks}\\ \text{+ better augmentations}\end{array}"] & \text{DetCon}_S \arrow[d] \\ \text{BYOL} \arrow[r, "\text{gain in efficiency?}"'] & \text{DetCon}_B \end{tikzcd}$$

	Detection COCO	
Pretrain epochs	300	1000
BYOL	41.2	41.6
DetCon$_B$	42.0	42.7
Efficiency Gain	$> 3\times$	

Experiments

$$\begin{tikzcd}[column sep=huge, row sep=huge]
\text{SimCLR} \arrow[r, "\text{2-10x gain in efficiency}"] \arrow[d, "\begin{array}{l}\text{+ prediction heads}\\ \text{+ target networks}\\ \text{+ better augmentations}\end{array}"] & \text{DetCon}_S \arrow[d] \\
\text{BYOL} \arrow[r, "\text{3-10x gain in efficiency!!!}"'] & \text{DetCon}_B
\end{tikzcd}$$

	Detection COCO		Instance Segmentation COCO		Semantic Segmentation PASCAL		Semantic Segmentation Cityscapes		Depth Estimation NYU v2	
Pretrain epochs	300	1000	300	1000	300	1000	300	1000	100	1000
BYOL	41.2	41.6	37.1	37.2	74.7	75.7	73.4	74.6	83.7	84.2
DetCon$_B$	42.0	42.7	37.8	38.2	75.6	77.3	75.1	77.0	85.1	86.3
Efficiency Gain	$> 3\times$		$> 3\times$		$\approx 3\times$		$> 3\times$		$> 10\times$	

Experiments

$$\begin{tikzcd}[column sep=huge, row sep=huge]
\text{SimCLR} \arrow[r, "\text{2-10x gain in efficiency}"] \arrow[d, "\begin{array}{l}\text{+ prediction heads}\\ \text{+ target networks}\\ \text{+ better augmentations}\end{array}"] & \text{DetCon}_S \arrow[d] \\
\text{BYOL} \arrow[r, "\text{3-10x gain in efficiency!!!}"'] & \text{DetCon}_B
\end{tikzcd}$$

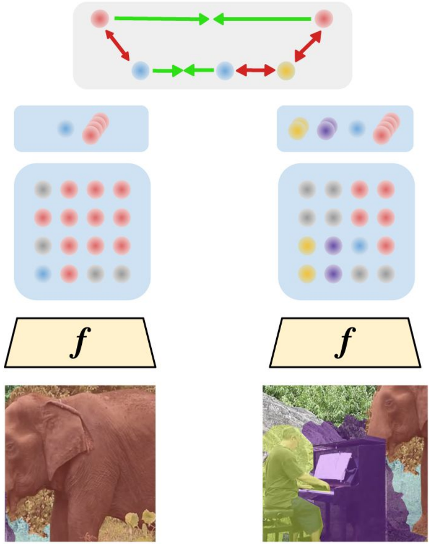
method	Fine-tune 1× AP^{bb}	Fine-tune 1× AP^{mk}	Fine-tune 2× AP^{bb}	Fine-tune 2× AP^{mk}
Supervised	39.6	35.6	41.6	37.6
VADeR	39.2	35.6	-	-
MoCo	39.4	35.6	41.7	37.5
SimCLR	39.7	35.8	41.6	37.4
MoCo v2	40.1	36.3	41.7	37.6
InfoMin	40.6	36.7	42.5	38.4
PixPro	41.4	-	-	-
BYOL	41.6	37.2	42.4	38.0
SwAV	41.6	37.8	-	-
DetCon$_S$	41.8	37.4	42.9	38.1
DetCon$_B$	**42.7**	**38.2**	**43.4**	**38.7**

	Detection COCO		Instance Segmentation COCO		Semantic Segmentation PASCAL		Semantic Segmentation Cityscapes		Depth Estimation NYU v2	
Pretrain epochs	300	1000	300	1000	300	1000	300	1000	100	1000
BYOL	41.2	41.6	37.1	37.2	74.7	75.7	73.4	74.6	83.7	84.2
DetCon$_B$	42.0	42.7	37.8	38.2	75.6	77.3	75.1	77.0	85.1	86.3
Efficiency Gain	$> 3\times$		$> 3\times$		$\approx 3\times$		$> 3\times$		$> 10\times$	

Knowledge of objects improves representation learning

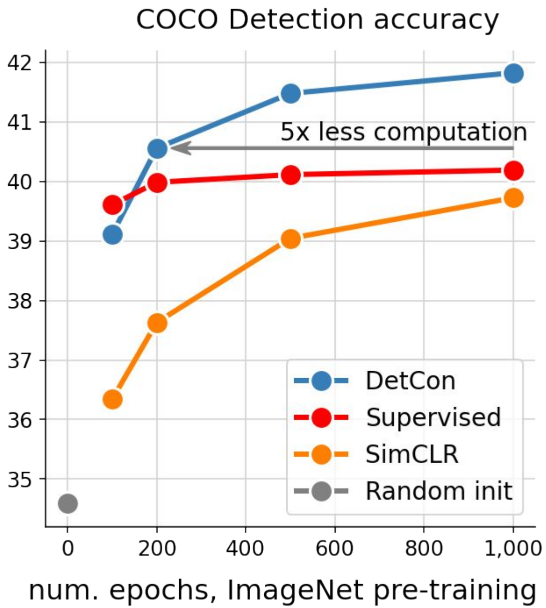
Pretrain a ResNet-50 on ImageNet

Fine-tune for object detection, segmentation



Neural representations

- Enable intelligent behavior
- Require minimal supervision
- Are generally applicable



Knowledge of objects improves representation learning

Pretrain a ResNet-50 on ImageNet

0. Sort E into $\pi = (o_1, \ldots, o_m)$, by non-decreasing
1. Start with a segmentation S^0, where each vert
2. Repeat step 3 for $q = 1, \ldots, m$.
3. Construct S^q given S^{q-1} as follows. Let v_i and by the q-th edge in the ordering, i.e., $o_q = (v_i$ components of S^{q-1} and $w(o_q)$ is small compa both those components, then merge the two co More formally, let C_i^{q-1} be the component of component containing v_j. If $C_i^{q-1} \neq C_j^{q-1}$ and S^q is obtained from S^{q-1} by merging C_i^{q-1} and
4. Return $S = S^m$.

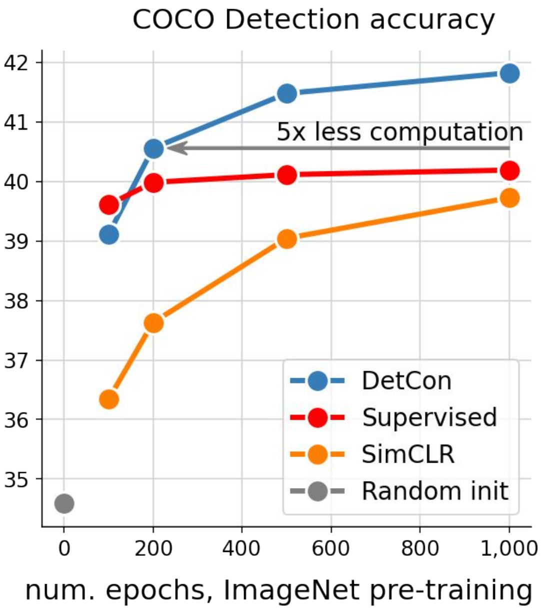
Neural representations

- Enable intelligent behavior

- Require minimal supervision

- Are generally applicable

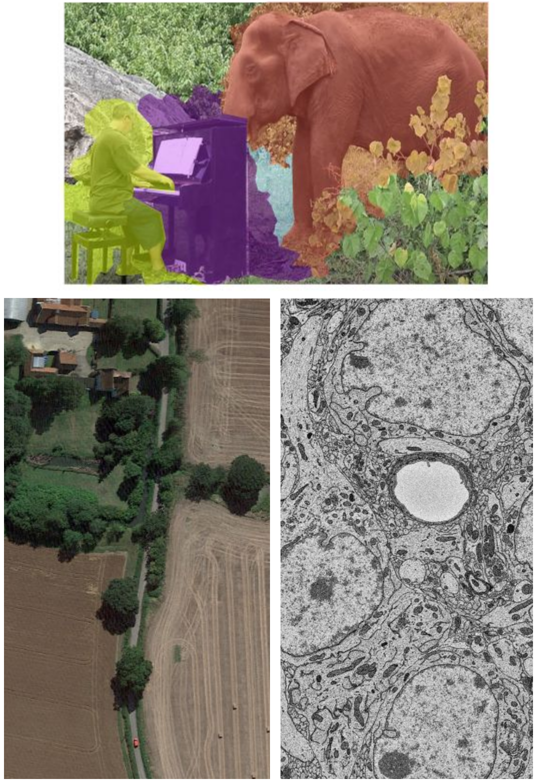
Fine-tune for object detection, segmentation



Can the self-supervised paradigm stay general?

Self-supervised pretraining across domains

Fine-tune for object detection, segmentation



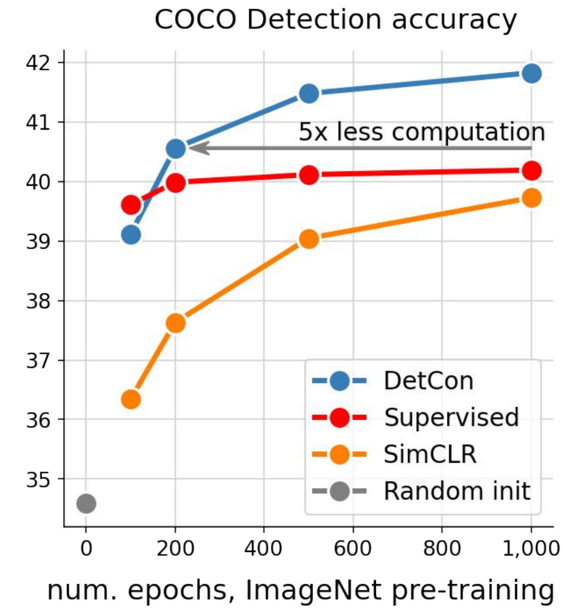
0. Sort E into $\pi = (o_1, \ldots, o_m)$, by non-decreasin
1. Start with a segmentation S^0, where each vert
2. Repeat step 3 for $q = 1, \ldots, m$.
3. Construct S^q given S^{q-1} as follows. Let v_i and by the q-th edge in the ordering, i.e., $o_q = (v_i$ components of S^{q-1} and $w(o_q)$ is small compa both those components, then merge the two co More formally, let C_i^{q-1} be the component of component containing v_j. If $C_i^{q-1} \neq C_j^{q-1}$ and S^q is obtained from S^{q-1} by merging C_i^{q-1} and
4. Return $S = S^m$.

Neural representations

– Enable intelligent behavior

– Require minimal supervision

– Are generally applicable

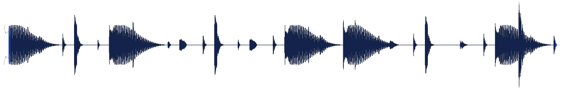


Can the self-supervised paradigm stay general?

Self-supervised pretraining across modalities

Fine-tune for object detection, segmentation

But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.



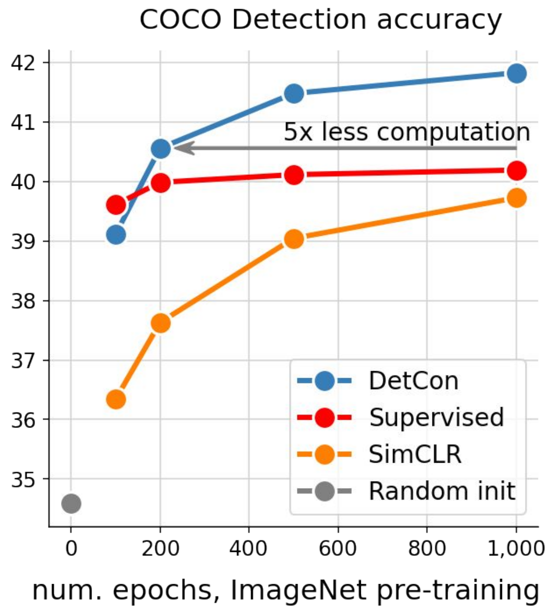
0. Sort E into $\pi = (o_1, \ldots, o_m)$, by non-decreasin
1. Start with a segmentation S^0, where each vert
2. Repeat step 3 for $q = 1, \ldots, m$.
3. Construct S^q given S^{q-1} as follows. Let v_i and by the q-th edge in the ordering, i.e., $o_q = (v_i$ components of S^{q-1} and $w(o_q)$ is small compa both those components, then merge the two co More formally, let C_i^{q-1} be the component of component containing v_j. If $C_i^{q-1} \neq C_j^{q-1}$ and S^q is obtained from S^{q-1} by merging C_i^{q-1} and
4. Return $S = S^m$.

Neural representations

- Enable intelligent behavior

- Require minimal supervision

- Are generally applicable



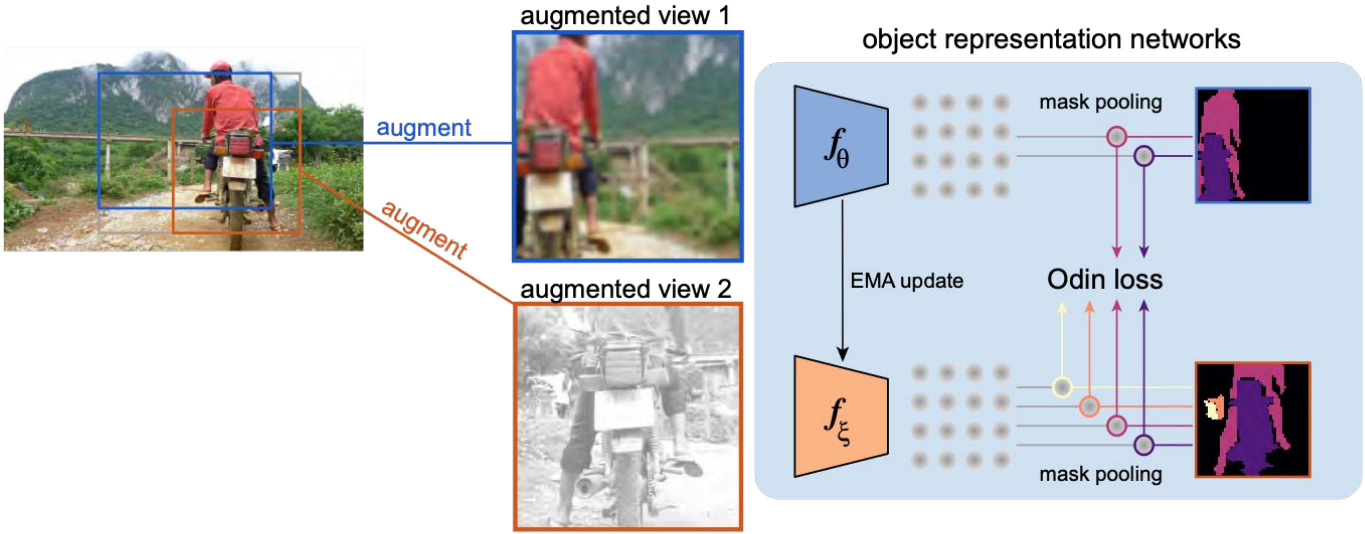
DeepMind

Object discovery and representation networks

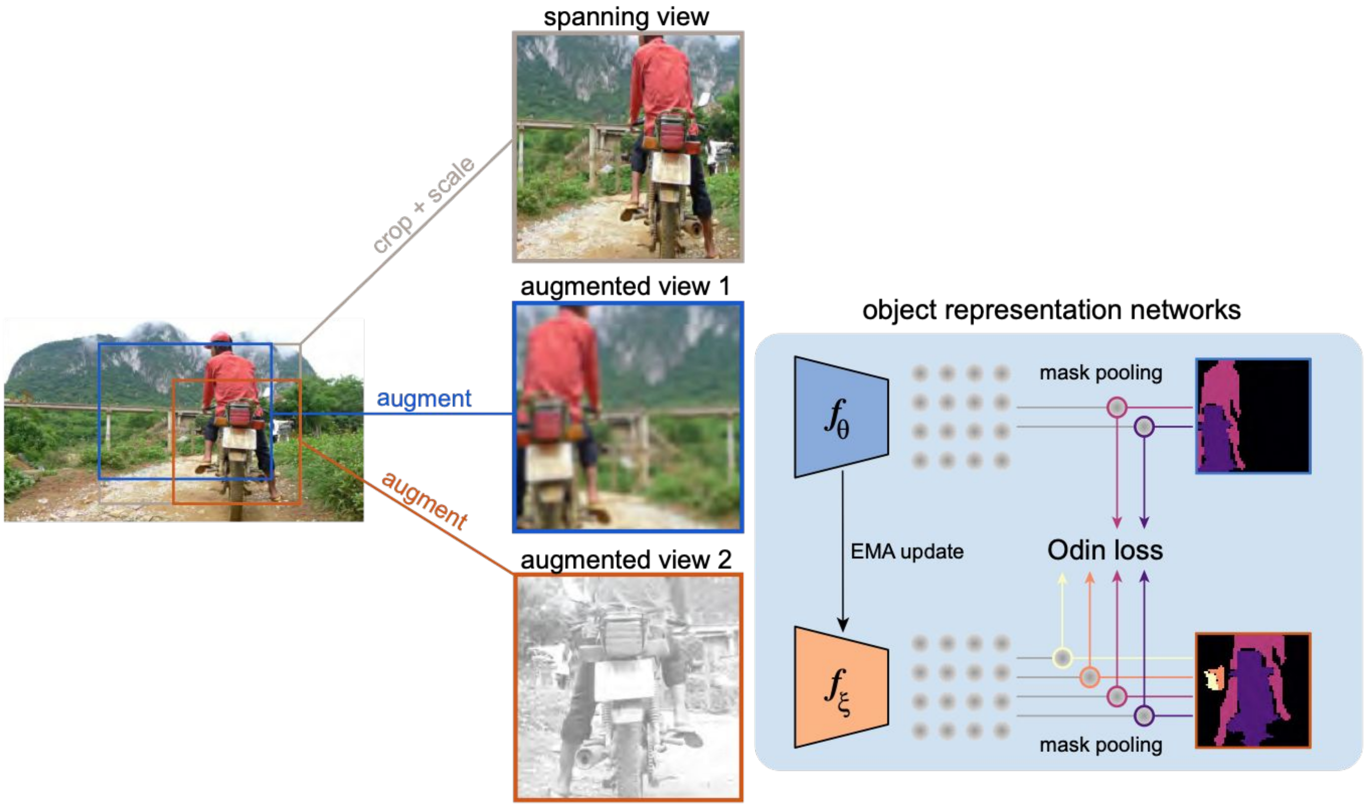
Olivier Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Drew Jaegle, Andrew Zisserman, João Carreira, Relja Arandjelović

ECCV 2022

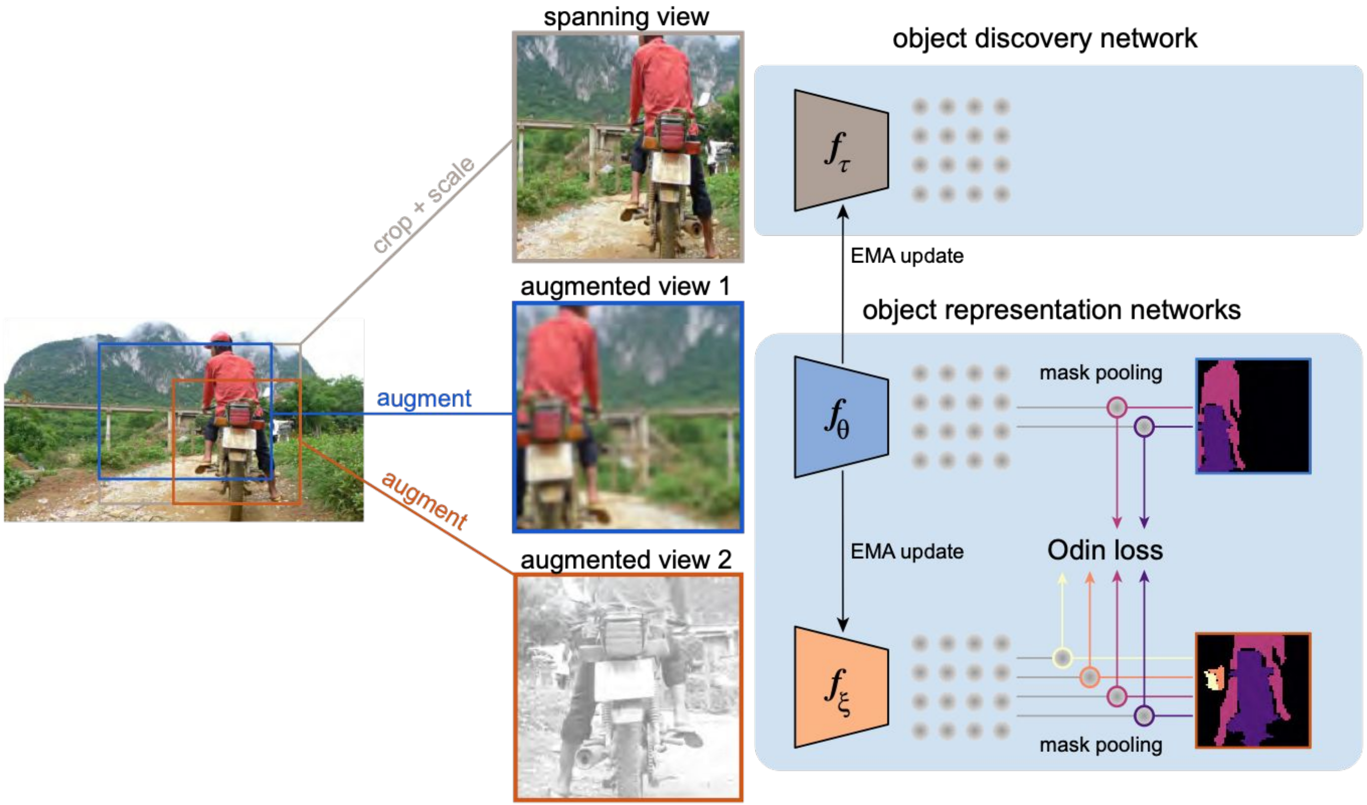
Odin: Object discovery and representation networks



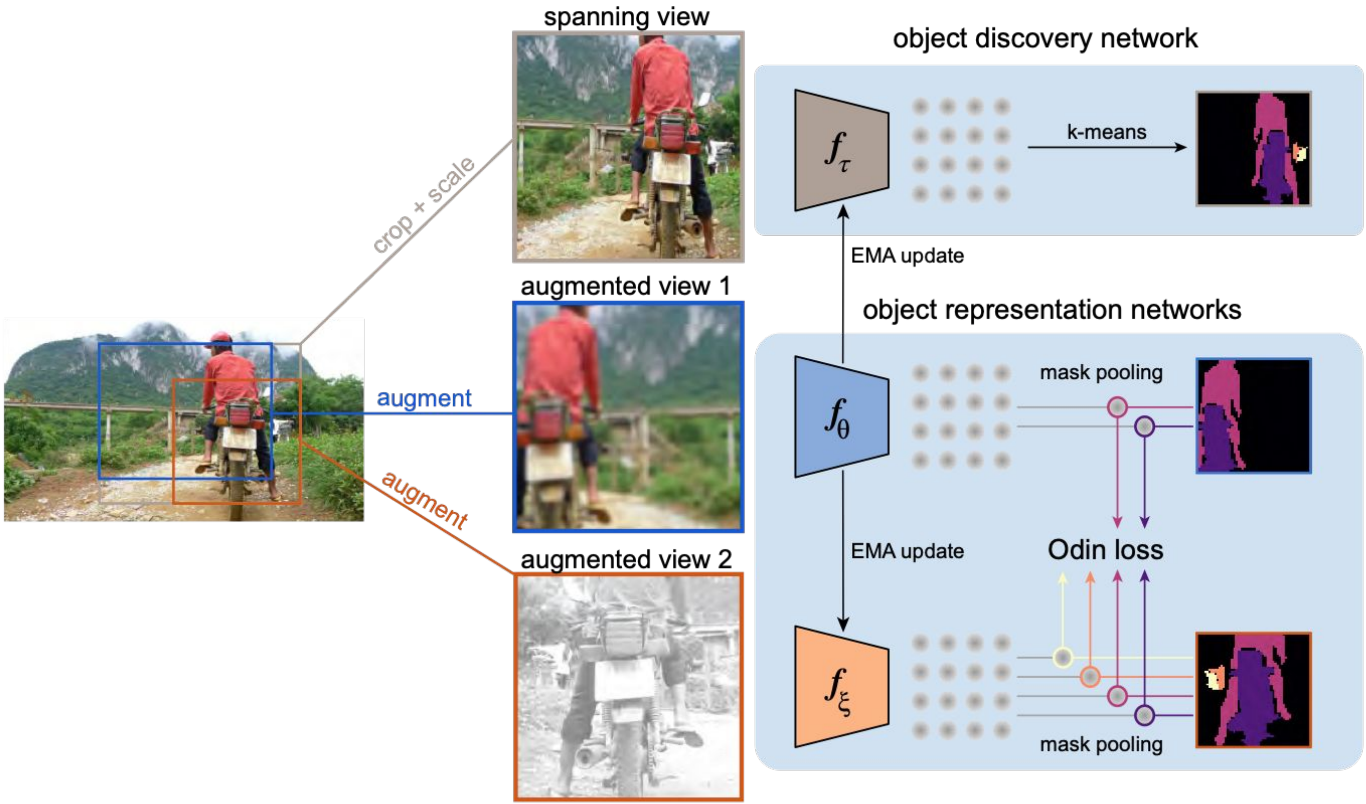
Odin: Object discovery and representation networks



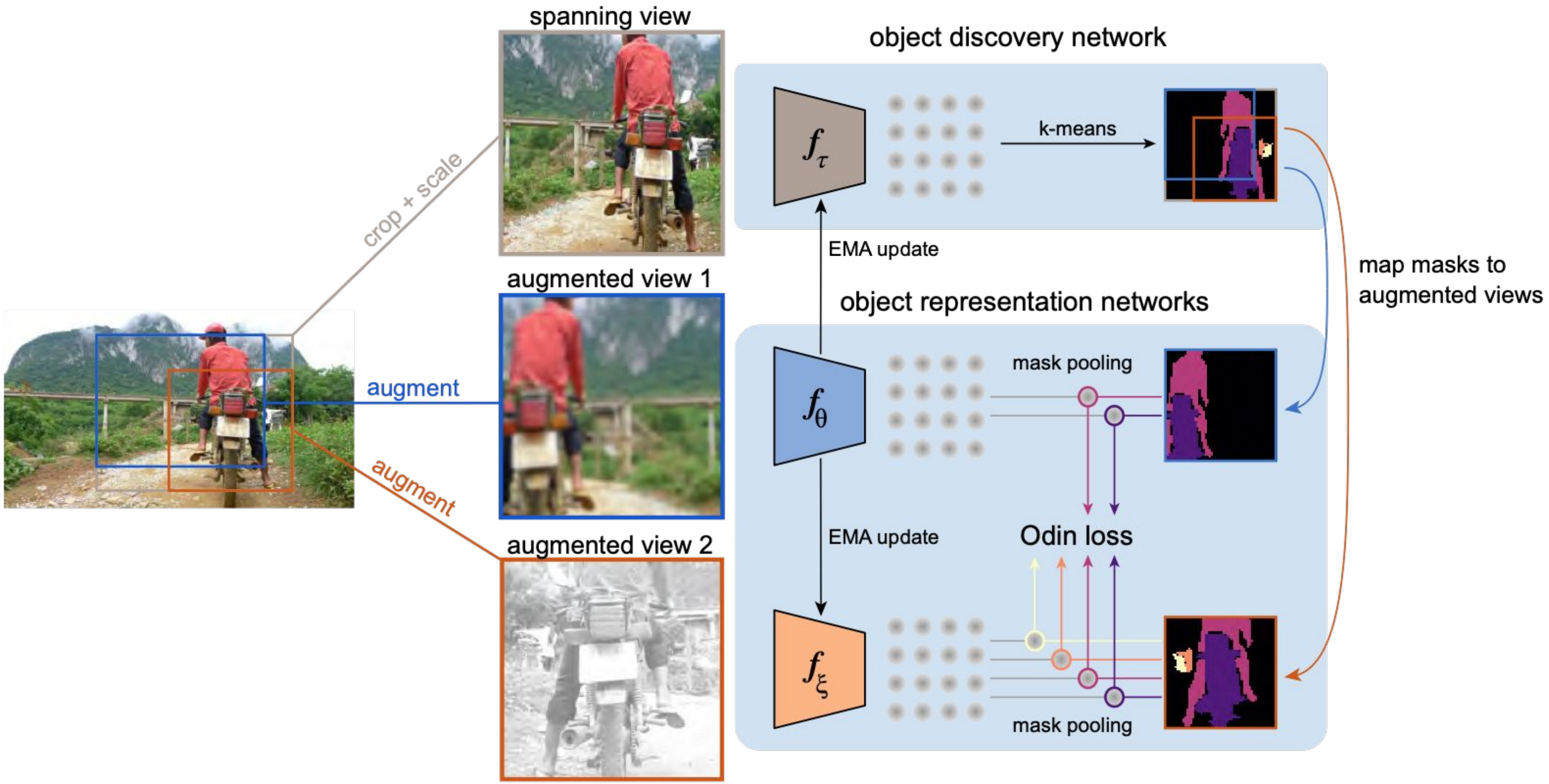
Odin: Object discovery and representation networks



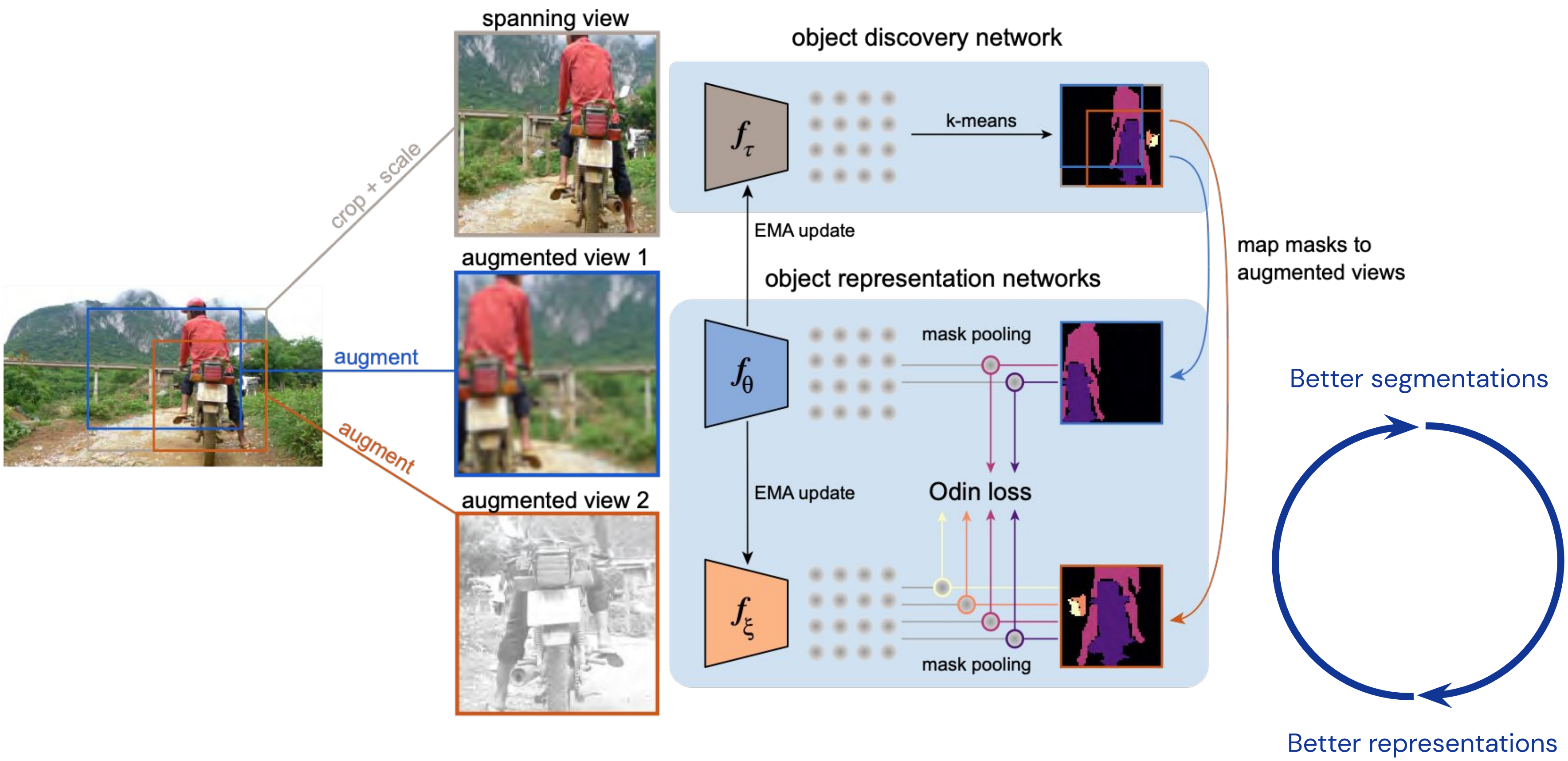
Odin: Object discovery and representation networks



Odin: Object discovery and representation networks



Odin: Object discovery and representation networks



Transfer learning

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Fine-tune for COCO object detection and instance segmentation using Mask-RCNN

pretraining		fine-tune 1×		fine-tune 2×	
Method	Knows obj?	AP^{bb}	AP^{mk}	AP^{bb}	AP^{mk}
Supervised	no	39.6	35.6	41.6	37.6
VADeR [68]	no	39.2	35.6	-	-
MoCo [37]	no	39.4	35.6	41.7	37.5
SimCLR [13]	no	39.7	35.8	41.6	37.4
MoCo v2 [14]	no	40.1	36.3	41.7	37.6
InfoMin [75]	no	40.6	36.7	42.5	38.4
DINO [12]	no	41.2	37.1	42.3	38.1
PixPro [86]	no	41.4	-	-	-
BYOL [33]	no	41.6	37.2	42.4	38.0
SwAV [11]	no	41.6	37.8	-	-
ReLIC v2 [77]	yes	42.5	38.0	43.3	38.6
$DetCon_B$ [40]	yes	42.7	38.2	43.4	38.7

Transfer learning

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Fine-tune for COCO object detection and instance segmentation using Mask-RCNN ✅

pretraining		fine-tune 1×		fine-tune 2×	
Method	Knows obj?	AP^{bb}	AP^{mk}	AP^{bb}	AP^{mk}
Supervised	no	39.6	35.6	41.6	37.6
VADeR [68]	no	39.2	35.6	-	-
MoCo [37]	no	39.4	35.6	41.7	37.5
SimCLR [13]	no	39.7	35.8	41.6	37.4
MoCo v2 [14]	no	40.1	36.3	41.7	37.6
InfoMin [75]	no	40.6	36.7	42.5	38.4
DINO [12]	no	41.2	37.1	42.3	38.1
PixPro [86]	no	41.4	-	-	-
BYOL [33]	no	41.6	37.2	42.4	38.0
SwAV [11]	no	41.6	37.8	-	-
ReLIC v2 [77]	yes	42.5	38.0	43.3	38.6
DetCon_B [40]	yes	42.7	38.2	43.4	38.7
Odin	no	**42.9**	**38.4**	**43.8**	**39.1**

Transfer learning

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Fine-tune for COCO object detection and instance segmentation using Mask-RCNN ✅
- Fine-tune for PASCAL and Cityscapes semantic segmentation ✅

Method	Knows obj?	PASCAL	Cityscapes
Supervised	no	74.4	74.9
BYOL [33]	no	75.7	74.6
DINO [12]	no	76.9	75.6
DetCon_B [40]	yes	77.3	77.0
ReLIC v2 [77]	yes	77.9	75.2
Odin	no	**78.6**	**77.1**

Transfer learning

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Fine-tune for COCO object detection and instance segmentation using Mask-RCNN ✅
- Fine-tune for PASCAL and Cityscapes semantic segmentation ✅
- Fine-tune for COCO object detection using FCOS* ✅

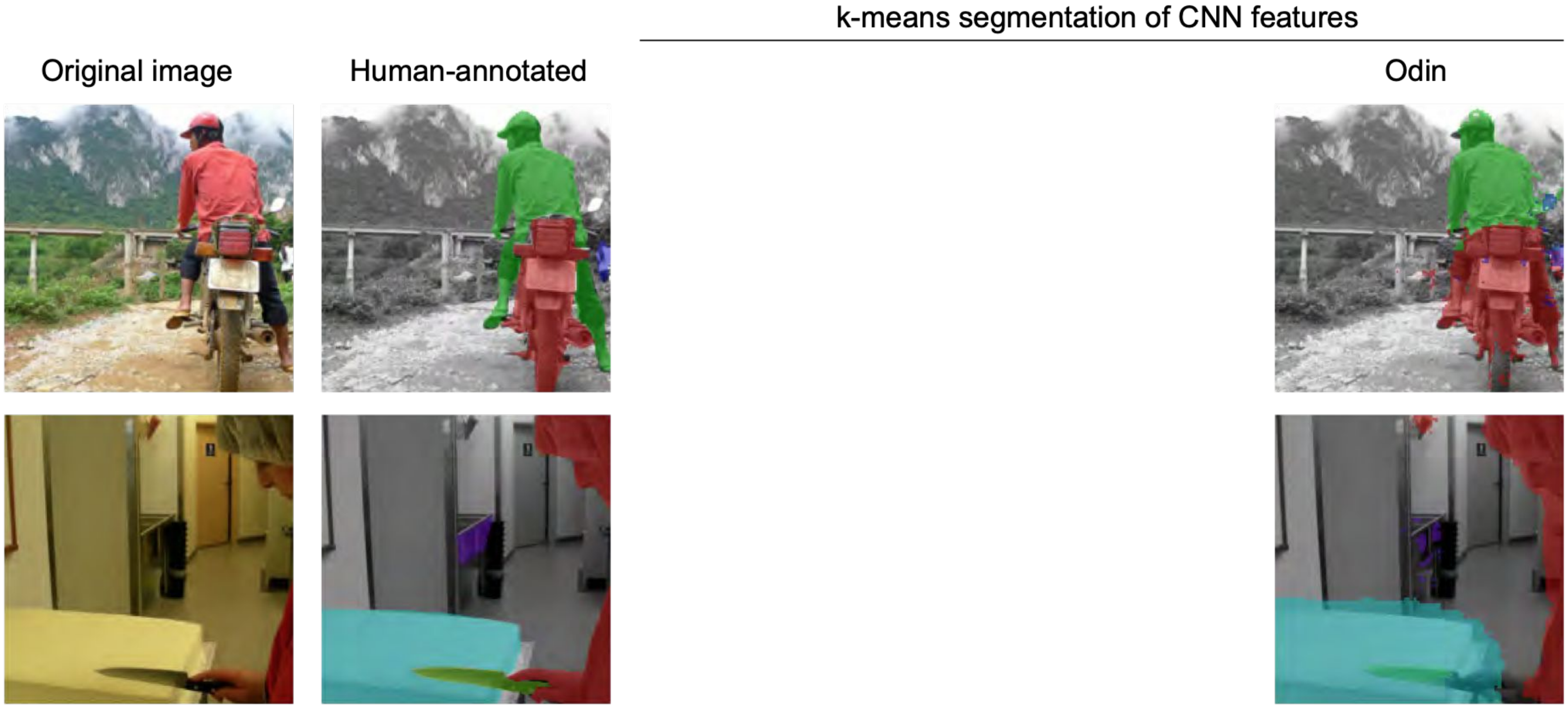
Method	Knows obj?	PASCAL	Cityscapes
Supervised	no	74.4	74.9
BYOL [33]	no	75.7	74.6
DINO [12]	no	76.9	75.6
DetCon$_B$ [40]	yes	77.3	77.0
ReLIC v2 [77]	yes	77.9	75.2
Odin	no	**78.6**	**77.1**

Pretraining	Knows obj?	ResNet-50	Swin-T	Swin-S
Supervised	no	44.2	46.7	48.3
DINO [12]	no	44.3	-	-
MOBY [85]	no	-	47.6	-
DetCon$_B$ [40]	yes	45.4	48.4	**50.4**
Odin	no	**45.6**	**48.5**	**50.4**

Object discovery

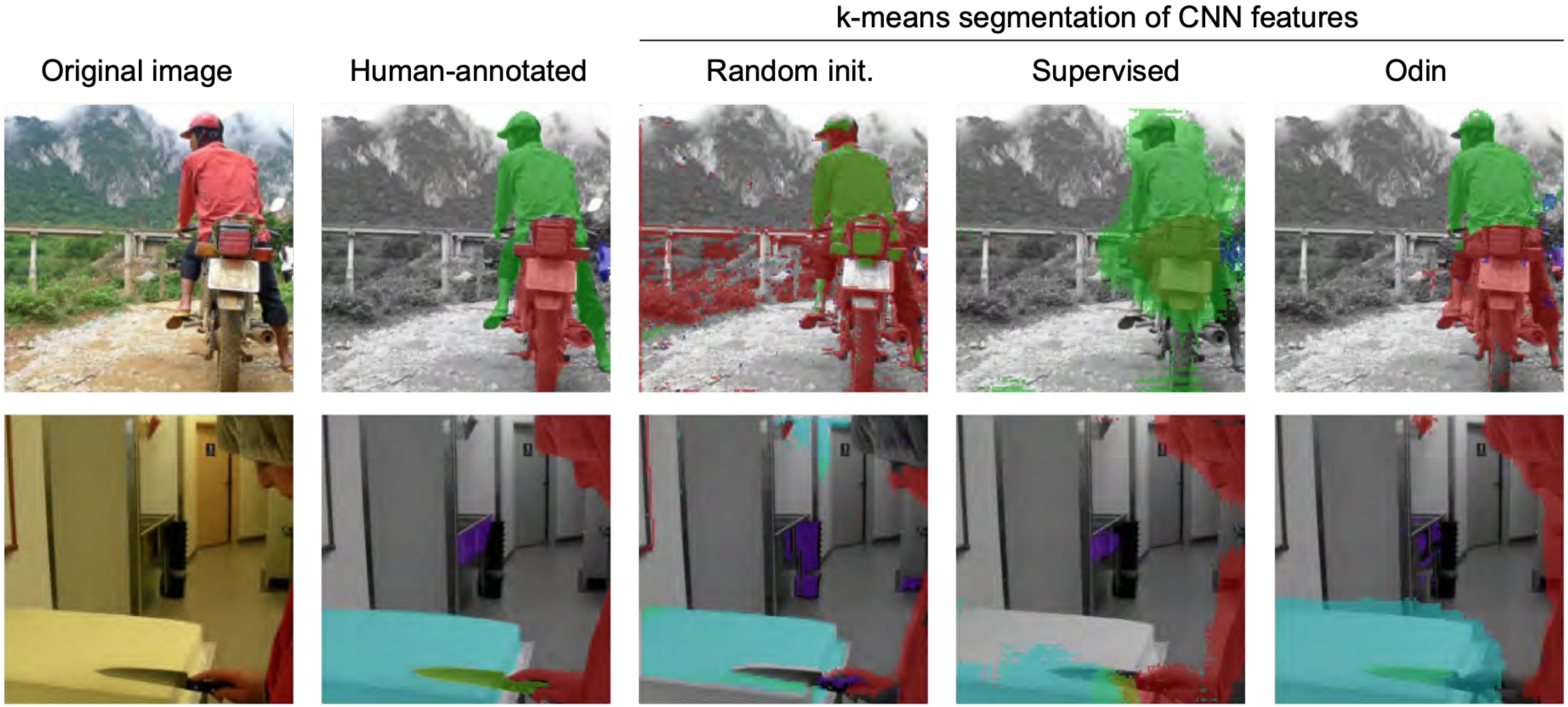
Object discovery

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap



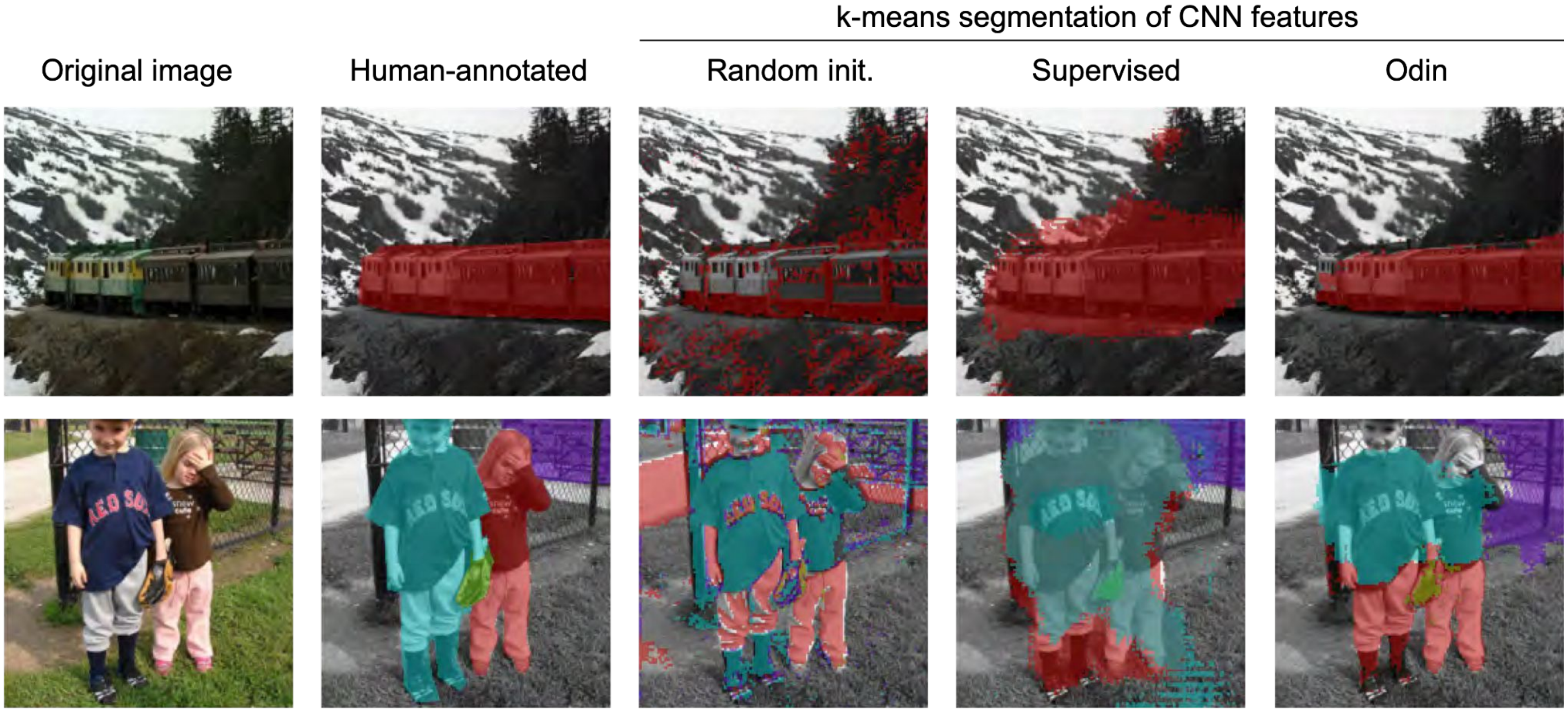
Object discovery

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap



Object discovery

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap



Object discovery

- Pretrain ResNet-50 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap

	ResNet-50		
Pretraining	ABO^i	ABO^c	OR
Random init	28.1	33.6	17.0
DetCon_B [42]	34.1	40.0	20.4
Supervised	35.8	41.1	23.8
DINO [12]	38.3	46.5	30.8
Odin	**41.5**	**48.6**	**36.5**
Odin†	**43.0**	**53.0**	**42.3**

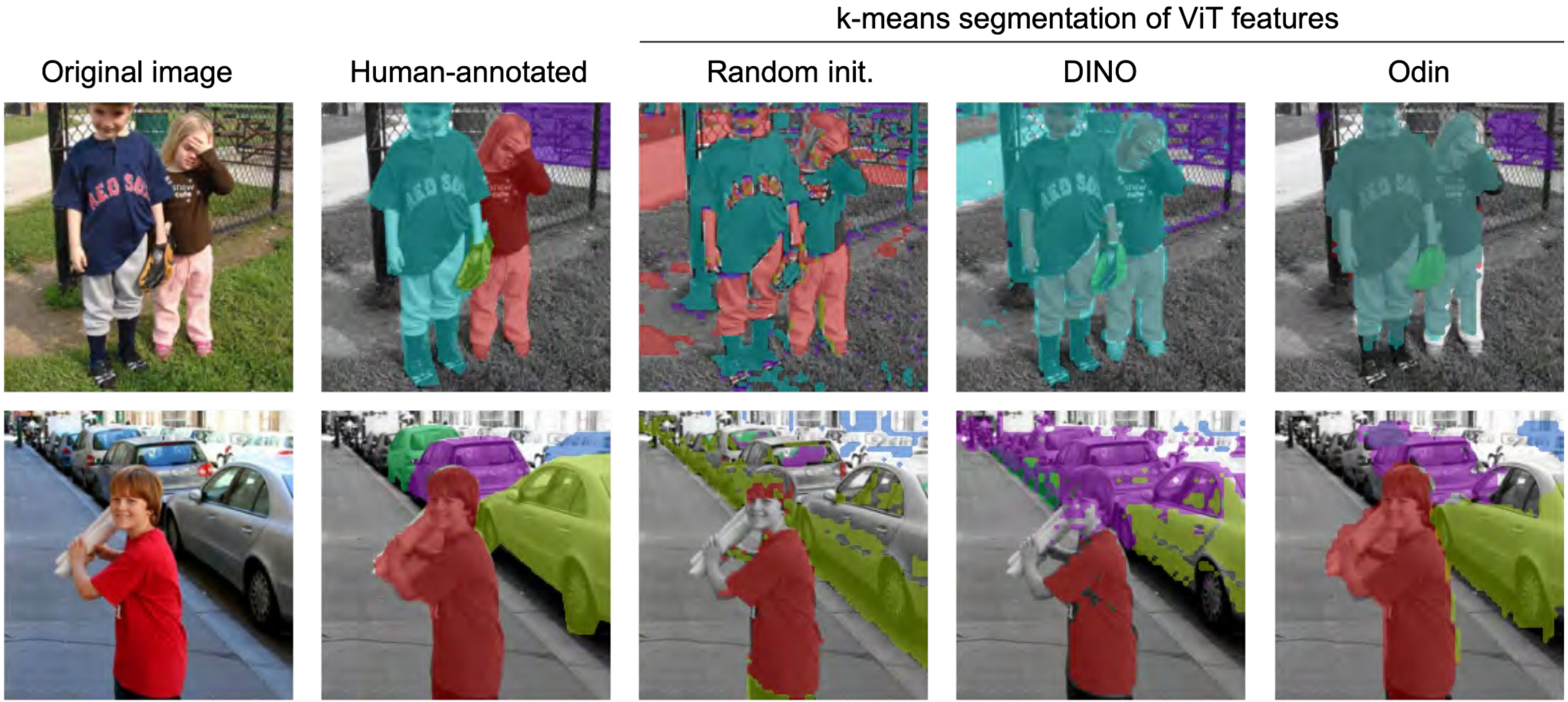
Object discovery

- Pretrain ViT-B/8 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap

	ResNet-50			ViT-B		
Pretraining	ABO^i	ABO^c	OR	ABO^i	ABO^c	OR
Random init	28.1	33.6	17.0	27.8	33.6	17.5
DetCon_B [42]	34.1	40.0	20.4	-	-	-
Supervised	35.8	41.1	23.8	43.9	53.6	41.9
DINO [12]	38.3	46.5	30.8	42.7	51.7	39.7
Odin	**41.5**	**48.6**	**36.5**	**45.9**	**53.9**	**44.1**
Odin†	**43.0**	**53.0**	**42.3**			

Object discovery

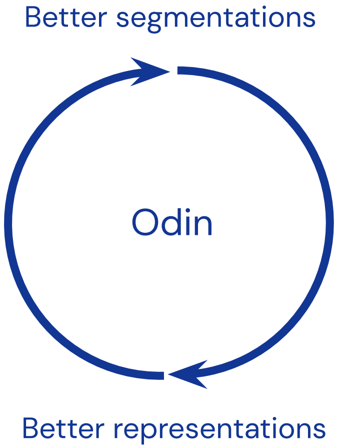
- Pretrain ViT-B/8 on ImageNet using self-supervised objective
- Evaluate on COCO, cluster features using k-means, report best overlap



Object knowledge can be extracted from learned representations

Self-supervised pretraining

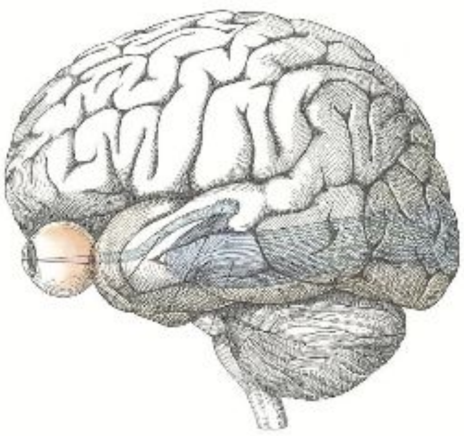
Fine-tune for object detection, segmentation



pretraining		fine-tune 1×	
Method	Knows obj?	AP^{bb}	AP^{mk}
Supervised	no	39.6	35.6
VADeR [68]	no	39.2	35.6
MoCo [37]	no	39.4	35.6
SimCLR [13]	no	39.7	35.8
MoCo v2 [14]	no	40.1	36.3
InfoMin [75]	no	40.6	36.7
DINO [12]	no	41.2	37.1
ReLIC v2 [77]	yes	42.5	38.0
DetCon_B [40]	yes	42.7	38.2
Odin	no	**42.9**	**38.4**

Can SSL leverage videos to learn good image representations?

Self-supervised pretraining



Natural videos provide

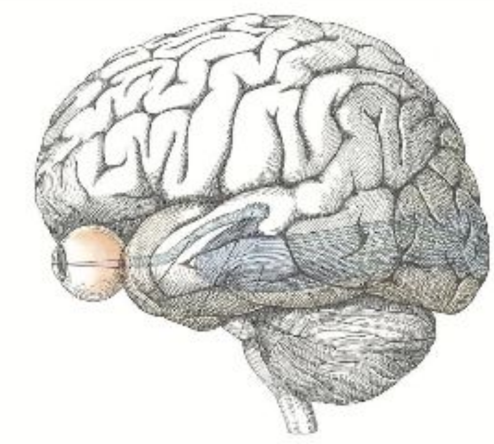
- Rich image augmentations

- Strong segmentation cues from motion

Can SSL leverage videos to learn good image representations?

Self-supervised pretraining

Fine-tune for object detection, segmentation



Natural videos provide

- Rich image augmentations
- Strong segmentation cues from motion

Yet they have yet to yield strong image representations!

(as evaluated by scene understanding tasks)

DeepMind

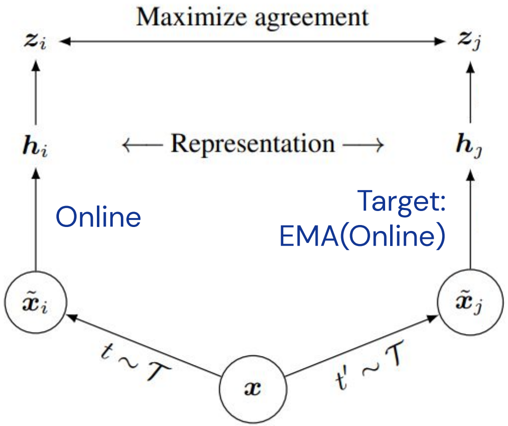
Self-supervised video pretraining yields strong image representations

Nikhil Parthasarathy, Ali Eslami, João Carreira, Olivier Hénaff

arXiv 2022

Can SSL leverage videos to learn good image representations?

Self-supervised pretraining on images or video



A strong contrastive baseline

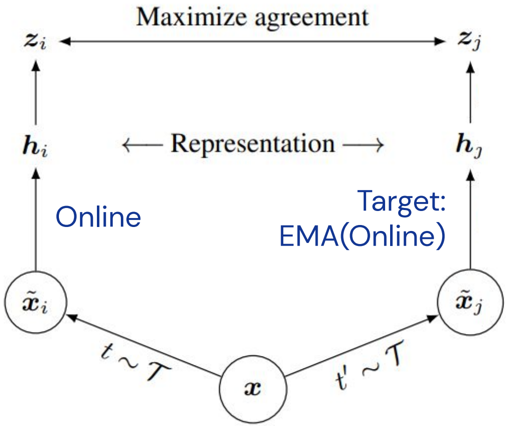
- MoCLR: best of SimCLR, MoCo, and BYOL (Tian, 2021)

- Frame x sampled from image datasets or videos

Can SSL leverage videos to learn good image representations?

Self-supervised pretraining on images or video

Fine-tune for object detection, segmentation



A strong contrastive baseline

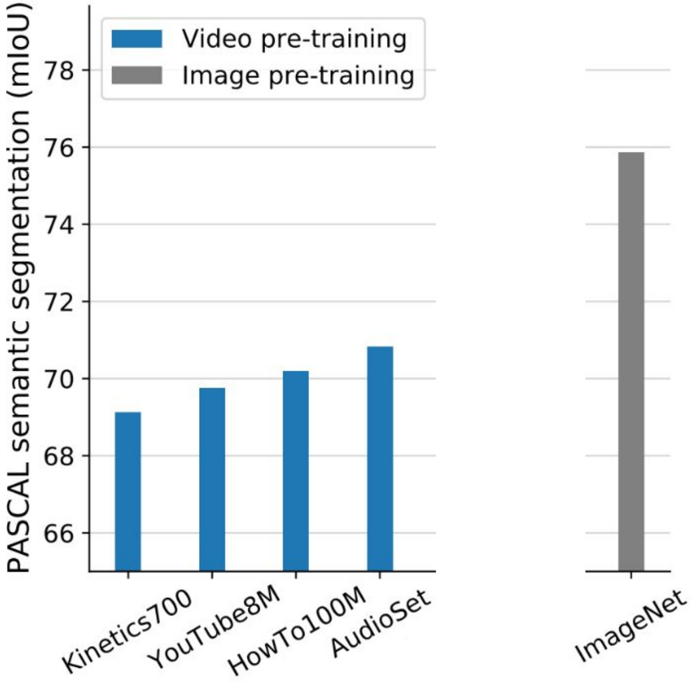
- MoCLR: best of SimCLR, MoCo, and BYOL (Tian, 2021)
- Frame x sampled from image datasets or videos

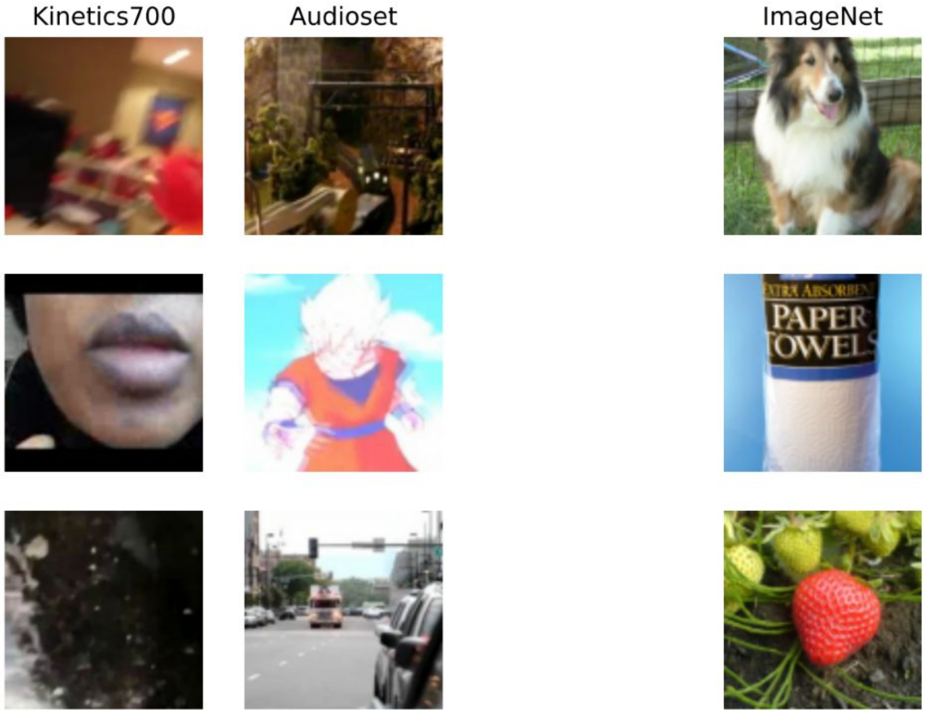
Fine-tune on

- Semantic segmentation (PASCAL or ADE20K)
- Object detection (COCO or LVIS)

Dataset curation matters

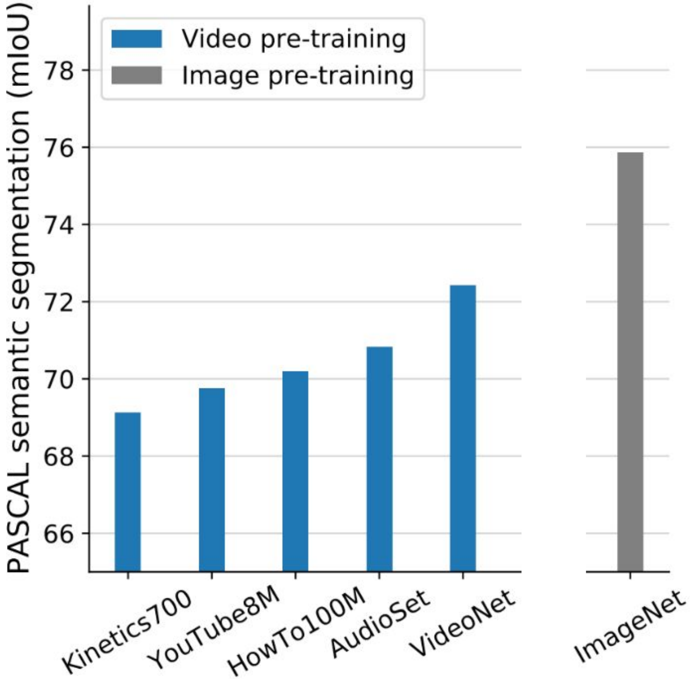
Treating video datasets as collections of independent frames

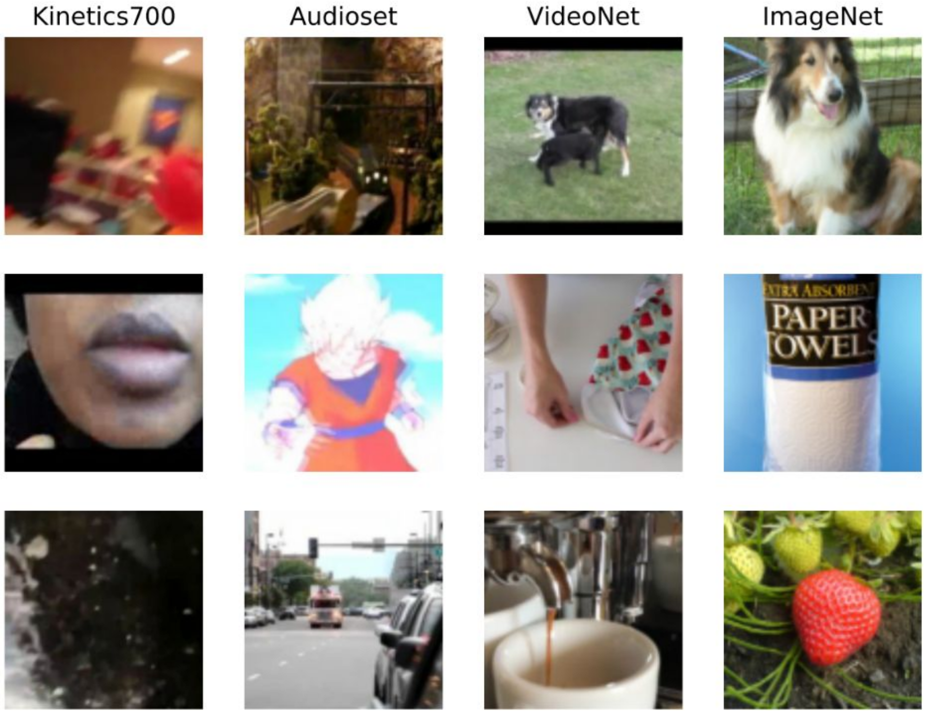




Dataset curation matters

Treating video datasets as collections of independent frames





VideoNet data pipeline

Procedure for filtering uncurated video datasets

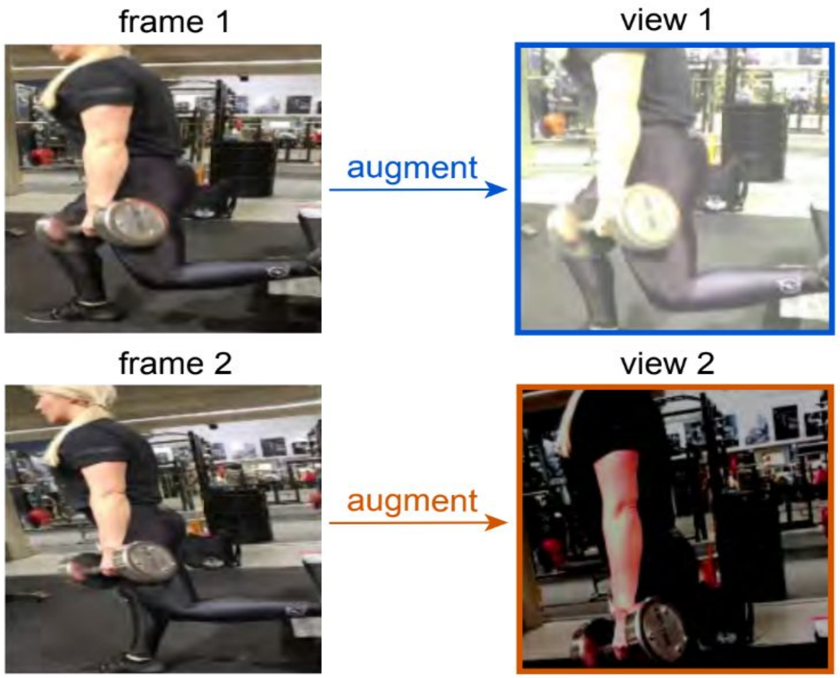
1) Query internet for ImageNet categories **(~5 million videos)**

2) Filter out videos less than 10s long

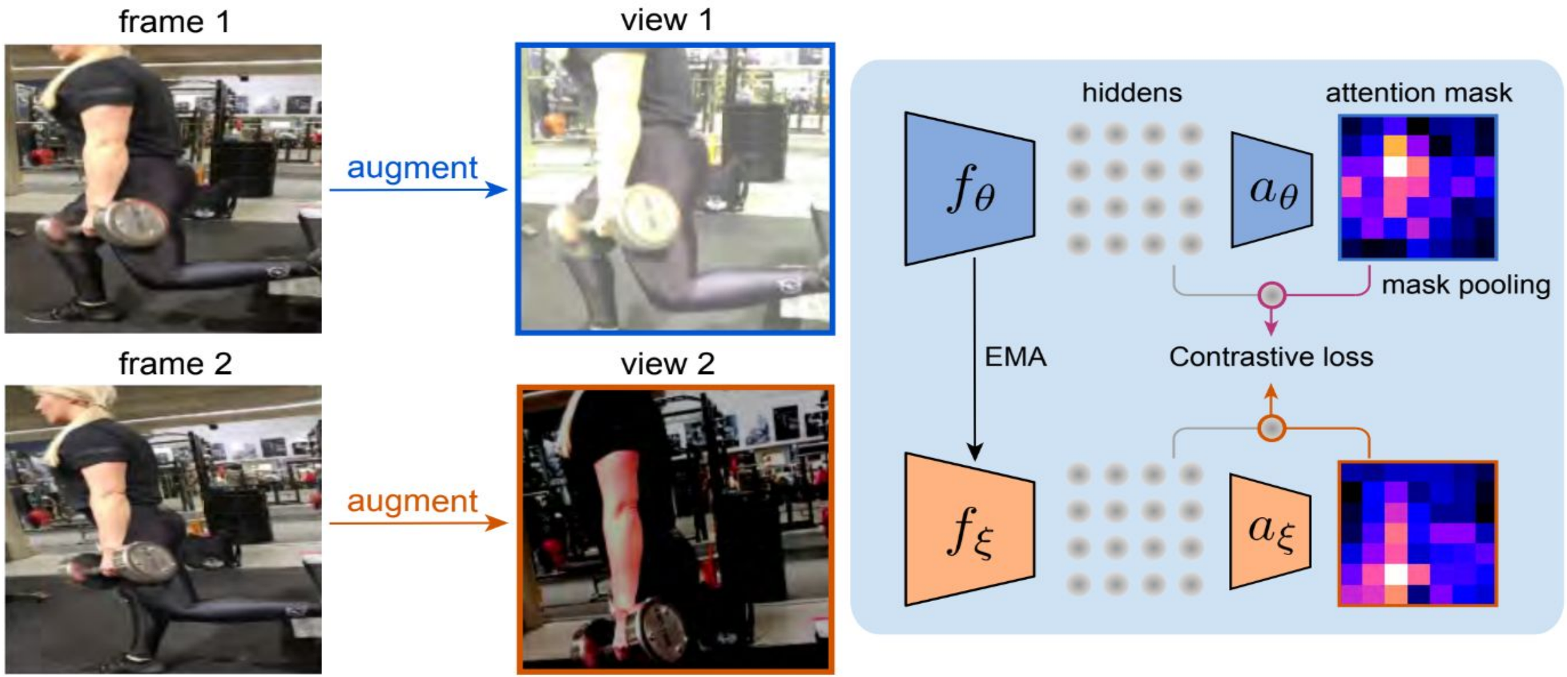
3) Run an ImageNet classifier on the first 100 frames of each video to verify they contain stated category **(~1.2 million videos)**

VITO: a better video-to-image model

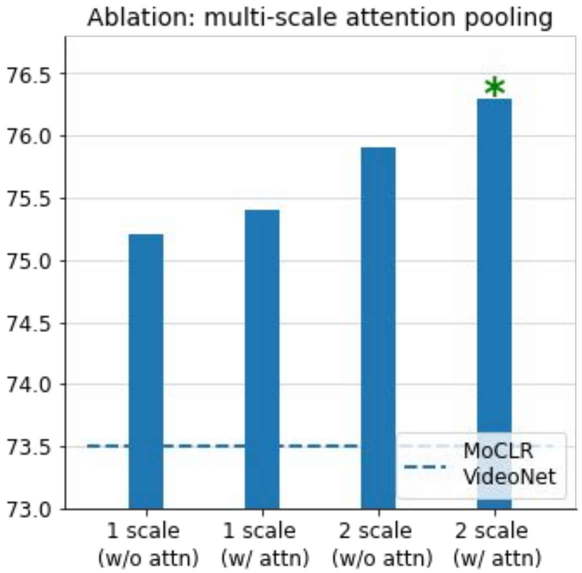
VITO: a better video-to-image model



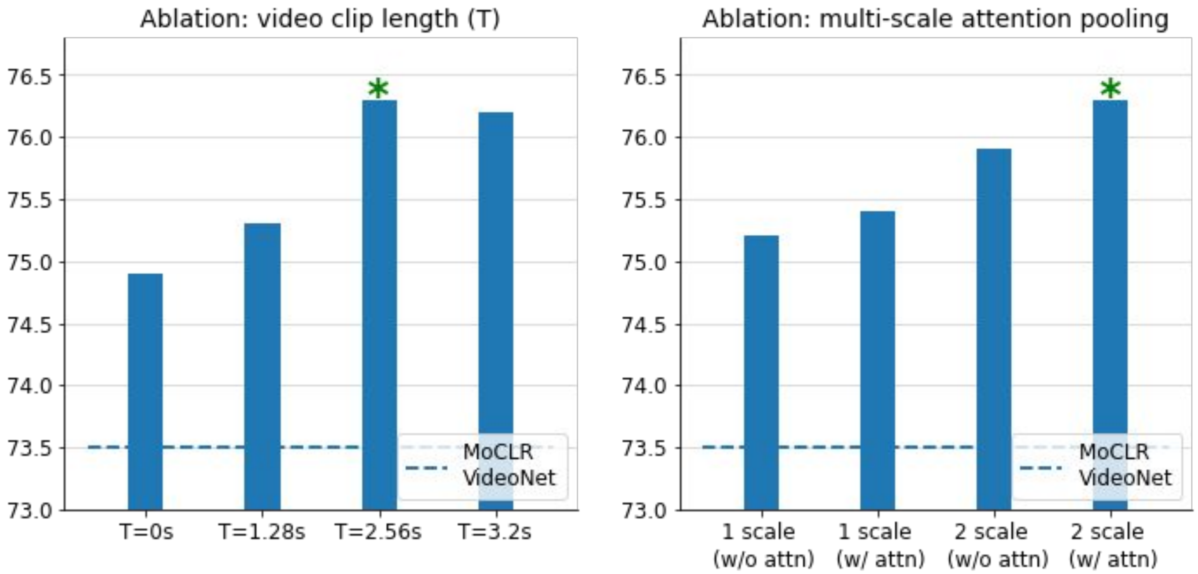
VITO: a better video-to-image model



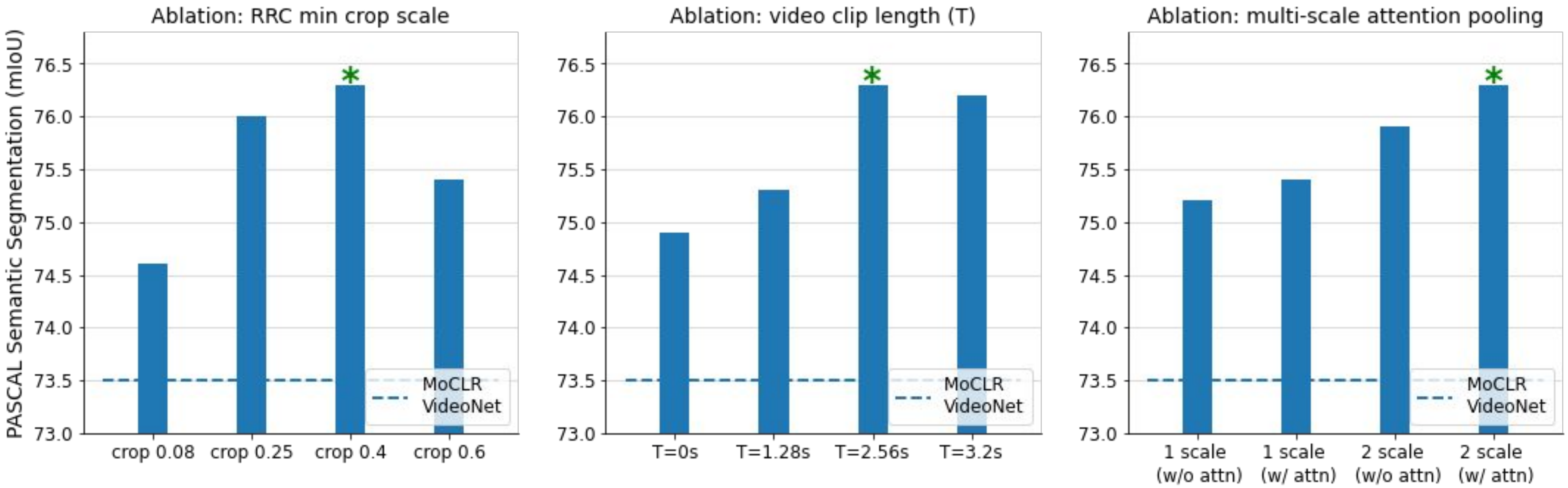
Results: ablating the components of VITO



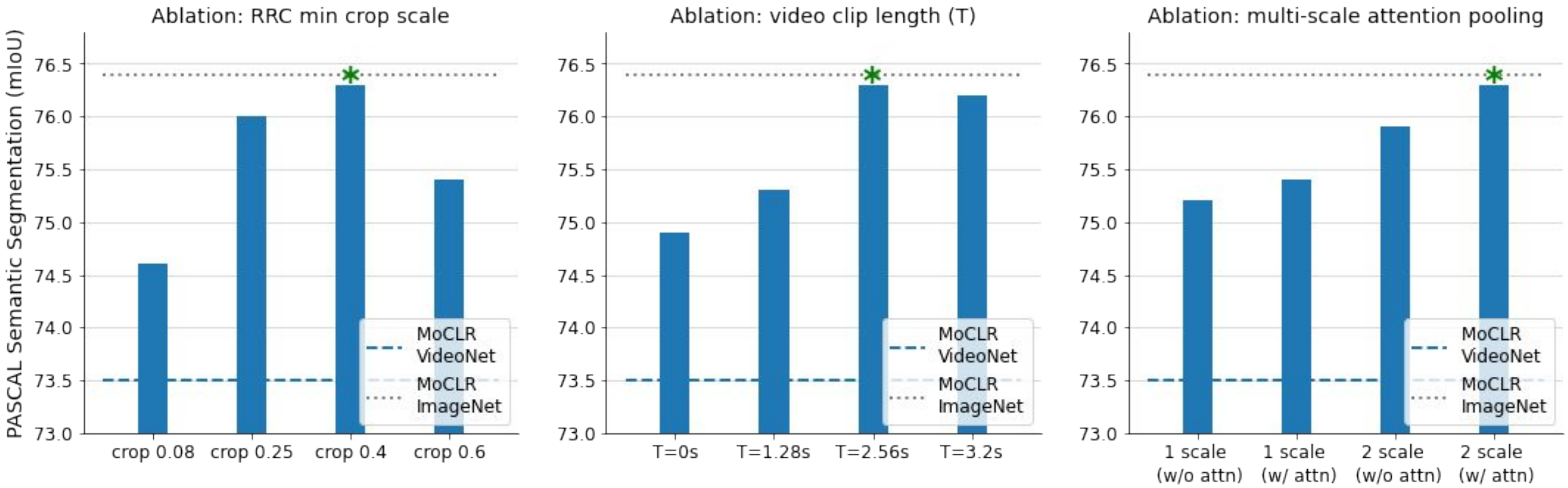
Results: ablating the components of VITO



Results: ablating the components of VITO



Results: ablating the components of VITO



VITO closes the gap with ImageNet MoCLR

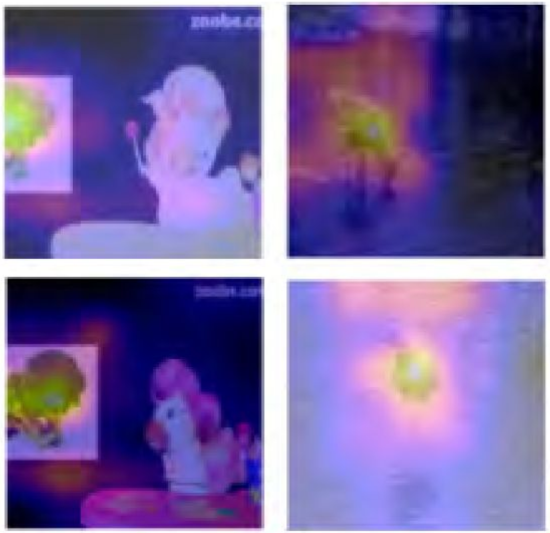
Results: what does the attention pooling learn?

Due to the contrastive loss, VITO must learn to attend to content that is:

- **Stable (or predictable)** over time
- **Unique or discriminative** relative to content from other videos

View 1

View 2



Results: what does the attention pooling learn?

Due to the contrastive loss, VITO must learn to attend to content that is:

- **Stable (or predictable)** over time
- **Unique or discriminative** relative to content from other videos

This also enables **semantic binding of content**

View 1

View 2

Results: what does the attention pooling learn?

Results: comparison to prior video-to-image methods

Pretraining	Dataset	Epochs	Semantic segmentation PASCAL	Semantic segmentation ADE20K	Object detection COCO	Object detection LVIS
Random Init			53.0	27.9	39.0	21.1
Methods pretraining on video datasets						
VFS (Xu & Wang, 2021)	K400	100	63.9	31.4	41.6	23.2
VIVI (Tschannen et al., 2020)	YouTube8M	192	65.8	34.2	41.3	23.2
VINCE (Gordon et al., 2020)	R2V2	200	69.0	35.7	42.4	24.4
CycleContrast (Wu & Wang, 2021)	R2V2	200	69.2	35.6	42.8	24.5
VITO	VideoNet	200	**75.5**	**39.2**	**43.6**	**25.6**

VITO vs. other frame-level SSL objectives highlights importance of VITO components

→ data curation, attention pooling, random-crop scale

Results: comparison to prior video-to-image methods

			Semantic segmentation		Object detection	
Pretraining	Dataset	Epochs	PASCAL	ADE20K	COCO	LVIS
Random Init			53.0	27.9	39.0	21.1
Methods pretraining on video datasets						
VFS (Xu & Wang, 2021)	K400	100	63.9	31.4	41.6	23.2
VIVI (Tschannen et al., 2020)	YouTube8M	192	65.8	34.2	41.3	23.2
VINCE (Gordon et al., 2020)	R2V2	200	69.0	35.7	42.4	24.4
CycleContrast (Wu & Wang, 2021)	R2V2	200	69.2	35.6	42.8	24.5
MMV (Alayrac et al., 2020)	AS + HT	1600	70.6	32.5	41.3	24.2
VITO	VideoNet	200	**75.5**	**39.2**	**43.6**	**25.6**

VITO vs. other frame-level SSL objectives highlights importance of VITO components

→ data curation, attention pooling, random-crop scale

VITO also outperforms recent multimodal methods

Results: comparison to prior video-to-image methods

Pretraining	Dataset	Epochs	Semantic segmentation		Object detection	
			PASCAL	ADE20K	COCO	LVIS
Random Init			53.0	27.9	39.0	21.1
Methods pretraining on video datasets						
VFS (Xu & Wang, 2021)	K400	100	63.9	31.4	41.6	23.2
VIVI (Tschannen et al., 2020)	YouTube8M	192	65.8	34.2	41.3	23.2
VINCE (Gordon et al., 2020)	R2V2	200	69.0	35.7	42.4	24.4
CycleContrast (Wu & Wang, 2021)	R2V2	200	69.2	35.6	42.8	24.5
MMV (Alayrac et al., 2020)	AS + HT	1600	70.6	32.5	41.3	24.2
VITO	VideoNet	200	**75.5**	**39.2**	**43.6**	**25.6**
VITO	AudioSet	300	73.6	38.5	43.2	25.0
VITO	VideoNet	300	**76.3**	**39.4**	**44.0**	**25.7**

VITO also outperforms prior art with existing video datasets (Audioset), but VideoNet provides further gains

Results: VITO closes the gap with ImageNet pretraining

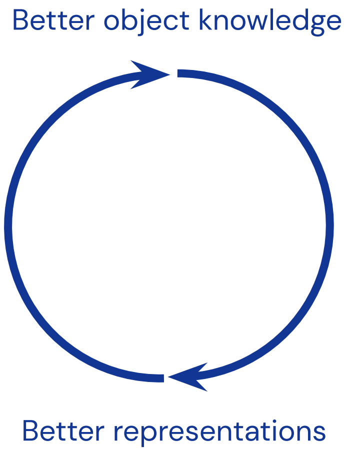
			Semantic segmentation		Object detection	
Pretraining	Dataset	Epochs	PASCAL	ADE20K	COCO	LVIS
Random Init			53.0	27.9	39.0	21.1
Methods pretraining on video datasets						
VFS (Xu & Wang, 2021)	K400	100	63.9	31.4	41.6	23.2
VIVI (Tschannen et al., 2020)	YouTube8M	192	65.8	34.2	41.3	23.2
VINCE (Gordon et al., 2020)	R2V2	200	69.0	35.7	42.4	24.4
CycleContrast (Wu & Wang, 2021)	R2V2	200	69.2	35.6	42.8	24.5
MMV (Alayrac et al., 2020)	AS + HT	1600	70.6	32.5	41.3	24.2
VITO	VideoNet	300	**76.3**	**39.4**	**44.0**	**25.7**
Methods pretraining on ImageNet						
Supervised	ImageNet	200	71.3	33.5	44.2	25.2
BYOL (Grill et al., 2020)	ImageNet	300	76.1	38.8	43.7	25.5
MoCLR (Tian et al., 2021)	ImageNet	300	76.4	39.2	43.9	25.8
DINO (Caron et al., 2021)	ImageNet	300	76.1	39.0	44.3	26.4

Conclusion

1. Knowledge of objects accelerates and improves representation learning
   → DetCon objective (ICCV, 2021)

2. Knowledge of objects can be extracted from learned representations
   → Odin framework (ECCV, 2022)

3. Videos can be used to learn strong image representations
   → VITO framework (arXiv, 2022)



DeepMind

Thanks!