Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill[*,1] **Florian Strub**[*,1] **Florent Altché**[*,1] **Corentin Tallec**[*,1] **Pierre H. Richemond**[*,1,2]
Elena Buchatskaya[1] **Carl Doersch**[1] **Bernardo Avila Pires**[1] **Zhaohan Daniel Guo**[1]
Mohammad Gheshlaghi Azar[1] **Bilal Piot**[1] **Koray Kavukcuoglu**[1] **Rémi Munos**[1] **Michal Valko**[1]

[1]DeepMind [2]Imperial College

arxiv.org/abs/2006.07733

Self-supverised Representation Learning

also called Unsupervised Representation Learning

The goal is to learn features that:

- Map similiar sematics closer
- Transferrable to downstream tasks

The key is to generate 'labels' from the data by *pretext tasks:*

- Predictive Pretext Tasks (Rotation, Jigsaws, Colorization, *etc*)
- Constrative Pretext Tasks (Instance discrimination)

Constrative Learning: *Similar to metric learning*

Minimize the distance between positive pairs
Maintain the distance between negative pairs

Query	Positive	Negative
Image A	**Augmented Image A**	**Image B**
Patch A	Tracked Patch A in Video	Random Patch B
Image A Channel A	Image A Channel B	Image B Channel B



Constrative Learning: *Similar to metric learning*

Avoid collapse
Suppose a constant representation

Minimize the distance between positive pairs
Maintain the distance between negative pairs

Query	Positive	Negative
Image A	**Augmented Image A**	**Image B**
Patch A	Tracked Patch A in Video	Random Patch B
Image A Channel A	Image A Channel B	Image B Channel B

Constrative Learning: *Similar to metric learning*

Minimize the distance between positive pairs
Maintain the distance between negative pairs

Query	Positive	Negative
Image A	**Augmented Image A**	**Image B**
Patch A	Tracked Patch A in Video	Random Patch B
Image A Channel A	Image A Channel B	Image B Channel B



BYOL achieves a new state-of-the-art without using negative pairs.

Constrative Learning: *Similar to metric learning*

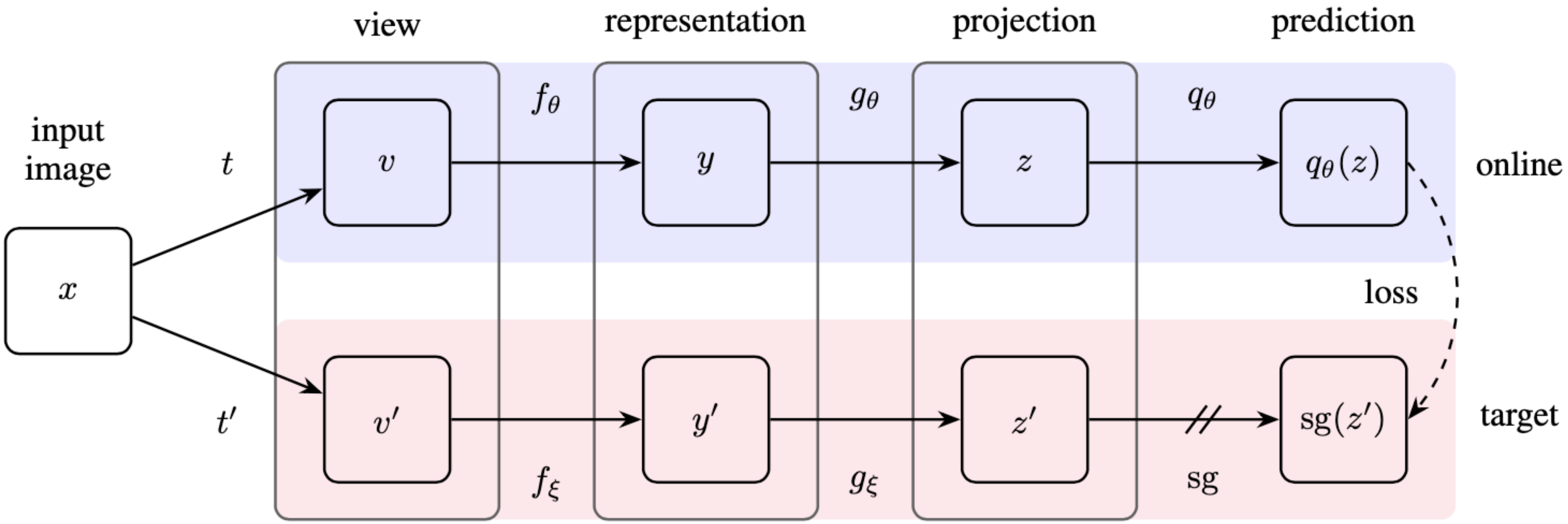
Minimize the distance between positive pairs
Maintain the distance between negative pairs

Query	Positive	Negative
Image A	**Augmented Image A**	**Image B**
Patch A	Tracked Patch A in Video	Random Patch B
Image A Channel A	Image A Channel B	Image B Channel B

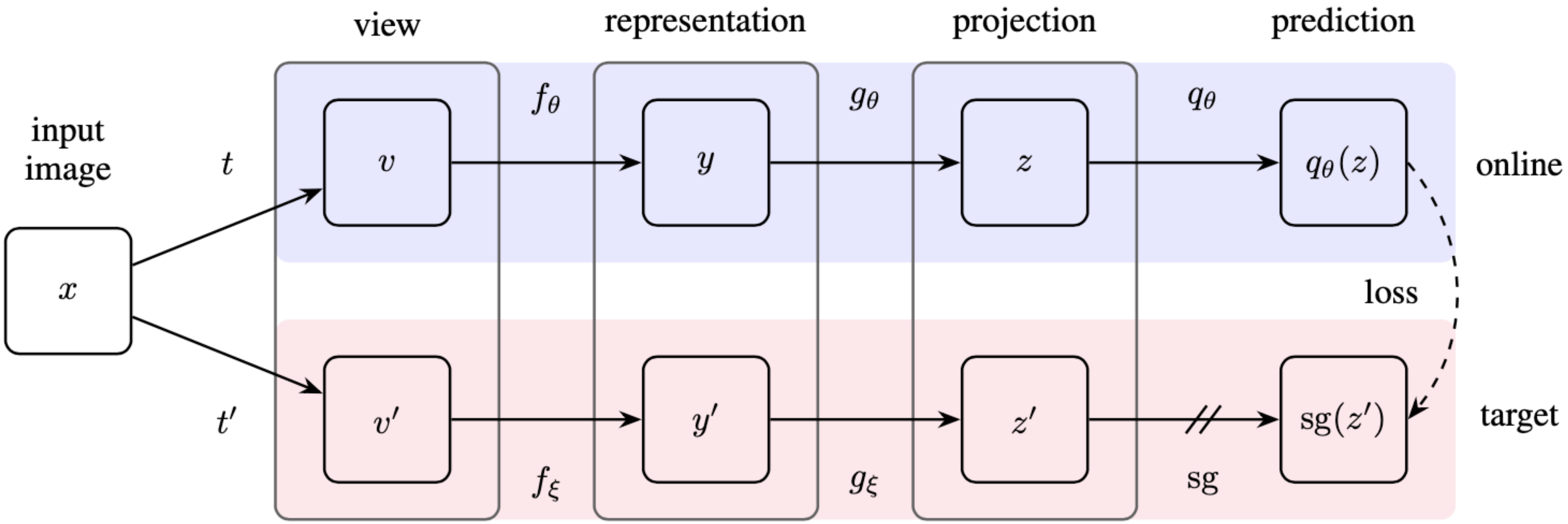
And no collapse! Like magic…

BYOL achieves a new state-of-the-art without using negative pairs.

Method: *Very Simple!*

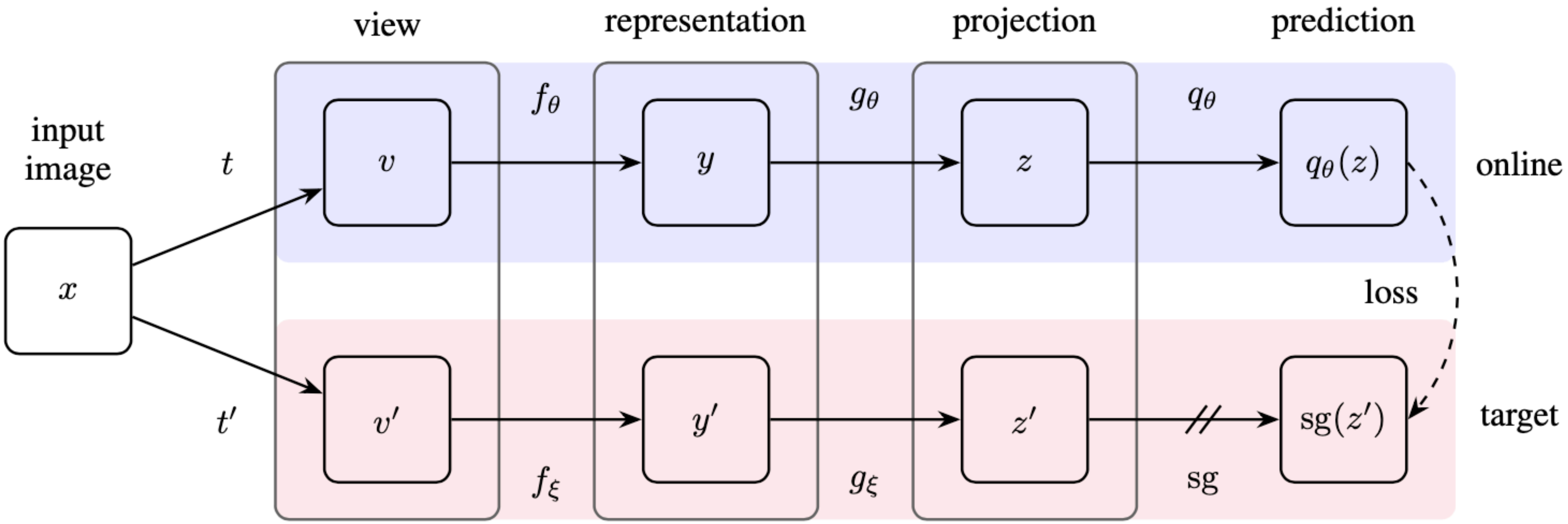


Method: *Very Simple!*



f *are CNNs,* ***g*** *and* ***q*** *are MLPs ,* ***sg*** *is stopping gradient*

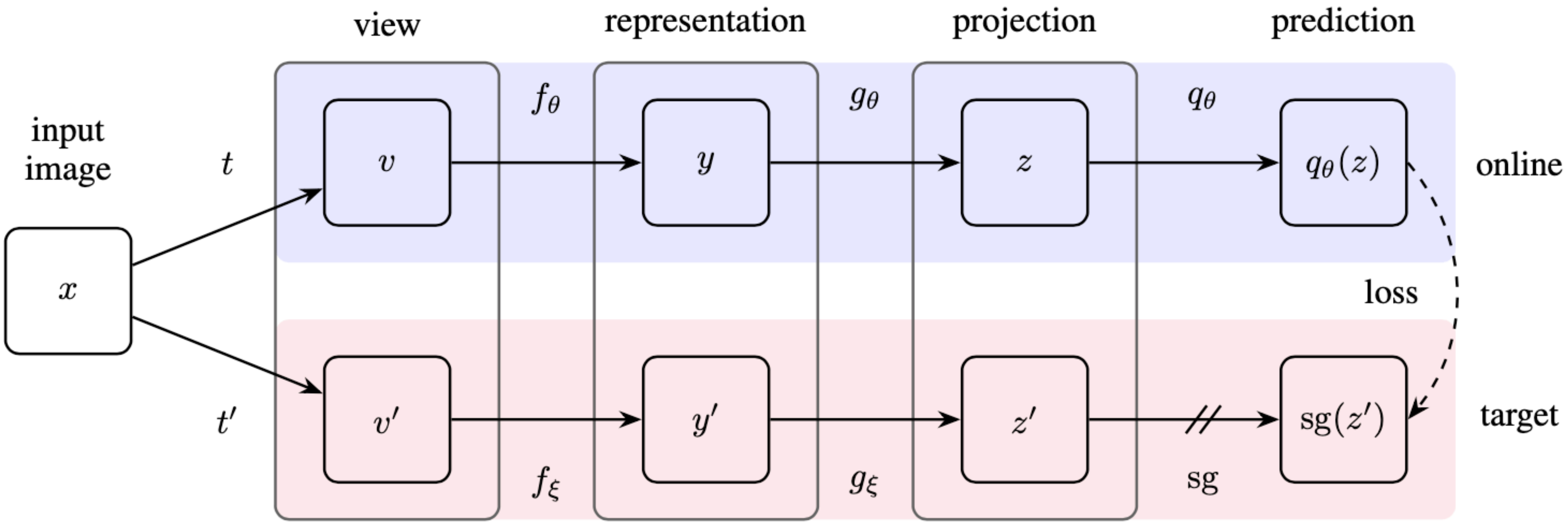
Method: *Very Simple!*



MSE Loss between normalized 'features'

$$\mathcal{L}_\theta^{\mathtt{BYOL}} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2$$

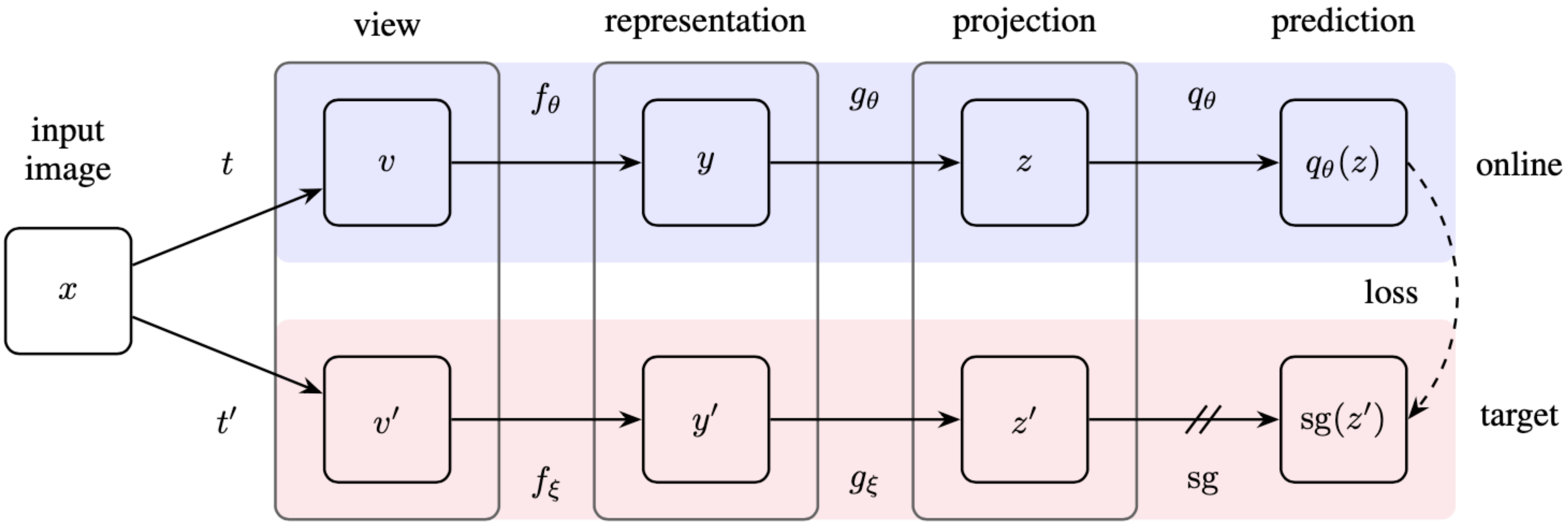
Method: *Very Simple!*



Mean teacher as target network

$$\xi \leftarrow \tau\xi + (1 - \tau)\theta.$$

Method: *Very Simple!*



Use f_theta as learned representation

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Randomly initialized network as target: ***18.8% accuracy on ImageNet***

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Randomly initialized network as target: ***18.8% accuracy on ImageNet***

Training a new network to predict a given target will produce enhenced representation…

So …

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Randomly initialized network as target: ***18.8% accuracy on ImageNet***

Training a new network to predict a given target will produce enhenced representation…

So what if build a sequence of representation, using the current network as the target of the subsequent network?

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Randomly initialized network as target: ***18.8% accuracy on ImageNet***

Training a new network to predict a given target will produce enhenced representation…

So what if build a sequence of representation, using the current network as the target of the subsequent network?

A >> B >> C >> D >> …

Motivation Behind

Randomly initialized network: ***1.4% accuracy on ImageNet***

Randomly initialized network as target: ***18.8% accuracy on ImageNet***

Training a new network to predict a given target will produce enhenced representation…

So what if build a sequence of representation, using the current network as the target of the subsequent network?

A >> B >> C >> D >> … *No experiment for this, Maybe hard to tune…*

Motivation Behind

Randomly initialized network: *1.4% accuracy on ImageNet*

Randomly initialized network as target: *18.8% accuracy on ImageNet*

Training a new network to predict a given target will produce enhenced representation…

So what if build a sequence of representation, using the current network as the target of the subsequent network?

A >> B >> C >> D >> … *No experiment for this, Maybe hard to tune…*

Use mean teacher as the target.

Sadly, why no collapse is not explained…

Experiments

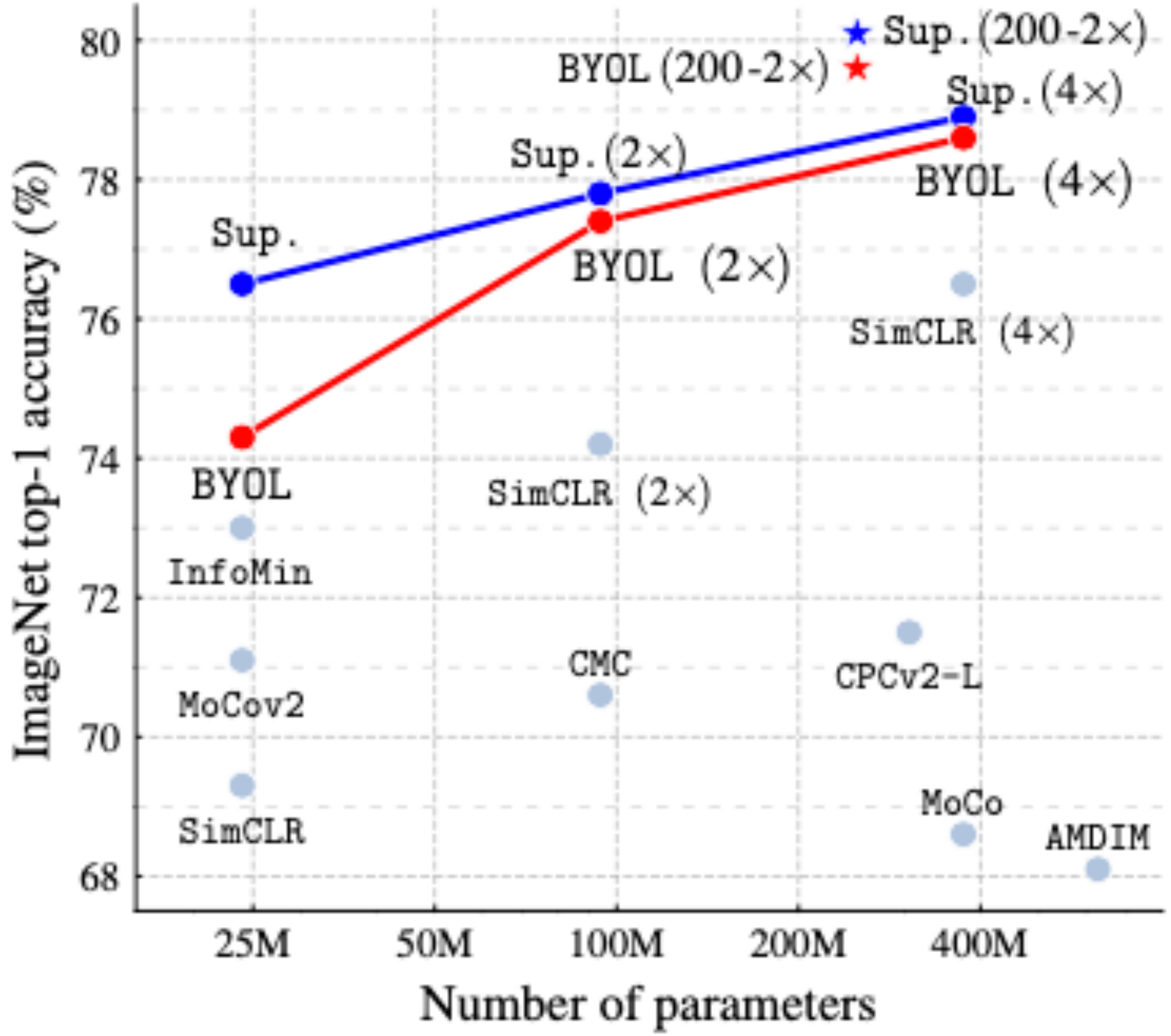


Figure 1: Performance of BYOL on ImageNet (linear evaluation) using ResNet-50 and our best architecture ResNet-200 (2×), compared to other unsupervised and supervised (Sup.) baselines [8].

Experiments

Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [32]	63.6	-
CPC v2 [29]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [34]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [29]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL (ours)	ResNet-50 (4×)	375M	78.6	94.2
BYOL (ours)	ResNet-200 (2×)	250M	**79.6**	**94.8**

(b) Other ResNet encoder architectures.

Table 1: Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet.

Experiments

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised [64]	25.4	56.4	48.4	80.4
InstDisc	-	-	39.2	77.4
PIRL [32]	-	-	57.2	83.8
SimCLR [8]	48.3	65.6	75.5	87.8
BYOL (ours)	**53.2**	**68.8**	**78.4**	**89.0**

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1		Top-5	
			1%	10%	1%	10%
CPC v2 [29]	ResNet-161	305M	-	-	77.9	91.2
SimCLR [8]	ResNet-50 (2×)	94M	58.5	71.7	83.0	91.2
BYOL (ours)	ResNet-50 (2×)	94M	62.2	73.5	84.1	91.7
SimCLR [8]	ResNet-50 (4×)	375M	63.0	74.4	85.8	92.6
BYOL (ours)	ResNet-50 (4×)	375M	69.1	75.7	87.9	92.5
BYOL (ours)	ResNet-200 (2×)	250M	**71.2**	**77.7**	**89.5**	**93.7**

(b) Other ResNet encoder architectures.

Table 2: Semi-supervised training with a fraction of ImageNet labels.

Experiments

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
SimCLR [8]	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
Supervised-IN [8]	72.3	**93.6**	78.3	53.7	61.9	66.7	**61.0**	**82.8**	74.9	**91.5**	**94.5**	94.7
Fine-tuned:												
BYOL (ours)	**88.5**	**97.8**	86.1	**76.3**	63.7	91.6	**88.1**	**85.4**	**76.2**	91.7	**93.8**	97.0
SimCLR (repro)	87.5	97.4	85.3	75.0	63.9	91.4	87.6	84.5	75.4	89.4	91.7	96.6
SimCLR [8]	88.2	97.7	85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97.0
Supervised-IN [8]	88.3	97.5	**86.4**	75.8	**64.3**	**92.1**	86.0	85.0	74.6	**92.1**	93.3	**97.6**
Random init [8]	86.9	95.9	80.2	76.1	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.

Method	AP_{50}	mIoU
Supervised-IN [9]	74.4	74.4
MoCo [9]	74.9	72.5
SimCLR (repro)	75.2	75.2
BYOL (ours)	**77.5**	**76.3**

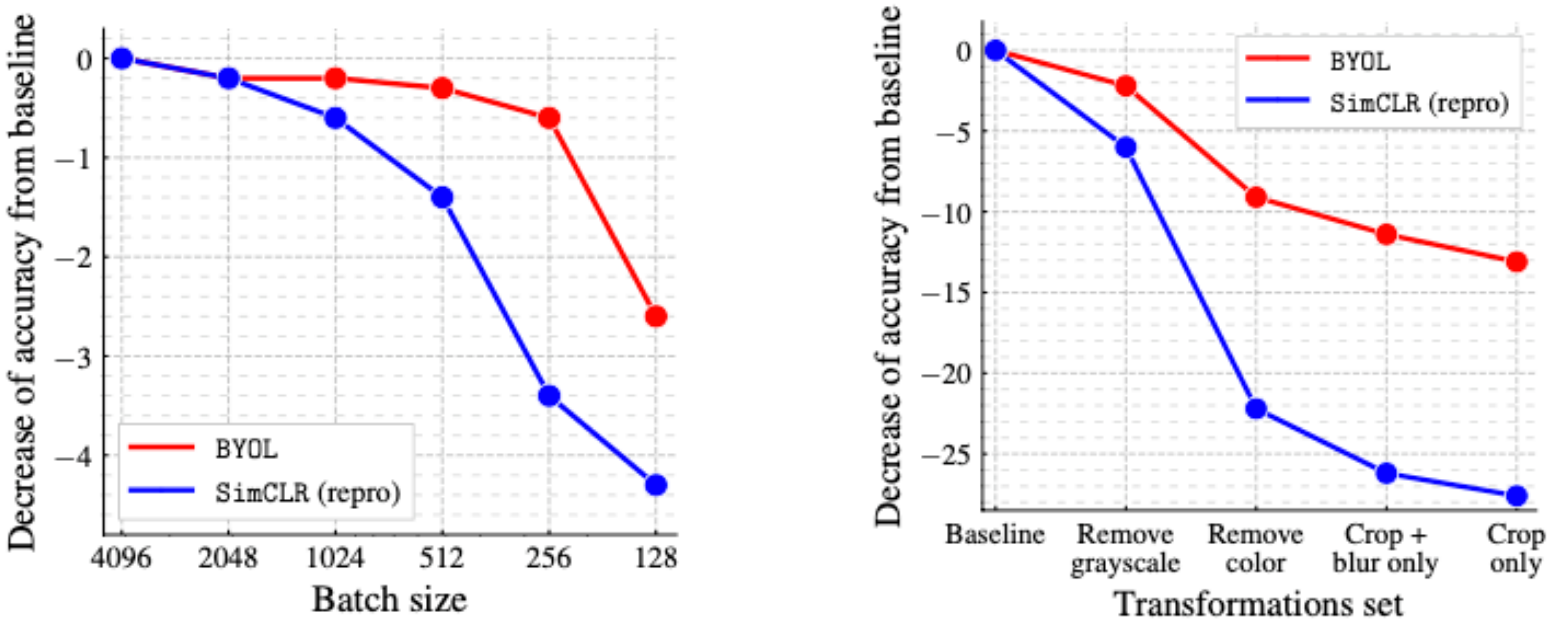
(a) Transfer results in semantic segmentation and object detection.

	Higher better			Lower better	
Method	pct.<1.25	pct.$<1.25^2$	pct.$<1.25^3$	rms	rel
Supervised-IN [70]	81.1	95.3	98.8	0.573	**0.127**
SimCLR (repro)	83.3	96.5	99.1	0.557	0.134
BYOL (ours)	**84.6**	**96.7**	**99.1**	**0.541**	0.129

(b) Transfer results on NYU v2 depth estimation.

Table 4: Results on transferring BYOL's representation to other vision tasks.

Experiments



(a) Impact of batch size

(b) Impact of progressively removing transformations

Figure 3: Decrease in top-1 accuracy (in % points) of BYOL and our own reproduction of SimCLR at 300 epochs, under linear evaluation on ImageNet.

8h x 512 TPUs…

Comment:

- Simple, Effective, maybe Delicate
- Unsupervised learning is the ***near*** future
- Augmentation matters

Guess why no collapse:

1) The initialization is closer to the better representation than the collapsed one.
Deep image prior.

Good representation = Deep image prior + Ignorng non-semantics?

*2) The **mean teacher** provides a super delicate balance to avoid collapse.*
Initialization is not collapsed and mean teacher maintains it well.

3) Batch norm scatters samples.
4) Bless of dimensionality.