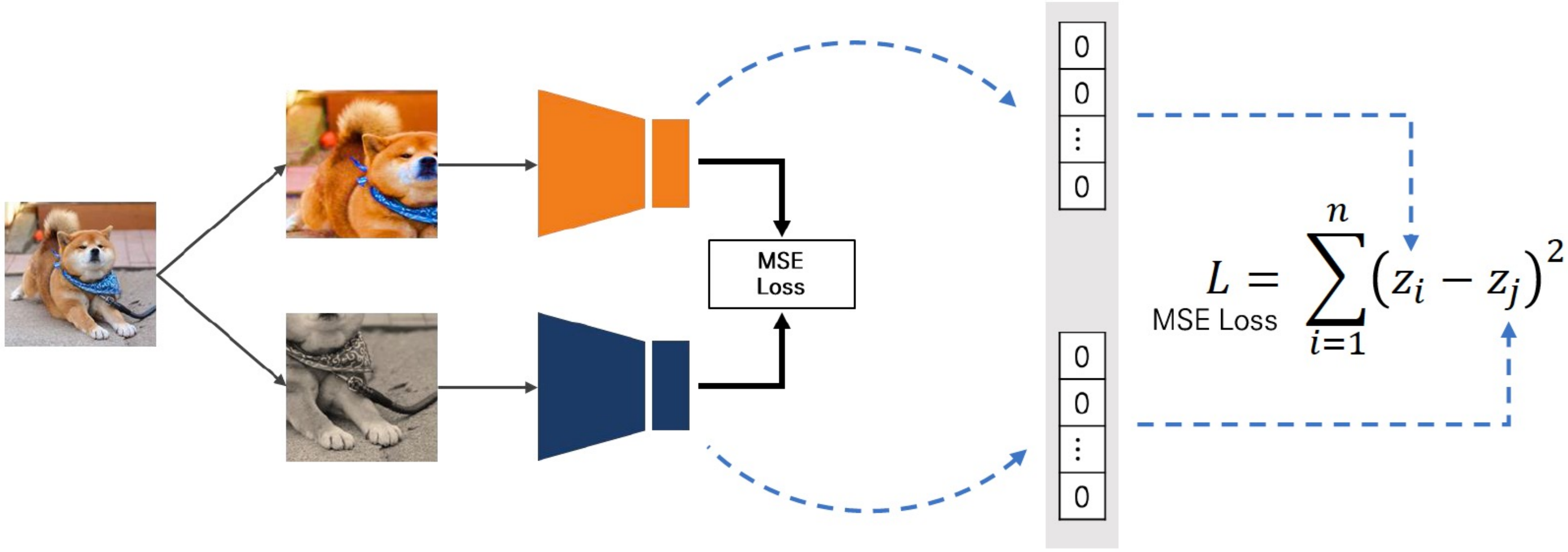
Bootstrap Your Own Latent

A New Approach to Self-Supervised Learning

조 원 양

Self-supervised learning

- Cross-view prediction framework에 기반



Cross-view prediction framework Constant Vector

- Collapsed representation 방지→ negative pairs

Contrastive loss

- collapased representation 방지

$$L_{i,j} = -\log \frac{\exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{\sum_{k=1\,[k\neq i]}^{N} \exp\left(\frac{sim(z_i, z_k)}{\tau}\right)}$$

Contrastive Loss

Cosine similarity
(Positive pair)

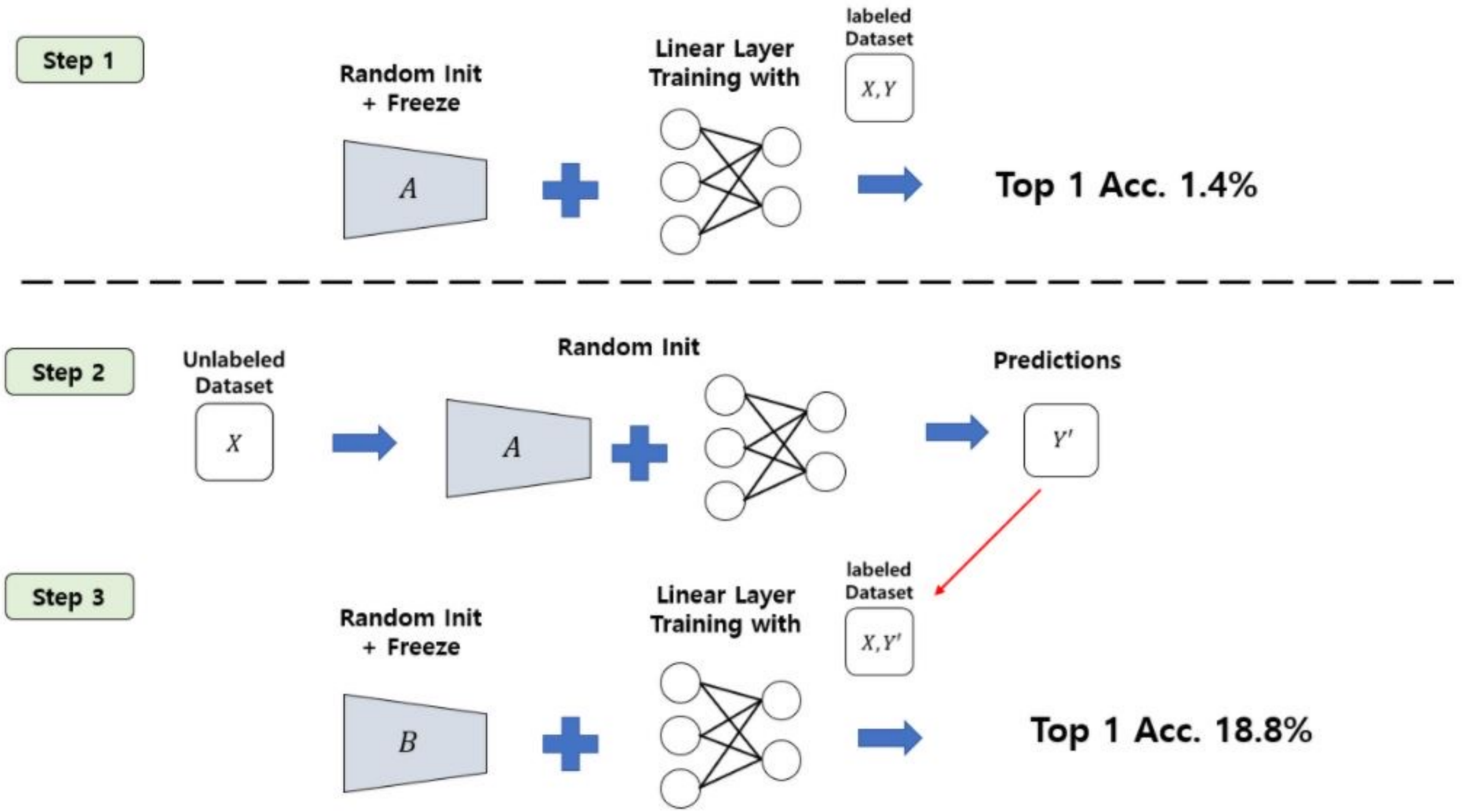
Cosine similarity
(Negative pair)

Contrastive Learning의 문제점

- Negative pairs 처리의 중요성
  - Large batch size(SimCLR), memory banks(MoCo) 그리고 customized mining strategies

- Data augmentation 선택의 중요성

Motivation



Step 1
Random Init
+ Freeze
A
Linear Layer
Training with
labeled
Dataset
X, Y
Top 1 Acc. 1.4%
Step 2
Unlabeled
Dataset
X
Random Init
A
Predictions
Y'
Step 3
Random Init
+ Freeze
B
Linear Layer
Training with
labeled
Dataset
X, Y'
Top 1 Acc. 18.8%

Architecture

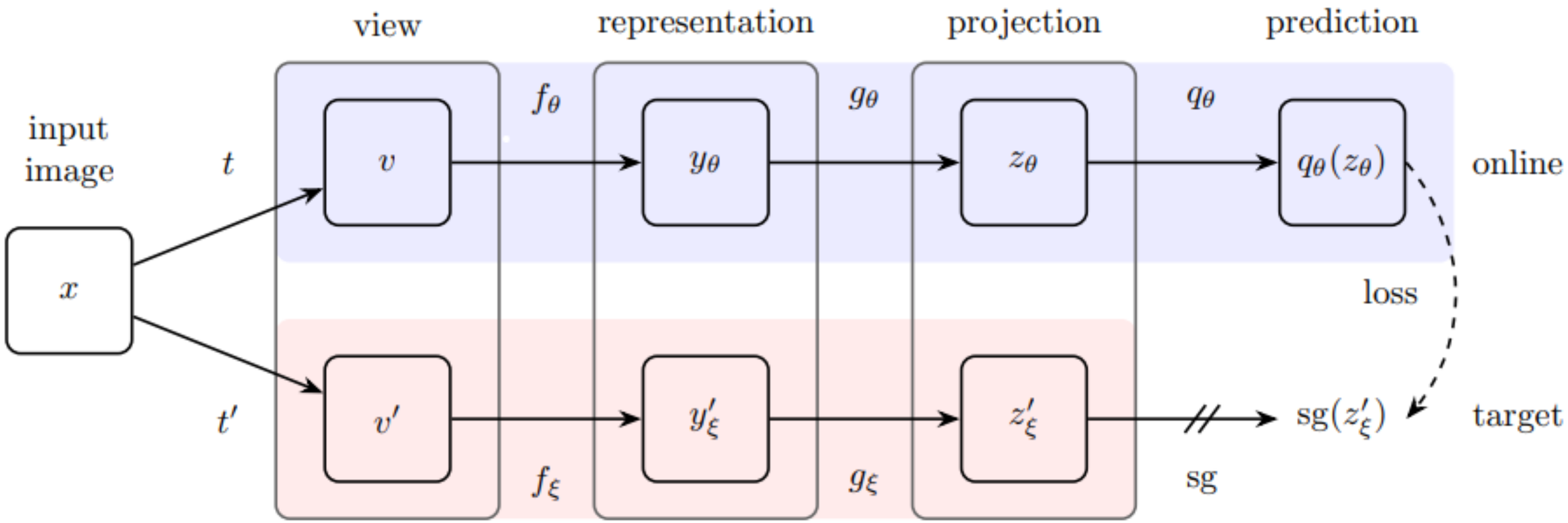


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

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

- Prediction
- slow-moving average

- encoding more information within online projection
- avoid collapsed representation

Architecture

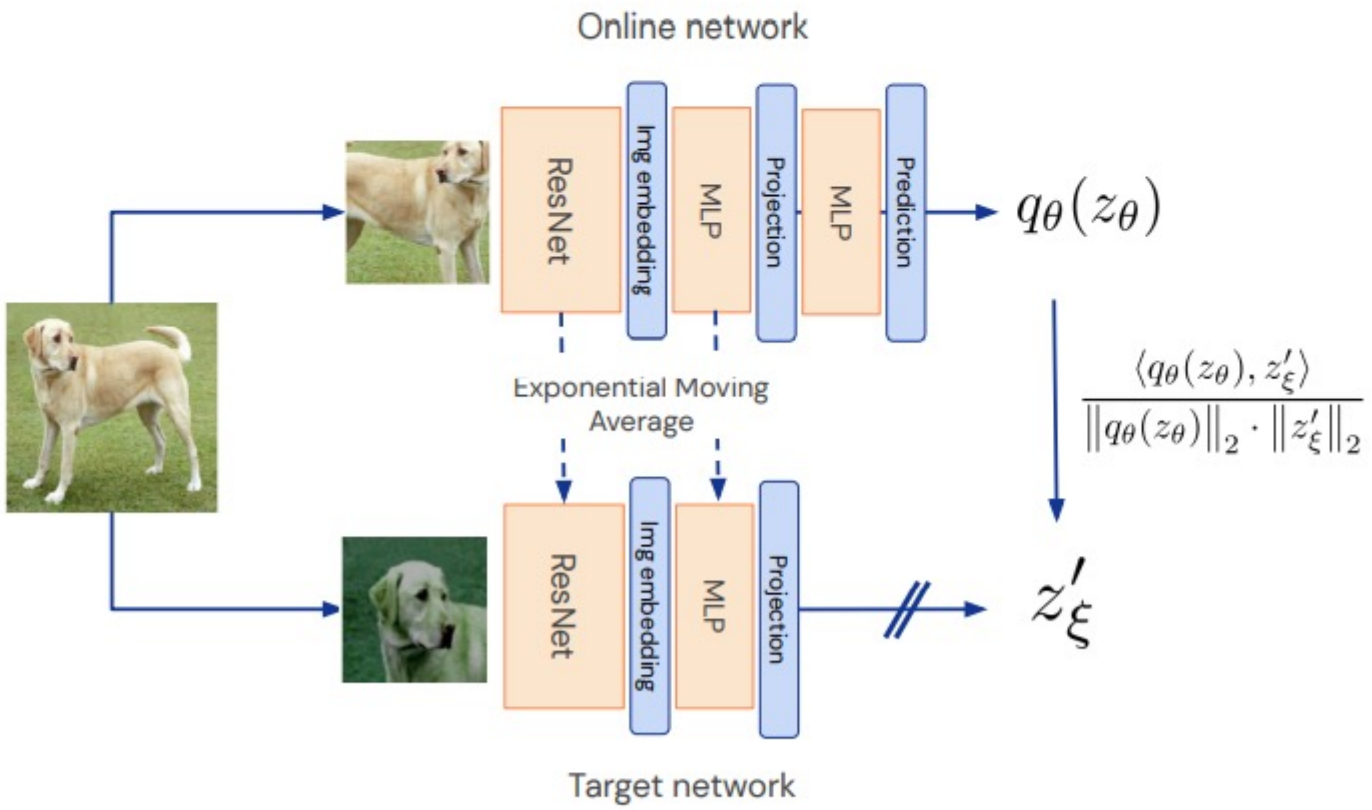


Figure 8: BYOL sketch summarizing the method by emphasizing the neural architecture.

Loss function

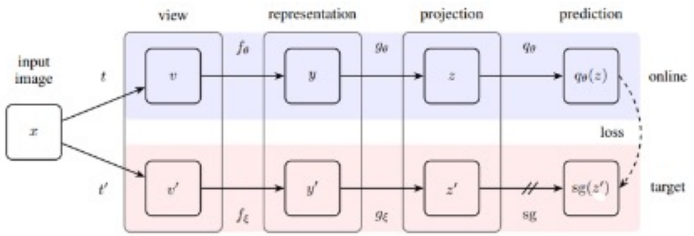
$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}$$

$$\overline{q_\theta}(z_\theta) \triangleq q_\theta(z_\theta)/\|q_\theta(z_\theta)\|_2$$

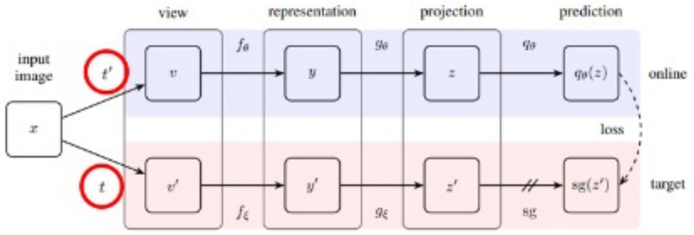
$$\overline{z}'_\xi \triangleq z'_\xi/\|z'_\xi\|_2$$

$$\mathcal{L}^{\texttt{BYOL}}_{\theta,\xi} = \mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi}$$

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\texttt{BYOL}}_{\theta,\xi}, \eta\right),$$
$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$



$\Rightarrow$ $\mathcal{L}_{\theta,\xi}$



$\Rightarrow$ $\widetilde{\mathcal{L}}_{\theta,\xi}$

Total Loss

$$\mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi}$$

Algorithm

Algorithm 1: BYOL: **B**ootstrap **Y**our **O**wn **L**atent

Inputs :

$\mathcal{D}$, $\mathcal{T}$, and $\mathcal{T}'$	set of images and distributions of transformations
θ, f_θ, g_θ, and q_θ	initial online parameters, encoder, projector, and predictor
ξ, f_ξ, g_ξ	initial target parameters, target encoder, and target projector
optimizer	optimizer, updates online parameters using the loss gradient
K and N	total number of optimization steps and batch size
$\{\tau_k\}_{k=1}^{K}$ and $\{\eta_k\}_{k=1}^{K}$	target network update schedule and learning rate schedule

for $k = 1$ **to** K **do**

  $\mathcal{B} \leftarrow \{x_i \sim \mathcal{D}\}_{i=1}^{N}$ `// sample a batch of` N `images`

  for $x_i \in \mathcal{B}$ **do**

    $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}'$ `// sample image transformations`

    $z_1 \leftarrow g_\theta(f_\theta(t(x_i)))$ and $z_2 \leftarrow g_\theta(f_\theta(t'(x_i)))$ `// compute projections`

    $z_1' \leftarrow g_\xi(f_\xi(t'(x_i)))$ and $z_2' \leftarrow g_\xi(f_\xi(t(x_i)))$ `// compute target projections`

    $l_i \leftarrow -2 \cdot \left(\frac{\langle q_\theta(z_1), z_1' \rangle}{\|q_\theta(z_1)\|_2 \cdot \|z_1'\|_2} + \frac{\langle q_\theta(z_2), z_2' \rangle}{\|q_\theta(z_2)\|_2 \cdot \|z_2'\|_2} \right)$ `// compute the loss for` x_i

  end

  $\delta\theta \leftarrow \frac{1}{N} \sum_{i=1}^{N} \partial_\theta l_i$ `// compute the total loss gradient w.r.t.` θ

  $\theta \leftarrow \text{optimizer}(\theta, \delta\theta, \eta_k)$ `// update online parameters`

  $\xi \leftarrow \tau_k \xi + (1 - \tau_k)\theta$ `// update target parameters`

end

Output : encoder f_θ

Intuitions on BYOL's behavior

- Undesirable equilibria가 unstable 하다고 가정함

$$\mathrm{Var}(X|Y,Z) \leq \mathrm{Var}(X|Y)$$

X : target projection
Y : current online projection
Z : additional variability on top of the online projection

$$\mathrm{Var}(z'_\xi|z_\theta) \leq \mathrm{Var}(z'_\xi|c)$$

How BYOL prevents representation collapse?

- Bootstrap Your Own Latent : A New Approach to Self-Supervised Learning
  - ✓ Addition of a predictor to the online network
  - ✓ Use of a moving average of online parameters

새로운 가설 제시

- Understanding self-supervised and contrastive learning with "Bootstrap Your Own Latent"
  - ✓ Batch Normalization 때문

반박

- BYOL works even without batch statistics
  - ✓ Batch Normalization 때문 아님

How BYOL prevents representation collapse?

Exploring Simple Siamese Representation Learning

Xinlei Chen   Kaiming He
Facebook AI Research (FAIR)

2020.11 arXiv

Understanding self-supervised Learning Dynamics without Contrastive Pairs

---

Yuandong Tian [1] Xinlei Chen [1] Surya Ganguli [1 2]

Facebook AI Research

Abstract

Contrastive approaches to self-supervised learning (SSL) learn representations by minimizing

man et al., 2019) whereby the hidden representations of two augmented views of the same object (positive pairs) are brought closer together, while those of different ob-

2021.02 arXiv

Experiment Results

- Linear evaluation on ImageNet

Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	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 [32]	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.

- Semi-supervised training on ImageNet

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised [77]	25.4	56.4	48.4	80.4
InstDisc	-	-	39.2	77.4
PIRL [35]	-	-	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 [32]	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.

Experiment Results

- Transfer to other classification tasks

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.

- Transfer to other vision tasks

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 [83]	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

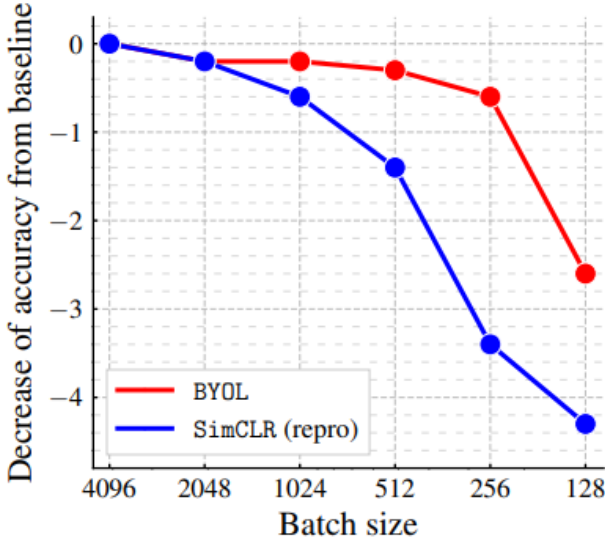
(b) Transfer results on NYU v2 depth estimation.

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

Ablation study

- Batch size

Batch size	Top-1 BYOL (ours)	Top-1 SimCLR (repro)	Top-5 BYOL (ours)	Top-5 SimCLR (repro)
4096	**72.5**	67.9	**90.8**	88.5
2048	72.4	67.8	90.7	88.5
1024	72.2	67.4	90.7	88.1
512	72.2	66.5	90.8	87.6
256	71.8	64.3±2.1	90.7	86.3±1.0
128	69.6±0.5	63.6	89.6	85.9
64	59.7±1.5	59.2±2.9	83.2±1.2	83.0±1.9

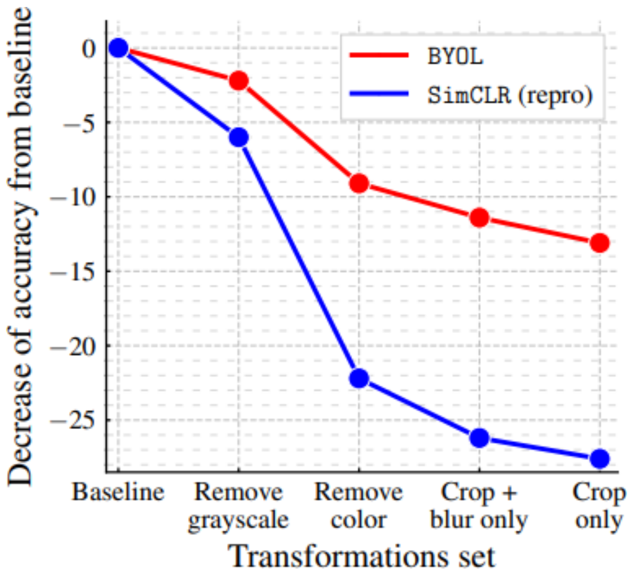


(a) Impact of batch size

Ablation study

- Image Augmentation

Image augmentation	Top-1 BYOL (ours)	Top-1 SimCLR (repro)	Top-5 BYOL (ours)	Top-5 SimCLR (repro)
Baseline	**72.5**	67.9	**90.8**	88.5
Remove flip	71.9	67.3	90.6	88.2
Remove blur	71.2	65.2	90.3	86.6
Remove color (jittering and grayscale)	63.4±0.7	45.7	85.3±0.5	70.6
Remove color jittering	71.8	63.7	90.7	85.9
Remove grayscale	70.3	61.9	89.8	84.1
Remove blur in $\mathcal{T}'$	72.4	67.5	90.8	88.4
Remove solarize in $\mathcal{T}'$	72.3	67.7	90.8	88.2
Remove blur and solarize in $\mathcal{T}'$	72.2	67.4	90.8	88.1
Symmetric blurring/solarization	72.5	68.1	90.8	88.4
Crop only	59.4±0.3	40.3±0.3	82.4	64.8±0.4
Crop and flip only	60.1±0.3	40.2	83.0±0.3	64.8
Crop and color only	70.7	64.2	90.0	86.2
Crop and blur only	61.1±0.3	41.7	83.9	66.4



참고

- https://arxiv.org/pdf/2006.07733.pdf
- https://2-chae.github.io/category/2.papers/26
- https://doubleby.github.io/self-supervised-learning/2021/01/27/BYOL/
- https://hoya012.github.io/blog/byol/
- https://blog.promedius.ai/ssl_byol/