Self-Supervised Learning

Marina Munkhoeva
Research Scientist, AIRI

HSE, 2025-10-16

00

Outline

Outline

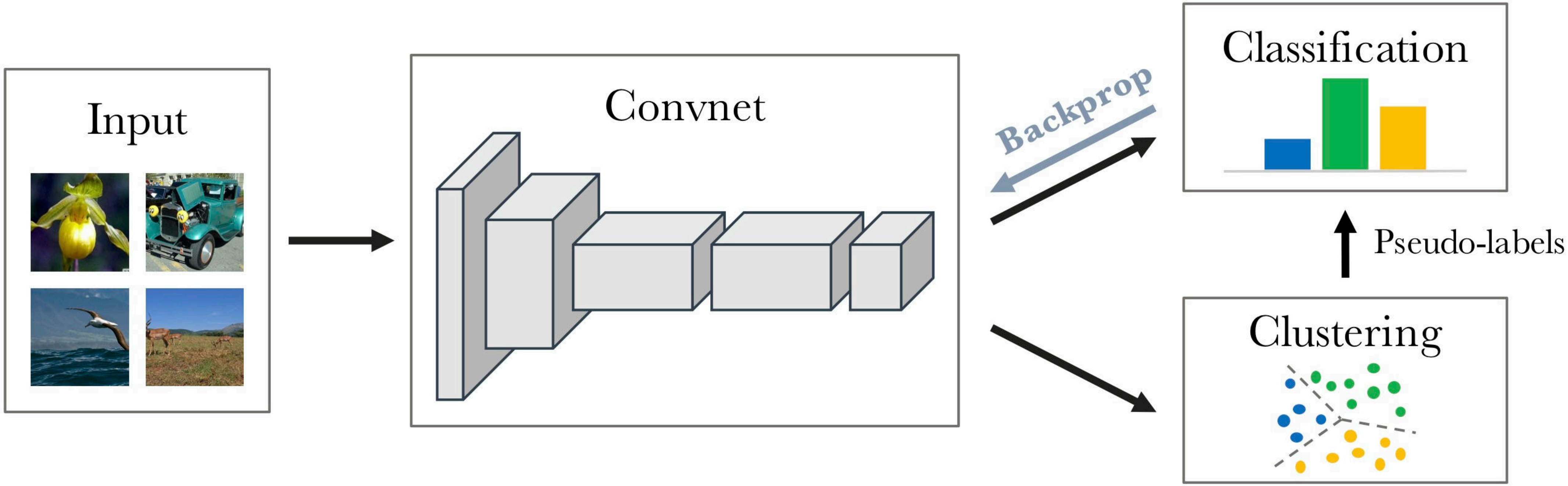
01

Clustering

Deep Cluster (Recap)

Just iteratively cluster features to get pseudo-labels for classification:

$$\min_{C \in \mathbb{R}_d \times k} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} \| f_{\theta(x_n)} - C y_n \|_2^2 \quad \text{such that} \quad y_n^\top 1_k = 1$$



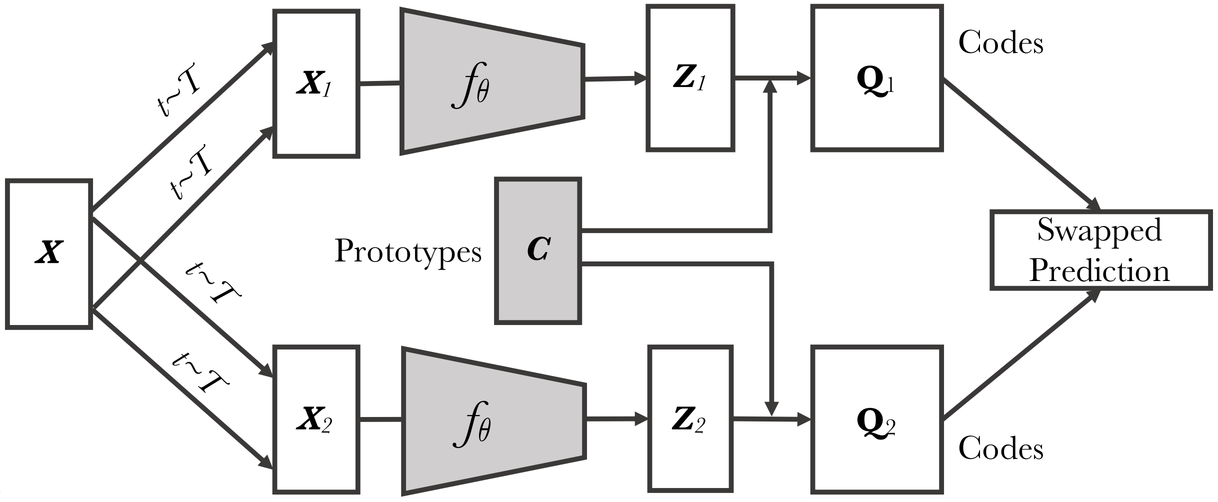
(Caron et al., 2018)

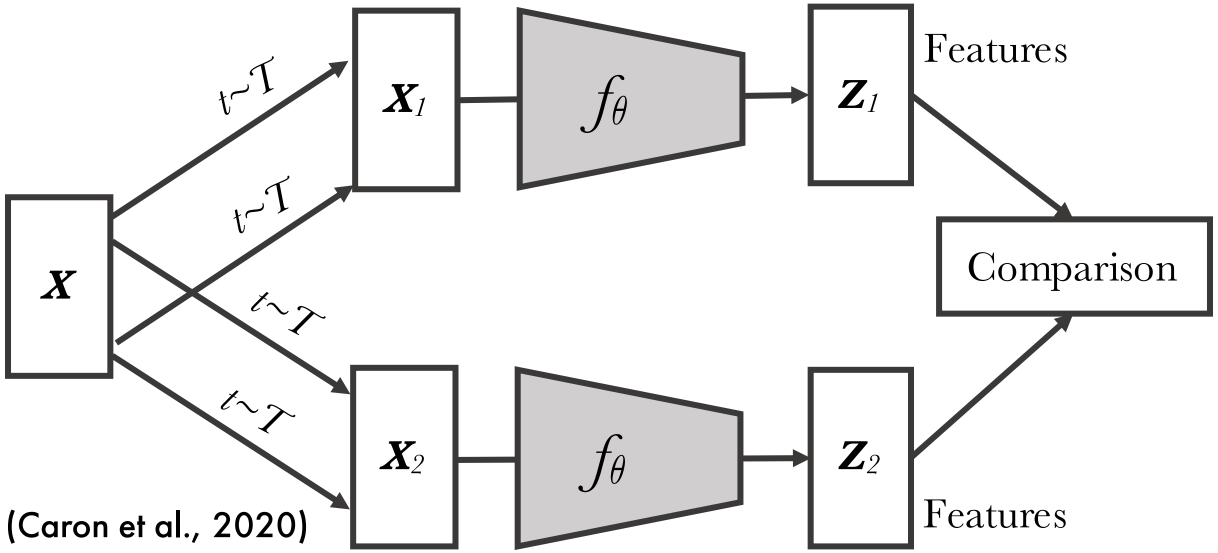
Clustering takes **third** of epoch time!

Swapping Assignments between Views

SwAV — Contrastive “DeepCluster”

- contrastive signal via swapping assignments
- learnable prototypes
- online clustering





(Caron et al., 2020)

Online Clustering

Map $Z = [z_1, ..., z_B]$ to $C = [c_1, ..., c_K]$ via codes matrix $Q = [q_1, ..., q_B]$

Similarity between clusters and representations $C^T Z$

Learn to equally partition codes in batch with $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$:

$$\max_{Q \in \mathcal{Q}} \mathrm{Tr}\ Q^\top C^T Z + \varepsilon H(Q)$$

Doubly stochastic matrices with positive entries

$\mathcal{Q} = \left\{ Q \in \mathbb{R}_+^{K \times B} \mid Q\mathbf{1}_B = \frac{1}{K}\mathbf{1}_K, Q^\top \mathbf{1}_K = \frac{1}{B}\mathbf{1}_B \right\}$

Enforce each prototypes picked to $\frac{B}{K}$ times on average

Sinkhorn-Knopp algorithm (iteratively normalize rows/columns):

$Q^* = \mathrm{Diag}(u) \exp\left(\frac{C^T Z}{\varepsilon}\right) \mathrm{Diag}(v)$, where u, v - renormalization vectors

SwAV

Once **soft** Cluster Assignment is done, we have codes Q

Now contrastive loss for image positive pair x_t, x_s:

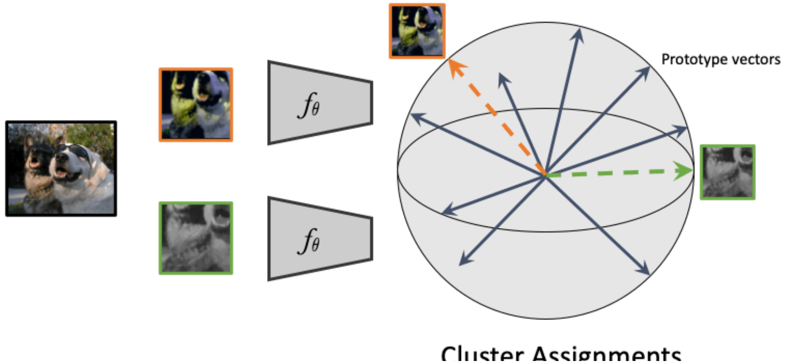
$$l(z_t, q_s) = -\sum_k q_s^{(k)} \log p_t^{(k)}$$

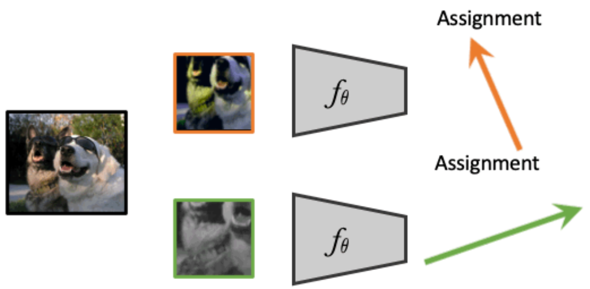
$$p_t^{(k)} = \frac{\exp\left(\frac{1}{\tau} z_t^\top c_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau} z_t^\top c_{k'}\right)}$$

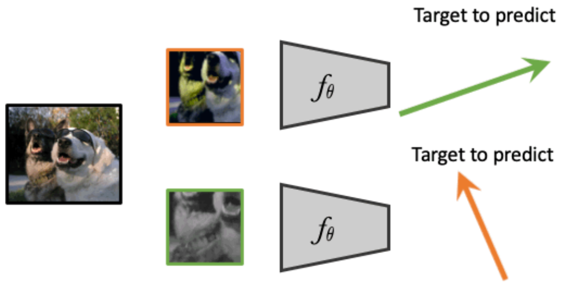
Symmetric loss $L(z_t, z_s) = l(z_t, q_s) + l(z_s, q_t)$

NB SwAV allows multi-crop

$$L(z_{t_1}, z_{t_2}, \ldots, z_{t_{V+2}} = \sum_{(i \in \{1,2\})} \sum_{v=1}^{V+2} \mathbf{1}_{v \neq i} l\left(z_{t_v}, q_{t_i}\right)$$







SwAV

Table 3: **Training in small batch setting.** Top-1 accuracy on ImageNet with a linear classifier trained on top of frozen features from a ResNet-50. All methods are trained with a batch size of 256. We also report the number of stored features, the type of cropping used and the number of epochs.

Method	Mom. Encoder	Stored Features	`multi-crop`	epoch	batch	Top-1
SimCLR		0	2×224	200	256	61.9
MoCov2	✓	$65,536$	2×224	200	256	67.5
MoCov2	✓	$65,536$	2×224	800	256	71.1
SwAV		$3,840$	$2\times160+4\times96$	200	256	72.0
SwAV		$3,840$	$2\times224+6\times96$	200	256	72.7
SwAV		$3,840$	$2\times224+6\times96$	400	256	**74.3**

SwAV

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [65]	R50	24	39.6
Jigsaw [46]	R50	24	45.7
NPID [58]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [68]	R50	24	58.8
NPID++ [44]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [44]	R50	24	63.6
CPC v2 [28]	R50	24	63.8
PCL [37]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	**75.3**

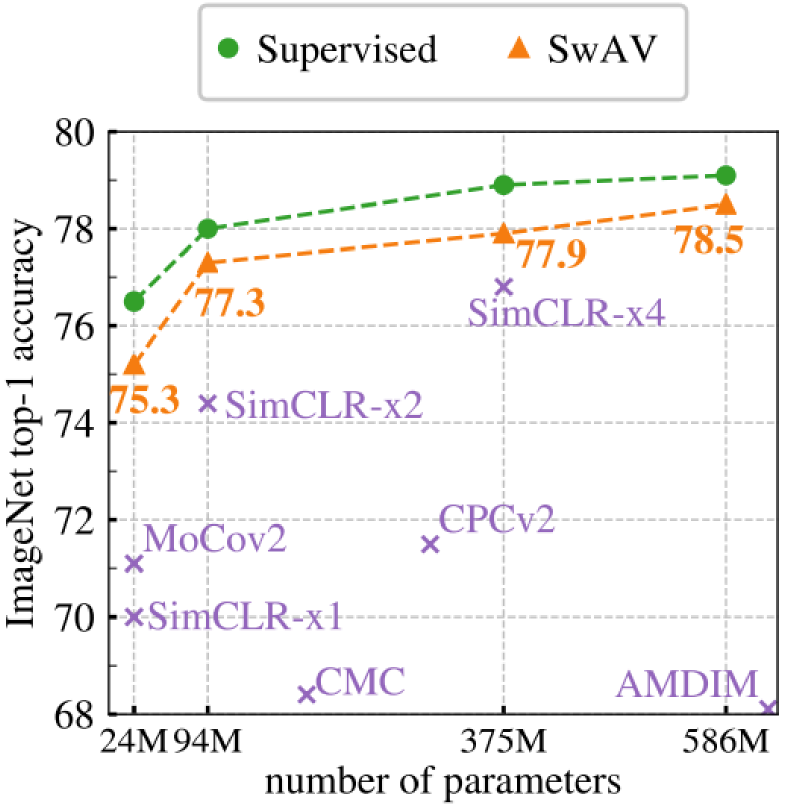
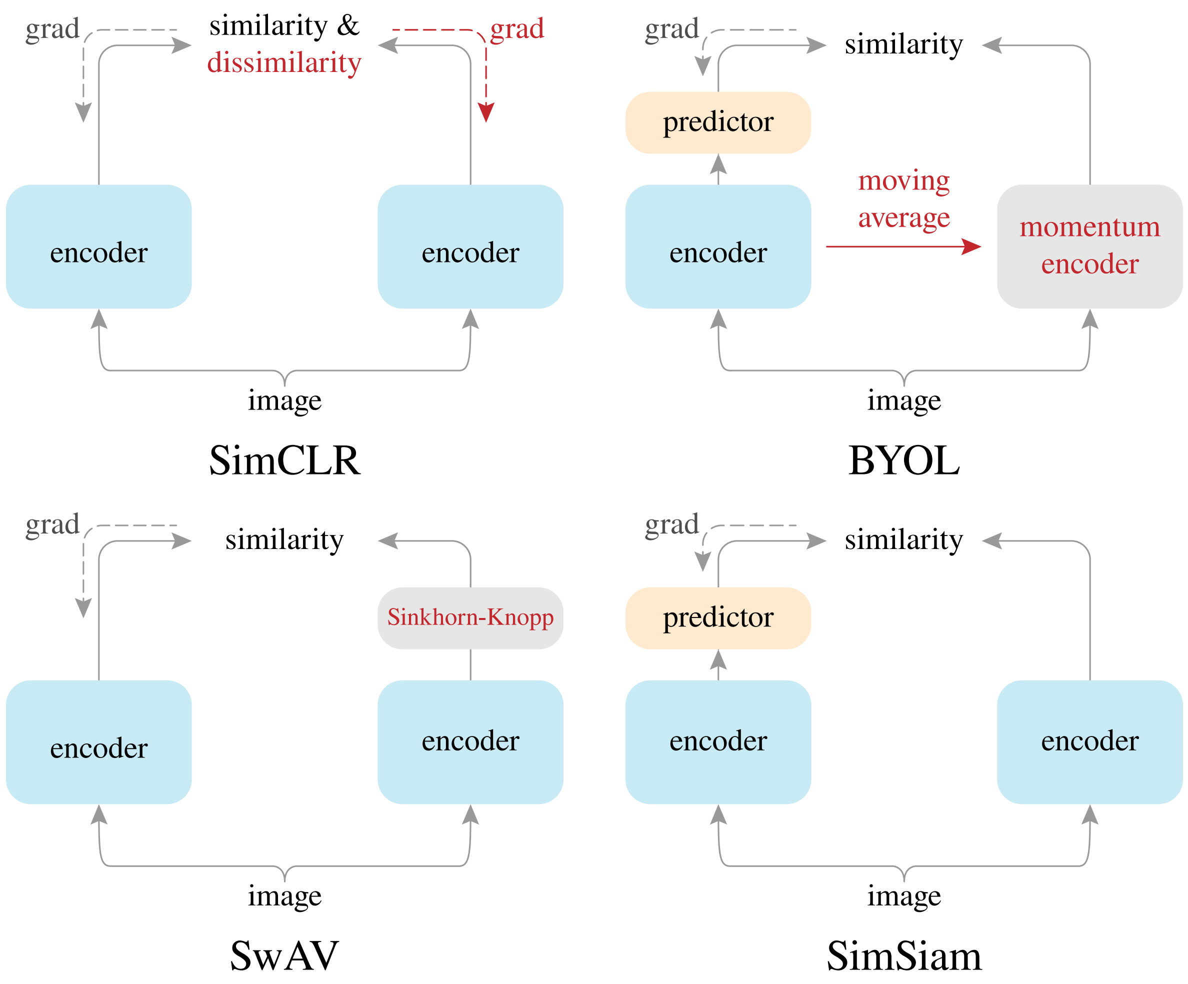


Figure 2: **Linear classification on ImageNet.** Top-1 accuracy for linear models trained on frozen features from different self-supervised methods. **(left)** Performance with a standard ResNet-50. **(right)** Performance as we multiply the width of a ResNet-50 by a factor ×2, ×4, and ×5.

SwAV

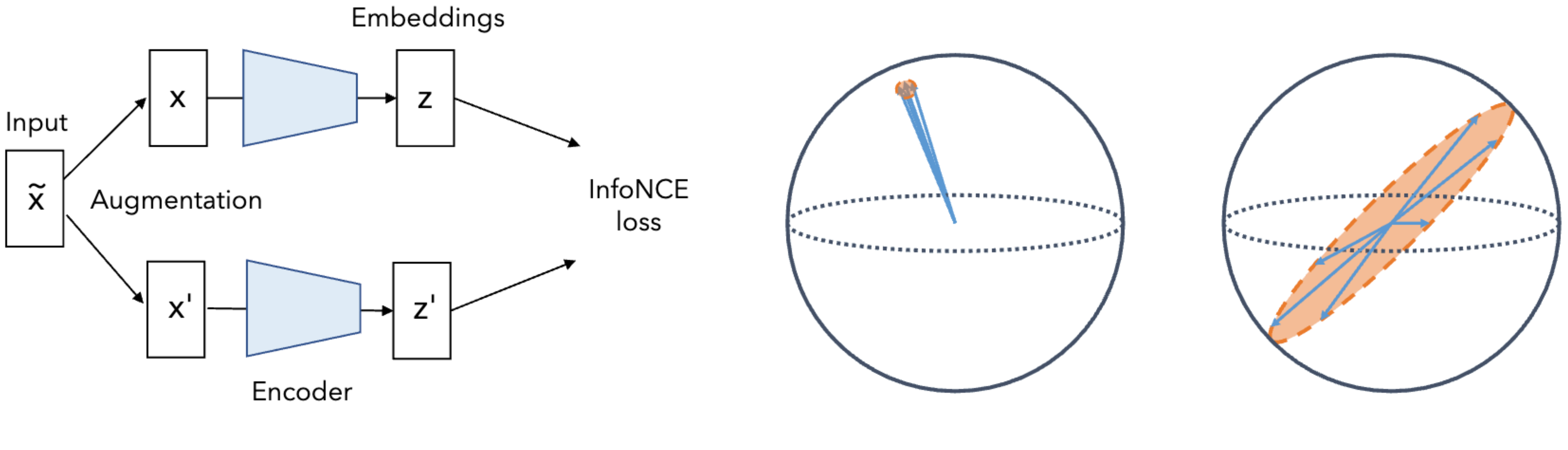
method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

Recap



grad
similarity &
dissimilarity
grad
encoder
encoder
image
SimCLR
grad
similarity
predictor
moving
average
encoder
momentum
encoder
image
BYOL
grad
similarity
Sinkhorn-Knopp
encoder
encoder
image
SwAV
grad
similarity
predictor
encoder
encoder
image
SimSiam

Collapse in Contrastive Learning



(a) embedding space (b) complete collapse (c) dimensional collapse

Figure 1: Illustration of the collapsing problem. For complete collapse, the embedding vectors collapse to same point. For dimensional collapse, the embedding vectors only span a lower dimensional space.

(Jing et al., 2021)

Dimensional Collapse

— strong correlation between dimensions

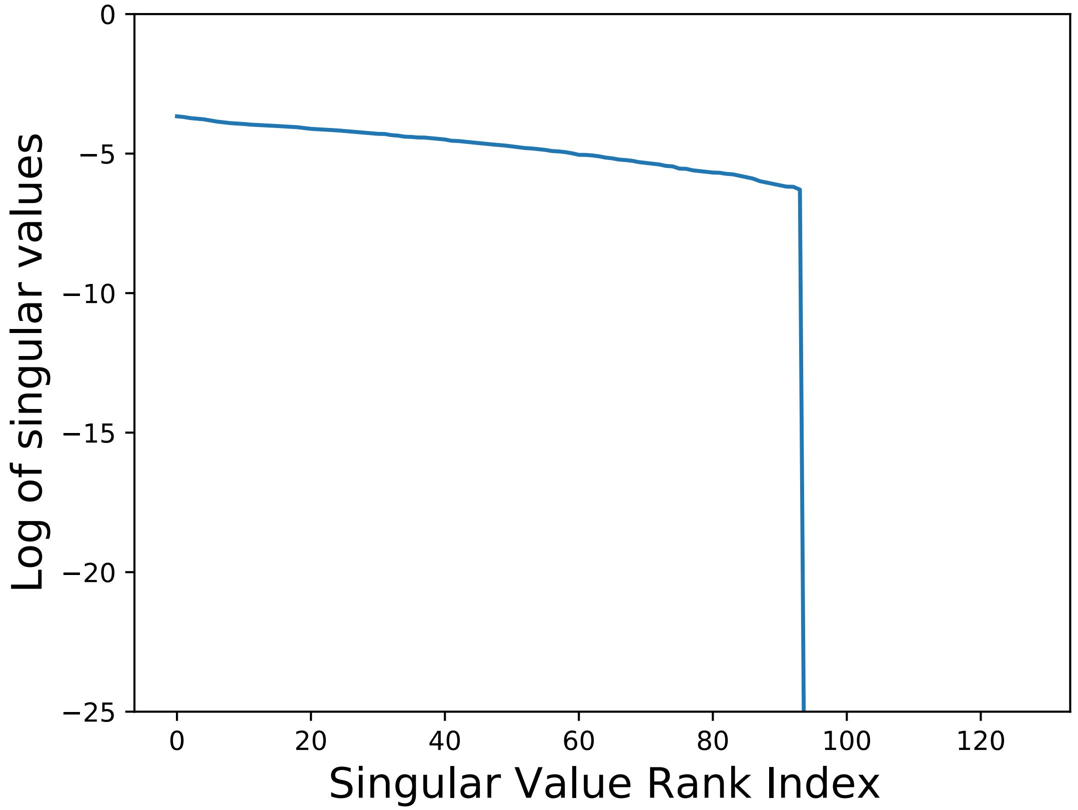
Covariance matrix of SimCLR embeddings

$$C = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T,$$

where $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$

$C = U\Sigma V^\top$

>20 singular values drop to zero

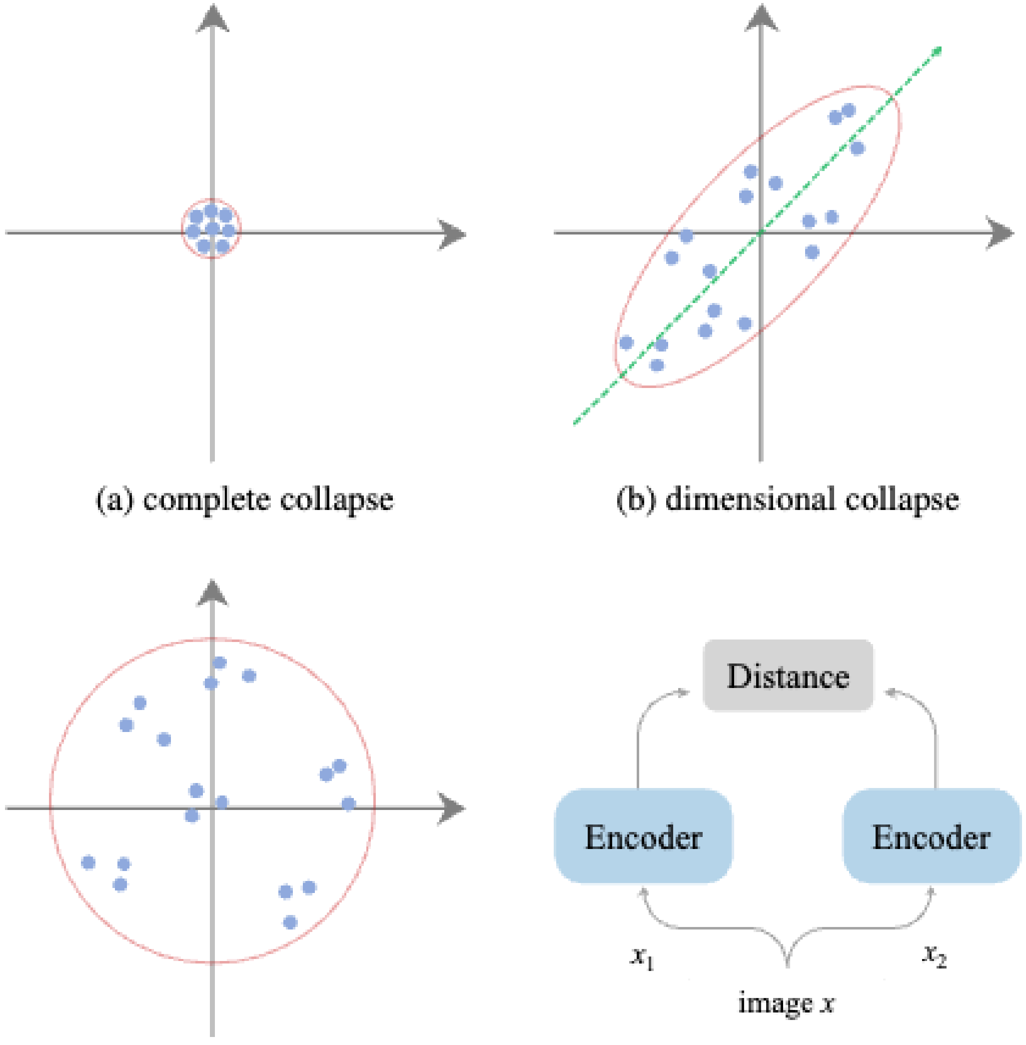


(Jing et al., 2021)

02

Decorrelation / Whitening

Feature Decorrelation



(a) complete collapse

(b) dimensional collapse

(c) decorrelated

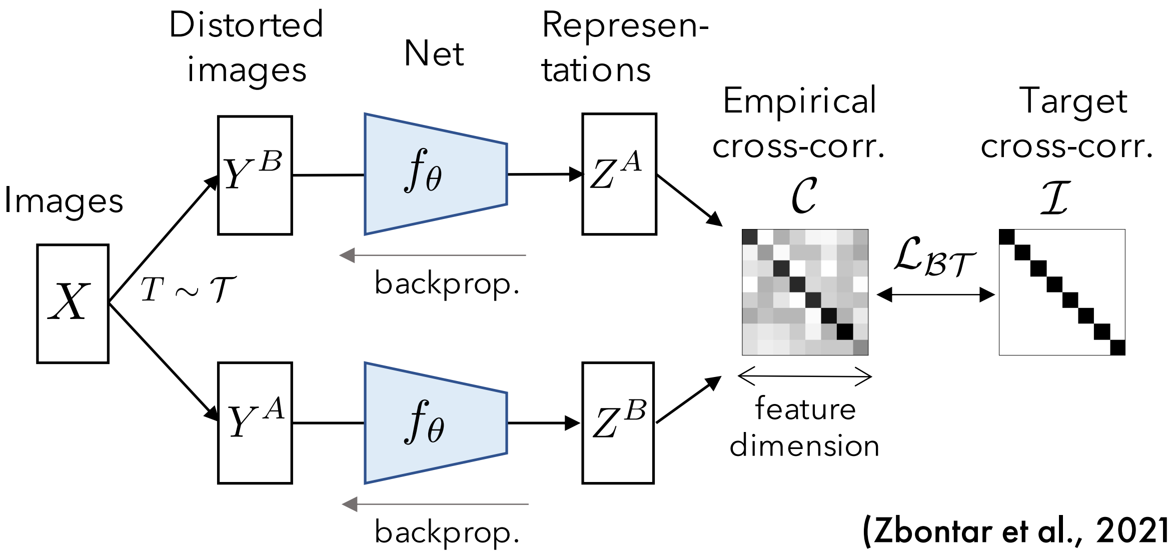
(d) the concise framework

Barlow Twins

Enforce statistically independent components

Cross-correlation matrix of twin embeddings

$$C_{ij} \triangleq \frac{\sum_b Z^A_{bi} Z^B_{bj}}{\sqrt{\sum_b \left(Z^A_{bi}\right)^2} \sqrt{\sum_b \left(Z^B_{bj}\right)^2}},$$



(Zbontar et al., 2021)

Z^A, Z^B — mean centered embedding matrices

for two data views A and B, i, j - index features, b - index of sample in a batch

invariance and redundancy-reduction terms: $\mathcal{L}_{\text{BT}} = \sum_i \left(1 - C_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2$

Decorrelating every pair of features maximizes information content preventing collapse

Information Bottleneck Principle

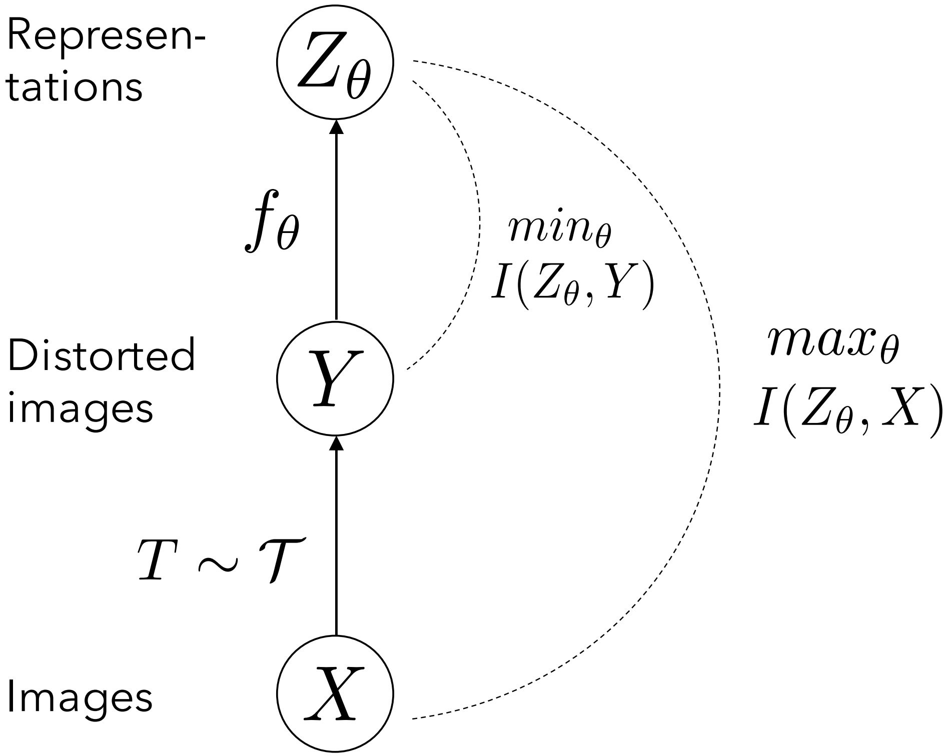
Information Bottleneck Principle applied to SSL

$$\mathrm{IB}_\theta \triangleq I(Z_\theta, Y) - \beta I(Z_\theta, X)$$

- representations informative about input
- representations invariant to distortions

$$\mathrm{IB}_\theta = [H(Z_\theta) - \cancel{H(Z_\theta|Y)}] - \beta[H(Z_\theta) - H(Z_\theta|X)]$$

$$= H(Z_\theta|X) + \frac{1-\beta}{\beta} H(Z_\theta)$$

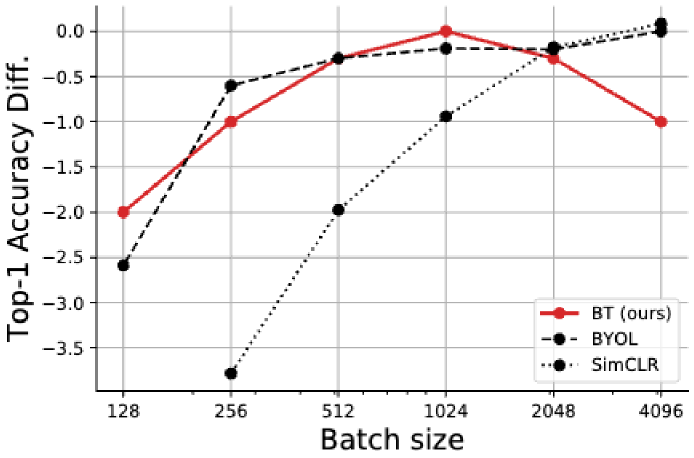


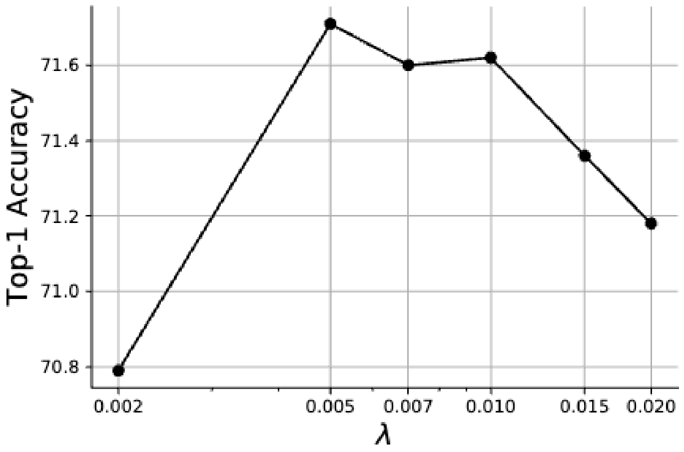
(Zbontar et al., 2021)

Assume Z is Gaussian $\Rightarrow \mathrm{IB}_\theta = \mathbb{E}_X \log|C_{Z_\theta|X}| + \frac{1-\beta}{\beta} \log|C_{Z_\theta}|$

$\beta > 1$, replace covariance with cross-correlation $\Rightarrow \mathcal{L}_{\mathrm{BT}}$

BT Ablations





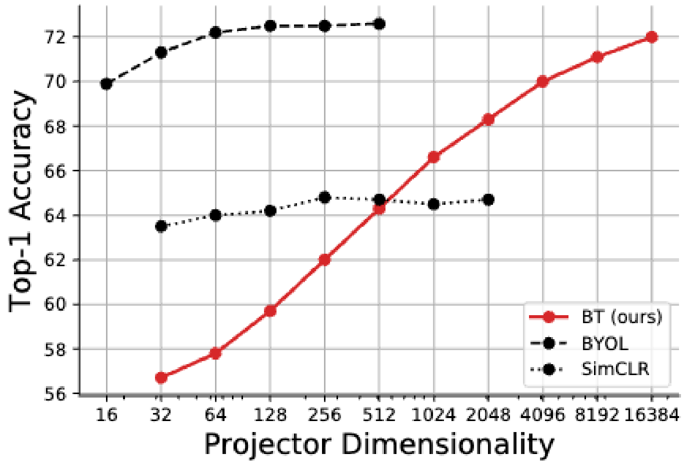


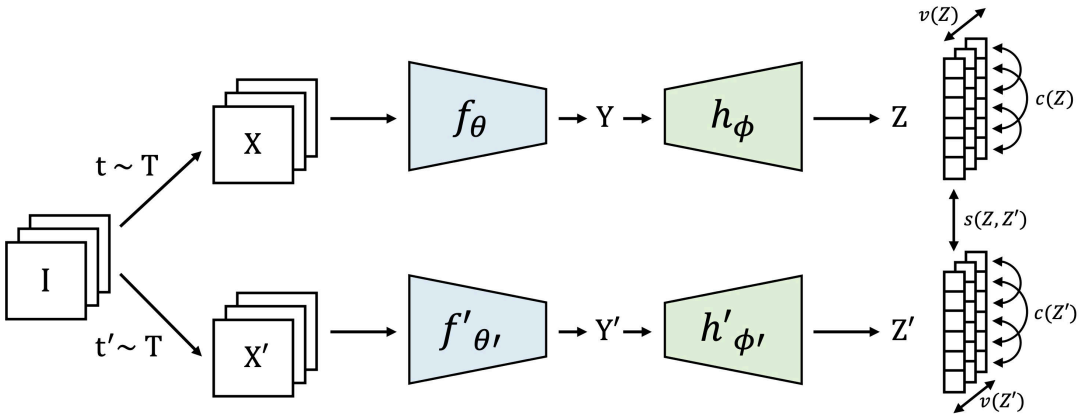
Table 3. **Transfer learning: image classification.** We benchmark learned representations on the image classification task by training linear classifiers on fixed features. We report top-1 accuracy on Places-205 and iNat18 datasets, and classification mAP on VOC07. Top-3 best self-supervised methods are underlined.

Method	Places-205	VOC07	iNat18
Supervised	53.2	87.5	46.7
SimCLR	52.5	85.5	37.2
MoCo-v2	51.8	<u>86.4</u>	38.6
SwAV (w/o multi-crop)	52.8	<u>86.4</u>	39.5
SwAV	<u>56.7</u>	<u>88.9</u>	<u>48.6</u>
BYOL	<u>54.0</u>	<u>86.6</u>	<u>47.6</u>
BARLOW TWINS (ours)	<u>54.1</u>	86.2	<u>46.5</u>

Variance-Invariance-Covariance

VICReg (Bardes et al., 2021):

- Variance: $v(Z) = \frac{1}{d}\sum_j^d \max(0, \gamma - S(z^j, \varepsilon))$
  $S(x, \varepsilon) = \sqrt{\mathrm{Var}(x) + \varepsilon}$
- Covariance: $c(Z) = \frac{1}{d}\sum_{i \neq j} [C(Z)]_{i,j}^2$
- Invariance: $s(Z) = \frac{1}{n}\sum_i \|z_i - z_i'\|_2^2$



invariance to different views + collapse prevention + information content maximization:

$$l(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]$$

Not much difference with other methods, what’s up?

Regularization

Table 4: **Effect of incorporating variance and covariance regularization in different methods.** Top-1 ImageNet accuracy with the linear evaluation protocol after 100 pretraining epochs. For all methods, pretraining follows the architecture, the optimization and the data augmentation protocol of the original method using our reimplementation. ME: Momentum Encoder. SG: stop-gradient. PR: predictor. BN: Batch normalization layers after input and inner linear layers in the expander. No Reg: No additional regularization. Var Reg: Variance regularization. Var/Cov Reg: Variance and Covariance regularization. Unmodified original setups are marked by a †.

Method	ME	SG	PR	BN	No Reg	Var Reg	Var/Cov Reg
BYOL	✓	✓	✓	✓	69.3†	70.2	69.5
SimSiam		✓	✓	✓	67.9†	68.1	67.6
SimSiam		✓	✓		35.1	67.3	67.1
SimSiam		✓			collapse	56.8	66.1
VICReg			✓		collapse	56.2	67.3
VICReg			✓	✓	collapse	57.1	68.7
VICReg				✓	collapse	57.5	68.6†
VICReg					collapse	56.5	67.4

Weights and Architecture

Table 5: **Impact of sharing weights or not between branches.** Top-1 accuracy on linear classification with 100 pretraining epochs. The encoder and expander of both branches can share the same architecture and share their weights (SW), share the same architecture with different weights (DW), or have different architectures (DA). The encoders can be ResNet-50, ResNet-101 or ViT-S.

	SW R50	DW R50	DA R50/R101	DA R50/ViT-S
BYOL	69.3	✗	✗	✗
SimCLR	64.4	63.1	63.9	63.5
Barlow Twins	68.7	64.2	65.3	63.9
VICReg	68.6	66.5	68.1	66.2

Preliminaries

Whitening converts $x = (x_1, ..., x_d)^\top$, $\mathbb{E}(x) = \mu = (\mu_1, ..., \mu_d)^\top$, $\mathrm{var}(x) = \Sigma$ into

$$z = (z_1, ..., z_d)^\top = Wx$$

that has unit diagonal "white" covariance $\mathrm{var}(z) = I$, and W — whitening matrix

How to choose W? $\quad W \Sigma W^\top = \mathrm{var}(z) = I \to W^\top W = \Sigma^{-1}$ W is not uniquely defined!

How to select optimal W? (Kessy et al., 2018)

Consider

- Soft-whitening (Barlow Twins, VICreg)
- Cholesky: $\Sigma = LL^T \to W_{\mathrm{Chol}} = L^{-1}$
- ZCA: $W^{\mathrm{ZCA}} = \Sigma^{-\frac{1}{2}}$

Table 1: Five natural whitening transformations and their properties.

	Sphering matrix $\boldsymbol{W}$	Cross-covariance $\boldsymbol{\Phi}$	Cross-correlation $\boldsymbol{\Psi}$	Rotation matrix $\boldsymbol{Q}_1$	Rotation matrix $\boldsymbol{Q}_2$
ZCA	$\boldsymbol{\Sigma}^{-1/2}$	$\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^{-1/2}$	$\boldsymbol{I}$	$\boldsymbol{A}^T$
PCA	$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$	$\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T$	$\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{V}^{-1/2}$	$\boldsymbol{U}^T$	$\boldsymbol{U}^T\boldsymbol{A}^T$
Cholesky	$\boldsymbol{L}^T$	$\boldsymbol{L}^T\boldsymbol{\Sigma}$	$\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}$	$\boldsymbol{L}^T\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{L}^T\boldsymbol{V}^{1/2}\boldsymbol{P}^{1/2}$
ZCA-cor	$\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}$	$\boldsymbol{P}^{1/2}\boldsymbol{V}^{1/2}$	$\boldsymbol{P}^{1/2}$	$\boldsymbol{A}$	$\boldsymbol{I}$
PCA-cor	$\boldsymbol{\Theta}^{-1/2}\boldsymbol{G}^T\boldsymbol{V}^{-1/2}$	$\boldsymbol{\Theta}^{1/2}\boldsymbol{G}^T\boldsymbol{V}^{1/2}$	$\boldsymbol{\Theta}^{1/2}\boldsymbol{G}^T$	$\boldsymbol{G}^T\boldsymbol{A}$	$\boldsymbol{G}^T$

Whitening for Self-Supervised Learning

W-MSE (Ermolov et al., 2021): Cholesky decomposition

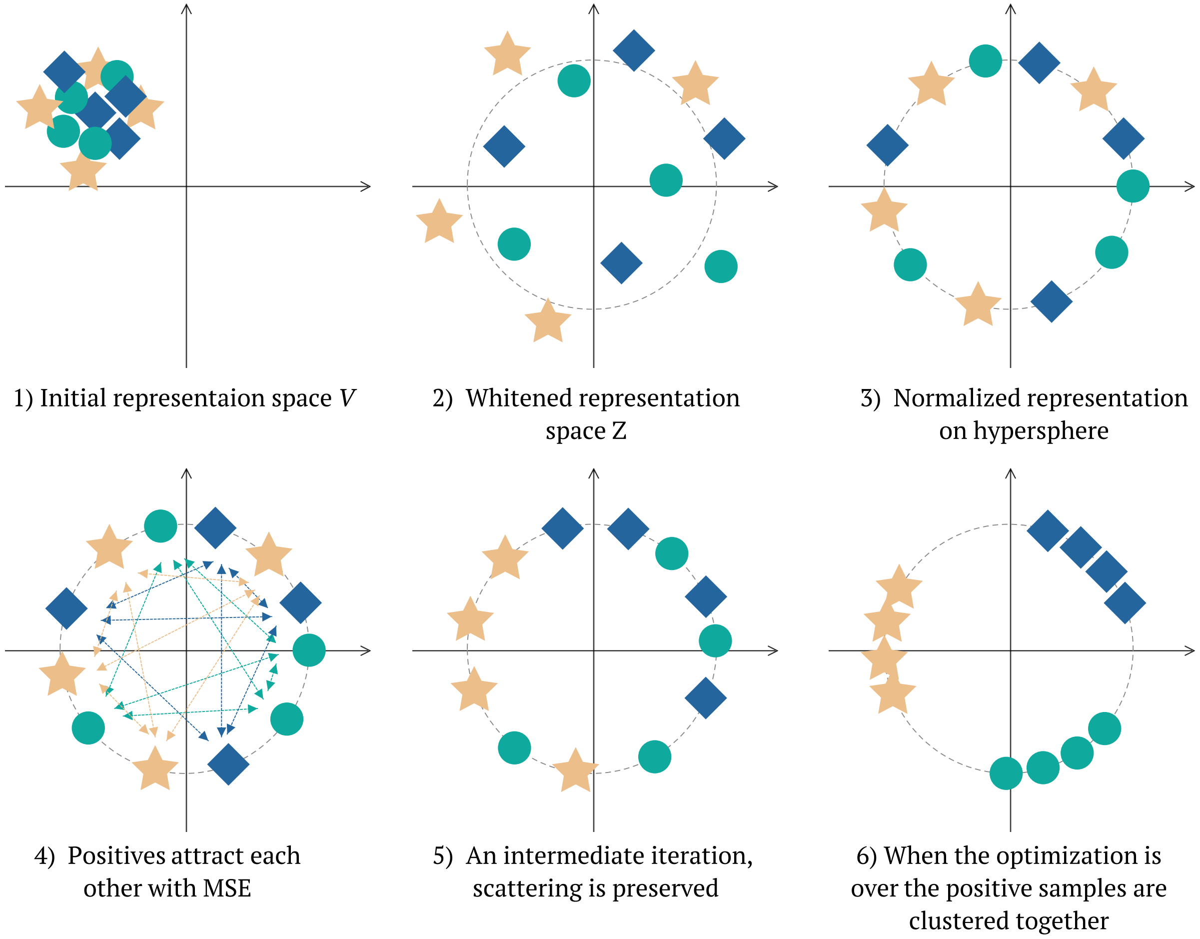
(x_i, x_j) — positive pairs

z_i, z_j — embeddings of positive pair

$$\min_\theta \mathbb{E}\left[\text{dist}\left(z_i, z_j\right)\right]$$

$$s.t. \ \text{cov}(z_i, z_i) = \text{cov}\left(z_j, z_j\right) = I$$

dist — cosine similarity



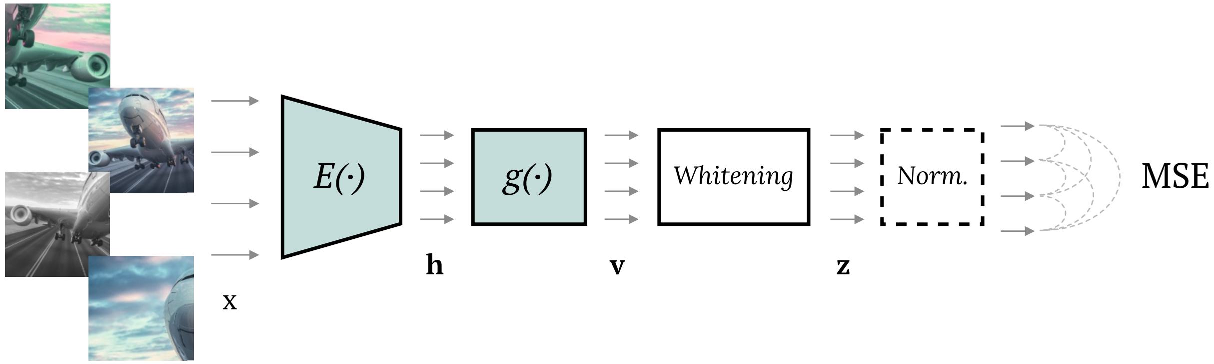
W-MSE

N unique images, d — augmentations,

$K = Nd$ — total batch size

$V = \{v_1, ..., v_K\}$ — embeddings of batch

$\Sigma_V = \frac{1}{K-1}\sum_k (v_k - \mu_V)(v_k - \mu_V)^\top$



W-MSE loss uses reparameterization of v to whitened z:

$$L_{\text{W-MSE}}(V) = \frac{2}{Nd(d-1)} \sum_{\text{pos}} \text{dist}\big(z_i, z_j\big),$$

$z = \text{Whitening}(v) = W_v(v - \mu_v)$ with $W_V^\top W_V = \Sigma_V^{-1}$

Compute Cholesky decomposition $\Sigma_V = LL^T$, take $W_V = L^{-1}$ on sub-batches of V

Complexity $O\big(k^3 + Mk^2\big)$ with k embedding dim, M slice size — comparable to forward pass

Decorrelated Batch Normalization

Consise framework to study collapse:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim D, t_1, t_2 \sim T} \, \|f_\theta(x_1) - f_\theta(x_2)\|_2^2$$

CIFAR-10 embedded with MSE loss (Hua et al., 2021):

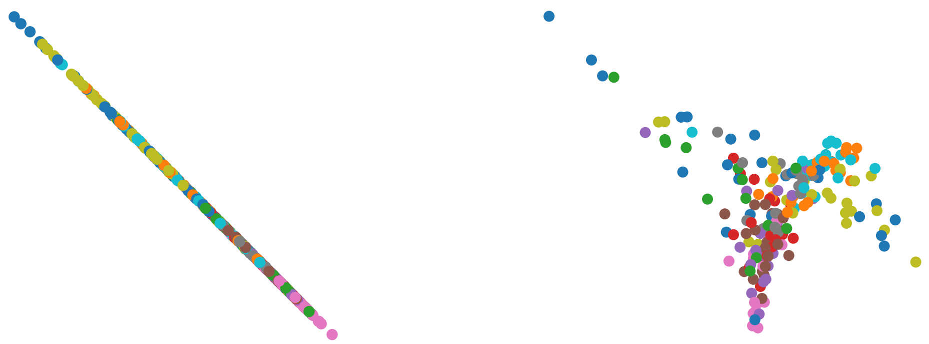
complete collapse | +batch normalization: dimensional collapse | +DBN: decorrelated features | SimCLR | Supervised

Batch Normalization

$X = (x_1, \ldots, x_B) \in \mathbb{R}^{D \times B}$ — input
$Y = (x_1, \ldots, x_B) \in \mathbb{R}^{D \times B}$ — output

Batch Normalization: $y_{b,d} = \frac{x_{b,d} - \mu_d}{\sqrt{\sigma_d^2 + \varepsilon}} \gamma_d + \beta_d$

- removes complete collapse



Decorrelated Batch Normalization (DBN):

$Y^{[h]} = \mathrm{ZCA}\left(X^{[h]}\right)$ with $\mathrm{ZCA} : Y = Q \Lambda^{-\frac{1}{2}} Q^\top \hat{X}$

- decorrelates covariance of feature groups

	acc. (%)	std.	corr.	loss
vanilla	35.44	**0.00**	0.13	0.00
BN	70.85	1.00	**0.99**	7.01
DBN	84.41	1.00	0.00	39.04

(Hua et al., 2021)

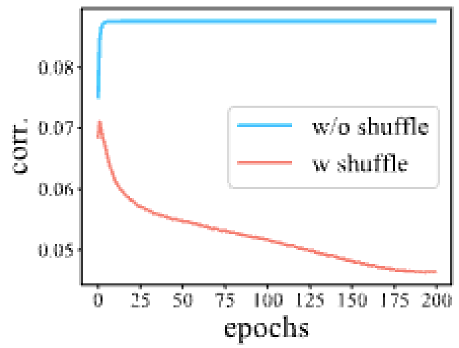
Shuffled Batch Normalization

Further decorrelation — permute the features randomly before grouping for DBN

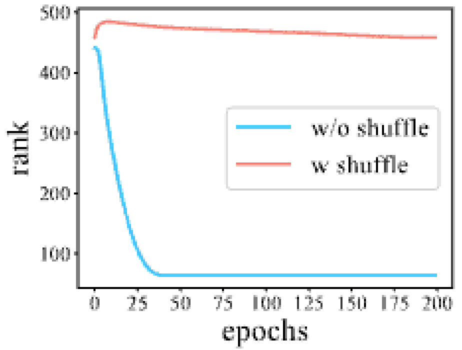
$\mathcal{P}$ — random D-order permutation

$$Y = \mathcal{P}^{-1}(\mathrm{DBN}_G(\mathcal{P}(X)))$$

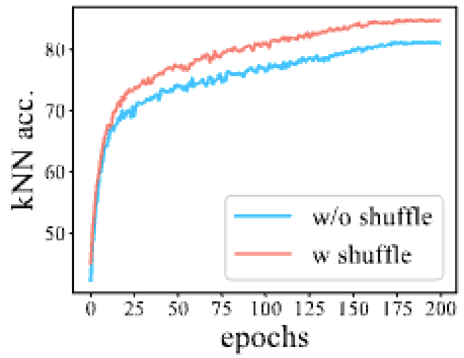
- cosine similarity interferes with grouping
- grouping required to satisfy Σ is full-rank



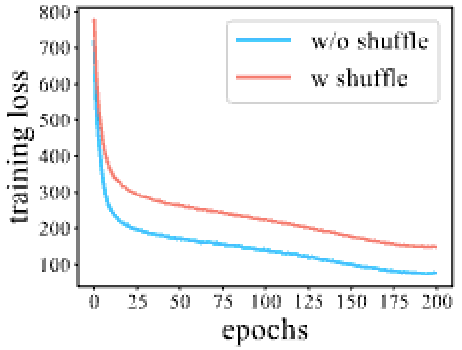
(a) **corr.** denotes the average correlation strength (*i.e.* the average of the absolute values of non-diagonal entries of the correlation matrix) of the projected features.



(b) **rank** denotes the (estimated) rank of spaces spanned by projected features of 512 samples, which is computed by checking singular values.



(c) **acc.** denotes accuracy in kNN classification.



(d) **loss** denotes the training loss.

Shuffled-DBN Ablation

robust to batch size change

batch size	32	64	128	256	512
Shuffled-DBN	88.25	**89.17**	**89.31**	**88.82**	87.92
Barlow Twins	86.89	87.98	88.21	87.57	85.19
BYOL	**88.37**	88.44	87.64	85.72	82.63
SimCLR	85.42	87.41	87.40	87.70	**87.98**
SimSiam	86.84	87.88	86.47	79.02	67.74

group size increase positively affects decorrelation ability

group size	16	32	64	128
kNN acc.	83.41	85.93	87.05	87.59
linear acc.	85.52	87.69	88.75	88.29

Shuffled-DBN Performance

	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet
SimCLR [8]	86.96	55.86	85.50	42.65
BYOL [37]	86.65	59.33	85.59	42.75
SimSiam [10]	86.31	59.44	**86.55**	41.58
Barlow Twins [51]	89.02	62.84	85.43	45.33
DBN	86.32	56.49	82.36	40.37
Shuffled-DBN	**89.50**	**62.95**	86.02	**45.96**

within concise framework

method	batch size	top-1
InstDisc [46]	256	58.5
LocalAgg [52]	128	58.8
MoCo [19]	256	60.6
SimCLR [8]	256	61.9
CPC v2 [35]	512	63.8
PCL v2 [33]	256	67.6
MoCo v2 [9]	256	67.5
MoCHi [29]	512	68.0
PIC [4]	512	67.6
AdCo [24]	256	68.6
Shuffled-DBN	512	65.18

outside concise framework

03

Bibliography

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. Arxiv Preprint Arxiv:2105.04906.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. Proceedings of the European Conference on Computer Vision (ECCV), 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems, 33, 9912–9924.

Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021). Whitening for self-supervised representation learning. International Conference on Machine Learning, 3015–3024.

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., & Zhao, H. (2021). On feature decorrelation in self-supervised learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 9598–9608.

Jing, L., Vincent, P., LeCun, Y., & Tian, Y. (2021). Understanding dimensional collapse in contrastive self-supervised learning. Arxiv Preprint Arxiv:2110.09348.

Kessy, A., Lewin, A., & Strimmer, K. (2018). Optimal whitening and decorrelation. The American Statistician, 72(4), 309–314.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. International Conference on Machine Learning, 12310–12320.

Thank you!