An Investigation into Whitening Loss for Self-supervised Learning

Xi Weng[1], Lei Huang[1, 2], Lei Zhao[1],
Rao Muhammad Anwer[2], Salman Khan[2], Fahad Shahbaz Khan[2]

[1]SKLSDE, Institute of Artificial Intelligence, Beihang University
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE

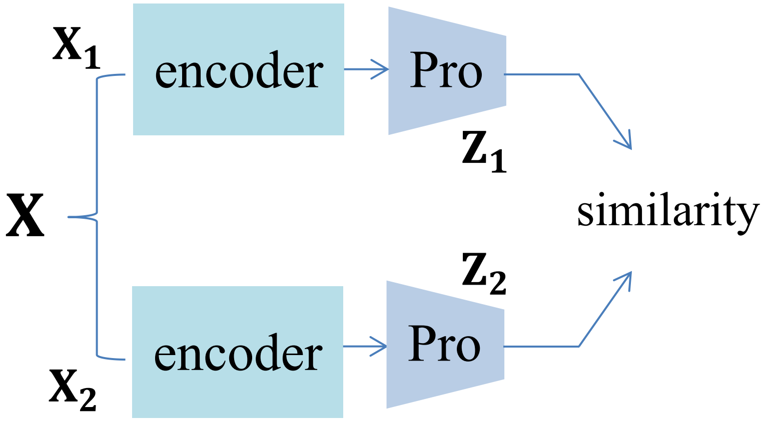
北京航空航天大学人工智能研究院
Institute of Artificial Intelligence, Beihang University

软件开发环境国家重点实验室
State Key Laboratory of Software Development Environment

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

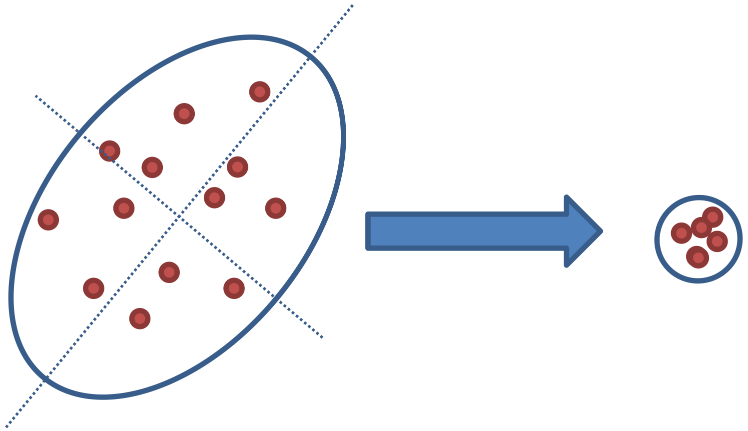
Siamese Network and Collapse

- Simaese Network

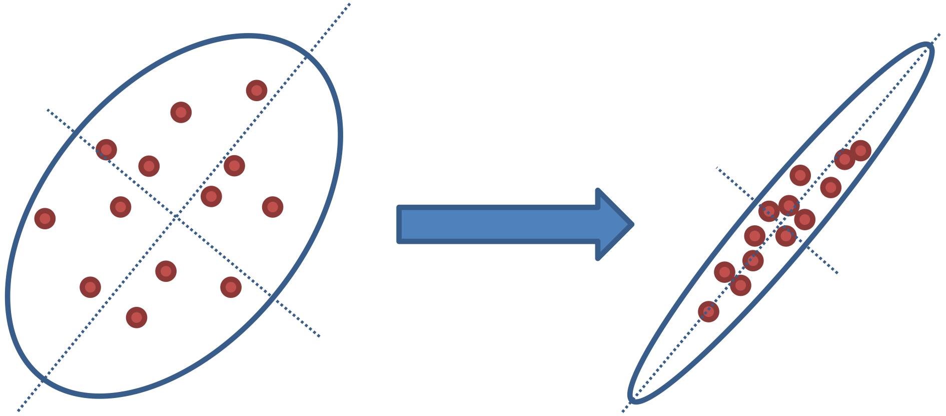


$$\mathcal{L}(\mathbf{x},\theta) = \mathbb{E}_{\mathbf{x}\sim\mathbb{D},\ \mathcal{T}_{1,2}\sim\mathbb{T}}\ \ell\big(f_\theta(\mathcal{T}_1(\mathbf{x})), f_\theta(\mathcal{T}_2(\mathbf{x}))\big)$$

- Complete Collapse

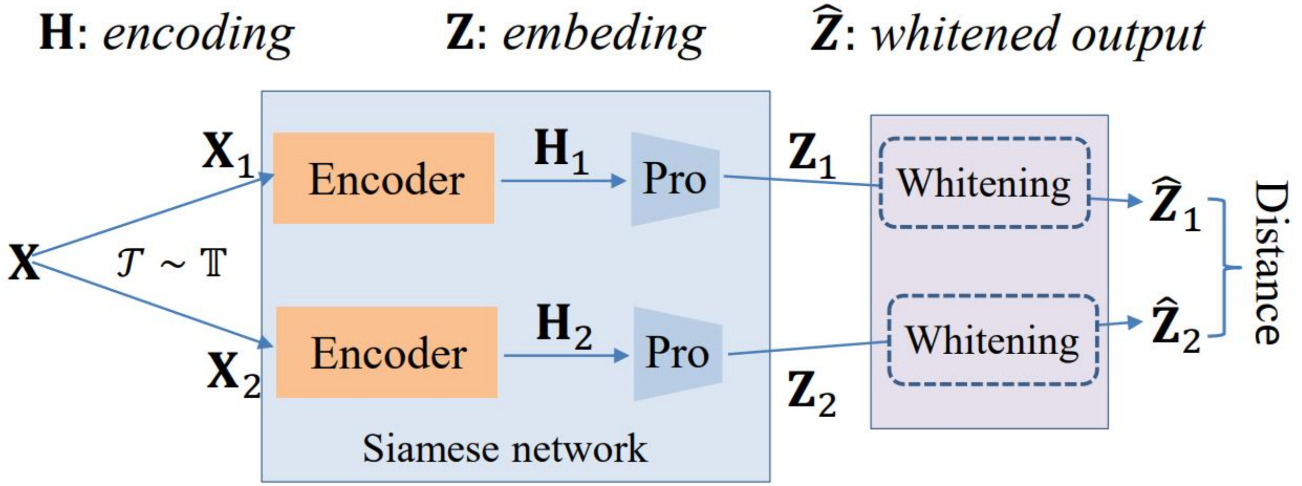


- Dimensional Collapse



Whitening loss

➢ Structure of whitening loss:



➢ Loss function:

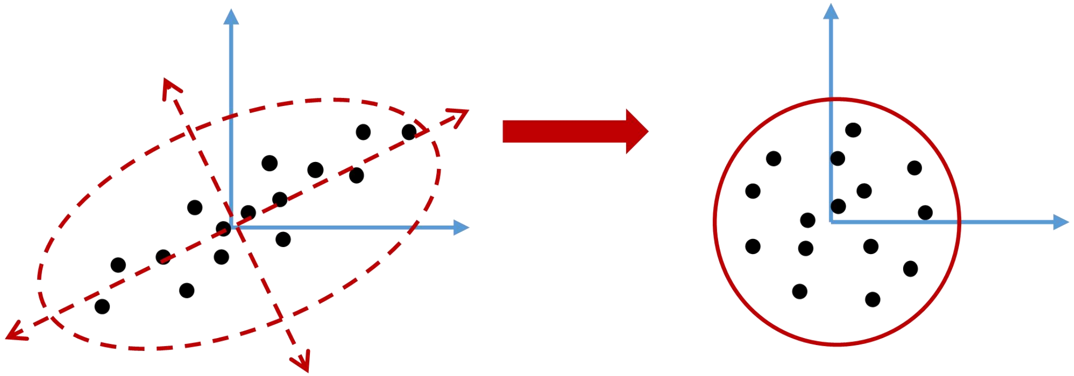
$$\min_{\theta} \mathcal{L}(\mathbf{x};\theta) = \mathbb{E}_{\mathbf{x}\sim\mathbb{D},\ \mathcal{T}_{1,2}\sim\mathbb{T}}\ \ell(\mathbf{z}_1, \mathbf{z}_2),$$
$$s.t.\ cov(\mathbf{z}_i, \mathbf{z}_i) = \mathbf{I},\ i \in \{1,2\}.$$

$$\min_{\theta} \mathcal{L}(\mathbf{X};\theta) = \mathbb{E}_{\mathbf{X}\sim\mathbb{D},\ \mathcal{T}_{1,2}\sim\mathbb{T}}\ \|\widehat{\mathbf{Z}}_1 - \widehat{\mathbf{Z}}_2\|_F^2$$
$$with\ \widehat{\mathbf{Z}}_i = \Phi(\mathbf{Z}_i),\ i \in \{1,2\},$$

Motivations of whitening loss

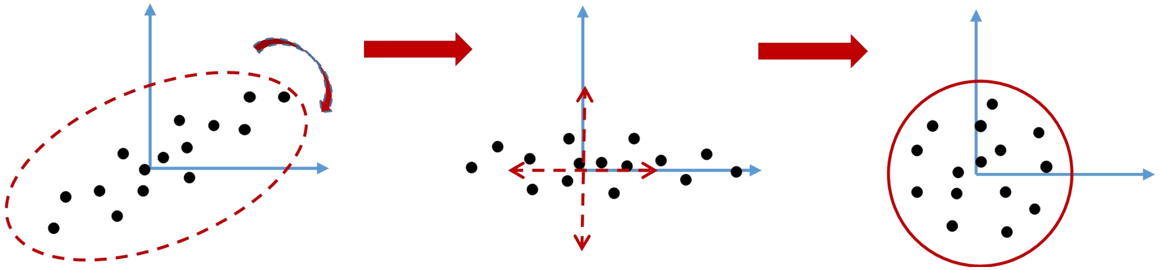
- Motivations of whitening loss:

1. Whitening operation can **remove the correlation among axes**
2. A whitened representation ensures the examples scattered in a **spherical distribution**



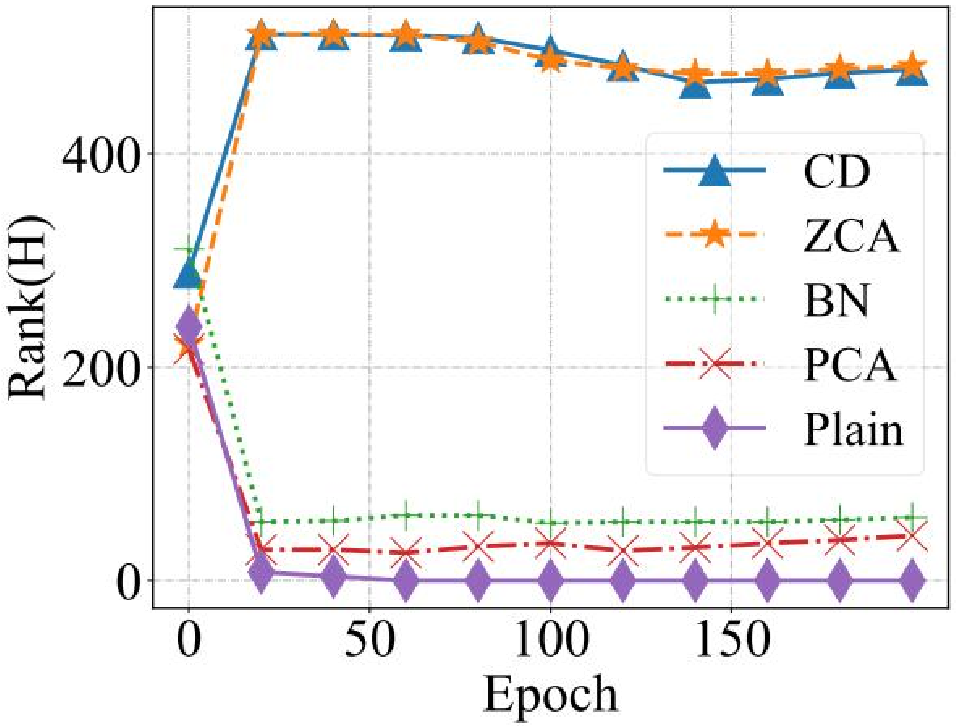
Are motivations of whitening loss correct?

➢ PCA Whitening (can also remove the correlation among axes)



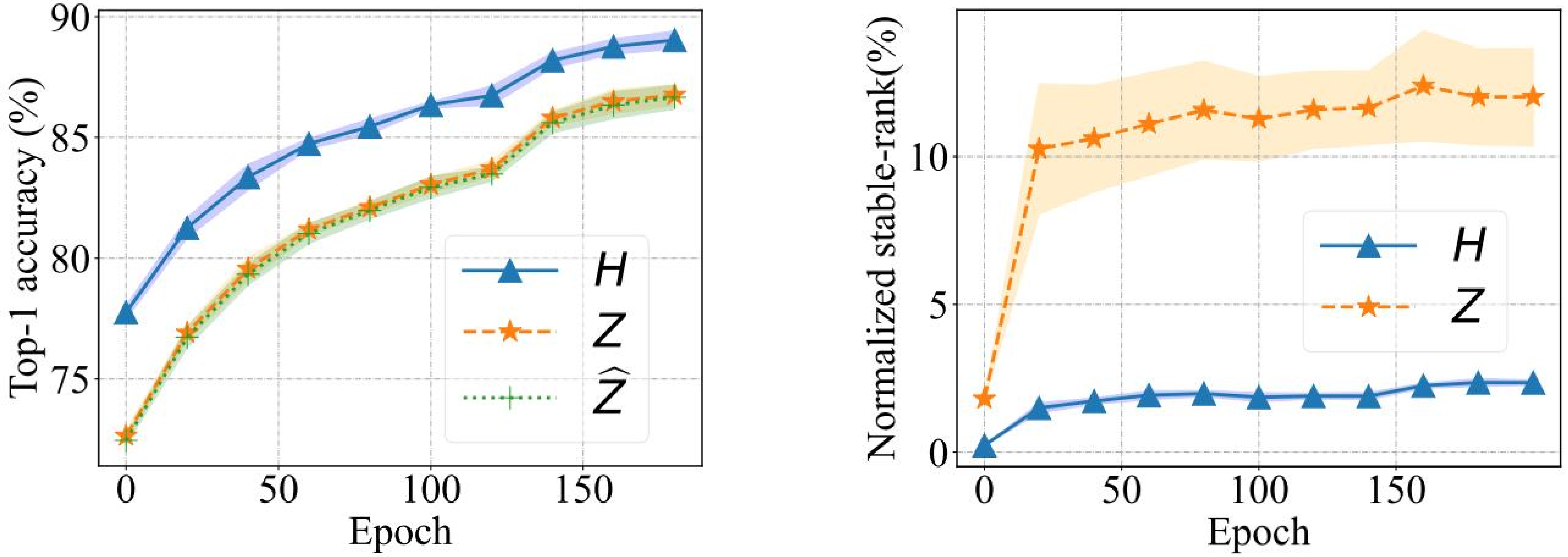
✓ A PCA whitened representation also ensures the examples scattered in a spherical distribution

➢ However, PCA Whitening Fails to Avoid Dimensional Collapse



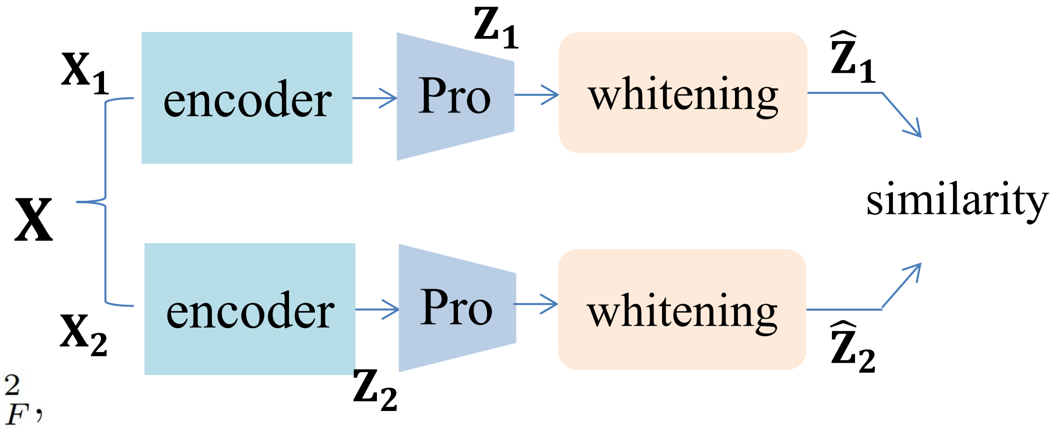
Are motivations of whitening loss correct?

➢Whitened Output is not a Good Representation.



The normalized stable-rank of $\widehat{Z}$ is always 100%

Analysing Decomposition of Whitening Loss

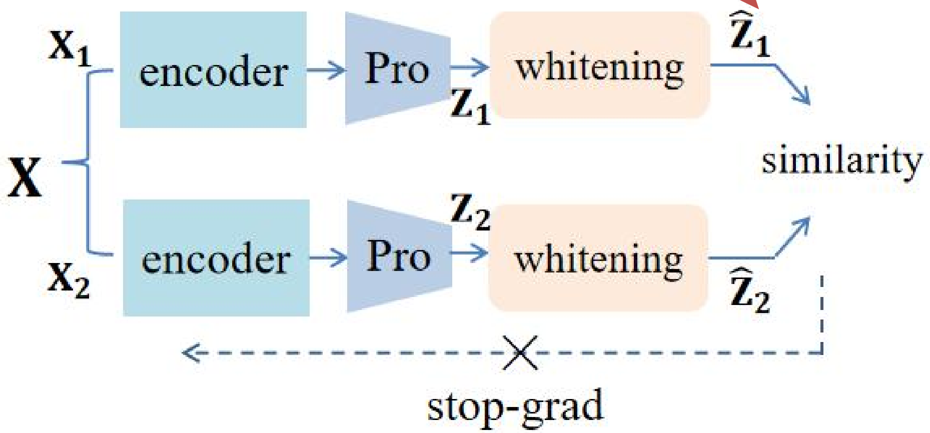


- Whitening loss:

$$\mathcal{L}(\mathbf{X}) = \frac{1}{m}\|\widehat{\mathbf{Z}}_1 - \widehat{\mathbf{Z}}_2\|_F^2.$$

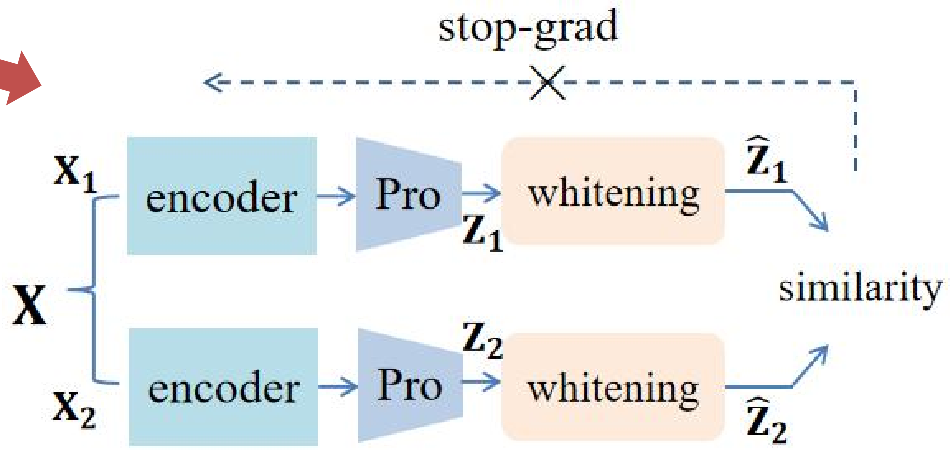
- A proxy loss:

$$\mathcal{L}'(\mathbf{X}) = \underbrace{\frac{1}{m}\|\widehat{\mathbf{Z}}_1 - (\widehat{\mathbf{Z}}_2)_{st}\|_F^2}_{\mathcal{L}'_1} + \underbrace{\frac{1}{m}\|(\widehat{\mathbf{Z}}_1)_{st} - \widehat{\mathbf{Z}}_2\|_F^2}_{\mathcal{L}'_2},$$



- We can prove

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}'}{\partial \theta}$$



Minimizing $\mathcal{L}'_1$ only require the embedding $\boldsymbol{Z}_1$ being full-rank, not whitened

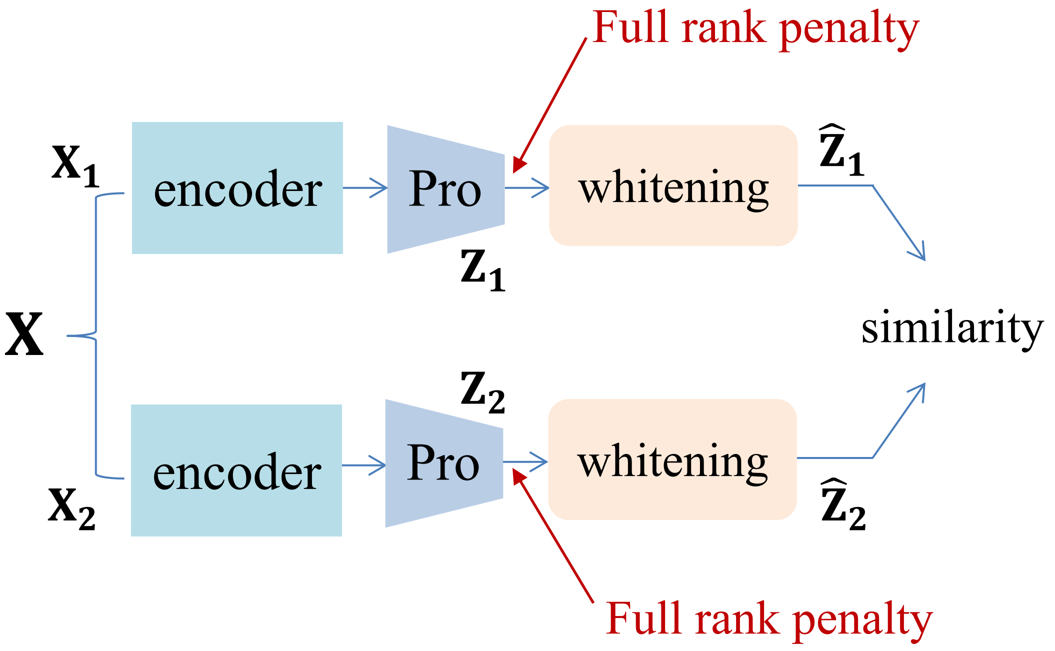
Connection to Soft Whitening

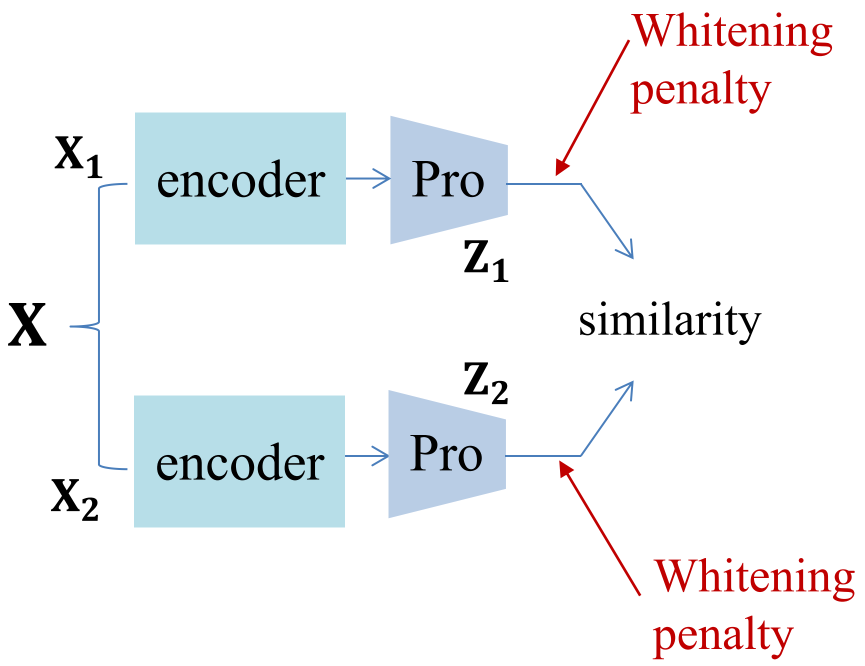
- Whitening loss:

$$\mathcal{L}(\mathbf{X}) = \frac{1}{m}\|\widehat{\mathbf{Z}}_1 - \widehat{\mathbf{Z}}_2\|_F^2.$$

- VICReg:

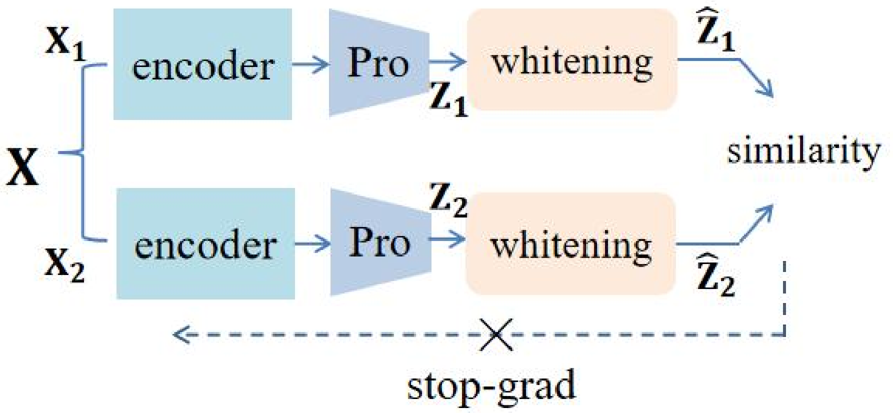
$$\mathcal{L}(\mathbf{X}) = \frac{1}{m}\|\mathbf{Z}_1 - \mathbf{Z}_2\|_F^2 + \alpha\sum_{i=1}^{2}(\|\frac{1}{m}\mathbf{Z}_i\mathbf{Z}_i^T - \lambda\mathbf{I}\|_F^2),$$

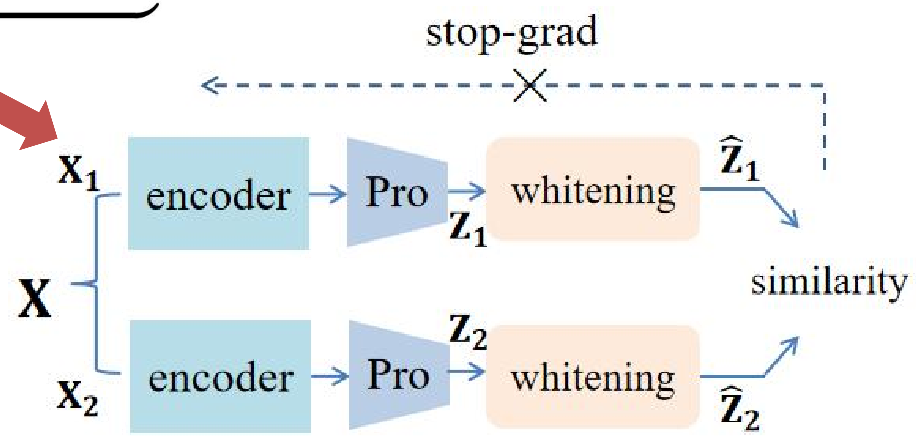


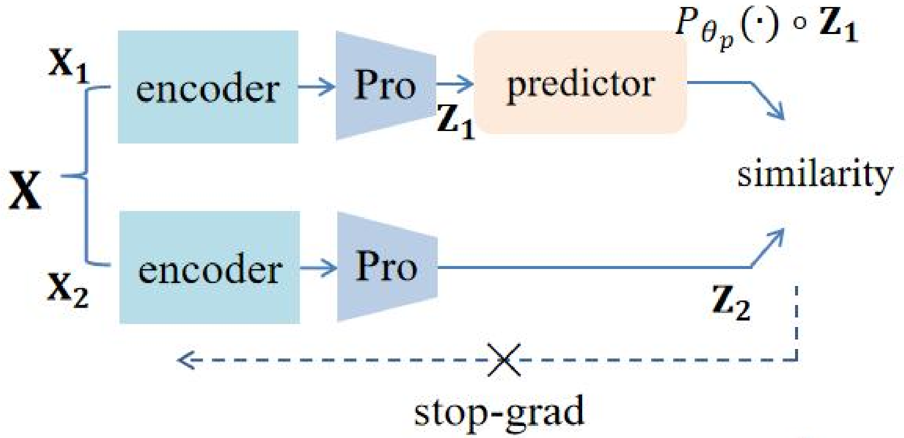


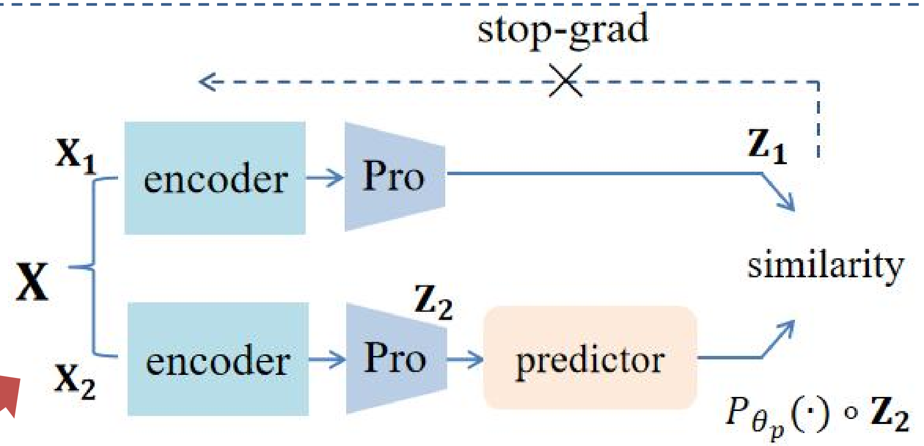
Connection to Asymmetirc Methods

- Whitening loss: $$\mathcal{L}^{'}(\mathbf{X}) = \underbrace{\frac{1}{m}\|\widehat{\mathbf{Z}}_1 - (\widehat{\mathbf{Z}}_2)_{st}\|_F^2}_{\mathcal{L}_1^{'}} + \underbrace{\frac{1}{m}\|(\widehat{\mathbf{Z}}_1)_{st} - \widehat{\mathbf{Z}}_2\|_F^2}_{\mathcal{L}_2^{'}},$$





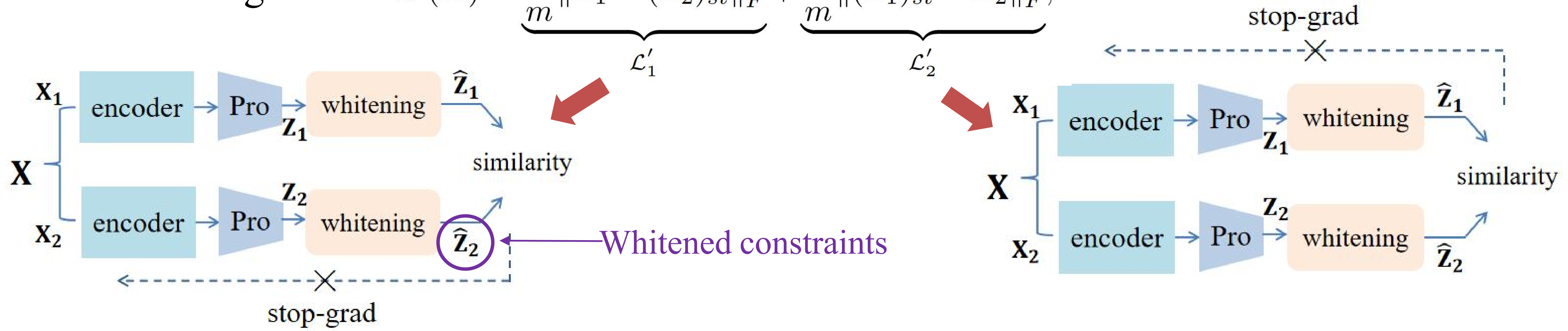


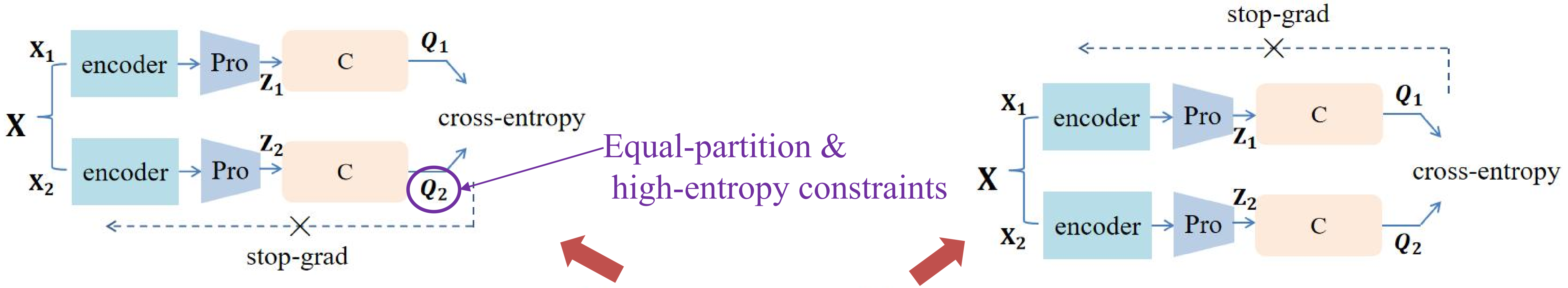


- SimSiam: $$\mathcal{L}(\mathbf{X}) = \frac{1}{m}\|P_{\theta_p}(\cdot) \circ \mathbf{Z}_1 - (\mathbf{Z}_2)_{st}\|_F^2 + \frac{1}{m}\|P_{\theta_p}(\cdot) \circ \mathbf{Z}_2 - (\mathbf{Z}_1)_{st}\|_F^2,$$

Connection to Other Non-contrastive Methods

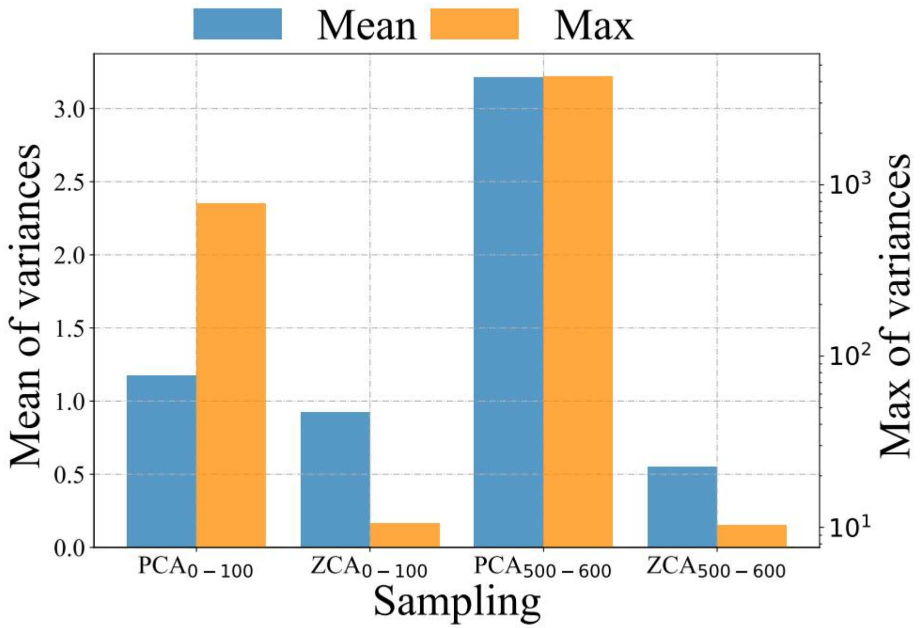
- Whitening loss: $$\mathcal{L}'(\mathbf{X}) = \underbrace{\frac{1}{m}\|\widehat{\mathbf{Z}}_1 - (\widehat{\mathbf{Z}}_2)_{st}\|_F^2}_{\mathcal{L}'_1} + \underbrace{\frac{1}{m}\|(\widehat{\mathbf{Z}}_1)_{st} - \widehat{\mathbf{Z}}_2\|_F^2}_{\mathcal{L}'_2},$$



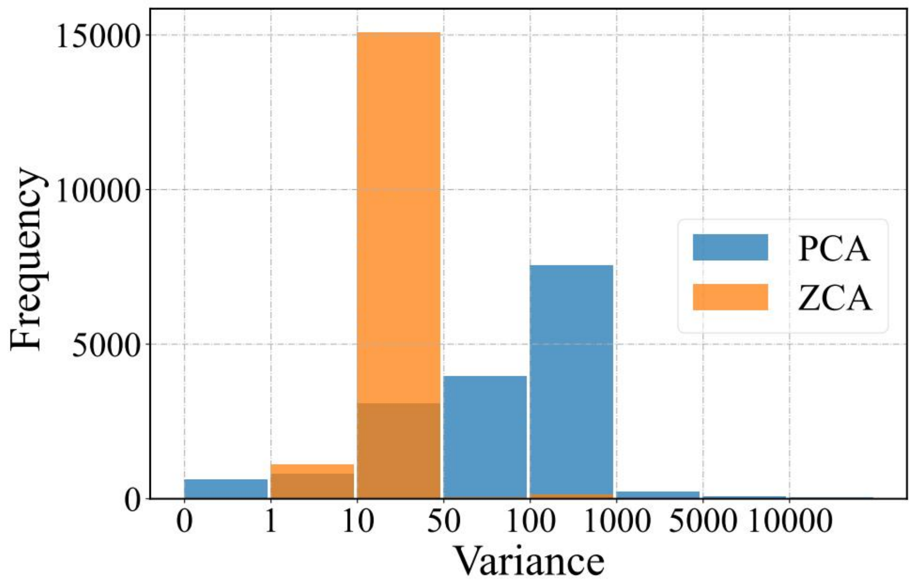


- SwAV: $$\mathcal{L}(\mathbf{X}) = \ell(\mathbf{C}^T\mathbf{Z}_1, (\mathbf{Q}_2)_{st}) + \ell(\mathbf{C}^T\mathbf{Z}_2, (\mathbf{Q}_1)_{st}).$$

Why PCA Whitening Fails to Avoid Dimensional Collapse?



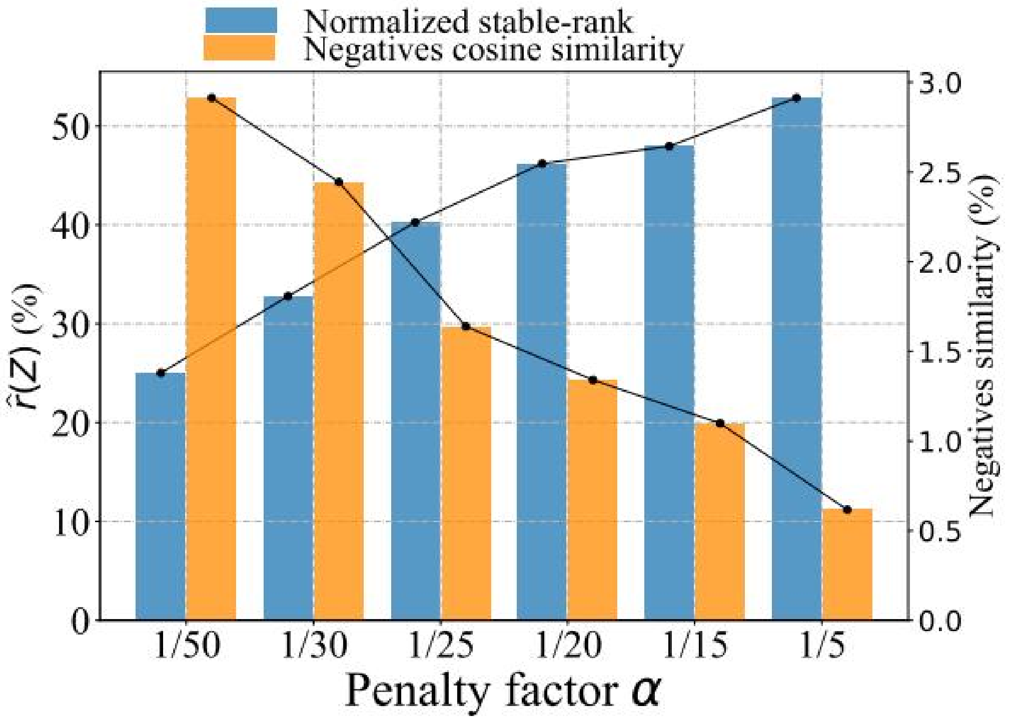
(a)



(b)

- PCA whitening: volatile sequence of whitened targets

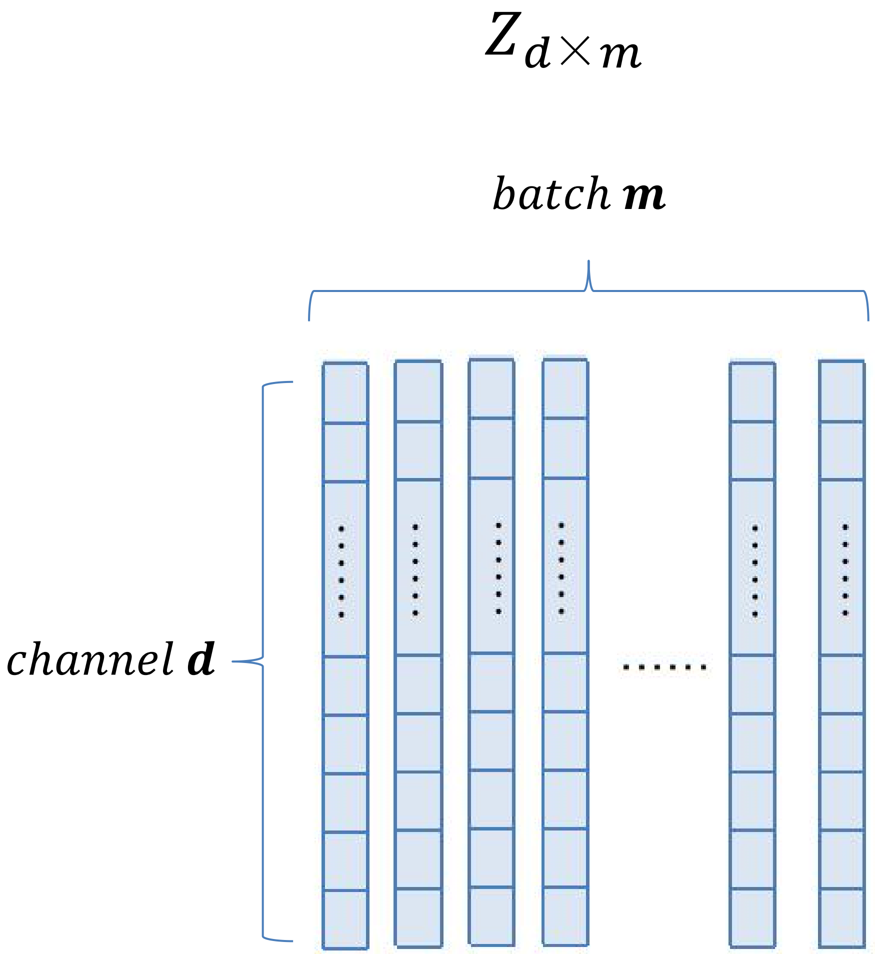
Why Whitened Output is not a Good Representation?



Similarity decreases when extent of whitening increases

- ➢ A whitened output leads to the state that can **break the potential manifold the examples in the same class belong to**

Channel Whitening (CW)



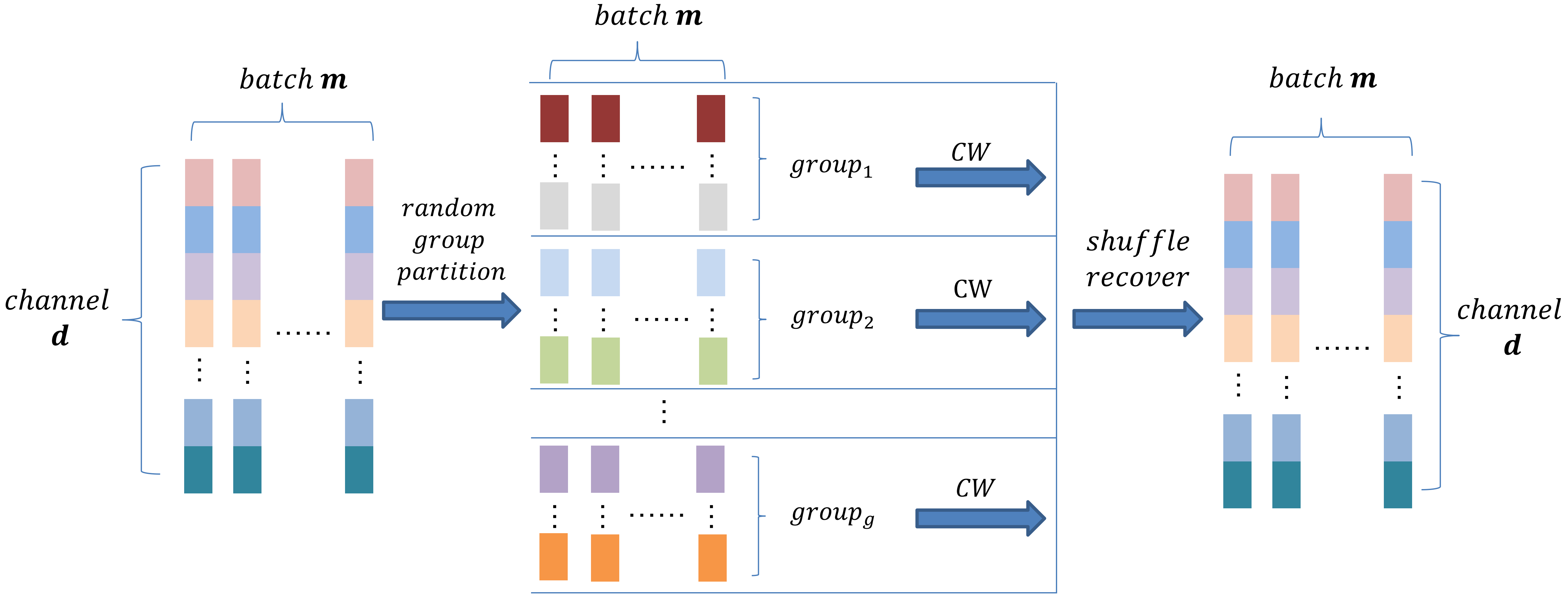
- Batch whitening (BW)
  - $centering: Z_B = Z \cdot (I - \frac{1}{m} 1 \cdot 1^T)$
  - $\Sigma = \frac{1}{m-1} Z_B \cdot {Z_B}^T$
  - $\widehat{Z} = \Phi \cdot Z_B$

requires **m > d** to avoid numerical instability.

- Channel whitening (CW)
  - $centering: Z_c = (I - \frac{1}{d} 1 \cdot 1^T) \cdot Z$
  - $\Sigma = \frac{1}{d-1} {Z_C}^T \cdot Z_C$
  - $\widehat{Z} = Z_C \cdot \Phi$

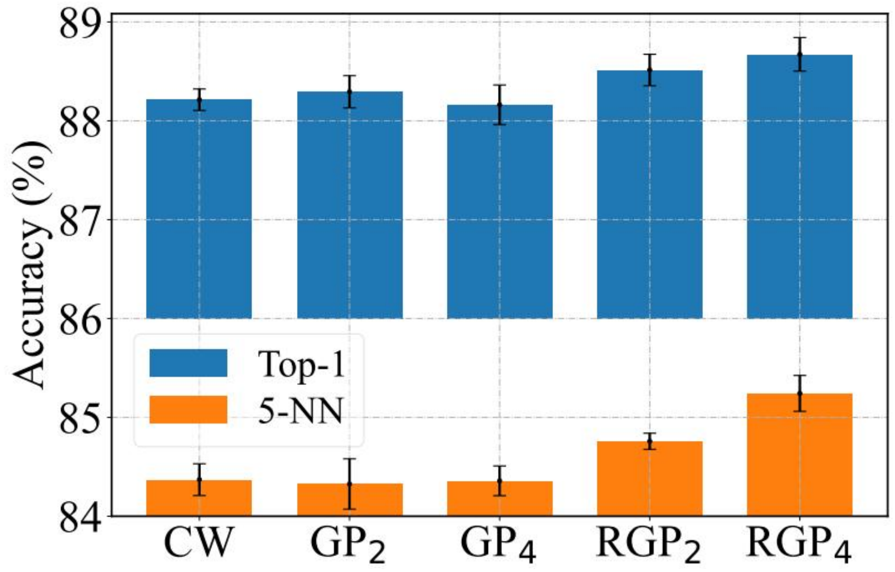
can obtain numerical stability **when the batch size is small**, since the condition that **d > m** can be obtained by design.

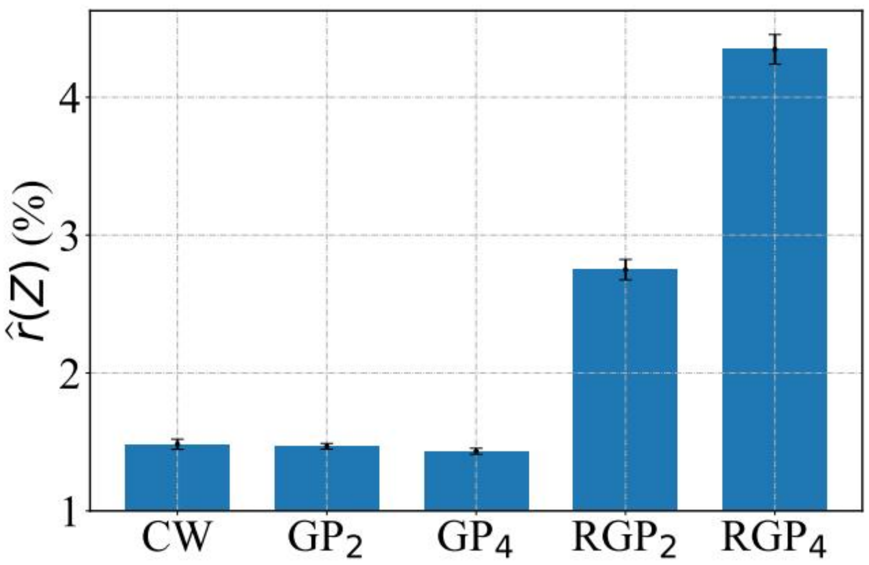
Random Group Partition (RGP)

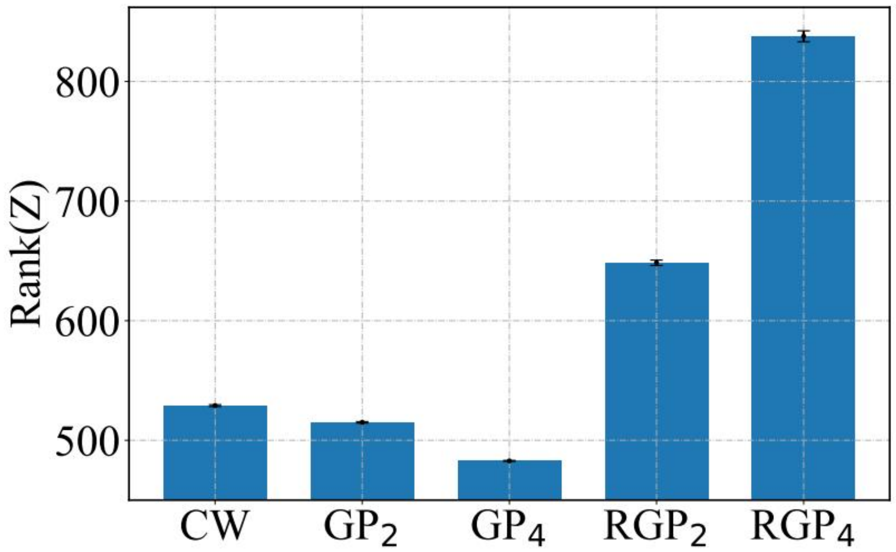


$$dimension\ of\ every\ group = \frac{d}{g}$$

Random Group Partition (RGP)







Method	CIFAR-10		CIFAR-100	
	linear	5-nn	linear	5-nn
CW 2	91.66	88.99	66.26	56.36
CW-GP 2	91.61	88.89	66.17	56.53
CW-RGP 2	**91.92**	**89.54**	**67.51**	**57.35**
CW 4	92.10	90.12	66.90	57.12
CW-GP 4	92.08	90.06	67.34	57.28
CW-RGP 4	**92.47**	**90.74**	**68.26**	**58.67**

Experiments for Empirical Study

➢ Experimental Setup for Comparison of Baselines

Table 1: Classification accuracy (top 1) of a linear classifier and a 5-nearest neighbors classifier for different loss functions and datasets with a ResNet-18 encoder.

Method	CIFAR-10		CIFAR-100		STL-10		Tiny-ImageNet	
	linear	5-nn	linear	5-nn	linear	5-nn	linear	5-nn
SimCLR [6]	91.80	88.42	66.83	56.56	90.51	85.68	48.84	32.86
BYOL [16]	91.73	89.45	66.60	56.82	91.99	88.64	**51.00**	**36.24**
SimSiam [8] (repro.)	90.51	86.82	66.04	55.79	88.91	84.84	48.29	34.21
Shuffled-DBN [21] (repro.)	90.45	88.15	66.07	56.97	89.20	84.51	48.60	32.14
Barlow Twins [45] (repro.)	88.51	86.53	65.78	55.76	88.36	83.71	47.44	32.65
VICReg [2] (repro.)	90.32	88.41	66.45	56.78	90.78	85.72	48.71	33.35
Zero-ICL [48] (repro.)	88.12	86.64	61.91	53.47	86.35	82.51	46.25	32.74
W-MSE 2 [12]	91.55	89.69	66.10	56.69	90.36	87.10	48.20	34.16
W-MSE 4 [12]	91.99	89.87	67.64	56.45	91.75	88.59	49.22	35.44
CW-RGP 2 (ours)	91.92	89.54	67.51	57.35	90.76	87.34	49.23	34.04
CW-RGP 4 (ours)	**92.47**	**90.74**	**68.26**	**58.67**	**92.04**	**88.95**	50.24	35.99

Experiments for Empirical Study

➢ Experimental Setup for Large-Scale Classification

Table 2: Comparisons on ImageNet linear classification. All are based on ResNet-50 encoder. The table is mostly inherited from [8].

Method	Batch size	100 eps	200 eps
SimCLR [6]	4096	66.5	68.3
MoCo v2 [7]	256	67.4	69.9
BYOL [16]	4096	66.5	70.6
SwAV [4]	4096	66.5	69.1
SimSiam [8]	256	68.1	70.0
W-MSE 4 [12]	4096	69.4	-
Zero-CL [48]	1024	68.9	-
BYOL [16] (repro.)	512	66.1	69.2
SwAV [4] (repro.)	512	65.8	67.9
W-MSE 4 [12] (repro.)	512	66.7	67.9
CW-RGP 4 (ours)	512	**69.7**	**71.0**

Experiments for Empirical Study

- Transfer to downstream tasks

Table 3: Transfer Learning. All competitive unsupervised methods are based on 200-epoch pre-training in ImageNet (IN). The table is mostly inherited from [8]. Our CW-RGP is performed with 3 random seeds, with mean and standard deviation reported.

Method	VOC 07+12 detection			COCO detection			COCO instance seg.		
	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
Scratch	60.2	33.8	33.1	44.0	26.4	27.8	46.9	29.3	30.8
IN-supervised	81.3	53.5	58.8	58.2	38.2	41.2	54.7	33.3	35.2
SimCLR [6]	81.8	55.5	61.4	57.7	37.9	40.9	54.6	33.3	35.3
MoCo v2 [7]	**82.3**	57.0	63.3	58.8	39.2	42.5	55.5	34.3	36.6
BYOL [16]	81.4	55.3	61.1	57.8	37.9	40.9	54.3	33.2	35.0
SwAV [4]	81.5	55.4	61.4	57.6	37.6	40.3	54.2	33.1	35.1
SimSiam [8]	82.0	56.4	62.8	57.5	37.9	40.9	54.2	33.2	35.2
CW-RGP (ours)	$82.2_{\pm 0.07}$	$\mathbf{57.2}_{\pm 0.10}$	$\mathbf{63.8}_{\pm 0.11}$	$\mathbf{60.5}_{\pm 0.28}$	$\mathbf{40.7}_{\pm 0.14}$	$\mathbf{44.1}_{\pm 0.14}$	$\mathbf{57.3}_{\pm 0.16}$	$\mathbf{35.5}_{\pm 0.12}$	$\mathbf{37.9}_{\pm 0.14}$

- Take Away
  - An in-depth analysis in whitening loss
  - An effective SSL method: CW-RGP

Thank you

https://github.com/winci-ai/CW-RGP