Recent Progress on
Self-Supervised Representation Learning

Self-Supervised Representation Learning

- Learn image features without human labels
- Map similar semantics closer
- Transferrable to downstream tasks

Instance discrimination:

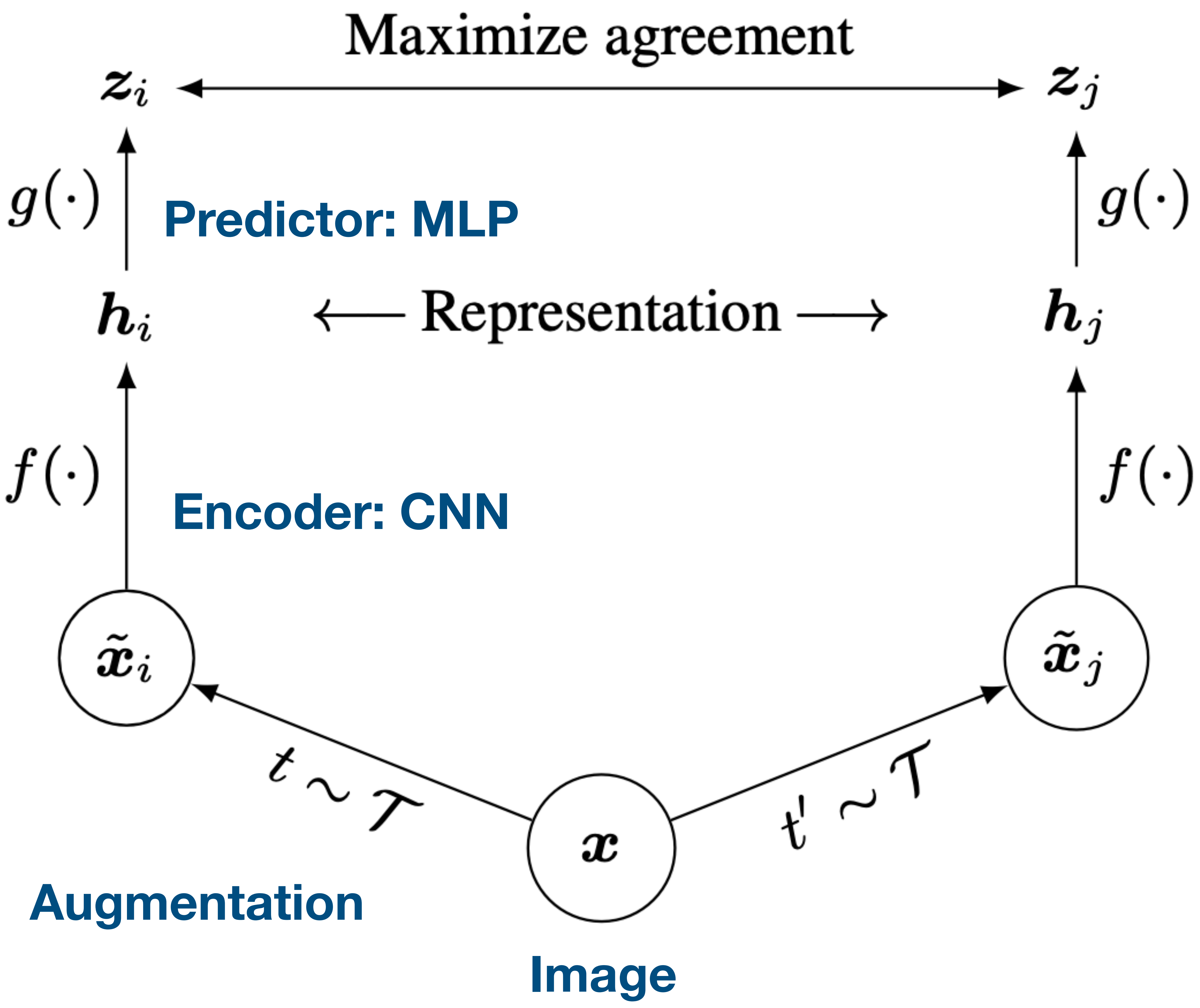
- Augmentations of the same image has similar features
- Augmentations of different images has distinct features

Recent papers:

- SimCLR
- SwAV
- BYOL
- SimSiam

A simple framework for contrastive learning of visual representations (SimCLR)

Chen, Ting, et al. ICML 2020



Positive pair: (i, j)
Augmentations of the same image

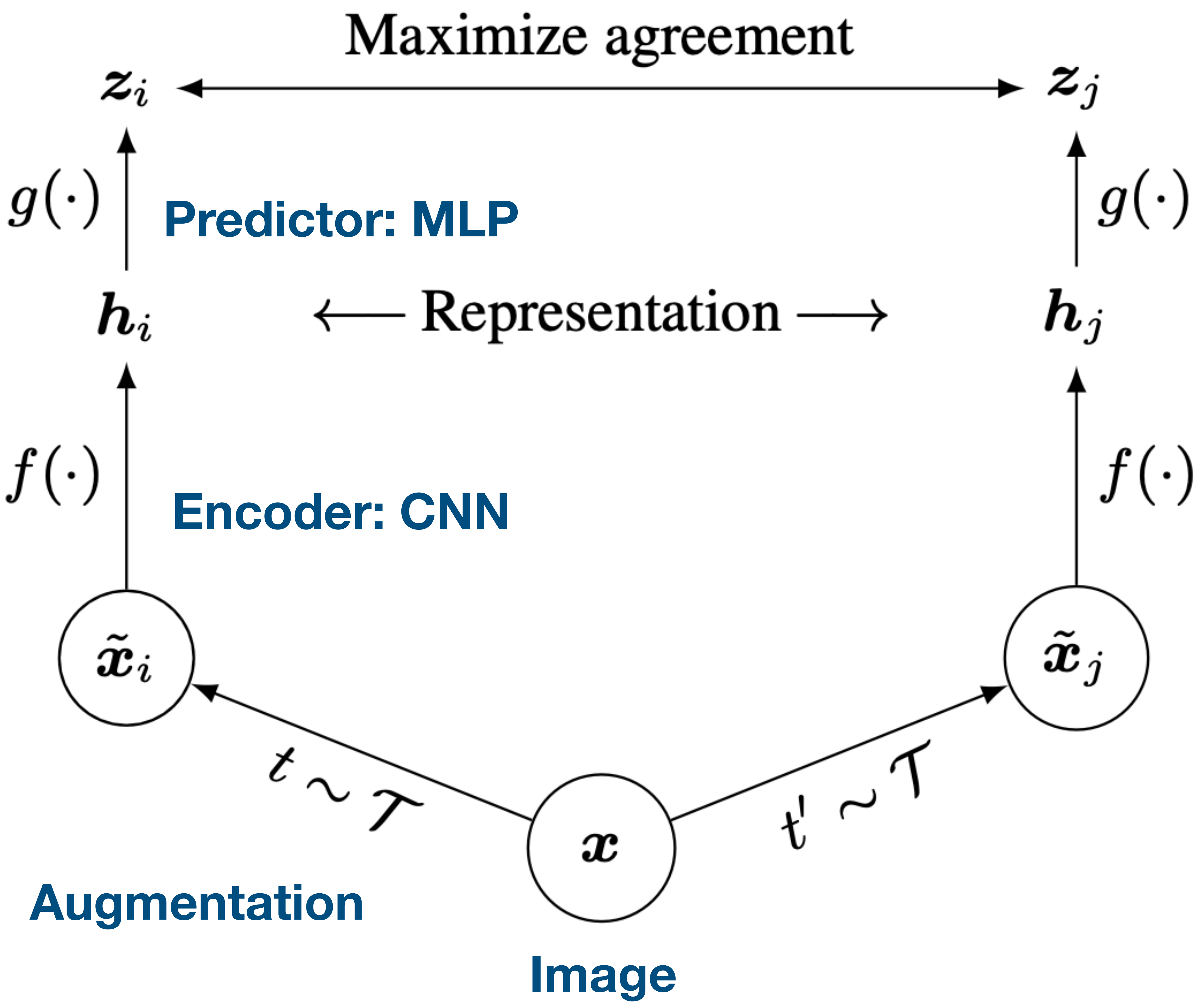
Negative pair:
Augmentations of different images

Contrastive loss (cosine similarity)**:**

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

A simple framework for contrastive learning of visual representations (SimCLR)

Chen, Ting, et al. ICML 2020



Positive pair: (i, j)
Augmentations of the same image

Negative pair:
Augmentations of different images

Contrastive loss (cosine similarity)**:**

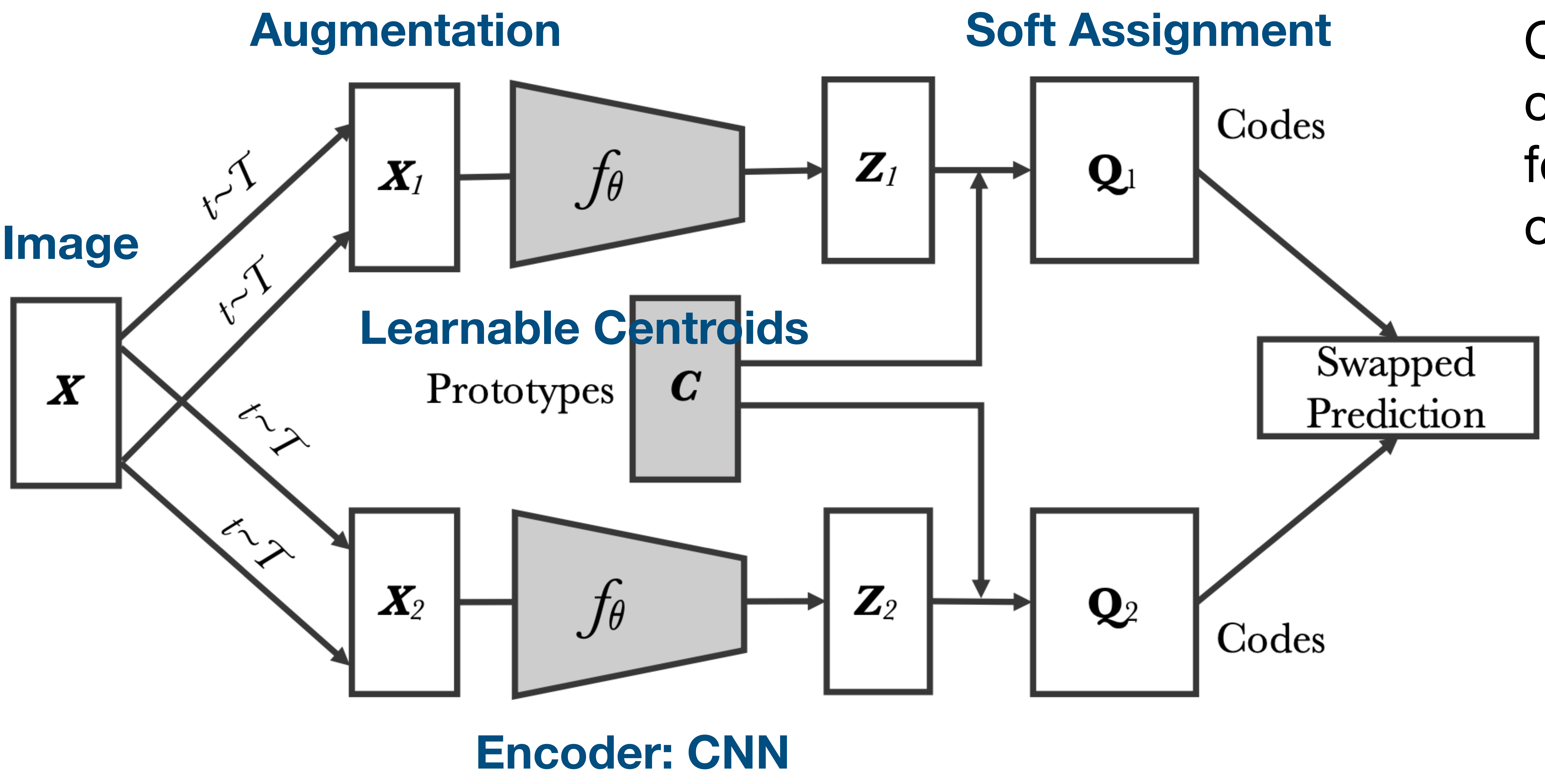
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\boxed{\sum_{k=1}^{2N}} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)},$$

*Negative pairs prevent features **collapse** to be constant*

but require expensive large batch size

Unsupervised learning of visual features by contrasting cluster assignments (SwAV)

Caron, Mathilde, et al. NIPS 2020



Online clustering and predict the codes of an augmentation using the features of the other augmentation of the same image

$$L(\mathbf{z}_t, \mathbf{z}_s) = \ell(\mathbf{z}_t, \mathbf{q}_s) + \ell(\mathbf{z}_s, \mathbf{q}_t),$$

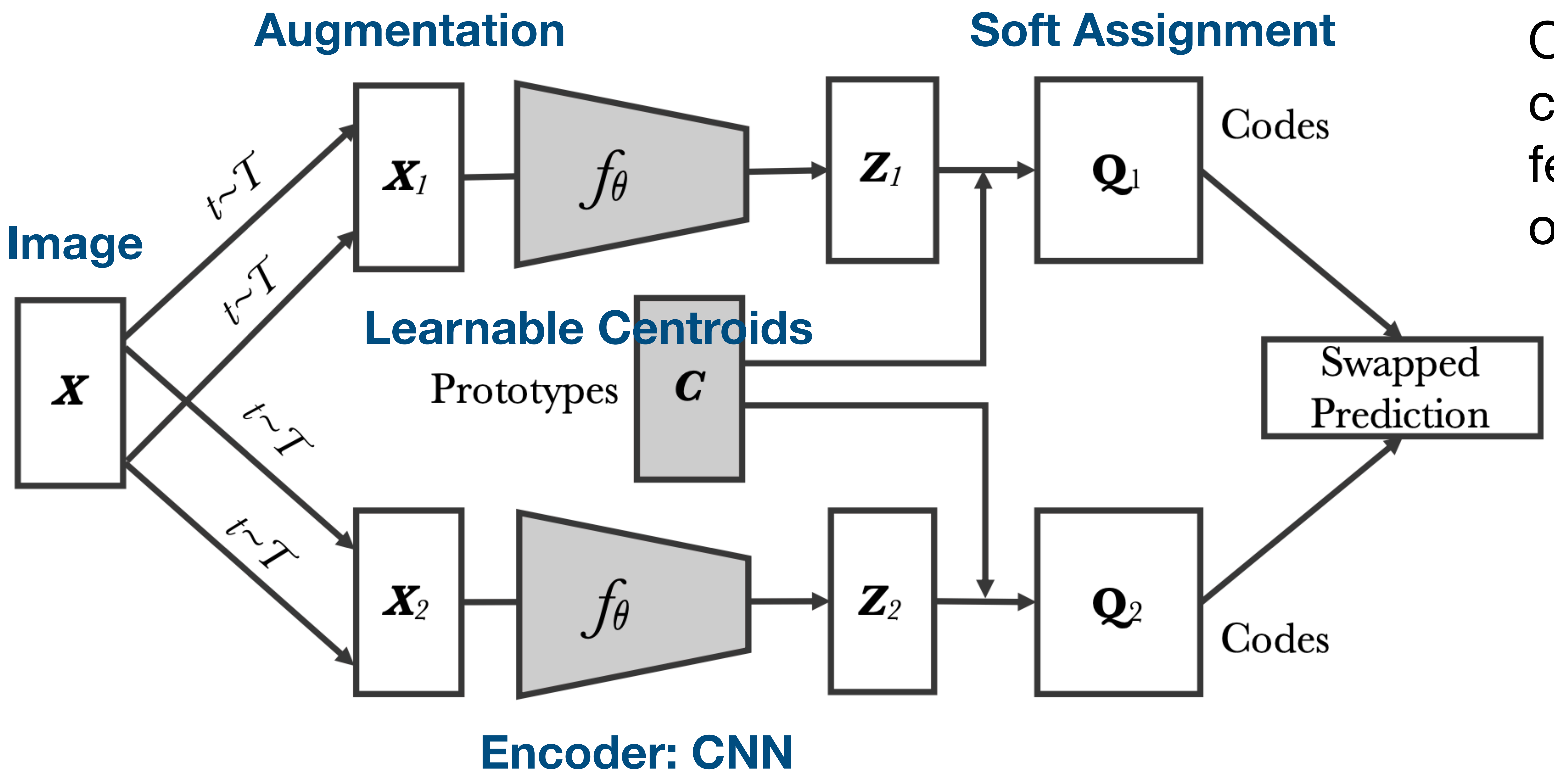
$$\ell(\mathbf{z}_t, \mathbf{q}_s) = -\sum_k \mathbf{q}_s^{(k)} \log \mathbf{p}_t^{(k)},$$

$$\text{where} \quad \mathbf{p}_t^{(k)} = \frac{\exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_{k'}\right)}.$$

In short, augmentations of the same image should belong to the same cluster

Unsupervised learning of visual features by contrasting cluster assignments (SwAV)

Caron, Mathilde, et al. NIPS 2020



Online clustering and predict the codes of an augmentation using the features of the other augmentation of the same image

$$L(\mathbf{z}_t, \mathbf{z}_s) \quad = \quad \ell(\mathbf{z}_t, \mathbf{q}_s) + \ell(\mathbf{z}_s, \mathbf{q}_t),$$

$$\ell(\mathbf{z}_t, \mathbf{q}_s) = -\sum_k \mathbf{q}_s^{(k)} \log \mathbf{p}_t^{(k)},$$

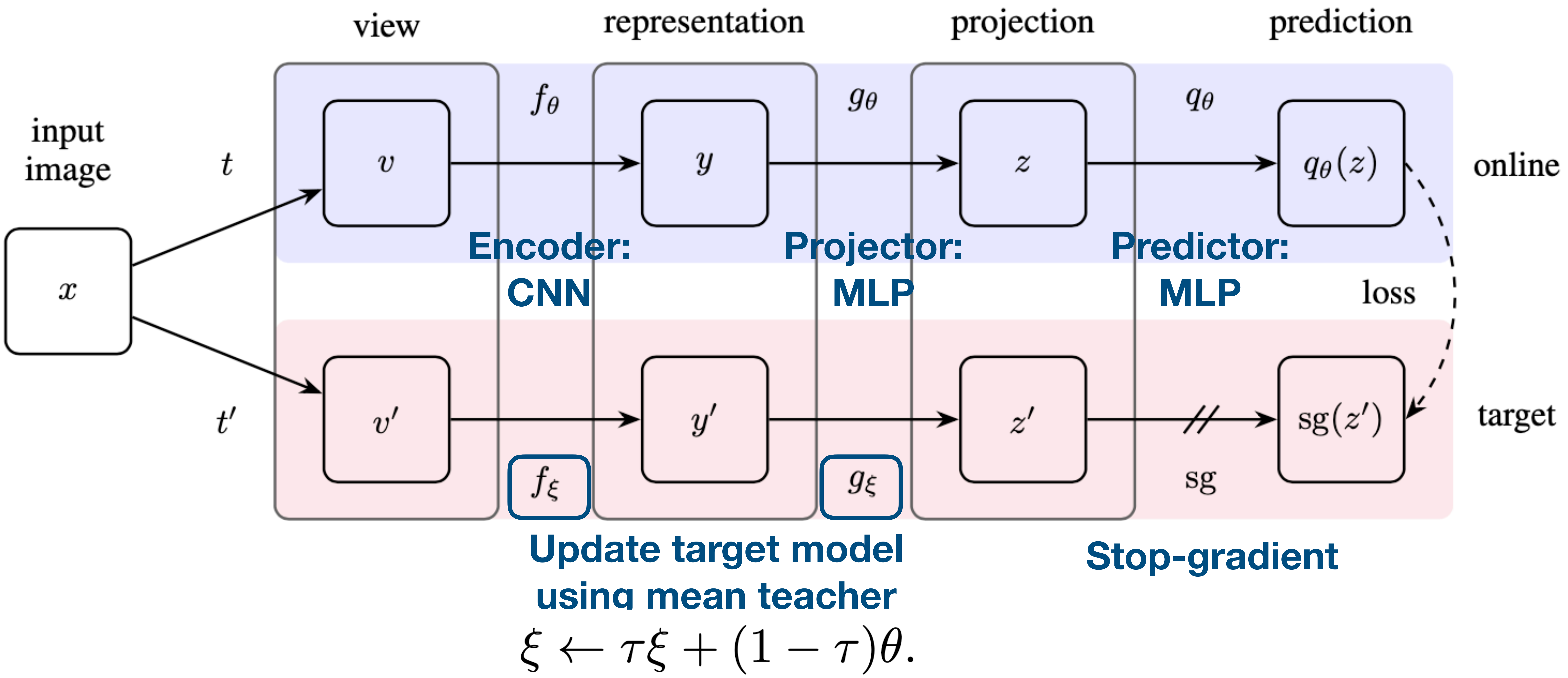
$$\text{where} \quad \mathbf{p}_t^{(k)} = \frac{\exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_{k'}\right)}.$$

SwAV has no negative pairs and prevent collapsing by constraining that the samples in a batch should be equally partitioned by the clusters

In short, augmentations of the same image should belong to the same cluster

Bootstrap your own latent: A new approach to self-supervised learning (BYOL)

Grill, Jean-Bastien, et al. Arxiv 2020

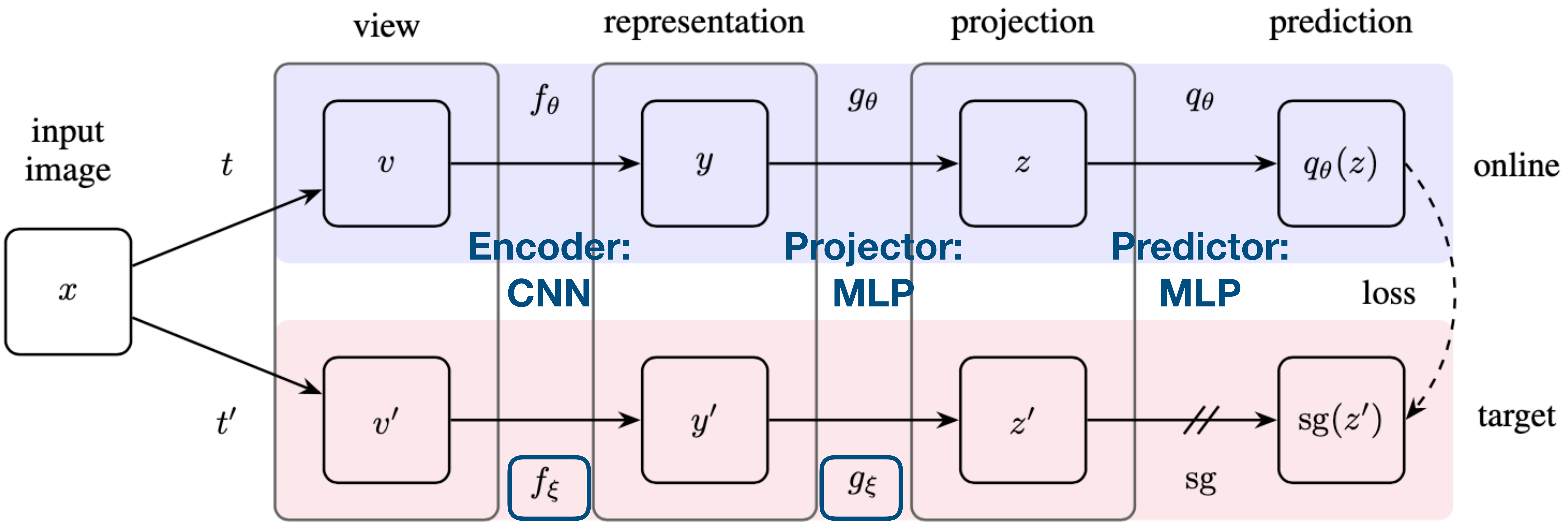


MSE Loss between final normalized features

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

Bootstrap your own latent: A new approach to self-supervised learning (BYOL)

Grill, Jean-Bastien, et al. Arxiv 2020



Update target model using mean teacher

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

Stop-gradient

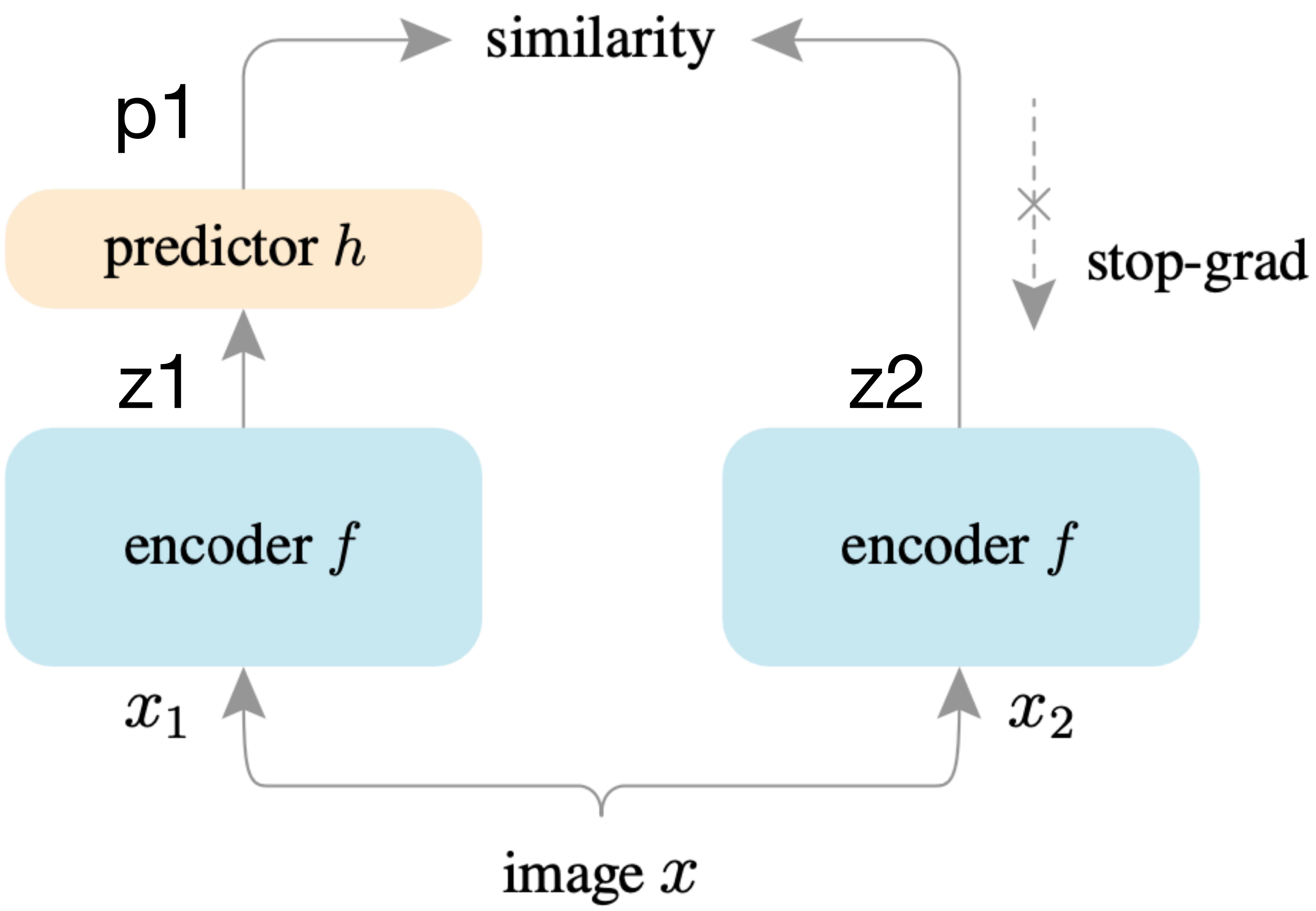
BYOL has no negative pairs and prevent collapsing by something unknown (probably the delicate balance provided by the mean teacher).

MSE Loss between final normalized features

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

Exploring Simple Siamese Representation Learning (SimSiam)

Chen, Xinlei, and Kaiming He. Arxiv 2020



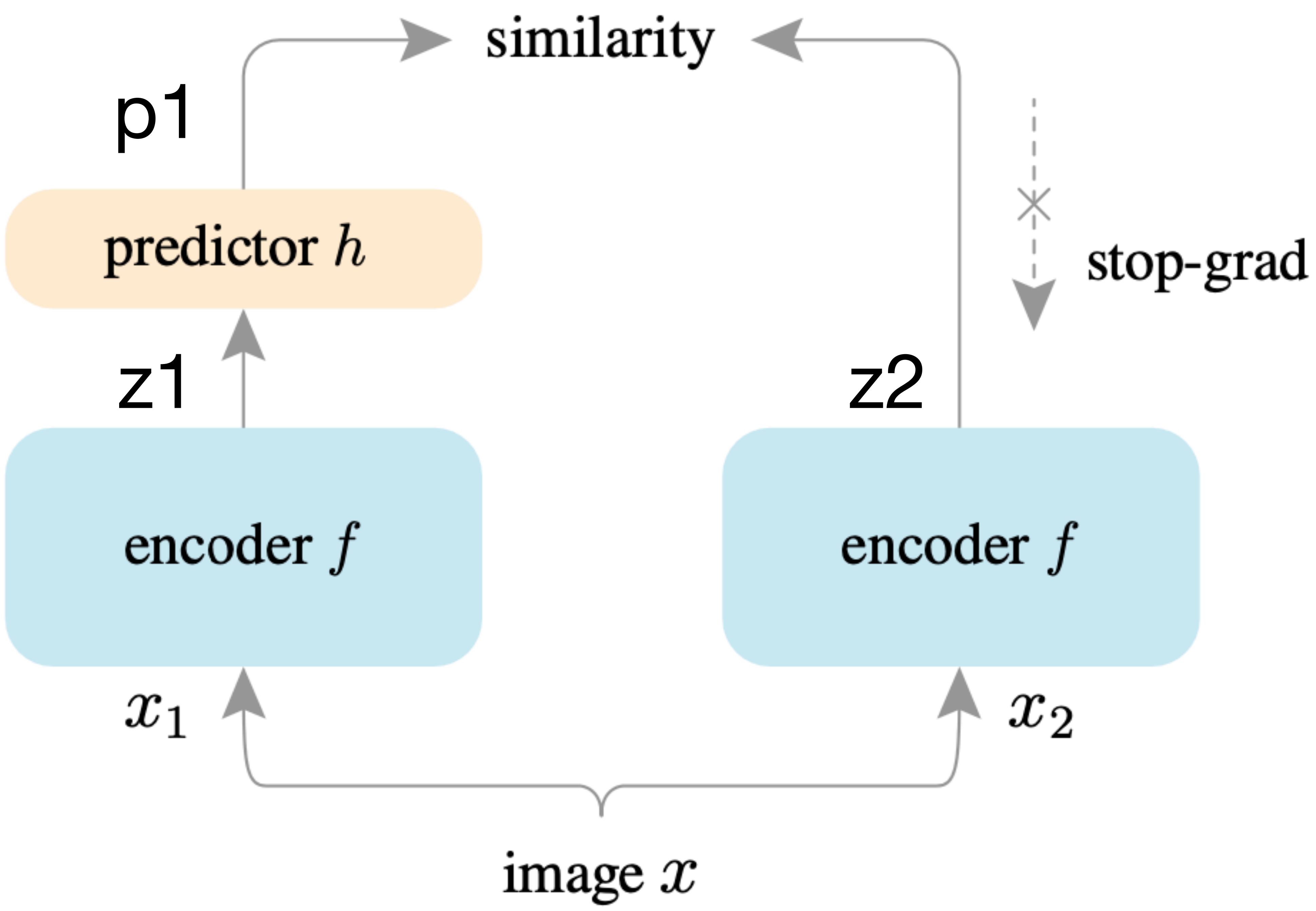
Maximize negative cosine similarities (*D*) with stop gradient:

$$\mathcal{L}=\frac{1}{2}\mathcal{D}(p_1, \texttt{stopgrad}(z_2))+\frac{1}{2}\mathcal{D}(p_2, \texttt{stopgrad}(z_1)).$$

Mean teacher with zero momentum

Exploring Simple Siamese Representation Learning (SimSiam)

Chen, Xinlei, and Kaiming He. Arxiv 2020



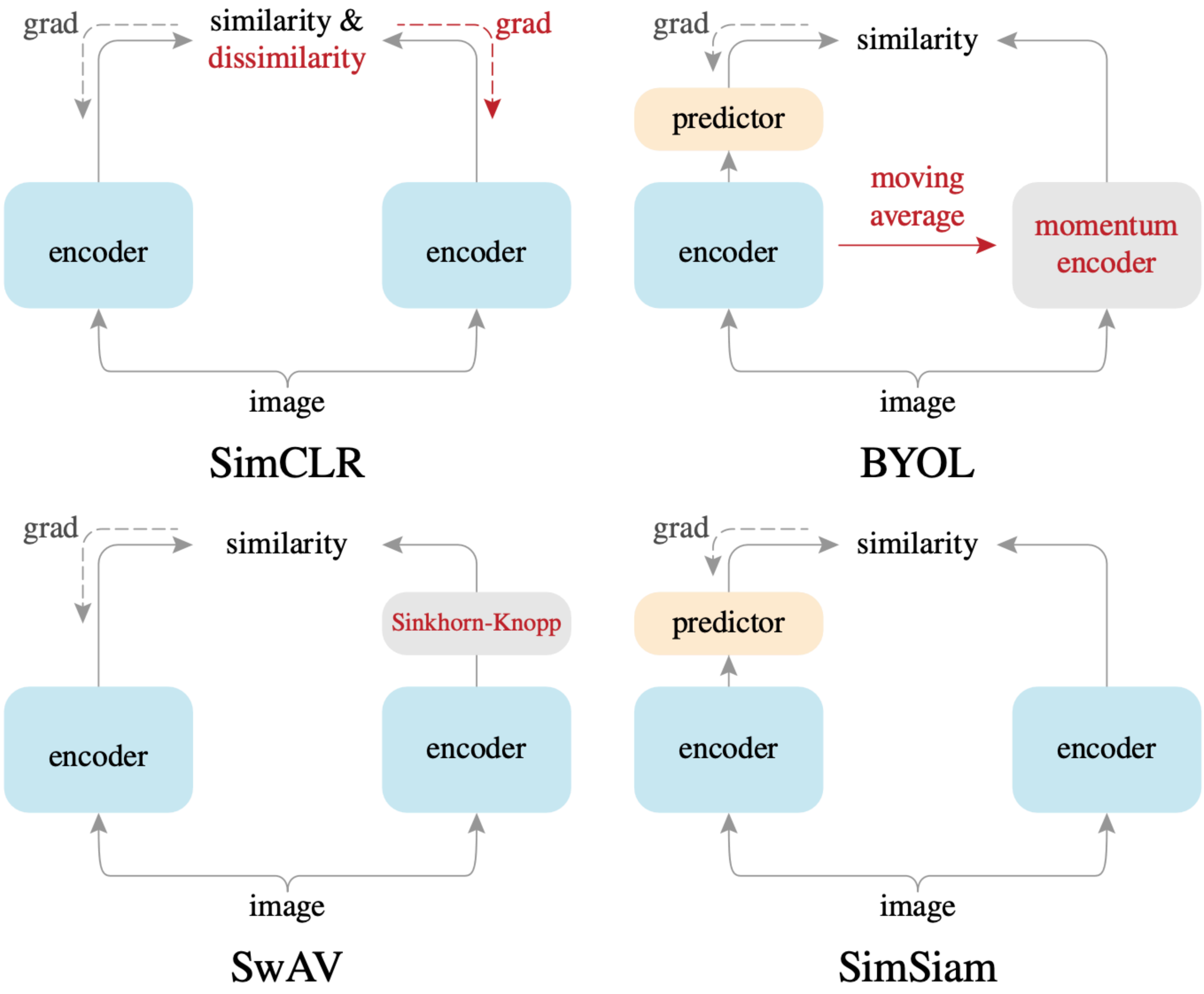
Maximize negative cosine similarities (*D*) with stop gradient:

$$\mathcal{L}=\frac{1}{2}\mathcal{D}(p_1, \texttt{stopgrad}(z_2))+\frac{1}{2}\mathcal{D}(p_2, \texttt{stopgrad}(z_1)).$$

Very simple

SimSiam has no negative pairs and prevent collapsing by something more unknown (empirically only the stop-gradient operation)

Comparison


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

Experiments

100+ GPUs

8 GPUs

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

Supervised ~ 76.5

Table 4. **Comparisons on ImageNet linear classification**. All are based on **ResNet-50** pre-trained with **two 224×224 views**. Evaluation is on a single crop. All competitors are from our reproduction, and "+" denotes *improved* reproduction *vs*. original papers (see supplement).

Leaning signal for self-supervised learning:

- Invariant to augmentations
- Dataset intrinsic structure, e.g., ImageNet is intrinsically for clustering
- Model inductive bias and prior
- Method inductive bias and prior

Future directions for self-supervised learning:

- Augmentation >> From video, From 3D world
- Data >> Active video collection
- Model >> Better model dedicated for self-supervised learning
- Method >> Underlying mechanism and math