01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Bo Pang[1], Yifan Zhang[1], Yaoyi Li[2], Jia Cai[2], and Cewu Lu[1⋆]

[1] Shanghai Jiao Tong University
{pangbo, zhangyf_sjtu, lucewu}@sjtu.edu.cn
[2] HuaWei Technologies Co., Ltd.
{liyaoyi, caijiai1}@huawei.com

Reporter： Hongguang Zhu

1.supervised learning:

- simple and efficient
- expensive and time-consuming annotating costs
- non-ideal generalization (annotation bias)

2. unsupervised learning:

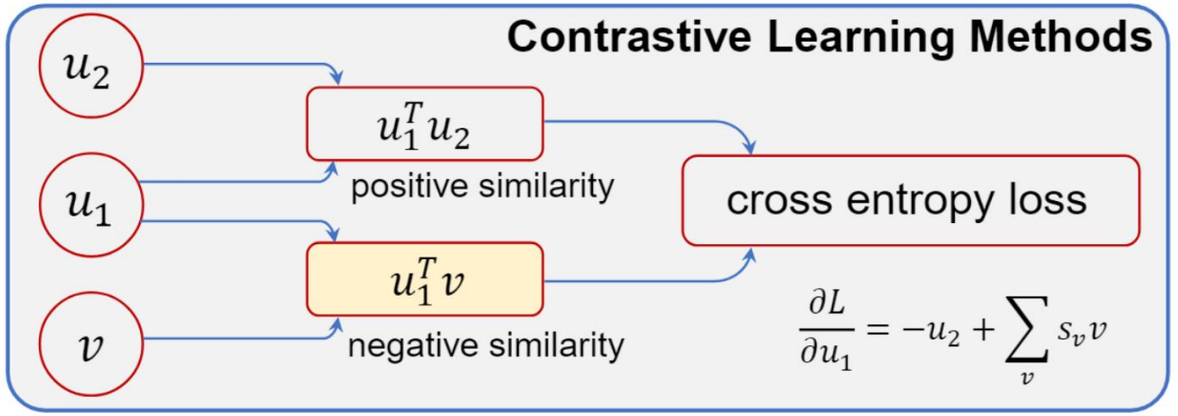
- Instance-level framework
- Clustering-based scheme

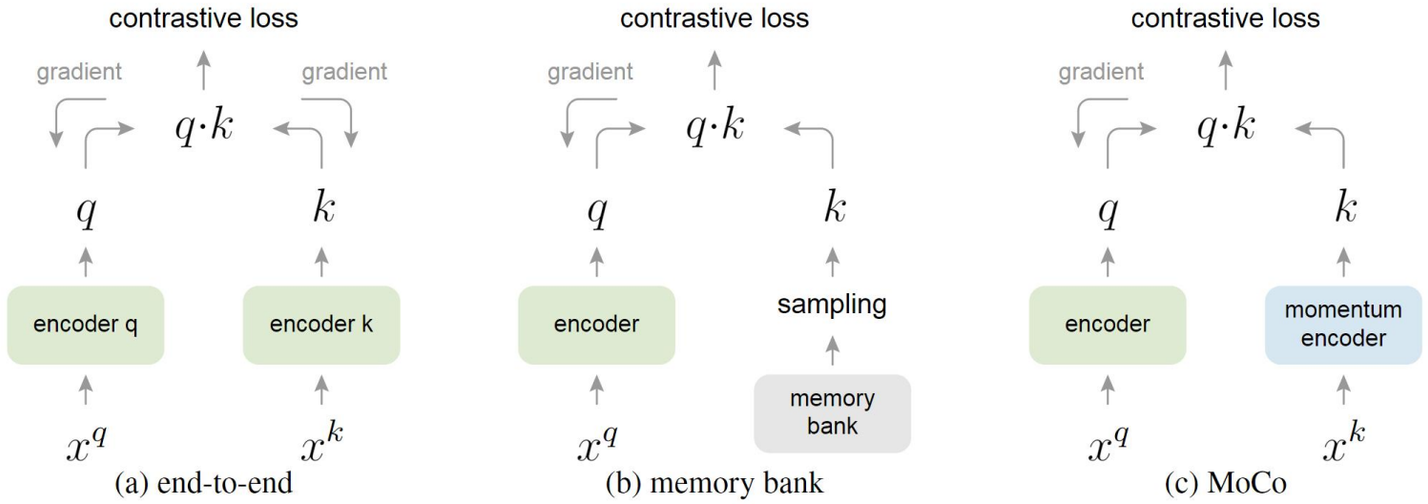
01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Instance-level framework

(1) contrastive learning methods:

- reduce the distance between two augmented views
- push apart view from negative samples





SimCLR, MOCOs...

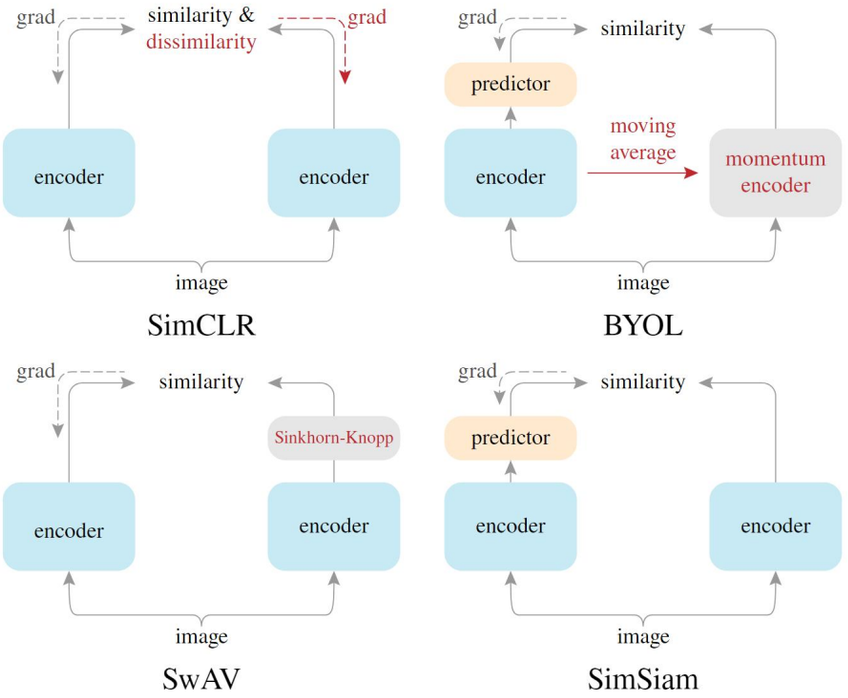
01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

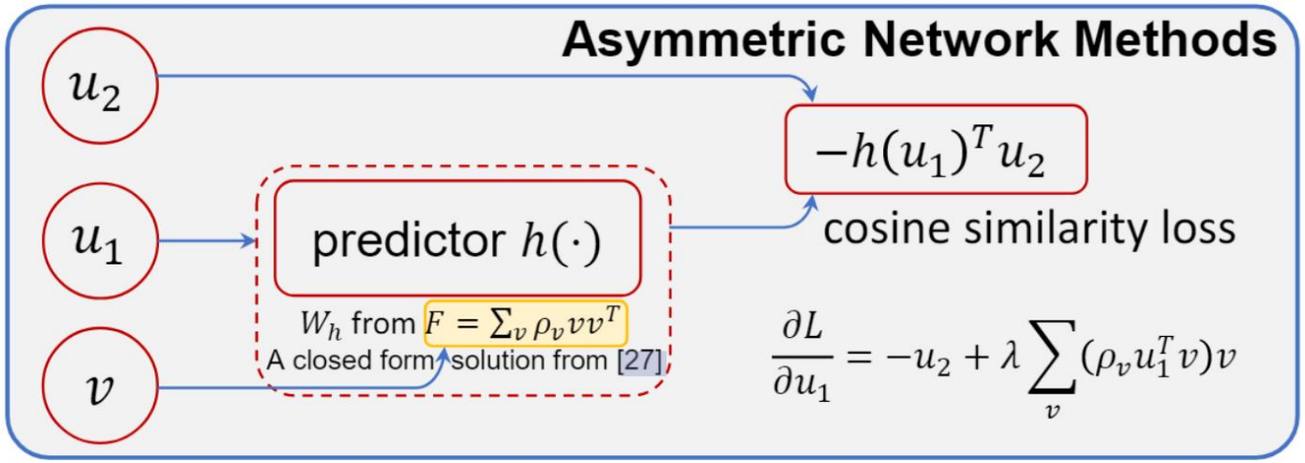
(2) asymmetric network methods

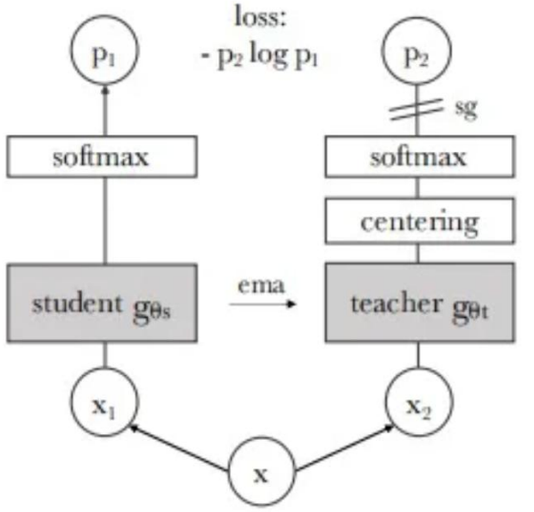
- get rid of negative sample
- introduce predictor network
- stop-gradient operation

idea: only adopt positive samples ↔ mode collapse

- encode pervious sample information in weight of predictor
- serve as negative gradient during back-propagation







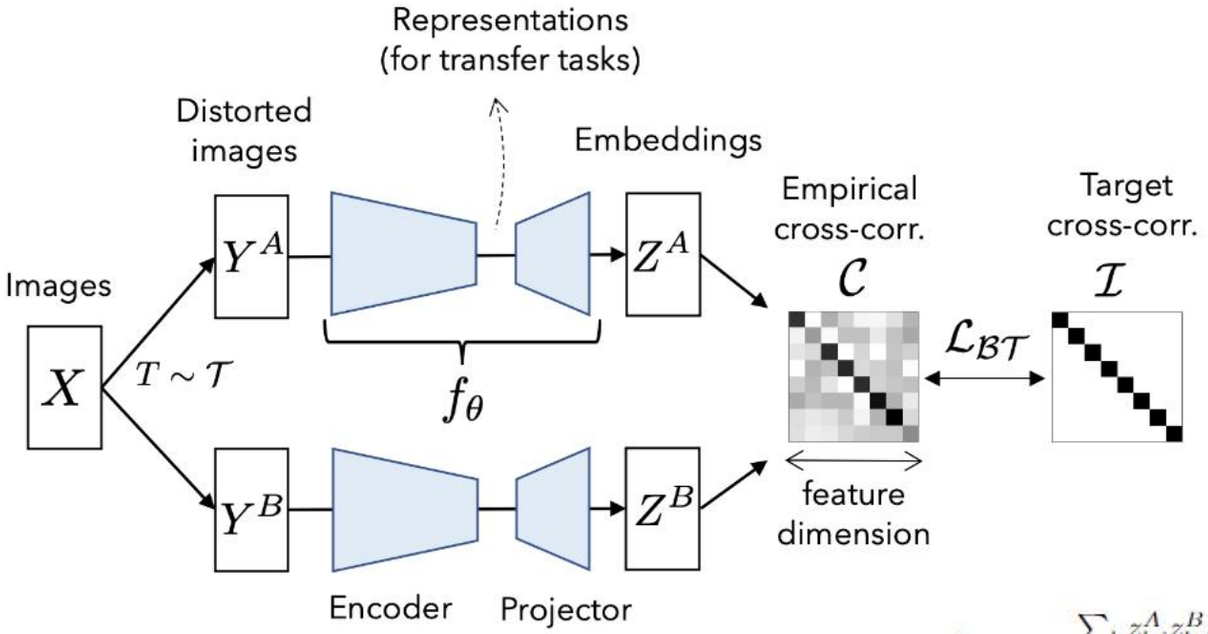
DINO: centering replaces the projection and predictor

Understanding Self-Supervised Learning Dynamics without Contrastive Pairs

01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

(3) feature decorrelation methods

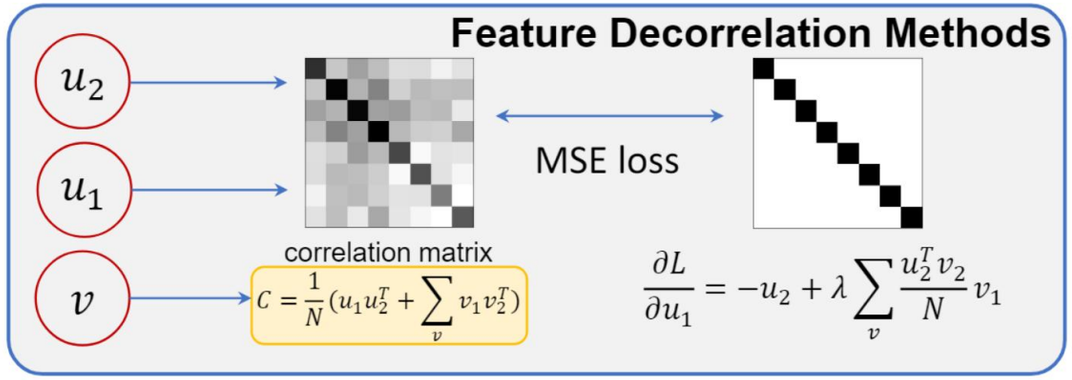
- reduce the redundancy between differnet feature dimensions



Barlow Twins

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2} \sqrt{\sum_b (z^B_{b,j})^2}}$$

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}}$$



Algorithm 1 PyTorch-style pseudocode for Barlow Twins.

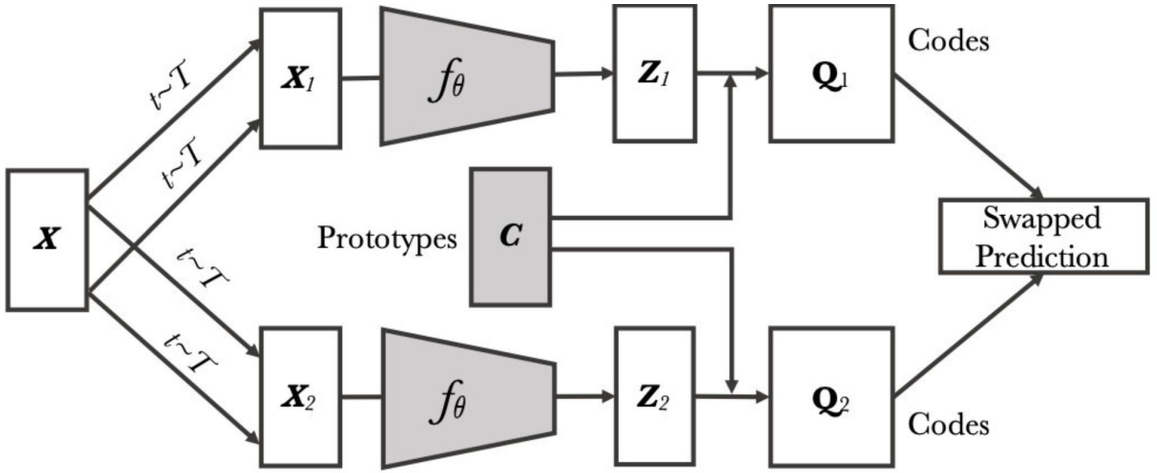
```
# f: encoder network
# lambda: weight on the off-diagonal terms
# N: batch size
# D: dimensionality of the embeddings
#
# mm: matrix-matrix multiplication
# off_diagonal: off-diagonal elements of a matrix
# eye: identity matrix

for x in loader: # load a batch with N samples
    # two randomly augmented versions of x
    y_a, y_b = augment(x)

    # compute embeddings
    z_a = f(y_a) # NxD
    z_b = f(y_b) # NxD

    # normalize repr. along the batch dimension
    z_a_norm = (z_a - z_a.mean(0)) / z_a.std(0) # NxD
    z_b_norm = (z_b - z_b.mean(0)) / z_b.std(0) # NxD

    # cross-correlation matrix
    c = mm(z_a_norm.T, z_b_norm) / N # DxD

    # loss
    c_diff = (c - eye(D)).pow(2) # DxD
    # multiply off-diagonal elems of c_diff by lambda
    off_diagonal(c_diff).mul_(lambda)
    loss = c_diff.sum()

    # optimization step
    loss.backward()
    optimizer.step()
```

01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Clustering-based scheme

- clustering: K-means / optimal transport
- but some restriction:
  - asynchronous two-satge training （DeepClustering）
  - no gradient propagate through group features
  - local equipartition constrainty to aviod degeneracy （SwAV）
  - not show an advantage in performance

SwAV: Swapping Assignments between multiple Views



code of view A → code of view B

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

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

instance-level discriminarion：

- supervisory signal:every sample as a category
- False negatives / Sampling Bias
- limited resource (large batch/ memory bank)

Group-level discrimination：

- supervisory signal hysteresis
- local equipartition decrease the validity of grouping
- not show an advantage in performance

01 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

This paper: SMoG

Novel:

- integrate the instance contrastive and clustering based methods
- contrastive group : allow gradient propagating through groups
- synchronous: instance grouping and representation learning

Achieve:

- avoid the false negatives and wipes off limiting factors of clustering methods
- surpasses the vanilla supervised-level performance for the first time
- multiple downstream tasks , CNN and Transformer backbones

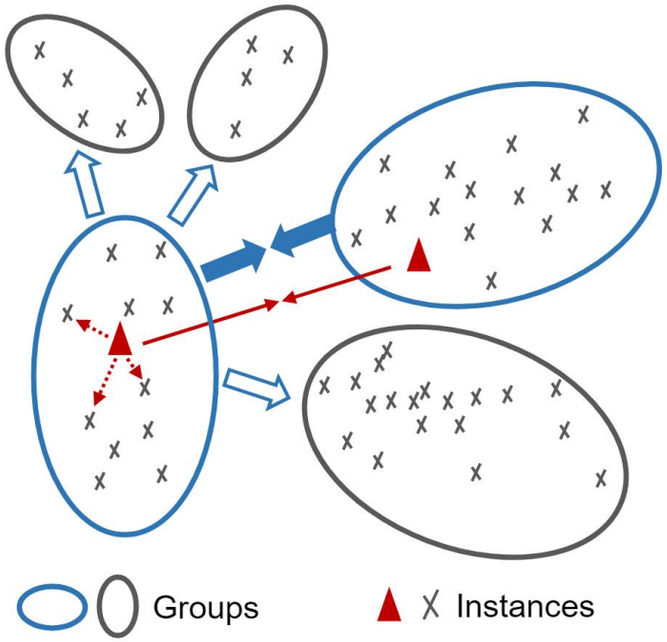


Fig. 1: Group contrastive *vs.* instance contrastive learning. Compared with instance contrasting (red parts), group contrasting (blue part) learns representations through higher-level semantics, which can reduce the chance that the already learned similar instances are still treated as negative pairs (false negatives). The significant reduction of contrast elements after the grouping process also makes global contrasting easier to calculate. The colorized instances or groups denote the positive pairs and the grey ones are negatives.

gap : 3% -> surpass surpervise

02 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

instance contrastive learning :

$$L_i = -\log \frac{\exp(\mathrm{sim}(f_\theta(\mathbf{x}_i^a), \hat{f}_\eta(\mathbf{x}_i^b)/\tau)}{\sum_{\mathbf{x}_j \neq \mathbf{x}_i^a, \mathbf{x}_j \in \mathbf{Y} \subseteq \mathbf{X}} \exp(\mathrm{sim}(f_\theta(\mathbf{x}_i^a), \hat{f}_\eta(\mathbf{x}_j))/\tau)},$$

(1) define group assigning function:

- generate instance group c_i
- update the groups {g}

$$\mathbf{c}_i^a, \{\mathbf{g}\} \leftarrow \phi(f_\theta(\mathbf{x}_i^a)|\{\mathbf{g}\})$$

How to design this function?

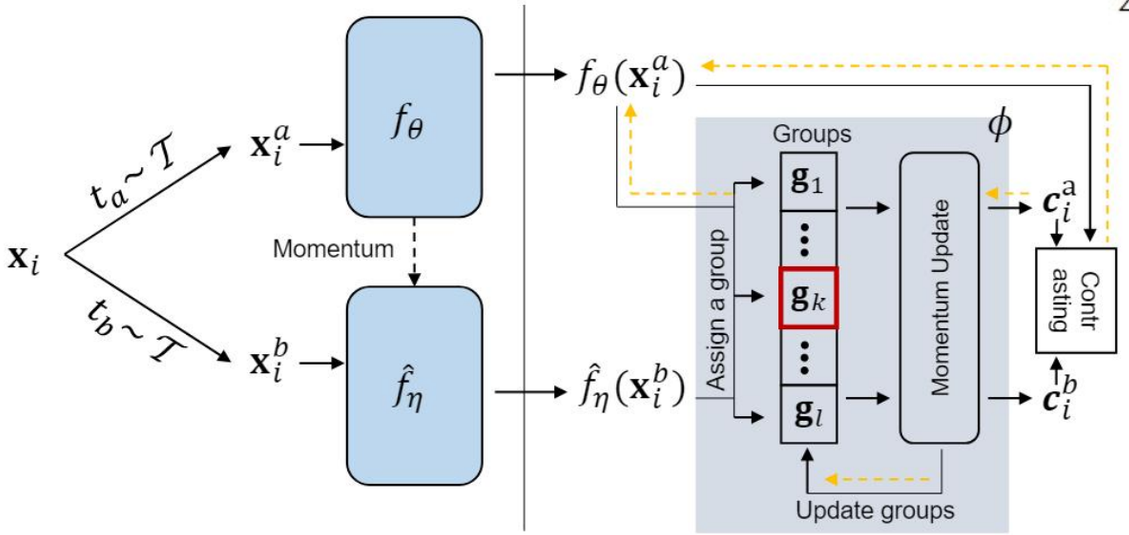
$$L_i = -\log \frac{\exp(\mathrm{sim}(\mathbf{c}_i^a, \mathbf{c}_i^b)/\tau)}{\sum_{\mathbf{g}_j \in \mathbf{G}} \exp(\mathrm{sim}(\mathbf{c}_i^a, \mathbf{g}_j)/\tau)},$$

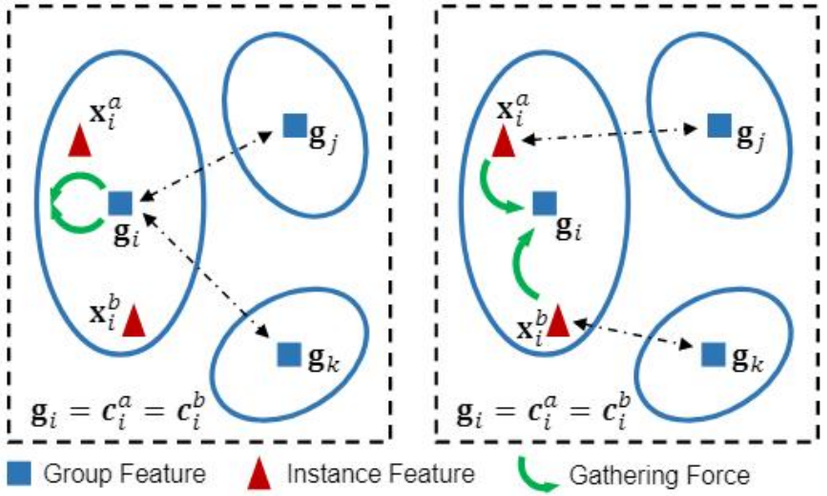
(1) c_i can replace the lastest $f_\theta(x_i)$ to participate in contrasting.

(2) group feature c_i should update synchronously with instance feature $f_\theta(x_i)$

c_i^a and c_i^b tend to be same group feature -> fail to gather similar instances and get them closer

$$L_i = -\log \frac{\exp(\mathrm{sim}(f_\theta(\mathbf{x}_i^a), \mathbf{c}_i^b)/\tau)}{\sum_{\mathbf{g}_j \in \mathbf{G}} \exp(\mathrm{sim}(f_\theta(\mathbf{x}_i^a), \mathbf{g}_j)/\tau)}.$$





a Network to get representation **b** Our SMoG

02 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

group assigning function:

$$\mathbf{c}_i^a, \{\mathbf{g}\} \leftarrow \phi(f_\theta(\mathbf{x}_i^a)|\{\mathbf{g}\})$$

global clustering each iteration: computing cost -> momentum grouping scheme

momentum grouping scheme:

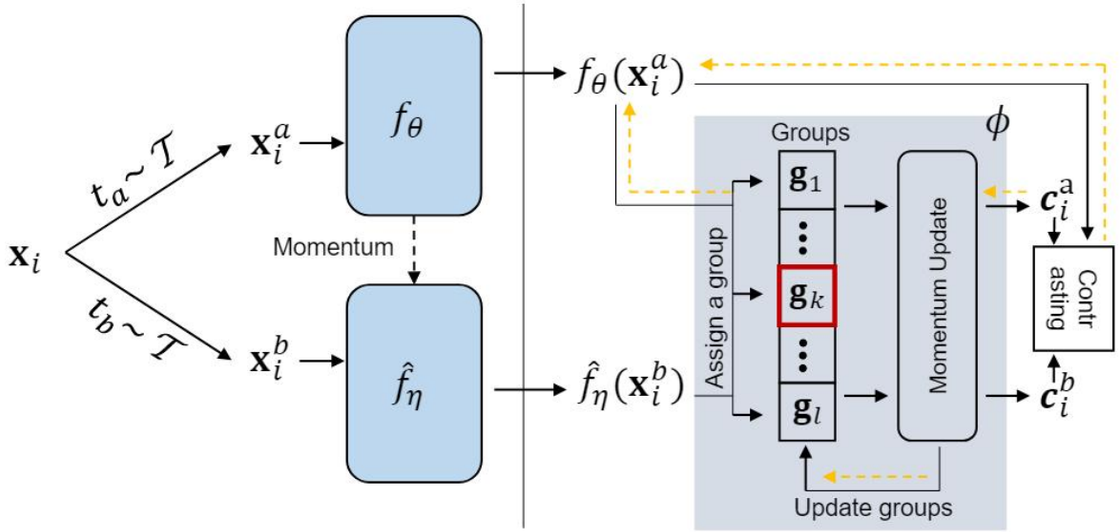
$$\mathbf{c}_i = \text{argmin}_{\mathbf{g}_k}(\text{sim}(f_\theta(\mathbf{x}_i), \mathbf{g}_k))$$

$$\mathbf{g}_k \leftarrow \beta * \mathbf{g}_k + (1-\beta) * \text{mean}_{\mathbf{c}_t=\mathbf{g}_k} f_\theta(\mathbf{x}_t),$$

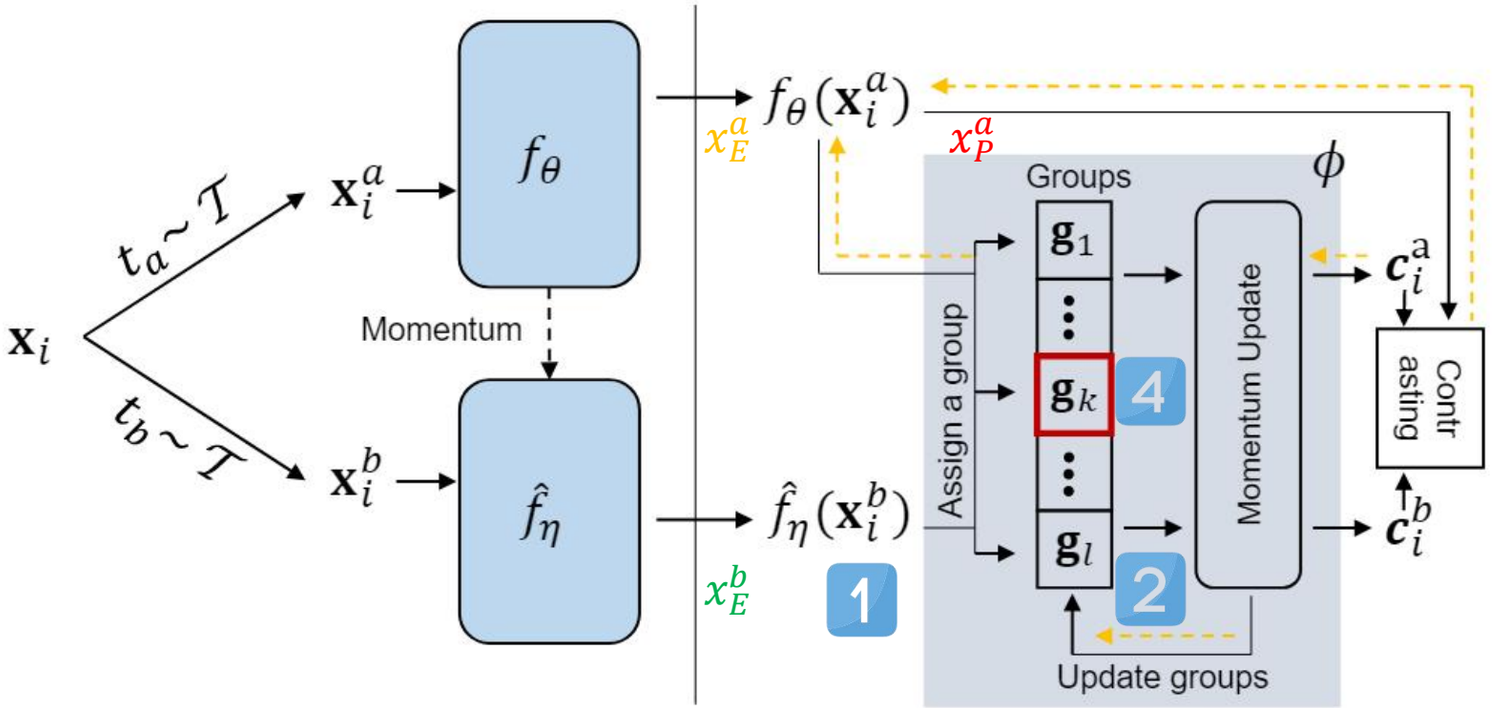
collapse:

- the scale of the groups are unbalanced
- the instance collapse into few groups

-> periodically grouping on the cached large feature set to relocate the groups



02 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping



a Network to get representation **b** Our SMoG

pervious:

- detach group features
- not contrast groups but classify into groups

grouping contrasting:

- contrasting among groups instead of classifying

(1) momentum encoder → label:

c_i^b =argmin(sim(x_E^b, g_k))

(2) x_E^a update the groups → new group features:

$$\mathbf{c}_i = \text{argmin}_{\mathbf{g}_k}(\text{sim}(f_\theta(\mathbf{x}_i), \mathbf{g}_k))$$

$$\mathbf{g}_k \leftarrow \beta * \mathbf{g}_k + (1-\beta) * \text{mean}_{\mathbf{c}_t=\mathbf{g}_k} f_\theta(\mathbf{x}_t),$$

(3) x_P^a calculate the loss:

$$L_i = -\log \frac{\exp(\text{sim}(f_\theta(\mathbf{x}_i^a), \mathbf{c}_i^b)/\tau)}{\sum_{\mathbf{g}_j \in \mathbf{G}} \exp(\text{sim}(f_\theta(\mathbf{x}_i^a), \mathbf{g}_j)/\tau)}.$$

(x_P^a, new group features) ↔(x_E^b, old group feature)

(4) replace group features:

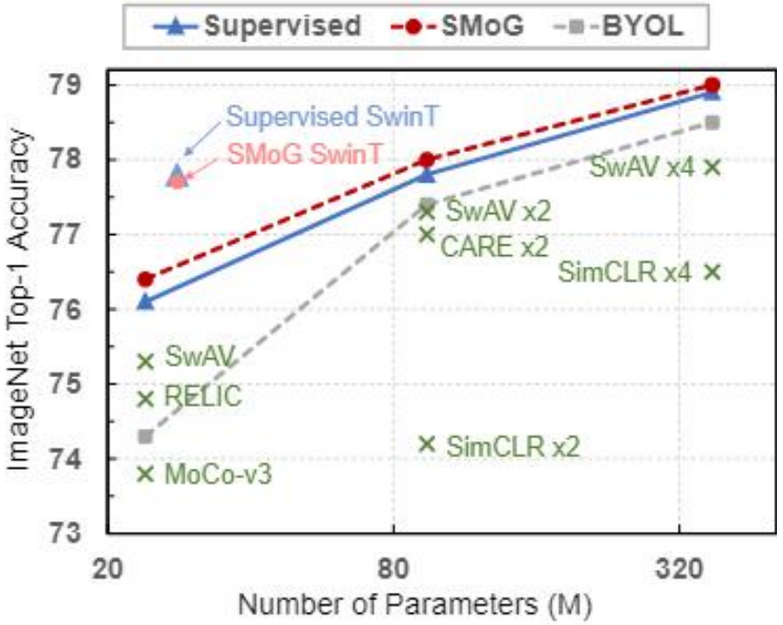
03 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Table 1: Linear protocol results on ImageNet. ResNet50 is adopted. † denotes the model adopts the multi-crop training strategy. "acc" means accuracy.

model	epoch	batchsize	top1 acc	top5 acc
supervised	100ep	256	76.1	92.7
instance contrastive method				
ReSSL [74]	200	256	69.6	-
MoCoV2 [23]	800	256	71.1	90.1
SimSiam [9]	800	256	71.3	-
InfoMin Aug. [60]	800	256	73.0	91.1
MoCoV3 [10]	400	4096	73.1	-
MoCoV3	800	4096	73.8	-
BYOL [21]	400	4096	73.2	-
BYOL	800	4096	74.3	91.6
Barlow Twins [70]	1000	2048	73.2	91.0
RELIC [47]	1000	4096	74.8	92.2
SSL-HSIC [37]	1000	4096	74.8	92.2
group-based method				
DeepCluster	400	256	52.2	-
ODC [71]	400	256	57.6	-
PCL [36]	200	256	67.6	-
DeepClusterV2	400	4096	70.2	-
SwAV	400	4096	70.1	-
CoKe [54]	800	1024	72.2	-
SMoG	400	2048	73.6	91.3
SMoG	800	4096	74.5	91.9
with multi-crop				
SwAV†	400	4096	74.6	-
SwAV†	800	4096	75.3	-
DC-v2†	800	4096	75.2	-
DINO † [7]	800	4096	75.3	-
UniGrad † [59]	800	4096	75.5	-
NNCLR † [15]	1000	4096	75.6	92.4
SMoG†	400	4096	**76.4**	**93.1**

Linear Evaluation:

model	backbone	throughput	param	top1 acc
supervised	SwinT	808	28	77.8 / 81.3
Supervised	DeiT-S/16	1007	21	77.5 / 79.8
MoBY	DeiT-S/16	1007	21	72.8
MoBY	SwinT	808	28	75.0
MoCoV3	DeiT-S/16	1007	21	72.5
MoCoV3	ViT-B/16	312	85	76.7
DINO	DeiT-S/16	1007	21	72.5
DINO †	DeiT-S/16	1007	21	77.0
EsViT [35]	SwinT	808	28	70.5
EsViT †	SwinT	808	28	77.0
SMoG	SwinT	808	28	74.5
SMoG †	SwinT	808	28	**77.7**



larger Resnet: not an increasing superiority over the supervised one

model	backbone	param	top1 acc	top5 acc
Supervised	Res50 (x2)	188	77.8	93.8
	Res50 (x4)	375	78.9	94.5
BYOL	Res50 (x2)	188	77.4	93.6
	Res50 (x4)	375	78.6	94.2
SwAV	Res50 (x2)	188	77.3	-
	Res50 (x4)	375	77.9	-
SMoG	Res50 (x2)	188	78.0	93.9
	Res50 (x4)	375	79.0	94.4

03 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Table 4: Semi-supervised results on ImageNet.

method	Top-1 Acc (%)			Top-5 Acc (%)		
	1%	10%	100%	1%	10%	100%
resnet-50 ×1						
Supervised	25.4	56.4	76.1	48.4	80.4	92.9
SimCLR	48.3	65.6	76.0	75.5	87.8	93.1
BYOL	53.2	68.8	77.7	78.4	89.0	93.9
SwAV	53.9	70.2	-	78.5	89.9	-
Barlow Twins	55.0	69.7	-	79.2	89.3	-
SSL-HSIC	52.1	67.9	77.2	77.7	88.6	93.6
NNCLR	56.4	69.8	-	80.7	89.3	-
SMoG	**58.0**	**71.2**	**78.3**	**81.6**	**90.5**	**94.2**
resnet-50 ×2						
Supervised	-	-	77.8	-	-	93.8
SimCLR	58.5	71.7	-	83.0	91.2	-
BYOL	62.2	73.5	-	84.1	91.7	-
SMoG	**63.6**	**74.4**	**80.2**	**85.6**	**92.4**	**95.2**

Table 5: Transfer learning results on semantic segmentation task. We fine-tune the representations on VOC2012 and Cityscapes dataset. The segmentation model is FCN with ResNet-50.

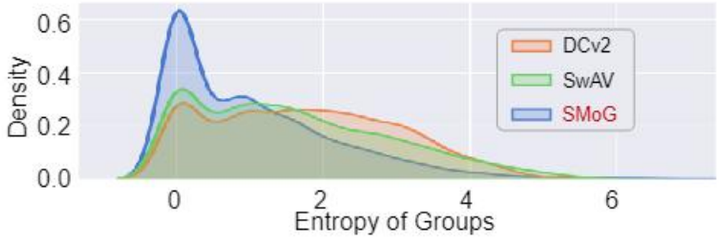
model	Cityscapes		VOC-2012	
	mIoU	mAcc	mIoU	mAcc
Supervised	73.83	82.56	73.59	83.74
MoCoV2	74.30	83.37	70.86	80.37
SwAV	74.80	83.01	74.97	84.27
BYOL	74.90	83.73	74.76	84.37
SMoG	**76.03**	**83.97**	**76.22**	**85.01**

Table 6: Transfer learning results on object detection and instance segmentation tasks. we adopt COCO as the fine-tuning dataset. Mask RCNN with ResNet-50-FPN is the detection and segmentation model. We report the AP metrics.

method	COCO det			COCO instance seg.		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Rand Init	31.0	49.5	33.2	28.5	46.8	30.4
Supervised	38.9	59.6	42.7	35.4	56.5	38.1
InsDis [67]	37.4	57.6	40.6	34.1	54.6	36.4
PIRL [45]	38.5	57.6	41.2	34.0	54.6	36.2
MoCoV2	39.4	59.9	43.0	35.8	56.9	38.4
SwAV	38.5	60.4	41.4	35.4	57.0	37.7
DC-v2 [6]	38.3	60.3	41.3	35.4	56.7	38.0
BYOL	39.4	59.9	43.0	35.8	56.8	38.5
Barlow Twins	39.2	58.7	42.6	34.3	55.4	36.5
SMoG	**40.1**	**61.6**	**43.7**	**36.9**	**58.7**	**39.3**

04 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Grouping quality：



each group, its entropy is $\sum_i -p_i * \log(p_i)$ where p_i is the ratio of instances that belong to class i (data annotation) in this group. A lower entropy means a group has more unitary semantics and is more meaningful. Thus, more groups with low

Fig. 5: Entropy of each groups. SMoG produces groups with much lower entropy which represents a better grouping quality in terms of high-level semantics.

Table 7: Ablation study on SMoG. We adopt ResNet50 as the backbone and train the unsupervised algorithm on ImageNet for 100 epochs without multi-crop training strategy. We report the linear evaluation results (Top-1 accuracy).

(a) Momentum ratio β. Linearly decreasing schedule performs better.

Schedule	Top1 acc
fixed 0.99 β	65.9
1.0 $\beta \rightarrow$ 0.9 β	67.0
1.0 $\beta \rightarrow$0.99 β	67.2
1.0 $\beta \rightarrow$0.999 β	67.1

(b) Tricks dealing with collapse. The periodical clustering is necessary.

Tricks	Top1 acc
None	0.1
+ periodical clustering (pd)	53.7
+ reset $\hat{f}_\eta$ periodically	53.4
+ pd & reset $\hat{f}_\eta$ periodically	67.2

(c) Number of groups. 3k is enough for SMoG.

# Groups	Top1 acc
300	65.2
1000	67.0
3000	67.2
10000	67.2

(d) Grouping algo. Momentum update better balances grouping and learning.

method	Top1 acc
Randomly select	30.2
Adopt latest	42.1
Averaging update	65.8
Momentum update	67.2

Periodical clustering is important！

Grouping number is not sensitive

Lastest feature as grouping feature is not good choice.

(RS): randomly select a group of latest instance features

(AL): $\mathbf{g}_k \leftarrow \text{mean}_{\mathbf{c}_t=\mathbf{g}_k} f_\theta(\mathbf{x}_t)$

(AU): $\mathbf{g}_k \leftarrow \mathbf{g}_k + (1/n) * (\text{mean}_{\mathbf{c}_t=\mathbf{g}_k} f_\theta(\mathbf{x}_t) - \mathbf{g}_k)$