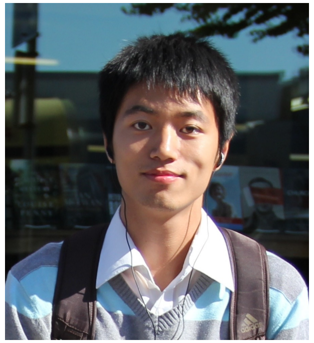
Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination

Xudong Wang
UC Berkeley / ICSI

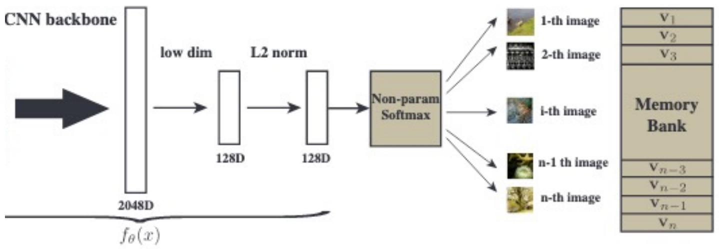
Ziwei Liu
NTU

Stella Yu
UC Berkeley / ICSI

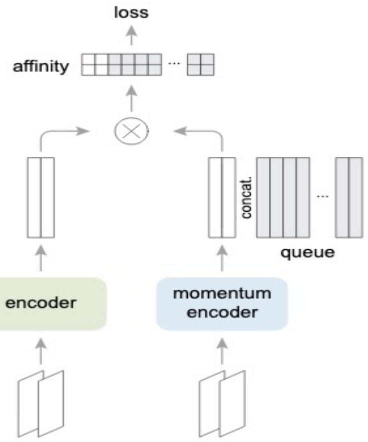


Previous Methods for Unsupervised Learning

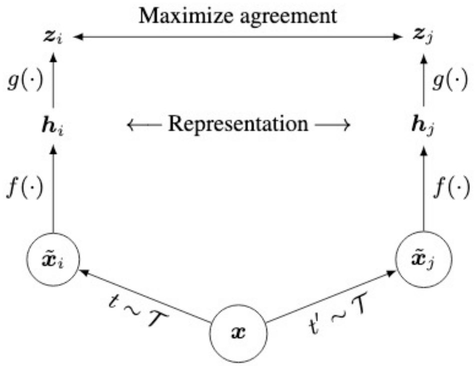
Instance Discrimination (Contrastive Learning)



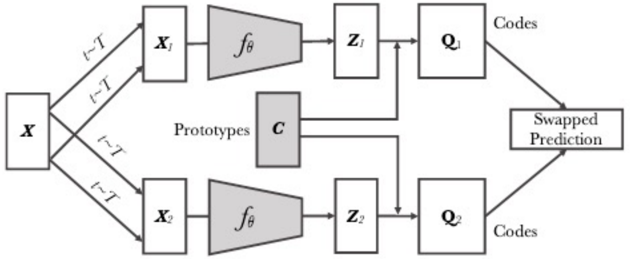
NPID [CVPR 2018]



MoCo [CVPR 2020]

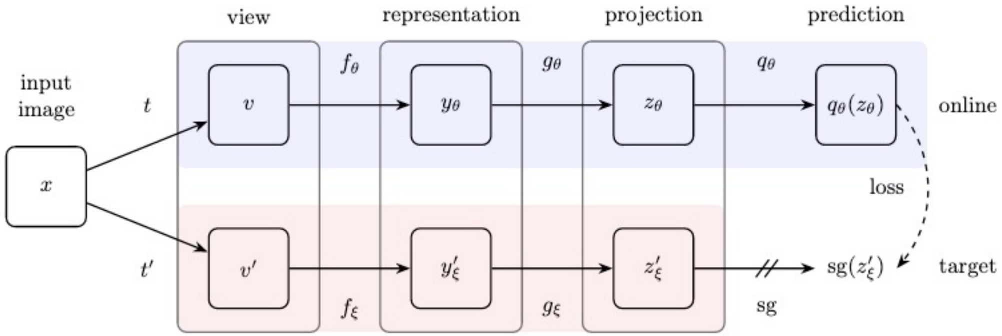


SimCLR [ICML 2020]

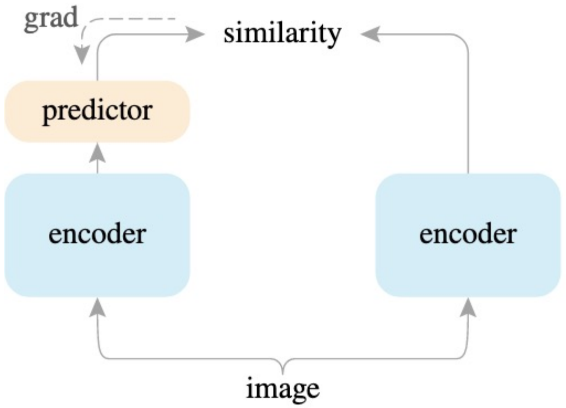


SwAV [NeurIPS 2020]

Instance Similarity (Positive Pairs Only)

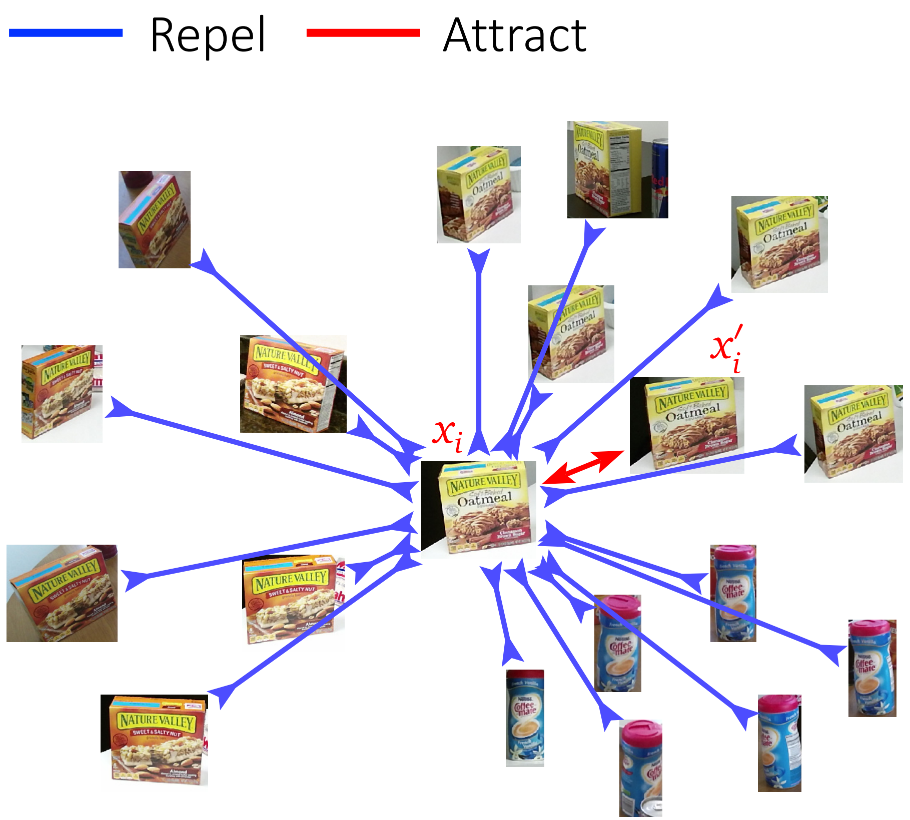


BYOL [NeurIPS 2020]



SimSiam [CVPR 2021]

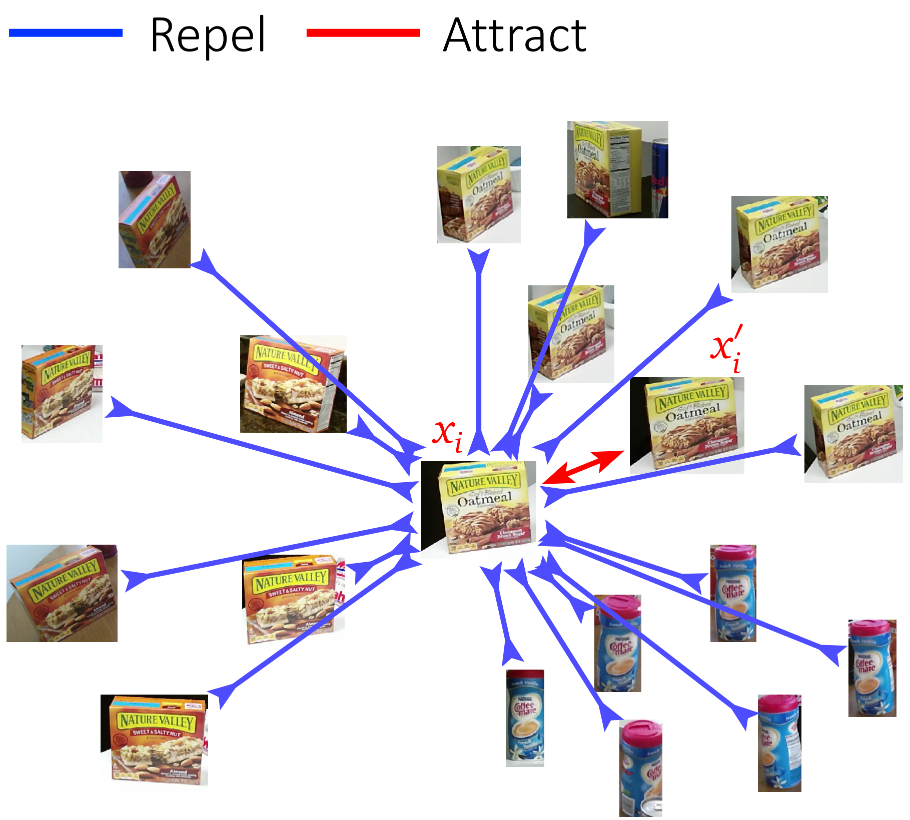
Caveats in Instance Discrimination



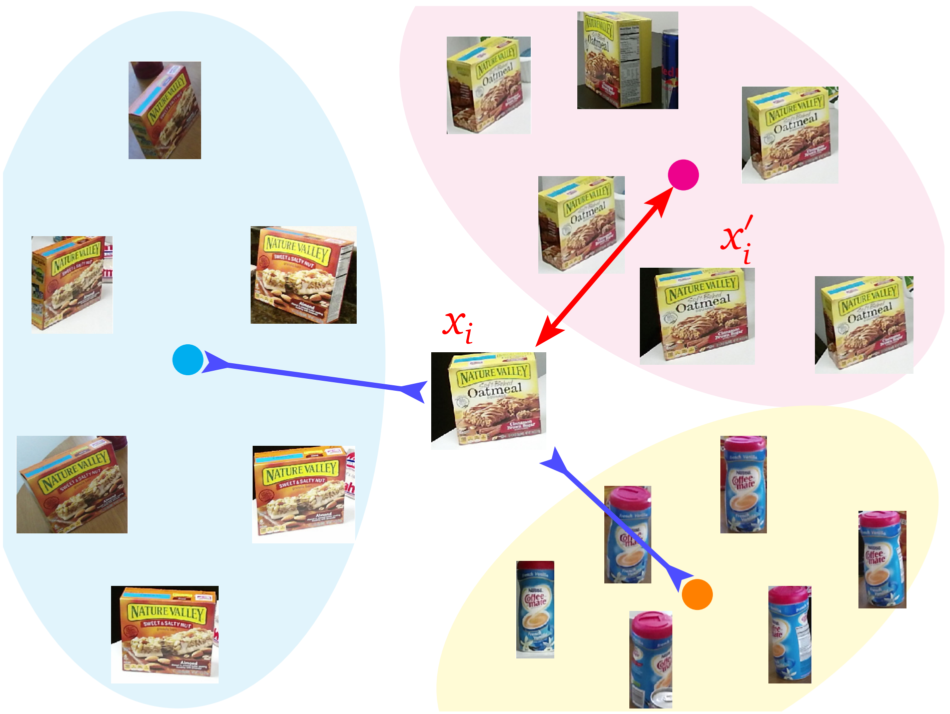
Instance Discrimination

- Ignores between-instance similarity
- Ignores natural groups which often underlie downstream tasks' discrimination at a coarser semantic level
- Repels all other instances including those highly similar ones
- Leans towards more instance discrimination than invariant mapping, reducing robustness

Instance Discrimination vs. Instance-Group Discrimination



Instance Discrimination

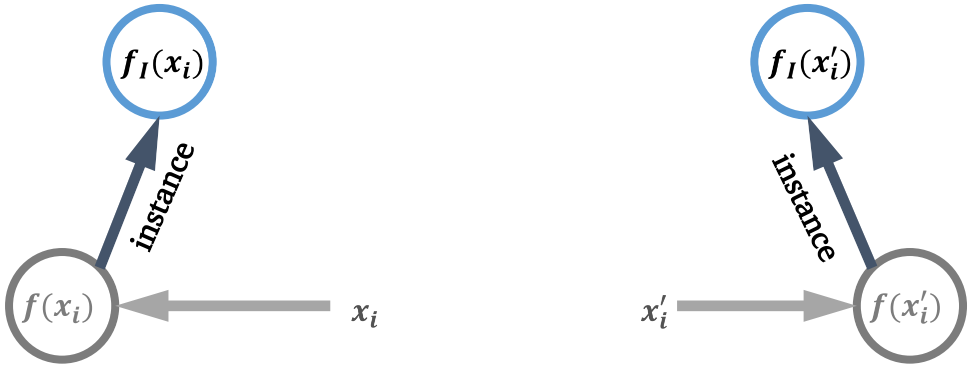


Instance-Group Discrimination

Cross-Level Instance-Group Discrimination (CLD)

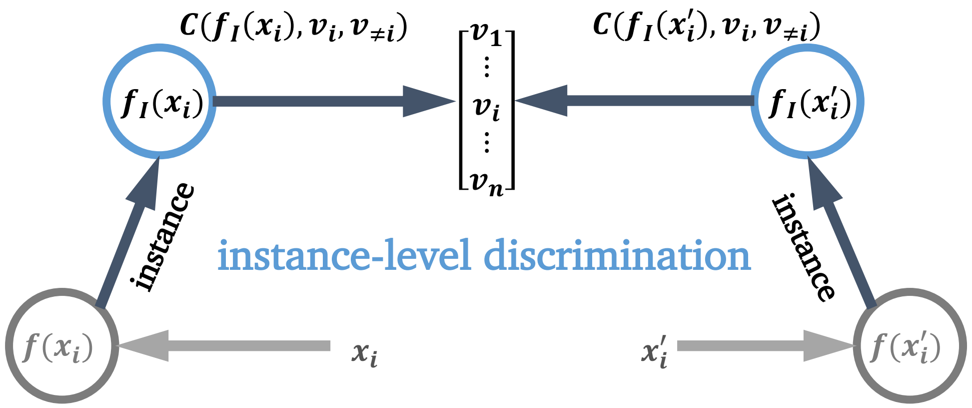
Two Augmented Views

Cross-Level Instance-Group Discrimination (CLD)



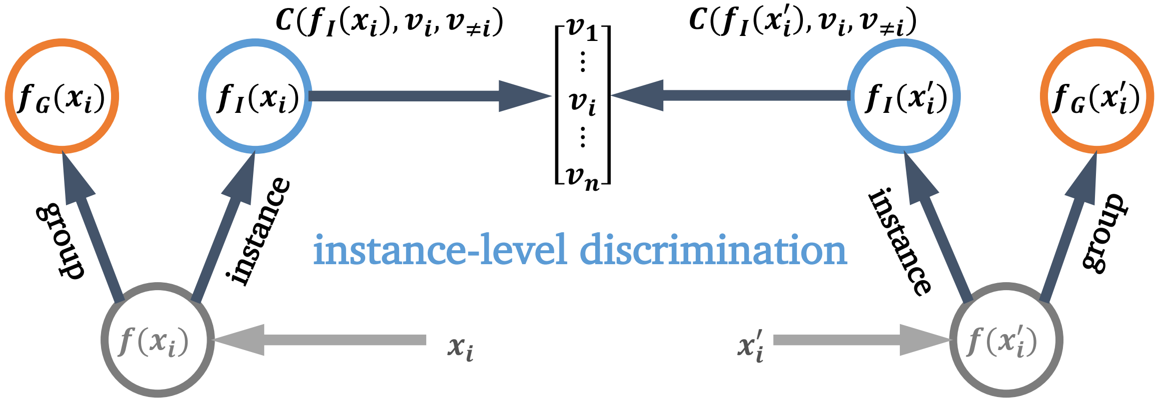
Feature Projection and Normalization

Cross-Level Instance-Group Discrimination (CLD)



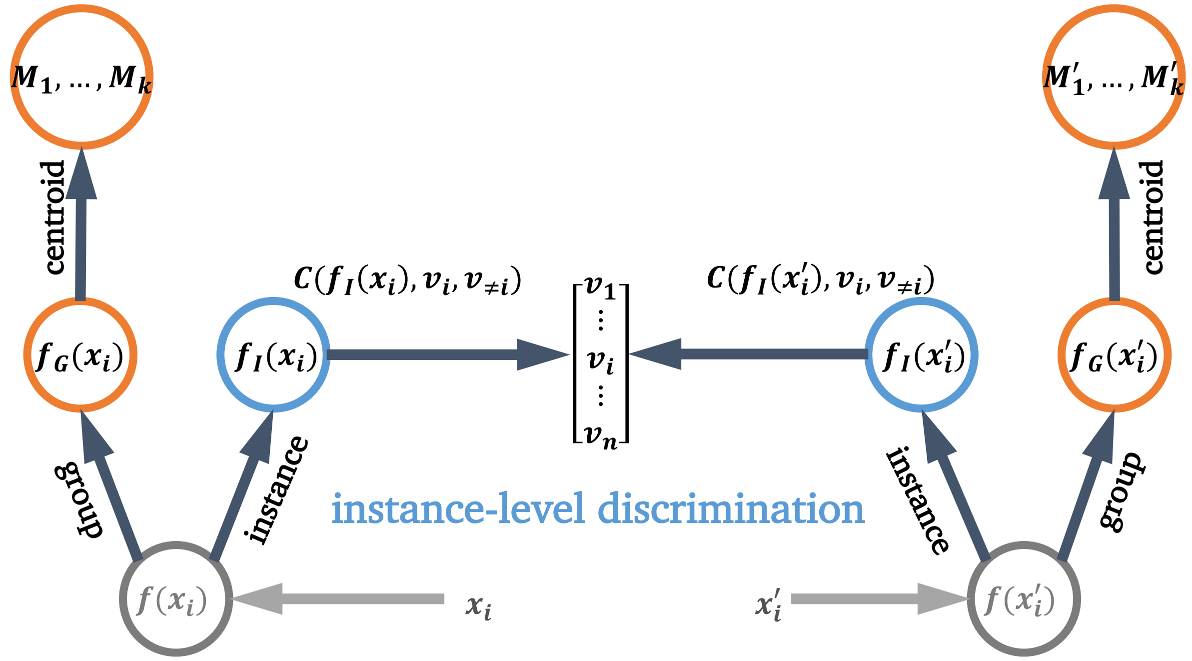
Instance Discrimination

Cross-Level Instance-Group Discrimination (CLD)



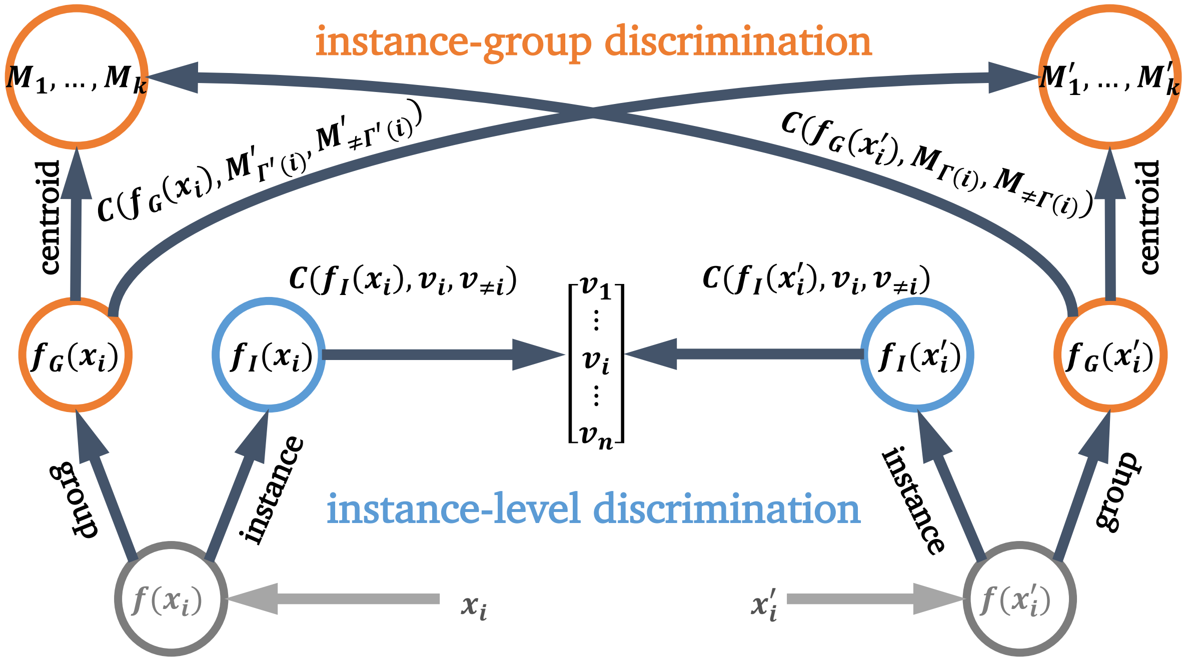
Group Branch
with Partial Shared Projection Head

Cross-Level Instance-Group Discrimination (CLD)



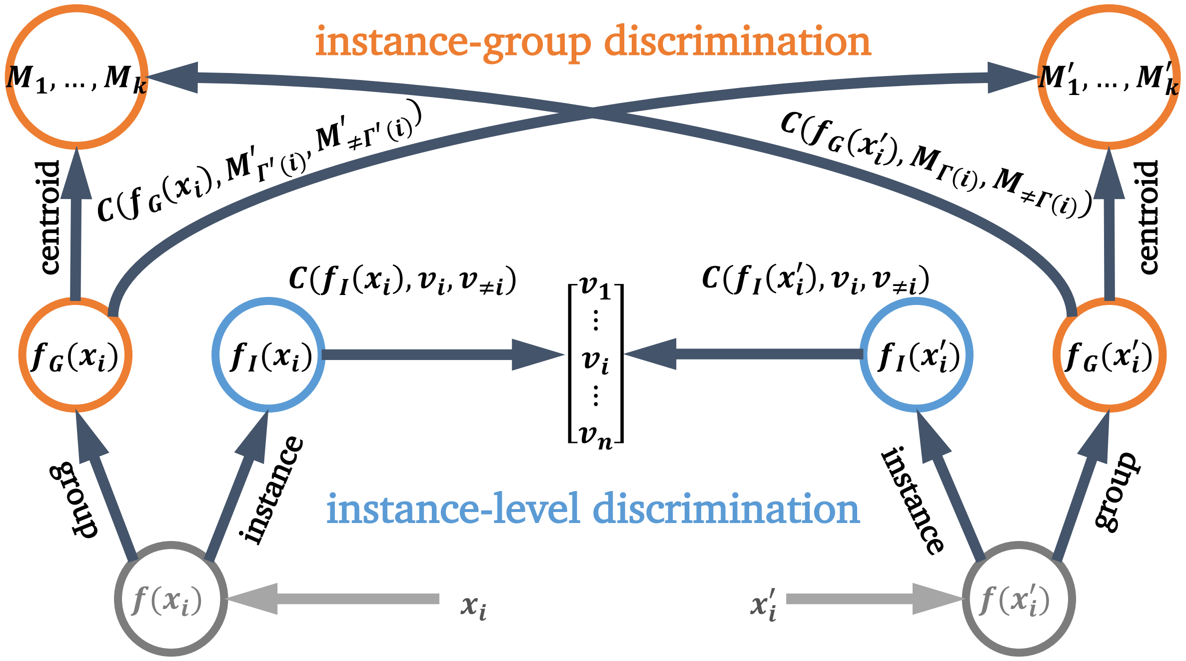
Local Clustering Centroids:
k-Means or Spectral Clustering

Cross-Level Instance-Group Discrimination (CLD)



Cross-view Instance-group Discrimination

CLD Objective



Consistent Cross-view Grouping

Minimizing the cross entropy between hard clustering assignment p_{ij} (as ground-truth) based on group branch feature $f_G(x_i)$ and soft assignment q_{ij} predicted from group branch feature $f_G(x_i')$ in a different view.

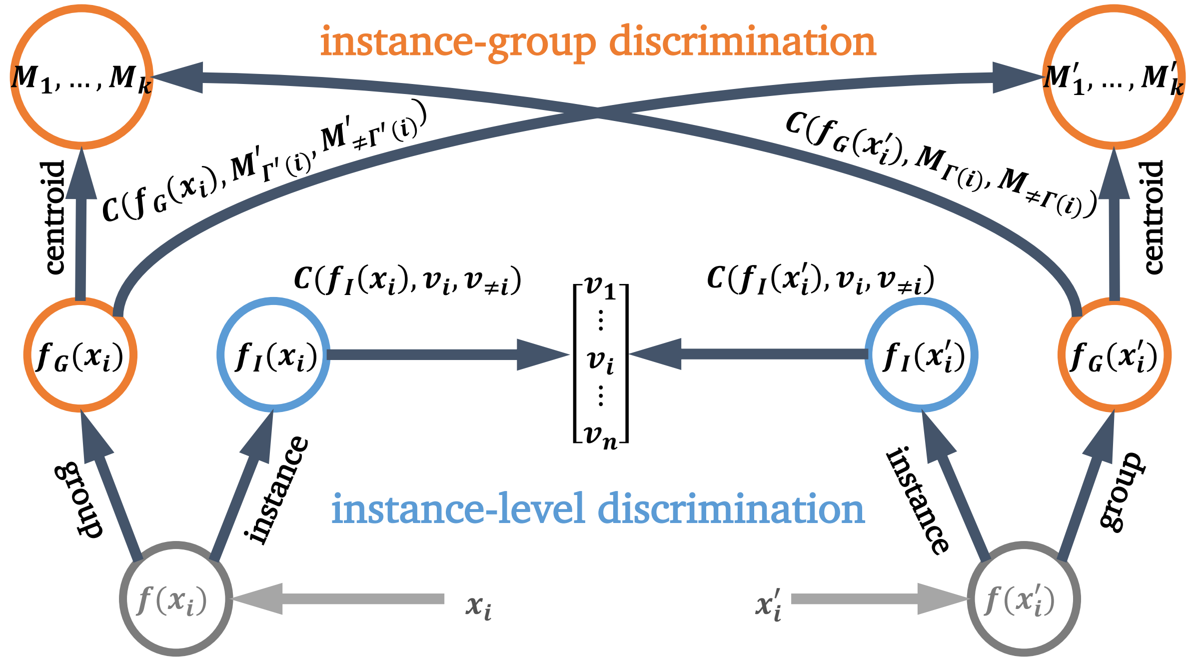
$$-E_p[\log q] = \sum_{i=1}^{n} C(f_G(x_i'), M_{\Gamma(i)}, M_{\neq\Gamma(i)}; T_G)$$

Total contrastive learning loss:

$$L(f;T_I,T_G,\lambda)=\sum_{i=1}^{n} \underbrace{C(f_I(x_i),v_i,v_{\neq i};T_I)+C(f_I(x_i'),v_i,v_{\neq i};T_I)}_{\text{instance-level discrimination}}$$

$$+\lambda\sum_{i=1}^{n} \underbrace{C(f_G(x_i'),M_{\Gamma(i)},M_{\neq\Gamma(i)};T_G)+C(f_G(x_i),M'_{\Gamma'(i)},M'_{\Gamma'(i)};T_G)}_{\text{cross-level discrimination}}$$

Cross-Level Instance-Group Discrimination (CLD)



- For instances x_i and x_j clustered in the same group:
  - *Instance feature $f_G(x_i)$ and $f_G(x_j)$ are attracted to the same group centroid M or M', and are thus drawn closer.*
- For similar instances x_i and x_j not in the same cluster:
  - *Repel common group centroids, thereby pulling instance features $f_G(x_i)$ and $f_G(x_j)$ closer*
- CLD discriminates at instance *and* group levels, more inline with coarser discrimination at downstream tasks.
- Greatly improves the positive/negative ratio for invariant mapping
  - *For example, the ratio on ImageNet is 1/65536 for MoCo's set-wise NCE vs. 1/255 for CLD's batch-wise NCE.*

Normalized Projection Head (NormLinear / NormMLP)

Existing methods:

project the feature to a unit hypersphere with L2 normalization.

Our methods:

Here, we normalize both the FC layer weights W and the shared feature vector f, so that projecting f on to W simply calculates their cosine similarity.

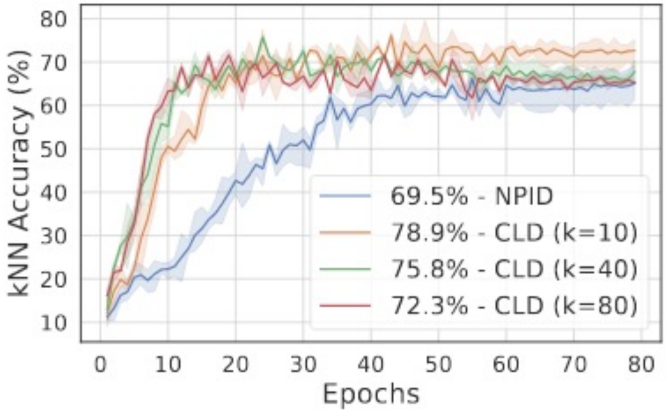
The final normalized d-dimensional feature $N(x_i)$ has t-th component:

$$N_t(x_i) =< \frac{W_t}{||W_t||}, \frac{f(x_i)}{||f(x_i)||} >$$

Simple yet effective with consistent performance gains!

High Correlation Datasets

- ✓ *More than 5-9% improvements with faster converging speed.*

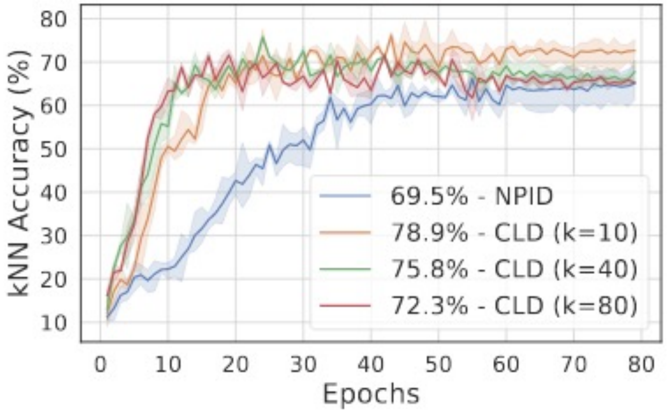


Kitchen-HC	kNN Accuracy
NPID	69.5
NPID + CLD	78.9 (+9.4)
MoCo	76.1
MoCo + CLD	**81.6** (+5.5)

kNN accuracies on Kitchen-HC

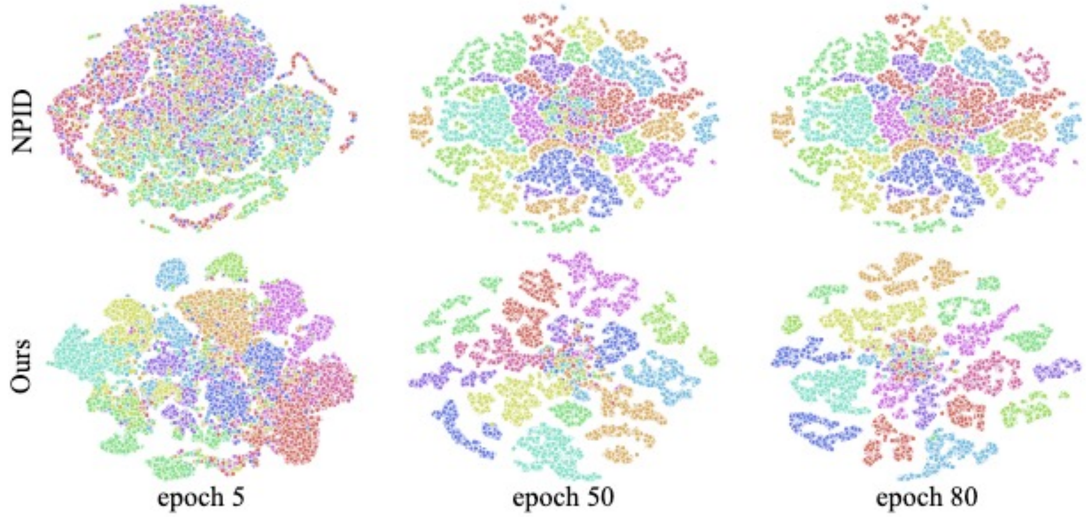
High Correlation Datasets

✓ *More than 5-9% improvements with faster converging speed.*



Kitchen-HC	kNN Accuracy
NPID	69.5
NPID + CLD	78.9 (+9.4)
MoCo	76.1
MoCo + CLD	**81.6** (+5.5)

kNN accuracies on Kitchen-HC



t-SNE visualization on different epochs

Having highly correlated instances breaks the instance discrimination presumption and causes slow or unstable training.

Long-tailed Datasets

- ✓ *6~11% improvements on CIFAR-LT*
- ✓ *3~5% improvements on ImageNet-LT*
- ✓ *Consistent improvements to MoCo and NPID*

	CIFAR10-LT		CIFAR100-LT		ImageNet-LT		
	top1	top5	top1	top5	many/med/few	top1	top5
Unsupervised							
NPID [53]	32.3	74.8	10.2	29.8	47.5/21.3/6.6	29.5	51.1
NPID + CLD	41.1	78.9	21.7	44.3	52.4/25.0/8.3	32.7	55.6
vs. baseline	**+8.8**	**+4.1**	**+11.5**	**+14.5**	**+4.9/+3.7/+1.7**	**+3.2**	**+4.5**
MoCo [24]	34.2	76.7	19.7	42.6	48.1/21.3/6.9	29.9	51.8
MoCo + CLD	**43.1**	**80.4**	**25.4**	**50.0**	**53.1/24.9/9.4**	**33.3**	**57.3**
vs. baseline	**+8.9**	**+3.7**	**+5.7**	**+7.4**	**+5.0/+3.6/+2.5**	**+3.4**	**+5.5**
Supervised							
CE	-	-	-	-	40.9/10.7/0.4	20.9	-
OLTR [37]	-	-	-	-	43.2/35.1/18.5	35.6	-

Consistent Performance Gains to Various Methods on ImageNet

ImageNet benchmark:

- Consistent improvements to various methods

Methods	Architecture	#epoch	#GPU	top-1
BYOL† [21]	R50-MLP (28M)	100	128	66.5
w/ CLD‡	R50-NormMLP (28M)	100	8	69.1
InfoMin [49]	R50-MLP (28M)	100	8	67.4
w/ CLD	R50-MLP (28M)	100	8	69.5
w/ CLD	R50-NormMLP (28M)	100	8	70.1
NPID [53]	R50-Linear (24M)	200	8	56.5
w/ CLD	R50-Linear (24M)	200	8	60.6
MoCo [24]	R50-Linear (24M)	200	8	60.6
w/ CLD	R50-Linear (24M)	200	8	63.4
w/ CLD	R50-NormLinear (24M)	200	8	63.8
MoCo v2 [7]	R50-MLP (28M)	200	8	67.5
w/ CLD	R50-MLP (28M)	200	8	69.2
w/ CLD	R50-NormMLP (28M)	200	8	70.0
InfoMin [49]	R50-MLP (28M)	200	8	70.1
w/ CLD	R50-MLP (28M)	200	8	70.6
w/ CLD	R50-NormMLP (28M)	200	8	**71.5**

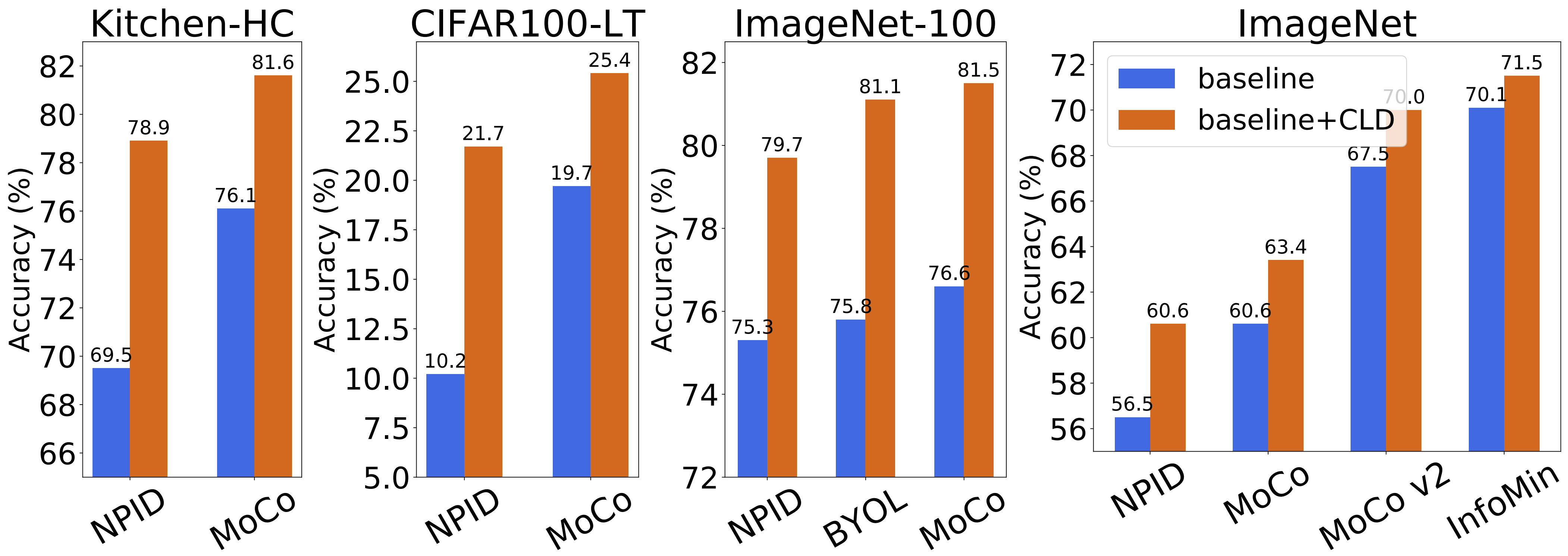
NormMLP is An Effective Alternative

ImageNet benchmark:

- Consistent improvements to various methods
- NormMLP is an effective alternative to vanilla MLP head.

Methods	Architecture	#epoch	#GPU	top-1
BYOL† [21]	R50-MLP (28M)	100	128	66.5
w/ CLD‡	R50-NormMLP (28M)	100	8	69.1
InfoMin [49]	R50-MLP (28M)	100	8	67.4
w/ CLD	R50-MLP (28M)	100	8	69.5
w/ CLD	R50-NormMLP (28M)	100	8	70.1
NPID [53]	R50-Linear (24M)	200	8	56.5
w/ CLD	R50-Linear (24M)	200	8	60.6
MoCo [24]	R50-Linear (24M)	200	8	60.6
w/ CLD	R50-Linear (24M)	200	8	63.4
w/ CLD	R50-NormLinear (24M)	200	8	63.8
MoCo v2 [7]	R50-MLP (28M)	200	8	67.5
w/ CLD	R50-MLP (28M)	200	8	69.2
w/ CLD	R50-NormMLP (28M)	200	8	70.0
InfoMin [49]	R50-MLP (28M)	200	8	70.1
w/ CLD	R50-MLP (28M)	200	8	70.6
w/ CLD	R50-NormMLP (28M)	200	8	**71.5**

Summary: Universal Add-on to Various Methods



↑ 5~9% ↑ 6~11% ↑ 4~5% ↑ 1.4~4.1%

Summary: SOTA Performance with a Much Smaller Compute!

CLD is the first method that achieves over 70% accuracy on ImageNet self-supervision benchmark, with affordable backbone (ResNet-50), batch size (=256), and training epochs (=100).

Methods	Architecture	#GPU	top-1 (#epoch=100)	top-1 (#epoch=200)
MoCo v2 [CVPR 2020]	R50-MLP (28M)	8	-	67.5
SimCLR [ICML 2020]	R50-MLP (28M)	128	66.5	68.3
SwAV [NeurIPS 2020]	R50-MLP (28M)	128	66.5	69.1
BYOL [NeurIPS 2020]	R50-MLP (28M)	128	66.5	70.6
SimSiam [Preprint]	R50-MLP (28M)	8	68.1	70.0
CLD	R50-NormMLP (28M)	8	**70.1**	**71.5**

Compare with state-of-the-arts under 100 and 200 training epochs

Summary: CLD Respects Semantics