Self-Supervised Learning for Classification and Beyond

Xie Enze

2nd year PhD student in University of Hong Kong

Supervisor: Prof. Ping Luo

SSL for **Classification**

- 1 Instance Discrimination (MoCo, SimCLR, BYOL)
- 2 Clustering and Classification (SwAV, DeepCluster)
- 3 Others (Solving Jigsaw, Rotation, ReSort, Location Prediction, GAN)

MoCo

Momentum Contrast for Unsupervised Visual Representation Learning

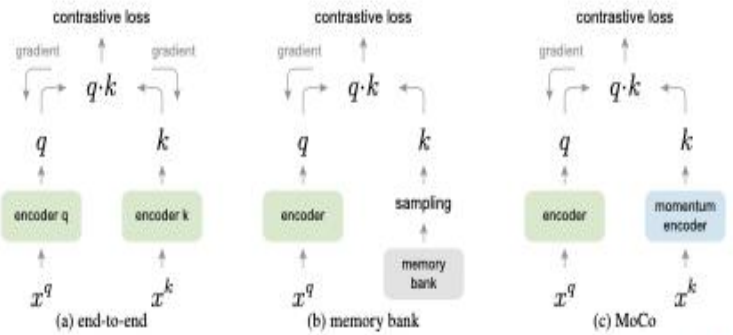
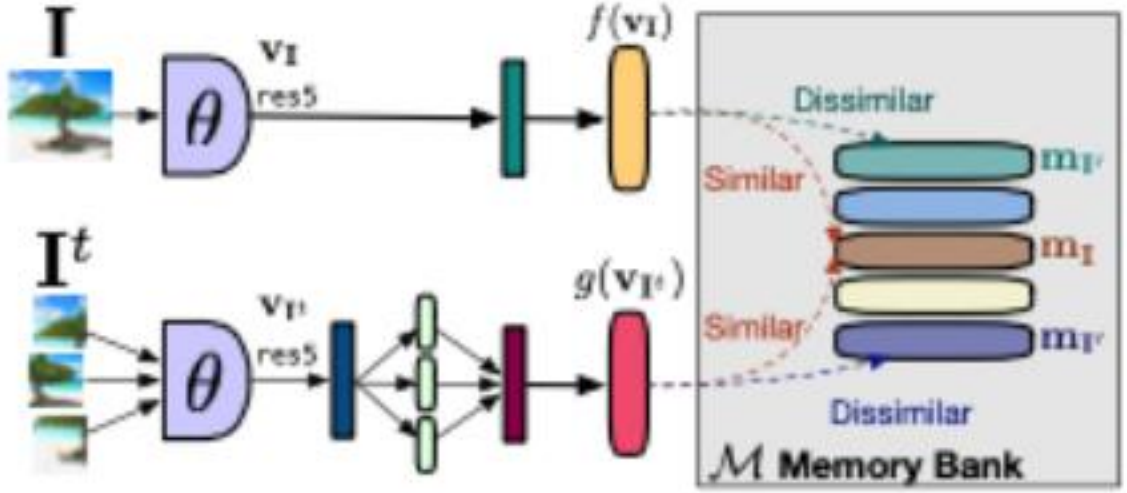


Figure 2. **Conceptual comparison of three contrastive loss mechanisms** (empirical comparisons are in Figure 3 and Table 3). Here we illustrate one pair of query and key. The three mechanisms differ in how the keys are maintained and how the key encoder is updated. **(a)**: The encoders for computing the query and key representations are updated ***end-to-end*** by back-propagation (the two encoders can be different). **(b)**: The key representations are sampled from a ***memory bank*** [61]. **(c)**: ***MoCo*** encodes the new keys on-the-fly by a momentum-updated encoder, and maintains a queue (not illustrated in this figure) of keys.

PIRL

Self–Supervised Learning of Pretext–Invariant Representations



BYOL

Bootstrap Your Own Latent
A New Approach to Self–Supervised Learning

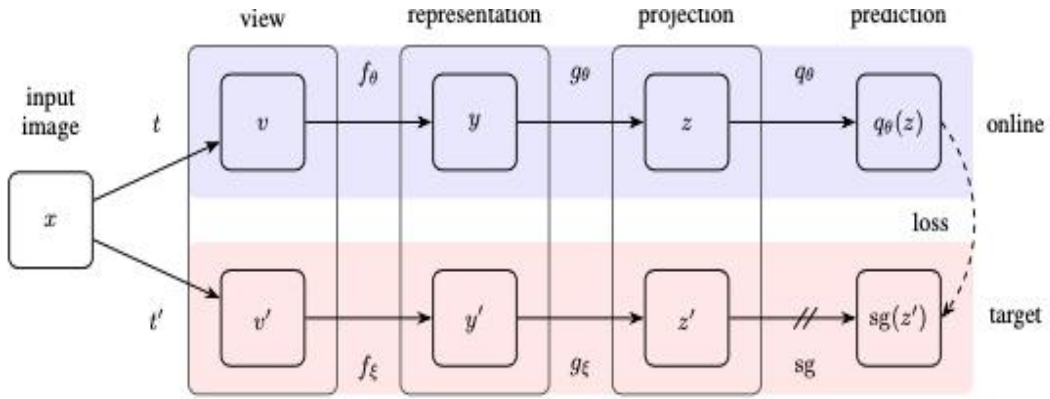


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z)$ and $\mathrm{sg}(z')$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded and y is used as the image representation.

SwAV

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

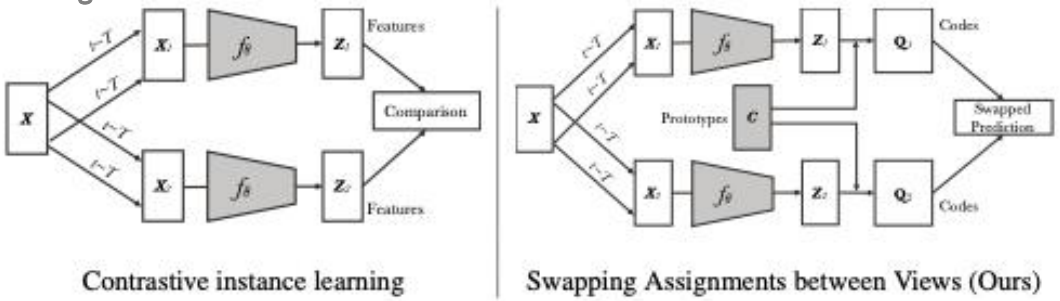


Figure 1: **Contrastive instance learning (left) *vs.* SwAV (right).** In contrastive learning methods applied to instance classification, the features from different transformations of the same images are compared directly to each other. In SwAV, we first obtain "codes" by assigning features to prototype vectors. We then solve a "swapped" prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Thus, SwAV does not directly compare image features. Prototype vectors are learned along with the ConvNet parameters by backpropragation.

SSL for **Detection**

Recently, Contrastive Learning and Clustering methods achieves great success on unsupervised representation learning, especially on ImageNet classification

However, it is interesting that their performance misalign on classification and detection…

Because classification and detection are different tasks, need different pretext task design

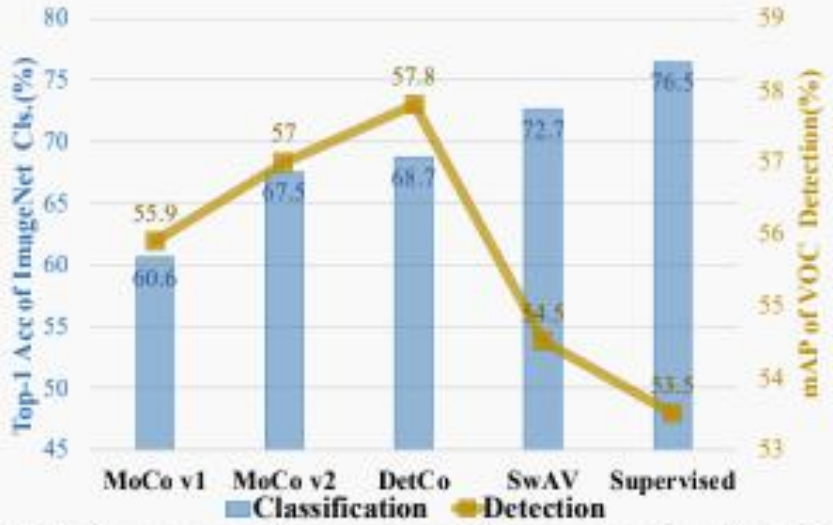


Figure 2. Performance of several self-supervised methods transferring to downstream tasks, ImageNet classification and PASCAL VOC detection. It shows the accuracy of **classification and detection are inconsistent and have low correlation**.

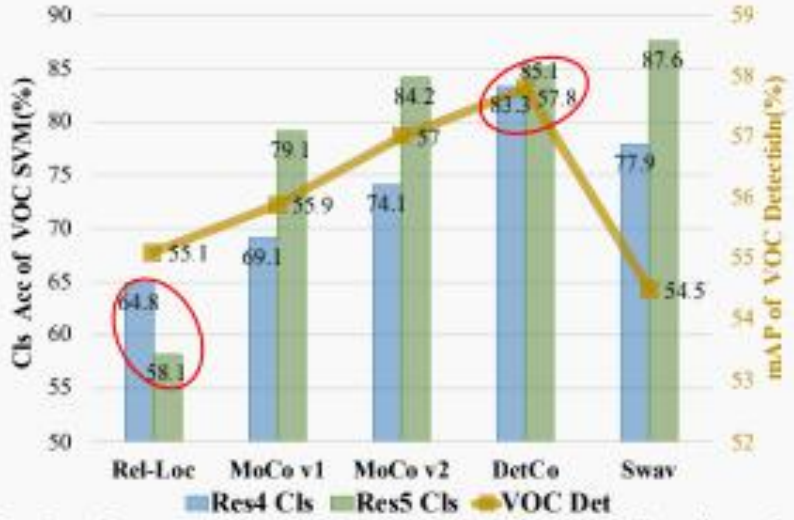
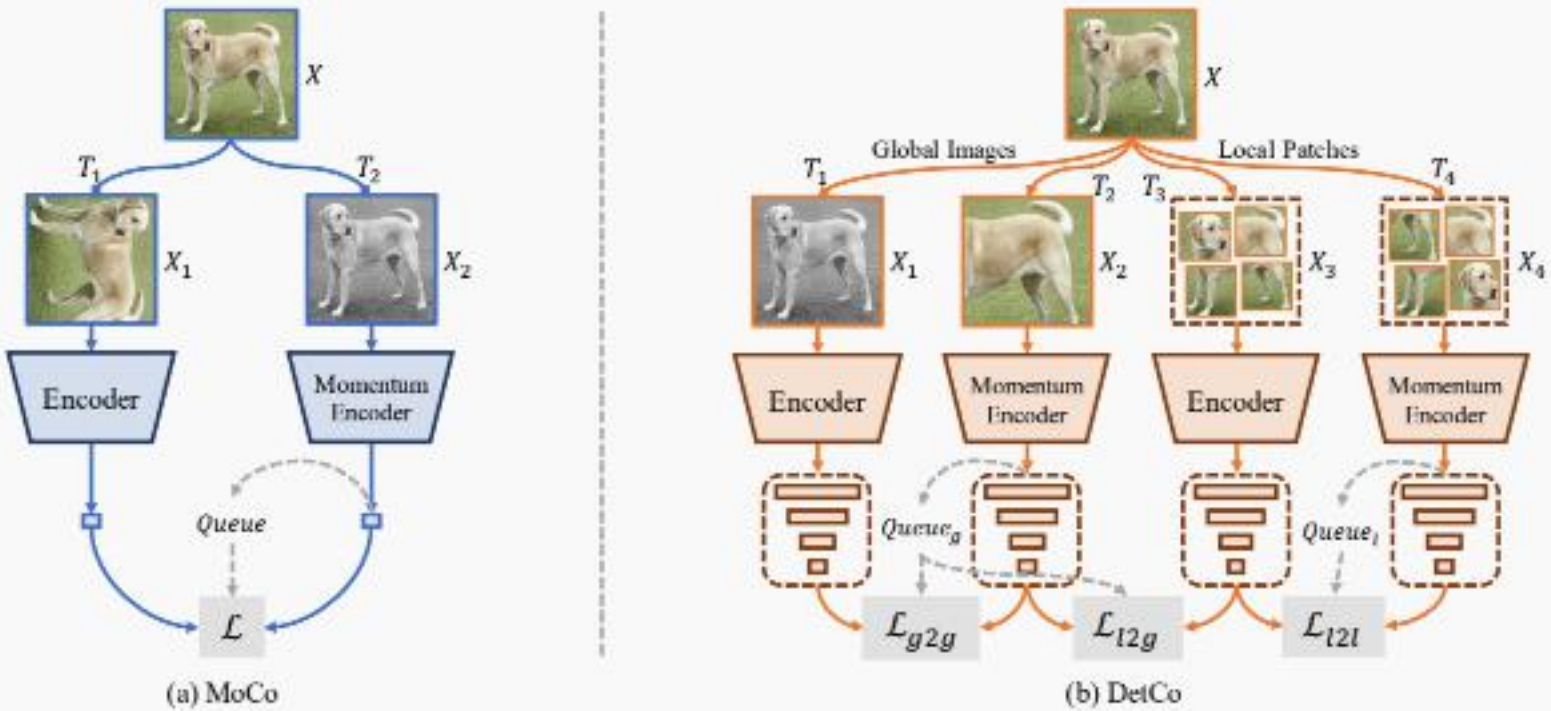


Figure 3. Performance of VOC SVM classification in `Res4`, `Res5` and detection. Although Relative-Loc is a non-contrastive method, it **keeps shallow layer feature discriminative and predicts position between local patches**, enabling competitive detection results.

DetCo：Contrastive Learning for Object Detection



(a) MoCo (b) DetCo

Figure 4. **The overall pipeline of DetCo compared with MoCo [18].** (a) is MoCo's framework, which only considers the high-level feature and learning contrast from a global perspective. (b) is our DetCo, which straightforwardly appends hierarchical intermediate contrast and two additional local patch views for input, building contrastive loss cross the global and local representation. Our DetCo improves the detection transferring ability by following the proposed three good practices. Note that "T" means image transforms and "L_{l2g}" means contrastive loss cross local and global features. "$Queue_{g/l}$" means different memory banks [38] for global/local features.

Improving Lower Bound of Mutual Information with Patch Representation

$$\mathbf{MI} \geqslant \log(K) - \mathcal{L}_{NCE} \triangleq \texttt{Lower Bound (LB)},$$

$$\texttt{LB}^{g\leftrightarrow l} - \texttt{LB}^{g\leftrightarrow g} = \mathcal{L}_{NCE}^{g\leftrightarrow g}(\mathbf{I}_1, \mathbf{I}_2) - \mathcal{L}_{NCE}^{g\leftrightarrow l}(\mathbf{P}_1, \mathbf{I}_2), \quad \text{(I)}$$

$$\mathcal{L}_{NCE}^{g\leftrightarrow g}(\mathbf{I}_q, \mathbf{I}_k) = -\log \frac{\texttt{Sim}_P^{g\leftrightarrow g}}{\texttt{Sim}_P^{g\leftrightarrow g} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g}}, \quad \text{(II)}$$

and

$$\mathcal{L}_{NCE}^{g\leftrightarrow l}(\mathbf{P}_q, \mathbf{I}_k) = -\log \frac{\texttt{Sim}_P^{g\leftrightarrow l}}{\texttt{Sim}_P^{g\leftrightarrow l} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l}}, \quad \text{(III)}$$

where $\texttt{Sim}_P$ means similarity between positive pairs and $\texttt{Sim}_N$ means similarity between negative pairs.

So $\texttt{LB}^{g\leftrightarrow l} - \texttt{LB}^{g\leftrightarrow g}$ is translated to Eqn. II − Eqn.III,

$$\begin{aligned}
\texttt{LB}^{g\leftrightarrow l} - \texttt{LB}^{g\leftrightarrow g} &= -\log \frac{\texttt{Sim}_P^{g\leftrightarrow g}}{\texttt{Sim}_P^{g\leftrightarrow g} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g}} \\
&\quad - \left(-\log \frac{\texttt{Sim}_P^{g\leftrightarrow l}}{\texttt{Sim}_P^{g\leftrightarrow l} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l}}\right) \\
&= \log \frac{\texttt{Sim}_P^{g\leftrightarrow g} \cdot (\texttt{Sim}_P^{g\leftrightarrow l} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l})}{\texttt{Sim}_P^{g\leftrightarrow l} \cdot (\texttt{Sim}_P^{g\leftrightarrow g} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g})}
\end{aligned} \quad \text{(IV)}$$

Intuitively, if we want to get $\texttt{LB}^{g\leftrightarrow l} > \texttt{LB}^{g\leftrightarrow g}$, we need to prove numerator (denote as A) > denominator(denote as B) in Eqn. IV,

$$\begin{aligned}
\texttt{LB}^{g\leftrightarrow l} - \texttt{LB}^{g\leftrightarrow g} &\approx \texttt{A} - \texttt{B} \\
&= \texttt{Sim}_P^{g\leftrightarrow g} \cdot (\texttt{Sim}_P^{g\leftrightarrow l} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l}) \\
&\quad - \texttt{Sim}_P^{g\leftrightarrow l} \cdot (\texttt{Sim}_P^{g\leftrightarrow g} + \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g}) \\
&= \texttt{Sim}_P^{g\leftrightarrow g} \cdot \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l} - \texttt{Sim}_P^{g\leftrightarrow l} \cdot \sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g} \\
&= P^g \cdot N^l - P^l \cdot N^g,
\end{aligned} \quad \text{(V)}$$

where we use P^g, N^l, P^l, N^g to denote $\texttt{Sim}_P^{g\leftrightarrow g}$, $\sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow l}$, $\texttt{Sim}_P^{g\leftrightarrow l}$, $\sum_{i=1}^{K} \texttt{Sim}_N^{g\leftrightarrow g}$ for simplicity. In summary, if we can prove $P^g \cdot N^l > P^l \cdot N^g$, then we can conclude that $\texttt{LB}^{g\leftrightarrow l} > \texttt{LB}^{g\leftrightarrow g}$.

Here we define $\Delta P = P^g - P^l$ and $\Delta N = N^g - N^l$, where ΔP and ΔN denotes the difference between global and local similarity. If we bring $\Delta P, \Delta N$ into Eqn. V, we can get:

$$\begin{aligned}
\texttt{LB}^{g\leftrightarrow l} - \texttt{LB}^{g\leftrightarrow g} &= P^g \cdot N^l - P^l \cdot N^g \\
&= P^g \cdot (N^g - \Delta N) - (P^g - \Delta P) \cdot N^g \\
&= \Delta P \cdot N^g - \Delta N \cdot P^g
\end{aligned} \quad \text{(VI)}$$

Here we can naturally know the $P^g, N^g \in (1, e)$. Then we give an empirically assumption that $\Delta P > 0$ and $\Delta N \to 0$. We verify the assumption is established in both theoretical analysis and experimental support. For experimental support, we collect the statistical data of 32000 samples, as shown in Figure IV, the experimental results match our assumption. The left figure of Figure IV shows that two positive pair's similarity distributions $\Delta P = 0.1$. The right figure shows that two negative pair's similarity distributions $\Delta N \to 0$. We will discuss the theoretical analysis in the next paragraph. *Combine with Eqn. VI and Figure IV*, we can easily conclude that $\Delta P \cdot N^g - \Delta N \cdot P^g > 0$, that is $\texttt{LB}^{g\leftrightarrow l} > \texttt{LB}^{g\leftrightarrow g}$.

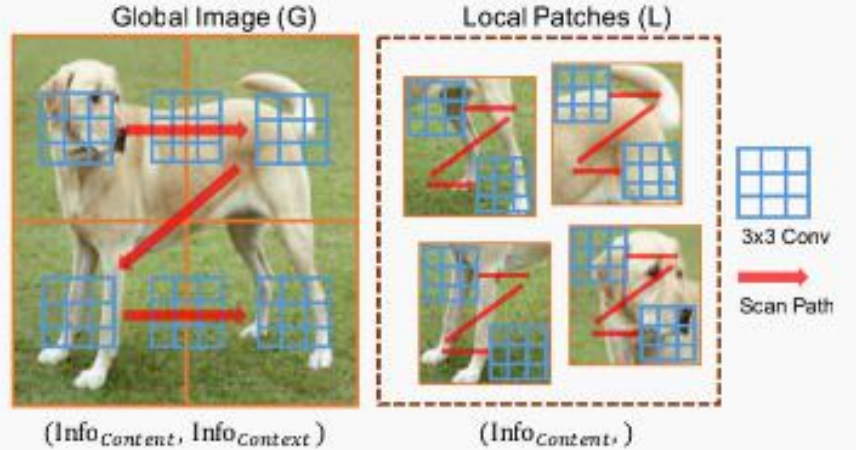


Figure III. **Illustration of the information of global image and local patches extracted by CNN.** For global image, both content information and the context information is extracted by CNN. For local patches, only the content information is extracted by CNN.

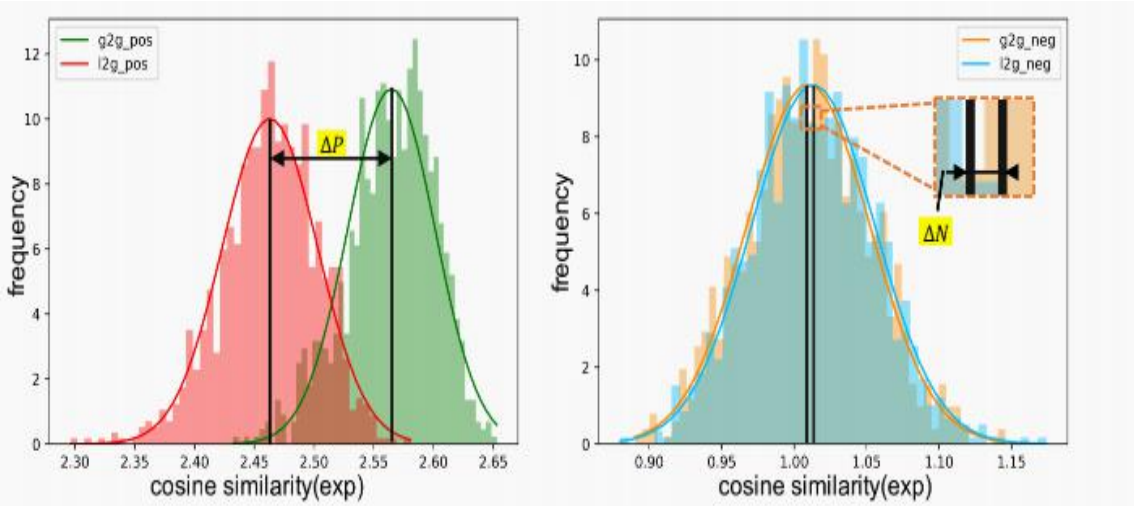


Figure IV. **The statistical distribution of exponential cosine similarity between positive pairs and negative pairs.** The left figure shows the similarity between positive pairs for global↔global and local↔glocal contrasts. The right figure shows the similarity between negative pairs.

Here, we propose three good practices for detection-friendly pretext tasks.

(1) Instance discrimination is better than classifica_x0002_tion or clustering to serve as a pretext task for object detection.
(2) Pretext tasks should keep both low-level and high_x0002_level features discriminative for object detection.
(3) Apart from global image features, local patch representation is also essential for object detection.

Ablation Study

	+HIC	+CGLC	Top1	Top5	mAP
(a)	×	×	64.3	85.6	56.3
(b)	✓	×	63.2 ↓	84.9 ↓	57.0 ↑
(c)	✓	✓	66.6 ↑	87.2 ↑	57.4 ↑

Table 1. **Ablation**: with / without hierarchical intermediate contrastive loss and cross global and local contrasts. The results are evaluated on ImageNet linear classification and PASCAL VOC07+12 detection.

	+HIC	+CGLC	Res2	Res3	Res4	Res5
(a)	×	×	47.1	58.2	70.9	82.1
(b)	✓	×	50.9 ↑	67.1 ↑	78.7 ↑	81.8 ↓
(c)	✓	✓	51.6 ↑	69.7 ↑	82.5 ↑	84.3 ↑

Table 2. **Ablation**: with / without hierarchical intermediate contrastive loss and cross global and local contrasts. Accuracy of feature in different stages are evaluated by PASCAL VOC07 SVM classification.

Transfer Performance

Method	Epoch	AP	AP_{50}	AP_{75}
Rand Init	-	33.8	60.2	33.1
Supervised	90	53.5	81.3	58.8
InsDis [38]	200	55.2(+1.7)	80.9(-0.4)	61.2(+2.4)
PIRL [29]	200	55.5(+2.0)	81.0(-0.3)	61.3(+2.5)
SwAV [3]	800	56.1(+2.6)	82.6(+1.3)	62.7(+3.9)
MoCo [18]	200	55.9(+2.4)	81.5(+0.2)	62.6(+3.8)
MoCov2 [5]	200	57.0(+3.5)	82.4(+1.1)	63.6(+4.8)
MoCov2 [5]	800	57.4(+3.9)	82.5(+1.2)	64.0(+5.2)
DetCo	100	57.4(+3.9)	82.5(+1.2)	63.9(+5.1)
	200	57.8(+4.3)	82.6(+1.3)	64.2(+5.4)
	800	**58.2**(+4.7)	**82.7**(+1.4)	**65.0**(+6.2)

Table 6. **Object Detection finetuned on PASCAL VOC07+12 using Faster RCNN-C4.** DetCo-100ep is on par with previous state-of-the-art, and DetCo-800ep achieves the best performance.

Methods	Instance Seg.		Semantic Seg.
	AP^{mk}	AP^{mk}_{50}	mIOU
Rand Init	25.4	51.1	65.3
supervised	32.9	59.6	74.6
InsDis [38]	33.0 (+0.1)	60.1 (+0.5)	73.3 (-1.3)
PIRL [29]	33.9 (+1.0)	61.7 (+2.1)	74.6 (0.0)
SwAV [3]	33.9 (+1.0)	62.4 (+2.8)	73.0 (-1.6)
MoCo [18]	32.3 (-0.6)	59.3 (-0.3)	75.3 (+0.7)
MoCov2 [5]	33.9 (+1.0)	60.8 (+1.2)	75.7 (+1.1)
DetCo	**34.7** (+1.8)	**63.2** (+3.6)	**76.5** (+1.9)

Table 7. **DetCo *vs.* supervised and other unsupervised methods on Cityscapes dataset.** All methods are pretrained 200 epochs on ImageNet. We evaluate instance segmentation and semantic segmentation tasks.

Method	Epoch	ImageNet		VOC07
		Top1	Top5	Acc
Jigsaw [30]	-	44.6	-	64.5
Rotation [15]	-	55.4	-	63.9
InsDis [38]	200	56.5	-	76.6
LocalAgg [41]	200	58.8	-	-
PIRL [29]	800	63.6	-	81.1
SimCLR [4]	1000	69.3	89.0	-
BYOL [17]	1000	74.3	91.6	-
SwAV [3]	200	72.7	-	87.6
MoCo [18]	200	60.6	-	79.2
MoCov2 [5]	200	67.5	-	84.1
DetCo	200	68.6	88.5	85.1

Table 8. **Comparison of ImageNet Linear Classification and VOC SVM Classification.** Although DetCo is designed for detection, it is also robust and competitive on classification task, and it substantially exceeds MoCov2 baseline by 1.5%.

Method	Epoch	AP^{dp}	AP^{dp}_{50}	AP^{dp}_{75}
Rand Init	-	40.8	78.6	37.3
Supervised	90	50.8	86.3	52.6
MoCo [18]	200	49.6(-1.2)	85.9(-0.4)	50.5(-2.1)
MoCo v2 [5]	200	50.9(+0.1)	87.2(+0.9)	52.9(+0.3)
DetCo	200	**51.3**(+0.5)	**87.7**(+1.4)	**53.3**(+0.7)

Table 9. **DetCo *vs.* other methods on Dense Pose task.** It also performs best on monocular 3D human shape prediction.

Transfer Performance

Method	Mask R-CNN R50-C4 COCO 90k						Mask R-CNN R50-FPN COCO 90k					
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Rand Init	26.4	44.0	27.8	29.3	46.9	30.8	31.0	49.5	33.2	28.5	46.8	30.4
Supervised	38.2	58.2	41.2	33.3	54.7	35.2	38.9	59.6	42.7	35.4	56.5	38.1
InsDis[38]	37.7(-0.5)	57.0(-1.2)	40.9(-0.3)	33.0(-0.3)	54.1(-0.6)	35.2(0.0)	37.4(-1.5)	57.6(-2.0)	40.6(-2.1)	34.1(-1.3)	54.6(-1.9)	36.4(-1.7)
PIRL[29]	37.4(-0.8)	56.5(-1.7)	40.2(-1.0)	32.7(-0.6)	53.4(-1.3)	34.7(-0.5)	37.5(-1.4)	57.6(-2.0)	41.0(-1.7)	34.0(-1.4)	54.6(-1.9)	36.2(-1.9)
SwAV[3]	32.9(-5.3)	54.3(-3.9)	34.5(-6.7)	29.5(-3.8)	50.4(-4.3)	30.4(-4.8)	38.5(-0.4)	**60.4**(+0.8)	41.4(-1.3)	35.4(0.0)	**57.0**(+0.5)	37.7(-0.4)
MoCo[18]	38.5(+0.3)	58.3(+0.1)	41.6(+0.4)	33.6(+0.3)	54.8(+0.1)	35.6(+0.4)	38.5(-0.4)	58.9(-0.7)	42.0(-0.7)	35.1(-0.3)	55.9(-0.6)	37.7(-0.4)
MoCov2[5]	38.9(+0.7)	58.4(+0.2)	42.0(+0.8)	34.2(+0.9)	55.2(+0.5)	36.5(+1.3)	38.9(0.0)	59.4(-0.2)	42.4(-0.3)	35.5(+0.1)	56.5(0.0)	38.1(0.0)
DetCo (ours)	**39.4**(+1.2)	**59.2**(+1.0)	**42.3**(+1.1)	**34.4**(+1.1)	**55.7**(+1.0)	**36.6**(+1.4)	**39.5**(+0.6)	60.3(+0.7)	**43.1**(+0.4)	**35.9**(+0.5)	56.9(+0.4)	**38.6**(+0.5)

Table 4. **Object detection and instance segmentation fine-tuned on COCO**. All methods are pretrained 200 epochs on ImageNet. Our DetCo is state-of-the-art, surpassing MoCov2 and the supervised method in all metrics.

Method	RetinaNet R50 12k			RetinaNet R50 90k			RetinaNet R50 180k			Keypoint RCNN R50 180k		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
Rand Init	4.0	7.9	3.5	24.5	39.0	25.7	32.2	49.4	34.2	65.9	86.5	71.7
Supervised	**24.3**	**40.7**	**25.1**	37.4	56.5	39.7	38.9	58.5	41.5	65.8	86.9	71.9
InsDis[38]	19.0(-5.3)	32.0(-8.7)	19.6(-5.5)	35.5(-1.9)	54.1(-2.4)	38.2(-1.5)	38.0(-0.9)	57.4(-1.1)	40.5(-1.0)	66.5(+0.7)	87.1(+0.2)	72.6(+0.7)
PIRL[29]	19.0(-5.3)	31.7(-9.0)	19.8(-5.3)	35.7(-1.7)	54.2(-2.3)	38.4(-1.3)	38.5(-0.4)	57.6(-0.9)	41.2(-0.3)	66.5(+0.7)	87.5(+0.6)	72.1(+0.2)
SwAV[3]	19.7(-4.6)	34.7(-6.0)	19.5(-5.6)	35.2(-2.2)	54.9(-1.6)	37.5(-2.2)	38.6(-0.3)	58.8(+0.3)	41.1(-0.4)	66.0(+0.2)	86.9(0.0)	71.5(-0.4)
MoCo[18]	20.2(-4.1)	33.9(-6.8)	20.8(-4.3)	36.3(-1.1)	55.0(-1.5)	39.0(-0.7)	38.7(-0.2)	57.9(-0.6)	41.5(0.0)	66.8(+1.0)	87.4(+0.5)	72.5(+0.6)
MoCov2[5]	22.2(-2.1)	36.9(-3.8)	23.0(-2.1)	37.2(-0.2)	56.2(-0.3)	39.6(-0.1)	39.3(+0.4)	58.9(+0.4)	42.1(+0.6)	66.8(+1.0)	87.3(+0.4)	73.1(+1.2)
DetCo (ours)	23.6(-0.7)	38.7(-2.0)	24.6(-0.5)	**38.0**(+0.6)	**57.4**(+0.9)	**40.7**(+1.0)	**39.8**(+0.9)	**59.5**(+1.0)	**42.4**(+0.9)	**67.2**(+1.4)	**87.5**(+0.6)	**73.4**(+1.5)

Table 5. **One-stage object detection and keypoint detection fine-tuned on COCO**. All methods are pretrained 200 epochs on ImageNet. DetCo outperforms all unsupervised counterparts.

Visualization

Figure 5. **Attention maps generated by DetCo and MoCov2 [5].** DetCo can activate more accurate object regions in the heatmap than MoCov2. More visualization results are in Appendix.

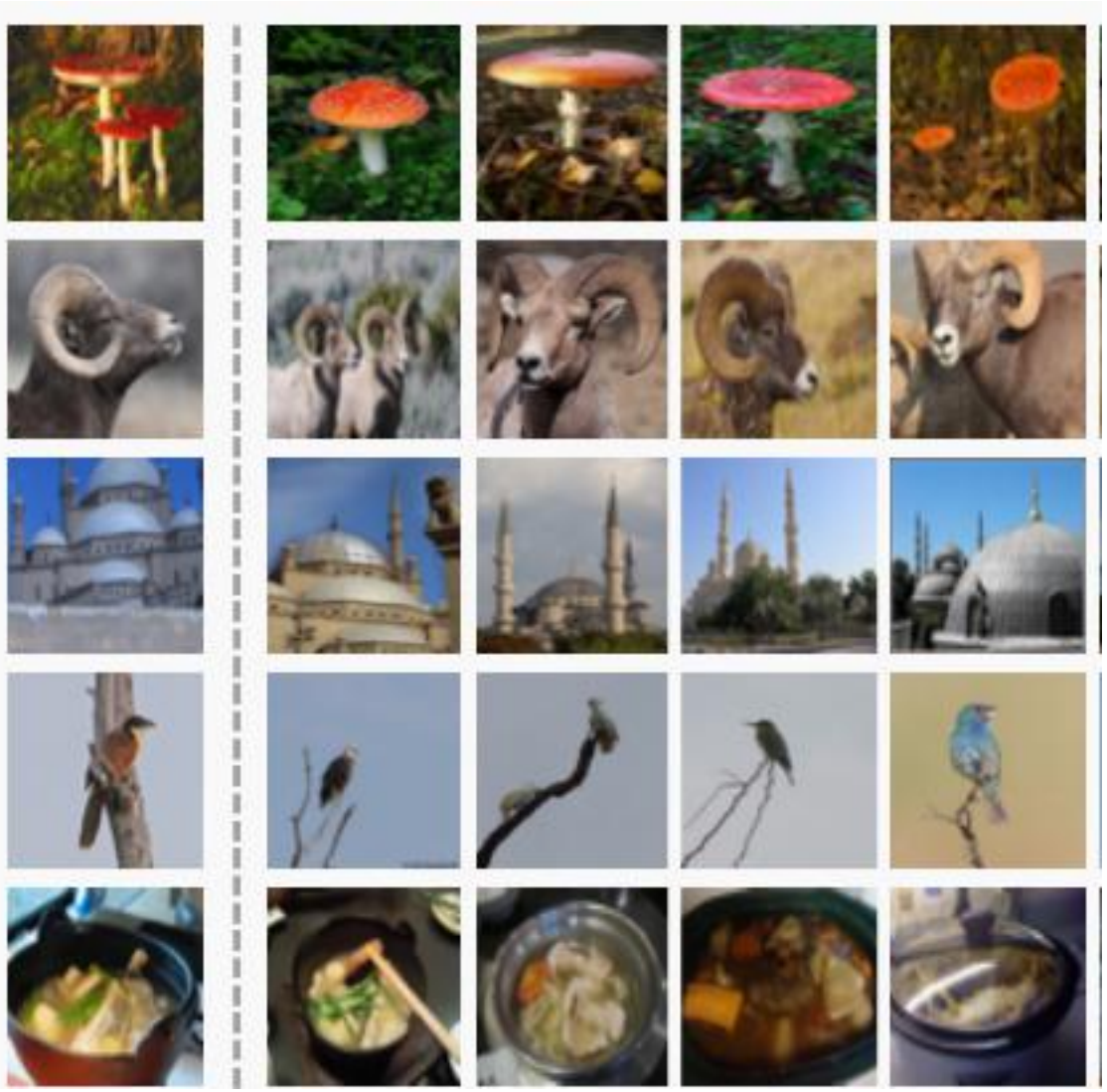


Image Retrieval

Summary

- Self-Supervised Learning has great potential to solve many problems, *e.g.*

(1)Object Localization,

(2)Image Clustering/Retrieval,

(3)Keypoint Matching

- The learned feature is also stronger than Supervised Methods when transfer to downstream tasks

- There is no single best pretext task for different downstream tasks.