DetCo : Unsupervised Contrastive Learning for Object Detection

3^{rd} Paper Study | 2022.08.03 Wed.

한 다 희 Han Dahee

Outline

1. Abstract
2. Introduction
3. Method
4. Experiments
5. Conclusion

Abstract

- ✓ Simple effective self-supervised approach for object detection
- ✓ 최근 unsupervised learning for object detection → image classification X
- ✓ DetCo → dense prediction task, image classification
  - ✓ Multi-level supervision
  - ✓ Contrastive learning between global and local patch
    - → Discriminative and consistent global and local representation

Introduction

- ✓ Self-supervised learning in computer vision
  - → To facilitate image classification, object detection, and semantic segmentation
  - → To provide models pre-trained on large-scale unlabeled data
- ✓ Previous → different pretext task → contrastive learning
- ✓ MoCo v1/v2, BYOL, SwAV : image classification O, object detection X
- ✓ DenseCL, InsLoc, PatchReID : image classification X, object detection O
- ✓ ***Challenging !***

Introduction

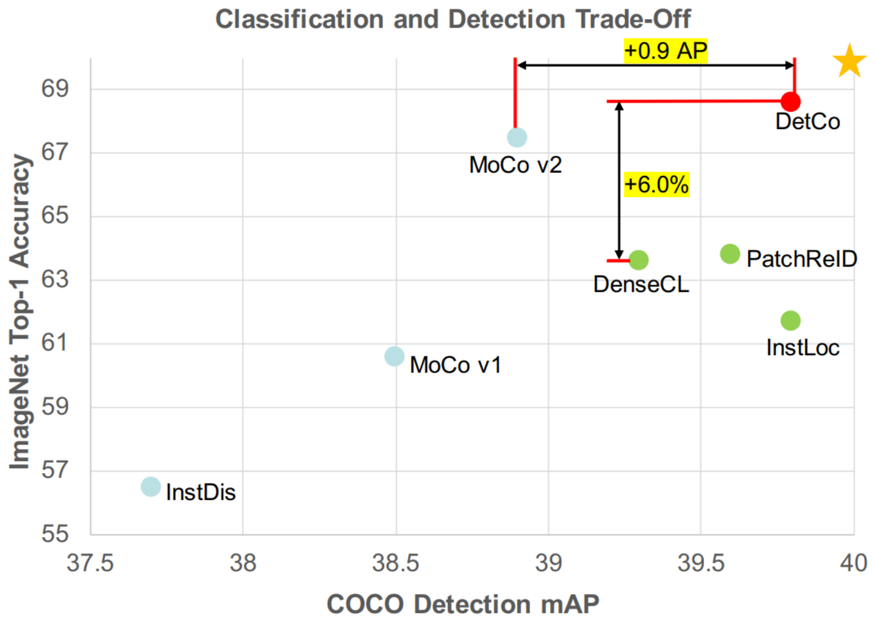

Figure 1. **Transfer accuracy on Classification and Detection.** DetCo achieves the best performance trade-off on both classification and detection. For example, DetCo outperforms its strong baseline, MoCo v2 [5], by 0.9 AP on COCO detection. Moreover, DetCo is significant better than recent work *e.g.* DenseCL [39], InsLoc [41], PatchReID [8] on ImageNet classification while also has advantages on object detection. Note that these three methods are concurrent work and specially designed for object detection (mark with **green**). The yellow asterisk indicates that a desired method should have both high performance in detection and classification.

Introduction

- ✓ Image classification → global instance 인식
- ✓ Object detection → local instance 인식
- ✓ Building instance representation
  - ✓ Discriminative at each level of feature pyramid
  - ✓ Consistent for both global image and local patch

Introduction

✓ DetCo : contrastive learning framework

→ Instance-level detection task & competitive image classification

→ Multi-level supervision : 각 stage에서 feature를 optimize

→ Contrastive learning between global image and local patches : image, patch 별 consistency

그리고 local patch끼리는 구별되게

Method

- ✓ Multi-level supervision → keeps features at multiple stages discriminative
- ✓ Global and local contrastive learning → global and local representation 향상
- ✓ DetCo Framework
  - ✓ MoCo v2 기반 (MLP heads, memory bank)
  - ✓ Image classification, instance-level detection task

$$\mathcal{L}(\mathbf{I}_q, \mathbf{I}_k, \mathbf{P}_q, \mathbf{P}_k) = \sum_{i=1}^{4} w_i \cdot (\mathcal{L}^i_{g\leftrightarrow g} + \mathcal{L}^i_{l\leftrightarrow l} + \mathcal{L}^i_{g\leftrightarrow l}), \quad (1)$$

$\sum_{i=1}^{4} w_i \cdot \mathcal{L}^i$: multi-level supervision

$\mathcal{L}^i_{g\leftrightarrow g} + \mathcal{L}^i_{l\leftrightarrow l} + \mathcal{L}^i_{g\leftrightarrow l}$: global and local contrastive learning

Method

- ✓ Multi-level Supervision
  - ✓ 각 level에서 strong discriminator
  - ✓ MoCo을 base로 수정
  - ✓ Backbone ResNet50 : Res2, Res3, Res4, Res5 모두 사용
  - ✓ 각 단계에서 contrastive loss 계산, discriminative representation 할 수 있도록
  - ✓ Momentum update
  - ✓ 각 level에서 feature 추출 위한 4개의 MLP heads
  - ✓ Multi-level feature → MLP heads → 4개의 global representation

Method

q, k encode의 global positive pair 유사하도록

$$\mathcal{L}_{g\leftrightarrow g}(\mathbf{I}_q, \mathbf{I}_k) = -\log \frac{\exp(q^g \cdot k_+^g / \tau)}{\sum_{i=0}^{K} \exp(q^g \cdot k_i^g / \tau)}, \qquad (2)$$

$$Loss = \sum_{i=1}^{4} w_i \cdot \mathcal{L}_{g\leftrightarrow g}^{i}, \qquad (3)$$

Method

- ✓ Global and Local Contrastive Learning
  - ✓ To keep consistent instance representation on both patch set and the whole image.
  - ✓ 9 patches → 9 local feature representations → concatenated → MLP head → final representation
  - ✓ Global ←→ local, local ←→ local

$$\mathcal{L}_{g \leftrightarrow l}(\mathbf{P}_q, \mathbf{I}_k) = -\log \frac{\exp(q^l \cdot k_+^g / \tau)}{\sum_{i=0}^{K} \exp(q^l \cdot k_i^g / \tau)}. \quad (4)$$

q의 local patch, k의 global positive pair

$$\mathcal{L}_{l \leftrightarrow l}(\mathbf{P}_q, \mathbf{P}_k) = -\log \frac{\exp(q^l \cdot k_+^l / \tau)}{\sum_{i=0}^{K} \exp(q^l \cdot k_i^l / \tau)}. \quad (5)$$

q의 local patch, k의 local patch positive pair

Method

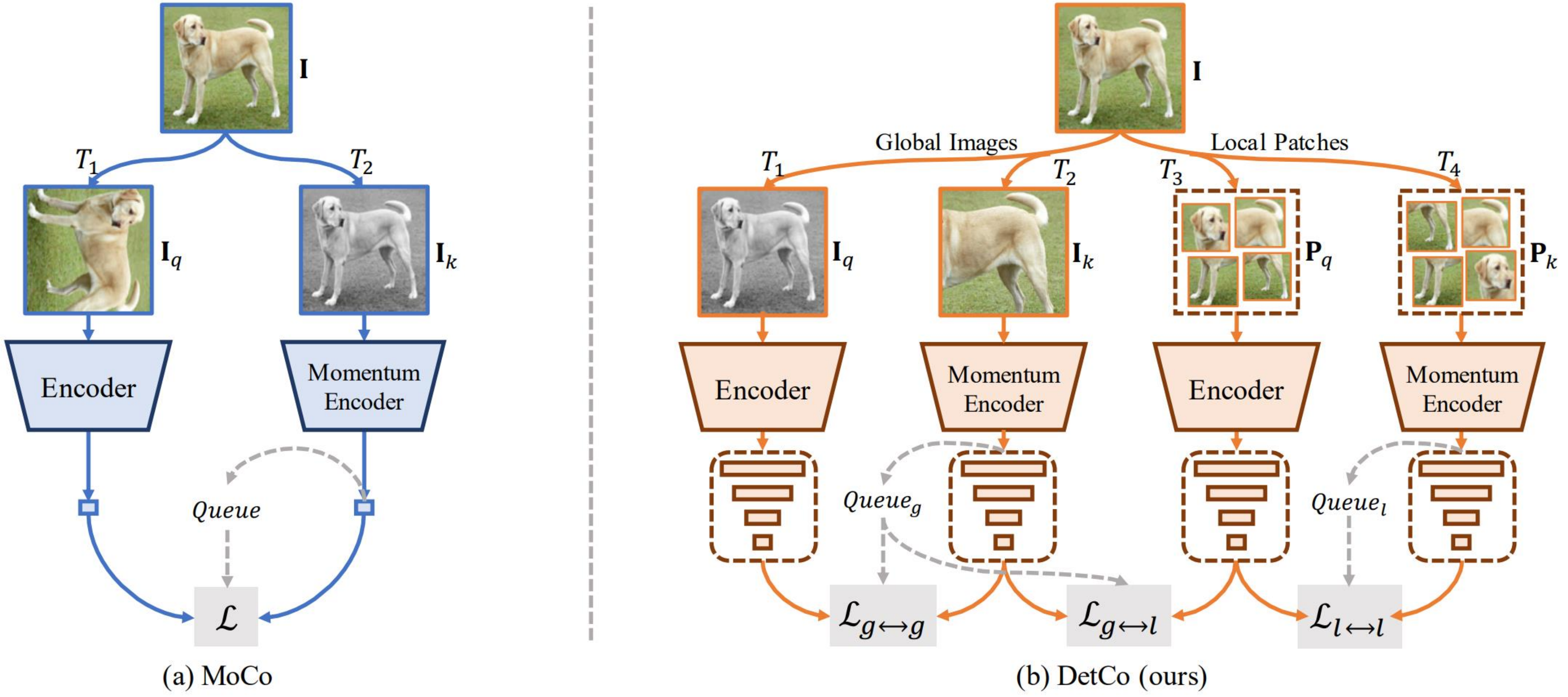

Figure 2. **The overall pipeline of DetCo compared with MoCo [19].** (a) is MoCo's framework, which only considers the single high-level feature and learning contrast from a global perspective. (b) is our DetCo, which learns representation with multi-level supervision and adds two additional local patch sets for input, building contrastive loss cross the global and local views. Note that "T" means image transforms. "$Queue_{g/l}$" means different memory banks [40] for global/local features.

Experiments

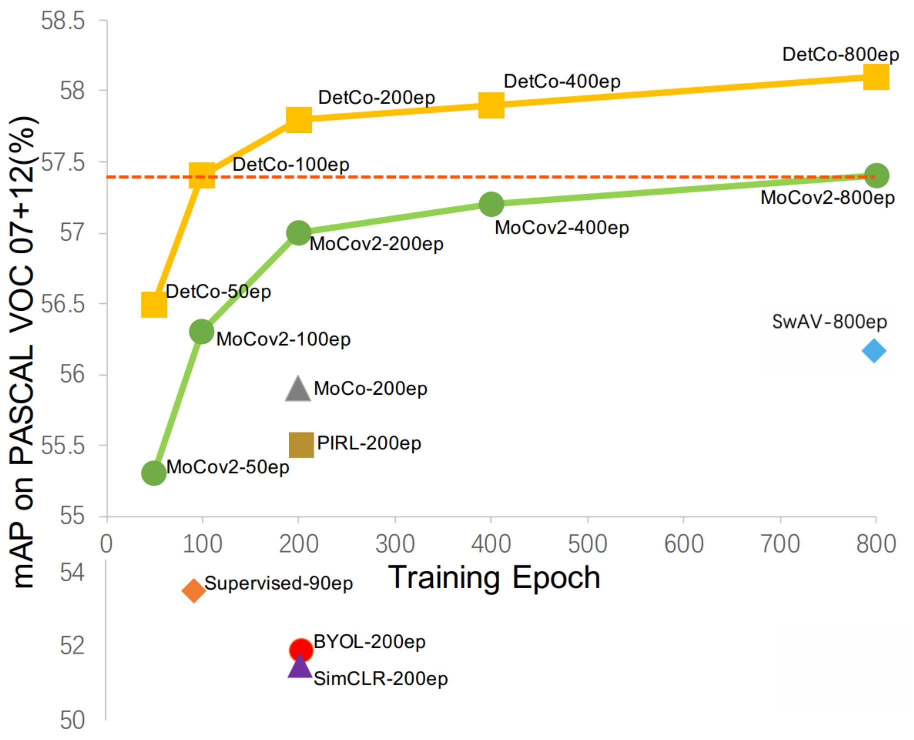

Figure 3. **Comparisons of mAP on PASCAL VOC 07+12 object detection.** For different pre-training epoches, we see that DetCo consistently outperforms MoCo v2[5], which is a strong competitor on VOC compared to other methods. For example, DetCo-100ep already achieves similar mAP compared to MoCov2-800ep. Moreover, DetCo-800ep achieves state-of-the-art and outperforms other counterparts.

Experiments

Method	Mask R-CNN R50-C4 COCO 12k						Mask R-CNN R50-FPN COCO 12k					
	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	7.9	16.4	6.9	7.6	14.8	7.2	10.7	20.7	9.9	10.3	19.3	9.6
Supervised	27.1	46.8	27.6	24.7	43.6	25.3	28.4	48.3	29.5	26.4	45.2	25.7
InsDis[40]	25.8(-1.3)	43.2(-3.6)	27.0(-0.6)	23.7(-1.0)	40.4(-3.2)	24.5(-0.8)	24.2(-4.2)	41.5(-6.8)	25.1(-4.4)	22.8(-3.6)	38.9(-6.3)	23.7(-2.0)
PIRL[30]	25.5(-1.6)	42.6(-4.2)	26.8(-0.8)	23.2(-1.5)	39.9(-3.7)	23.9(-1.4)	23.7(-4.7)	40.4(-7.9)	24.4(-5.1)	22.1(-4.3)	37.9(-7.3)	22.7(-3.0)
SwAV[3]	16.5(-10.6)	35.2(-11.6)	13.5(-14.1)	16.1(-8.6)	32.0(-11.6)	14.6(-10.7)	25.5(-2.9)	46.2(-2.1)	25.4(-4.1)	24.8(-1.6)	43.5(-1.7)	25.3(-0.4)
MoCo[19]	26.9(-0.2)	44.5(-2.3)	28.2(+0.6)	24.6(-0.1)	41.8(-1.8)	25.6(+0.3)	25.6(-2.8)	43.4(-4.9)	26.6(-2.9)	23.9(-2.5)	40.8(-4.4)	24.8(-0.9)
MoCov2[5]	27.6(+0.5)	45.3(-1.5)	28.9(+1.3)	25.1(+0.4)	42.6(-1.0)	26.3(+1.0)	26.6(-1.8)	44.9(-3.4)	27.7(-1.8)	24.8(-1.6)	42.1(-3.1)	25.7(0.0)
DetCo	**29.8**(+2.7)	**49.1**(+2.3)	**31.4**(+3.8)	**26.9**(+2.2)	**46.0**(+2.4)	**27.9**(+2.6)	**29.6**(+1.2)	**49.4**(+1.1)	**31.0**(+1.5)	**27.6**(+1.2)	**46.6**(+1.4)	**28.7**(+3.0)

Table 2. **Object detection and instance segmentation fine-tuned on COCO**. All methods are pretrained 200 epochs on ImageNet. **Green** means increase and **gray** means decrease. DetCo outperforms all supervised and unsupervised counterparts.

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[40]	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[30]	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[19]	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	**39.8**(+1.6)	**59.7**(+1.5)	**43.0**(+1.8)	**34.7**(+1.4)	**56.3**(+1.6)	**36.7**(+1.5)	**40.1**(+1.2)	**61.0**(+1.4)	**43.9**(+1.2)	**36.4**(+1.0)	**58.0**(+1.5)	**38.9**(+0.8)

Table 3. **Object detection and instance segmentation fine-tuned on COCO**. All methods are pretrained 200 epochs on ImageNet. DetCo outperforms all supervised and unsupervised counterparts.

STRADVISION

Experiments

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[40]	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[30]	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[19]	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	**25.3**(+1.0)	**41.6**(+0.9)	**26.5**(+1.4)	**38.4**(+1.0)	**57.8**(+1.3)	**41.2**(+1.5)	**39.7**(+0.8)	**59.3**(+0.8)	**42.6**(+1.1)	**67.2**(+1.4)	**87.5**(+0.6)	**73.4**(+1.5)

Table 7. **One-stage object detection and keypoint detection fine-tuned on COCO**. All methods are pretrained 200 epochs on ImageNet. DetCo outperforms all supervised and unsupervised counterparts.

Method	RetinaNet R50 COCO 1% Data			RetinaNet R50 COCO 2% Data			RetinaNet R50 COCO 5% Data			RetinaNet R50 COCO 10% Data		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP50	AP75
Rand Init	1.4	3.5	1.0	2.5	5.6	2.0	3.6	7.4	3.0	3.7	7.5	3.2
Supervised	8.2	16.2	7.2	11.2	21.7	10.3	16.5	30.3	15.9	19.6	34.5	19.7
MoCo[19]	7.0(-1.2)	13.5(-2.7)	6.5(-0.7)	10.3(-0.9)	19.2(-2.5)	9.7(-0.6)	15.0(-1.5)	27.0(-3.3)	14.9(-1.0)	18.2(-1.4)	31.6(-2.9)	18.4(-1.3)
MoCo v2[5]	8.4(+0.2)	15.8(-0.4)	8.0(+0.8)	12.0(+0.8)	21.8(+0.1)	11.5(+1.2)	16.8(+0.3)	29.6(-0.7)	16.8(+0.9)	20.0(+0.4)	34.3(-0.2)	20.2(+0.5)
DetCo	**9.9**(+1.7)	**19.3**(+3.1)	**9.1**(+1.9)	**13.5**(+2.3)	**25.1**(+3.4)	**12.7**(+2.4)	**18.7**(+2.2)	**32.9**(+2.6)	**18.7**(+2.8)	**21.9**(+2.3)	**37.6**(+3.1)	**22.3**(+2.6)

Table 8. **Semi-Supervised one-stage detection fine-tuned on COCO 1%, 2%, 5% and 10% data**. All methods are pretrained 200 epochs on ImageNet. DetCo is significant better than supervised / unsupervised counterparts in all metrics.

Experiments

	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised	45.0	64.1	49.0	27.7	47.5	59.6
DetCo	**46.5**	**65.7**	**50.8**	**30.8**	**49.5**	**59.7**

Table 4. **DetCo *vs.* Supervised pre-train** on Sparse R-CNN DetCo largely improves 1.5 mAP and 3.1 AP_s.

Method	Epoch	ImageNet		VOC07
		Top1	Top5	Acc
Jigsaw [31]	-	44.6	-	64.5
Rotation [16]	-	55.4	-	63.9
InsDis [40]	200	56.5	-	76.6
LocalAgg [44]	200	58.8	-	-
PIRL [30]	800	63.6	-	81.1
SimCLR [4]	1000	69.3	89.0	-
BYOL [18]	1000	74.3	91.6	-
SwAV [3]	200	72.7	-	87.6
MoCo [19]	200	60.6	-	79.2
MoCov2 [5]	200	67.5	-	84.1
DetCo	200	68.6	88.5	85.1

Table 10. **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 1.1%.

Experiments

Method	Epoch	AP	AP_{50}	AP_{75}
Rand Init	-	33.8	60.2	33.1
Supervised	90	53.5	81.3	58.8
InsDis [40]	200	55.2(+1.7)	80.9(-0.4)	61.2(+2.4)
PIRL [30]	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 [19]	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 9. **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.

Experiments

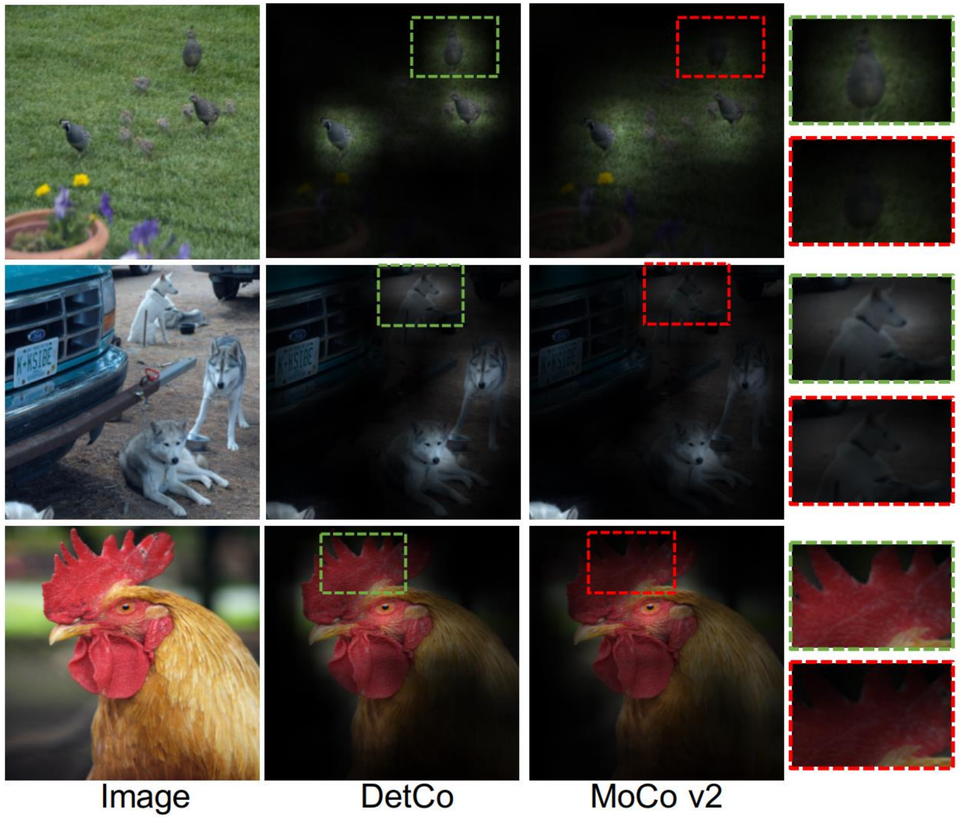

Figure 4. **Attention maps generated by DetCo and MoCov2 [5].** DetCo can activate more accurate object regions in the heatmap than MoCov2. More visualization results are in Appendix.

Conclusion

- ✓ MoCo v2 baseline
  - ✓ (1) multi-level supervision
  - ✓ (2) global and local contrastive learning
- ✓ Detection, image classification

References

[Paper]

Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., ... & Luo, P. (2021). Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8392-8401).

Q & A

Thank you ☺