Mosaic Representation Learning for Self-supervised Visual Pre-training

Zhaoqing Wang[1,4] **Ziyu Chen**[4] **Yaqian Li**[4] **Yandong Guo**[4] **Jun Yu**[3]
Mingming Gong[2,*] **Tongliang Liu**[1,*]
[1] Sydney AI Centre, The University of Sydney [2] The University of Melbourne
[3] University of Science and Technology of China [4] OPPO Research Institute
zwan6779@uni.sydney.edu.au; mingming.gong@unimelb.edu.au
tongliang.liu@sydney.edu.au

ICLR 2023

Background

Self-Supervised Learning (SSL)

High-quality representation learning is a fundamental task in machine learning. Tremendous number of visual recognition models have achieved promising performance by learning from large-scale annotated datasets. However, a great deal of challenges exist in collecting large-scale datasets with annotations, e.g., label noise, high cost and privacy concerns. To address these issues, self-supervised learning (SSL) is proposed to learn generic representations without manual annotation. Recent progress in visual self-supervised learning shows remarkable potential and achieves comparable results with supervised learning.

Related work

Among these SSL methods, a common underlying idea is to extract invariant feature representations from different augmented views of the same input image.

Contrastive learning is one of the most popular self-supervised learning frameworks and has achieved great success in recent years.

$$z_i^q = g(F(x_i^q)),\ z_i^k = g(F(x_i^k))$$

$$\mathcal{L}_{contrast} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(z_i^q, z_i^k)/\tau)}{\exp(sim(z_i^q, z_i^k)/\tau) + \sum_{n^- \in \mathcal{N}} \exp(sim(z_i^q, n^-)/\tau)}$$

Related work

Without the need of negatives, BYOL and Simsiam adopt an extra predictor to map the embedding z to the prediction p and minimize their negative cosine similarity of them.

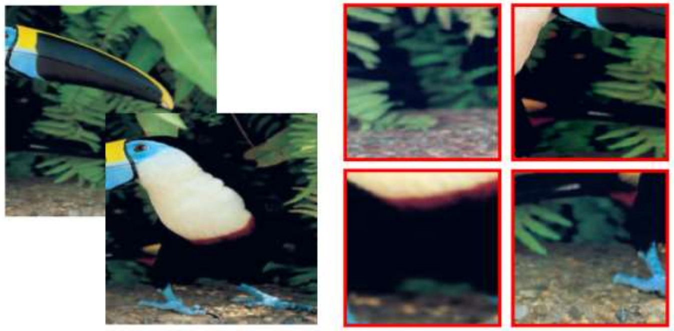
$$\mathcal{L}_{cos} = \frac{1}{N}\sum_{i=1}^{N} -sim(p_i^q, sg(z_i^k)) - sim(p_i^k, sg(z_i^q))$$

Motivation

A carefully-designed data augmentation strategy is an essential part of the above self-supervised learning frameworks.

SimCLR and InfoMin empirically investigate the impact of different data augmentations and observe that SSL benefits more from strong data augmentations than supervised learning.

SwAV proposes the multi-crop strategy, which achieves significant performances on downstream tasks.



(a) Standard view (b) Multi-crop view

Motivation

we propose a mosaic representation learning framework (MosRep) consisting of a new data augmentation strategy, which can enrich the contextual background of each small crop and encourage the “local-to-global” correspondences.

Method

Given an input image x_i, we generate two standard views and M small crops by three separate augmentation operators t^q, t^k and t^s.

$$x_i^q = t^q(x_i), \quad x_i^k = t^k(x_i), \quad \mathcal{X}_i^s = t^s(x_i), \qquad t^q, t^k, t^s \sim T$$

We randomly shuffle all small crops from images in a batch and divide them into groups. We set up M crops in each group and ensure that the crops in each group come from different input images. Then, we compose the crops of each group into a single view, termed as the mosaic view x_i^v, and record the coordinates (t, l, b, r) of each small crop relative to x_i^v.

$$x_i^v = Compose(\mathcal{M}_i), \quad \mathcal{M}_i = \{x_{ij}^s \mid i \in \mathbf{N}\}_{j=1}^M$$

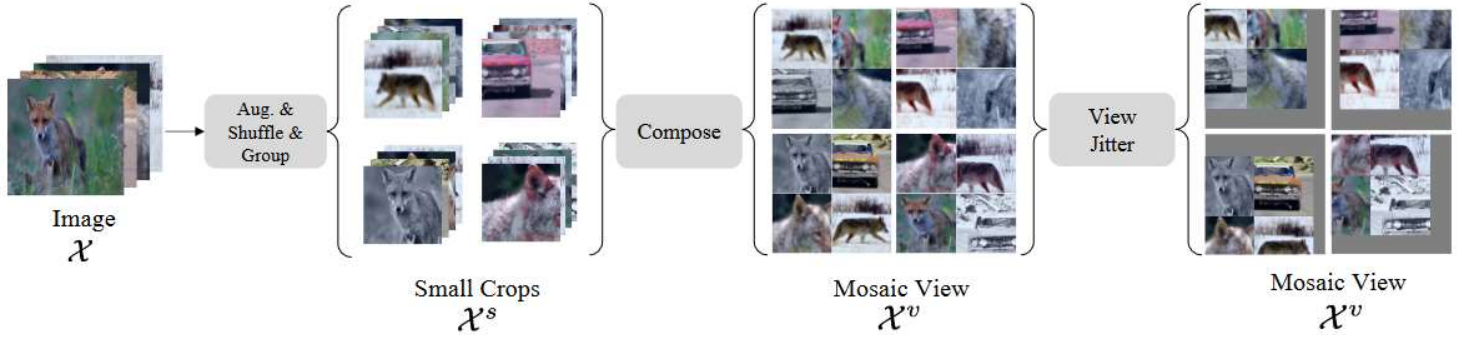
Method

Since the spatial position of each crop is fixed in the mosaic view x_i^v , the model can easily memorize the position, resulting in over-fitting. In order to tackle this dilemma, we conduct the view jitter operation on the mosaic view. We first sample offsets of the mosaic view from a beta distribution $\beta(\alpha, \alpha)$ with two identical parameters α.

$$\Delta x = \theta \cdot u, \quad \Delta y = \theta \cdot v, \qquad u, v \sim \beta(\alpha, \alpha)$$

We set $\alpha < 1$, which indicates a U-shaped distribution. In this way the mosaic view is more likely to be jittered in a relatively wider range with larger offsets.

$$(t', l', b', r') = (t + \Delta y, l + \Delta x, b + \Delta y, r + \Delta x)$$



Method

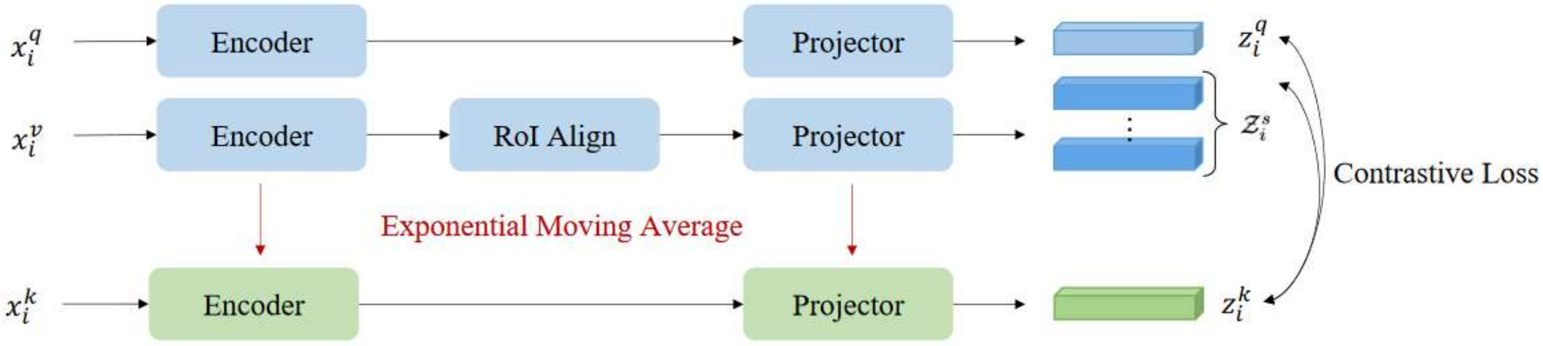
Given two standard views x_i^q , x_i^k and a mosaic view x_i^v , an encoder $F(\cdot)$ is used to extract the feature h of them.

$$h_i^q = F(x_i^q), \quad h_i^k = F(x_i^k), \quad h_i^v = F(x_i^v)$$

We adopt a RoI Align operator to extract the feature of each small crop in the mosaic view.

All features are mapped into an embedding space by a projector $g(\cdot)$.

$$z_i^q = g(h_i^q), \quad z_i^k = g(h_i^k), \quad \mathcal{Z}_i^s = g(\mathcal{H}_i^s)$$



Method

MoCo-v2

$$
\begin{aligned}
\mathcal{L}_{contrast} = & -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(z_i^q, z_i^k)/\tau)}{\exp(sim(z_i^q, z_i^k)/\tau) + \sum_{n_j\in\mathcal{N}}\exp(sim(z_i^q, n_j)/\tau)} \\
& -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\log\frac{\exp(sim(z_{ij}^s, z_i^k)/\tau)}{\sum_{n^-\in\mathcal{Z}^k\cup\mathcal{N}}\exp(sim(z_{ij}^s, n^-)/\tau)},
\end{aligned}
$$

BYOL

$$
\begin{aligned}
Loss = & -\frac{1}{B}\sum_{j\in B}\sum_{k\in B'}\frac{\exp(p_{c_{jk}}^{i2t}/\tau^T)}{\sum_{k\in B'}\exp(p_{c_{jk}}^{i2t})}\cdot\log\frac{\exp(p_{s_{jk}}^{i2t}/\tau^S)}{\sum_{k\in B'}\exp(p_{s_{jk}}^{i2t})} \\
& -\frac{1}{B}\sum_{j\in B}\sum_{k\in B'}\frac{\exp(p_{c_{jk}}^{t2i}/\tau^T)}{\sum_{k\in B'}\exp(p_{c_{jk}}^{t2i})}\cdot\log\frac{\exp(p_{s_{jk}}^{i2i}/\tau^S)}{\sum_{k\in B'}\exp(p_{s_{jk}}^{i2i})},
\end{aligned}
$$

Experiment

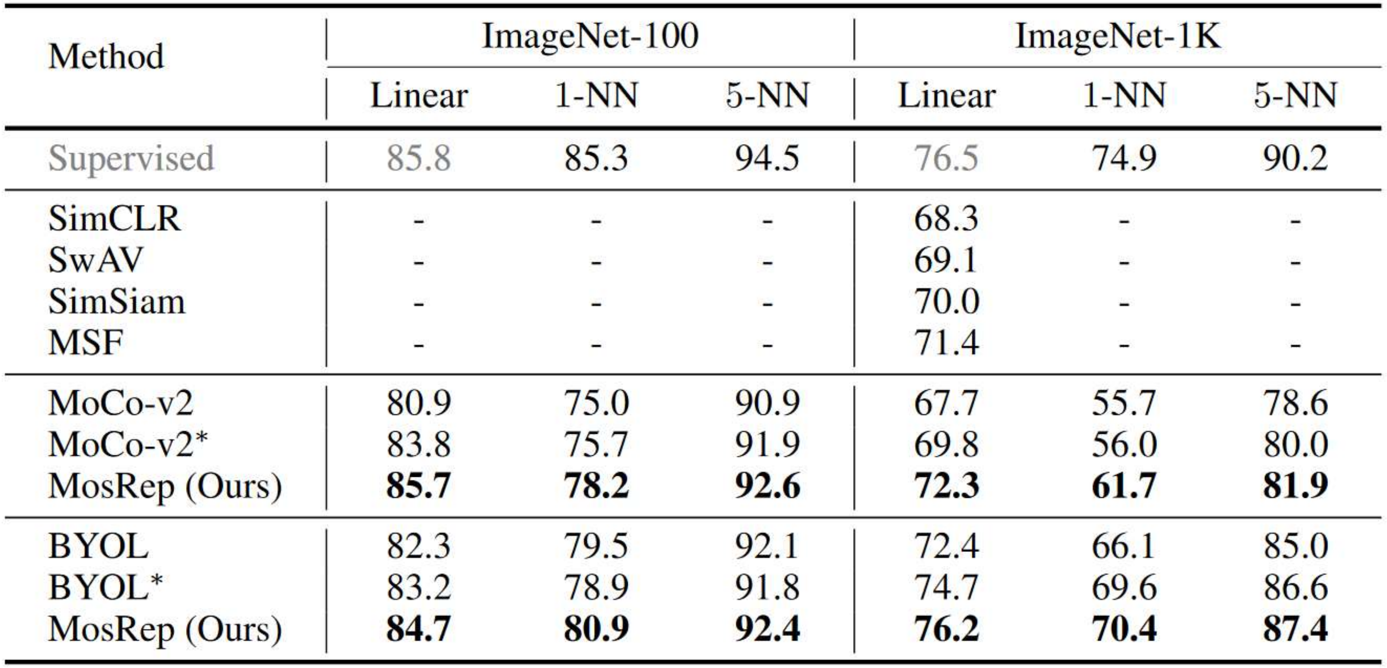
Datasets 1) ImageNet-100 2) ImageNet-1K

Architectures We adopt the ResNet-50 model as our encoder. We build on two different frameworks, MoCo-v2 and BYOL.

Data Augmentation During pre-training, we adopt the data augmentation used in MoCo-v2 and BYOL for the standard view. As for the mosaic view, we generate M = 4 small crops with 112 × 112 input size, and other data augmentations are the same with MoCo-v2 and BYOL.

Experiment

Linear Probing, Nearest Neighbor

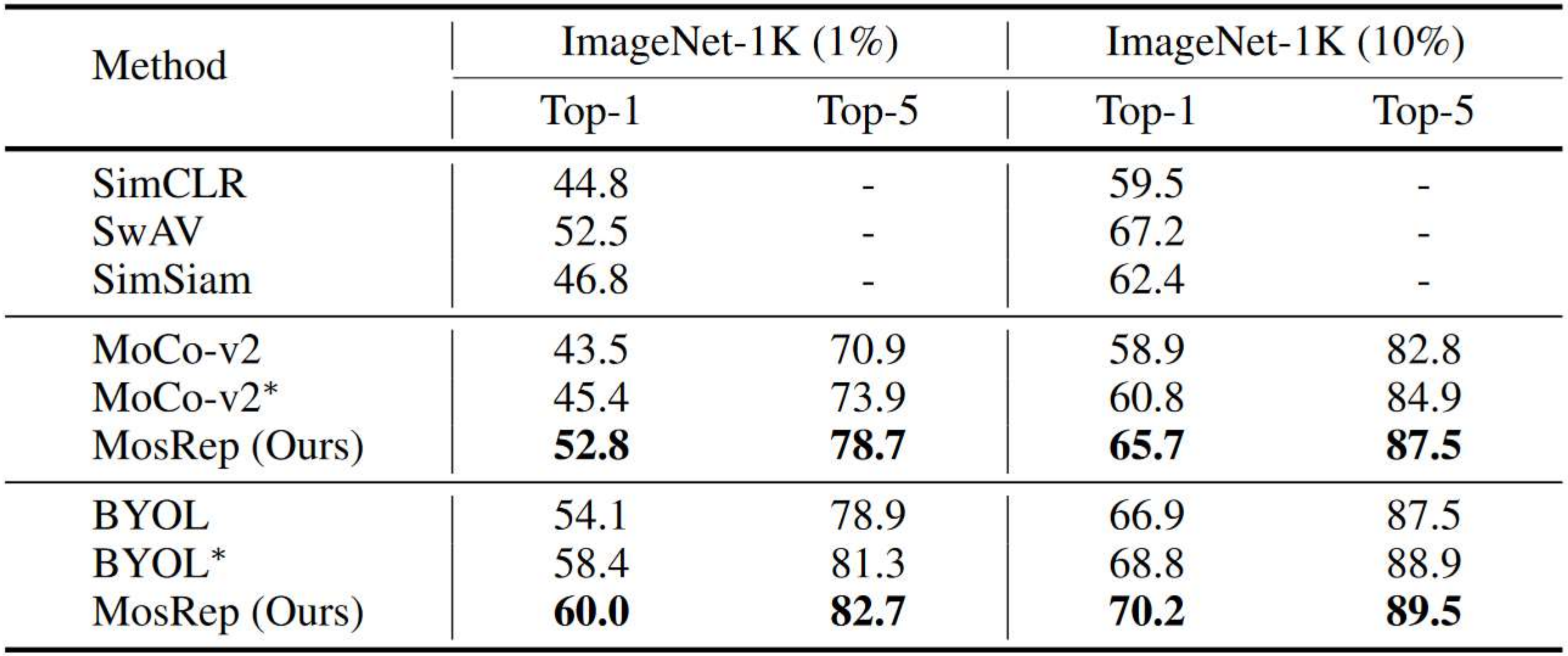


Method	ImageNet-100			ImageNet-1K		
	Linear	1-NN	5-NN	Linear	1-NN	5-NN
Supervised	85.8	85.3	94.5	76.5	74.9	90.2
SimCLR	-	-	-	68.3	-	-
SwAV	-	-	-	69.1	-	-
SimSiam	-	-	-	70.0	-	-
MSF	-	-	-	71.4	-	-
MoCo-v2	80.9	75.0	90.9	67.7	55.7	78.6
MoCo-v2*	83.8	75.7	91.9	69.8	56.0	80.0
MosRep (Ours)	**85.7**	**78.2**	**92.6**	**72.3**	**61.7**	**81.9**
BYOL	82.3	79.5	92.1	72.4	66.1	85.0
BYOL*	83.2	78.9	91.8	74.7	69.6	86.6
MosRep (Ours)	**84.7**	**80.9**	**92.4**	**76.2**	**70.4**	**87.4**

MosRep shows a considerable improvement over the strong baseline on both IN-100 and IN-1K datasets, which demonstrates that the performance gain is not simply from more small crops.

Experiment

Semi-supervised Learning

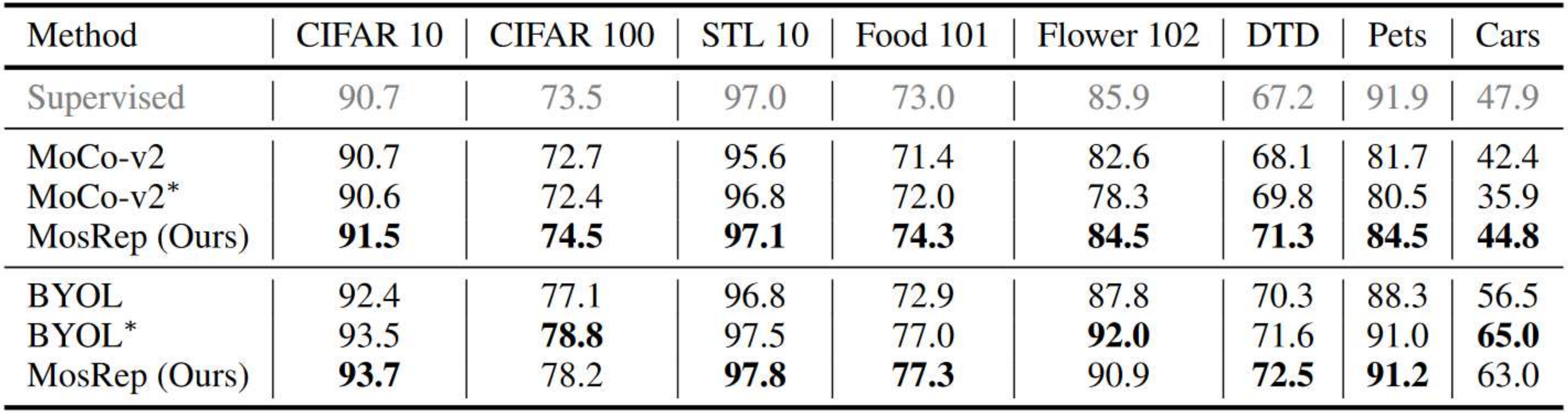


Method	ImageNet-1K (1%)		ImageNet-1K (10%)	
	Top-1	Top-5	Top-1	Top-5
SimCLR	44.8	-	59.5	-
SwAV	52.5	-	67.2	-
SimSiam	46.8	-	62.4	-
MoCo-v2	43.5	70.9	58.9	82.8
MoCo-v2*	45.4	73.9	60.8	84.9
MosRep (Ours)	**52.8**	**78.7**	**65.7**	**87.5**
BYOL	54.1	78.9	66.9	87.5
BYOL*	58.4	81.3	68.8	88.9
MosRep (Ours)	**60.0**	**82.7**	**70.2**	**89.5**

Impressively, when we build our proposed MosRep on the MoCo-v2, we achieves considerable improvements over the strong baseline with both 1% and 10% labels.

Experiment

Linear probing transfer learning

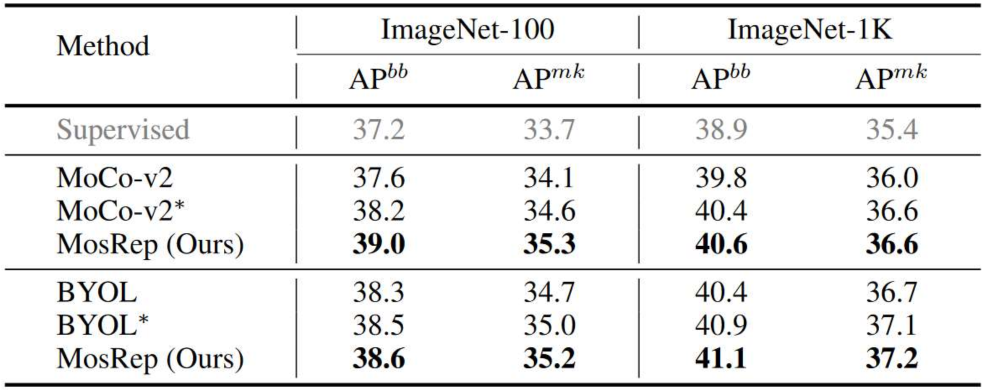


Method	CIFAR 10	CIFAR 100	STL 10	Food 101	Flower 102	DTD	Pets	Cars
Supervised	90.7	73.5	97.0	73.0	85.9	67.2	91.9	47.9
MoCo-v2	90.7	72.7	95.6	71.4	82.6	68.1	81.7	42.4
MoCo-v2*	90.6	72.4	96.8	72.0	78.3	69.8	80.5	35.9
MosRep (Ours)	**91.5**	**74.5**	**97.1**	**74.3**	**84.5**	**71.3**	**84.5**	**44.8**
BYOL	92.4	77.1	96.8	72.9	87.8	70.3	88.3	56.5
BYOL*	93.5	**78.8**	97.5	77.0	**92.0**	71.6	91.0	**65.0**
MosRep (Ours)	**93.7**	78.2	**97.8**	**77.3**	90.9	**72.5**	**91.2**	63.0

Although BYOL is a state-of-the-art self-supervised learning framework, which achieves excellent transfer performances on eight datasets, our approach can improve the generalization performance on most datasets, even outperforming the IN-1K supervised model across the board.

Experiment

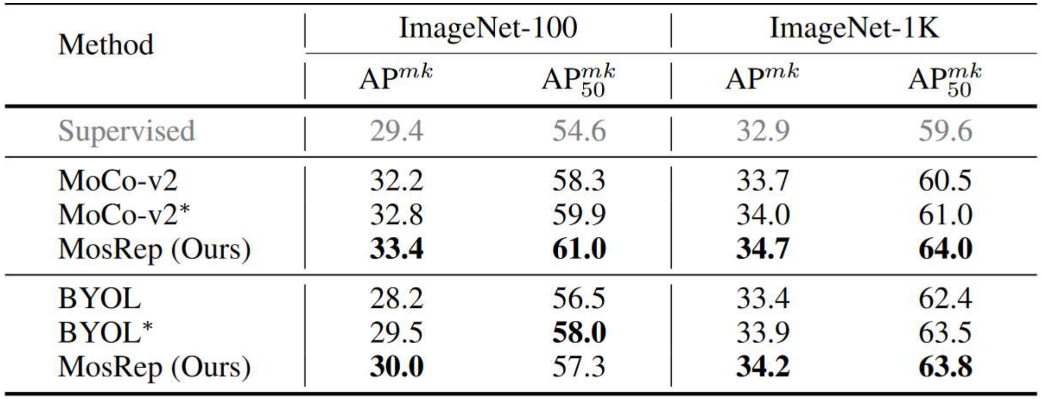
COCO Object Detection & Instance Segmentation



Method	ImageNet-100		ImageNet-1K	
	AP^{bb}	AP^{mk}	AP^{bb}	AP^{mk}
Supervised	37.2	33.7	38.9	35.4
MoCo-v2	37.6	34.1	39.8	36.0
MoCo-v2*	38.2	34.6	40.4	36.6
MosRep (Ours)	**39.0**	**35.3**	**40.6**	**36.6**
BYOL	38.3	34.7	40.4	36.7
BYOL*	38.5	35.0	40.9	37.1
MosRep (Ours)	**38.6**	**35.2**	**41.1**	**37.2**

Our MosRep can consistently increase the ability of detection and segmentation on the MoCo-v2 and BYOL frameworks.

Cityscapes Instance Segmentation

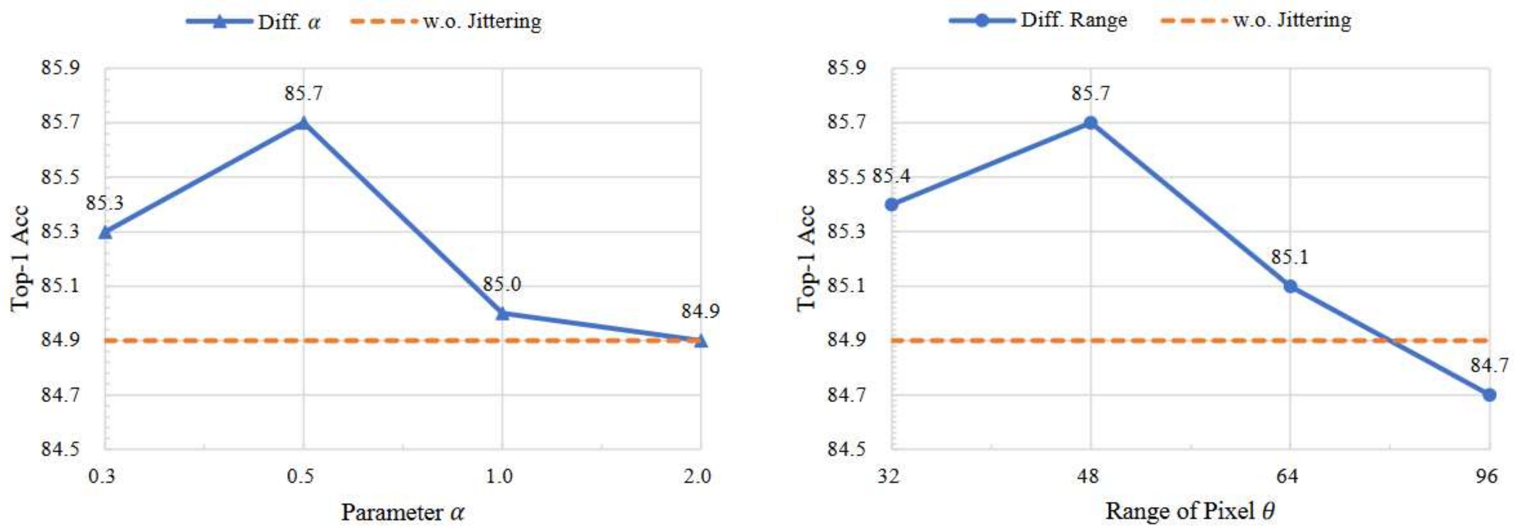


Method	ImageNet-100		ImageNet-1K	
	AP^{mk}	AP^{mk}_{50}	AP^{mk}	AP^{mk}_{50}
Supervised	29.4	54.6	32.9	59.6
MoCo-v2	32.2	58.3	33.7	60.5
MoCo-v2*	32.8	59.9	34.0	61.0
MosRep (Ours)	**33.4**	**61.0**	**34.7**	**64.0**
BYOL	28.2	56.5	33.4	62.4
BYOL*	29.5	**58.0**	33.9	63.5
MosRep (Ours)	**30.0**	57.3	**34.2**	**63.8**

Each pretrained model is transferred to Mask R-CNN R50-FPN model, subsequently finetuned on Cityscapes train set and evaluated on Cityscapes val set.

Experiment

Ablations



The orange dashed line denotes the MosRep without the jitter operation. jittering operation achieves better transfer performance than ones without the jittering operation, which demonstrates that this operation can increase the variance of the mosaic view to prevent the model from memorizing the spatial location of each small crop.

Thanks