Spatially Consistent Representation Learning

Byungseok Roh* Wuhyun Shin* Ildoo Kim Sungwoong Kim
Kakao Brain
{peter.roh, aiden.hsin, ildoo.kim, swkim}@kakaobrain.com

Abstract

Self-supervised learning has been widely used to obtain transferrable representations from unlabeled images. Especially, recent contrastive learning methods have shown impressive performances on downstream image classification tasks. While these contrastive methods mainly focus on generating invariant global representations at the image-level under semantic-preserving transformations, they are prone to overlook spatial consistency of local representations and therefore have a limitation in pretraining for localization tasks such as object detection and instance segmentation. Moreover, aggressively cropped views used in existing contrastive methods can minimize representation distances between the semantically different regions of a single image.

In this paper, we propose a spatially consistent representation learning algorithm (SCRL) for multi-object and location-specific tasks. In particular, we devise a novel self-supervised objective that tries to produce coherent spatial representations of a randomly cropped local region according to geometric translations and zooming operations. On various downstream localization tasks with benchmark datasets, the proposed SCRL shows significant performance improvements over the image-level supervised pretraining as well as the state-of-the-art self-supervised learning methods. Code is available at https://github.com/kakaobrain/scrl.

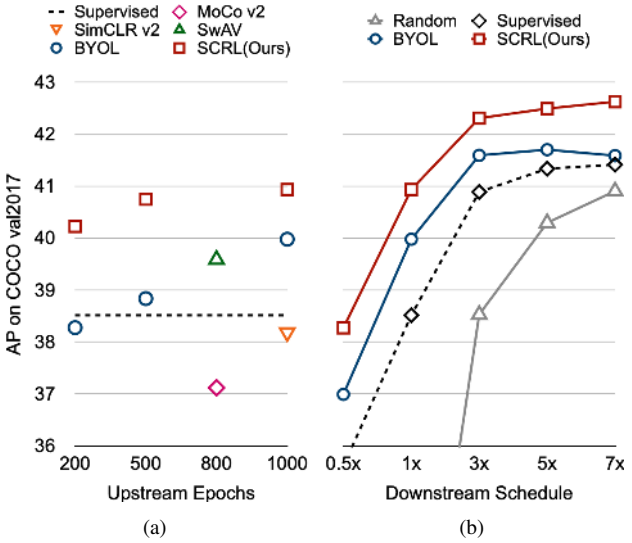

Figure 1. **(a)** AP on downstream of COCO detection task w.r.t. the upstream epochs on ImageNet. We use a ResNet-50-FPN backbone with Faster R-CNN, using default training configuration used in [42]. Only with 200 epochs of upstream, SCRL outperforms the ImageNet pre-trained counterpart as well as the state-of-the-art self-supervised learning methods. **(b)** AP on COCO detection task under varied downstream schedules from 0.5× (45k iterations) to 7× (630k iterations). SCRL consistently outperforms random initialization, supervised pretraining, and BYOL in all the training schedules.

1. Introduction

In computer vision, unsupervised representation learning from a large amount of unlabeled images has been shown to be effective in improving the performances of neural networks for unknown downstream tasks, especially with few labeled data [8, 26]. While conventional generative modeling algorithms are difficult to obtain semantically meaningful representations from high-resolution natural images due to their focus on low-level details [9, 2], self-supervised learning algorithms have recently shown promising results in obtaining semantic representations via the use of proxy tasks on unsupervised data [10, 30, 22, 12, 28, 16, 3, 41, 1, 14]. Among them, contrastive learning methods with discriminative models have particularly achieved remarkable performances on most downstream tasks related to image classification problems [28, 16, 5, 3, 4, 41, 1, 14, 20, 39].

Contrastive self-supervised learning aims to obtain discriminative representations based on the semantically positive and negative image pairs. Specifically, it tries to produce invariant representations from semantic-preserving augmentations of the same image while making representations dissimilar from different images. However, most existing contrastive methods exploit consistent global rep-

*Equal contribution

resentations on a per image basis, specific for image classification, and therefore they are likely to generate inconsistent local representations with respect to the same spatial regions after image transformations. For example, when a certain object in an image is geometrically shifted or scaled, previous global contrastive methods can produce a similar global representation, even if the local feature of that object ends up losing consistency[36], since they use global pooling by which they can attend to other discriminative areas instead. This can consequently lead to performance degradation on localization tasks based on spatial representations. In addition, previous contrastive methods often utilize heavily cropped views from an image to make a positive pair, and hence the representations between the semantically different regions are rather induced to be matched [34].

In order to resolve these issues on the existing global contrastive learning methods, we propose a spatially consistent representation learning algorithm, *SCRL*, that can leverage lots of unlabeled images, specifically for multi-object and location-specific downstream tasks including object detection and instance segmentation. In specific, we develop a new self-supervised objective to realize the invariant spatial representation corresponding to the same cropped region under augmentations of a given image. Since we are able to figure out the two exactly matched spatial locations for each cropped region on the two transformed images, each positive pair of cropped regions necessarily has a common semantic information. From a positive pair of cropped feature maps, we apply RoIAlign [18] to the respective maps and obtain equally-sized local representations. We optimize the encoding network to minimize the distance between these two local representations. Since BYOL [14] has shown to be an efficient contrastive learning method without requiring negative pairs, we adapt its learning framework for producing our spatially coherent representations.

We perform extensive experiments and analysis on several benchmark datasets to empirically demonstrate the effectiveness of the proposed SCRL in significantly improving the performances of fine-tuned models on various downstream localization tasks. Namely, SCRL consistently outperforms the random initialization, the previous image-level supervised pretraining and the state-of-the-art self-supervised methods, on the tasks of object detection and instance segmentation, with the PASCAL VOC, COCO and Cityscapes datasets. In particular, SCRL leads to regress object boundaries more precisely owing to accurate spatial representations before being fed into the task-specific head networks. Importantly, as shown in Figure 1, SCRL outperforms the other pretraining methods even with a small number of epochs during upstream training on unlabeled images. In addition, the improvements in fine-tuned downstream performance obtained by SCRL are consistently maintained under longer schedules as well as small data regime(*i.e.*, 1/10 of COCO training data), which validates the benefits of transferred spatial representations by SCRL.

Our main contributions can be summarized as follows:

- We take into account spatial consistency rather than global consistency on image representations and propose a novel self-supervised learning algorithm, SCRL, on unlabeled images, especially for multi-object and location-aware downstream tasks.
- We generate multiple diverse pairs of semantically-consistent cropped spatial feature maps and apply an efficient contrastive learning method with a dedicated local pooling and projection.
- A variety of experimental results show clear advantages of SCRL over the existing state-of-the-art methods as a transferrable representation pretraining in obtaining better performances on localization tasks.

2. Related Work

Early approaches for self-supervised learning rely on hand-crafted proxy tasks [10, 30, 22, 12] from which the models can extract meaningful information that are beneficial to the considered downstream tasks. However, the representation obtained by those works are prone to lose generality due to the strong prior knowledge reflected to the design choice of pretext tasks.

Recently, contrastive methods [28, 16, 5, 3, 4] have made a lot of progress in the field of self-supervised learning. The goal of contrastive learning is to minimize the distances between the positive pairs, namely, two different augmented views of a single image. At the same time, negative pairs should be pushed apart, which can be directly encouraged by training objectives, such as InfoNCE [41]. PIRL [28] tries to learn invariant features under semantic-preserving transformations. MoCo [16, 5] focuses on constructing a minibatch with a large number of negative samples by utilizing the dynamic queuing and the moving-averaged encoder. SimCLR [3, 4] improves the quality of representation by finding a more proper composition of transformations and an adequate size of non-linear heads at the top of the network. Those methods, however, generally require a larger batch size compared to the supervised counterpart in order to avoid mode collapse problems.

More recently, SwAV [1] modifies previous pairwise representation comparisons by introducing cluster assignment and swapped prediction. They also propose a novel augmentation strategy, multi-crop, which appears to be similar to our SCRL in that they compare multiple smaller patches cropped from the same image, but substantially different in that the spatial consistency on the feature map is

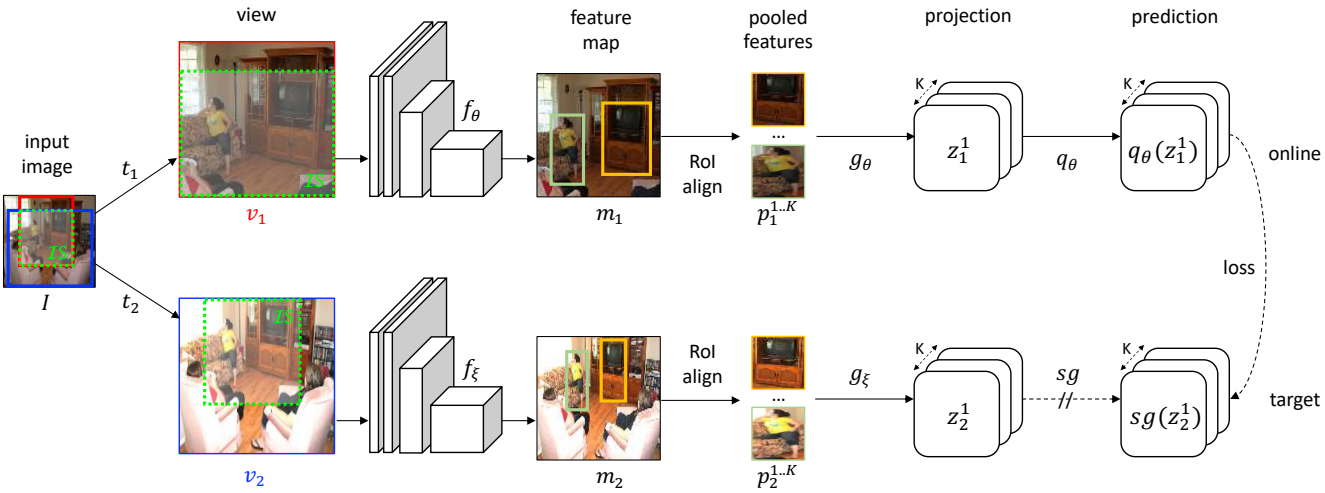

Figure 2. An illustrative view of our method. We first find the intersection region $\mathcal{IS}$ between v_1 and v_2 to randomly generate K number of RoIs within $\mathcal{IS}$. SCRL minimizes a similarity loss between the predictions of the pooled RoIs $q_\theta(z_1^k)$ for v_1 and the projections of the pooled RoIs $sg(z_2^k)$ for v_2, where the online network's parameters θ are trained, parameters of the target network ξ are updated by an exponential moving average of θ, and sg stands for stop-gradient. At the end of training, everything but f_θ is discarded. (Images located in 'feature map' and 'pooled features' are not real features, we use v_1 and v_2 images for a better understanding.)

not directly considered. Grill *et al*. [14] devises the method named BYOL where the network bootstraps its own representation by keeping up with the moving averaged version of itself. BYOL removes the necessity of negative pairs and have shown to be more robust against changes in batch size. However, they also still overlook the object-level local consistency, and moreover, there always exists, as pointed out in [34], a chance that aggressive views taken from a single image can have different semantic meanings, especially in the object-level.

There also has been a large body of works that leverage geometric correspondence for learning dense representation. While the early efforts [15, 44, 21, 40, 31] mostly rely on explicit supervisory signals, some of recent works adopt self-supervised methods to learn the parts or landmarks of the data. Similar to our work, Thewlis *et al*. [38] employ siamese framework to match label maps from the different views of an image by making use of equivariance relation between them. DVE [37] extends this work to even take into account the correspondence across different instances that shares the same object category. In this vein of works, their interests are only limited to dense representation itself that comes with object structure learning, whereas our work aims to transfer the learned representation to a wide variety of localization downstream tasks.

Concurrent to our work, VADeR [33] also similarly learns pixel-level representation in order to transfer it to multiple dense prediction tasks. While VADeR constructs the positive pairs from a discrete set of pixels, our method, SCRL, can possibly sample infinite number of pairs by pooling the variable-sized random regions with bilinear interpolation. This means that VADeR can be viewed as a specific instance of our method where the sizes of the pooled box is fixed to a single pixel without a sophisticated pooling technique. Furthermore, VADeR utilizes an extra decoder architecture, while SCRL exploits an encoder-only structure.

3. Method

This section describes the proposed SCRL for downstream localization tasks in detail. We first present the proposed pair of locally cropped boxes to be matched from different views of the same image. Then, the proposed self-supervised objective based on the spatial consistency loss is defined. The details in implementation of our self-supervised learning are finally presented.

3.1. Spatially Consistent Representation Learning

Motivated by BYOL [14], we use two neural networks: the *online network* which is defined by a set of parameters θ and *target network* parameterized by ξ. The target network provides the regression target to train the online network while the target network's parameter set ξ follows the online network's parameter set θ by using an exponential moving average with a decay parameter τ, *i.e.*, $\xi \leftarrow \tau\xi + (1-\tau)\theta$.

Let $I \in \mathbb{R}^{W\times H\times C}$, $\mathbb{T}_1$ and $\mathbb{T}_2$ denote a training image and two sets of image augmentation strategies, respectively. Our method generates two augmented views $v_1 = t_1(I)$ and $v_2 = t_2(I)$ from I by applying different image augmentations $t_1 \in \mathbb{T}_1$ and $t_2 \in \mathbb{T}_2$. These augmented images (v_1, v_2) are respectively fed into the two encoder networks (f_θ, f_ξ) having the last Global Average Pooling (GAP) layer removed to obtain spatial feature maps $m_1 = f_\theta(v_1) \in \mathbb{R}^{\widetilde{W}\times\widetilde{H}\times\widetilde{C}}$ and the same size of $m_2 = f_\xi(v_2)$.

Unlike previous methods that minimize the global representation distance between aggressive augmented v_1 and v_2 regardless of semantic information, we propose a method to minimize the local representation distance between the two local regions only if they associated with the same spatial regions and thus the same semantic meanings.

To do so, as shown in Figure 2, we first find the intersection regions $\mathcal{IS}(v_1, v_2) \in \mathbb{R}^{\widehat{W}\times\widehat{H}\times C}$ on I, where $\mathcal{IS}(\cdot)$ denotes an operation that generates a spatially corresponding region between v_1 and v_2, and $\widehat{W}$ and $\widehat{H}$ are width and height of it, respectively.

After finding the intersection region, we randomly sample an arbitrary box $B = (x, y, w, h)$ in $\mathcal{IS}$ such that

$$\begin{aligned} w \sim \text{Unif}\,(\widehat{W}_{min}, \widehat{W}),\ \ x \sim \text{Unif}\,(0, \widehat{W} - w), \\ h \sim \text{Unif}\,(\widehat{H}_{min}, \widehat{H}),\ \ y \sim \text{Unif}\,(0, \widehat{H} - h), \end{aligned} \tag{1}$$

where $\widehat{W}_{min} = W/\widetilde{W}$ and $\widehat{H}_{min} = H/\widetilde{H}$ (*e.g.*, we use $\widehat{W}_{min} = \widehat{H}_{min} = 32$ for ImageNet training with ResNet-50 and ResNet-101). Then, to be utilized for spatial representation matching, B has to be translated to the coordinates in each view $v_{i\in\{1,2\}}$, which can be denoted as $B_i = (x_i, y_i, w_i, h_i)$.

Due to aggressive image augmentations, the size, location, and internal color of each box (B_1, B_2) may be different for each views (v_1, v_2), however the semantic meaning in the cropped box area does not change between B_1 and B_2. Here, we crop a local region by a rectangular box. Therefore, in order to exactly map one rectangular box to another rectangular box after geometrical transformations, we exclude certain affine transformations such as shear operations and rotations.

Even without internal color changes, in general, the generated spatial feature maps from conventional CNNs are not coherently changed in scale and internal object translation of an input image [36]. We observe that the previous self-supervised learning methods based on the global consistency loss also have the same limitation in the spatial consistency.

Subsequently, to obtain the equally-sized local representations from B_1 and B_2, we crop the corresponding sample regions, called region-of-interests (RoIs), not on the input images but on the spatial feature maps and locally pool the cropped feature maps by 1x1 RoIAlign [18].

If the number of the boxes that are wanted to be sampled is more than one, we can efficiently obtain multiple pairs of local representations simultaneously by this cropping and pooling on a given spatial feature map, *i.e.*, $p_i^k = \text{RoIAlign}(B_i^k, m_i)$, where $k = \{1, .., K\}$ and K is the total number of generated boxes in an image.

Within the online network, we then perform the projection, $z_1^k = g_\theta(p_1^k)$, from the pooled representation p_1^k, followed by the prediction, $q_\theta(z_1^k)$. At the same time, the target network outputs the target projection from p_2^k such that $z_2^k = g_\xi(p_2^k)$. Our spatial consistency loss is then defined as a mean squared error between the normalized prediction and the normalized target projection as follows

$$\mathcal{L}_\theta^{SCRL} = \frac{1}{K}\sum_{k=1}^{K} \|\overline{q_\theta(z_1^k)} - \overline{z_2^k}\|_2^2, \tag{2}$$

where $\overline{q_\theta(z_1^k)} = q_\theta(z_1^k)/\|q_\theta(z_1^k)\|_2$ and $\overline{z_2^k} = z_2^k/\|z_2^k\|_2$. To symmetrize the loss $\mathcal{L}_\theta^{SCRL}$ in Eq. 2, we also feed v_2 to the online network and v_1 to the target network respectively to compute $\widetilde{\mathcal{L}}_\theta^{SCRL}$ and the total loss is defined as

$$\mathcal{L}_\theta = \mathcal{L}_\theta^{SCRL} + \widetilde{\mathcal{L}}_\theta^{SCRL}. \tag{3}$$

During the self-supervised learning, we only optimize the online network to minimize $\mathcal{L}_\theta$ with respect to θ. It is noted that we follow the BYOL framework for its simplicity in the use of only positive pairs without mode collapse. However, our spatial consistency strategy can be combined with general contrastive learning that also makes use of the negative pairs. We leave the explicit use of negative pairs for future works.

There are lots of possible positive RoI pairs in a single image, and more diversely generated boxes can lead to efficient training as well as performance improvement. Thus, we promote the diversity of box instances by taking overlapped area among them into consideration. In detail, we compute the IoU (Intersection-over-Union) among the sampled RoIs and reject a candidate box if the IoU with previously generated boxes is larger than 50%. We repeat this until the number of survived samples reaches to K. By default, we set $K = 10$. The performance variations according to K and whether the use of IoU thresholding will be presented in Section 4.

3.2. Implementation Details

Dataset For the task of self-supervised pretraining, we make use of 1.2 million training images on ImageNet [7] as unlabeled data.

Image Augmentations SCRL uses the same set of image augmentations in SimCLR [3] and BYOL [14]. We perform simple random cropping [1] with 224×224 resizing, horizontal random flip, followed by a color distortion, and an optional grayscale conversion. Then, Gaussian blur and solarization are applied randomly to the images.

Network Architecture We use a residual network [19] with 50 layers as our base networks f_θ and f_ξ. We also use ResNet-101 as a deeper network. Specifically, the feature map m_i corresponds to the output of the last convolution block in ResNet, which has a feature dimension of (7, 7,

[1] A random patch of the image is selected, with an area uniformly sampled between 20% and 100% of that of the original image, and an aspect ratio logarithmically sampled between 3/4 and 4/3.

pretrain	AP	AP_{50}	AP_{75}
random	32.0	56.7	31.3
supervised-IN	53.2	81.7	58.2
BYOL	55.0	83.1	61.1
SCRL	**57.2**	**83.8**	**63.9**

Table 1. VOC detection using Faster R-CNN w/ FPN, ResNet-50. Supervised-IN denotes the representation trained using image labels on ImageNet (IN) dataset.

2048). After 1x1 RoIAlign [18] of the randomly generated 10 RoIs, a feature dimension of 2048 is fed into projection g_θ that consists of a linear layer with output size 4096 followed by batch normalization, rectified linear units (ReLU) [29], and a final layer with output dimension 256 as in BYOL [14]. The architecture of the predictor q_θ is the same as g_θ.

Optimization We use the same optimization method as in SimCLR [3] and BYOL [14] (LARS optimizer [43] with a cosine learning rate decay [27] over 1000 epochs, warm-up period of 10 epochs, linearly scaled initial learning rate [13]). We use the initial learning rate as 0.45. The exponential moving average parameter τ is initialized as 0.97 and is increased to one during training. We use a batch size of 8192 on 32 V100 GPUs.

4. Experiments

In this section, we elaborate the experiments on transferability to the localization tasks with various pre-trained models including ours. In addition, we conduct extensive ablation studies to understand the key factors in the proposed algorithm.

4.1. Transfer to Localization Vision Tasks

A main goal of representation learning is to learn features that are transferrable to downstream tasks. In this section, we compare SCRL with ImageNet supervised pretraining as well as the state-of-the-art self-supervised learning methods using ImageNet dataset without labels, transferred to various downstream localization tasks on PASCAL VOC [11], COCO [25], and CityScapes [6].

As in MoCo [16], we fine-tune with synchronized BN [32] that is trained, instead of freezing it [19]. We also use SyncBN in the newly initialized layers (*e.g.*, FPN [23]). Weight normalization is performed when fine-tuning supervised as well as unsupervised pretraining models. We use a batch size of 16 on 8 V100 GPUs. Unless otherwise noted, all of the following experiments use the default hyper-parameters introduced in `Detectron2` [42], which are more favorable hyper-parameters for the ImageNet supervised pretraining. Nonetheless, SCRL shows significant performance improvements over the ImageNet supervised pretraining as well as the state-of-the-art unsupervised learning methods. Considering the performance gaps among the previous self-supervised learning methods including supervised image-level pretraining, the obtained performance improvements by SCRL are relatively large and significant across various tasks and networks.

downstream	pretrain	AP	AP_{50}	AP_{75}
FRCNN w/ FPN	random	29.8	48.3	31.8
	supervised-IN	38.5	59.8	41.5
	MoCo v2†	37.1	57.2	40.2
	SimCLR v2†	38.1	58.9	41.3
	SwAV†	39.6	61.3	43.2
	BYOL	40.0	61.3	43.6
	SCRL	**40.9**	**62.5**	**44.5**
RetinaNet w/ FPN	random	24.5	39.1	26.0
	supervised-IN	37.5	57.0	39.9
	MoCo v2†	37.0	55.8	39.5
	SimCLR v2†	37.4	56.5	40.3
	SwAV†	36.7	56.3	39.3
	BYOL	37.7	57.7	40.3
	SCRL	**39.0**	**58.7**	**41.9**

Table 2. COCO detection using ResNet-50. †: We use publicly available checkpoints released by the paper authors. For the upstream self-supervised pretraining task on ImageNet, MoCo v2 and SwAV are trained with *800 epochs* while we run *1000 epochs* for SimCLR v2, BYOL, and SCRL.

4.1.1 PASCAL VOC Object Detection

We first evaluate SCRL on PASCAL VOC [11] object detection task with ResNet-50-FPN [19, 23] and Faster R-CNN [35]. We fine-tune all layers end-to-end with SyncBN [32] and an input image is in [480, 800] pixels during training and 800 at inference. The reported results in Table 1 use the same experimental setting. We use VOC `trainval07+12` as a training set and evaluate on VOC `test2007` set. We evaluate more rigorous metrics of COCO-style AP and AP_{75} including the default VOC metric of AP_{50}.

As shown in Table 1, SCRL substantially outperforms on both supervised pretraining and BYOL, *e.g.*, by 4.0 points and 2.2 points in AP respectively. Moreover, considering that the performance gap is more increased in terms of AP_{75}, SCRL contributes more to the correct box regression than the correct object classification. We conjecture that spatially consistent matching in the upstream task improves its localization performance without any auxiliary technique and results in better box regression.

4.1.2 COCO Object Detection

We evaluate SCRL on COCO object detection task. Faster R-CNN [35] and RetinaNet [24] with FPN [23] are used to measure the transferability within both two-stage and

pretrain	AP	AP_{50}	AP_{75}
random	31.2	49.6	33.7
supervised-IN	40.4	61.6	44.1
BYOL	41.1	62.3	45.0
SCRL	**42.9**	**64.4**	**46.8**

Table 3. COCO detection using Faster R-CNN, ResNet-101-FPN.

pretrain	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
random	28.7	46.9	30.6	30.9	49.7	33.2
supervised-IN	35.4	56.7	38.1	39.0	59.9	42.9
MoCo [16]	35.1	55.9	37.7	38.5	58.9	42.0
VADeR [33]	35.6	56.7	38.2	39.2	59.7	42.7
BYOL	37.2	58.8	39.8	40.4	61.6	44.1
SCRL	**37.7**	**59.6**	**40.7**	**41.3**	**62.4**	**45.0**

Table 4. COCO instance segmentation using Mask R-CNN w/ FPN, ResNet-50: bounding-box AP (AP^{bb}) and mask AP (AP^{mk}) evaluated on `val2017`.

pretrain	COCO keypoints detection			Cityscapes segmentation	
	AP	AP_{50}	AP_{75}	AP	AP_{50}
random	63.2	85.3	68.9	26.2	52.1
supervised-IN	65.7	87.1	71.7	32.7	60.1
BYOL	65.8	87.0	72.0	34.2	62.1
SCRL	**66.5**	**87.8**	**72.3**	**34.7**	**63.6**

Table 5. COCO keypoints detection and Cityscapes instance segmentation using Mask R-CNN w/ FPN, ResNet-50

single-stage object detection frameworks. An input image is in [640, 800] pixels during training and is 800 at inference. We fine-tune all layers end-to-end and train on `train2017` set and evaluate on `val2017` set.

As shown in Table 2, SCRL outperforms with a significant margin not only the supervised pretraining on ImageNet but also the state-of-the-art self-supervised counterparts including MoCo v2 [5], SimCLR v2 [4], SwAV [1], and BYOL [14], in all metrics with both two-stage and single-stage detection architectures.

We also perform evaluation using ResNet-101-FPN with Faster R-CNN. The evaluation results on COCO `val2017` in Table 3 show the same trend even if the encoder network is changed. While the upstream tasks using ResNet-50 are trained with 1000 epochs, both BYOL and SCRL of ResNet-101 are trained with 200 epochs for upstream tasks to measure the relative performance gap.

4.1.3 Other Localization Downstream Tasks

We perform other localization downstream tasks with the default hyper-parameters as in [42] unless otherwise specified.

COCO Instance Segmentation. We fine-tune ResNet-50 model with Mask R-CNN [18]. All experimental settings for Mask R-CNN are same as Section 4.1.2. Similar to COCO detection, SCRL performs the best in all metrics not only AP^{mk} but also AP^{bb} in Mask R-CNN, as shown in Table 4.

COCO Keypoints Detection. We use Mask R-CNN (keypoints version) with ResNet-50-FPN implemented in [42], fine-tuned on COCO `train2017` and evaluated on `val2017`. Table 5 shows that SCRL outperforms in all metrics over the supervised pretraining as well as BYOL.

Cityscapes Instance Segmentation. On Cityscapes instance segmentation task [6], we fine-tune a model with Mask R-CNN. An input image is in [800, 1024] pixels during training and 1024 at inference. Unlike above downstream tasks, we use a batch of 8 which is the default training setting in [42]. As shown in Table 5, SCRL outperforms the supervised ImageNet pretraining and BYOL.

pretrain	# of RoIs	AP	AP_{50}	AP_{75}
BYOL	N/A	38.3	59.7	41.2
SCRL	1	39.2	60.3	42.4
SCRL	5	39.8	61.2	43.4
SCRL	10	40.2	61.3	43.7

Table 6. Performances according to the number of boxes used for spatial matching‡.

# of RoIs	box jitter	IoU thr.	AP	AP_{50}	AP_{75}
1	none	0.5	39.2	60.3	42.4
1	10%	0.5	39.1	60.1	42.5
1	20%	0.5	38.8	60.0	42.2
10	none	0.5	40.2	61.3	43.7
10	10%	0.5	40.0	61.0	43.5
10	20%	0.5	39.6	60.8	43.1
10	∞	0.5	38.8	59.8	42.2
10	none	none	39.8	61.0	43.5

Table 7. Performances according to different box generations in SCRL‡.

‡: We run all experiments with *200 epochs* for upstream task and fine-tune on COCO detection with the default parameters in [42]

4.2. Ablation Studies

In this subsection, we delve deeper into our SCRL by performing various ablation studies. We run all ablations using 1000 epochs during upstream pretraining if not specified, and fine-tuned on COCO detection with Faster RCNN and ResNet-50-FPN.

4.2.1 Number of Boxes Used for Spatial Matching

As shown in Table 6, increasing the number of RoI pairs to be matched between the two views improves the downstream performances. In particular, SCRL outperforms BYOL even when using a single pair for the spatial matching. This ensures a fair comparison between the two meth-

pretrain	AP	AP_{50}	AP_{75}
random	17.8	32.0	17.9
supervised-IN	22.6	38.4	23.5
MoCo v2	20.9	34.8	21.7
SimCLR v2	22.1	37.3	23.0
SwAV	25.5	**43.3**	26.4
BYOL	25.5	42.3	26.9
SCRL	**26.4**	43.2	**28.0**

Table 8. COCO detection with 10% training dataset using Faster R-CNN w/ FPN, ResNet-50.

ods from the perspective of the number of matched pairs per an image. It demonstrates that our method brings more effect than that from simply increasing the number of augmented samples. Another benefit from our approach is that increasing the number of matched pairs from a single image only adds negligible computational costs since they share the same feature map, not requiring multiple forward feature extractions as in multi-crop [1].

4.2.2 Importance of SCRL's Box Generation

To evaluate the benefit of the precise box matching scheme, we perform experiments on what happens when relaxing the exact matching rule with respect to the two cropped regions in a pair. In other words, we randomly jitter the matched boxes to different degrees, where each box can shrink or expand its width and height, and move its position within the margin of a specific percentage. As shown in Table 7, the heavier jittering incurs larger performance degradation, which supports the importance of exact spatial coherence induced to the feature space in SCRL. This is especially compared with BYOL which generates a randomly cropped pair on the input space regardless of its semantic coherence. We also confirm the effect of the technique, IoU thresholding, among the sampled regions in that it prevents the creation of redundant boxes and produces more diverse non-overlapping boxes, which results in better spatial representations.

4.2.3 Representation Power in Small Data Regime

To confirm that the representation itself learned with SCRL contains a lot of more useful information for an object detection task, we evaluate the object detection performance fine-tuned with only 10% of COCO `train2017` dataset while keeping the same hyper-parameters in [42]. As shown in Table 8, SCRL outperforms all the baselines with noticeable margins. This indicates that SCRL provides a more transferable representation that can be harnessed for localization tasks even in a small data regime.

pretrain	LR schedule				
	0.5×	1×	3×	5×	7×
random	23.2	29.8	38.5	40.3	40.9
supervised-IN	35.6	38.5	40.9	41.3	41.4
BYOL	37.0	40.0	41.6	41.7	41.6
SCRL	**38.3**	**40.9**	**42.3**	**42.5**	**42.6**

Table 9. Object detection AP on COCO `val2017` with training schedules from 0.5× (45k iterations) to 7× (630k iterations).

pretrain	upstream epochs	global linear eval. (ImageNet)	RoI linear eval. (COCO GT-boxes)
supervised-IN	90	74.3	72.7
MoCo v2	800	71.1	69.3
SimCLR v2	1000	71.7	71.7
SwAV	800	**75.3**	72.6
BYOL	1000	74.3	71.5
SCRL	1000	70.3	**74.8**

Table 10. Linear evaluation accuracy on ImageNet and COCO GT boxes. Note that, for RoI linear evaluation, we start all over tracking the running statistics of the batch normalization layers during the linear head training to adapt to the unseen distribution of COCO dataset.

4.2.4 Performance on Varied Downstream Schedules

Table 9 illustrates downstream performance under various lengths of the training schedule *i.e.*, 0.5×, 1×, 3×, 5×, 7×, in comparison to other baselines. As discussed in [17], ImageNet pretraining shows rapid convergence than random initialization at the early stage of training but the final performance is not any better than the model trained from scratch. Similarly, BYOL appears to provide a slightly better initial point but the gap between the aforementioned baselines wears off at last. On the other hand, SCRL goes beyond that limit and the noticeable gain is preserved even in longer schedules. We argue that SCRL provides task-specific representations of quality that the previous pretraining methods have not yet achieved. It also implies that one should rethink, in fact, not the ImageNet pretraining itself but the right way of doing it to the specific downstream task.

4.2.5 RoI Linear Evaluation using GT Boxes

In this section, we propose a simple protocol for localization tasks, similar to the linear classification evaluation, but focusing only on the region-of-interests given by the ground-truth boxes of localization datasets. Specifically, we train a single linear layer, on the top of the frozen backbone followed by a RoIAlign layer delivering the localized features of ground-truth boxes to the learner[2].

[2]Considering the small-sized object, 800×800 resized-images are used to train a single linear layer. We only perform horizontal random flip and use a batch size of 512 on 8 V100 GPUs. Initial LR is 0.1 with a cosine learning rate decay over 80 epochs, warm-up period of 5 epochs.

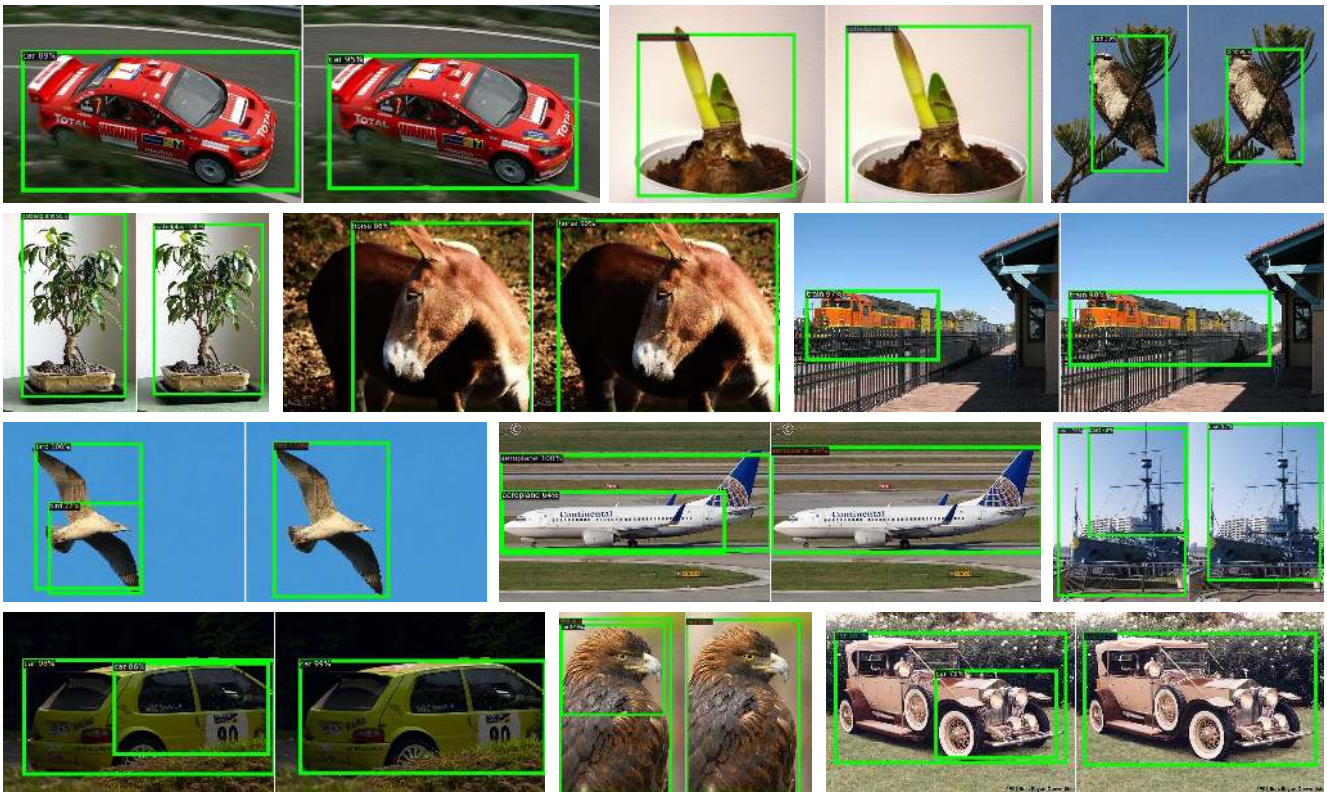

Figure 3. Qualitative comparison between BYOL (left) and SCRL (right) on PASCAL VOC detection w/ Faster R-CNN, ResNet-50-FPN. The first two rows show the regression capability of bounding boxes. The latter two rows present the malfunction of BYOL.

Table 10 shows linear evaluation results, using the pre-GAP features of ResNet-50, under the both standard and the proposed protocol. SCRL outperforms other baselines on our protocol. Interestingly, trade-off is observed between the two protocols when the performance is saturated, which suggests that, for object detection tasks, learned representations have to be linearly separable under spatial constraints rather than in a global fashion.

4.3. Qualitative Analysis

We found that the performance gain of our method comes more from the box regression than the category classification as we described in 4.1.1. To understand exactly what aspect of the regression is improved, we scrutinize the detection result in qualitative perspective. Figure 3 illustrates the detected boxes with correct class prediction, where the left and right figure of each pair represent the outcomes from the model having been initialized with BYOL and SCRL, respectively.

The first two rows are the cases in which SCRL predicts precise and tight box boundaries while BYOL fails to do so. The latter two rows represent the ones where BYOL detects a smaller box embracing only a mere *part* of the entire object and confidently predicts it as a whole whereas SCRL successfully resists it.

We conjecture that this misbehavior of BYOL pretraining attributes to its translation-invariant representations by matching features between two randomly cropped views, while SCRL does learn the features sensitive to positional variation. Both methods learn scale-invariant features but significantly different in that BYOL has a limited vision of already-cropped and resized-to-the-same-scale patches while SCRL sees the patches on the feature map along with the global context thanks to the peripheral vision from receptive field. Thereby, it may help SCRL to capture a full view of objects considering entire context regardless of its size and position.

5. Conclusion

In this paper, we have presented a novel self-supervised learning algorithm, SCRL, that tries to produce consistent spatial representations by minimizing the distance between spatially matched regions. By doing so, the proposed SCRL outperforms, with a significant margin, the state-of-the-art self-supervised learning methods on various downstream localization tasks.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020.

[2] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. *CoRR*, abs/2006.10029, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223. IEEE Computer Society, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017.

[9] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *NeurIPS*, 2019.

[10] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9):1734–1747, 2016.

[11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR (Poster)*. OpenReview.net, 2018.

[13] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020.

[15] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, pages 3279–3286. IEEE Computer Society, 2015.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. IEEE, 2020.

[17] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, pages 4917–4926. IEEE, 2019.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[20] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[21] Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, pages 3253–3261. IEEE Computer Society, 2016.

[22] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 577–593. Springer, 2016.

[23] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944. IEEE Computer Society, 2017.

[24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007. IEEE Computer Society, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[26] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

[27] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net, 2017.

[28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6706–6716. IEEE, 2020.

[29] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814. Omnipress, 2010.

[30] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV (6)*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016.

[31] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. In *NeurIPS*, pages 6237–6247, 2018.

[32] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *CVPR*, pages 6181–6189. IEEE Computer Society, 2018.

[33] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations, 2020.

[34] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *CoRR*, abs/2007.13916, 2020.

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[36] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection SNIP. In *CVPR*, pages 3578–3587. IEEE Computer Society, 2018.

[37] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, pages 6360–6370. IEEE, 2019.

[38] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, pages 844–855, 2017.

[39] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[40] Nikolai Ufer and Björn Ommer. Deep semantic feature matching. In *CVPR*, pages 5929–5938. IEEE Computer Society, 2017.

[41] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[43] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017.

[44] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, pages 4353–4361. IEEE Computer Society, 2015.