

# Self-Supervised Visual Representations Learning by Contrastive Mask Prediction

Yucheng Zhao<sup>\*†1</sup>Guangting Wang<sup>\*‡1</sup>Chong Luo<sup>2</sup>Wenjun Zeng<sup>2</sup>Zheng-Jun Zha<sup>‡1</sup>University of Science and Technology of China<sup>1</sup>Microsoft Research Asia<sup>2</sup>

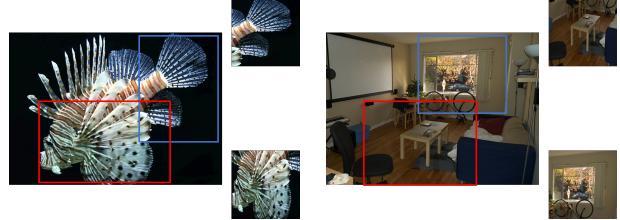
{lnc, flylight}@mail.ustc.edu.cn {cluo, wezeng}@microsoft.com zhazj@ustc.edu.cn

## Abstract

*Advanced self-supervised visual representation learning methods rely on the instance discrimination (ID) pretext task. We point out that the ID task has an implicit semantic consistency (SC) assumption, which may not hold in unconstrained datasets. In this paper, we propose a novel contrastive mask prediction (CMP) task for visual representation learning and design a mask contrast (MaskCo) framework to implement the idea. MaskCo contrasts region-level features instead of view-level features, which makes it possible to identify the positive sample without any assumptions. To solve the domain gap between masked and unmasked features, we design a dedicated mask prediction head in MaskCo. This module is shown to be the key to the success of the CMP. We evaluated MaskCo on training datasets beyond ImageNet and compare its performance with MoCo V2 [4]. Results show that MaskCo achieves comparable performance with MoCo V2 using ImageNet training dataset, but demonstrates a stronger performance across a range of downstream tasks when COCO or Conceptual Captions are used for training. MaskCo provides a promising alternative to the ID-based methods for self-supervised learning in the wild.*

## 1. Introduction

Self-supervised learning (SSL) of visual representation has been a great success in recent years [11, 3, 10, 2, 1], facilitating a wide range of downstream tasks, including both visual tasks [13, 22, 12, 15, 29], and cross-modality tasks [34, 31]. Recently, several contrastive learning based SSL methods have demonstrated strong results on downstream image classification task [10], closing the performance gap between them and supervised learning methods. As SSL methods do not require human annotation of training datasets, people are optimistic about the day when they surpass supervised feature learning methods when virtually



(a) An example image that satisfies the SC assumption. (b) An example image that does not satisfy the SC assumption.

Figure 1. Illustration of semantic consistency (SC) assumption by visualizing two cropped views of an image. (a) An image from the ImageNet dataset satisfies the SC assumption, as the two cropped views sample different parts of the same fish. (b) An image from the COCO dataset does not satisfy the SC assumption, as the two cropped views contain different objects belonging to different semantic categories.

unlimited data are made use of.

The state-of-the-art (SOTA) SSL methods for visual feature learning are mostly based on a pretext task called instance discrimination (ID) [28]. The idea is illustrated in Figure 2. The intuition behind is that we can learn a good feature representation by merely asking the feature to be discriminative of individual instances. To realize this idea, SOTA SSL methods [11, 3] optimize a contrastive loss function which enforces the model to output sufficiently close representations for different views of a single image. Since semantic information is a principal component in visual features, an implicit assumption here is that a training image should have consistent semantic meaning across views, which we call semantic consistency (SC) assumption.

The SC assumption is almost always satisfied in the ImageNet dataset [5], which exhibits an object-centric bias [21]. An example image from the ImageNet dataset and its two cropped views are shown in Figure 1(a). The two crops sample different parts of the same fish, so it is quite reasonable to let them have similar feature representations. However, if the ID-based SSL methods are to be extended to datasets beyond ImageNet, they will need to handle images

\* Equal contribution. † Interns at MSRA. ‡ Corresponding author.

like the one as shown in Figure 1(b). The two cropped views of this example image contain entirely different objects belonging to different semantic categories. Is it still reasonable or helpful to enforce similar feature representations for such crops? In other words, are ID-based SSL methods readily applicable to the unconstrained datasets without SC guarantee?

The initial studies are not encouraging. Purushwalkam *et al.* [21] notice that the ID pretext task takes advantage of the object-centric bias in the ImageNet dataset. In addition, some preliminary work [11, 10, 17] tries to directly extend ID-based SSL methods from ImageNet to other datasets, including Instagram-1B [16], Places365 [35], and YFCC100M [24], but they do not get satisfactory results. These facts cast a shadow over the future of ID-based SSL methods. After all, being able to make effective use of large amounts of unconstrained data is the most competitive feature of SSL methods.

We are therefore motivated to explore alternative pretext tasks which do not rely on the SC assumption. A task that once achieved great success in natural language pre-training has entered our sight. It is the mask prediction task, also known as the masked language model (MLM) in BERT [6]. In the vision domain, image inpainting is also a mask prediction task and has been used as a pretext task for visual representation learning [19]. However, image inpainting is a generative method operating in the pixel space. As Grill *et al.* pointed out, such method is computationally expensive, and the high level of detail required for image generation may not be necessary for representation learning [10].

In this paper, we propose to use contrastive mask prediction as a pretext task for self-supervised visual representation learning. The idea is illustrated in Figure 2 (bottom). We make the task contrastive so that the extracted features can focus on high-level semantic meanings instead of pixel-level details. In order to realize this idea, we design a novel SSL method named Mask Contrast (MaskCo). In the design of MaskCo, we answer two basic questions as to what features to contrast and how to bridge the domain gap between features of masked and unmasked regions. First, we propose to use region-level features instead of view-level features to compute contrastive loss. This choice is the key to get rid of the SC assumption because we compare the masked and unmasked versions of the exact region instead of comparing two views. Second, as a domain gap exists between the predicted features from the masked and unmasked view, we insert a mask prediction head (MPH) into the network to bridge the gap. Both quantitative and qualitative results have demonstrated that MPH is an indispensable component in MaskCo.

Evaluations of MaskCo on a range of datasets show promising results. While MaskCo achieves comparable performance with MoCo V2 [4] when using ImageNet as the

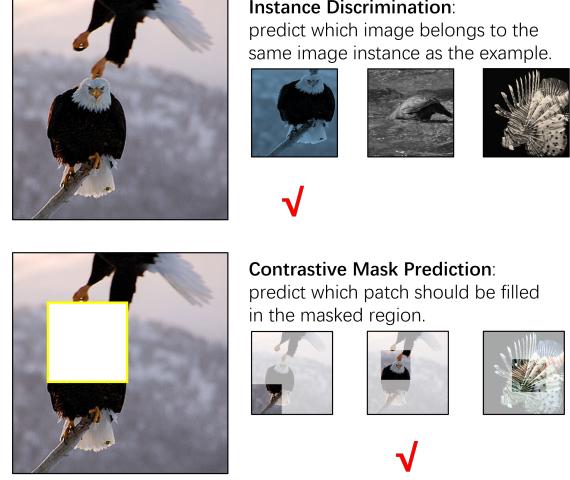


Figure 2. Illustration of pretext tasks for SSL: the commonly used instance discrimination task and the proposed contrastive mask prediction task.

pre-training dataset, it achieves much stronger performance than MoCo on multiple downstream tasks when the unconstrained datasets COCO [14] or Conceptual Captions [23] are used for pre-training. These results verify that MaskCo has relaxed the SC assumption, and has high potential to be used for learning image representations in the wild.

## 2. Background and Related Work

In this section, we first provide background on self-supervised visual representation learning and the instance discrimination task. Then we discuss related work on pre-training beyond ImageNet and the context prediction task.

### 2.1. SSL and the ID task

SSL methods try to learn generalizable and transferable representations from unlabeled data using pretext tasks that provide free supervision. With well-designed pretext tasks [7, 18, 19, 32, 1, 8, 28], current SSL methods [2, 10] are able to achieve comparable performance on a range of downstream tasks with their supervised counterparts. The most attractive feature of SSL is its potential to make use of unlimited data. The community are generally optimistic about its future performance when a much larger dataset than ImageNet is used for training.

Recently, the instance discrimination (ID) task [28] has dominated SSL for its strong performance on almost all testing benchmarks [11, 3, 2, 10, 4]. This task aims to discriminate each image instance by viewing an image as a distinct class of its own. The common practice to solve the ID task is to use contrastive learning. Chen *et al.* [3] proposed a simple framework for contrastive learning, named SimCLR. It learns representations by contrasting images after the com-

position of data augmentations, showing that SOTA performance can be achieved by contrastive learning without specialized architectures or a memory bank. MoCo [11] and its improved version MoCo V2 [4] proposed to use momentum encoder to release the requirement of large batch-size or memory bank. More recently, BYOL [10] removed negative samples from contrastive learning by iteratively bootstrapping the outputs of a network to serve as targets for an enhanced representation. Meanwhile, ID task is also extended to pixel-level contrastive learning [30, 26, 20] for dense prediction tasks. These methods, however, are beyond the scope of our discussion as they generally do not perform well in image classification task.

## 2.2. Pre-training beyond ImageNet

Although ID-based SSL methods have achieved revolutionary progress, their success is mainly confirmed on the ImageNet pre-training dataset, a labeled and curated dataset. Purushwalkam *et al.* [21] noticed that the advance of current ID-based SSL methods moderately comes from their usage of dataset bias of ImageNet. They also found that training MoCo on the less-biased COCO dataset does not get encouraging results. Xiao and Michael [33] also reported degraded performance when training MoCo on COCO and Pascal datasets.

Other work [2, 11] reported preliminary results on non-ImageNet datasets. Using larger uncurated datasets of Internet images, directly applying current SSL methods has shown marginal gains, even though the dataset size is orders of magnitude larger than ImageNet. Furthermore, the current SOTA method BYOL [10] even shows a significant performance drop when replacing the pre-training dataset from ImageNet to Place365. While it is still not clear whether the performance drop is due to the domain gap between pre-training and downstream datasets, it is clear that the ID task makes a strong SC assumption on the training images, which can hardly hold in unconstrained datasets.

## 2.3. Context prediction tasks

In this paper, we propose a novel pretext task named Contrastive Mask Prediction. It is inspired by the MLM in BERT [6] and has a strong correlation with some context-based prediction tasks [7, 18, 19] in visual representation learning. BERT [6] is a milestone pre-training method in natural language processing. In the core of BERT is the mask prediction task, also known as the mask language model (MLM). From the input sentence, some tokens are randomly masked out, and the training objective is to predict the vocabulary ID of the masked word based on its context. MLM cannot be directly extended to handle images as images are continuous signals without a finite dictionary.

In the computer vision domain, RelativePosition [7] is a context-based pretext task whose goal is to predict the

relative position between two image patches. The authors hypothesized that doing well on this task requires the understanding of scenes and objects. Following this idea, Noroozi *et al.* [18] proposed a self-supervised learning method based on solving jigsaw puzzles, reducing the ambiguity of the pretext task, and achieving better performance on multiple downstream tasks. However, these tasks have relied on heuristics which could limit the generality of the learned representations [3].

Image inpainting [19] is also used as a pretext task in SSL by hypothesizing that deep semantic understanding is essential to recover the missing content. As we have pointed out earlier, this work falls into the category of generative methods. In this work, we embrace the contrastive methods and the reasons will be detailed in the next section.

## 3. Learning by Contrastive Mask Prediction

In this section, we first introduce the CMP task, which relaxes the SC assumption in SSL and enables the representation learning in the wild. Then we describe our approach called mask contrast (MaskCo), which concretely implements the CMP task. Last, the implementation details are presented.

### 3.1. The Contrastive Mask Prediction Task

The mask prediction task explores the data structure through the context correlation. Given a natural signal, a small region is masked out before it is processed by a neural network model. The training objective is to obtain a feature representation of the input signal, from which the exact signal of the masked region or its feature can be predicted. The mask prediction task involves a fundamental assumption, that is, correlation exists between content and its context, which we believe is generally valid for natural signals, including language and images. However, the prediction process is easy to implement for tokenized data like text but not so straightforward for continuous data like images. To adapt the mask prediction on the image, we modify the objective from predicting a permanent category ID to distinguishing the true candidates from distractions and implement it by the contrastive loss. We thus name the modified task as contrastive mask prediction (CMP).

An illustration of CMP is shown in Figure 2 (bottom). The image contains a main object which is a bird, and the bird’s top body is masked out. Looking at the masked image, people can reasonably guess that the masked region should include a bird’s head. We come up with this guess by looking at the surrounding context where a twig and the bird’s legs appear. This ability comes from understanding the semantic content, and it is indeed what we want a pre-trained representation to possess. However, it will be hard for a person to draw the exact pixel of the missing part since the details of a bird could vary. Likewise for a network, fill-

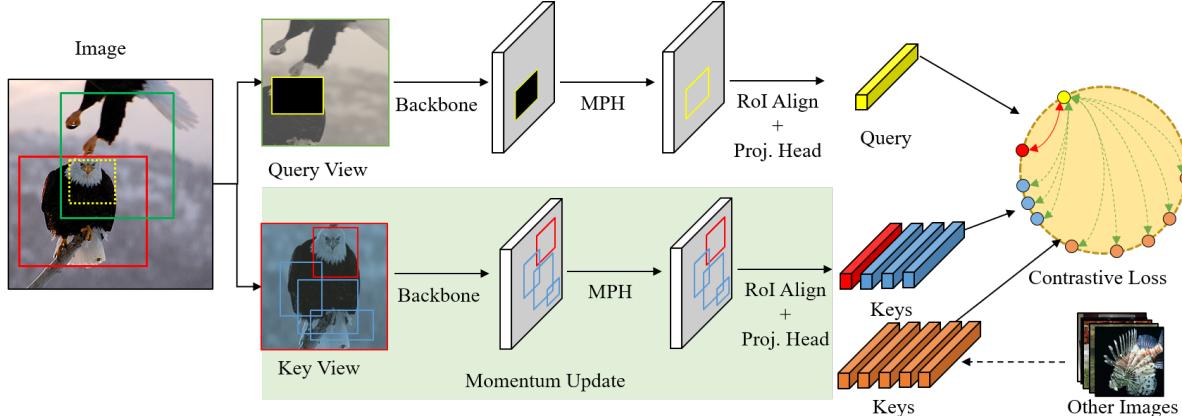


Figure 3. Illustration of **Mask Contrast (MaskCo)**: Given a training image, we first acquire two views by data augmentations. A selected region is masked out in the query view while the key view stays unaltered. We obtain the query and key features by a forward pass through the backbone network, the mask prediction head (MPH), and the RoI align layer with a projection head. We perform region-level contrastive learning, in which the query feature for the masked region is pulled closer to the positive key feature, and pushed away from the negative key features. Most negative keys are from other images while we could also include a few negative keys from other regions of the same image.

ing in the missing part in the pixel-level will be computationally expensive and unnecessary. This is why we choose to use a contrastive version of mask prediction instead of a generative version as used in previous work [19].

### 3.2. The Mask Contrast Approach

This subsection describes the Mask Contrast (MaskCo) approach we developed to realize the CMP task. Figure 3 provides an overview of the approach. Similar to MoCo, we adopt a Siamese network structure in pre-training, the online network processes the query view and the target network processes the key view. The parameters in the online network are updated through back-propagation, while those in the target network are updated with momentum.

The first prominent feature of MaskCo is that we perform region-level contrast instead of view-level contrast. The region refers to the masked region in the image and all of the positive and negative samples. While this idea is quite straightforward to the mask prediction task, we would like to emphasize that it is the key to release the SC assumption used in ID tasks. This is because, for the predicted feature of the masked region, the representation of the same region drawn from an unmasked view is an indisputable positive sample that does not depend on any hypothesis.

Given an image instance, we first create query view and key view through random cropping and other data augmentation techniques. The constraint here is that the two views should have a sufficiently large overlap. Then, a masked region is identified within the overlap of the two views. In the query view, the selected region is masked to zero. In the key view, the selected region is the only positive key box. We additionally pick several negative key boxes, which have a

small overlap with the positive key box. More negative key boxes are sampled from other images. The query view and the key view are passed through the backbone network, and the region-level feature is extracted by RoI align operation after the features pass through an additional mask prediction head (MPH). The final feature to compute contrastive loss is the RoI feature passing a non-linear projection head [4, 3].

The second prominent feature of MaskCo is the design of MPH. We rely on MPH to bridge the domain gap between features of masked and unmasked regions. In most of the downstream tasks, the backbone network is used to acquire representations for unmasked images. Therefore, it does not need to have the filling-in-the-blank capability. However, in the pre-training stage, we contrast the feature of the masked region with all other features of unmasked regions. The network will naturally learn how to do mask prediction. There is a domain gap between the generally useful visual features (of unmasked regions) and the task-specific features (of masked regions). We therefore design a specialized module for mask prediction and insert it between the backbone network and the RoI align operation. The MPH is simply implemented by a few convolution layers. Later, our visualization and quantitative evaluation in Section 4.4 will confirm its effectiveness.

We use contrastive loss [11, 25] as our training objective, which can be thought of as a *dictionary look-up* task. Consider an encoded feature  $q$  as the query and a set of encoded features  $\{k_i | i = 0, \dots, K\}$  as keys of a dictionary. Assume there is a single key (denoted as  $k^+$ ) that matches the query  $q$ . The contrastive loss is a function whose value is low when  $q$  is similar to the positive key and dissimilar to

all other keys. With similarity measured by L2-normalized dot product, the loss function can be formulated in the following:

$$L = -\log \frac{\exp(q \cdot k^+ / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)} \quad (1)$$

where  $\tau$  is a temperature hyper-parameter [3, 11]. This form is the exact one appeared in [11], while it can also take other forms [3, 25].

In MaskCo, the query  $q$  is the region-level feature of the masked box, the positive key  $k^+$  is the feature of the corresponding region in the key view. The negative keys  $k_i | i = 0, \dots, K, k_i \neq k^+$  are the feature vectors of negative key boxes drawn from other images or the same image. It is worth noting that contrasting region-level features allows us to include negative samples within the same image. This is not possible in ID-based methods because they use view-level features. We conjecture that negative keys from the same image can be viewed as hard negatives which might enhance the localization capability of the trained features. We therefore present two variants of MaskCo. The default version only uses inter-image negatives and the MaskCo(+in) uses both inter-image and intra-image negatives. We show that MaskCo(+in) has a small advantage in some detection benchmarks.

### 3.3. Implementation Details

Next we provide implementation details of MaskCo.

**Data augmentation details:** The global image augmentation on the key view and the query view is the same as MoCo v2 [4], except the random crop is constrained as described in Section 3.2. Given an image, a random region is cropped with at least 20% of the image and resized to  $224 \times 224$  with a random horizontal flip, followed by a random color jittering related to brightness, contrast, saturation, hue, and grayscale. Gaussian Blur is also used for augmentation. After global augmentation, we randomly generate 17 bounding boxes in the key view with the size from  $32 \times 32$  to  $128 \times 128$ . The first one is viewed as the positive key box, and others are viewed as the negative key boxes. The projection of the positive key box in the query view is viewed as the masked box. We enforce the negative key boxes in the generation process to have an IoU lower than 0.2 with the positive key box.

**Model details:** We use the OpenSelfSup<sup>1</sup> codebase to implement MaskCo. The backbone model is standard ResNet-50 [13]. The mask prediction head is three 3-layer blocks of residual layers ( $1 \times 1$  conv- $3 \times 3$  conv- $1 \times 1$  conv), which is the same as conv5 layer in the ResNet-50. We use the same projection head as MoCo v2 [4].

**Training details:** The training epoch is set to 100 in most experiments on ImageNet dataset, and set to the same

<sup>1</sup> <https://github.com/open-mmlab/OpenSelfSup>

iterations on other datasets. For comparison with other methods on ImageNet (in Table 3), we set the pre-training epoch to 200. Most training hyper-parameters are the same as MoCo V2 [4]. Specifically, we use the SGD optimizer and set the base learning rate as 0.03, momentum as 0.9, weight decay as 0.001, and total batch size as 256 distributed across 8 V100 GPU. The cosine learning rate schedule is also used. The training details for the downstream tasks can be found in Appendix.

## 4. Experiments

### 4.1. Experiment Setup

We use four datasets in our experiment. The first one is the COCO dataset [14], which has 328k images and is commonly used for complicated scene understanding tasks, including object detection and semantic segmentation. The second one is the Conceptual Captions (CC) dataset [23], which is designed for image captioning. CC dataset contains 3.3 million images acquired from the Internet. Both COCO and CC do not exhibit the object-centric bias as ImageNet. For completeness and comparison with other SSL methods, we also include ImageNet (IN) dataset [5], which has 1.3 million images belonging to 1000 classes. IN is well-balanced in its class distribution and each image generally contains iconic views of objects. The last dataset is ImageNet 10% (IN-10%), which is constructed by sampling 1/10 of the images per class on the original ImageNet.

We use four downstream tasks for evaluation.

**ImageNet Linear Classification:** We follow the linear evaluation protocol [9]. When we train on the ImageNet dataset, the pre-trained model is fixed, and only the additional linear classification layer is fine-tuned. We report the top-1 accuracy of pre-trained features from conv4 and conv5, since most pretext tasks achieve the best performance on these two layers. The full results can be found in the Appendix.

**Pascal VOC Object Detection:** We follow the setting introduced in MoCo [11]. A Faster R-CNN detector [22] with the ResNet50-C4 backbone is implemented in Detectron2 [27]. We fine-tune all layers end-to-end and synchronize all batch normalization layers. The fine-tuning is performed on *trainval07+12* set (16.5k images) and the evaluation is on *test2007* set. We report evaluation metrics AP, AP50, and AP75.

**COCO Object Detection and Instance Segmentation:** We follow the setting introduced in MoCo [11], where a Mask RCNN detector with the ResNet50-C4 backbone is used, implemented in Detectron2 [27]. In the optimization, we fine-tune all layers end-to-end and synchronize all batch normalization layers. We use the 2x schedule in Detectron2. The fine-tuning is performed on *train2017* set (118k images) and evaluation is on *val2017* set. In the evaluation,

Dataset	IN	IN-10%	COCO	CC
ResNet Dist.	0.1546	0.1547	0.1814	0.1782

Table 1. ResNet distance for multiple pre-training datasets. The distances of non-ImageNet datasets are significantly larger than ones in ImageNet dataset.

we report AP, AP50, and AP75 for detection as well as AP, AP50, and AP75 for segmentation.

## 4.2. Dataset Analysis

Previous work [21] has suggested that ID-based SSL methods take advantage of the object-centric bias of the ImageNet dataset. This is highly related to the semantic consistency assumption we pointed out. In order to better understand this problem, we define an objective metric, called the *ResNet distance*, to reflect the semantic consistency in an image. The major tool we use is the PyTorch official ResNet-50 model. Specifically, the model is pre-trained with the image classification task on the ImageNet dataset in a supervised way.

Given an image, we generate two views  $x_1, x_2$  using exactly the same random crop operation as in MoCo, and compute the ResNet distance as follows:

$$d(x_1, x_2) = \|\hat{y}^1 - \hat{y}^2\|^2, \quad (2)$$

where  $\hat{y}_1$  and  $\hat{y}_2$  are conv5 features, with unit-normalization in the channel dimension, of  $x_1$  and  $x_2$  generated by the aforementioned ResNet-50 model.

We calculate the average ResNet distance for various pre-training datasets, including ImageNet, ImageNet 10%, COCO and CC. We can find from Table 1 that the ResNet distances of COCO and CC are significantly larger than those of ImageNet datasets, suggesting that the SC assumption does not hold well on non-ImageNet datasets.

Furthermore, we show that the performance of ID-based SSL methods drops significantly on pre-training datasets with a larger ResNet distance. In order to eliminate the dataset size factor, we carry out the comparison using COCO dataset and ImageNet 10% dataset, whose sizes are almost identical. Three SOTA ID-based SSL methods, including NPID, SimCLR, and MoCo v2, are evaluated. The results are shown in Figure 4. We can find that all ID-based methods have significant performance drop when transferring to non-ImageNet datasets.

## 4.3. Main Results

We provide our experimental results for object detection, image classification, and instance segmentation tasks, compare MaskCo’s performance with state-of-the-art SSL methods. All compared methods use the same training settings and are implemented by the OpenSelfSup.

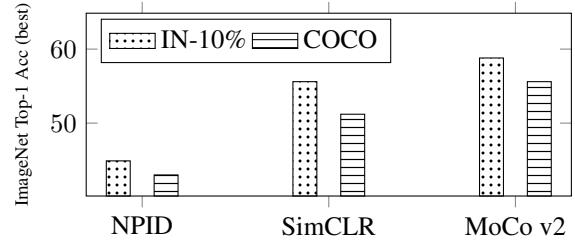


Figure 4. Pre-training three ID-based SSL methods on IN-10% and COCO. The ImageNet linear classification top-1 accuracy from best layers are reported. Significant performance drop is observed when training on COCO.

The main advantage of MaskCo over ID-based methods is that it does not rely on the SC assumption. In order to verify this, we train MaskCo and MoCo v2 using different pre-training datasets and compare their performance on VOC detection and ImageNet linear classification. The results are shown in Table 2. When pre-trained on ImageNet, MaskCo achieves slightly lower performance on three benchmark tasks. However, for the other two pre-training datasets, MaskCo achieves consistently higher performance than MoCo v2. For COCO dataset, which has the largest ResNet distance among the three datasets, MaskCo outperforms MoCo v2 with 0.6 AP on Pascal detection, and 1.1% on ImageNet linear classification (Comparison based on the best result from conv4 and conv5). For Conceptual Captions, which is another unconstrained dataset, MaskCo achieves a similar performance gain over MoCo on all three tasks.

We noticed that although MaskCo shows relative advantage over MoCo v2 on COCO and CC datasets, there is a performance gap between the model trained on COCO/CC and the model trained on ImageNet. One possible reason is that we keep all the hyper-parameters tuned on ImageNet for a fair comparison among methods, but this is definitely unfair when it comes to the performance comparison among pre-training datasets. In the future, we plan to explore automatic hyper-parameter adjustment policy for unconstrained dataset, but it is beyond the scope of this work.

Besides, although CC is a much larger dataset than COCO, the results of MaskCo trained on CC do not show much advantage over the results trained on COCO in Table 2. This is reasonable because we are showing the results achieved at the same iteration steps. To make the point clear, we conduct side-by-side experiments on these two datasets and show the results at multiple training iterations from 200k to 800k (the default setting reported in Table 2 uses 500k iterations). As shown in Figure 5, training on COCO starts to have a saturated performance after 600k iterations, but training on the larger CC dataset does not show any sign of saturation. The performance still grows almost linearly at 800k iterations. It matches the conventional wis-

Dataset	Method	VOC			ImageNet	
		AP	AP <sub>50</sub>	AP <sub>75</sub>	conv4	conv5
IN	MoCo v2	<b>56.4</b>	<b>82.0</b>	62.8	<b>60.0</b>	<b>64.4</b>
	MaskCo	56.0	81.3	<b>63.1</b>	57.6	63.1
COCO	MoCo v2	55.0	80.7	60.9	55.6	51.4
	MaskCo	<b>55.6</b>	<b>81.0</b>	<b>61.5</b>	<b>55.7</b>	<b>56.7</b>
CC	MoCo v2	55.1	80.8	60.6	<b>56.0</b>	56.5
	MaskCo	<b>55.7</b>	<b>81.0</b>	<b>61.5</b>	55.4	<b>57.1</b>

Table 2. Comparison of MoCo v2 and MaskCo on multiple pre-training datasets. Evaluations are conducted on Pascal VOC object detection and ImageNet linear classification. All evaluated models use the same training steps, which is about 500k (equal to 100 epoch on ImageNet).

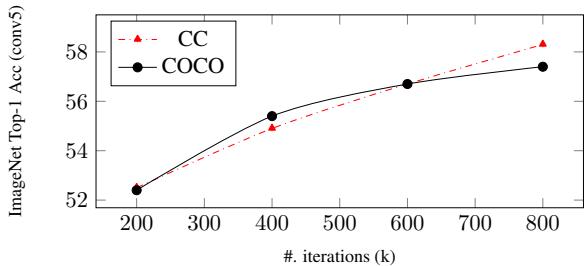


Figure 5. Comparison of pre-trained MaskCo with ImageNet linear classification top-1 accuracy across different pre-training iterations on COCO and CC datasets. Pre-training on COCO gets saturated performance after 600k iterations, while the model pre-trained on CC can continuously improve.

dom that training on larger datasets needs more iteration steps to converge.

Last, we present a comparison between MaskCo and most SOTA SSL methods in Table 3. The pre-training dataset is ImageNet, so it is not a setting to show the advantage of MaskCo. But even in such a setting, MaskCo manages to outperform most other SSL methods on the detection and segmentation tasks. Only on the ImageNet classification task, MaskCo is slightly inferior to MoCo v2 and BYOL. Another variant of MaskCo, which uses the intra-image negative samples, achieves an AP of 57.2 on Pascal VOC, which is slightly better than MaskCo. But on the image classification task, it incurs some losses.

#### 4.4. Ablation Study

We conduct ablation studies using the Pascal VOC object detection and ImageNet linear classification tasks. The pre-training dataset is ImageNet.

**Effect of Mask Prediction Head:** We evaluate the effectiveness of the mask prediction head introduced in Section 3.2. MPH is implemented by three 3-layer blocks of residual layers in the default setting, and here, we vary the number of 3-layer blocks to inspect this design. In Table 4, the first row reports the results when the block number is

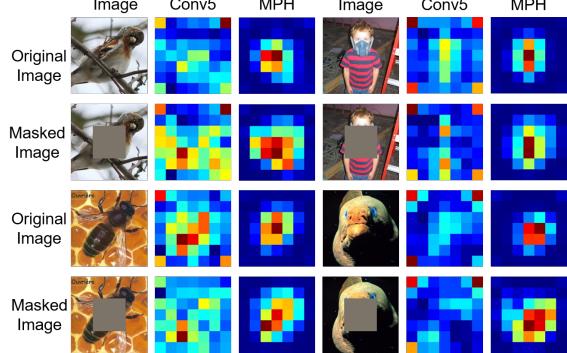


Figure 6. Visualization of MPH. Given an image and its masked counterpart, we visualize their feature maps of conv5 and MPH. The features in the masked region is noticeable low in conv5 while being recovered in MPH. More visualization results can be found in Appendix.

0, or no MPH is used. Compared with other rows that use MPH, we can notice a significant performance drop in the ImageNet classification task. This confirms our conjecture that a domain gap exists between the image classification task and the CMP pretext task. Another observation is the object detection APs on Pascal VOC are not influenced by MPH. We believe this is caused by the architecture of the detector we used, whose backbone ends with conv4. Other results with {1, 2, 3, 4} 3-layer blocks show a trend that a sufficiently large network is necessary for filling the domain gap, and three blocks are enough in our scenario.

We also visualize features before and after MPH to demonstrate the impact of MPH. We feed example image pairs into a pre-trained MaskCo model and visualize the feature maps of conv5 and MPH. In an image pair, the first is the original image, and the second is a masked image where the 64x64 region in the center is masked by zero. We find the features corresponding to the masked region in conv5 exist noticeable low response while they are recovered to a certain extent after MPH.

**Effect of Mask:** We verify that the masking operation defined in CMP is essential to learn features that transfer well. Without the mask, the pretext task will degenerate to predict a region-level feature from the exact same region with non-geometric data augmentations. This is somewhat equal to the instance discrimination task removing random cropping. Intuitively, it is an effortless task and may not drive the model to learn semantically meaningful representations. In our practice, removing the mask indeed weakens the pretext task, makes the loss decrease from 2.51 to 1.83, and causes performance drop in the downstream tasks, as shown in Table 5.

**Negative Sampling Strategy:** We design a novel negative sampling strategy in Section 3.2, where not only the inter-image negatives are taken into account but intra-image negatives may also be included. The inner-image negatives

Method	Pascal VOC			MS COCO			MS COCO			ImageNet	
	$AP$	$AP_{50}$	$AP_{75}$	$AP^{bb}$	$AP_{50}^{bb}$	$AP_{75}^{bb}$	$AP^{seg}$	$AP_{50}^{seg}$	$AP_{75}^{seg}$	conv4	conv5
Rand Init	32.8	59.0	31.6	35.6	54.6	38.2	31.4	51.5	33.5	9.1	6.5
Supervised	54.2	81.6	59.8	40.0	59.9	43.1	34.7	56.5	36.9	67.6	76.2
Relative-Pos [7]	55.1	80.4	61.2	40.0	59.6	43.5	35.0	56.5	37.3	49.3	40.2
Rotation-Pred [8]	55.5	80.9	61.4	40.0	59.3	43.6	34.9	56.0	37.4	55.0	49.1
NPID [28]	54.1	80.0	59.5	39.4	59.0	42.8	34.5	55.9	36.6	57.7	51.7
MoCo v2 [4]	<b>57.0</b>	<b>82.2</b>	<u>63.4</u>	<b>41.0</b>	<b>60.6</b>	<b>44.5</b>	<b>35.6</b>	<b>57.2</b>	<b>38.0</b>	<u>61.6</u>	<u>66.7</u>
SimCLR [3]	51.5	79.4	55.6	39.6	59.1	42.9	34.6	55.9	<u>37.1</u>	54.4	61.6
BYOL [10]	49.0	49.6	52.8	40.2	<b>60.6</b>	43.3	34.9	57.0	36.7	<b>62.8</b>	<b>71.6</b>
MaskCo	<b>56.7</b>	<b>82.1</b>	<b>63.9</b>	<b>40.8</b>	<b>60.5</b>	<b>44.2</b>	<b>35.5</b>	<b>57.1</b>	<b>38.0</b>	59.6	65.1

Table 3. Comparison of MaskCo to previous supervised/self-supervised pre-training methods on ImageNet pre-training datasets. Evaluations are conducted on Pascal VOC object detection, COCO object detection and instance segmentation, and ImageNet linear classification. All models are pre-trained with 200 epochs and batch size of 256 using an open-source implementation, OpenSelfSup.

#. 3-layer blocks	Pascal VOC			ImageNet	
	$AP$	$AP_{50}$	$AP_{75}$	conv4	conv5
0	<b>56.1</b>	81.7	62.2	57.2	41.9
1	<b>56.1</b>	81.8	62.4	58.1	59.0
2	<b>56.4</b>	81.8	62.8	57.9	62.1
3	<b>56.0</b>	81.3	63.1	57.6	63.1
4	<b>56.0</b>	81.6	62.0	57.3	63.0

Table 4. Ablation studies on mask prediction head. The first row shows the results of removing MPH and other rows show the results of using different number of 3-layer blocks.

Mask	Pascal VOC			ImageNet	
	$AP$	$AP_{50}$	$AP_{75}$	conv4	conv5
✓	<b>56.1</b>	81.7	62.2	57.2	41.9
	<b>53.6</b>	79.4	59.3	51.9	52.3

Table 5. Ablation studies on the mask strategy. MPH is not used in these experiments.

are designed to enhance pre-trained representations’ localization capability since they are harder to discriminate. We have mentioned that adding intra-image negatives improves the object detection AP, but degrades the image classification accuracy. We believe this is caused by the inconsistency between the two types of negatives, confusing the network about how much difference with the positive is sufficient to be classified into negative. We would like to explore a more elegant fusion strategy for the two types of negatives in future work. Moreover, we show that only using intra-image negative is not enough to learn good representations, which is caused by insufficient diversity of intra-image negatives.

## 5. Conclusion

In this work, we propose a novel SSL pretext task, named contrastive mask prediction (CMP), and the framework to

inter	intra	Pascal VOC			ImageNet	
		$AP$	$AP_{50}$	$AP_{75}$	conv4	conv5
✓		56.0	81.3	63.1	57.6	63.1
	✓	55.6	80.7	61.5	50.8	48.6
✓	✓	56.4	81.6	63.1	57.5	61.3

Table 6. Ablation studies on the negative sampling strategy. We try the inter-image negatives and intra-image negatives as well as their combination. Results reported at 100 epoch.

implement CMP, named mask contrast (MaskCo). CMP is motivated by the analysis of the limitation in current dominant pretext task, called instance discrimination (ID). We point out that ID is based on a semantic consistency assumption of curated data, leading to deteriorated performance when trained on unconstrained datasets. Our approach is based on a more fundamental assumption that contexts are correlated. It achieves consistent performance gain on multiple pre-training datasets over an ID-based method MoCo v2. As a generic pretext task, CMP has shown great promise towards training on unconstrained datasets, leading a new path to the ultimate goal of self-supervised learning: training with unlimited data in the wild.

## References

- [1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 139–156. Springer, 2018. [1](#), [2](#)
- [2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. [1](#), [2](#), [3](#)
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Pro-*

- ceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4, 5, 8
- [4] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 1, 2, 3, 4, 5, 8
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 1, 5
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2, 3
- [7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430. IEEE Computer Society, 2015. 2, 3, 8
- [8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR (Poster)*. OpenReview.net, 2018. 2, 8
- [9] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, pages 6390–6399. IEEE, 2019. 5
- [10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 2, 3, 8
- [11] 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. 1, 2, 3, 4, 5
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. IEEE Computer Society, 2017. 1
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 1, 5
- [14] 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. 2, 5
- [15] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Trans. Multim. Comput. Commun. Appl.*, 15(1s):8:1–8:19, 2019. 1
- [16] Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV (2)*, volume 11206 of *Lecture Notes in Computer Science*, pages 185–201. Springer, 2018. 2
- [17] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6706–6716. IEEE, 2020. 2
- [18] 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. 2, 3
- [19] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544. IEEE Computer Society, 2016. 2, 3, 4
- [20] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020. 3
- [21] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *NeurIPS*, 2020. 1, 2, 3, 6
- [22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 1, 5
- [23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL (1)*, pages 2556–2565. Association for Computational Linguistics, 2018. 2, 5
- [24] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. 2
- [25] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 4, 5
- [26] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *CoRR*, abs/2011.09157, 2020. 3
- [27] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 5
- [28] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742. IEEE Computer Society, 2018. 1, 2, 8
- [29] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang. Convolutional attention networks for scene text recognition. *ACM Trans. Multim. Comput. Commun. Appl.*, 15(1s):3:1–3:17, 2019. 1
- [30] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *CoRR*, abs/2011.10043, 2020. 3
- [31] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

- [32] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV (3)*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer, 2016.  
2
- [33] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *NeurIPS*, 2020. 3
- [34] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, pages 13275–13285. IEEE, 2020. 1
- [35] Bolei Zhou, Ágata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 2