Multi-Label Self-Supervised Learning with Scene Images

Ke Zhu Minghao Fu Jianxin Wu*
State Key Laboratory for Novel Software Technology
Nanjing University, China
zhuk@lamda.nju.edu.cn, fumh@lamda.nju.edu.cn, wujx2001@nju.edu.cn

Abstract

Self-supervised learning (SSL) methods targeting scene images have seen a rapid growth recently, and they mostly rely on either a dedicated dense matching mechanism or a costly unsupervised object discovery module. This paper shows that instead of hinging on these strenuous operations, quality image representations can be learned by treating scene/multi-label image SSL simply as a multi-label classification problem, which greatly simplifies the learning framework. Specifically, multiple binary pseudo-labels are assigned for each input image by comparing its embeddings with those in two dictionaries, and the network is optimized using the binary cross entropy loss. The proposed method is named Multi-Label Self-supervised learning (MLS). Visualizations qualitatively show that clearly the pseudo-labels by MLS can automatically find semantically similar pseudo-positive pairs across different images to facilitate contrastive learning. MLS learns high quality representations on MS-COCO and achieves state-of-the-art results on classification, detection and segmentation benchmarks. At the same time, MLS is much simpler than existing methods, making it easier to deploy and for further exploration.

1. Introduction

Self-supervised learning (SSL) methods based on contrastive learning [4, 3] have facilitated numerous downstream tasks. Those [16, 6, 14, 13] that target object-centric images (e.g., ImageNet) are already relatively mature, but inventing SSL methods for scene images (e.g., MS-COCO [23]) gains popularity recently. Since unlabeled scene images (or multi-label images) are more natural [46] and richer in semantics, various SSL methods [39, 45, 19, 40, 20, 44] have successively emerged.

SSL methods focusing on scene/multi-label images can be summarized into two categories. One is dense matching, such as DenseCL [39], MaskCo [51] and Self-EMD [24]. They take the features’ locations into account to improve the performance on dense prediction tasks. They mostly differ in how the heuristic matching metric is designed. Another branch of work like SoCo [41] and ORL [46] resort to unsupervised object discovery to find local object contents, and learn quality representation with both object- and scene-level semantics. However, they usually involve multi-stage SSL pretraining on top of expensive box generation [35].

These scene image SSL methods are generally based on the contrastive loss (e.g., InfoNCE [27]), where two randomly augmented views of the same image are forced to be close to each other, and optionally push away views from different images. The loss assumes single-label images [27], but the input are in fact multi-label: hence there is a mismatch between the loss and the data. On one hand, it can be difficult for two views randomly cropped from the same scene image to be exactly matched [39, 47]. On the other hand, there is only one matched positive pair in InfoNCE, while more positive pairs are naturally preferred: a view cropped from a multi-label image likely contains mul-
multiple semantic concepts or objects.

Therefore, this paper proposes a simple yet direct approach towards scene image SSL, named as Multi-Label Self-supervised learning (MLS). We treat each image (or randomly cropped patch) as a semantic bag with multiple objects, then retrieve images sharing similar semantics with any object in the bag from a large image dictionary. Note that an object in a bag is not necessarily within the set of any object in the bag from a large image dictionary. Note that an object in a bag is not necessarily within the set of human-annotated categories. As illustrated in Fig. 1, the cropped patch contains person, horse, road and bus, which will be pulled closer to similar images containing any of these objects and be pushed away from those dissimilar ones. Specifically, the patch’s embedding produced by a backbone network will select top k similar embeddings as k (pseudo) positive images from a dictionary, and the rest will be negatives. In another dictionary containing images in the same order as the first one, the BCE (binary cross entropy) loss plus these binary pseudo-labels will classify the patch’s embedding after an MLP projector using all the images in this second dictionary as classifiers, and generates gradients that optimize the backbone network. This framework is illustrated in Fig. 2.

Our framework has two benefits. First, the large dictionary has diverse positive samples for any given input, hence provides many quality positive pairs with deformations or intra-class variations [11]. Second, unlike InfoNCE, the BCE loss is not mutually exclusive among classes, hence allows the co-occurrence of multiple classes in one scene image. By applying our SSL method to scene images (e.g., MS-COCO), we achieve state-of-the-art results on object detection, instance segmentation and various classification benchmarks. Our contributions are summarized as follows:

1. For the first time, we formulate scene image SSL as a multi-label classification, and propose our Multi-Label Self-supervised (MLS) learning approach.

2. Unlike previous methods that adopt dense matching or unsupervised object discovery, MLS is simple in concept, and enjoys intuitive visualizations (c.f. Fig. 4) which clearly verifies our motivation.

3. Extensive experiments of object detection, instance segmentation and classification on various benchmark datasets, together with ablation studies, clearly demonstrate the effectiveness of our method.

2. Related Work

Traditional SSL methods. Self-supervised learning (SSL) has emerged as a promising direction towards unsupervised representation learning [6]. Early SSL methods [27, 33] are mainly derived from noise contrastive estimation [15]. Later, SSL methods based on contrastive learning and clustering paradigms have been proved effective, too. Representative methods such as SimCLR [4], MoCo [16], BYOL [14] and SwAV [3] have exhibited both simplicity and generalization ability [36] to various downstream tasks. A set of variants such as Simsiam [7], InfoMin [34] and NNCLR [11] try to simplify, analyze and optimize traditional SSL methods from different aspects. However, all of them mainly focus on pretraining on object-centric images like the ImageNet dataset [32] for image classification (e.g., ImageNet linear evaluation), but pay less attention to dense prediction tasks.

Scene-image SSL methods. The development of scene (multi-label) image SSL is a diversified process, which mainly contains two branches: dedicated dense feature matching and multi-stage pretraining with unsupervised object discovery. DenseCL [39], PixPro [47], ReSim [44], LEWEL [20] and SetSim [40] all apply the InfoNCE loss in a dense matching manner. They propose dense loss functions in addition to the traditional InfoNCE loss, and their pipelines differ in how the matching metrics are (probably manually) selected. Another branch of methods such as SoCo [41] and ORL [46] try to learn quality feature representation with both scene- and object-level information: they utilize multi-stage pretraining to find correspondence across images with the help of unsupervised object generation methods [35, 54], which suffers from huge computation costs. A recent SSL method [36] that is close to our work proposed kNN-MoCo as an extra module. But, this method is still in a single-label InfoNCE manner and the improvement is small compared to our multi-label MLS (c.f. Table 6). Compared to existing methods, our MLS is novel in concept and simple in implementation.

Multi-label classification. Multi-label recognition aims to predict the presence or absence of each object class in an input image [48]. Traditional methods are based on three aspects: attention modules [53, 52], correlation matrix [8, 5] and unsupervised box generation [42, 38]. The loss function adopted in this task is mostly binary cross entropy (BCE), which is not mutually exclusive among classes [48], such that the presence of one class does not suppress the existence of others. In this paper, we inherit this unique property of multi-label classification and try to integrate BCE into self-supervised learning on scene images. We adopt the joint image embedding to generate k pseudo-labels, together with the classification logits to calculate the final loss term, which effectively guides the multi-label self-supervised learning.

3. Method

In this section, we start by describing the preliminaries of self-supervised learning, including the InfoNCE loss and its variants. We will then describe the proposed Multi-Label Self-supervised (MLS) learning methods and how the over-
all loss function is formulated and optimized.

3.1. Preliminaries

Self-supervised learning methods, whether contrastive based [16, 14] or clustering based [3], usually rely on the InfoNCE loss [27] or its variants [33]:

\[
\mathcal{L}_{nce} = -\sum_{i=1}^{N_{pos}} \log \frac{\exp(q \cdot k_i^+ / \tau)}{\sum_{k_-} \exp(q \cdot k_- / \tau)},
\]

where \(q\) and \(k_+\) are positive embeddings (similar to \(z_1\) and \(z_2\) in Fig. 2) after the multi-layer perception (MLP) projector of two encoders. The negative features \(k_-\) may come from a memory bank [43], a large dictionary or a queue [6], or the current mini-batch [4]. Negatives may be removed by using additional techniques (e.g., the stop gradient adopted in BYOL [14] and SimSiam [7]).

Similar with the form of a normal softmax function [53], the total \(N_{pos}\) positives are enumerated over \(k_+\).

\[
\text{in which the total } \frac{N_{pos}}{N_{pos}} \text{ positives are enumerated over } k_+.
\]

It can indeed alleviate the dilemma of the traditional form by involving more positives, but all these methods mainly focus on supervised learning [21] or lack extensive experiments on large dense prediction tasks [36].

More importantly, the concept of multi-label learning is not introduced, that is, the classes are still mutually exclusive, which hinders the model from learning quality representations from scene image self-supervised learning.

3.2. Multi-label self-supervised learning

Now, we will introduce the proposed MLS method in detail. As discussed above, previous SSL methods who targeted scene images all adopt loss functions in the single-label form, while in this paper, we aim to reduce the gap between the optimization objective and the unique property of multi-label images.

We adopt MoCo-v2 [6] as our base structure. Specifically, given an input image \(I\), it is first cropped and augmented into two different views:

\[
v_1 = \mathcal{T}(I), \quad v_2 = \mathcal{T}'(I),
\]

where \(\mathcal{T}\) and \(\mathcal{T}'\) are two randomly sampled data augmentations. These two views are then passed through the base encoder \(\phi(\cdot)\) and a momentum encoder \(\phi^m(\cdot)\) to get the backbone features \(g_1\) and \(g_2\):

\[
g_1 = \phi(v_1), \quad g_2 = \phi^m(v_2),
\]

in which \(g_1, g_2 \in \mathbb{R}^{d_g}\) (e.g., \(d_g = 2048\) in ResNet-50 [18]).

Next, the backbone features are both passed through a multi-layer perceptron (MLP) projector \(f(\cdot)\) or \(f^m(\cdot)\):

\[
z_1 = f(g_1), \quad z_2 = f^m(g_2)
\]

to get the final embeddings \(z_1, z_2 \in \mathbb{R}^{d_z}\). Similar to MoCo-v2 that retains a large dictionary \(Q_g\) (shown in Fig. 2), here we keep two queues \(Q_g \in \mathbb{R}^{D\times d_g}\) and \(Q_z \in \mathbb{R}^{D\times d_z}\), with normalized embeddings coming from \(g_2\) and \(z_2\), respectively.

\[
\begin{align*}
g_2 \text{ Enqueue} & \quad \Rightarrow \quad Q_g, \quad \frac{g_2}{\|g_2\|} \quad \Rightarrow \quad Q_g, \\
\frac{z_2}{\|z_2\|} \quad \Rightarrow \quad Q_z.
\end{align*}
\]

Note that for simplicity, the enqueue and dequeue operations are not shown in Fig. 2, and that features stored in \(Q_g\) and \(Q_z\) are all kept in the same order. Embeddings in both queues are L2 normalized.

Then, the backbone feature \(g_1\) of view \(v_1\) picks its top \(k\) nearest neighbors from \(Q_g\) and treats them as containing similar semantics as that in the input image’s view \(v_1\) (e.g., containing at least one object that appears in the input view \(v_1\)). These nearest neighbors are given pseudo positive labels and all the rest items in \(Q_g\) are pseudo negatives. This step forms multiple binary pseudo-labels \(y \in \mathbb{R}^{D}\):

\[
y = \text{IsTopk}(g_1 \odot Q_g),
\]

where \(\odot\) is the matrix multiplication operation.

In the next step, the normalized embeddings inside \(Q_z\) are treated as normalized classifiers. They are multiplied by the feature \(z_1\) to produce the logits \(p \in \mathbb{R}^{D}\), which classify \(z_1\) into the \(D\) pseudo categories:

\[
p = z_1 \odot Q_z.
\]

Finally, the pseudo labels \(y\) and the classified logits interact with each other, and are optimized using the binary cross entropy (BCE) loss,

\[
\mathcal{L}_{ml} = \frac{-1}{D} \sum_{i=1}^{D} \left[y_i \log \sigma \left(\frac{p_i}{\tau}\right) + (1 - y_i) \log (1 - \sigma \left(\frac{p_i}{\tau}\right))\right],
\]

in which \(\tau\) is a temperature hyperparameter (following MoCo-v2’s), and \(\sigma(\cdot)\) is the sigmoid function [52] that maps the scores to the range \([0, 1]\).
3.3. Optimization

When we use the multi-label loss $L_{ml}$ as the only loss function, we found that empirically the optimization process often faces an unstable training issue. We believe this unstable optimization should be attributed to the adverse interaction between assigning pseudo-labels and learning representations from a poor starting point. High quality embeddings from the queues $Q_g$ and $Q_z$ cannot be obtained if the pseudo-labels are incorrect. At the same time picking nearest neighbors and classification will surely fail if the enqueued embeddings are misleading (a point which was mentioned in DenseCL [39], too).

Two strategies easily stabilize the multi-label SSL learning. One is to use InfoNCE alone in the first few epochs to warmup the dictionary $Q_g$ and $Q_z$. The other remedy is to combine InfoNCE with our multi-label BCE loss during the whole training process, which is widely used in previous methods [39, 1, 40] and easier in implementation. Hence, we adopt this strategy and our overall loss function is formulated as:

$$L = L_{nce} + \lambda L_{ml},$$

(10)

where $L_{nce}$ is the InfoNCE loss that is adopted in MoCo-v2 (c.f. Eq. 1) and along with many multi-label SSL methods. $L_{ml}$ is our multi-label BCE loss (c.f. Eq. 9). The value of combination weight $\lambda$ is fixed at $\lambda = 0.5$.

The proposed method is abbreviated as MLS (multi-label self-supervised) learning.

4. Experiments

Now we validate the effectiveness of MLS through extensive experiments. We first describe the settings of our experiments, including datasets for upstream pretraining and downstream finetuning, training details, and the architecture for object detection, instance/semantic segmentation and image classification.

4.1. Experimental settings

Datasets. We adopted the MS-COCO train2017 set [23] for SSL pretraining, which is widely applied in the study of scene image SSL [46, 41, 39]. Note that we did not use ImageNet. For downstream tasks, we conducted experiments on MS-COCO, VOC0712 [12], CityScapes [10] as well as 7 small classification datasets (CUB200 [37], Flowers [26], Cars [22], Aircraft [25], Indoor67 [30], Pets [28] and DTD [9]) following previous SSL methods [2]. Specifically, MS-COCO contains 118,287 and 5000 images for training and validation, respectively. VOC2007 has 5011 and 4952 images in the trainval and test sets, respectively. CityScapes is a semantic segmentation dataset containing 2975 and 500 images for training and validation, respectively. For more information about the 7 small classification datasets, please refer to [2] for a more detailed description.

Pretraining. We adopted ResNet-50 [18] as our base encoder (c.f. Fig. 2) and mostly follow the structure of MoCo-v2. We pretrained our models for 400 or 800 epochs on MS-COCO, set learning rate as 0.3, weight decay as 0.0001. For data augmentation, we adopted random resized crop, color distortion, random gray scale, random flip, etc, following [36, 3]. The queue size $D$ of $Q_g$ and $Q_z$ was 4096. The temperature $\tau$ was set as 0.2 and was shared between the InfoNCE loss and our multi-label BCE loss. The dimensionalities of the embeddings $g_1$ and $z_1$ were 2048 and 256, respectively. For faster convergence, we adjusted the momentum update coefficient to 0.995. With regard to our multi-label BCE loss, the parameter $k$ was set as 20, and the combination coefficient $\lambda$ was 0.5.

Downstream. We finetuned the pretrained model on various benchmarks, including MS-COCO object detection and instance segmentation, CityScapes semantic segmentation, VOC0712 detection and other classification benchmarks. Specifically, we finetuned on MS-COCO using the Mask-RCNN R50-FPN/C4 [17] architecture for 90k itera-
4.2. COCO detection and segmentation results

We first evaluated downstream performance on MS-COCO object detection and instance segmentation. Following previous work, we used Mask R-CNN R50-FPN for finetuning. The results is in Table 1. DenseCL [39] and Self-EMD [24] are those adopting dense matching, while SoCo [41] and ORL [46] both resort to the unsupervised object discovery method Selective Search [35]. For a fair comparison, we reproduced SoCo’s results since it was originally pretrained on COCO+ [41] (which is based on COCO but has more unlabeled data). Note that ReSim-C4 [44] and SetSim [40] are suitable for scene image pretraining, but only ImageNet results were reported. As shown in Table 1, with only 400 epochs of pretraining, our MLS is almost better than all counterparts (e.g., DenseCL, SoCo, LEWEL [20]) and surpasses supervised results by 1.2% APbbox and 0.8% APseg. With the standard 800 epochs of pretraining, MLS is better than all previous scene image SSL methods. This further verifies our motivation: formulating scene image self-supervised learning as a multi-label classification is both simpler in concept and effective in performance.

We also tried the Mask R-CNN R50-C4 structure. Since there are relatively less SSL results based on this structure, we reproduced supervised learning, BYOL [14], MoCo-v2 [6], SoCo [41] and DenseCL [39] methods by pretraining and finetuning on MS-COCO. To compare fairly, all these SSL models were pretrained for 400 epochs on MS-COCO.

As can be found in Table 2, our MLS surpasses all previous self-supervised learning methods like SoCo [41] and DenseCL [39], despite being much simpler than them. Note that SoCo needs unsupervised box generation and multiple auxiliary loss functions. Our MLS also achieves better results than supervised ImageNet pretraining, with an
improvement of 0.5% AP and 0.6% AP\textsuperscript{seg} for object detection and instance segmentation, respectively.

4.3. Transfer learning

Next, we test how our MLS method performs on various transfer learning tasks, including CityScapes semantic segmentation, VOC0712 object detection and (multi-label) classification.

**CityScapes semantic segmentation.** We first transferred pretrained models to CityScapes [10] semantic segmentation. Because relatively few SSL methods worked on this benchmark and the segmentation architecture generally differed, we first reproduced several SSL methods (BYOL, MoCo-v2, DenseCL and SoCo), then finetuned them in the same setting. It is clearly demonstrated in Table 3 that our multi-label self-supervised learning method is superior than all of them, surpassing state-of-the-art dense matching approach DenseCL [39] and object discovery method SoCo [41], both by significant margins. Our MLS also improves supervised learning from 77.5% mIoU to 79.0% mIoU and from 77.8% mIoU to 78.5% mIoU with PSANet [50] and PSPNet [49], respectively.

**VOC0712 object detection.** Moving on to other dense prediction tasks, we evaluated the effectiveness of MLS on VOC0712 object detection with Faster R-CNN R50-C4. Since there were missing results for BYOL and MoCo-v2 in this experimental setting, we reproduced both of them for a fair comparison. As shown in Table 4, our MLS surpasses all previous object-centric SSL methods by a large margin. For example, our method is higher than SimCLR by 3.5% AP and 5.6% AP\textsubscript{75} metrics. MLS is also better than supervised ImageNet results (an improvement of 1.7% AP and 2.4% AP\textsubscript{75}), showing again the great potential of multi-label self-supervised learning.

**Single- and multi-label classification.** Then, we explore how our proposed method behaves on image classification benchmarks. As pointed out by a previous SSL research work [45], various dense matching or unsupervised object discovery based SSL methods usually sacrifice the performance of image classification, in return for high accuracy in dense prediction tasks. In other words, classification is tough for these SSL methods. Therefore, we evaluated MLS on both multi-label (e.g., VOC2007) and single-label classification tasks. As discussed above, we also reproduced the results of MoCo-v2 and DenseCL because MLS has the same architecture with them. It can be found in Table 5 that MLS is consistently better than previous SSL methods, surpassing our baseline MoCo-v2 by a noticeable margin—an improvement of 5.3% mAP for VOC07 and 3.0% accuracy for Indoor67. We attribute this gain to the joint embedding property of MLS: since our method is optimized with global features, it can be naturally easier for us to capture image-level information, thus suitable for image classification tasks. However, all SSL methods are generally worse than supervised ImageNet pretraining, showing that supervised pretraining with more pretraining data (ImageNet vs. COCO) still remains a powerful technique for general image classification.

### 4.4. Ablation studies

**Compare with other loss functions.** We first explore if there are other alternatives besides multi-label BCE that can also achieve similar or even better results. An alternative, however, needs to produce multiple positive paradigms in the loss term. We consider kNN softmax adopted in [21, 36], and propose a variant of BCE by disregarding the negative term in Eq. 9 (calling it ‘BCE-pos’). As can be seen in Table 6, adding kNN softmax and the BCE-pos loss function besides the InfoNCE loss used in [36] both lead to consistent improvements in all metrics, demonstrating the effectiveness of our hypothesis. Still, they are both inferior to our proposed MLS, showing that pure BCE works better than the variant in scene image self-supervised learning.

**Why keeping two dictionaries?** There are two dictionaries Q\textsubscript{g} and Q\textsubscript{z} in our pipeline with different functionalities: one for pseudo-labeling and the other for classification.
Table 5. Results on downstream classification, including VOC2007 multi-label and 7 small single-label classification datasets. We reproduced the baseline MoCo-v2 and dense matching approach DenseCL for a fair comparison. The evaluation metric for VOC2007 is mAP while for others is top-1 accuracy. All models were pretrained for 400 epochs on the MS-COCO dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>VOC07</th>
<th>CUB200</th>
<th>Flowers</th>
<th>Cars</th>
<th>Aircraft</th>
<th>Indoor</th>
<th>Pets</th>
<th>DTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>ImageNet</td>
<td>89.0</td>
<td>81.3</td>
<td>96.7</td>
<td>90.6</td>
<td>86.7</td>
<td>58.1</td>
<td>64.4</td>
<td>74.7</td>
</tr>
<tr>
<td>MoCo-v2</td>
<td>COCO</td>
<td>80.5</td>
<td>68.8</td>
<td>89.6</td>
<td>87.5</td>
<td>80.4</td>
<td>64.1</td>
<td>79.4</td>
<td>60.9</td>
</tr>
<tr>
<td>DenseCL</td>
<td>COCO</td>
<td>83.7</td>
<td>69.3</td>
<td>88.6</td>
<td>88.3</td>
<td>79.9</td>
<td>65.4</td>
<td>80.0</td>
<td>61.8</td>
</tr>
<tr>
<td>MLS</td>
<td>COCO</td>
<td>85.8</td>
<td>71.7</td>
<td>91.2</td>
<td>88.5</td>
<td>81.2</td>
<td>67.1</td>
<td>81.9</td>
<td>63.0</td>
</tr>
</tbody>
</table>

Table 6. MS-COCO detection and segmentation results using different pretraining loss functions. The ‘kNN’ loss was adopted in [36] and the ‘BCE-pos’ loss means only positive terms of binary cross entropy in Eq. 9 were considered.

| Method       | Detection   |  | Segmentation  |  |
|--------------|-------------|-----------------|-----------------|
|              | APbbox      | APbbox50 | APbbox75 | APseg   | APseg50 | APseg75 |
| Baseline     | 38.8        | 58.3     | 42.5     | 35.5    | 55.8    | 37.5    |
| +kNN         | 39.1        | 58.7     | 42.6     | 35.3    | 55.8    | 37.9    |
| +BCE-pos     | 39.6        | 59.4     | 43.1     | 35.7    | 56.5    | 38.1    |
| +BCE (ours)  | 40.1        | 60.2     | 43.9     | 36.2    | 57.3    | 38.6    |

Table 7. Effect of using different dictionaries. ‘Q_g only’ means the embeddings in Q_g and g_1 are used twice to both generate the pseudo labels y and the logits p (c.f. Fig. 2). ‘Q_z only’ means only using Q_z and z_1.

<table>
<thead>
<tr>
<th>Method</th>
<th>APbbox</th>
<th>APseg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_g only</td>
<td>39.5</td>
<td>35.7</td>
</tr>
<tr>
<td>Q_z only</td>
<td>39.8</td>
<td>36.0</td>
</tr>
<tr>
<td>Both Q_g and Q_z (ours)</td>
<td>40.1</td>
<td>36.2</td>
</tr>
</tbody>
</table>

Why not keeping only one for simplicity? We explain this problem through both visualization and quantitative experiments. First, we plot the histogram of scores in Fig. 3, with g_1 ⊙ Q_g and z_1 ⊙ Q_z representing the backbone and MLP scores, respectively. The backbone scores have witnessed a dramatic change after training, indicating some information the network potentially obtained by SSL. On the other hand, the MLP scores always follow a normal distribution, with only slight changes in the mean and variance statistics. Hence, we choose to use the backbone queue Q_g for pseudo labeling. The MLP dictionary Q_z is used for classification, following the common practice in previous SSL methods [16, 39]. Our choice is also verified by the experiments in Table 7, where using only Q_g or Q_z is clearly sub-optimal, and worse than our MLS where they are jointly adopted.

Effect of k. Now we examine how different k values (the number of pseudo positive labels generated) affect the final results of our MLS method. The downstream results of MS-COCO detection and segmentation can be found in Table 8, where a too small and a too large k value both lead to a performance drop. A small k means semantically related images may become false negatives in BCE classification, while a large k will similarly lead to false positives. Our MLS is relatively more robust when k is small. But when k is too large, the accuracy drop is significant. Based on this observation, we choose k = 20 by default.

Effect of queue size and λ. The fourth module we evaluated is the dictionary (queue) size and the combination weight λ in Eq. 10. We pretrained the models on the MS-COCO dataset using MLS and finetuned them for COCO object detection with Mask R-CNN R50-FPN. As demonstrated in Fig. 5 (a), a too small queue size (e.g., 1024) is detrimental for SSL pretraining, since a small queue means limited representations stored, which is not enough for an image to retrieve its positive neighbors. An overly large dictionary also obtains sub-optimal results, because dictionaries that are too large probably contain many out-of-date embeddings (due to the dequeue and enqueue mechanism [6]) and might lead to incorrect pseudo labels. And, we also validate the robustness of our MLS by trying multiple λ values. It can be seen in Figure 5 (b) that all λ leads to consistent
improvement over the baseline (c.f. Table 6), and there is a sweet spot in the choice of λ. Thus, in this paper we choose λ = 0.5 by default in our experiment.

**Visualization of positive pseudo-labels.** Last but not least, we validate the effectiveness of our MLS pipeline by visualization: are the top k nearest neighbors (NN) picked up from the dictionary $Q_g$ indeed semantically similar to the input? Because random resized crop is applied in the SSL pipeline (it is the cropped patch that is sent to the network), we record the cropped location of each image during data augmentation. Then, for each query $g_1$ in Fig. 2, we find its top-4 NNs stored in the queue $Q_g$. All these queries and NNs are associated with a certain cropped patch, and thus we visualize them with red and yellow rectangles, respectively. As shown in Fig. 4, our MLS backbone effectively captures semantically similar correspondences across the dataset. Specifically, intra-class variance (the first row) and multiple positive partial concepts (the second and third rows) are clearly demonstrated, showing that our motivation in Sec. 1 is valid and that MLS is not only suitable for scene image SSL, but might also be adaptable to image retrieval, as well.

### 5. Conclusions and Limitations

In this paper, we argued that scene image self-supervised learning does not necessarily rely on dense matching or unsupervised object discovery methods, and instead proposed our Multi-Label Self-supervised (MLS) learning approach. The key idea in MLS is that a multi-label image contains multiple concepts or objects, hence we need to have multiple positive pairs corresponding to different objects in the input image. Specifically, we create two dictionaries, one for creating pseudo-labels (positive or negative pairs) and the other for distinguishing between them (i.e., a multi-label classification problem). This is the first time that multi-label SSL is cast as a multi-label classification problem, and MLS has been validated by extensive experiments on various benchmarks (e.g., MS-COCO, CityScapes, VOC0712, etc.) and tasks (object detection, image segmentation, single- and multi-label classification).

As for limitations, it still remains unclear why the BCE loss alone will cause an unstable training issue. This is a long-lasting and open question as numerous scene image SSL methods [39, 44, 40, 1] also struggle if the InfoNCE
loss is removed. Hence, these methods (including our MLS) always use their proposed modules in combination with InfoNCE. In the future, we will explore how and why the collapsing phenomenon exists without the normal contrastive loss, and will apply our MLS to single-label image SSL to further explore the adaptability of our method.

References


[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In ICCV Workshop on 3D Representation and Recognition, 2013.


[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Rai- son, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,


