Self-Supervised Pyramid Representation Learning for Multi-Label Visual Analysis and Beyond

Cheng-Yen Hsieh* Chih-Jung Chang* Fu-En Yang Yu-Chiang Frank Wang
National Taiwan University
chengyeh@andrew.cmu.edu, {b06201018, f07942077, ycwang}@ntu.edu.tw

Abstract

While self-supervised learning has been shown to benefit a number of vision tasks, existing techniques mainly focus on image-level manipulation, which may not generalize well to downstream tasks at patch or pixel levels. Moreover, existing SSL methods might not sufficiently describe and associate the above representations within and across image scales. In this paper, we propose a Self-Supervised Pyramid Representation Learning (SS-PRL) framework. The proposed SS-PRL is designed to derive pyramid representations at patch levels via learning proper prototypes, with additional learners to observe and relate inherent semantic information within an image. In particular, we present a cross-scale patch-level correlation learning in SS-PRL, which allows the model to aggregate and associate information learned across patch scales. We show that, with our proposed SS-PRL for model pre-training, one can easily adapt and fine-tune the models for a variety of applications including multi-label classification, object detection, and instance segmentation.

1. Introduction

To understand the complex relations in natural scenes or explore rich information from an image, many real-world visual recognition tasks (*e.g.*, semantic scene classification [41], or medical diagnosis [1]) require the learned model to predict *more than one* semantic label given a single input image. The conventional single-label classification methods mainly focus on assigning *single* class label to each image without considering the multiple-object scenarios in one image or handling the relations among distinct label semantics. More particularly, the derived features are required to describe the presence of multiple objects and semantic label dependencies in an image for tackling multi-label visual analysis tasks. While existing [27, 32–34, 40, 43, 50, 52] methods perform promising performance, they still acquire a large amount of multi-label annotated data for training. Considering the labeling cost, collecting fully annotated data for learning a model for multi-label tasks would be computationally expensive.

To alleviate the huge burdens of collecting and annotating large-scale multi-label datasets, an effective approach is to pre-train a general-purpose model in the self-supervised learning (SSL) manner, followed by the fine-tuning process to facilitate the learning of downstream tasks of interest. Recent SSL pre-training approaches [3, 6, 9, 10, 14, 17, 19, 23, 30, 38] learn discriminative representations based on *image*-level contrastive learning scheme, which pulls the views from the same image together and pushes the features from different images away. While such training fashion significantly improves the performance on single-label image classification, the above SSL methods are only trained at *image*-level, which lacks the ability to describe the multiple objects in an image. Hence, transferring the learned knowledge from such SSL pre-trained models to downstream multi-label visual analysis tasks remains underexplored.

To perform pre-training for downstream multi-label tasks, we aim at exploiting inherent semantic label dependencies in a *self-supervised* manner. In this paper, we propose a unique *self-supervised pyramid representation learning (SS-PRL)* framework. Without observing any ground truth labels at either image or object levels, our SS-PRL is learned in a *cross-scale patch-level SSL* manner that derives pyramid representations and semantic prototypes at *patch* levels. This allows one to explore the presence of objects and label dependencies in an image while leveraging the correlation across multiple patch scales to associate and aggregate the knowledge learned from different patch scales.

With the particular aim of exploiting fine-grained information within an image for mimicking objects presented at various scales, our proposed SS-PRL constructs multiple branches to extract global image-level and local patch-level features from the input image for learning the pyramid representations and associated prototypes. These prototypes are designed to serve as semantic cues for describ-

*Authors contributed equally

ing label dependencies and thus are expected to improve the model capability for downstream multi-label tasks (*e.g.,* multi-label image classification, or object detection).

To further integrate the information from different patch-level representations, we present *cross-scale patch-level correlation learning* in SS-PRL. This enforces the correspondence of output predictions from global image and local patches, which guides the model to leverage multi-grained information. To verify the effectiveness of our SS-PRL for diverse downstream tasks, we consider multi-label image classification, object detection, and segmentation benchmarks in our experiments. We confirm that our SS-PRL performs favorably against SOTA methods and achieve promising performances.

The contributions of our work are highlighted below:

- To the best of our knowledge, we are among the first to design pretext tasks in a self-supervised manner for facilitating downstream multi-label visual analysis tasks.
- We propose Self-Supervised Pyramid Representation Learning (SS-PRL), deriving multi-scale patch-level pyramid representations with semantic prototypes discovered to exploit their inherent correlation.
- A unique cross-scale patch-level correlation is introduced in our SS-PRL to leverage the learned knowledge across multiple and distinct spatial scales, ensuring sufficient representation ability of our model.
- In addition to a wide range of downstream tasks at object instance and pixel levels, we qualitatively demonstrate that the learned prototypes at different scales would describe the associated visual concepts.

2. Related work

2.1. Multi-Label Image Classification

Multi-label image classification aims at assigning a set of labels to each image. Due to the fact that pictures in everyday life are inherently multi-labeled and contain more complex visual appearances and diverse label semantics, multi-label visual analysis is more practical yet challenging compared with conventional single-label classification tasks. Associating local image regions to labels has been proven to be beneficial in multi-label classification since an image is usually composed of objects with different scales located in arbitrary regions. SRN [54] learns an attention map that associates related image regions to each label in order to portray the underlying spatial relation between semantic labels. Gau *et al.* [18] improves the performance of multi-label classification by introducing a consistency objective on visual attention regions under image transformations. In addition, Ridnik *et al.* [33] and Wu *et al.* [43] propose asymmetric loss and distribution-balanced loss respectively to mitigate the accuracy degradation from the positive-negative imbalance. While promising, most existing works [18, 27, 32–34, 43, 54] generally learn the correspondence between image regions and labels in a fully-supervised fashion.

To mitigate the costly process of collecting and annotating large-scale multi-labeled datasets, various settings of multi-label classification with limited supervision have been proposed. For example, multi-label learning with missing labels [36] considers the case in which only a partial set of labels is available; semi-supervised multi-label classification [8] admits a few fully-labeled data and a large amount of unlabeled data; partial multi-label learning [48] discusses the setting that each instance is annotated with a set of candidate labels. Different from the above settings, we aim to tackle multi-label visual analysis in a *self-supervised* fashion that pre-trains on unlabeled data while only using a few labeled samples for further fine-tuning.

2.2. Self-Supervised Learning

Recently, self-supervised learning methods [2–4, 6, 7, 9–11, 14, 16, 17, 19, 23, 24, 29, 30, 38, 39, 51] achieve remarkable progress on single-label image classification and narrow the performance gap compared with fully-supervised counterparts. One group of SSL approaches adopt the contrastive objective to perform instance discrimination on a large amount of unlabeled data. For instance, PIRL [29], SimCLR [9], and MoCo v1/v2 [10, 19] share the same concept of pulling multiple views of an image close while pushing different instances apart to derive the compact yet discriminative representations. BYOL [17] and SimSiam [11] claim that the use of asymmetry network architecture and exponential moving average update strategy are the crucial factors in preventing mode collapsing when the training process only rely on positive pairs. Barlow Twins [51] tries to align the corresponding entities between the two embedded features from each positive pair by Siamese network.

Another group of SSL works can be viewed as clustering-based methods, which learn visual representations via pseudo-label prediction. DeepCluster [5] and SeLa [2] apply k-means clustering and optimal transport respectively to produce pseudo labels. In contrast to [2, 5], SwAV [6] proposes an online clustering method that assigns the soft labels to the input image via the learned prototype vectors. We note that the aforementioned SSL methods simply extract a single feature to represent an image and thus do not handle the presence of multiple objects from an image well. The ability to transfer the learned knowledge from such pre-training tasks (*i.e.,* image-level contrastive learning or clustering) to downstream tasks with multiple labels (*e.g.,* semantic segmentation, object detec-

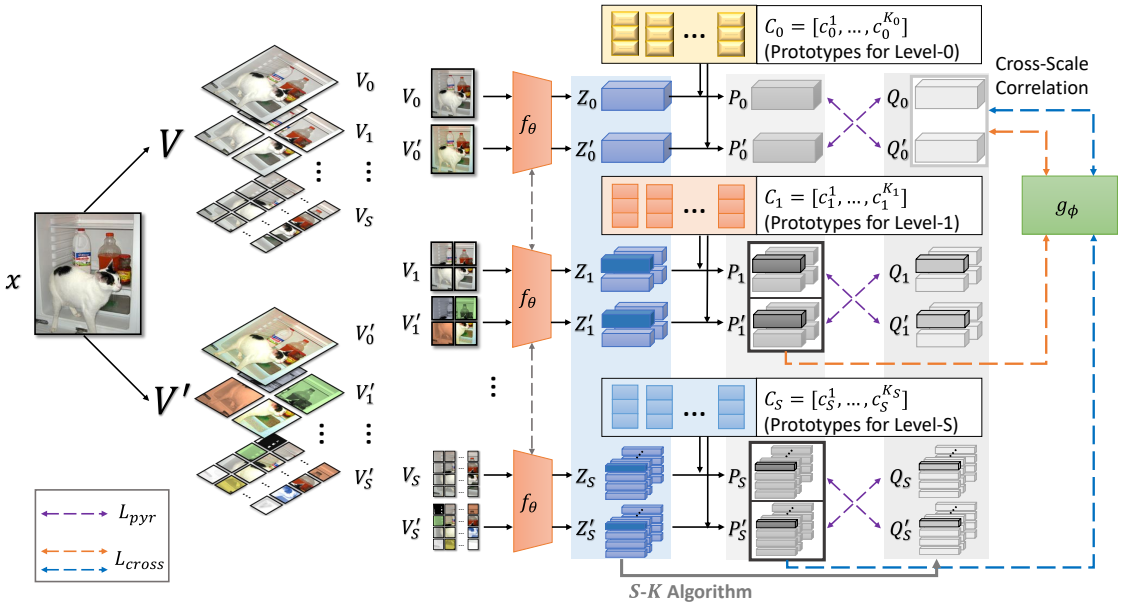


Figure 1. **Self-Supervised Pyramid Representation Learning.** The input x is augmented into two pyramid views $V = \{V_s\}_{s=0}^{S}$ and $V' = \{V'_s\}_{s=0}^{S}$ with patch sets obtained at each scale. For scale s, we have f_θ derive the pyramid representations Z, which are further transformed into prototype-based representations P based on the learned/assigned prototypes C_s at that scale. With the prototype assignments Q inferred from Z via S-K algorithm [12], we observe correlation between Q and the aggregated P across each scale via g_ϕ as cross-scale patch-level self-supervision.

tion, or multi-label classification) remains challenging and still underexplored.

To better finetune the pre-training models for facilitating downstream visual classification tasks, a number of works [22, 28, 31, 35, 42, 45–47, 49, 53] design specific pretext tasks which are consistent with the characteristics of downstream tasks of interest. These methods are generally dedicated to constructing the pretext tasks that benefit dense prediction like semantic segmentation, object detection, or keypoints detection. For example, DenseCL [42] introduces a pairwise contrastive loss at the pixel-level features between two views of an input image. DetCo [46] jointly learns discriminative representations from global images and local patches via contrastive learning across multiple scales and network layers. InsLoc [49] proposes a localization pretext task with the contrastive loss by taking crops of foreground images pasted onto different background images. MaskCo [53] contrasts region-level features with the contrastive mask prediction task.

We note that, while [22, 28, 31, 35, 42, 46, 47, 49, 53] integrate the local information into instance discrimination scheme, they are *not* designed to observe the inherent relations among objects and thus are sub-optimal for downstream multi-label visual analysis tasks. In this paper, we design the pretext task for multi-label image classification by deriving the pyramid representations across multiple scales, producing multi-level semantic prototypes to exploit the label relations from the observed training data.

3. Proposed Method

3.1. Problem Formulation

For the sake of completeness, we first define the problem setting considered in this work. Given an unlabeled dataset $\mathcal{D}_u = \{x_1, x_2, ..., x_N\}$ of N images, we aim to learn a feature extractor f_θ on $\mathcal{D}_u$, facilitating downstream tasks associated with multi-labels. As depicted in Fig. 1, we present a Self-Supervised Pyramid Representation Learning (SS-PRL) framework, which consists of a feature extractor f_θ and a cross-scale correlation learner $g_\phi = \{g_{\phi,s}\}_{s=1}^{S}$ at each patch scale s. We apply f_θ to derive pyramid representations Z from the pyramid of views V, which are then transformed into prototype-based representations P based on semantic prototypes C_s learned at each scale. To further leverage multi-grained information from different scales, cross-scale patch-level correlation is enforced between S-K based prototype assignments Q and the aggregated P across scales via g_ϕ. Once the learning is complete, one can apply and fine-tune f_θ for downstream tasks like multi-label image classification, object detection, or segmentation.

3.2. Self-Supervised Pyramid Representation Learning

As illustrated in Fig. 1, the framework of our proposed Self-Supervised Pyramid Representation Learning contains the stage of *patch-based pyramid representation learning* and *cross-scale patch-level correlation learning*. The

former is to derive pyramid representations via prototypes learned at each *patch* level, aiming to handle the presence of multiple objects while exploring the label dependencies from unlabeled data. As for the latter stage, we further associate and aggregate the learned knowledge across different patch scales by enforcing the coherence between the prediction of local patches and global images. We now detail the designs for the above two stages below.

3.2.1 Learning of patch-based pyramid representation

As noted in Section 2.2, prior SSL works [2, 6, 9, 19, 29] generally embed an image into a single feature and are not designed for observing multiple objects presented in an image. Hence, such derived models and representations cannot be easily transferred to downstream multi-label visual analysis tasks. To handle images with multi-objects/labels, we derive pyramid representations at the patch level instead of producing image-level features. This allows the model to observe and capture more fine-grained information from an image. In addition, our SS-PRL is designed to learn prototypes at each patch level, which exploits potential label dependencies in an unsupervised fashion.

As shown in Figure 1, we first build two pyramids of views $V = \{V_s\}_{s=0}^{S}$ and $V' = \{V'_s\}_{s=0}^{S}$, which are generated with different augmentations from the input image x. For each patch scale s, the image patch group $V_s = [v_{s,1}, \ldots, v_{s,M_s}]$ is produced by dividing the image x into M_s non-overlapping patches and randomly transforming each patch with data augmentations. Similar remarks can be applied to the derivation of $\{V'_s\}_{s=0}^{S}$. To derive the patch-level pyramid representations $Z_s = [z_{s,1}, \ldots, z_{s,M_s}] \in \mathbb{R}^{D \times M_s}$ and $Z'_s = [z'_{s,1}, \ldots, z'_{s,M_s}] \in \mathbb{R}^{D \times M_s}$, we feed the two pyramid views $\{V_s\}_{s=0}^{S}$ and $\{V'_s\}_{s=0}^{S}$ into feature extractor f_θ, which contains a shared backbone network and $S+1$ independent projection heads corresponding to patch scale $s = 0, 1, ..., S$.

Prototype-based self-supervised learning. With pyramid representations Z_s and Z'_s obtained, we require our SS-PRL to produce representations that are discriminative and be capable of capturing the inherent semantic dependencies observed from training data, which is thus beneficial to downstream multi-labeled tasks. Inspired by [6], we learn a group of patch-level semantic prototypes $C_s \in \mathbb{R}^{D \times K_s}$ at each patch scale s (where K_s denotes the number of prototypes at scale s) to mine and reflect the label semantics observed from unlabeled training data. To allow the feature extractor f_θ and semantic prototypes C_s to be learned jointly in an online fashion, we utilize the consistency between the probability distribution of $z_{s,m}$ and $z'_{s,m}$ as self-supervision [6]. To be more specific, such prototypes C_s can be viewed as clustering centroids, and we then transform $z_{s,m}$ to prototype-based representations P_s by assigning each representation $z_{s,m}$ to prototypes $C_s = [c_s^1, \ldots, c_s^{K_s}]$ at each scale s. We derive the prototype-based representations $P_s = [p_{s,1}, \ldots, p_{s,M_s}] \in \mathbb{R}^{K_s \times M_s}$ from $z_{s,m}$ and C_s to represent probability distribution as follows:

$$p_{s,m}^{\top} = softmax(\frac{1}{\tau} z_{s,m}^{\top} C_s), \tag{1}$$

where τ is a temperature parameter as noted in [44].

However, simply aligning the prototype-based representations P_s and P'_s might lead to mode collapse problems [6]. To alleviate this issue, we further utilize the iterative Sinkhorn-Knopp algorithm [12], denoted by $S\text{-}K(\cdot,\cdot)$, to compute the prototype assignment vector $q_{s,m} = S\text{-}K(z_{s,m}, C_s)$ for two S-K based prototype assignments $Q_s = [q_{s,1}, \ldots, q_{s,M_s}]$ and $Q'_s = [q'_{s,1}, \ldots, q'_{s,M_s}]$, which serve as the target of prediction by P_s. With the equal partition property imposed by Sinkhorn-Knopp algorithm, the consistency enforced between $p_{s,m}$ and $q'_{s,m}$ is capable of alleviating the mode collapse problems [6]. As a result, the objective for our pyramid representation learning L_{pyr} is defined as:

$$L_{pyr} = \sum_{s=0}^{S} \sum_{m=1}^{M_s} \frac{\alpha_s}{M_s} (CE(q'_{s,m}, p_{s,m}) + CE(q_{s,m}, p'_{s,m})), \tag{2}$$

where CE denotes the cross-entropy loss, and α_s balances each loss term at different patch scales s.

Although the above pyramid representations can be learned without label supervision, self-supervision at each scale is observed separately. As later verified in Table 4, this would lack the ability to associate patch-level prototypes across image scales and thus limit the downstream classification tasks associated with multi-labels. This is why the additional self-supervision across patch scales needs to be enforced, as we introduce below.

3.2.2 Cross-scale patch-level correlation learning

As noted above, it would be desirable to train deep learning models which exploit semantic dependencies not only at each patch scale but also discover such properties with information properly aggregated and leveraged across scales. To achieve this goal and to benefit downstream multi-label classification tasks, we uniquely observe the correlation observed between the prototype/cluster assignments derived at the coarsest image scale (*i.e.*, Q_0 or Q'_0) and the prototype-based representations P_s aggregated at each scale s. With a deployed cross-scale correlation learner g_ϕ, the above correlation can be enforced and be served as cross-scale patch-level self-supervision for training purposes.

More specifically, we perform average pooling across all M_s representation vectors in P_s and P'_s from level s, resulting in $\mu(P_s)$ and $\mu(P'_s)$, respectively. We then apply a

Pre-training Method		Multi-Label Classification (mAP)			
		Pretrained on COCO		Pretrained on ImageNet	
		COCO	VOC	COCO	VOC
Supervised		62.5	81.8	68.5	86.7
MoCo v2 [19]	*general-purpose SSL*	50.2	67.9	54.3	82.5
SwAV [6]	*general-purpose SSL*	60.3	79.2	60.1	83.2
BYOL [17]	*general-purpose SSL*	52.6	70.1	58.4	80.2
DenseCL [42]	*dense prediction SSL*	57.0	75.2	60.5	82.9
DetCo [46]	*dense prediction SSL*	52.7	70.6	60.0	81.3
MaskCo [53]	*dense prediction SSL*	51.9	70.2	50.3	75.1
InsLoc[49]	*dense prediction SSL*	45.0	61.8	49.5	74.8
SS-PRL (ours)		**61.3**	**80.5**	**63.8**	**85.4**

Table 1. **Performance on multi-label classification tasks with fine-tuned linear classifiers on VOC and COCO.** With the backbone network (*i.e*. ResNet-50) pre-trained with different supervised/self-supervised methods, we report the mAP on COCO and VOC with *fine-tuned linear classifiers*. All methods are pre-trained on COCO with 200 epochs or ImageNet with 100 epochs, respectively.

Pre-training Method		Multi-Label Classification on COCO (mAP)					
		Pretrained on COCO			Pretrained on ImageNet		
		1% labels	10% labels	100% labels	1% labels	10% labels	100% labels
Random Init.		4.6	10.7	42.5	4.6	10.7	42.5
MoCo v2 [19]	*general-purpose SSL*	34.0	46.9	54.2	26.4	55.8	63.7
SwAV [6]	*general-purpose SSL*	43.6	56.2	61.4	39.3	58.6	66.9
BYOL [17]	*general-purpose SSL*	35.1	48.0	54.8	38.7	53.1	62.5
DenseCL [42]	*dense prediction SSL*	42.9	54.8	62.2	**43.4**	59.4	65.8
DetCo [46]	*dense prediction SSL*	32.0	48.3	54.7	37.9	56.2	62.7
MaskCo [53]	*dense prediction SSL*	31.6	48.0	57.4	24.0	53.2	62.1
InsLoc[49]	*dense prediction SSL*	29.0	43.9	53.5	36.1	56.6	66.5
SS-PRL (ours)		**45.1**	**57.0**	**62.9**	41.0	**60.9**	**67.4**

Table 2. **Performance on multi-label classification tasks in semi-supervised settings on COCO.** Methods listed are pre-trained on COCO for 200 epochs or ImageNet for 100 epochs, respectively. Models are then fine-tuned on 1%, 10%, and 100% of labeled data randomly chosen from COCO for 20 epochs. Note that, *Random Init.* denotes the model trained from scratch.

set of cross-scale correlation learners $g_\phi = \{g_{\phi,s}\}_{s=1}^{S}$, one for each scale, to project $\mu(P_s)$ and $\mu(P'_s)$ onto the representation space of p_0 and p'_0, i.e., at the global image level. As a result, our cross-scale correlation loss L_{cross} can be formulated as:

$$L_{cross} = \sum_{s=1}^{S} \beta_s (CE(Q_0, g_{\phi,s}(\mu(P_s))) + CE(Q'_0, g_{\phi,s}(\mu(P'_s)))), \quad (3)$$

where CE is the cross-entropy loss, and β_s balances cross-scale correlation losses across different scales.

It is worth noting that, learning pyramid representations for different patch-level scale pairs not only encourages the feature extractor f_θ to exploit the patch-level information in an image, it also aggregates the fine-grained semantics for matching the global ones (via g_ϕ) presented in an image. As confirmed in our experiments, this self-supervised learning strategy allows us to fine-tune f_θ for downstream tasks associated with multi-labeled images.

3.3. Pre-Training and Fine-Tuning Stages

Self-supervised pre-training of f_θ and g_ϕ. Overall, the full objective function L for pre-training feature extractor f_θ and the cross-scale correlation learner g_ϕ can be summarized below:

$$L = L_{pyr} + \lambda L_{cross}, \quad (4)$$

where λ acts as the weight to balance the two terms, and is set as 1.0 throughout our work. On the other hand, we select the same values for α_s and β_s in (2) and (3) for simplicity (we set these hyperparameters as 1.0 for $s = 0$ and 0.25 for other scales to balance the influence of different levels). The effectiveness of each loss is later confirmed by the ablation study in Section 4.3, and the pseudo-code is summarized in the supplementary material.

Supervised fine-tuning for f_θ. Once the feature extractor f_θ is pre-trained via our proposed SS-PRL, we then fine-tune it to downstream tasks associated with multi-label images in a supervised fashion. For example, as presented in Section 4, we adapt the pre-trained f_θ (*e.g*., with the architecture as ResNet-50 [21]) to multi-label image classification, object detection, and segmentation tasks using different amounts of images with ground truth annotation. Please see the next section for the thorough experiments on these tasks and comparisons to state-of-the-art SSL methods.

Method		Mask R-CNN R50-FPN COCO 15k											
		Pretrained on COCO						Pretrained on ImageNet					
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Random Init.		11.5	21.3	11.3	10.8	19.7	10.7	11.5	21.3	11.3	10.8	19.7	10.7
MoCo v2 [19]	*general-purpose SSL*	17.0	30.6	17.2	15.9	28.4	15.9	21.1	36.7	21.8	19.9	34.4	20.4
SwAV [6]	*general-purpose SSL*	18.1	33.7	17.6	17.3	31.5	17.1	23.1	41.2	23.4	22.1	38.6	22.5
BYOL [17]	*general-purpose SSL*	17.4	31.5	17.4	16.2	29.2	16.1	21.4	37.5	22.1	20.1	35.1	20.6
DenseCL [42]	*dense prediction SSL*	**20.2**	35.4	**20.8**	18.9	33.0	**19.3**	21.9	38.0	22.9	20.7	35.8	21.3
DetCo [46]	*dense prediction SSL*	15.6	29.7	14.8	14.8	27.3	14.4	20.9	38.1	20.9	19.9	35.3	19.9
MaskCo [53]	*dense prediction SSL*	18.5	32.9	18.7	17.3	30.7	17.4	20.6	35.6	21.5	19.5	33.4	20.0
InsLoc [49]	*dense prediction SSL*	17.5	31.5	17.6	16.5	29.3	16.6	23.5	40.5	**24.7**	22.2	38.1	22.9
SS-PRL (ours)		**20.2**	**36.7**	20.2	**19.1**	**34.3**	19.0	**23.6**	**42.5**	24.0	**22.7**	**39.7**	**23.1**

Table 3. **Downstream object detection and instance segmentation tasks on COCO.** We report the bounding box AP (AP^{bb}) for object detection and the mask AP (AP^{mk}) for instance segmentation on COCO. All methods are pre-trained on COCO for 200 epochs or ImageNet for 100 epochs and then fine-tuned for the above tasks on COCO for 15k iterations. Note that *Random Init.* denotes the detector trained from scratch (*i.e.* the encoder is randomly initialized without any pre-training). The best results in each category are in **bold**, and the second-best ones are underlined.

Prototype	mAP
Baseline	79.2
Shared across all scales	79.4
Learned & correlated across scales	**80.5**

Method	mAP
Baseline	79.2
SS-PRL w/ L_{pyr} only	79.5
SS-PRL w/ L_{cross} only	79.8
Full SS-PRL (L_{pyr} + L_{cross})	**80.5**

Table 4. **Ablation Studies on the derived patch-level prototypes (top) and the proposed loss functions (down).** Note that *Shared across all scales* indicates the same prototypes learned across patch scales (*i.e.*, same C_s at different patch scales in Fig. 1). We see that prototypes learned from each scale and enforced by our cross-scale correlation would be desirable. And, SS-PRL achieves the best results when both L_{pyr} and L_{cross} are introduced.

4. Experiments

4.1. Datasets and Experimental Setups

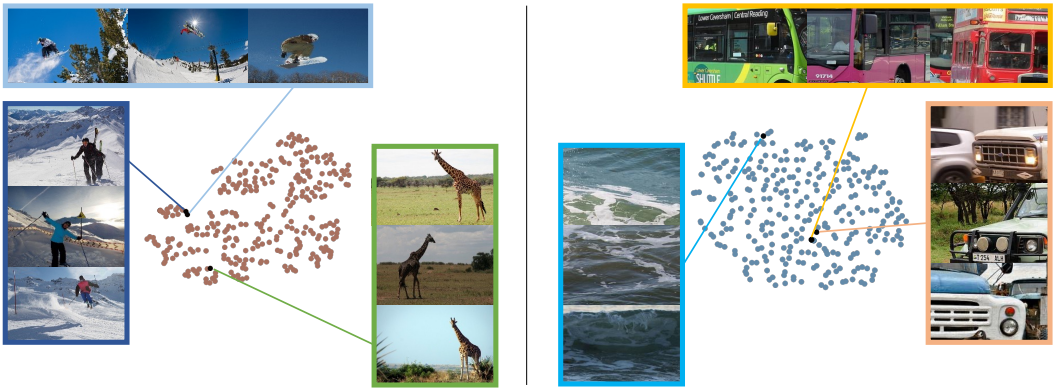
Pre-training Dataset. We consider MSCOCO [26] and ImageNet [13]. For MSCOCO [26], COCO `train2014` [26], which contains ∼83k images, is used for SSL pre-training, and we train all methods for 200 epochs with a batch size of 128. As for ImageNet [13], we utilize the training set with ∼1.28M training images for SSL pre-training, and train methods for 100 epochs with a batch size of 256. Our image pyramids contain three patch scales (*i.e.*, $s = 0, 1, 2$) in all experiments. The patch sets consist of 4 patches ($M_1 = 4$) at scale $s = 1$ and 9 patches ($M_2 = 9$) at scale $s = 2$. Further training details such as data generation and hyperparameter selection are provided in our supplementary material.

Evaluation protocol. We evaluate the pre-trained models by fine-tuning on downstream multi-label classification, object detection, and segmentation tasks using MSCOCO `train2014` [26] and PASCAL VOC [15]. For multi-label classification task, we follow the linear evaluation setting [29] to train a linear multi-label classifier on top of the fixed pre-trained backbone network (*e.g.*, Resnet-50) on COCO `train2014` [26] and VOC `trainval07` [15], and then report mean average precision (mAP) on COCO `val2014` [26] and VOC `test2007` [15]. We also follow the semi-supervised setting and randomly sample 1%, 10%, and 100% labeled data from COCO `train2014` [26] (which is ∼0.8k, ∼8k, and ∼83k images) to fine-tune the whole network for 20 epochs, and then report mAP on COCO `val2014` [26].

As for object detection and instance segmentation tasks, we pre-train and fine-tune a Mask R-CNN [20] detector with FPN [25] backbone on COCO `train2014` [26] and evaluating on COCO `val2014` [26]. Note that, synchronized batch normalization is applied in the backbone network, FPN, and prediction heads during training. We report the results of detectors with $15k$ training iterations to compare the transfer ability of each SSL pre-training method. Due to page limitations, we provide quantitative comparisons on the downstream semantic segmentation task in our supplementary material.

4.2. Quantitative Evaluation

Multi-label classification with fine-tuned linear classifiers. We first perform downstream multi-label image classification with fine-tuned linear classifiers and compare our results with existing general-purpose [6, 17, 19] and dense prediction based [42, 46, 49, 53] self-supervised learning methods on two commonly-used public benchmarks, COCO `train2014` [26] and VOC [15]. In Table 1, we observe that our SS-PRL outperforms state-of-the-art SSL approaches on multi-label classification benchmarks



(a) Image Level (Scale 0) **(b) Patch Level (Scale 2)**

Figure 2. **t-SNE visualization of the learned prototypes on COCO.** We visualize the learned prototype at the corresponding scale, with selected images associated with each prototype illustrated. (a) At scale $s = 0$, nearby prototypes show similar semantic meanings of scenes (*e.g.* snowfield). (b) At scale $s = 2$, nearby prototypes are semantically related object-level information (*e.g.*, cars).

when pre-trained on COCO dataset [26]. Moreover, SS-PRL surpasses all SSL methods by a significant margin when pre-trained on ImageNet [13] by obtaining 63.8% and 85.4% mAP on COCO [26] and VOC [15], respectively. With the proposed pyramid representation learning, we are able to obtain better results than previous SSL methods [6, 17, 19] that are not designed to handle patch or object-level information. It can be seen that our method also outperforms SSL methods that integrate local information for exploiting data discrimination [42, 46, 49, 53] with large margins.

Multi-label classification in semi-supervised settings. Table 2 compares SS-PRL results with previous SSL methods in the semi-supervised settings of multi-label classification by sampling 1% and 10% labeled data. SS-PRL significantly improves over the state-of-the-art in most settings, showing the prowess when transferred to datasets with limited annotation. We also provide results when fine-tuned with 100% labeled data, where we outperform the randomly initialized model by 20.4% and 24.9% mAP. From this experiment, the effectiveness of our model for multi-label image classification can be successfully confirmed.

Object detection and instance segmentation. The results of object detection and instance segmentation tasks on COCO [26] with $15k$ training iterations are reported in Table 3. SS-PRL outperforms existing general-purpose SSL methods and achieves comparable or even better results with dense prediction based SSL methods when pre-trained on both COCO [26] and ImageNet [13]. The above results exhibit the impressive ability of SS-PRL for downstream dense prediction tasks at object or instance levels.

4.3. Ablation Study

We now conduct ablation studies and parameter analysis to better understand how each component of SS-PRL contributes to the overall performance in downstream multi-label classification tasks. We pre-train models on the COCO [26] dataset and report the mAP on VOC [15] for evaluation. We adopt SS-PRL trained with global images only (*i.e.*, $s = 0$) as our baseline.

Learning of patch-level prototypes. The patch-level prototypes C_s introduced in Section 3.2 provide semantic cues of inherited label dependencies observed in training data, and ensure the feature extractor f_θ to exploit meaningful regional information at each patch scale from an image. In Table 4, we report the linear evaluation results of SS-PRL trained with prototypes C_s learned within and across scales s. It can be seen that mAP drops by 1.1% when prototypes are shared across different scales. This indicates that the prototypes at different patch scales capture the hierarchical semantic/label dependencies of the dataset that is crucial to the downstream tasks with multi-labeled data. Additional visualization for such learned prototype sets will be shown in Figure 2, 3 and discussed in Section 4.4.

Loss functions. To analyze the effectiveness of each developed loss function (*i.e.*, the pyramid representation learning loss L_{pyr} and the cross-scale correlation loss L_{cross}), we conduct an ablation study on the VOC dataset [15]. Table 4 reports the performance of multi-label image classification tasks using linear evaluation protocol. We observe limited performance improvement (+0.3%) when the model is only trained with the pyramid loss L_{pyr}. This is due to the fact that the objectives enforced at each scale are not guaranteed to be mutually related, restricting the discriminative capability. When cross-scale correlation loss L_{cross} is included,

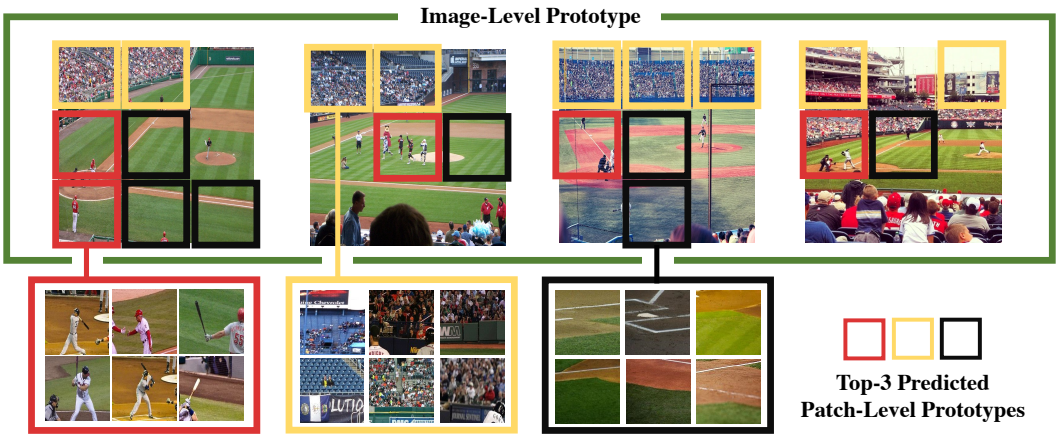


Figure 3. **Correlation between prototypes across different levels.** We randomly choose an image-level (scale 0) prototype from COCO (marked in green) and visualize its top-3 corresponding patch-level prototype predictions at scale 2 (marked in red, yellow and black). With examples of the three selected patch-level prototypes shown at the bottom row, we observe that the patch-level prototypes distinctively represent fine-grained visual concepts which are related to those of the image-level prototypes.

we observe the performance boosts up 0.6% mAP compared to the baseline. This indicates the importance of exploring the correspondence of pyramid features across each level to derive discriminative yet coherent representations. The best results (+1.3%) are obtained by our full SS-PRL which considers both L_{pyr} and L_{cross}, exploiting semantic concepts and correspondence within and across patch scales.

4.4. Visualization

Prototypes learned at each scale. To further visualize and relate the prototypes learned at different scales, we visualize the learned global image-level and local patch-level prototypes using t-SNE [37] and show example results in Figures 2 (a) and (b), respectively. In both cases, nearby prototypes show semantically related visual concepts compared to prototypes far apart. At the image level (a), nearby prototypes share similar semantics of *scenes* (*e.g.* skiing and snowboarding). At the patch level (b), prototypes close to each other show related semantic concepts of *objects* (*e.g.* two different parts of a car). On the contrary, two prototypes far apart represent different semantic meanings at both levels (*e.g.* snowfield vs. grassland and cars vs. ocean). This demonstrates that our method would be able to discover semantic dependencies at different patch scales.

Dependency of prototypes across different levels. Finally, we visualize the correlation dependency between an image-level prototype and the associated patch-level prototypes in Figure 3. Specifically, we divide all the images from a randomly chosen image-level prototype into patches and generate their corresponding patch-level prototype predictions by SS-PRL. All the results of predictions are counted and the top-3 patch-level prototypes that are most frequently predicted will be visualized. The patch-level prototypes correspond to three different iconic elements (*i.e.*, fields, audience, and players) that further mine the fine-grained semantics from the image-level prototype (*i.e.*, baseball games), showing the semantic correspondence of predictions across image and patch levels.

5. Conclusion

In this paper, we presented Self-Supervised Pyramid Representation Learning (SS-PRL) for pre-training deep neural networks, with the goal to facilitate downstream vision tasks at object, instance, or pixel levels. By deriving pyramid representations and learning prototypes at each patch level, our SS-PRL is able to exploit the inherent semantic information within and across image scales via self-supervision. This is achieved by our introduced cross-scale patch-level correlation learning, which aggregates and associates the knowledge across different scales, observing and enforcing the dependency between pyramid representations across patch levels. We conduct a wide range of experiments, including the tasks of multi-label image classification, object detection, and instance segmentation, which support the use of our SS-PRL as a desirable pre-training strategy. With visualization of the learned representations and ablation studies, the design of the proposed SS-PRL can be properly verified.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 110-2634-F-002-052. We also thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] Qaisar Abbas, M Emre Celebi, Carmen Serrano, Irene Fondon Garcia, and Guangzhi Ma. Pattern classification of dermoscopy images: A perceptually uniform model. *Pattern Recognition*, 46(1):86–97, 2013.

[2] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[8] Gang Chen, Yangqiu Song, Fei Wang, and Changshui Zhang. Semi-supervised multi-label learning by solving a sylvester equation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 410–419. SIAM, 2008.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[16] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila 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. 2020.

[18] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021.

[23] Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[28] Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*, 2021.

[29] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6707–6717, 2020.

[30] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

[31] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

[32] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[33] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[34] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. *arXiv preprint arXiv:2111.12933*, 2021.

[35] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021.

[36] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[38] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020.

[40] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[41] Mei Wang, Xiangdong Zhou, and Tat-Seng Chua. Automatic image annotation via local multi-label classification. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 17–26, 2008.

[42] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[43] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[45] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021.

[46] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

[47] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[48] Ning Xu, Yun-Peng Liu, and Xin Geng. Partial multi-label learning with label distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6510–6517, 2020.

[49] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021.

[50] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[52] Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008, 2010.

[53] Yucheng Zhao, Guangting Wang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. Self-supervised visual representations learning by contrastive mask prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10160–10169, 2021.

[54] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017.