

# PatchML: Patch Based Learning for Multi-label Image Classification

Anonymous CVPR submission

## Abstract

Multi-label image classification models are typically trained with image level supervision. In this paper, we investigate whether localization annotations (points and bounding boxes) can improve the performance of image classification on partially annotated datasets. Localization annotations are more expensive than image-level annotations. Thus, at a fixed annotation budget, a model trained on localization annotations will have access to fewer training annotations. Despite this reduction, our **PatchML** approach shows that we can use the localization annotations to provide a better supervision signal. This improves the performance of multi-label image classification models on partially annotated datasets. Furthermore, we propose **PatchCluster (PC)**, which uses class-specific patch “prototypical” features to pseudo-label visually similar unannotated patches. **PatchML** along with (PC) can improve the performance of multi-label image classification model by 4-10% across different budget settings on MS-COCO [31] and Objects-365 v2 [40] datasets.

## 1. Introduction

Over the past few years, deep learning models have made significant progress in the task of multi-label image classification [2, 8, 9, 36]. A multi-label image classifier predicts all classes present in an image. For the image shown in Fig. 1(a), a multi-label classifier should predict cat, sofa, remote, wall and plant. The state of the art approaches *pool* the spatial features from an image encoder into a fixed length feature representation. A *fully connected layer* converts the feature representation into an image-level class prediction. Image level annotations are used as supervision to train accurate class predictions. The fixed length feature representation for this image is expected to learn discriminative features to recognize the above categories while being invariant to changes in appearance, size, spatial context, combinations of different categories, *etc.* Given a large amount of training data for each category, deep learning models [9, 32] have shown the ability to accurately recog-

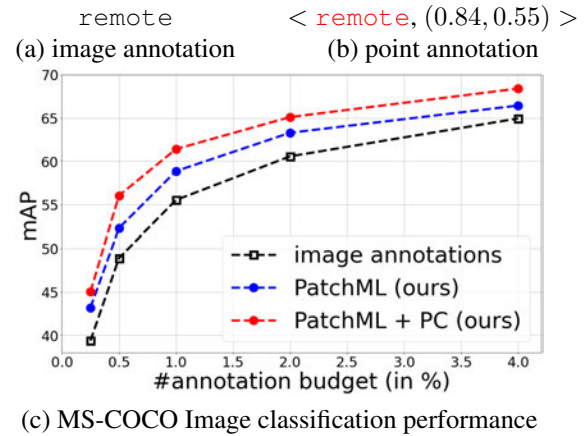


Figure 1. Point supervision outperforms image level supervision for image classification in partial label settings.

nize multiple concepts in an images. However, real world large scale multi-label datasets [16, 27] show that as the scale of images and categories increases, the long tail of rare classes becomes unavoidable [42, 49, 56]. Moreover, it is infeasible to fully annotate these large-scale datasets. For example, OpenImages [27] dataset comprises of 9600 categories across 9M training images, which corresponds to a total of 86B possible annotations [2]. In reality, OpenImages has 60M training annotations.

To ease the bottleneck of learning invariance in spatial contexts and sizes, [13, 55] have proposed to compute the class prediction directly by computing spatial logits. The spatial logits are *pooled* to get the final image level predictions. This has been shown to improve the performance of multi-label image classification model. However, since existing works use image annotations, the supervision in such approaches is at the image level. In our work, we investigate adding supervision directly at each of the patch cells of

the spatial logits grid. At the same annotation budget<sup>1</sup>, we study whether this direct patch-level supervision improves the performance of multi-label image classification models that are trained on partially annotated datasets.

[35, 47] report that localization annotations (in the form of points, bounding boxes, *etc.*) come at a higher cost than annotating image-level labels. For example, in Fig. 1(a), an image level annotation for the `remote` category costs 1s on average. Meanwhile, a point annotation (Fig. 1(b)) in MS-COCO [31] costs 2.4s (1s to annotate at the image level and 1.4s to annotate the actual point). The use of more expensive localization annotations while building large scale datasets would reduce the total number of annotations. We propose **PatchML**, which shows that adding more expensive localization annotations can help multi-label image classification performance at the same annotation budget setting as the image level annotations. In Fig. 1(c), for the partially annotated MS-COCO image classification dataset [2, 31], adding point level annotations (blue) improves performance w.r.t. image annotations (black) at the same annotation budgets. Training with location annotations also improves the ability to localize using spatial feature maps. We use the features from the annotated patches in the spatial grid to create a patch “prototype” for each class, and subsequently use this prototype to pseudo-label visually similar patches in unlabeled images. **PatchCluster (PC)** leads to a further boost in performance (red curve in Fig. 1(c)).

In summary, our main contributions are as follows:

1. We investigate whether the more expensive localization annotations can improve multi-label image classification performance on partially annotated datasets at the same annotation budget as the image level annotations.
2. We propose a loss function to incorporate localization annotations while training multi-label image classification models.
3. We create patch “prototypical” features for each class to pseudo-label un-annotated patches. We do extensive experiments to combine the approaches from semi-supervised and partial label literature.
4. We show that with using single point annotation for a positive class, our approaches can outperform models trained with image annotations by 4-10% on the relative scale across different budget settings on MS-COCO [31] and Objects-365 v2 [40].

## 2. Related Work

**Multi-label image classification on partially annotated datasets.** With the increase in scale of datasets, learning

with **partial annotations** has become a popular area of research in recent years. Partially annotated datasets have a combination of positive, negative and unknown classes. Existing approaches treat the unknowns as *negative* [4, 43], *ignore* [12, 29] or as *noise* [20, 25]. There is a large body of work which looks into modeling class relationships [7, 22, 28] to predict the labels of the unknown classes. Class relationships can be modeled explicitly via graph networks [8, 12, 48] and cross-modal attention [51]. There are also implicit ways to model class relationships. Approaches such as [5, 50] perform matrix completion to reconstruct labels. [10, 24] use generative models to reason about the unknown classes. Recently, [2] proposed a discriminative model to determine the likelihood of an unknown label being present in an image. Combined with a prior, it has a filtering step which chooses to add the unknown label to the negative set. This is the current state of the art model to train multi-label image classification models on partially annotated datasets. The concept of **positive unlabeled** datasets is similar to that of partially annotated datasets. The positive unlabeled datasets have only positive or unknown classes. Most PU learning works have focused on the single label image classification problem, while [11, 23] have explored PU learning for multi-label classification problems. The approaches in this direction assign positive or negative labels to the unknowns by explicitly modeling class relationships [19] and matrix completion [17]. In our work, we train multi-label image classification models on partially annotated datasets. The novelty of our approach lies in the use of localization annotations to train on partially annotated datasets at same annotation budget settings.

**Semi-supervised multi-label classification** [15, 33, 34, 46] is a special case of partially annotated datasets where the model trains on images which have no annotations. Recent advances in single-label semi supervised learning use consistency regularization [39] to create pseudo-labels to train image classification models [6, 41, 52]. These approaches have been modified to work for multi-label semi-supervised learning [21, 37] as well.

**Use of image and localization annotations.** Our approach uses a combination of image labels and localization annotations for the image classification task. **Omni-supervised object detection** approaches [35, 45, 47, 54] also use different annotation types such as image labels, points, scribbles, bounding boxes. However, the goal of such approaches is to improve the performance of object detectors, which requires focusing on a metric that depends on both classification and the bounding box localization. This is a key difference from our work because in image classification, our main focus is on image classification performance and not about localization accuracy. Our work is arguably the closest to the seminal idea of [18] which showed that object localization and image classification can help each other. In

<sup>1</sup>Annotation budget is the time to annotate the dataset.

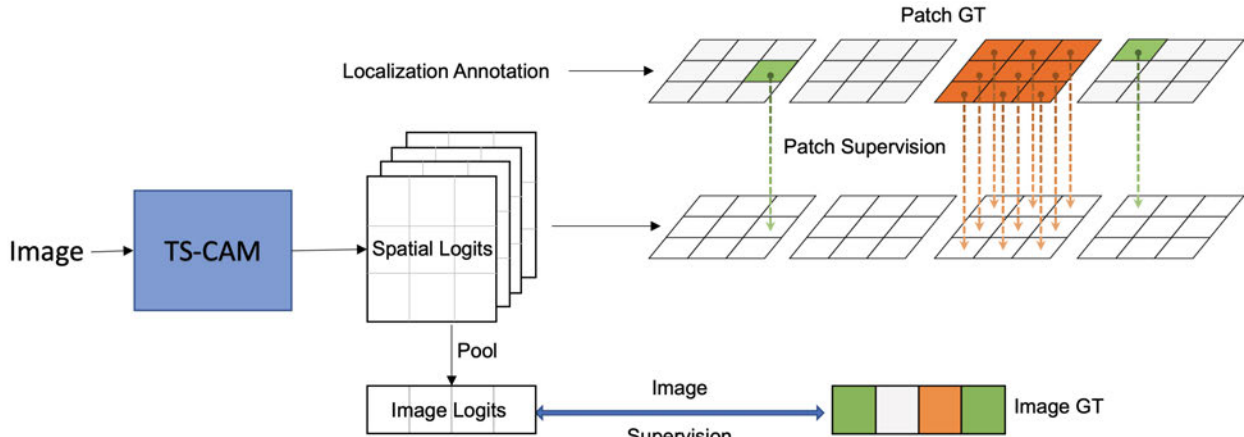


Figure 2. We propose **PatchML** to train with both image and localization annotations. We pass the image through TS-CAM [14] to get spatial logits for each class. Localization annotations are converted to patch level ground truth and supervision is applied to each patch separately. Image level annotations are used to provide supervision to pooled spatial logits. Green, brown and grey denote positive, negative and unknown ground truth, respectively.




			
[31]	80s	84.9s	101s
[40]	365s	372s	395s
	(a) Image	(b) Point	(c) Bounding Box

Figure 3. **Annotation types.** Average annotation costs for MSCOCO [31] and Objects-365 v2 [40] (in s/image).

contrast to their approach, we use large scale partially annotated datasets to train deep learning networks.

**Class activation maps (CAM)** [14, 53] and use of spatial logits [13, 55] has shown that there exists a correlation between image classification performance and localization quality.

### 3. Image classification with localization annotations

Image classification datasets are annotated with labels at the image level where each label is marked as either present or absent. The goal of our approach is to investigate whether adding localization annotations (such as points and boxes) can improve multi-label image classification performance at the same annotation budget (as image level annotations) on partially annotated datasets. We show that image and localization annotations help in achieving better image classification performance. Moreover, we use the features from the annotated patches to create “prototypical” feature representations for each class. We use this feature representation to find visually similar patches from un-annotated images and assign pseudo labels to them.

We discuss the annotation types that we use in Sec. 3.1. We also discuss how we convert the localization annotations

into a discretized patch GT, which can be used to provide supervision to the spatial logits. In Sec. 3.2, we discuss how the patch GT is used to train an image classification model with both image-level and patch-level supervision. Finally, in Sec. 3.3, we discuss details of how we have used visually similar patches from annotated images to generate pseudo-labels on un-annotated images.

#### 3.1. Localization annotations → Patch GT

**Image annotation (Fig. 3a).** Most image classification annotations come in the form of a label for an image. Let the annotation matrix for a dataset,  $\mathcal{D}$  be denoted by  $Y \in \{-1, 0, 1\}^{N \times C}$  for  $N$  images and  $C$  classes. The annotation  $y_{x,c}$  denotes that a class  $c$  can either be present (1), absent (0) or uncertain ( $-1$ ) in an image  $x$ . Each  $y_{x,c}$  is an image-level annotation. Let  $\mathbb{P}_x, \mathbb{N}_x$  and  $\mathbb{U}_x$  denote the set of positive, negative and uncertain classes in  $x$ , respectively.

**Point annotation (Fig. 3b).** Point based annotations are denoted by  $(u_x, v_x, c)$ . This indicates that the class  $c$  is present in image  $x$  at location  $(u_x, v_x)$ .  $u_x, v_x \in [0, 1]$  denote the coordinates of the image relative to image width and height along the X and Y-axes, respectively. The discretized patch in the spatial grid,  $(j, k)$  where the point,  $(u_x, v_x)$  lies in, is assigned the positive (1) label. as shown with green color in Fig. 2. Unless there are other points annotated in the image, the other patches are ignored and assigned the value of -1 (grey colored patch gt in Fig. 2).

**Bounding box annotation (Fig. 3c).** A bounding box is denoted by  $(u_x, v_x, w_x, h_x, c)$ . It is used to denote an image  $x$  has atleast one instance of  $c$  with its bounding box centered at  $(u_x, v_x)$  with its width and height as  $w_x$  and  $h_x$  respectively (where,  $u_x, v_x, w_x, h_x \in [0, 1]$  w.r.t. the image dimensions). We use intersection of the box with the patches to determine which patches to annotate positive for each box annotation.



Please note that the point and bounding box annotation annotations only exist for positive annotations. The negative annotations are only at the image level. Fig. 3 shows the relative costs for different annotation types on MS-COCO [31] and Objects-365 v2 [40] and are based using cost computations from related literature [1, 35, 47]. Based on the patch level GT assignment, for the image, class tuple  $(\mathbf{x}, c)$  we obtain a set of positive patches,  $\mathcal{P}_{\mathbf{x},c}^r$ , negative patches,  $\mathcal{N}_{\mathbf{x},c}^r$  or uncertain patches,  $\mathcal{U}_{\mathbf{x},c}^r$ . We use this notation to define the loss function in our case.

### 3.2. Training with Patch GT

Given a dataset,  $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ , we have a labeled set,  $\mathcal{D}^l = \{\mathbf{x}_i^l, \mathbf{y}_i^l\}_{i=1}^N$  and an unlabeled set,  $\mathcal{D}^u = \{\mathbf{x}_i^u, \mathbf{y}_i^u\}_{i=1}^N$ , s.t.  $\mathcal{D}^l \cup \mathcal{D}^u = \mathcal{D}$ ,  $\mathcal{D}^l \cap \mathcal{D}^u = \emptyset$ . The labeled set,  $\mathcal{D}^l$  consists of images which have at least one GT label annotated. The unlabeled set,  $\mathcal{D}^u$  consists of images which do not have any annotations. Thus,  $y_{\mathbf{x},c} = -1, \forall c, \forall (\mathbf{x}^u, \mathbf{y}^u) \in \mathcal{D}^u$ . Let  $f(\cdot; \theta) \in [0, 1]^{\mathcal{H} \times \mathcal{W} \times C}$  denote a multi-label image classification network, whose weights are denoted by  $\theta$ . The output of this function has dimensions of  $\mathcal{H} \times \mathcal{W} \times C$ , where  $\mathcal{H}, \mathcal{W}$  is the height and width of the spatial output grid from  $f$  respectively.  $C$  is the number of classes for the image classification model. The image level prediction of an image,  $\mathbf{x}$  and class  $c$  is represented as  $\hat{y}_{\mathbf{x},c}^{img} = \text{pool}(f_c(\mathbf{x}; \theta)) \in [0, 1]$ .  $\text{pool}$  function converts the spatial logits to image level logits,  $\hat{\mathbf{y}}_{\mathbf{x}}^{img}$ . Following are the loss functions we have employed in our approaches.

**Image level loss function.** The training data in our case is partially annotated. We use the state of the art class-aware selective loss (CSL-I) [2] for training our multi-label image classification model. The image level class-aware selective loss is as follows:

$$L_{image}^{sup}(\mathbf{x}^l, \mathbf{y}^l, \theta) = \sum_{c^+ \in \mathbb{P}_{\mathbf{x}^l}} \mathcal{L}_F(\hat{y}_{\mathbf{x}^l, c^+}^{img}, \gamma_{image}^+) + \sum_{c^- \in \mathbb{N}_{\mathbf{x}^l}} \mathcal{L}_F(1 - \hat{y}_{\mathbf{x}^l, c^-}^{img}, \gamma_{image}^-) + \sum_{c^* \in \mathbb{U}_{\mathbf{x}^l}} \omega_{c^*} \mathcal{L}_F(1 - \hat{y}_{\mathbf{x}^l, c^*}^{img}, \gamma_{image}^u) \quad (1)$$

where  $\mathcal{L}_F$  is the basic loss term of focal loss defined as  $\mathcal{L}_F(\hat{y}, \gamma) = (1 - \hat{y})^\gamma \log \hat{y}$  [2, 30] and  $\gamma_{image}^+, \gamma_{image}^-, \gamma_{image}^u$  are focus parameters for the positive, negative and un-annotated labels, respectively.  $\omega_{c^*}$  is the selectivity parameter introduced in [2]. The focus parameters are used to alleviate positive, negative and uncertainty class imbalance of partially annotated multi-label datasets. Along with this, the selective loss has a mechanism to selectively ignore unknown annotations, i.e.,  $y_{i,c} =$

$-1$  based on a prior and a label likelihood and treat the rest of the unknown annotations as negatives. All of our experiments except those specifically shown in Tab. 3 are done in ignore mode.

**Patch-level loss function.** We propose to add a patch-level loss,

$$L_{patch}^{sup}(\mathbf{x}^l, \mathbf{y}^l) = \sum_{c^+ \in \mathbb{P}_{\mathbf{x}^l}} \sum_{r \in \mathcal{P}_{i,c^+}^r} w_+^p \mathcal{L}_F(f_{c^+}^r(\mathbf{x}; \theta), \gamma_{patch}^+) + \sum_{c^- \in \mathbb{N}_{\mathbf{x}^l}} \sum_{r \in \mathcal{P}_{i,c^-}^r} w_-^p \mathcal{L}_F(1 - f_{c^-}^r(\mathbf{x}; \theta), \gamma_{patch}^-) + \sum_{c^* \in \mathbb{U}_{\mathbf{x}^l}} \sum_{r \in \mathcal{P}_{i,c^*}^r} w_u^p \mathcal{L}_F(1 - f_{c^*}^r(\mathbf{x}; \theta), \gamma_{patch}^*) \quad (2)$$

where,  $\gamma_{patch}^+, \gamma_{patch}^-, \gamma_{patch}^u$  are focus parameters for the patch-based positive, negative and un-annotated labels, respectively.  $w_+^p, w_-^p, w_u^p$  are the respective weights of the patch based positive, negative and uncertain losses.

When we convert a single point or box annotation to patch based ground truth, we mark all the unannotated patches as unknowns. But, some of these unannotated patches very likely are negatives unless the object covers all the patches. Therefore, we take inspiration from label likelihood in selective loss which effectively ignores the top  $k$  unknowns and treats the rest as negatives. We call this mode the patch selective mode in our experiments (CSL-P).

The overall loss function is as follows:

$$L^{sup}(\mathbf{x}^l, \mathbf{y}^l) = w_{image} L_{image}^{sup}(\mathbf{x}^l, \mathbf{y}^l) + w_{patch} L_{patch}^{sup}(\mathbf{x}^l, \mathbf{y}^l) \quad (3)$$

### 3.3. Training with visually similar patches

Using the loss function defined in (3), we are able to improved the localization quality of the spatial logits. We take an average of the features across with the patches,  $r \in \mathcal{P}_{\mathbf{x}^l, c}^r, \forall \mathbf{x}^l$  s.t.  $c \in \mathbb{P}_{\mathbf{x}^l}$ . This creates a prototype feature for each class,  $c$ . Given an un-annotated image,  $\mathbf{x}^u$ , we compute the cosine distance of the model features corresponding to each patch in the grid with each of the  $C$  prototypes. The patches which have a cosine distance  $\leq d^+$  are pseudo-labeled as positives for the corresponding category. If all the patches in an image have cosine distance  $> d^-$  (where,  $0 \leq d^+ < d^- \leq 1$ ) with a particular class prototype, then that image is annotated with a negative pseudo-label.

With the positive and negative pseudo-labels,  $\mathbf{y}^{u, pl}$ , we formulate the semi-supervised loss as follows:

$$L^{pl}(\mathbf{x}^u, \mathbf{y}^{u, pl}) = w_{image}^{pl} L_{image}^{pl}(\mathbf{x}^u, \mathbf{y}^{u, pl}) + w_{patch}^{pl} L_{patch}^{pl}(\mathbf{x}^u, \mathbf{y}^{u, pl}) \quad (4)$$

In practice, we observe that,  $w_{patch}^{pl} = 0$  gives us best results. While we use patch features to find visually similar patches, we assign the pseudo-label conservatively at the image level. When training with pseudo-labels generated from the visually similar patches, we combine the loss functions in (3) and (4) to have the overall semi-supervised loss function as:

$$L^{ssl}(\mathbf{x}, \mathbf{y}) = \begin{cases} L^{sup}(\mathbf{x}, \mathbf{y}), & (\mathbf{x}, \mathbf{y}) \in \mathcal{D}^l \\ L^{pl}(\mathbf{x}, \mathbf{y}^{pl}), & (\mathbf{x}, \mathbf{y}) \in \mathcal{D}^u \end{cases} \quad (5)$$

## 4. Experiments

We investigate our question of whether localization annotations are useful for training image classification models on challenging multi-label datasets: MS-COCO [31] and Objects-365 v2 [40]. In the subsequent paragraphs, we discuss how we used annotations in the respective datasets to create image-level, point-level and box-level annotations. For all our experiments, we use TS-CAM [14] with a DeiT base [44] backbone. We use a 224x224 image resolution as input to the image classification network. We initialized the weights of the DeiT backbone by training on ImageNet [38] and trained our models for up to 20 epochs using Adam optimizer [26] with early stopping.

**Datasets.** We evaluate the efficacy of our patch based learning approach on simulated settings in MS-COCO [31] and Objects-365 v2 [40] datasets. We follow [2] and use the Random per Annotation (RPA) approach to create partially labeled datasets over various annotation budgets. Roughly 1% and 2% of the data is annotated in realistic partial label datasets such as OpenImages [3] and LVIS [16] respectively. We follow a similar setting while creating partial annotations on MS-COCO and Object-365 datasets. For MS-COCO, we run our image classification models on 0.25%, 0.5%, 1%, 2% and 4% annotation budgets. For Objects365 dataset, we run our classification models on 0.05%, 0.1%, 0.2%, 0.5%, 1% annotation budgets. 100% corresponds to annotating each image in the dataset with all classes at the image level. We use lower percentages in our study to understand the process of building large-scale partially annotated datasets for image classification. To do a study comparing image and localization annotations, we use the annotation cost computation from [35, 47] which are shown in Fig. 3.

For MS-COCO, we generate point level annotations by randomly selecting an instance of a sampled positive class and then randomly selecting a point in its segmentation mask. For Objects-365, we select the center of the bounding box of a randomly selected instance and sample a point from a 2D Gaussian distribution with its mean lying at the center and standard deviation of 0.02 (which roughly corresponds to a quarter of a patch) in each dimension.

**Classification metric.** We used mean average precision to

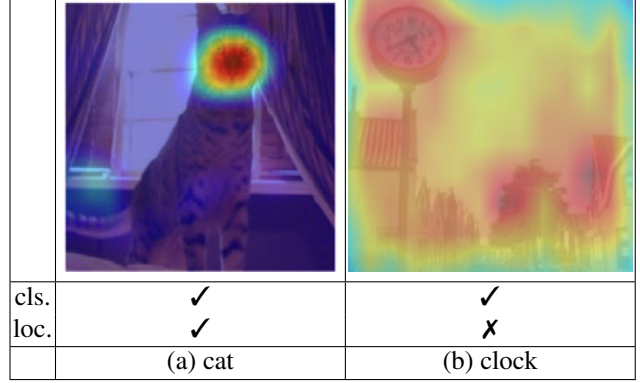


Figure 4. Spatial logits w.r.t. the classification (mAP-cl) and localization (mAP-loc) metrics. ✓: correct, X: incorrect.

evaluate the performance of the multi-label image classification models (mAP-cl).

**Localization Metric.** As opposed to image segmentation or object detection tasks, the goal of the localization metric is to gauge whether the model has localized discriminative parts of the object. As shown in the cat image in Fig. 4(a), the discriminative regions produced by the classification model have localized the head of the cat accurately. The high scoring regions of the attention map are localized well within the body of the cat and does not spill much outside the body of the cat. This is a “good” match with the ground truth mask. In Fig. 4(b), the model predictions spill onto the parts of the image which are not relevant to the clock category. Thus, this is a “bad” match with the ground truth mask. We use this intuition to define the intersection over area of predicted mask (IoA-P) metric. IoA-P between a binary mask of the predicted (thresholded) attention map, **pred** and its GT binary mask, **gt** and is defined as follows:

$$\text{IoA-P}(\text{pred}, \text{gt}) = \frac{\text{pred} \cap \text{gt}}{\text{area of pred}} \quad (6)$$

Please note that unlike the IoU metric, IoA-P metric is not commutative for the **pred** and **gt** arguments. We use different thresholds in the range of [0.6, 0.7, 0.8, 0.9] to binarize the predicted attention masks. We use mAP metric (mAP-loc) at IoA threshold of 0.5 to determine the localization quality.

We show our results and ablation on the MS-COCO dataset in Sec. 4.1, Sec. 4.2, Sec. 4.3 and Objects-365 v2 dataset in Sec. 4.4. Finally, we show visualizations of our model scores with attention heatmaps from the spatial logits in Sec. 4.5.

### 4.1. Localization annotations to classify images

**Image vs single-point vs single-bbox.** In this section, we empirically evaluate whether localization annotations improve the performance of multi-label image classification

Budget (in %)	Ann. Type	Ann. Statistics			mAP-cls $\uparrow$ (in %)	mAP-loc $\uparrow$ (in %)
		Pos.	Neg.	Total		
0.25	I	602	16K	17K	39.37	21.24
	SP	559	15K	16K	<b>43.19</b>	<b>28.54</b>
	SB	471	13K	13K	<u>42.27</u>	<u>26.35</u>
0.50	I	1221	32K	33K	48.89	25.14
	SP	1147	30K	31K	<b>52.30</b>	<b>33.73</b>
	SB	944	25K	26K	<u>50.12</u>	<u>33.43</u>
1.00	I	2440	64K	66K	55.53	29.96
	SP	2314	60K	62K	<b>58.86</b>	<b>36.54</b>
	SB	1910	51K	52K	<u>57.36</u>	<u>33.53</u>
2.00	I	4942	128K	132K	60.57	32.89
	SP	4634	120K	125K	<b>63.29</b>	<b>38.74</b>
	SB	3949	101K	105K	<u>61.45</u>	<u>34.95</u>
4.00	I	9687	255K	265K	64.94	33.31
	SP	9142	240k	250k	<b>66.43</b>	<b>40.59</b>
	SB	7677	202k	210k	<u>65.20</u>	<u>37.21</u>

Table 1. **Annotation types.** Ablation study of *Image* (I), *Single point* (SP) and *Single box* (SB) with varying annotation budgets over the labeled set. The metrics are averaged over 3 different dataset snapshots. (**Bold:** Best, Underline: 2<sup>nd</sup> Best)

results on partially annotated MS-COCO dataset. In Tab. 1, we compare different localization annotation types: single-point (SP) and single-box (SB) with the image-level annotation type (I) on the partially annotated MS-COCO dataset [2, 31]. For each annotation type and budget, we show the annotation statistics such as the number of positive, negative and total annotations. These numbers are used from the ratios mentioned in Fig. 3. In the case of single points (SP), we can observe that the cost of 1.06s reduces the the positive and negative counts w.r.t. the image level annotation scenario (I). But the mAP metric is better by 2-5% for SP across different budget settings. In the case of single boxes (SB), the improvement in performance is 1-2% w.r.t. I. The higher annotation cost of the bounding boxes (1.27s) significantly reduces the number of positive and negative annotations. The reduced number of annotations cause the improvement gap of SB to be less than SP. We demonstrate in the supplementary material that with relaxing the annotation budget constraints and using the same number of positive and negative annotations, SB performs similar to SP. This shows that point annotations can be an inexpensive substitute for the bounding box annotations, while bringing significant improvements in the multi-label image classification performance under the partial label setting. We also observe that SP consistently outperforms the other annotation methods on the mAP-loc metric as well. Interestingly, mAP-loc metric for SP is better than SB. We hypothesize that there could be two reasons for this: (a) point annotations are cheaper than box annotations, (b) point annotations come from the segmentations masks of the object and are hence accurate. Due to its axis-aligned rectangle shape, it can cover non relevant parts of an object. We show some

	Budget (in %)	Approach	Pos. Ann.	Neg. Ann.	Total. Ann.	#Img.	mAP-cls $\uparrow$ (in %)
Point annotations	0.25	SP	559	15K	16K	14.2k	<b>42.04</b>
		AP	1257	14K	16K	13.7k	40.36
	0.50	SP	1147	30k	31k	26.0k	<b>53.22</b>
		AP	2738	29k	32k	25.1k	52.82
	1.00	SP	2314	60k	62k	44.0k	<b>59.50</b>
		AP	5388	58k	63k	42.7k	59.09
	2.00	SP	4634	120k	125k	64.9k	63.41
		AP	11328	115k	126k	63.7k	<b>63.52</b>
	4.00	SP	9142	240k	250k	78.9k	65.93
		AP	14783	157k	171k	78.4k	<b>66.04</b>
Box annotations	0.25	SB	471	13K	13K	12.2k	<b>40.63</b>
		AB	838	10k	11K	9.7k	36.44
	0.50	SB	944	25k	26k	22.6k	<b>49.84</b>
		AB	1832	19k	21k	18.2k	49.12
	1.00	SB	1910	51k	52k	39.1k	<b>58.39</b>
		AB	3752	339k	43k	32.6k	54.75
	2.00	SB	3949	101k	105k	60.0k	<b>61.78</b>
		AB	7715	78k	86k	52.6k	59.92
	4.00	SB	7677	202k	210k	76.5k	<b>64.48</b>
		AB	14783	157k	171k	71.8k	64.17

Table 2. **Single vs multi instance annotations.** Single vs all instances per image for point annotations (SP vs AP) and box annotations (SB vs AB) (**Bold:** Best)

examples of such cases in Fig. 5.

**Single vs multiple instance annotations.** Instead of using single points (SP) and single boxes (SB), we use all points (AP) and all bounding box (AB) instances to annotate each positive class. We show the results of this comparison in Tab. 2. We show the comparison between single point (SP) and all points (AP) in the upper half of the table and single box (SB) vs all box (AB) settings in the second half of the table. In the 0.25% budget setting, the AB setting does not cause a big enough reduction in the negative annotations while increasing the number of positive annotations. This helps in achieving better mAP performance than the SB setting. However, at higher budget settings, the additional cost of annotating all positive instances in an image causes a reduction in the total of annotations and images available during training. This causes a reduction in the mAP metric. This observation is in contrast to the concept of federated datasets [16], where all positive instances are annotated for a label present in the image. We note that this is particular to the task of image classification as demonstrated here, whereas federated datasets [16] have annotation support for higher level semantic tasks such as panoptic segmentation, object detection, etc.

We use PatchML to refer to the single point (SP) setting from now on.

## 4.2. Training with partial and missing annotations

In this section, we study how we can use the un-annotated annotations during training of image classifica-

Budget (in %)	Ann. Type	Supervised Learning	Partial Label Learning		Semi-Supervised Learning	
			+CSL-I [2]	+CSL-P	+FixMatch [37]	+PatchCluster
0.25	I	38.30	38.28	-	39.07	31.23
	SP	42.04	42.06	41.83	42.48	<b>45.02</b>
0.50	I	49.2	49.85	-	51.12	42.85
	SP	53.22	53.19	53.34	<u>53.91</u>	<b>56.04</b>
1.00	I	56.8	55.97	-	57.61	53.83
	SP	59.6	<u>59.76</u>	59.6	59.29	<b>61.41</b>
2.00	I	60.68	60.71	-	61.35	61.38
	SP	63.41	63.41	<u>63.61</u>	63.25	<b>65.1</b>
4.00	I	65.23	65.13	-	66.28	<u>67.29</u>
	SP	65.93	66.39	66.63	66.21	<b>68.39</b>

Table 3. Training with partial annotations and semi supervised learning on unlabeled images. (Bold: Best, Underline: 2<sup>nd</sup> Best).

Budget (in %)	Ann. Type	mAP-cls		
		DeiT (224)	DeiT-TSCAM (224)	DeiT-TSCAM (384)
0.25	I	35.08	38.30	42.48
	SP	-	42.04	45.34
0.50	I	45.59	49.2	53.04
	SP	-	53.22	57.53
1.0	I	53.30	56.80	58.39
	SP	-	59.60	63.63
2.00	I	58.82	60.68	64.50
	SP	-	63.41	68.16
4.00	I	63.76	65.23	69.97
	SP	-	65.93	70.74

Table 4. Ablation on model architectures and input resolution.

tion models in partially annotated settings. In Tab. 3, we show results on two settings: (a) image annotations (I), (b) PatchML. We show results of using ignore mode for the unknown classes. We implement partial label learning baselines from [2], CSL-I along with its patch version, CSL-P. Among semi-supervised methods, we implemented FixMatch [37]. For each of the baselines, we perform grid search across all the hyper-parameters.

Our PatchML+PatchCluster (last column of even rows) consistently outperforms PatchML and PatchML with any other variant of partial label learning (CSL-I/CSL-P) or semi supervised learning (FixMatch). While CSL-I/CSL-P usually gives  $\leq 1\%$  improvement over the ignore mode, FixMatch gives an improvements of 2-4% across different budget settings. However, PatchML+PatchCluster consistently gives an improvement of 3% over the ignore mode across different budget settings.

I+PC setting corresponds to using the image level features instead of the patch features in PatchML+PC. At lower budget setting, I+PC performs worse than even supervised approaches. We suspect that due to lack of enough training samples, the image level feature are not discriminative enough to correctly find its visually similar images. However, at higher budgets, the I+PC outperforms other variants

of partial label learning and FixMatch. For PatchML+PC implementation, we use cosine distance values for the positive pseudo labels assignment as  $d^+$  in the range 0.30–0.35 to get the best performance. Meanwhile, for I+PC implementation,  $d^+ = 0.10$  gives the best performance. A higher value of  $d^+$  indicates that PatchML representations are more robust than features from I.

### 4.3. Ablation on Model architectures

To incorporate localization annotations into an image classification network, we have used the TS-CAM version [14] on top of the DeiT backbone [44]. Since TS-CAM has been designed for weakly supervised object localization, we investigate whether this modification hurts the performance on vanilla DeiT for multi-label classification. The results are shown in Tab. 4. TSCAM version of DeiT outperforms vanilla DeiT by 2-3% across different budget settings. This result corroborates the findings in [55], which have shown that reasoning about semantic attention is useful for training multi-label classification. We also run our experiments with a higher input image resolution (384 instead of 224). At a higher resolution, training with single point helps achieve better classification performance across different budget settings.

### 4.4. Objects-365 v2

In Tab. 5, we show supervised training experiments on Objects-365 v2 [40] dataset which has 365 classes and around 2 million images. Notably, our results from MS-COCO transfer here as well and single point annotations significantly outperform image level annotations within the same annotation budget.

### 4.5. Qualitative results

We show qualitative examples of the semantic attention maps (spatial logits) and compare them across models trained with image and single point annotations in Fig. 5. Point based annotations show superior localization performance in general. In Fig. 5(a-c), we show that point an-



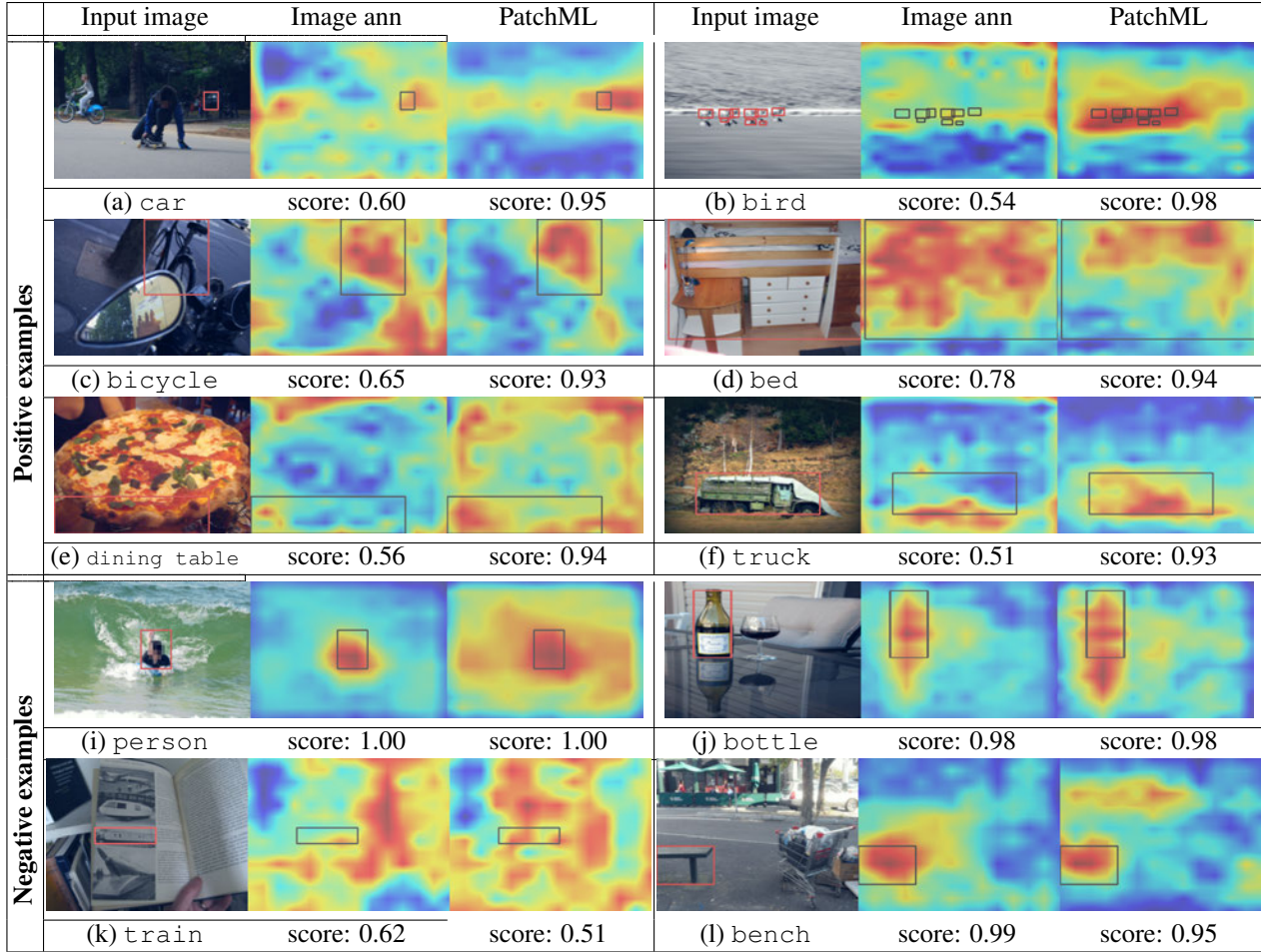


Figure 5. **Qualitative examples.** Each subfigure shows the original input image with the ground truth bounding box (left), the semantic heat map generated from the predictions of a model trained with image annotations (middle) and the semantic heatmap generated from the predictions of a model trained with single point annotations (right).

Budget (%)	Annotation Type	#Img. (M)	mAP-cls $\uparrow$
0.05	I	0.29	5.95
	SP	0.29	<b>10.01</b>
0.10	I	0.53	7.25
	SP	0.53	<b>11.70</b>
0.20	I	0.90	7.76
	SP	0.90	<b>11.38</b>
0.50	I	1.46	10.60
	SP	1.45	<b>15.63</b>
1.00	I	1.70	14.09
	SP	1.70	<b>18.47</b>

Table 5. **PatchML on Objects365v2 (Bold: Best, M=Million)**

notation based model can recognize and localize small objects. In comparison, the image annotation based model performs poorly on these images. In Fig. 5(d), point based models can accurately and confidently predict a heavily occluded bed. Fig. 5(e) is an interesting scenario where the dining table is barely seen. Fig. 5(f) shows PatchML can better capture the extents of an object, whereas image annotation based models seemingly capture only certain parts.

However, for image classification scenarios, this does not cause any issues. In Fig. 5(i-l), we show scenarios where PatchML localization is poor.

## 5. Conclusion

We studied the use of more expensive location annotations, while maintaining the total annotation budget, to train multi-label image classification networks on partially annotated datasets. Our PatchML approach provides direct supervision to patches. Consequently, patch level features can be used to pseudo-label un-annotated images using Patch-Cluster. We show that using a combination of these approaches can improve multi-label image classification performances on partially annotated MS-COCO and Objects365 v2 datasets. Our extensive experiments and ablations show the efficacy of our approach, both numerically and qualitatively in terms of better attention maps.



## References

- [1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 4
- [2] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7
- [3] Rodrigo Benenson and Vittorio Ferrari. From colouring-in to pointillism: revisiting semantic segmentation supervision. *arXiv preprint arXiv:2210.14142*, 2022. 5
- [4] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR*, pages 2801–2808. IEEE, 2011. 2
- [5] Ricardo Cabral, Fernando Torre, Joao P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. *Advances in neural information processing systems*, 24, 2011. 2
- [6] Zhaowei Cai, Avinash Ravichandran, Subhansu Maji, Charles Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021. 2
- [7] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 2
- [8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 1, 2
- [9] Xing Cheng, Hezheng Lin, Xiangyu Wu, Dong Shen, Fan Yang, Honglin Liu, and Nian Shi. Mltr: Multi-label classification with transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 1
- [10] Hong-Min Chu, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Deep generative models for weakly-supervised multi-label classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 400–415, 2018. 2
- [11] Elijah Cole, Oisín Mac Aodha, Titouan Llorca, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. 2
- [12] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, 2019. 2
- [13] Thibaut Durand, Nicolas Thome, and Matthieu Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *CVPR*, pages 4743–4752, 2016. 1, 3
- [14] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 3, 5, 7
- [15] Yuhong Guo and Dale Schuurmans. Semi-supervised multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 355–370. Springer, 2012. 2
- [16] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 5, 6
- [17] Yufei Han, Guolei Sun, Yun Shen, and Xiangliang Zhang. Multi-label learning with highly incomplete data via collaborative embedding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1494–1503, 2018. 2
- [18] Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *2009 IEEE 12th international conference on computer vision*, pages 237–244. IEEE, 2009. 2
- [19] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. Pu learning for matrix completion. In *International conference on machine learning*, pages 2445–2453. PMLR, 2015. 2
- [20] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label learning from noisy labels with non-linear feature transformation. In *Asian Conference on Computer Vision*, pages 404–419. Springer, 2018. 2
- [21] Junxiang Huang, Alexander Huang, Beatriz C Guerra, and Yen-Yun Yu. Percentmatch: Percentile-based dynamic thresholding for multi-label semi-supervised classification. *arXiv preprint arXiv:2208.13946*, 2022. 2
- [22] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *CVPR*, 2020. 2
- [23] Atsushi Kanehira and Tatsuya Harada. Multi-label ranking from positive and unlabeled data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5138–5146, 2016. 2
- [24] Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. *Advances in neural information processing systems*, 25, 2012. 2
- [25] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *CVPR*, 2022. 2
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [27] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Mallocci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label

- and multi-class image classification. *Dataset available from <https://storage.googleapis.com/openimages/web/index.html>*, 2017. 1
- [28] Kaustav Kundu and Joseph Tighe. Exploiting weakly supervised visual patterns to learn from partial annotations. *Advances in Neural Information Processing Systems*, 33:561–572, 2020. 2
- [29] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 2
- [30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4
- [31] 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 *ECCV*, 2014. 1, 2, 3, 4, 5, 6
- [32] 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. 1
- [33] Yi Liu, Rong Jin, and Liu Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, volume 6, pages 421–426, 2006. 2
- [34] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. *Advances in neural information processing systems*, 32, 2019. 2
- [35] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo2: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020. 2, 4, 5
- [36] 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. 1
- [37] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 2, 7
- [38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 5
- [39] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2
- [40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1, 2, 3, 4, 5, 7
- [41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2
- [42] Merrielle Spain and Pietro Perona. Measuring and predicting importance of objects in our visual world. 2007. 1
- [43] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010. 2
- [44] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 5, 7
- [45] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018. 2
- [46] Bo Wang, Zhuowen Tu, and John K Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *Proceedings of the IEEE international conference on computer vision*, pages 425–432, 2013. 2
- [47] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9367–9376, 2022. 2, 4, 5
- [48] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020. 2
- [49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1
- [50] Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. *Advances in neural information processing systems*, 26, 2013. 2
- [51] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12709–12716, 2020. 2
- [52] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2
- [53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on*

*computer vision and pattern recognition*, pages 2921–2929, 2016. [3](#)

- [54] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. [2](#)
- [55] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021. [1](#), [3](#), [7](#)
- [56] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013. [1](#)

1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187