Acknowledging the Unknown for Multi-label Learning with Single Positive Labels

Donghao Zhou[1,2], Pengfei Chen[3], Qiong Wang[1], Guangyong Chen[4(✉)], and
Pheng-Ann Heng[1,5]

[1] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality
Technology, Shenzhen Institute of Advanced Technology,
Chinese Academy of Sciences, Shenzhen, China
`dh.zhou@siat.ac.cn`
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Tencent Technology, Shenzhen, China
[4] Zhejiang Lab, Hangzhou, China
`gychen@zhejianglab.com`
[5] The Chinese University of Hong Kong, Hong Kong, China

Abstract. Due to the difficulty of collecting exhaustive multi-label annotations, multi-label datasets often contain partial labels. We consider an extreme of this weakly supervised learning problem, called single positive multi-label learning (SPML), where each multi-label training image has only one positive label. Traditionally, all unannotated labels are assumed as negative labels in SPML, which introduces false negative labels and causes model training to be dominated by assumed negative labels. In this work, we choose to treat all unannotated labels from an alternative perspective, i.e. acknowledging they are unknown. Hence, we propose entropy-maximization (EM) loss to attain a special gradient regime for providing proper supervision signals. Moreover, we propose asymmetric pseudo-labeling (APL), which adopts asymmetric-tolerance strategies and a self-paced procedure, to cooperate with EM loss and then provide more precise supervision. Experiments show that our method significantly improves performance and achieves state-of-the-art results on all four benchmarks. Code is available at https://github.com/Correr-Zhou/SPML-AckTheUnknown.

Keywords: Weakly supervised learning · Single positive multi-label learning · Entropy maximization · Pseudo-labeling

1 Introduction

Each image is assumed to associate with only one label in *multi-class* classification, while the real world is closer to being *multi-label*, since multiple objects are inherently contained in a realistic scene [51]. Along with the wide adoption of deep learning, recent years have witnessed great progress in multi-label classification [33,52,26]. Unfortunately, it is excessively laborious to collect exhaustive multi-label annotations for a large-scale image dataset [8]. The major reasons

	person	dog	bus	bicycle	apple	boat	laptop	couch
(a)	✓	✗	✓	✓	✗	✗	✗	✗
(b)	✓	✗	?	✓	?	✗	?	?
(c)	✓	?	?	?	?	?	?	?

Fig. 1. Example of the image with (a) full annotations, (b) partial annotations, and (c) "singe positive" annotations. In single positive multi-label learning (SPML), for this image, only one positive label (*person*) is annotated, and the other labels (*dog*, *bus*, *bicycle*, etc) remain unannotated

are that the number of potential labels for an image can be large [27] and human annotators tend to ignore rare or small objects [48], which could cause a large amount of label noise. In fact, some publicly available multi-label datasets, including MS-COCO [25] and Open Images [19], are considered to contain only partial labels of images [14]. Therefore, it is of scientific interest to explore multi-label learning with partially labeled datasets.

In this paper, we focus on an extreme of this weakly supervised learning problem in multi-label classification called *single positive multi-label learning* (*SPML*) [7], where only one single positive label (and no other positive or negative labels) is annotated for each training image (see Fig. 1). SPML is a worth-exploring variant of multi-label learning for the following reasons: First, this is a common setting in numerous practical scenarios. For instance, when collecting images from the web, the only annotated positive label of an image is from its query [22]. Second, some multi-class datasets like ImageNet [35] actually contain images that associates with more than one label [2,38,42], while the multi-class training fashion penalizes any predictions beyond the single annotated label [51]. Third, in-depth research of multi-label learning with this minimal supervision could significantly relax the annotation requirement of large-scale multi-label datasets, which helps to reduce expensive annotation costs [9].

Due to the lack of supervision from all negative labels and most positive labels, SPML remains a challenging problem. Unfortunately, if ignoring unannotated labels, the model trained with only positive labels would collapse to a trivial solution. Since negative labels are generally the overwhelming majority of multi-label annotations, the traditional method is to assume all unannotated labels are negative to provide supervision signals, which is commonly regarded as the baseline of SPML [7]. However, this assumption would introduce false negative labels and make model training be dominated by assumed negative labels, which damages model generalization and causes a substantial performance drop. Therefore, unannotated labels should be *properly treated* in model training. In this work, instead of making any unrealistic assumptions, we choose to treat all unannotated labels from a quite different perspective in SPML, i.e. *acknowledging the fact that they are unknown*, and propose a simple but efficient method in response to the issues above. Our main contributions are summarized as follows:

1. Motivated by the idea of acknowledging the unknown, we propose a novel loss function called *entropy-maximization (EM) loss* for SPML, which aims to maximize the entropy of predicted probabilities for unannotated labels. Besides, we reveal the mechanism of EM loss with gradient-based analysis, which demonstrates that EM loss is capable of tackling inherent issues of the traditional method due to its special gradient regime.
2. For more precise supervision, we propose *asymmetric pseudo-labeling (APL)* to cooperate with EM loss. Considering the positive-negative label imbalance of unannotated labels, APL adopts asymmetric-tolerance strategies for positive and negative pseudo-labels with a self-paced procedure, which can generate highly accurate pseudo-labels to further boost performance.
3. Experiments performed on four popular multi-label datasets, i.e. PASCAL VOC [10], MS-COCO [25], NUS-WIDE [6], and CUB [43], show that our method achieves state-of-the-art results on all four benchmarks. Moreover, detailed ablation study and further analysis verified the effectiveness and rationality of the proposed method.

2 Related Work

Multi-label Learning with Weak Supervision. There are several multi-label tasks similar to single positive multi-label learning (SPML). Multi-label learning with missing labels (MLML) assumes that some labels are annotated and the others are "missing" for each training image [40]. Different methods have been proposed for MLML, including treating missing labels as negative ones [44,29,47,39], label matrix completion [4,50,3], learning image-label similarities [14], etc. Semi-supervised multi-label learning (SSML) assumes training data is a subset of images with exhaustive labels and a large number of images without any labels [27]. Common solutions to SSML include non-negative matrix factorization [28], label propagation [45], aligning image features [46], etc. However, these two tasks are quite different from SPML, where all training images are labeled but each one only contains one positive label. Within the limit of this special setting, most standard methods of these two tasks, like consistency regularization [37] and learning label correlations [9,53], are not applicable to SPML.

Entropy Min-/Maximization. The idea of entropy min-/maximization has been widely exploited in various fields. Grandvalet et al. proposed entropy minimization regularization for semi-supervised learning [11], and Pereyra et al. have shown that penalizing low entropy also acts as a strong regularizer in multi-class classification [32]. In domain adaptation, entropy min-/maximization has been adopted to adversarially optimize the model [36] and learn from target domain data [49]. Besides, entropy maximization has been shown to improve exploration in reinforcement learning [30]. To the best of our knowledge, our entropy-maximization (EM) loss is the *first* loss function to utilize entropy maximization for unannotated labels and demonstrate its effectiveness.

Pseudo-Labeling. The goal of pseudo-labeling is to assign labels to under-labeled samples with a trained model [20]. There are many methods to implement pseudo-labeling, including exploiting neighborhood graphs [15], performing

clustering [41], estimating prediction uncertainty [9,34], etc. However, existing pseudo-labeling techniques fail to consider the positive-negative label imbalance of unannotated labels in SPML, which would result in a large amount of label noise. By adopting asymmetric-tolerance strategies for positive and negative pseudo-labels, our asymmetric pseudo-labeling (APL) can significantly reduce wrong pseudo-labels, allowing for better performance.

3 Methodology

In this section, we start by giving a formal definition of single positive multi-label learning (SPML) in Sec. 3.1. Then, we would introduce our entropy-maximization (EM) loss designed for SPML in Sec. 3.2. Finally, we would describe the proposed asymmetric pseudo-labeling (APL) in Sec. 3.3.

3.1 Problem Definition

Let $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ denotes a partially labeled multi-label dataset with N images, where each image $\mathbf{x}^{(n)}$ from input space $\mathcal{X}$ is associated with a vector of labels $\mathbf{y}^{(n)}$ from the label space $\mathcal{Y} = \{-1, 0, 1\}^C$ with C classes. Let $y_c^{(n)}$ be the c-th entry of $\mathbf{y}^{(n)}$, where $y_c^{(n)} = 1$ indicates $\mathbf{x}^{(n)}$ is relevant to the c-th class (i.e. $\mathbf{x}^{(n)}$ contains a positive label of the c-th class) and $y_c^{(n)} = -1$ otherwise. Besides, $y_c^{(n)} = 0$ indicates that the label of the c-th class is unannotated for $\mathbf{x}^{(n)}$. In SPML, $\mathbf{y}^{(n)}$ satisfies $y_c^{(n)} \in \{0, 1\}$ and $\sum_{c=1}^{C} \mathbb{1}_{[y_c^{(n)}=1]} = 1$, where $\mathbb{1}_{[\cdot]}$ denotes the indicator function, meaning that only one positive label (and no other positive or negative labels) is annotated for each training image.

The goal of SPML is to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}'$ from $\mathcal{D}$, where $\mathcal{Y}' = \{-1, 1\}^C$ is the ground-truth label space for $\mathcal{X}$. Defining f as a deep model $f(\cdot; \mathbf{w})$ with the weights $\mathbf{w}$, the standard training approach is to treat each label prediction as an independent binary classification and solve $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{f}^{(n)}, \mathbf{y}^{(n)})$, where $\mathbf{f}^{(n)} = f(\mathbf{x}^{(n)}; \mathbf{w}) \in [0, 1]^C$ denotes the predicted probabilities for $\mathbf{x}^{(n)}$ activated by the sigmoid function, and $\mathcal{L} : [0, 1]^C \times \mathcal{Y} \to \mathbb{R}$ denotes a loss function that can handle labels $\mathbf{y}^{(n)}$. Let $f_c^{(n)}$ denotes the c-th entry of $\mathbf{f}^{(n)}$.

3.2 Entropy-Maximization Loss

Since negative labels are generally the overwhelming majority of multi-label annotations, the widely recognized baseline of SPML is *assuming-negative (AN) loss* [7], which assumes all unannotated labels as negative ones for training. Given a image-label pair $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$, AN loss is defined as

$$\mathcal{L}_{\text{AN}}(\mathbf{f}^{(n)}, \mathbf{y}^{(n)}) = -\frac{1}{C} \sum_{c=1}^{C} [\mathbb{1}_{[y_c^{(n)}=1]} \log(f_c^{(n)}) + \mathbb{1}_{[y_c^{(n)}=0]} \log(1 - f_c^{(n)})]. \quad (1)$$
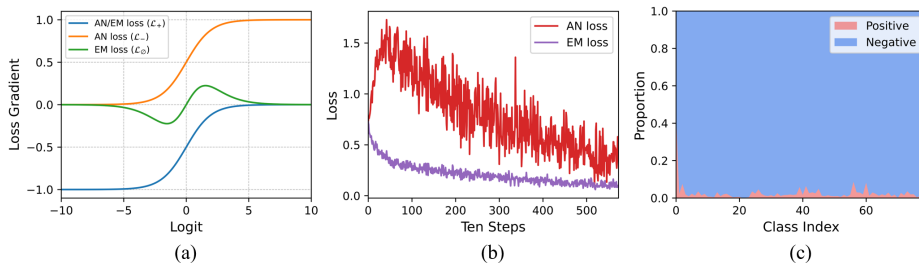
Fig. 2. (a): Gradients of $\mathcal{L}_+$, $\mathcal{L}_-$ and $\mathcal{L}_\varnothing$. For fair comparison, α of $\mathcal{L}_\varnothing$ is set to 1. (b): Training losses (averaged in 3 runs) of annotated positive labels (i.e. $\mathcal{L}_+$) on VOC from the model trained with AN and EM loss, where α is also set to 1. Note that $\mathcal{L}_+$ of AN loss is more unstable and would increase in early training. (c): Proportions of unannotated positive and negative labels of each class on COCO

To better understand the influence of AN loss on model training, we perform gradient-based analysis for AN loss, which is commonly used to deeply study loss functions [24,33]. For convenience, let $g = g_c^{(n)}$ denotes the output logit of the c-th class for $\mathbf{x}^{(n)}$, $p = 1/(1 + e^{-g})$ (i.e. the sigmoid function) denotes the predicted probability, and $\mathcal{L}_+ = -\log(p)$ (resp. $\mathcal{L}_- = -\log(1 - p)$) denotes the loss of the annotated positive label (resp. assumed negative label) in AN loss. The gradients of AN loss for the logit g are

$$\begin{cases} \dfrac{\partial \mathcal{L}_+}{\partial g} = \dfrac{\partial \mathcal{L}_+}{\partial p}\dfrac{\partial p}{\partial g} = \dfrac{-e^{-g}}{1 + e^{-g}}, & y_c^{(n)} = 1, \\ \dfrac{\partial \mathcal{L}_-}{\partial g} = \dfrac{\partial \mathcal{L}_-}{\partial p}\dfrac{\partial p}{\partial g} = \dfrac{1}{1 + e^{-g}}, & y_c^{(n)} = 0, \end{cases} \tag{2}$$

which shows that AN loss utilizes the same gradient regime for $\mathcal{L}_+$ and $\mathcal{L}_-$ (see Fig. 2(a)). Thus, the model trained with AN loss would treat annotated positive labels and assumed negative labels equally during training, which would result in bad performance in SPML due to the following issues:

1. *Dominance of Assumed Negative Labels*: For each training sample, assumed negative labels are $C - 1$ times more than annotated positive ones. Hence, affected by the same gradient regime, the model training would be dominated by assumed negative labels until they are well fitted, which hinders the model from learning from annotated positive labels (see Fig. 2(b)).
2. *Introduced Label Noise*: It is unavoidable that there are positive labels wrongly assumed as negative ones in AN loss. Due to the same gradient regime, false negative labels and true positive labels would severely confuse the model.
3. *Over-Suppression for Confident Positive Predictions*: During training, the model may output a confident positive prediction for an assumed negative label that may be a ground-truth positive one. Unfortunately, AN loss would provide a large gradient for it (see Fig. 2(a)), aiming to attain a smaller logit.

Despite there are some common approaches to reducing the impacts of these issues (e.g. adopting regularization techniques), the gradient regime of AN loss, caused by its unrealistic assumption, would still damage model generalization. Therefore, unannotated labels need to be properly treated during training, or more specifically, be treated with *a better gradient regime*. Instead of assuming they are negative, we choose to treat them from an alternative perspective, i.e. acknowledging they are unknown. Motivated by that, we propose *entropy-maximization (EM) loss* for SPML, which is defined as

$$\mathcal{L}_{\text{EM}}(\mathbf{f}^{(n)}, \mathbf{y}^{(n)}) = -\frac{1}{C} \sum_{c=1}^{C} [\mathbb{1}_{[y_c^{(n)}=1]} \log(f_c^{(n)}) + \mathbb{1}_{[y_c^{(n)}=0]} \alpha H(f_c^{(n)})], \quad (3)$$

$$H(f_c^{(n)}) = -[f_c^{(n)} \log(f_c^{(n)}) + (1 - f_c^{(n)}) \log(1 - f_c^{(n)})], \quad (4)$$

where α is a hyperparameter to down-weight the strength of entropy maximization. Note that EM loss adopts entropy maximization as a *separate* loss term (instead of an extra auxiliary regularizer) for unannotated labels to independently provide supervision signals for them, which is a unique loss design.

Likewise, we reveal the mechanism of EM loss by analyzing its gradients. Let $\mathcal{L}_{\varnothing} = \alpha[p \log p + (1 - p) \log (1 - p)]$ denotes the loss of the unannotated label in EM loss. The gradients of EM loss for the logit g are

$$\begin{cases} \dfrac{\partial \mathcal{L}_+}{\partial g} = \dfrac{\partial \mathcal{L}_+}{\partial p} \dfrac{\partial p}{\partial g} = \dfrac{-e^{-g}}{1 + e^{-g}}, & y_c^{(n)} = 1, \\ \dfrac{\partial \mathcal{L}_{\varnothing}}{\partial g} = \dfrac{\partial \mathcal{L}_{\varnothing}}{\partial p} \dfrac{\partial p}{\partial g} = \dfrac{-\alpha e^{-g} \log e^{-g}}{(1 + e^{-g})^2}, & y_c^{(n)} = 0, \end{cases} \quad (5)$$

which shows that EM loss utilizes the same gradient regime as AN loss for annotated positive labels and adopts a quite different one for unannotated labels (see Fig. 2(a)). Compared with AN loss, the special gradient regime of EM loss can lead to the following beneficial training behaviours (see Appendix B for additional empirical evidence of these claims):

1. *Learning from Annotated Labels Preferentially*: In early training, the model would produce ambiguous predictions before being well trained, i.e. predicted probabilities (resp. output logits) would be near 0.5 (resp. 0). At this time, the gradients of $\mathcal{L}_+$ (resp. $\mathcal{L}_{\varnothing}$) would be relatively large (resp. small) (see Fig. 2(a)), leading the model to preferentially learn from annotated positive labels (see Fig. 2(b)). Unlike simply down-weighting $\mathcal{L}_-$, EM loss tends to keep the predictions of unannotated labels ambiguous, and thus is capable of providing small gradients for them throughout training.
2. *Mitigating the Effect of Label Noise*: There are no positive labels wrongly regarded as negative ones in EM loss, which prevents the model from producing incorrect negative predictions. Moreover, though unannotated positive labels still exist, the model trained with EM loss would mainly focus on the annotated ones due to the special gradient regime.

Algorithm 1 Asymmetric Pseudo-Labeling

Input: Training set $\mathcal{D}$ and model f_{T_w} trained with Eq. 3 for T_w epochs
Parameter: Total training epoch T_t, sample proportion $\theta\%$ and loss weight β
Output: Well-trained model f_i

1: $i \leftarrow T_\mathrm{w}$, $\theta'\% \leftarrow \theta\%/(T_\mathrm{t} - T_\mathrm{w})$
2: **repeat**
3:    Generate pseudo-labels using f_i by following Eq. 6
4:    Train f_{i+1} from f_i with Eq. 8
5:    $i \leftarrow i + 1$
6: **until** early stopping **or** $i = T_\mathrm{t}$
7: **return** f_i

---

3. *Maintaining Confident Positive Predictions*: After preferentially learning from annotated positive labels, the model is more likely to output confident positive predictions for potential positive labels. When the logit is large enough, the gradient of $\mathcal{L}_\varnothing$ would decline and even approach 0 as the logit goes larger (see Fig. 2(a)), which helps to maintain these confident positive predictions.

3.3   Asymmetric Pseudo-Labeling

Exploiting a better gradient regime, the model trained with EM loss can produce more distinguishable predictions for unannotated positive and negative labels (see Fig. 4(a)&(b)). According to these informative predictions, we can assign pseudo-labels to partially labeled images, aiming to provide more precise supervision for the model and then further improve its performance. Generally, pseudo-labeling is performed by setting a predefined rule to select potential positive or negative labels, e.g. setting a score threshold [20] or sample proportion [5]. In pseudo-labeling, there is a natural trade-off between the provided supervision and the introduced noise, since generating more pseudo-labels is often accompanied by introducing more noise. Specifically, adopting a *low-tolerance* strategy (high score threshold or low sample proportion) would provide less accurate pseudo-labels, whereas adopting a *high-tolerance* strategy (low score threshold or high sample proportion) would introduce more noisy pseudo-labels.

Different from the other tasks, in SPML, the amounts of unannotated positive and negative labels are quite imbalanced on most classes (see Fig. 2(c), refer to Appendix F.1 for more examples). Considering this enormous imbalance, we choose to treat positive and negative pseudo-labels with *asymmetric tolerance* and thus propose *asymmetric pseudo-labeling (APL)* to cooperate with EM loss, aiming to generate relatively sufficient and accurate pseudo-labels. For negative pseudo-labels, APL adopts a *high-tolerance* strategy with a high sample proportion $\theta\% = 90\%$, which means that 90% of unannotated labels on a class could be selected as negative pseudo-labels at most during training. For positive ones, APL directly ignores them to avoid introducing any noisy positive pseudo-labels that would significantly reduce performance (see Fig. 3(b)), which can be regarded as an extreme *low-tolerance* strategy.

As shown in Algorithm 1, after warm-up for T_{w} epochs, pseudo-labels would be generated for subsequent epochs until training is over, which is equivalent to exploiting them to gradually fine-tune the model. Moreover, inspired by self-paced learning [18], instead of generating all the pseudo-labels at once time, APL would progressively generate more confident ones of them. Specifically, negative pseudo-labels are generated with a sample proportion $\theta'\% = \theta\%/(T_{\text{t}} - T_{\text{w}})$ in each epoch, where T_{t} is the total training epoch. For the c-th class, APL firstly sorts the predicted probability set $U_c = \{f_c^{(n)} | y_c^{(n)} = 0, n = 1, 2, ..., N\}$ in ascending order. Then, APL assigns negative pseudo-labels of the c-th class to the images with the $\theta'\%$ lowest predicted probabilities in U_c. For an partially labeled image $\mathbf{x}^{(n)}$, negative pseudo-labels are generated by following

$$y_c^{(n)} \triangleq -\mathbb{1}_{[f_c^{(n)} \in U_c']}, \quad \forall c \in \{c | y_c^{(n)} = 0\}, \tag{6}$$

where $U_c' \subset U_c$ denotes the $\theta'\%$ lowest predicted probability subset of U_c, and $y_c^{(n)} \triangleq -1$ means that $\mathbf{x}^{(n)}$ is assigned the negative pseudo-label of the c-th class. Moreover, instead of assigning hard labels to images, APL records $s_c^{(n)} \triangleq f_c^{(n)}$ as soft labels for them to mitigate the effect of noisy negative pseudo-labels. Thus, the loss term of the negative pseudo-label is

$$\mathcal{L}_*(f_c^{(n)}, s_c^{(n)}) = s_c^{(n)} \log(f_c^{(n)}) + (1 - s_c^{(n)}) \log(1 - f_c^{(n)}). \tag{7}$$

As for the remaining unannotated labels, we still provide supervision signals for them by entropy maximization. Finally, adopting β as a hyperparameter to down-weight $\mathcal{L}_*$, EM loss combined with APL is formulated as

$$\mathcal{L}_{\text{EM+APL}}(\mathbf{f}^{(n)}, \mathbf{y}^{(n)}) = -\frac{1}{C} \sum_{c=1}^{C} [\mathbb{1}_{[y_c^{(n)}=1]} \log(f_c^{(n)}) + \mathbb{1}_{[y_c^{(n)}=0]} \alpha H(f_c^{(n)})$$
$$+ \mathbb{1}_{[y_c^{(n)}=-1]} \beta \mathcal{L}_*(f_c^{(n)}, s_c^{(n)})]. \tag{8}$$

4   Experiments

4.1   Experimental Setup

Datasets. Since there are no existing datasets totally in line with the setting of single positive multi-label learning (SPML), following [7], we use popular large-scale multi-label datasets to simulate "single positive" datasets by discarding annotations, which can guarantee access to all ground-truth labels for analyzing training phenomena and evaluating performance. Specifically, after withholding 20% of the training images for validation, we randomly select one positive label to keep and treat the other labels are unannotated ones for each training image, which is performed once for each dataset. Note that the validation and test sets remain fully labeled. We use PASCAL VOC 2012 (VOC) [10], MS-COCO 2014 (COCO) [25], NUS-WIDE (NUS) [6], and CUB-200-2011 (CUB) [43] in our experiments. More dataset descriptions are contained in Appendix C.1.

Implementation Details. For fair comparison, we follow the main implementation of [7]. Specifically, ResNet-50 [12] pretrained on the ImageNet [35] is adopted as the backbone, followed by global average pooling [23] and a $2048 \times C$ fully connection layer that outputs the predicted probabilities of C classes. All input images are resized to 448×448, and the training images are horizontally flipped with a probability of 0.5 for data augmentation. For each method, we conduct a grid search of batch sizes in $\{8, 16\}$ and learning rates in $\{1e - 2, 1e - 3, 1e - 4, 1e - 5\}$, and then select the hyperparameters with the best mean average precision (mAP) on the validation set. Moreover, we use Adam [17] as our optimizer to train all models for 10 epochs, and early stopping is performed when mAP on the validation set descends. Each experiment runs three times, and the mean and standard deviations of mAP are reported.

Comparing Methods. We compare our method to the following methods: AN loss (assuming-negative loss), EntMin (entropy minimization regularization) [11], Focal loss [24], ASL (asymmetric loss) [33], ROLE (regularized online label estimation) [7], and ROLE+LI (ROLE combined with the "LinearInit") [7]. Besides, we also compare our method to the baseline (i.e. AN loss) with the following improvement: DW (down-weighting $\mathcal{L}_-$), L1R/L2R (l_1/l_2 regularization), LS (label smoothing), and N-LS (label smoothing for only $\mathcal{L}_-$). Moreover, we also report the performance of the models trained with binary cross-entropy (BCE) loss on the fully labeled dataset and on a special partially labeled dataset (i.e. 1 P. & All N. in Table 1). To simulate the lacking of positive supervision, BCE loss would ignore all unannotated labels (i.e. their losses are not computed) in the latter case. Refer to Appendix C.2 and C.3 for the details of the comparing methods and the hyperparameters of each method respectively.

4.2 Results and Discussion

The experimental results on four SPML benchmarks are reported in Table 1. It can be observed that the assumption of AN loss (i.e. assuming all unannotated labels are negative) causes a significant performance drop, e.g. 9.81% and 12.59% mAP decrements than being trained with full annotations (i.e. All P. & All N.) on NUS and CUB respectively. When adopting some improvement on AN loss, the impact of this assumption can be mitigated. Down-weighting $\mathcal{L}_-$ of Eq. 1 can help to achieve competitive performance on CUB, and label smoothing is more useful on VOC. Note that adopting label smoothing for only assumed negative labels can achieve better performance, which verifies the importance of learning from annotated positive labels. Whereas, a common regularization technique, i.e. l_1/l_2 regularization, seems to under-perform in the SPML setting.

As for the other comparing methods, Focal loss and ASL can achieve good performance on NUS and CUB, which can reduce the effect of the dominance of assumed negative labels but still can not address this issue well. The state-of-the-art method of SPML (i.e. ROLE), which adopts a jointly trained label estimator to perform label estimation during training, has competitive performance in SPML, e.g. achieving 88.77% mAP on VOC. Moreover, the "LinearInit" training fashion can be integrated with ROLE to further improve its performance.

Table 1. Experimental results of our method and the comparing methods on four SPML benchmarks. The best performance of the methods in the SPML setting (i.e. 1 P. & 0 N.) is marked in bold and the second best is marked in italic

Ann. Labels	Methods	VOC	COCO	NUS	CUB
All P. & All N.	BCE loss	89.42±0.27	76.78±0.13	52.08±0.20	30.90±0.64
1 P. & All N.	BCE loss	87.60±0.31	71.39±0.19	46.45±0.27	20.65±1.11
	AN loss	85.89±0.38	64.92±0.19	42.27±0.56	18.31±0.47
	DW	86.98±0.36	67.59±0.11	45.71±0.23	19.15±0.56
	L1R	85.97±0.31	64.44±0.20	42.15±0.46	17.59±1.82
	L2R	85.96±0.36	64.41±0.24	42.72±0.12	17.71±1.79
	LS	87.90±0.21	67.15±0.13	43.77±0.29	16.26±0.45
1 P. & 0 N.	N-LS	88.12±0.32	67.15±0.10	43.86±0.54	16.82±0.42
	EntMin	53.16±2.81	32.52±5.55	19.38±3.64	13.08±0.15
	Focal loss	87.59±0.58	68.79±0.14	47.00±0.14	19.80±0.30
	ASL	87.76±0.51	68.78±0.32	46.93±0.30	18.81±0.48
	ROLE	87.77±0.22	67.04±0.19	41.63±0.35	13.66±0.24
	ROLE+LI	88.26±0.21	69.12±0.13	45.98±0.26	14.86±0.72
1 P. & 0 N.	EM loss	*89.09±0.17*	*70.70±0.31*	*47.15±0.11*	*20.85±0.42*
	EM loss+APL	**89.19±0.31**	**70.87±0.23**	**47.59±0.22**	**21.84±0.34**

However, this state-of-the-art method still achieves a poor classification result on CUB. It is worth noting that entropy minimization regularization, which is the opposite of our entropy maximization (EM) loss, achieves badly poor and unstable results on all datasets. It demonstrates that this widely adopted method of semi-supervised learning does not applicable to SPML, since only positive labels are annotated for multi-label training images in this special setting.

It can be observed that our EM loss, which adopts entropy maximization for unannotated labels, outperforms the existing methods on all four SPML benchmarks, e.g. achieving 89.09% and 70.70% mAP on VOC and COCO respectively. Note that ROLE uses a learnable label matrix to act as the label estimator, which significantly increases the memory consumption during training. Moreover, this method requires access to the average number of positive labels, which is practically unavailable in the SPML setting. Compared with ROLE, our EM loss is cost-free and does not introduce any extra learnable parameters in light of its simplicity. Besides, the model trained with EM loss can even outperform the model trained with additional supervision (i.e. 1 P. & All N.) in three benchmarks (i.e. VOC, NUS, and CUB). Moreover, it can also be observed that asymmetric pseudo-labeling (APL) can further boost the performance of the model trained with EM loss, and finally our EM loss+APL achieves state-of-the-art results on all four benchmarks. Especially, EM loss+APL can achieve a 6.98% mAP increment than ROLE+LI on CUB, and achieves 89.19% mAP on VOC, which even approaches the result of being trained with full annotations.

Table 2. Left: Experimental results of AN loss, ROLE, and EM loss on VOC with different backbones. Right: Precision (averaged in 3 runs) of pseudo-labels generated by APL cooperating with AN and EM loss on four multi-label datasets

Methods	ResNet-34	ResNet-50	ResNet-101
AN loss	84.60±0.29	85.89±0.38	86.59±0.19
ROLE	85.44±0.16	87.77±0.22	88.37±0.07
EM loss	**86.58±0.25**	**89.09±0.17**	**88.80±0.11**

Losses	VOC	COCO	NUS	CUB
AN loss	96.44%	96.32%	98.16%	89.82%
EM loss	**99.57%**	**99.67%**	**99.89%**	**91.58%**

Table 3. Experimental results of APL and the pseudo-labeling variants on CUB. ✓ indicates that the corresponding technique are adopted. Note that the first row denotes that pseudo-labeling is not performed and the last row is our APL

High Prop.	Soft Label	DW	Pos. PL	mAP
				20.85±0.42
	✓	✓		20.95±0.34
✓		✓		20.99±0.38
✓	✓			21.00±0.44
✓	✓	✓	✓	20.82±0.68
✓	✓	✓		**21.84±0.34**

4.3   Ablation Study

Robustness of EM Loss to α. To investigate how performance is affected by α, we present the experimental results of EM loss with different α in Fig. 3(a). The performance has a peak at $\alpha = 0.2$ on VOC and an improper α might cause a performance drop. However, the variation range of mAP remains slight and EM loss can still achieve relatively high performance with different α. Especially, EM loss with all different α can still outperform AN loss, and even approaches the result of being trained with full annotations (i.e. BCE loss) when $\alpha = 0.2$, showing that performance can be improved by selecting a proper α.

Effectiveness of EM Loss on Different Backbones. To study the impact of the special gradient regime of EM loss when the scale of the deep model varies, we consider ResNet with different depths (i.e. ResNet-34, ResNet-50, and ResNet-101) as backbones to be trained with EM loss. As reported in Table. 2, EM loss achieves the best performance on these backbones as we expect. However, the performance of EM loss on ResNet-50 is even better than on ResNet-101, which can be reserved as an interesting work for further exploration.

Evaluation of APL. We perform detailed ablation study for APL (see Fig. 3(b) and Table 3). Specifically, we remove or add several key components of APL to obtain the following five variants: 1) w/o pseudo-labeling: the model is trained only with EM loss. 2) w/o a high sample proportion: $\theta\%$ is set to 10%. 3) w/o soft labels: the pseudo-labeled samples are assigned hard pseudo-labels. 4) w/o down-weighting: β is set to 1. 5) w/ positive pseudo-labeling: positive pseudo-labeling is also performed in the same way as negative pseudo-labeling, and its sample proportion is set to 10%. The results on the test set of CUB show that APL can achieve more stable training and each component adopted by APL contributes
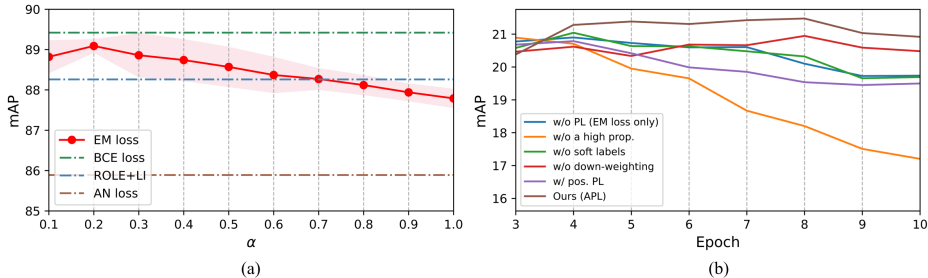
Fig. 3. (a): Experimental results (averaged in 3 runs) of EM loss with different α on VOC. (b): Performance curves (averaged in 3 runs) of APL and its variants on CUB. Note that the x-axis ticks start by 3 since the first 3 epochs are for warm-up

to performance improvement (see Appendix D for detailed analysis). Moreover, we report the precision of pseudo-labels generated by APL (see Table. 2, we disable early stopping for fair comparison), which shows that APL can generate pseudo-labels with high precision ($\approx 99\%$) in EM loss on most datasets. Though unannotated labels are already assumed as negative ones in AN loss, we also try adopting APL in AN loss to select potential ground-truth negative labels for reference, showing the effectiveness of asymmetric-tolerance strategies and the capability of EM loss to reduce false negative predictions.

4.4   Further Analysis

Class-wise Performance Improvement. To investigate whether EM loss is beneficial to most classes of the "single positive" datasets, we report class-wise average precision (AP) improvement of EM loss (over AN loss) on the test sets of VOC and COCO. As shown in Fig. 4(c), our EM loss can improve performance on most classes, or more specifically, on 95% classes of VOC and 93.75% classes of COCO, which demonstrates its general effectiveness.

Distinguishability of Model Predictions. A well-generalizing model should be able to produce informative predictions for unannotated labels, i.e. the predicted probabilities of unannotated positive and negative labels are more distinguishable. The divergences between these predicted probabilities can be quantitatively calculated by the Wasserstein distance. As shown in Fig. 4(a), the model trained with EM loss usually achieves a much larger Wasserstein distance, which means that it can yield more distinguishable predictions on most classes. Furthermore, we also visualize the predicted probabilities on the test set. For the most common "person" class of COCO, as shown in Fig. 4(b), EM loss also contributes to more distinguishable predictions for the test images. We also provide more visualization results of other classes in Appendix F.2.

Generalization Evaluation by Loss Landscapes. Considering the sharpness of loss minima, we can further explore why EM loss helps to better generalization. Given established principles of statistics and information theory, the study in [1]
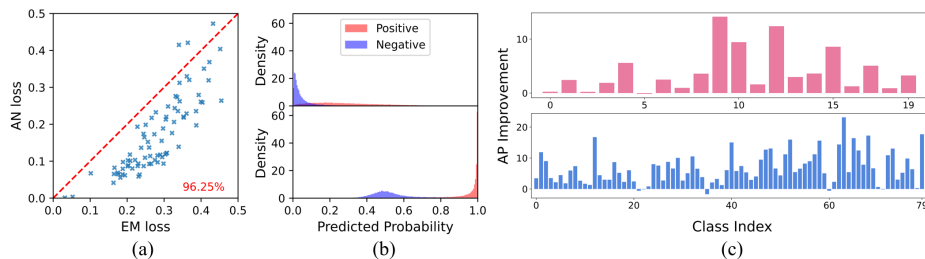
Fig. 4. (a): Wasserstein distances between the distributions of the predicted probabilities for unannotated positive and negative labels on the training set of COCO, where each point indicates one class. We present the proportion of points that are under the red line. (b): Densities of predicted probabilities on the most common "person" class of the test set of COCO, produced by the models trained with AN (*top*) and EM (*bottom*) loss. (c): Class-wise AP improvement of EM loss on VOC (*top*) and COCO (*bottom*)
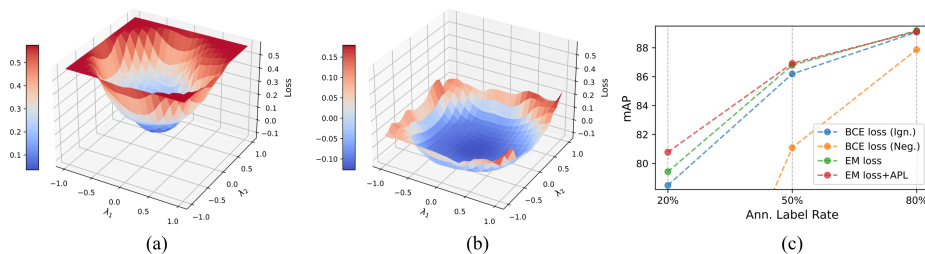
Fig. 5. (a)&(b): 3D visualization of loss landscapes around the minima of the models trained with AN and EM loss on VOC respectively. Note that we set the z-axis to the same scale for comparing the sharpness. (c): Comparison of our method and two baselines in the MLML setting. We randomly discard a percentage of annotations per image as done in [9] to simulate MLML datasets

shows that a flat minimum has lower mutual information between training data and model parameters and thus leads to better generalization. Several previous works [13,16,31] also show that a flat minimum can make the converged model generalize well. Using the technique in [21], we visualize the loss landscapes in Fig. 5. It can be observed that the model trained with AN (resp. EM) loss would converge to a sharp (resp. flatter) minimum, which further explains why the model trained with EM loss can acquire better generalization.

Qualitative results. As shown in Fig. 6, we can observe that the model trained with AN loss would produce low predicted probabilities for almost all unannotated positive labels, especially for small and rare objects in the images. Besides, the predictions of the test images also show that AN loss would severely damage model generalization. However, the model trained with EM loss can produce relatively confident positive predictions for positive labels of training and test images, even for those small and rare objects, which further demonstrates the remarkable effectiveness of EM loss.
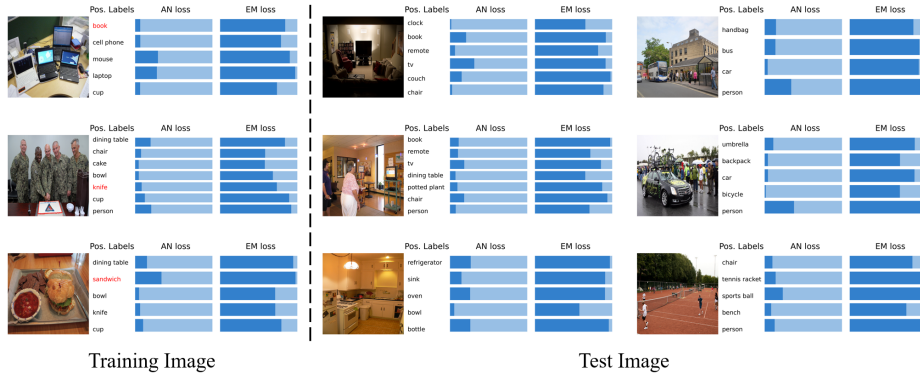
Fig. 6. Examples of training and test images of COCO. Note that the annotated positive labels are marked in red for each training image, and the blue bars indicate the predicted probabilities produced by the models trained with AN and EM loss

Performance in a More General Scenario. Our work focuses on learning from single positive labels, but we are curious about whether it can perform well in a more general setting, i.e. multi-label learning with missing labels (MLML). Thus, we report the preliminary experimental results in Fig. 5. Note that the baselines are BCE loss with unannotated labels being ignored or regarded as negative ones and EM loss acts like BCE loss with unannotated labels being supervised by entropy maximization in MLML. We find that our method can still achieve performance improvement, which shows that it can be generalized to other similar tasks and is well worth further exploration in the future.

5 Conclusions

In this paper, we aim to solve an extreme variant of multi-label learning called single positive multi-label learning (SPML) which is a challenging but under-explored problem. Instead of assuming all unannotated labels are negative as traditionally done, we choose to treat them from an alternative perspective, i.e. acknowledging they are unknown. Hence, we propose entropy-maximization (EM) loss and asymmetric pseudo-labeling (APL) for SPML. Experiments show that our method achieves state-of-the-art results on all four benchmarks. We believe that our method can serve as a stronger baseline for SPML and expect this paper can motivate more future work of weakly supervised learning.

References

1. Achille, A., Soatto, S.: Emergence of invariance and disentanglement in deep representations. The Journal of Machine Learning Research **19**(1), 1947–1980 (2018)
2. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020)
3. Cabral, R., De la Torre, F., Costeira, J.P., Bernardino, A.: Matrix completion for weakly-supervised multi-label image classification. IEEE transactions on pattern analysis and machine intelligence **37**(1), 121–135 (2014)
4. Cabral, R.S., Torre, F., Costeira, J.P., Bernardino, A.: Matrix completion for multi-label image classification. In: Advances in neural information processing systems. pp. 190–198 (2011)
5. Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., Chen, K., Wang, Z.: Self-pu: Self boosted and calibrated positive-unlabeled training. In: International Conference on Machine Learning. pp. 1510–1519. PMLR (2020)
6. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
7. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
8. Deng, J., Russakovsky, O., Krause, J., Bernstein, M.S., Berg, A., Fei-Fei, L.: Scalable multi-label annotation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3099–3102 (2014)
9. Durand, T., Mehrasa, N., Mori, G.: Learning a deep convnet for multi-label classification with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 647–657 (2019)
10. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep **8**, 5 (2011)
11. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hochreiter, S., Schmidhuber, J.: Flat minima. Neural computation **9**(1), 1–42 (1997)
14. Huynh, D., Elhamifar, E.: Interactive multi-label cnn learning with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9423–9432 (2020)
15. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
16. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kumar, M., Packer, B., Koller, D.: Self-paced learning for latent variable models. Advances in neural information processing systems **23**, 1189–1197 (2010)

19. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)

20. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)

21. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. arXiv preprint arXiv:1712.09913 (2017)

22. Li, W., Wang, L., Li, W., Agustsson, E., Berent, J., Gupta, A., Sukthankar, R., Van Gool, L.: Webvision challenge: Visual learning and understanding with web data. arXiv preprint arXiv:1705.05640 (2017)

23. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)

24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

26. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)

27. Liu, W., Wang, H., Shen, X., Tsang, I.: The emerging trends of multi-label learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

28. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: AAAI. vol. 6, pp. 421–426 (2006)

29. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)

30. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. pp. 1928–1937. PMLR (2016)

31. Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N.: Exploring generalization in deep learning. arXiv preprint arXiv:1706.08947 (2017)

32. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)

33. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)

34. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329 (2021)

35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)

36. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

37. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)
38. Stock, P., Cisse, M.: Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 498–512 (2018)
39. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
40. Sun, Y.Y., Zhang, Y., Zhou, Z.H.: Multi-label learning with weak label. In: Twenty-fourth AAAI conference on artificial intelligence (2010)
41. Taherkhani, F., Dabouei, A., Soleymani, S., Dawson, J., Nasrabadi, N.M.: Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12267–12277 (June 2021)
42. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., Madry, A.: From imagenet to image classification: Contextualizing progress on benchmarks. In: International Conference on Machine Learning. pp. 9625–9635. PMLR (2020)
43. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., California Institute of Technology (2011)
44. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2016)
45. Wang, L., Ding, Z., Fu, Y.: Adaptive graph guided embedding for multi-label annotation. In: IJCAI (2018)
46. Wang, L., Liu, Y., Qin, C., Sun, G., Fu, Y.: Dual relation semi-supervised multi-label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6227–6234 (2020)
47. Wang, Q., Shen, B., Wang, S., Li, L., Si, L.: Binary codes embedding for fast image tagging with incomplete labels. In: European Conference on Computer Vision. pp. 425–439. Springer (2014)
48. Wolfe, J.M., Horowitz, T.S., Kenner, N.M.: Rare items often missed in visual searches. Nature **435**(7041), 439–440 (2005)
49. Wu, X., Zhou, Q., Yang, Z., Zhao, C., Latecki, L.J., et al.: Entropy minimization vs. diversity maximization for domain adaptation. arXiv preprint arXiv:2002.01690 (2020)
50. Xu, M., Jin, R., Zhou, Z.H.: Speedup matrix completion with side information: Application to multi-label learning. In: Advances in neural information processing systems. pp. 2301–2309 (2013)
51. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2340–2350 (2021)
52. Zhu, K., Wu, J.: Residual attention: A simple but effective method for multi-label recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 184–193 (2021)
53. Zhu, Y., Kwok, J.T., Zhou, Z.H.: Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering **30**(6), 1081–1094 (2017)