The following publication K. Wang, X. Yan, D. Zhang, L. Zhang and L. Lin, "Towards Human-Machine Cooperation: Self-Supervised Sample Mining for Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1605-1613 is available at https://doi.org/10.1109/CVPR.2018.00173.

Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection

Keze Wang[1,2] Xiaopeng Yan[1] Dongyu Zhang[1*] Lei Zhang[2] Liang Lin[1,3]
[1]Sun Yat-sen University [2]The Hong Kong Polytechnic University [3]SenseTime Research
kezewang@gmail.com; yanxp3@mail2.sysu.edu.cn; zhangdy27@mail.sysu.edu.cn
cslzhang@comp.polyu.edu.hk; linliang@ieee.org

Abstract

Though quite challenging, leveraging large-scale unlabeled or partially labeled images in a cost-effective way has increasingly attracted interests for its great importance to computer vision. To tackle this problem, many Active Learning (AL) methods have been developed. However, these methods mainly define their sample selection criteria within a single image context, leading to the suboptimal robustness and impractical solution for large-scale object detection. In this paper, aiming to remedy the drawbacks of existing AL methods, we present a principled Self-supervised Sample Mining (SSM) process accounting for the real challenges in object detection. Specifically, our SSM process concentrates on automatically discovering and pseudo-labeling reliable region proposals for enhancing the object detector via the introduced cross image validation, i.e., pasting these proposals into different labeled images to comprehensively measure their values under different image contexts. By resorting to the SSM process, we propose a new AL framework for gradually incorporating unlabeled or partially labeled data into the model learning while minimizing the annotating effort of users. Extensive experiments on two public benchmarks clearly demonstrate our proposed framework can achieve the comparable performance to the state-of-the-art methods with significantly fewer annotations.

1. Introduction

In the past decade, object detection has gained incredible improvements both in accuracy and efficiency, benefiting from the remarkable success of deep Convolutional Neural Nets (CNNs) [19][33][14]. Through producing candidate object regions of input images, object detection is converted into the region classification task, e.g., R-CNN [12]. Recently, more powerful neural network architectures such as ResNet [14] have further pushed the object detection performance into new records. Behind these successes, massive data collection and annotation such as MS-COCO [25] are indispensable yet quite expensive. Under such a circumstance, there is an increasing demand of leveraging large-scale unlabeled data to promote the detection performance in an incremental learning manner. However, to achieve this goal, there remain two technical issues: i) Object annotation for training is usually labor-intensive. Compared with other visual recognition task (e.g., image classification), annotating object requests to provide both the category label and bounding box of an object. In order to reduce the burden of active users, it is highly required to develop human-machine cooperation based approaches to self-annotate most of the unlabeled data; ii) Picking out the training samples that are advantageous to boost the detection performance is a nontrivial task. As figured out in [32, 15], existing detection benchmarks usually contain an overwhelming number of "easy" examples and a small number of "hard" ones (i.e., informative samples with various illuminations, deformations, occlusions and other intra-class variations). Utilizing these "hard" samples is a key to train the model more effectively and efficiently. However, as pointed out in [39], due to following a long-tail distribution, these examples are quite uncommon. Hence, it is a sophisticated task to find "hard" yet informative samples.

To address the aforementioned issues, we investigate sample mining techniques to incrementally improve object detectors with minimal user effort. Recently, Active Learning (AL) [9, 22, 31, 23, 26] proposed to progressively select and annotate most informative unlabeled samples for user annotation to boost the model. Hence, the sample selection criteria play a crucial role in AL, and are typically defined according to the prediction certainty (confidence) or other informative criteria like diversity of samples. Recently, many AL methods [37, 24, 36] have been developed for training deep convolutional neural networks (CNNs). However, their sample selection criteria are dominantly performed under a single sample context. This make them sen-

[*] Corresponding author is Dongyu Zhang.

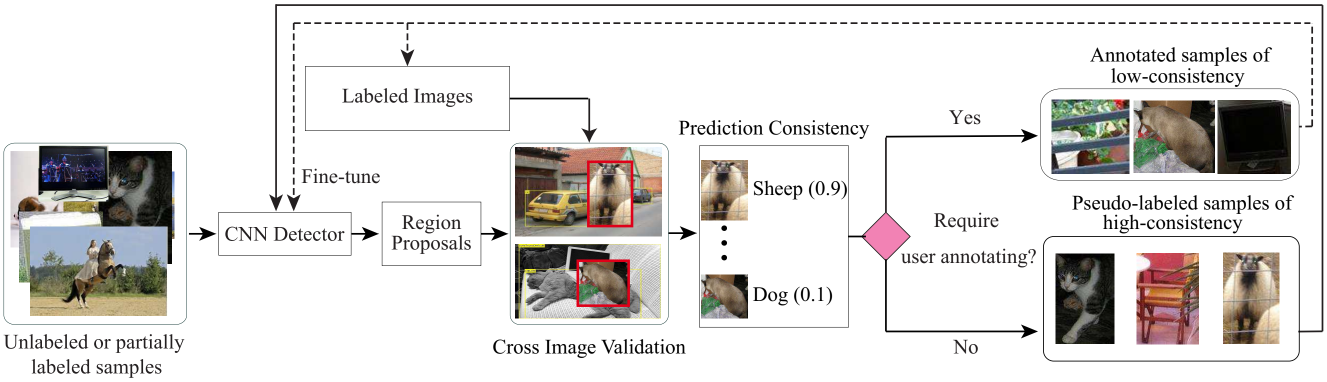


Figure 1. The pipeline of the proposed framework with Self-supervised Sample Mining (SSM) process for object detection. Our framework includes stages of high-consistency sample pseudo-labeling via the SSM and low-consistency sample selecting via the AL, where the arrows represent the work-flow, the full lines denote data flow in each mini-batch based training iteration, and the dash lines represent data are processed intermittently. As shown, our framework presents a rational pipeline for improving object detection from unlabeled and partially labeled images by automatically distinguishing high-consistency region proposals, which can be easily and faithfully recognized by computers after the cross image validation, and low-consistency ones, which can be labeled by active users in an interactive manner.

sitive to the bias of classifiers and the unbalance of samples.

Attempting to overcome the above-mentioned drawback of the existing AL approaches, this paper develops a Self-supervised Sample Mining (SSM) process for automatically mining valuable samples in terms of boosting performance of object detectors. Our developed SSM process is motivated by the recently popular self-supervised learning [10, 5, 38] technique. In stead of designing to learn a optimal visual representation as [38, 5], we concentrate on developing a rational pipeline of using the self-supervision to significantly decrease the amount of user annotations for improving detection performance. In the proposed SSM process, given the region proposals from unlabeled or partially labeled images, we evaluate their estimation consistency by performing the cross image validation, i.e., pasting them into different annotated images to validate its prediction consistency by the up-to-date object detector. Note that, to avoid ambiguity, the images for validation are randomly picked out from the labeled samples that do not belong to the estimated category of the under-processing proposal. Through a simple ranking mechanism, those incorrectly pseudo-labeled proposals have a large chance to be filtered due to the challenges inside various image contexts. In this way, the bias of classifiers and unbalance of samples can be effectively alleviated. Then, we propose to provisionally assign disposable pseudo-labels to the ones with high estimation consistency, and retrain the detectors within each mini-batch iteration. Since the pseudo-annotations may still contain errors, small amount of user interactions is necessary to keep our SSM under well control.

By resorting to our SSM, we further propose a novel incremental learning framework to gradually incorporate unlabeled samples to enhance object detectors, as illustrated in Fig. 1. In our framework, inspired by the recently proposed techniques: Curriculum Learning (CL) [2] and Self-paced Learning (SPL) [20][18], we formulate the combining of the SSM and the AL as a concise optimization problem. Specifically, the SSM or AL process in our framework can jointly collaborate with each other. This is done by imposing a set of latent variables to progressively include samples into training. These variables determine whether a sample should be selected for pseudo-labeling or annotating. Meanwhile, the misleading of pseudo-labeled errors can be suppressed since the sample selection criterion is progressively optimized together with the batch-based incremental learning. In fact, the ambiguity of incorrectly annotated samples by users can also be eliminated, thanks to the correction of the majority pseudo-labeled samples. Hence, our SSM can further improve the robustness of classifiers against noisy samples/outliers in the pursuit of detection accuracy.

The **main contributions** of this work are two-fold. First, we propose a novel self-supervised process for automatically discovering and pseudo-labeling reliable region proposals via the cross image validation, which is compatible with the mini-batch based training for large-scale practical scenarios. Second, through fusing the proposed SSM process with the AL, we propose a principled framework with a concise optimization formulation and an alternative optimization algorithm. Extensive experiments demonstrate that our proposed framework can not only achieve a clear performance gain by mining additional unlabeled data, but also outperform the dominantly state-of-the-art methods with significantly fewer annotations.

2. Related Work

Active Learning. This branch of works mainly focuses on the sample selection strategy, i.e., how to pick out the most informative unlabeled samples for annotation. One of the most common strategies is the certainty-based selection [22, 34], in which the certainties are measured according to the prediction confidence on new unlabeled samples. The diversity of the selected instance over the unlabeled data has been also considered in [3]. Recently, Elhamifar *et al.* [6] proposed to consider both the uncertainty and diversity measure via convex programming. Freytag *et al.* [8] presented a concept that generalizes previous methods based on the expected model change and incorporates the underlying data distribution. Vijayanarasimhan *et al.* [36] proposed a novel active learning approach in crowdsourcing settings for live learning of object detectors, in which the system autonomously identifies the most uncertain instances via a hashing based solution. Rhee et al. [29] proposed to improve object detection performance by leveraging a collaborative sampling strategy, which integrates the uncertainty and diversity criteria from the AL and the feature similarity measurement of semi-supervised learning philosophy. However, these mentioned AL approaches usually emphasize those low-confidence samples (e.g., uncertain or diverse samples) while ignoring the rest majority of high-confidence samples. More recently, attempting to leverage these ignored samples, several works [24, 37] have been proposed to progressively select the minority of most informative samples and pseudo-label the majority of high prediction confidence samples for network fine-tuning. Though these approaches have achieved promising performances, they have limitations due to that their defined hyper-parameters for pseudo-labeling is empirically set and updated within a single image context. Furthermore, these methods do not support mini-batch based training. Therefore, none of them has successfully proved their capability on handling large-scale object detection task.

Self-paced Learning. Inspired by the cognitive principle of humans/animals, Bengio *et al.* [2] initialized the concept of curriculum learning (CL), in which a model is learned by gradually including samples into training from easy to complex. To make it more implementable, Kumar *et al.* [20] substantially prompted this learning philosophy by formulating the CL principle as a concise optimization model named self-paced learning (SPL). Recently, several works [18, 16, 41] provided more comprehensive understanding of the learning insight underlying CL/SPL, and formulated the learning model as a general optimization problem.

Based on this framework, multiple SPL variants [16, 17, 41, 18] have been proposed for object detection. Lee *et al.* [21] introduced a self-paced approach to focus on the easiest instances first, and progressively expands its repertoire to include more complex objects. Sangineto *et al.* [30] presented a self-paced learning protocol for object detection that iteratively selects the most reliable images and boxes according to class-specific confidence levels and inter-classifier competitions. Dong *et al.* [1] proposed an object detection framework that uses only a few bounding box labels per category by consistently alternating between detector amelioration and reliable sample selection.

Self-supervised Learning. Aiming at training the feature representation without additional manual labeling, self-supervised learning (SSL) has first been introduced in [28] for vowel class recognition, and further extended for object extraction in [10]. Recently, plenty of SSL methods[5, 38] have been proposed, e.g., Wang *et al.* [38] proposed to employ visual tracking to provide the supervision for learning visual representations among thousands of unlabeled videos. Doersch *et al.* [5] investigated multiple self-supervised methods to encourage the network to factorize the information in its representation. Different from these methods that focus on learning an optimal visual representation, our SSM intends to use the self-supervision to mine valuable information from unlabeled and partially labeled data.

3. Self-supervised Sample Mining

3.1. Formulation

In the context of object detection, suppose that we have n region proposals from m object categories. Denote the training sample set as $\mathcal{D} = \{x_i\}_{i=1}^n \subset R^d$, where x_i is the i-th region proposal generated from the training images. We have m detectors/classifiers (including the background) for recognizing each proposal by the one-vs-rest strategy. Correspondingly, we denote the label set of x_i as $\mathbf{y}_i = \{y_i\}_{j=1}^m$, where $y_i^{(j)}$ corresponds to the label of x_i for the j-th object category. i.e., if $y_i^{(j)} = 1$, this means that x_i is categorized as an instance of the j-th object category. We should give two necessary remarks on our problem setting. One is that most of sample labels $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ are unknown and need to be completed in the learning process. The initially annotated images are denoted by $\mathbf{I}$. The other remark is that the data $\{x_i\}_{i=1}^n$ might possibly be fed into the system in an incremental way. This means that the data scale might be consistently growing. The whole loss function for our proposed framework with SSM process is formulated as follows:

$$Loss = \mathcal{L}_{loc}(\mathbf{W}) + \mathcal{L}_{cls}^{\text{AL}}(\mathbf{W}) + \mathcal{L}_{cls}^{\text{SSM}}(\mathbf{W}, \mathbf{V}), \qquad (1)$$

where $\mathcal{L}_{loc}(\mathbf{W})$ denotes the bounding box regression loss defined in [11]. $\mathcal{L}_{cls}^{\text{AL}}(\mathbf{W})$ and $\mathcal{L}_{cls}^{\text{SSM}}(\mathbf{W}, \mathbf{V})$ imply the classification loss for the AL and SSM processes, respectively. We define the AL process as $\mathcal{L}_{cls}^{\text{AL}}(\mathbf{W}) = \frac{1}{|\Omega_I|}\sum_{i\in\Omega_I}\sum_{j=1}^m \ell_j(x_i, \mathbf{W})$, where Ω_I denotes the labeled proposals from the currently annotated image I ($I \in \mathbf{I}$). $\ell_j(x_i, \mathbf{W})$ means the softmax loss of the proposal

x_i in the j-th classifier:

$$\ell_j(x_i, \mathbf{W}) = -\Big(\frac{1+y_i^{(j)}}{2}\log\phi_j(x_i;\mathbf{W}) + \frac{1-y_i^{(j)}}{2}\log(1-\phi_j(x_i;\mathbf{W}))\Big),$$

where $\mathbf{W}$ represents the parameter of the CNN for all m categories (including background), $\phi_j(x_i;\mathbf{W})$ denotes the probability of belonging to the j-th category for each region proposal x_i.

To adaptively select x_i for pseudo-labeling to update its $\mathbf{y}_i$, our SSM process introduces a set of latent weight variables, i.e., $\mathbf{V} = \{\mathbf{v}^{(j)}\}_{j=1}^m = \{[v_1^{(j)}, \cdots, v_n^{(j)}]^T\}_{j=1}^m$, and is formulated as:

$$\mathcal{L}_{cls}^{\text{SSM}}(\mathbf{W},\mathbf{V}) = \frac{1}{|\overline{\Omega_I}|}\sum_{i\in\overline{\Omega_I}}\sum_{j=1}^{m} v_i^{(j)}\ell_j(x_i,\mathbf{W}) + R(x_i, v_i^{(j)},\mathbf{W})$$
$$\text{s.t.} \quad \sum_{j=1}^{m}|y_i^{(j)}+1| \le 2, y_i^{(j)} \in \{-1,1\}, \tag{2}$$

where $\overline{\Omega_I}$ denotes the unlabeled proposals from the unlabeled or partially labeled image I. The regularization function $R(\cdot)$ penalizes the sample weights linearly in terms of the loss. In this paper, we utilize the hard weighting regularizer due to its well adaptability to complex scenarios. The hard weighting regularizer is defined as:

$$R(x_i, v_i^{(j)}, \mathbf{W}) = -f(x_i,\mathbf{W})v_i^{(j)}. \tag{3}$$

Finally, we can directly calculate $v_i^{(j)}$ as:

$$v_i^{(j)} = \begin{cases} 1, & \ell_j(x_i,\mathbf{W}) \le f(x_i,\mathbf{W}), \\ 0, & otherwise. \end{cases} \tag{4}$$

In contrast to the existing works [24, 37] that rely on only an empirical hyper-parameter for each category to control the loss tolerance, we exploit a sample-dependent manner, i.e., the cross image validation $f(\cdot)$, to include samples into training. Therefore, our model can be considered as a self-supervised self-paced learning framework. As illustrated in Fig. 2, the cross image validation is regarded as the estimation consistency of reliable region proposals. Specifically, $f(\cdot)$ is defined as:

$$f(x_i,\mathbf{W}) = \frac{\lambda}{|\Omega_I^{\bar{j}}|}\sum_{p\in\Omega_I^{\bar{j}}}\mathbf{1}\big(\text{IoU}\big(B_I(x_i), B_I(x_p)\big) \ge \gamma\big)\phi_j(x_p;\mathbf{W}), \tag{5}$$

where $\Omega_I^{\bar{j}}$ represents the labeled region proposals from the annotated image I without j-th category objects for consistency evaluation. λ denotes the pace parameter. $B_I(x_i)$ denotes the bounding box of the proposal x_i in the selected image I, while $\text{IoU}\big(B_I(x_i), B_I(x_p)\big)$ implies the intersection of union between two bounding boxes $B_I(x_i)$ and $B_I(x_p)$. Note that, γ represents the threshold parameter to identify whether these two bounding boxes correspond to the same object, and it is set to 0.5 in all experiments according to the evaluation protocol of object detection. $\mathbf{1}(\cdot)$ is the indicator function. If the proposal x_p generated by the up-to-date detector from the newly pasted image includes the same object as x_i, i.e., $\text{IoU}\big(B_I(x_i), B_I(x_p)\big) \ge \gamma$, then we calculate its estimation consistency value $\phi_j(x_p;\mathbf{W})$, which denotes the possibility of x_p being the j-th category during the cross image validation.

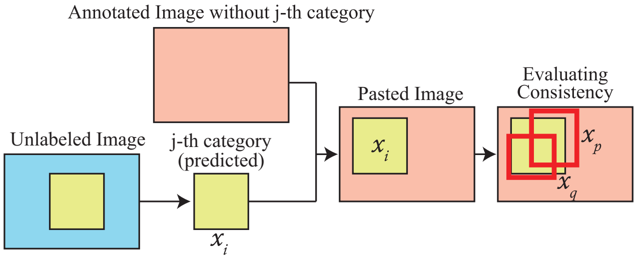


Figure 2. The illustration of the proposed cross image validation. The proposal x_i from the unlabeled image, predicted to belong to the j-th category, is randomly pasted into a certain annotated image without j-th category objects for consistency evaluation. The red bounding boxes $B_I(x_p)$ and $B_I(x_q)$ denote the newly generate region proposals x_p and x_q from the newly pasted image by the up-to-date classifier, respectively. As shown, the unlabeled and annotated images have entirely different context information (denoted in different color).

3.2. Alternative Optimization Strategy

The alternative minimization is readily employed to solve this optimization. Specifically, the algorithm is designed by alternatively updating the sample importance weight set $\mathbf{V}$ via the SSM process, the label set $\mathbf{Y}$ via pseudo-labeling and the AL process, and the network parameters $\mathbf{W}$ via standard backpropagation. In the following, we introduce the details of these optimization steps. The convergence of this algorithm to the practical implementation of our framework will also be discussed.

Updating $\mathbf{V}$: Fixing the $\{\mathbf{Y},\mathbf{X},\mathbf{W}\}$, we can directly calculate $f(x_i,\mathbf{W})$ via Eqn. (5), and further obtain $\mathbf{V}$ via Eqn. (4).

Updating $\mathbf{Y}$: After obtaining $\mathbf{V}$, we calculate the consistency score s_i for sample selection as:

$$j^* = \arg\max_{j\in[m]} \phi_j(x_p;\mathbf{W}),$$
$$s_i = \frac{1}{|\mathbf{I}|}\sum_{I\in\mathbf{I}}\frac{1}{|\Omega_I^{\bar{j}^*}|}\sum_{p\in\Omega_I^{\bar{j}^*}}\phi_{j^*}(x_p;\mathbf{W}). \tag{6}$$

where j^* represents the predicted category by the current detector with highest confidence. Note that, to reduce the computation cost, we only randomly pick out at most N annotated images for evaluating the consistency of the proposal x_i. In all the experiments, we empirically set $N = 5$

for the trade-off between the accuracy and efficiency. Given $\mathbf{S} = \{s_i\}_{i=1}^n$, we rank all unlabeled samples in a descending order for each classifier as [16], and pick out top-k nonzero ones at most for each object category. $\mathcal{H}$ is regarded as high-consistency samples with assigned pseudo-labels. Specifically, these important samples for m categories are defined as $\mathcal{H} = [H_1, ..., H_j, ..., H_m](|H_j| \leq k)$, where k is an empirical parameter to control the number of selected samples for each category.

Fixing $\{\mathbf{W}, \mathbf{V}, \{x_i\}_{i=1}^{\mathcal{H}}\}$, we optimize $\mathbf{y}_i$ of Eqn. (2) which corresponds to solve the following problem for each high-consistency sample $i \in \mathcal{H}$ with its important weight vector $\mathbf{v}_i \neq \mathbf{0}$:

$$\min_{\mathbf{y}_i \in \{-1,1\}^m, i \in \mathcal{H}} \sum_{j=1}^{m} v_i^{(j)} \ell_j(x_i, \mathbf{W}), \quad \text{s.t.} \sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2, \tag{7}$$

where $\mathbf{v}_i$ is fixed, and can be treated as constant. The constraint $\sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2$ largely excludes all samples for pseudo-labeling except under the following two conditions: i) when $y_i^{(j)}$ is predicted to be positive by one classifier but all other classifiers produce negative predictions, or ii) when all classifiers predict $y_i^{(j)}$ to be negative, i.e., x_i is rejected by all classifiers and identified as belonging to an undefined object category. These are the rational cases for practical object detection in large-scale scenarios. Note that we optimize $\mathbf{Y}$ by exhaustively attempting to assign -1 or 1 to each sample for all m categories to minimize Eqn. (7). The computational cost of this process is acceptable because we only need to take $m + 1$ attempts. Through this fashion, Eqn. (7) always has a clear solution by enforcing pseudo-labels on those top-ranked high-consistency sample set $\mathcal{H}$. This is exactly the mechanism underlying a re-ranking technique [16]. Compared with the previous methods [37, 24], our framework can effectively suppress the error accumulation during the incremental pseudo-labeling via the following two advantages: (i) The cross image validation can provide more accurate and robust estimations under various challenging image contexts; (ii) All the pseudo-labels are disposable. They will be discarded after each mini-batch iteration. These advantages are beneficial for the detector to avoid being misled by the accumulate errors.

Low-consistency Sample Annotating: After pseudo-labeling high-consistency samples in such a self-supervised manner, we employ the AL process to update the annotated image set $\mathbf{I}$ by providing more informative guidance based on human knowledge. The AL process aims to select most informative unlabeled samples and to label them as either positive or negative by requesting user annotation. Our selection criteria are based on the classical uncertainty-based strategy [22, 34]. Specifically, we collect those samples with quite small s_i after performing the cross image validation. Then, we utilize the current classifiers to predict their labels. Those predicted as more than two positive labels (i.e., predicted as the corresponding object category) actually represent these samples making the current classifiers ambiguous. We thus adopt them as so called "low-consistency" samples and randomly add z of them into low-consistency sample set $\mathcal{U}$ for further manually annotation by active users. Actually, other similar criterion can be utilized in this step. We employed this simple strategy just due to its intuitive rationality and efficiency.

Algorithm 1 Alternative Optimization Strategy

```
Input: Input dataset $\{x_i\}_{i=1}^n$
Output: Output model parameters $\{\mathbf{W}\}$
Initialize network parameters $\mathbf{W}$ with pre-trained CNN, initially annotated samples $\mathbf{I}$, sample weight set $\mathbf{V}$ and the corresponding consistency score set $\mathbf{S}$;
while true do
    for all mini-batch = $1, ..., T$ do
        Update $\mathbf{V}$ and $\mathbf{S}$ by the SSM process via Eqn. (4) and Eqn. (6), respectively;
        Update $\mathcal{H}$ by the re-ranking;
        Update $\{\mathbf{y}_i\}_{i\in\mathcal{H}}$ by pseudo-labeling via Eqn. (7);
        Update $\mathbf{W}$ by standard backpropagation Eqn.(8);
    end for
    Update low-consistency sample set $\mathcal{U}$;
    if $\mathcal{U}$ is not empty do
        Update the annotated region proposals $\{\Omega_I\}_{I\in\mathbf{I}}$ with $\{\mathbf{y}_i\}_{i\in\mathcal{U}}$ by the AL;
    else
        break;
    end if
end while
return $\mathbf{W}$;
```

Updating W: Fixing $\{\mathcal{D}, \mathbf{V}, \mathbf{Y}\}$, the original model in Eqn. (1) can be approximated as:

$$\min_{\mathbf{W}} \frac{1}{|\mathcal{H} \cup \{\Omega_I\}_{I \in \mathbf{I}}|} \sum_{i \in \mathcal{H} \cup \{\Omega_I\}_{I \in \mathbf{I}}} \sum_{j=1}^{m} \ell_j(x_i, \mathbf{W}) + \mathcal{L}_{loc}(\mathbf{W}). \tag{8}$$

Thus, we can update the network parameters $\mathbf{W}$ by standard backpropagation. Note that, we do not consider the regularization function $R(\cdot)$, which is introduced to regularize the latent variable set $\mathbf{V}$.

The entire algorithm can then be summarized into Algorithm 1. It is easy to see that this algorithm finely accords with the pipeline of our proposed framework in Fig. 1.

Convergence Analysis: Our framework can guarantee the convergence to a local optimum based on its implementation. The reason is three-fold: (i) the objective function Eqn. (2) w.r.t $\mathbf{V}$ is convex; (ii) network fine-tuning in Eqn. (8) via backpropagation can converge to a local optimal; (iii) as the model becomes mature, the AL process stops when no low-consistency samples are found.

4. Experiments

To justify the effectiveness of our proposed SSM process and framework, we have conducted a number of experiments on the public VOC 2007/2012 benchmarks [7], whose data are usually divided into two-fold: trainval and test. To evaluate the model performance on the VOC 2007 benchmark, we regard the VOC 2007 trainval as the initial annotated samples, and consider the VOC 2012 trainval data as unlabeled data that need to be mined. Therefore, the active user annotating process equals to fetch the VOC 2012 annotations. As for the VOC 2012 benchmark, we employ the VOC 2007 trainval and test set for initialization. Moreover, we regard the large-scale object detection dataset COCO [25] as the 'secondary' unlabeled data. In other words, we will perform sample mining on it only when all the VOC 2012 trainval annotations have been used. As for the annotation key 'annotated' and 'pseudo', the first one represents the proportion of the user annotations appended/fetched during the training over the pre-given annotations (i.e., the VOC 2007 trainval), which are used for initializing the object detectors. 'pseudo' implies the percentage of pseudo-labeled object proposals from unlabeled data over the pre-given annotations. The lower 'annotated' value is, the less user efforts for annotating are required. The higher 'pseudo' value is, the more pseudo-labeled samples are obtained. Hence, when achieving the same performance, a superior method should have lower 'annotated' but higher 'pseudo' values.

We adopt the PASCAL Challenge protocol, i.e., a correct detection should has more than 0.5 IoU with the ground truth bounding-box, and exploit the mean Average Precision (mAP) as for the evaluation metric. In all experiments, we set parameters $\{\lambda, k, N, \gamma, z\} = \{0.9, 500, 5, 0.5, 100\}$. The fine-tuning manner for the RFCN pipeline is the same as [4], except that we treat the COCO trainval set as unlabeled data for mining rather than pre-train the network. In the testing phase, we use a single scale of 600 pixels as input except for the VOC 2012 test set, which we use a multi-scale test as [13]. All the experiments are conducted on a desktop with Intel CPU 3.4GHz and four Titan Xp GPUs. The testing time of our framework is 120 millisecond/image, and our training time is 620 millisecond/image. As for the base line method (i.e., RFCN), its testing time is 120 millisecond/image and training time is 150 millisecond/image.

In order to demonstrate that our proposed framework is general to different network architecture and object recognition framework, we have incorporated our framework into the Fast R-CNN (FRCN) pipeline with AlexNet [19] and the new state-of-the-art RFCN [4] with ResNet-101 [14]. We denote these variants as "FRCN+Ours" and "RFCN+Ours", respectively. To validate the superior performance of the proposed framework, we have compared it with the CEAL [37] and K-EM [40] approaches. Note that, since the CEAL is designed for image classification and is not mini-batch friendly, we extended it for object detection by alternatively performing sample selection and fine-tuning the CNN. Since the source code of K-EM [40] is not publicly available, we have obtained the results from their paper [40] directly. We denote these methods as "FRCN+CEAL," and "FRCN+K-EM", respectively. Moreover, we have also included five baseline methods that ignore the pseudo-labeling of unlabeled samples: "FRCN+RAND", which randomly selects region proposals for user annotations, FRCN+EMC (Expected-Model-Change) [8], FRCN+Entropy [31], FRCN+ELC (Expected-Labeling-Change) [27] and FRCN+Density [31]. For a fair comparison, we initialize all the methods by providing only 5% annotations, and allow FRCN+Ours and FRCN+CEAL to mine unlabeled samples only from the VOC 2007 train/val set. Moreover, we repeat the testing by five trails to report the average mAP with standard variance to present a comprehensive evaluation.

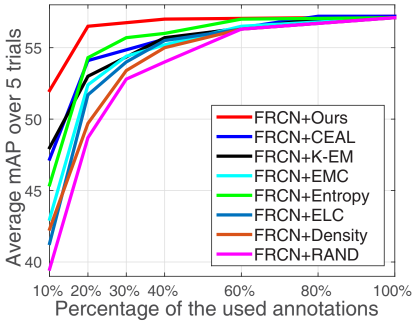


Figure 3. Quantitative comparison of detection performance (mAP) on the PASCAL VOC 2007 test set

4.1. Results and Comparisons

The Fig. 3 demonstrates the comparison of detection performance using the Fast-RCNN (FRCN) pipeline with AlexNet [19] on the VOC 2007 test set. As illustrated in Fig. 3, our proposed framework FRCN+Ours consistently outperforms all the compared methods by clear margins under all annotation settings. Specifically, FRCN+Ours can achieve the equivalent of a fully supervised performance (i.e., FRCN with 100% user annotations) when fetching only approximately 30% user annotations, while most of the compared methods require nearly 60%. These results indicate the superior performance of our framework.

To demonstrate the feasibility and great potential of our framework, we have conducted amount of experiments to fine-tune RFCN with ResNet-101 (well pre-trained on ImageNet) on the VOC 2007/2012 trainval set by using our framework and the compared baseline RFCN+RAND. The compared results on the VOC 2007 and 2012 benchmarks are illustrated in Tab. 1 (a)(b), respectively. By controlling the number of training iterations, RFCN+Ours and RFCN+RAND can fetch different amounts of annotations

Table 1. Test set mAP for VOC 2007/2012 under the RFCN [4] pipeline. The entries with the best APs with each sub-table are bold-faced. Annotation key: 'annotated' denotes the proportion of the user annotations appended/fetched from the VOC 2012 trainval set during the training over the initial annotations (i.e., 07 denotes the VOC 2007 trainval set, while 07+ denotes the VOC 2007 trainval and test sets), which are used for initializing the object detectors; 'pseudo' implies the percentage of pseudo-labeled object proposals from unlabeled data (i.e., the VOC 2012/COCO trainval set) over the pre-given annotations.

	Method	initial	test	annotated	pseudo	mAP
(a)	RFCN	07	07	0%	0%	73.9
	RFCN+RAND	07	07	20%	0%	75.6±1.0
	RFCN+RAND	07	07	60%	0%	76.5±1.1
	RFCN+RAND	07	07	100%	0%	77.2±0.9
	RFCN+RAND	07	07	200%	0%	79.1±0.4
	RFCN+Ours	07	07	20%	300%	76.0±0.1
	RFCN+Ours	07	07	60%	400%	77.4±0.2
	RFCN+Ours	07	07	100%	500%	78.3±0.2
	RFCN+Ours	07	07	200%	800%	79.7±0.2
	RFCN+Ours	07	07	200%	1000%	**80.6±0.2**
(b)	RFCN	07+	12	0%	0%	69.1
	RFCN+RAND	07+	12	10%	0%	71.5±1.1
	RFCN+RAND	07+	12	30%	0%	72.7±1.3
	RFCN+RAND	07+	12	50%	0%	74.4±1.0
	RFCN+RAND	07+	12	100%	0%	76.8±0.4
	RFCN+Ours	07+	12	10%	100%	72.6±0.1
	RFCN+Ours	07+	12	30%	150%	73.6±0.1
	RFCN+Ours	07+	12	50%	200%	75.5±0.2
	RFCN+Ours	07+	12	100%	200%	77.3±0.2
	RFCN+Ours	07+	12	100%	800%	**78.1±0.2**
(c)	RFCN	07	07	0%	0%	73.9
	RFCN+SPL	07	07	0%	300%	74.1±0.5
	RFCN+SPL	07	07	0%	400%	74.7±0.6
	RFCN+SSM	07	07	0%	300%	75.6±0.2
	RFCN+SSM	07	07	0%	400%	76.7±0.3
	RFCN+AL	07	07	20%	0%	75.5±0.1
	RFCN+AL	07	07	60%	0%	77.0±0.2
	RFCN+AL	07	07	100%	0%	**77.5±0.2**
(d)	RFCN	07+	12	0%	0%	69.1
	RFCN+SPL	07+	12	0%	100%	70.9±0.5
	RFCN+SSM	07+	12	0%	100%	72.1±0.3
	RFCN+AL	07+	12	10%	0%	71.8±0.1
	RFCN+AL	07+	12	30%	0%	73.0±0.2
	RFCN+AL	07+	12	50%	0%	**74.7±0.2**

from the VOC 2012 trainval set.

As one can see from Tab. 1 (a)(b), both RFCN+RAND and RFCN+Ours gradually obtain increased detection accuracy when the number of annotations increases. Our framework consistently outperforms the compared baseline RFCN+RAND under all appending conditions on both the VOC 2007 and 2012 benchmarks by a clear margin. Moreover, our framework surpasses RFCN+RAND by nearly 1.5% (80.6% vs 79.1%) on the VOC 2007 benchmark and 1.3% (78.1% vs 76.8%) on the VOC 2012 benchmark with significantly small variations when sufficient pseudo-labeled object proposals are provided. This validates the effectiveness of our framework. Some examples of the selected high-consistency and low-consistency region proposals via the cross image validation are depicted in Fig. 5.

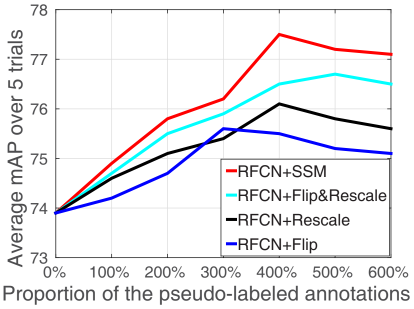


Figure 4. Quantitative comparison of average mAP on the VOC 2007 test set

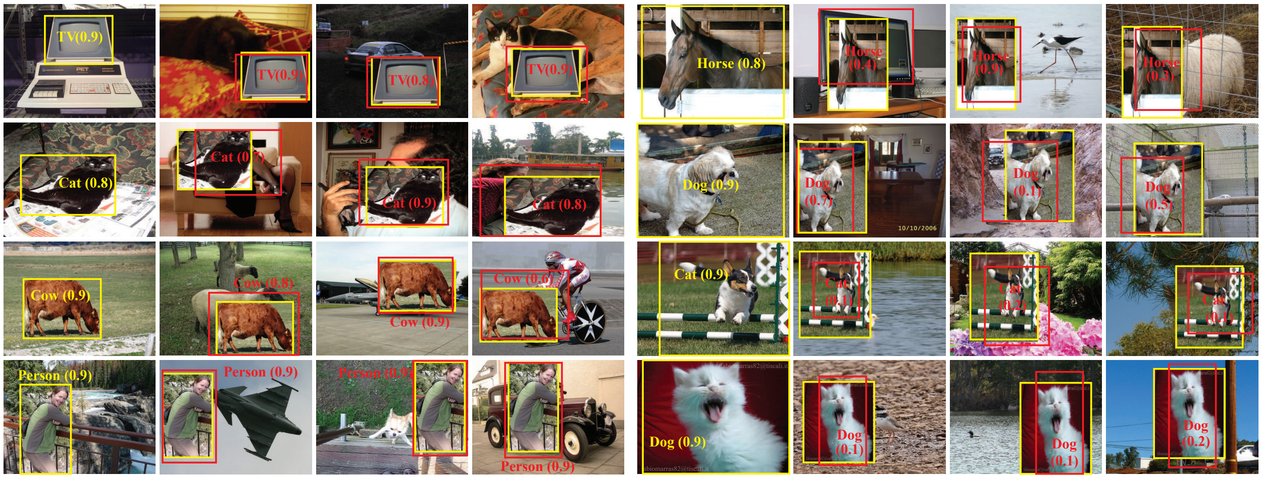
4.2. Ablation Study

To validate the contribution of each component inside our framework, we have also conducted sufficient experiments for empirical analysis. The variant of our framework that discarding the active learning process is denoted as "FRCN+SSM" / "RFCN+SSM". Similarly, these pipelines with Active Learning (AL), Self-paced Learning (SPL) are denoted by "RFCN+AL", "RFCN+SPL", respectively. RFCN+AL adaptively collects low-confidence proposals to request the annotations and stops when no low-confidence samples are found. RFCN+SPL is implemented according to [37].

As illustrated in Tab. 1 (c)(d), given the same amount of annotations during initialization, RFCN+SSM performs significantly better than RFCN+SPL on both VOC 2007 and the VOC 2012 test set. Specifically, RFCN+SSM achieves a nearly 2% performance improvement (76.7% vs 74.7%) with small variations over RFCN+SPL by pseudo-labeling the same amount of high-consistency region proposals for training on the VOC 2007 benchmark. A consistent performance gain of approximately 1.2% (72.1% vs 70.9%) is obtained on the VOC 2012 benchmark by RFCN+SSM. These results validate the significant contribution of the proposed SSM process on mining reliable region proposals for improving object detection.

We have also compared our self-supervised sample mining (SSM) process with three baseline methods under the AL process disabled setting. RFCN+Flip implies horizontally flipping images for validation, while RFCN+Rescale represents randomly rescaling an image from 50% to 200% of its original size. RFCN+Flip&Rescale means the fusion of them. As shown in Fig. 4, RFCN+SSM consistently outperforms all the competing methods by clear margins at all pseudo-labeling proportions. Moreover, compared to these baselines, RFCN+SSM obtains a slighter performance drop (caused by the accumulated pseudo-labeling errors) after reaching its peak performance. This also proves the effectiveness of our SSM for suppressing the error accumulation.

Tab. 1 (c)(d) also demonstrates that RFCN+AL consistently outperforms the baseline RFCN and RFCN+SSM. Though the improvements are minor compared to the best



(a) High-consistent Region Proposals

(b) Low-consistent Region Proposals

Figure 5. Selected (a) high-consistent and (b) low-consistent region proposals with pseudo-labels in yellow via cross image validation. The first column lists the region proposal with predicted label and high prediction confidence from the unlabeled images of the PASCAL VOC 2012 dataset. The region proposal is randomly put into the validation images for object detection. The predicted bounding box that has more than 0.5 IoU with this proposal is illustrated in red. The corresponding prediction confidence (ranging from 0 to 1) for being the pseudo-labeled category is also in red. It is obvious from (a) that those high-consistent proposals are of high quality for network fine-tuning. As one can see from (b), the confidence of those bounding boxes on the validation images is low-consistent due to inaccurate bounding box (first two rows) or incorrect pseudo-labels (last two rows)

performance of the RFCN+SSM, our proposed AL stage is still beneficial for promoting object detection. This slight improvement occurs because the informative samples with great potential for improving performance follow a long-tail distribution, as reported in [39]. Therefore, it is necessary to employ abundant training samples by asking active users to provide labels or finding other assistance. Fortunately, our proposed high-consistency sample pseudo-labeling via the SSM process is an effective way to address this issue.

The results of weaker initialization (5% and 10% annotations from the VOC 2007 train/val set) are listed in Tab. 2. As shown, our RFCN+SSM achieves a consistent performance gain of about 2% over the original RFCN. Compared to RFCN+RAND, RFCN+Ours obtains about 1.5% higher average mAP with much lower variances. However, compared to RFCN+Ours, FRCN+Ours in Fig. 3 obtains a higher mAP gain. The reason is FRCN and RFCN use different algorithms (Selective Search [35] vs. Region Proposal Network [11]) to generate object proposals. Since our objective is to mine samples from these proposals, our model obtains various performance boosts based on the quality of proposals. This shows the generality and effectiveness of our model over different proposals.

Acknowledgment

This work was supported in part by the Hong Kong Polytechnic Universitys Joint Supervision Scheme with the Chinese Mainland, Taiwan and Macao Universities (Grant no. G-SB20). This work was also supported by HK RGC General Research Fund (PolyU 152135/16E), in part by Guangdong "Climbing Program" Special Funds under Grant pdjhb0010, in part by National Natural Science Foundation of China (NSFC) under Grant U1611461 and Grant 61702565, and in part by Science and Technology Planning Project of Guangdong Province of No.2017B010116001.

Table 2. Test set mAP for VOC 2007 under the RFCN pipeline with ResNet-101.

Method	initial	annotated	mAP	initial	annotated	mAP
RFCN	5%	0%	49.4±1.4	10%	0%	56.8±1.3
RFCN+SSM	5%	0%	51.6±0.3	10%	0%	59.4±0.2
RFCN+RAND	5%	30%	53.3±1.1	10%	30%	60.4±1.5
RFCN+Ours	5%	30%	55.1±0.1	10%	30%	62.9±0.2

5. Conclusion

In this paper, we have introduced a principled Self-supervised Sample Mining (SSM) process, and justified its effectiveness in mining valuable information from unlabeled or partially labeled data to boost object detection. We further involve this process in the AL pipeline with a concise formulation, which is developed for retraining object detectors via faithfully pseudo-labeled high-consistency object proposals after our proposed cross image validation. The proposed SSM process contributes to effectively improve the detection accuracy and the robustness against noisy samples. Meanwhile, the rest samples, being low consistency (high uncertainty) by the current detectors, can be handled by the AL, which benefits to generate reliable and diverse samples gradually. In the future, we will apply our SSM to improve other specific visual detection task with unlabeled web images/videos.

References

[1] X. D. amd Liang Zheng, F. Ma, Y. Yang, and D. Meng. Few-example object detection with model communication. In *arXiv:1706.08249 [cs.CV]*, 2017. 3

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 2, 3

[3] K. Brinker. Incorporating diversity in active learning with support vector machines. In *ICML*, 2003. 3

[4] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 6, 7

[5] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017. 2, 3

[6] Elhamifar, Ehsan, S. Guillermo, Y. Allen, and S. S. Shankar. A convex optimization framework for active learning. In *ICCV*, 2013. 3

[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) results, 2007. 6

[8] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*, pages 562–577, 2014. 3, 6

[9] C. Gao, D. Meng, W. Tong, Y. Yang, Y. Cai, H. Shen, G. Liu, S. Xu, and A. Hauptmann. Interactive surveillance event detection through mid-level discriminative representation. In *ACM MM*, 2014. 1

[10] A. Ghosh, N. R. Pal, and S. K. Pal. Self-organization for object extraction using a multilayer neural network and fuzziness measures. *IEEE Trans. On Fuzzy Systems*, pages 54–68, 1993. 2, 3

[11] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 3, 8

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *T-PAMI*, 2015. 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 1, 6

[15] R. T. Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari. How hard can it be? estimating the difficulty of visual search in an image. In *CVPR*, 2016. 1

[16] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: self-paced reranking for zero-example multimedia search. In *ACM MM*, 2014. 3, 5

[17] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014. 3

[18] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015. 2, 3

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 6

[20] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 2, 3

[21] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011. 3

[22] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *ACM SIGIR*, 1994. 1, 3, 5

[23] L. Liang and K. Grauman. Beyond comparing image pairs: Setwise active learning for relative attributes. In *CVPR*, 2014. 1

[24] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang. Active self-paced learning for cost-effective and progressive face identification. *T-PAMI*, 2017. 1, 3, 4, 5

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 6

[26] B. Liu and V. Ferrari. Active learning for human pose estimation. In *ICCV*, 2017. 1

[27] P. Melville and R. J. Mooney. Diverse ensembles for active learning. In *ICML*, 2004. 6

[28] S. K. Pal, A. K. Datta, and D. D. Majumder. Computer recognition of vowel sounds using a self-supervised learning algorithm. *JASI*, VI:117–123, 1978. 3

[29] P. Rhee, E. Erdenee, S. D. Kyun, M. U. Ahmed, and S. Jin. Active and semi-supervised learning for object detection with imperfect data. In *Cognitive Systems Research*, 2017. 3

[30] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe. Self paced deep learning for weakly supervised object detection. *CoRR*, abs/1605.07651, 2016. 3

[31] B. Settles. Active learning literature survey. In *University of Wisconsin, Madison, 2010, 52(55-66): 11*. 1, 6

[32] A. Shrivastava, A. Gupta, and R. Girshic. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 1

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2014. 1

[34] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2, 2002. 3, 5

[35] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. In *T-PAMI*, 2013. 8

[36] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, pages 1449–1456, 2011. 1, 3

[37] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *T-CVST*, 2016. 1, 3, 4, 5, 6, 7

[38] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2, 3

[39] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017. 1, 8

[40] Z. Yan, J. Liang, W. Pan, J. Li, and C. Zhang. Weakly- and semi-supervised object detection with expectation-maximization algorithm. In *arXiv:1702.08740 [cs.CV]*, 2017. 6

[41] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015. 3