Adaptively Denoising Proposal Collection for Weakly Supervised Object Localization

Wenju Xu[1] · **Yuanwei Wu[1]** · **Wenchi Ma[1]** · **Guanghui Wang[1]**

Abstract
In this paper, we address the problem of weakly supervised object localization, which trains a detection network on the dataset with only image-level annotations. The proposed approach is built on the observation that the proposal set from the training dataset is a collection of background, object parts, and objects. Several strategies are taken to adaptively eliminate the noisy proposals and generate pseudo object-level annotations for the weakly labeled dataset. A multiple instance learning algorithm enhanced by mask-out strategy is adopted to collect the class-specific object proposals, which are then utilized to adapt a pre-trained classification network to a detection network. In addition, the detection results from the detection network are re-weighted by jointly considering the detection scores and the overlap ratio of proposals in a proposal subset optimization framework. The optimal proposals work as object-level labels that enable a pseudo-strongly supervised dataset for training the detection network. Consequently, we establish a fully adaptive detection network. Extensive evaluations on the PASCAL VOC 2007 and 2012 datasets demonstrate a significant improvement compared with the state-of-the-art methods.

Keywords Weakly supervised object localization · Proposal subset optimization · Re-weighting · Retraining

1 Introduction

Object detection, which attempts to place a tight bounding box around every object of a given image, is an important problem for image understanding. This problem has been extensively studied in recent years [1–6], and the state-of-the-art detection performance promotes a variety of applications, including human pose estimation [7] and crowd counting [8]. One key step for object detection is to learn a distinctive representation of the objects from a large quantity of labeled data. Most existing methods rely on object-level labeled dataset [9] so that

✉ Wenju Xu
  xuwenju@ku.edu

  Guanghui Wang
  ghwang@ku.edu

[1] School of Engineering, University of Kansas, Lawrence, KS 66049, USA

their models learn visual features from those specified regions. However, data annotation is an exhaustive and error prone work. In order to reduce the annotation cost, a common strategy is to learn the detector in a weakly supervised manner that only binary image-level labels representing the overall presence or absence of an object category are added to the images for training.

Multiple instance learning (MIL) [10–12] is an intensively used strategy in dealing with the task of weakly supervised object localization (WSL). It selects object regions of interest (proposals) from the positive images that contains the object, and learns an appearance model of the object from the features in the selected regions. This method has the tendency to get stuck in local optima. Therefore, a re-localization and re-training strategy is typically taken to push the solution close to the global optima. Pentina et al. [13] forms a curriculum learning strategy to feed the training process from easy images with big objects to hard images with many small objects. Shi et al. [5] propose a strategy that re-weights the proposals' scores according to the consistence between the proposal size and the estimated object size. Even though these strategies attempt to improve the MIL, finding positive image bags containing certain class object for MIL classifier, in some senses, depends on guessing and it is possible to take a negative bag as a positive one. It is also difficult to get tight bounding boxes exactly containing the objects. These drawbacks require strategies that adaptively refine the estimated bounding boxes to tightly contain the objects.

Another line of this research is based on convolutional neural networks (CNNs) [14,15] that are capable of learning generic visual features generalized to many tasks. Methods in this category are inspired by the facts that, without location annotation, a pre-trained image classification CNN learns representative information of objects and object parts. Many efforts leverage CNN to extract discriminative appearance features, then train a MIL appearance model for object detection [16]. Recent efforts [3,4] achieve significant performance improvement with proposed end-to-end methods, which adopt a pre-trained classification network to mine location information and transfer the problem from weakly supervised object localization to pseudo-strongly supervised object detection. However, generating instance-level labels from the image-level labels is nontrivial since the objects from the same category may appear with different shapes and background. A pre-trained classifier makes predictions on salient features. The extracted appearance features represent object parts, which lack information on the instance as a whole. Moreover, it is different to determine the size of bounding boxes that exactly contain the objects in the feature-level searching. As a result, the obtained instance-level labels are inexact.

In this paper, we propose a new framework based on two observations: (i) the proposals are a collection of background, object parts, and objects; and (ii) it is hard to train object detectors directly under weakly labeled dataset due to the substantial amount of noise in the object proposal collection and the size variation of the objects. Our method integrates several strategies to adaptively eliminate the noise in the object proposal collection. We take an enhanced MIL algorithm, which is proceeded by a mask-out strategy to mine the proposal collection and fine-tune a pre-trained classification network through re-weighting and re-training, which exploits proposal subset optimization [19] to further re-weight the detection results.

Our re-weighting and re-training strategy aims at determining the optimal proposals automatically. To this end, we take a subset optimization method to select object proposals. It is based on both the detection scores from the pre-trained detection network and the overlaps between the candidate bounding boxes. This strategy puts higher weights on proposals that have large overlap area with others. Specially, We reweight those object proposals with high detection scores according to how much the bounding box overlaps with other bounding

boxes. Iteratively, we utilize this subset optimization method to improve the re-localization step.

This re-weighting scheme reduces the uncertainty in the proposal distribution, making the re-weighting step more likely to pick a proposal correctly covering the object. Figure 2 shows an example of how the subset optimization changes the proposal score induced by the current object detector, leading to a more accurate localization.

Our contributions are as follows: (i) we propose a novel work flow to collect confident proposals, which integrates the mask-out strategy, MIL, and subset proposal optimization. The MIL model is trained on the selected proposals of mask-out strategy and mines confident proposals to reduce the background clutters and potential confusion from similar objects cases. The subset proposal optimization further refines the proposals by re-scoring the bounding box; (ii) following the idea of re-localization and re-training, the candidate proposals are refined based on both the detection scores and the overlap ratios between the proposals. We then iteratively adapt a pre-trained classification network to a detection network with those quality enhanced proposals. This is a new pipeline for improving object proposals; and (iii) detailed evaluations on the PASCAL VOC 2007 and 2012 datasets [20] demonstrate that our weakly supervised object detection with adaptively denoised proposal collection performs favorably against the state-of-the-art methods. The proposed model and trained parameters will be available on the authors website.

2 Related Work

Extracting meaningful information from the environment is a challenging task [21,22]. In recent years, deep neural networks are becoming more and more popular for knowledge discovering in many computer vision tasks, such as object detection [23,24], visual question answering [25], pose estimation [26–28], image synthesis [29–31], face recognition [32], and depth estimation [33,34]. Object detection is the task of recognizing and localizing the objects in the images with the deep model trained on labelled ground truth [35]. However, labelling the images with bounding box for each object is a nontrivial work. In the scenario of weakly supervised localization, the training images are known to containing instances of a certain object class but their locations. There is no ground truth bounding box available for each object in the training dataset. The task is both to localize the objects (estimate the bounding boxes tightly containing the instances) and to classify the objects. What we have are the image-level annotations which are weak supervision for localizing the objects. To train a detection network with image-level supervision, we need first to localize objects in all the images of the training dataset based on image-level annotations, and then use the localization results to train a detector for the test set. The WSL problem is often handled with multiple instance learning (MIL) [2,3,6,12,36,37], where the images are treated as bags of object proposals [38,39] (which are bounding boxes estimated to localize the objects). A negative image dose not contain instances of certain category. A positive image contains at least one positive instance, mixed in with a majority of negative ones. The goal is to find the true positive instances from which to learn a classifier for proposal classification.

Previous works achieve significant improvement by exploring ways to enhance the MIL. Siva et al. [40] propose an effective negative mining approach combined with discriminative saliency measures. Song et al. [6] formulate an initialization strategy for WSL as a discriminative submodular cover problem in a graph-based framework, and develop a negative mining technique to increase robustness against incorrectly localized boxes [41]. Bilen et al.

[2,3] propose a relaxed version of MIL that softly labels object instances instead of choosing the highest scoring ones. They also propose a discriminative convex clustering algorithm to jointly learn a discriminative object model and enforce the similarity of the localized object regions.

As CNNs have turned out to be surprisingly effective in many vision tasks including classification and detection, recent state-of-the-art WSL approaches also build on CNN architectures [3,42] or CNN features [36]. Bilen et al. [3] modify a region-based CNN architecture [43] and propose a CNN with two streams, one focuses on recognition and the other one on localization, which performs simultaneously region selection and classification. Similarly, Li et al. [4] use the MIL to obtain the initial detection results and propose a domain adaption method [44,45] to fine-tune a classification network into a detection network with the initial detection results. The results show a performance improvement on the detection accuracy. Shi et al. [5] attempt to score the proposals by the size and retrain the detection network with the re-weighted proposals according to an easy to hard order, based on the assumption that the proposals with bigger size provide more information to train the network than the those with smaller size. Our work is related to these CNN-based MIL approaches that perform WSL by end-to-end training from image-level labels. In contrast to the above methods, however, we focus on a CNN architecture that is re-trained in an order for detection accuracy improvement with denoised proposals.

The concept of adaptive learning in an order was also studied in computer vision [5,13,46–48]. These works focus on a key question: how to re-weight the proposals? Sharmanska et al. [47] employ some privileged information to distinguish between easy and hard examples in an image classification task. The privileged information are additional cues available at training time, but not at test time. They employ several additional cues, such as object bounding boxes [49,50], image tags and rationales to define their concept of easiness [46]. Lai et al. [51] select highly confident object proposals under the guidance of class-specific saliency maps. Pentina et al. [13] consider learning the visual attributes of objects. Shi et al. [5] propose a size estimator to re-weight the proposals based on the size of the instances in the image. They use curriculum learning in a WSL setting and propose object size as an "easiness" measure. Shi et al. [52] consider the task of discovering object classes in an unordered image collection. their model is initialized with regions of "stuff" categories, and is then used to support discovering "thing" categories in unlabelled images with the help of a fully supervised segmentation model. Bodla et al. [53] propose a soft method to select the bounding boxes. They decay the classification score of a box which has a high overlap with top-scoring boxes, rather than suppressing it. Jie et al. [14] explore the Fast RCNN model [43] and propose a self-taught learning method for proposal selection. The most related work to ours is the very recent study [15], which designs an on-line classifier refinement pipeline to progressively locate the most discriminative region of an image. By contrast, we propose a novel work flow to adaptively refine the proposals, i.e., to iteratively collect a more confident subset of proposals. In addition, we take the re-training strategy to fine-tune the model with the denoised proposal subset. The proposed work flow, by integrating several novel proposal mining strategies, makes it adaptable to a variety of weakly supervised object detection tasks.

3 Adaptively Denoised Proposal Collection

The proposed weakly supervised object detection method is illustrated in Fig. 1. This model consists of three major components, namely confident proposal learning, object detector
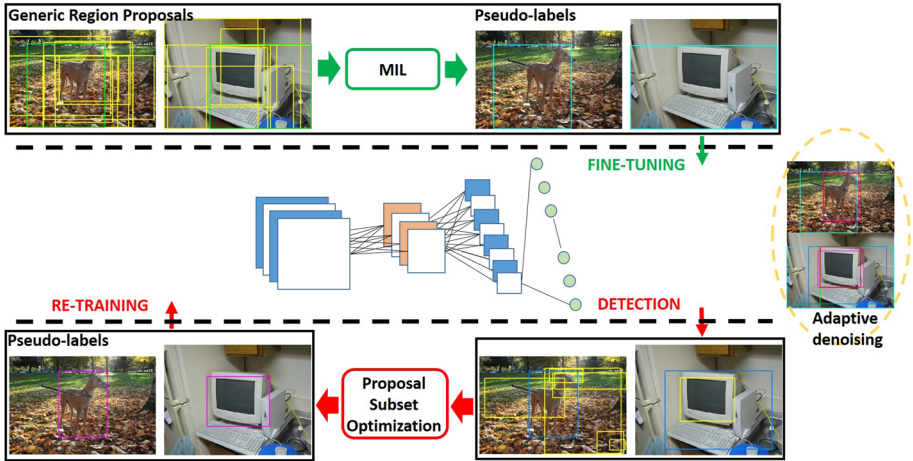
Fig. 1 Overview of our method. We use mask-out strategy to collect the generic region proposals and take the MIL to generate pseudo labeled training set. This dataset is then fed to a WSL loop, so that the object detector is re-trained progressively. We also take the re-localization [17,18] step by re-weighting object proposals according to the detection scores and the overlap of the proposals. Bounding boxes (in yellow) represent the confident proposals; while the bounding box in other colors in each block represents the highest confident proposal. (Color figure online)

learning and proposal subset optimization. They are successively employed to adaptively refine the proposal collection. The remainder of this section discusses these three components in details.

3.1 Confident Proposal Mining

We consider the weakly supervised object localization problem as an adaptively proposal denoising procedure that gradually refines the proposal collection. At the end, we transfer the problem from the weakly supervised object localization to a pseudo-strongly supervised object detection. Based on a pre-trained CNN classification network and a MIL model, our work flow adaptively selects confident proposals other than those comprised of background or object parts from the candidate proposals generated by EdgeBoxes [39].

Assisted by the classification network, we first utilize the mask-out strategy to collect object proposals. The idea of masking out the input of CNN has been previously explored in [54], which replaces the pixel values of the proposals with fixed mean pixel values; and compares the classification scores of feeding the real image and its mask-out images into the classification network. Intuitively, if the mask-out image introduces a notable drop in the classification score for the cth class, the region can be considered as containing an object belonging to the cth class. Inspired by [4,54], we apply the mask-out strategy to select the proposals containing a certain object. We denote the classification network as f_c that maps an image to a confidence vector of cth classes. The confident proposals B_c are selected by investigating the difference of classification score between the selected image $I(x)$ and its mask-out image $I(x/b)$. This is formulated as

$$B_c = arg \max_b (f_c(I(x)) - f_c(I(x/b)))$$ (1)

where b represents the masked-out region. To select confident proposals, we first set a threshold on the classification score. The region b is considered discriminative for cth class based on two aspects: the score of classifying the image $I(x)$ to the cth class is beyond the threshold and the classification score drop between the image and corresponding mask-out images is maximum.

Once the proposals are obtained by applying the mask-out strategy, we separately learn one MIL model for each category. Taking the purified proposals selected by the mast-out strategy as training dataset makes the basic MIL initialized from a higher baseline, which not only stabilizes the training process, but also reduces the time for training [4]. In the MIL model, each instance is described by a feature vector. More specifically, each feature vector is regarded as an instance and each image is represented by a bag of instances. For instance, the training image x_i is considered as a bag of proposals with pseudo strong labels $y_i \in \{-1, 1\}$ indicating whether the bag contains an instance in the specific category. A bag is considered to be negative if there is no instances or all its instances are not in that category, while it is positive if there is at least one of its instances in that category. Given feature representation $\phi(x_i, z)$, we iteratively train the MIL model with the objective written as

$$\min_{w \in R} \frac{1}{2}||w||_2^2 - \sum_{i=1}^{n} log \left(\left(y_i \max_{z \in \mathcal{Z}} w^T \phi(x_i, z) - \frac{1}{2} \right) + \frac{1}{2} \right) \tag{2}$$

where w represents the parameters of the MIL model and z is called the "latent variable" chosen from the set Z, which is typically a set of bounding boxes. The top-scoring proposals given by the mask-out strategy are taken as positive samples for each category, which are used to train the MIL model. Among the initial bounding boxes, the set Z contains all possible candidate instances. Maximizing the objective function over Z amounts to choosing a bounding box containing the whole object. The proposal, in this work, is represented by a 4096-dimensional feature vector from the second-last layer of the classification network.

The top row in Fig.1 demonstrates the idea of the confident proposal mining, which starts from the mask-out strategy and ends with the high confident output from MIL.

3.2 Proposal Subset Optimization

Proposal selection based object detection method has one severely issue of overlapping among the bounding boxes that correspond to the same object. To select the best bounding box for each object, greedy non-maximum suppression (NMS) is widely employed as the latest strategy which selects the top-scoring bounding box b_i and discards other bounding boxes $\mathcal{M}$ that have overlaps with the chosen one larger than a threshold T. Due to simplicity, this NMS mainly focus on the detection score s_i. By taking the Intersection over Union (IoU) as the measure of overlapping, this non-maximum suppression process can be described as

$$s_i = \begin{cases} s_i & IoU(\mathcal{M}, b_i) < T; \\ 0 & IoU(\mathcal{M}, b_i) \geq T. \end{cases} \tag{3}$$

However, there are no instance-level labels available for network training in the weakly supervised localization task, even the bounding boxes estimated with top score are tended to be noisy. To overcome this issue, we propose a subset optimization scheme. It is realized by re-weighting the detection scores among the bounding boxes with high but noisy initial scores, where greedy NMS is not able to adjust the estimated bounding box accordingly. The proposed approach is similar to that described in [19]. However, we employ the method to

solve the weakly supervised learning problem. The confident proposals with high detection scores are grouped into clusters by jointly considering the scores and the spatial overlaps between the proposals. The bounding box set is represented by $B = \{b_i : i = 1 : n\}$. We denote the group membership as $X = (x_i)_{i=1}^n$, where $x_i = j$ if b_i belongs to a cluster b_j. Then one exemplary bounding box o is selected from each cluster B as the final output. This is formulated as finding the maximum a posterior (MAP) solution of the joint distribution $P(O, X|I; B, S)$, which tends to assign big value to the bounding boxes that have large overlap with more confident proposals. After taking the log of the posterior, the objective function becomes:

$$O = X^* = arg \max_X \sum_{i=1}^n \omega_i(x_i) \tag{4}$$

where $\omega_i(x_i = j) = log P(x_i = j|I)$

$$P(x_i = j|I) = \begin{cases} Z_2^i \lambda & if\ j = 0; \\ Z_2^i K(b_i, b_j) s_i & otherwise. \end{cases} \tag{5}$$

$K(b_i, b_j)$ is the window IoU used to measure the spatial overlap between b_i and b_j, $S = \{s_i : i = 1 : n\}$ is the score set containing the detection scores of all the bounding boxes, and Z_2^i is the normalization constant. Parameter β and γ control the penalty level. Note that our proposal subset optimization method takes both the scores and the overlapping into consideration since the detection scores in the weakly supervised task are not always reliable.

The proposal subset optimization problem is defined as:

$$O^* = arg \max_O \beta \sum_{i \in \hat{O}} s_i - \gamma \sum_{i, j \in \hat{O}: i \neq j} K(b_i, b_j) \tag{6}$$

In this setting, we first maximize the objective function over X according to Eq. (4), which will select the cluster centers. Then, a greedy algorithm is used to choose a minimal number of bounding boxes as the outputs based on Eq. (6). More details of the method can be found in [19].

3.3 Object Detector Learning

In this step, we adapt the pre-trained classification network to an object detection network. This neural network is trained with the pseudo labeled proposals obtained from the proposal subset optimization strategy. We employ the re-weighting and re-training strategy for network adaption. The network parameters are fine-tuned for object localization, as illustrated in the bottom of the Fig. 1. We organize it as adaptively refining the proposal subset, which is similar to the curriculum learning. However, we do not separate the training dataset into easy and hard parts. We start by running MIL, which is initialized with the results from mask-out strategy. This leads to a reasonable first detection model A_1. We move forward by running proposal subset optimization on the proposals subset with high detection scores, which produce a re-weighted proposal subset. The process then moves on to the second training iteration, where the training dataset consists of re-weighted proposals with more confident pseudo labels. As a result, the refined model A_2 will localize the objects better than A_1, as it is trained with better supervision in the re-training step. The process iteratively moves on to the next round, starting from the detection model A_k and yielding a better one A_{k+1}. The whole training procedure is described in Algorithm 1.

Fig. 2 Detection results from NMS (red line in left) and subset optimization (center). Bounding boxes (BB) (right) represent the highest confident proposals got from different steps (blue BB: CNN, green BB: maskout, pink BB: re-train, cyan BB: MIL). We compare the detection results by bounding boxes in different colors, which shows our re-training strategy is able to get the denoising proposals by re-weighting object proposals according to the detection scores and the overlap ratios of the proposals. (Color figure online)

The selected results from each strategy are shown in Fig. 2. It is demonstrated that the bounding boxes selected by the fully adapted detection network exactly contain the objects, while the bounding boxes selected by mask-out strategy and MIL contain the object but with a large margin. By re-weighting the confident proposals according to the detection scores and the overlap of the proposals, the re-training strategy is able to generate more confident proposals.

Algorithm 1 The training pipeline of the proposed algorithm.

Input:
    Images $x \in X$; $B = \{b_i : i = 1 : n\}$ candidate boxes; $S = \{s_i : i = 1 : n\}$ the corresponding scores; K, the refinement times; M, the network iteration times; θ_E, network parameters.
Output:
    A fully adaptive detection network.
1: Classify the real images and the mask-out images with the classification network; select the top M proposals by Eq. (1) as the initial proposal set $P_0 \leftarrow \{x, s_0, B_0\}$.
2: For each category, construct positive and negative bags within S_0; train the MIL model and collect the detection results from the trained MIL model as proposal set $P_1 \leftarrow \{x, s_1, B_1\}$.
3: **for** $k = 1$ **to** $K - 1$ **do**
4:     Set $P \leftarrow P_k$.
5:     **for** $m = 1$ **to** $M - 1$ **do**
6:         Sample $P \rightarrow \{x, s, b\}$ as a minibatch.
7:         Network forward propagation and get loss ℓ.
8:         Network backward propagation, $\theta_E \xleftarrow{+} -\nabla_{\theta_E}(\ell)$.
9:     **end**
10:     Collect the detection results from the trained detection network, $P'_k \leftarrow \{x_k, s_k, B_k\}$.
11:     Choose the proposals with subset optimization by Eq. (6); update the proposal set $P_{k+1} \leftarrow \{x, s'_k, B'_k\}$.
12: **end**

---

4 Experimental Evaluation

Dataset and Settings   The proposed approach is extensively evaluated on two publicly available datasets: PASCAL VOC 2007 and 2012 datasets. Both of them have 20 classes of different images. We employ both the AlexNet [55] and VGGNet [56] as our base CNN models, initialized with parameters transferred from the classification network, which is pre-
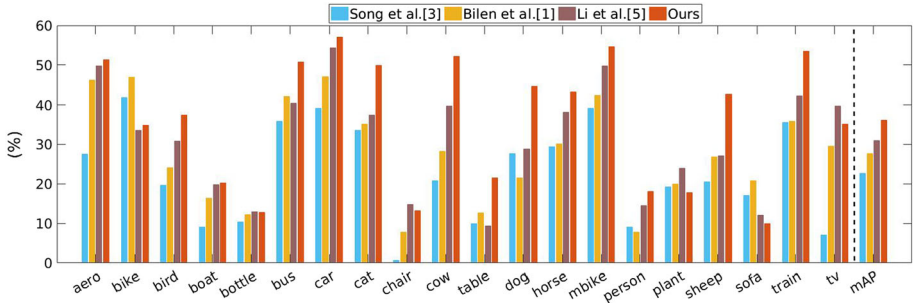
Fig. 3 A comparison of our method (AlexNet) of detection mean average precision (mAP) on the PASCAL VOC 2007 dataset. Our method with the mAP (36.1%) significantly outperforms other methods for most of the categories

trained on the ImageNet dataset. As an initialization step for class-specific proposal mining, we use Edge Boxes [39] to generate 2000 object proposals for each image. The mask-out strategy is first utilized to remove most of the noisy proposals and return top 50 confident proposals. These selected proposals work as the input for multiple instance learning. At the re-training stage, network is trained by employing the SGD solver with the learning rate of 0.0001 for 40k iterations.

Evaluation Metrics   To quantitatively evaluate the performance of the proposed method, we take two types of metrics, which are applied at the training and testing stage respectively. In the training dataset, we compute the percentage of images from which we obtain correct localization (CorLoc) [11]. In the test dataset, we evaluate the performance of the object detector using mean average precision (mAP), a standard metric used in PASCAL VOC. Within both the metrics, we consider that a bounding box is correct if it has an IoU ratio of at least 50% with the ground-truth object annotation.

Comparison with the State-of-the-Art Algorithms   We compared the proposed algorithm with the state-of-the-art methods dealing with the weakly supervised object localization problem [2,4–6]. None of them use strong labels for training.

Figure 3 shows the performance comparison between our proposed method developed with the AlexNet as baseline and the state-of-the-art WSL works [2,4,6] on the VOC 2007 dataset. Models from Song et al. [6] and Bilen et al. [2] are MIL-based approaches with advanced model initialization. Our method is developed based on that from Li et al. [4]. Moreover, Tang et al. [15] propose an on-line instance classifier refinement, which classifies a fixed-size conv feature produced by some convolutional (conv) layers with spatial pyramid pooling (SPP) layer. As the classifier is trained with the features from the SPP net, this model takes the advantage of a better initialization. In an entirely different way, we progressively adapt a classification network to an object detection network with denoised proposals as the pseudo strong labels. Such domain adaptation helps to learn a better object detector from image-level annotated data. Unlike previous works that rely on noisy proposals to localize the object candidates, we mine finer and class-specific proposals from the proposed work flow, which integrates the mask-out, MIL and subset proposal optimization. In addition, a fully model adaption is guaranteed with the re-training and re-weighting strategy.

By incorporating the proposal subset optimization, the proposed model significantly outperforms other methods in terms of mAP for most of the categories. In Table 1, we make

Table 1 Quantitative comparison in terms of detection mean average precision (mAP) on the PASCAL VOC 2007 test set and correct localization (CorLoc) on the PASCAL VOC 2007 trainval set using AlexNet or VGGNet

Methods (VOC 2007)	CorLoc	mAP
Li et al. [4]	49.8	31.0
Shi et al. [5] (AlexNet)	**60.9**	<u>36.0</u>
Our scheme	<u>53.4</u>	**37.2**
Li et al. [4]	52.4	39.5
Shi et al. [5]	**64.7**	37.2
Bilen et al. [3] (VGGNet)	53.5	34.8
Jie et al. [14]	56.1	40.8
Tang et al. [15]	<u>60.6</u>	**41.2**
Our scheme	55.9	<u>40.9</u>
Methods (VOC 2012)	CorLoc	mAP
Li et al. [4]	–	22.4
Our scheme (AlexNet)	–	**25.3**
Li et al. [4]	–	29.1
Jie et al. [14]	54.8	**38.5**
Tang et al. [15] (VGGNet)	**62.1**	<u>37.9</u>
Our scheme	<u>55.2</u>	35.2

The last rows show the mAP on the PASCAL VOC 2012 val set. We highlight the best performances and underline the 2nd best performances

comparisons in terms of both the CorLoc and mAP on the training and testing set of the VOC 2007 dataset, respectively. In addition, we present the mAP on the val set of the VOC 2012. For other baseline methods, we list the best performances of the AlexNet and VGGNet models, which are reported in the paper. Based on the VGGNet, our method achieves 40.9% mAP on VOC 2007 test set and 35.2% mAP on VOC 2012 val set. It is also evident from Table 1 that the detection performance is significantly improved by using a deeper network. Note that the method introduced by jie et al. [14] is a regional CNN detector (Fast R-CNN [43]). This model trained on seed samples is sufficiently powerful for selecting the most confident tight positives and is able to further train itself with the optimized proposals. We compare our method against this Fast RCNN based method by listing the results in Table 1. A similar performance is obtained by our model as the one on VOC 2007.

In addition to the standard IoU for evaluation, we analyze the influence over different IoU threshold in Fig. 4. It is evident that setting IoU = 0.5 achieves the best performance, and the results are not very sensitive to different values: when changing it from 0.5 to 0.6, the performance only drops a little bit.

Impact of Re-training Strategy The re-training strategy we utilized so far is straightforward. The process is to establish an order that adaptively optimize the refined proposals, and then fine-tune the detection network with the confident proposals. We notice that the proposals used to fine-tune the network are critical to train the baseline for detection. So it is promising to improve the annotation through an adaptive way.

We use the same settings during the re-training stage as we adapt the classification network to a detection network. After training the detection network, we select the top 30 detection results and optimize them with the proposal subset optimization. Consequently, the training

Fig. 4 Performance over different IoU threshold of the VGG16 version on PASCAL VOC 2007
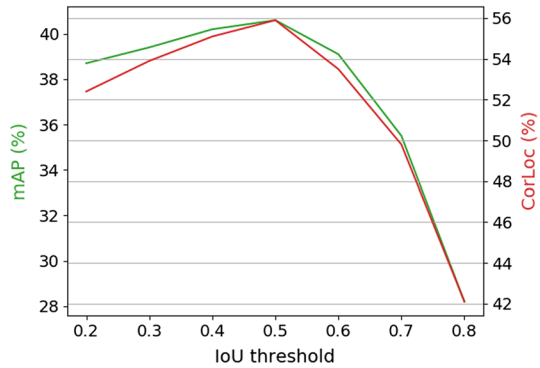

Table 2 Quantitative comparison in terms of detection mean average precision (mAP) on the PASCAL VOC 2007 test set for different re-training steps with AlexNet or VGGNet

Re-train (mAP)	0th	1st	2nd	3rd	4rd
AlexNet	31.0	34.1	36.8	37.2	37.2
VGGNet	38.5	39.9	40.3	40.9	40.9

Table 3 Quantitative comparison in terms of computational time (hour) on the PASCAL VOC 2007 and 2012 training sets for different strategies

Strategy	BBox-initialization		Re-weighting	Re-training		
	Mask-out	MIL	Subset optimization	AlexNet	VGG16	VGG19
VOC 2007	3	24	2	4	7	9
VOC 2012	7	36	3	7	14	17

dataset is adaptively denoised and we obtain a better detection network. Table 2 demonstrates that the mAP is increased from 31.0 to 37.2% for the AlexNet and from 38.5 to 40.9% for the VGGNet.

Computational Time Analysis We report the evaluation results on PASCAL VAL 2007 and PASCAL VAL 2012 in the paper. The re-training is conducted under AlexNet, VGG16, and VGG19. The training time of the experiments largely depends on the hardware resource. We train and evaluate the proposed method using the Intel Xeon(R) CPU E5-1607 v2 @ 3.00 GHz × 4 and four K80 GPUs with 12 GB memory on a cluster. To reduce the training time of MIL, we employ 12 CPUs to separately train the MIL for each category of 21 classes. The training time of the experiment is shown in Table 3.

Error analysis Figure 5 shows some samples with accurate detections and Fig. 6 shows several examples with wrong detections. Our model often detects the correct objects in the image since we train the detector by incorporating a proposal subset optimization to improve the inaccuracy of the localization. Most of the model for WSL task may fail to predict a sufficient tight bounding box [4]. The adaptive denoising part of Fig. 1 illustrates the procedure that the proposals are adaptively selected so that they gradually converge to the ground-truth of annotations. Nonetheless, the proposed model still has limitation as shown in
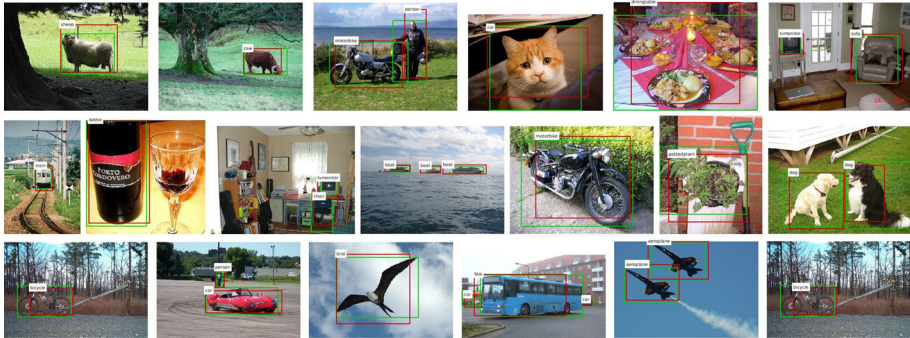
Fig. 5 Sample detection results. Green boxes indicate ground-truth annotation. Red boxes indicate correct detections (with IoU $\geq$ 0.5). The sample images show the correct detections from different classes. (Color figure online)
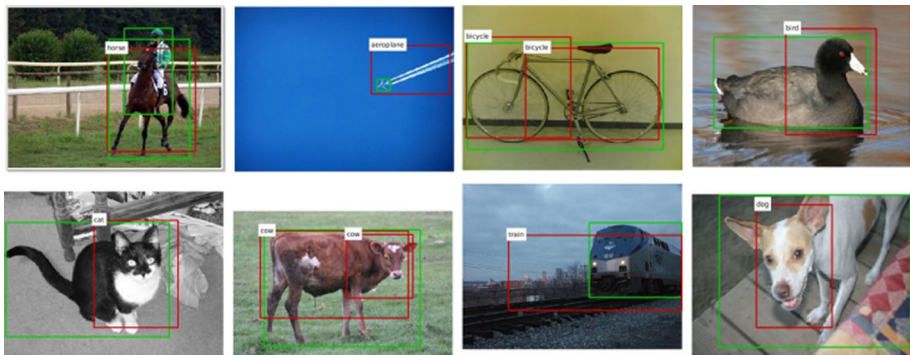

Fig. 6 The sample images shows the wrong detections due to imprecise detection. Green boxes indicate ground-truth annotation. Red boxes indicate imprecise detections (with IoU < 0.5). (Color figure online)

the wrong detections in Fig. 6. This is because our proposal subset optimization also depends on the detection scores even though it incorporates the overlaps of the proposals.

5 Conclusion

We have proposed a novel model by integrating adaptive proposal denoising strategies to handle the weakly supervised object localization problem. This approach first selects confident proposals by utilizing the output of the MIL framework as the starting point of training the detection network. At the training stage, we first adapt a pre-trained classification network with high confident proposals to a detection network, then re-weight the detection results with the proposal subset optimization method. The re-weighted proposals are taken to re-train the network, resulting a detection network that achieves competitive performance on PASCAL VOC datasets. As a follow-up study, it is desire to adapt a new feature extraction method for the weakly supervised localization task. It is interesting to add the attention mechanism that assists to obtain attended features. We would like to introduce a module that effectively and efficiently extracts purified features.

References

1. Bazzani L, Bergamo A, Anguelov D, Torresani L (2016) Self-taught object localization with deep networks. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–9
2. Bilen H, Pedersoli M, Tuytelaars T (2015) Weakly supervised object detection with convex clustering. In: CVPR
3. Bilen H, Vedaldi A (2015) Weakly supervised deep detection networks. arXiv preprint arXiv:1511.02853
4. Li D, Huang J-B, Li Y, Wang S, Yang M-H (2016) Weakly supervised object localization with progressive domain adaptation. In: CVPR
5. Shi M, Ferrari V (2016) Weakly supervised object localization using size estimates. In: ECCV
6. Song HO, Girshick RB, Jegelka S, Mairal J, Harchaoui Z, Darrell T et al (2014) On learning to localize objects with minimal supervision. In: ICML
7. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: CVPR
8. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: CVPR
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: CVPR. IEEE
10. Cinbis RG, Verbeek J, Schmid C (2014) Multi-fold mil training for weakly supervised object localization. In: CVPR
11. Deselaers T, Alexe B, Ferrari V (2012) Weakly supervised localization and learning with generic knowledge. IJCV 100(3):275–293
12. Gokberk Cinbis R, Verbeek J, Schmid C (2014) Multi-fold mil training for weakly supervised object localization. In: CVPR
13. Pentina A, Sharmanska V, Lampert CH (2015) Curriculum learning of multiple tasks. In: CVPR
14. Jie Z, Wei Y, Jin X, Feng J, Liu W (2017) Deep self-taught learning for weakly supervised object localization. In: CVPR
15. Tang P, Wang X, Bai X, Liu W (2017) Multiple instance detection network with online instance classifier refinement. In: CVPR
16. Wu J, Yu Y, Huang C, Yu K (2015) Deep multiple instance learning for image classification and auto-annotation. In: CVPR
17. Zhang Y, Sohn K, Villegas R, Pan G, Lee H (2015) Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction. In: CVPR
18. Zhu Y, Urtasun R, Salakhutdinov R, Fidler S (2015) Segdeepm: exploiting segmentation and context in deep neural networks for object detection. In: CVPR
19. Zhang J, Sclaroff S, Lin Z, Shen XH, Price B, Mech R (2016) Unconstrained salient object detection via proposal subset optimization. In: CVPR
20. Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A (2015) The Pascal visual object classes challenge: a retrospective. IJCV 111(1):98–136
21. Xu W, Choi D, Wang G (2018) Direct visual-inertial odometry with semi-dense mapping. Comput. Electr. Eng. 67:761–775
22. Xu W, Choi D (2016) Direct visual-inertial odometry and mapping for unmanned vehicle. In: International symposium on visual computing. Springer, pp 595–604
23. Wang G, Wang X, Fan B, Pan C (2017) Feature extraction by rotation-invariant matrix representation for object detection in aerial image. IEEE Geosci Remote Sens Lett 14(6):851–855
24. Liu X, Song M, Zhang L, Tao D, Bu J, Chen C (2012) Pedestrian detection using a mixture mask model. In: Proceedings of 2012 9th IEEE international conference on networking, sensing and control. IEEE, pp 271–276
25. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 99:1–13
26. Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoencoder for human pose recovery. IEEE Trans Image Process 24(12):5659–5670
27. Hong C, Chen X, Wang X, Tang C (2016) Hypergraph regularized autoencoder for image-based 3D human pose recovery. Signal Process 124:132–140
28. Yu J, Hong C, Rui Y, Tao D (2017) Multitask autoencoder model for recovering human poses. IEEE Trans Ind Electron 65(6):5060–5068

29. Xu W, Keshmiri S, Wang G (2019) Adversarially approximated autoencoder for image generation and manipulation. IEEE Trans Multimed 21(9):2387–2396
30. Xu W, Keshmiri S, Wang G (2019) Toward learning a unified many-to-many mapping for diverse image translation. Pattern Recognit 93:570–580
31. Xu W, Shawn K, Wang G (2019) Stacked wasserstein autoencoder. Neurocomputing 363:195–204
32. Cen F, Wang G (2019) Dictionary representation of deep features for occlusion-robust face recognition. IEEE Access 7:26595–26605
33. He L, Wang G, Hu Z (2018) Learning depth from single images with deep neural network embedding focal length. IEEE Trans Image Process 27(9):4676–4689
34. He L, Yu M, Wang G (2018) Spindle-net: CNNs for monocular depth inference with dilation kernel method. In: 2018 24th International conference on pattern recognition (ICPR). IEEE, pp 2504–2509
35. Ma W, Wu Y, Wang Z, Wang G (2018) Mdcn: multi-scale, deep inception convolutional neural networks for efficient object detection. In: 2018 24th International conference on pattern recognition (ICPR). IEEE, pp 2510–2515
36. Cinbis RG, Verbeek J, Schmid C (2017) Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans Pattern Anal Mach Intell 39(1):189–203
37. Shi Z, Siva P, Xiang T (2017) Transfer learning by ranking for weakly supervised object annotation. arXiv preprint arXiv:1705.00873
38. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. IJCV 104(2):154–171
39. Zitnick CL, Dollár P (2014) Edge boxes: locating object proposals from edges. In: ECCV
40. Siva P, Russell C, Xiang T (2012) In defence of negative mining for annotating weakly labelled data. In: ECCV
41. Song HO, Lee YJ, Jegelka S, Darrell T (2014) Weakly-supervised discovery of visual pattern configurations. In: NIPS
42. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: CVPR
43. Girshick R (2015) Fast R-CNN. In: ICCV
44. Hoffman J, Guadarrama S, Tzeng ES, Hu R, Donahue J, Girshick R, Darrell T, Saenko K (2014) LSDA: large scale detection through adaptation. In: NIPS
45. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR
46. Lapin M, Hein M, Schiele B (2014) Learning using privileged information: SVM+ and weighted SVM. Neural Netw 53:95–108
47. Sharmanska V, Quadrianto N, Lampert CH (2013) Learning to rank using privileged information. In: ICCV
48. Tudor Ionescu R, Alexe B, Leordeanu M, Popescu M, Papadopoulos DP, Ferrari V (2016) How hard can it be? Estimating the difficulty of visual search in an image. In: CVPR
49. Guillaumin M, Ferrari V (2012) Large-scale knowledge transfer for object localization in imagenet. In: CVPR
50. Rochan M, Wang Y (2015) Weakly supervised localization of novel objects using appearance transfer. In: CVPR
51. Lai B, Gong X (2017) Saliency guided end-to-end learning for weakly supervised object detection. arXiv preprint arXiv:1706.06768
52. Shi M, Caesar H, Ferrari V (2017) Weakly supervised object localization using things and stuff transfer. arXiv preprint arXiv:1703.08000
53. Bodla N, Singh B, Chellappa R, Davis LS (2017) Improving object detection with one line of code. arXiv preprint arXiv:1704.04503
54. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: ECCV
55. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556