Boosting Weakly Supervised Object Detection with Progressive Knowledge Transfer

Yuanyi Zhong[1] *, Jianfeng Wang[2], Jian Peng[1], and Lei Zhang[2]

[1] University of Illinois at Urbana-Champaign {yuanyiz2,jianpeng}@illinois.edu
[2] Microsoft {jianfw,leizhang}@microsoft.com

Abstract. In this paper, we propose an effective knowledge transfer framework to boost the weakly supervised object detection accuracy with the help of an external fully-annotated source dataset, whose categories may not overlap with the target domain. This setting is of great practical value due to the existence of many off-the-shelf detection datasets. To more effectively utilize the source dataset, we propose to iteratively transfer the knowledge from the source domain by a one-class universal detector and learn the target-domain detector. The box-level pseudo ground truths mined by the target-domain detector in each iteration effectively improve the one-class universal detector. Therefore, the knowledge in the source dataset is more thoroughly exploited and leveraged. Extensive experiments are conducted with Pascal VOC 2007 as the target weakly-annotated dataset and COCO/ImageNet as the source fully-annotated dataset. With the proposed solution, we achieved an mAP of 59.7% detection performance on the VOC test set and an mAP of 60.2% after retraining a fully supervised Faster RCNN with the mined pseudo ground truths. This is significantly better than any previously known results in related literature and sets a new state-of-the-art of weakly supervised object detection under the knowledge transfer setting. Code: https://github.com/mikuhatsune/wsod_transfer.

Keywords: weakly supervised, object detection, transfer learning, semi-supervised

1 Introduction

Thanks to the development of powerful CNNs and novel architectures, the performance of object detectors has been dramatically improved in recent years [9,7,21,37]. However, such successes heavily rely on supervised learning with fully annotated detection datasets which can be costly to obtain, since annotating locations and category labels of all object instances is time-consuming and sometimes prohibitively expensive. This issue has motivated many prior works on weakly supervised object detection (WSOD), where only image-level labels are available and normally much cheaper to obtain than box-level labels.

* Part of this work was done when the first author was an intern at Microsoft.

Existing WSOD methods [25,2,27,26] are mostly based on multiple instance learning (MIL), in which an image is represented as a bag of regions, e.g., generated by selective search [31]. The training algorithm needs to infer which instances in a bag are positive for a positive image-level class. Thus, the problem of learning a detector is converted into training an MIL classifier.

Compared to fully supervised detectors, a large performance gap exists for weakly supervised detectors. For example, on the Pascal VOC 2007 dataset [6], a fully supervised Faster RCNN can achieve an mAP of 69.9% [21], while the state-of-the-art weakly supervised detector, to the best of our knowledge, can only reach to an mAP of 53.6% [33].

One direction to bridge the performance gap is to utilize in a domain transfer learning setting the well-annotated external source datasets, many of which are publicly available on the web, e.g., COCO [18], ImageNet [4], Open Images [15], and Object 365 [23]. Due to the existence of these off-the-shelf detection datasets, this domain transfer setting is of great practical value and has motivated many prior works, under the name transfer learning [5,29,24,34,14,16], domain adaptation [10,17,3,11,12], and mixed supervised detection [35]. For example, [30] proposes to train a generic proposal generator on the source domain and an MIL classifier on the target domain in a one-step transfer manner. In [16], a universal bounding box regressor is trained on the source domain and used to refine bounding boxes for a weakly supervised detector. In [35], a domain-invariant objectness predictor is utilized to filter distracting regions before applying the MIL classifier. Other related works include [5,29,24,34,14,17,3,12].

Although the domain transfer idea is very promising, it is worth noting that the top pure weakly supervised detector [33] actually outperforms the best transfer-learned weakly supervised detector [35,16] on VOC in the literature. Despite many challenges in domain transfer, one technical deficiency particularly related to object detection lies in imperfect annotations, where the source images may contain objects of the target domain categories but unannotated. In such cases, the object instances will be treated as background regions (or false negatives) in the source data, which is known as the incomplete label problem in object detection [32]. As a result, detectors trained with the source data will likely have a low recall of objects of interest in the target domain.

To address this problem, we propose to transfer progressively so that the knowledge can be extracted more thoroughly by taking into account the target domain. Specifically, we iterate between extracting knowledge by a one-class universal detector (OCUD) and learning a target domain object detector through MIL. The target domain detector is used to mine the pseudo ground truth annotations in both the source and target datasets to refine the OCUD. Compared with existing works, the key novelty is to extract knowledge in multi-steps rather than one-step. Technically, by adding pseudo ground truths in the source data, we effectively alleviate the problem of false negatives as aforementioned. By adding pseudo ground truths in and including the target dataset in fine-tuning, the refined OCUD is more adapted to the target domain data distribution. Empirically, we observe significant gains, e.g., from 54.93% mAP with one-step

transfer to 59.71% with multi-step transfer (5 refinements) on Pascal VOC 2007 test data by leveraging COCO-60 as source (removing the VOC 20 categories). By retraining a fully supervised Faster RCNN with the mined pseudo ground truths, we can achieve 60.24% mAP, which again surpasses the pure WSOD method [33] remarkably and sets a new state of the art under the transfer setting. Finally, as a reference, the detection performance also surpasses the original fully supervised Faster RCNN with the ZF net backbone (59.9% mAP) [21].

2 Related Work

Weakly Supervised Object Detection (WSOD). WSOD is extensively studied in the literature [25,2,27,26]. The problem is often formulated as an image classification with multi-instance learning. Typically, candidate bounding boxes are first generated by independent proposal methods such as Edge Boxes [39] and Selective Search [31]. Then the proposals on one image are treated as a bag with the image labels as bag-level labels. WSDDN [2] utilizes a two-stream architecture that separates the detection and classification scores, which are then aggregated through softmax pooling to predict the image labels. OICR [27] and the subsequent PCL [26] transform the image-level labels into instance-level labels by multiple online classifier refinement steps. Class activation maps can also be used to localize objects [38]. WSOD2 [33] exploits the bottom-up and top-down objectness to improve performance. Among existing works, pseudo ground truth mining is heavily used as a tool for iterative refinement [28,26,36].

Classifier refinement methods such as OICR [27] and PCL [26] are related in that they conduct refinement steps. Our method is similar to them when restricted to operating on the target data only. However, there are several notable differences. We study the WSOD-with-transfer rather than the pure WSOD setting. Our pseudo ground truth mining is conducted on both the source and target data. We refine both the classifier and the box proposals by retraining the OCUD rather than the instance classifier only.

WSOD with Knowledge Transfer. One way to improve the accuracy is to utilize a source dataset and transfer the knowledge to the target domain through semi-supervised or transfer learning. Visual or semantic information in the category labels or images is often exploited to help solve the problem. For example, the word embeddings of category texts are employed in [29,1] to represent class semantic relationships. The appearance model learned on the source classes are transferred to the target classes in [24,22,17]. Many methods leverage weight prediction to effectively turn a novel category classifier into a detector [14,10,29]. For example, LSDA [10] and [29] transfer the classifier-to-detector weight differences. Recent works [5,30,16,35] share with us in spirit learning general object knowledge from the source data. The knowledge can either be the objectness predictor [5,35], the object proposals [30] or the universal bounding box regressor [16]. In particular, [30] also trains a universal detector (in their case, SSD [19]) on the source dataset, and uses the detection results from this detector as proposals during MIL on the target dataset. The process can be seen as a

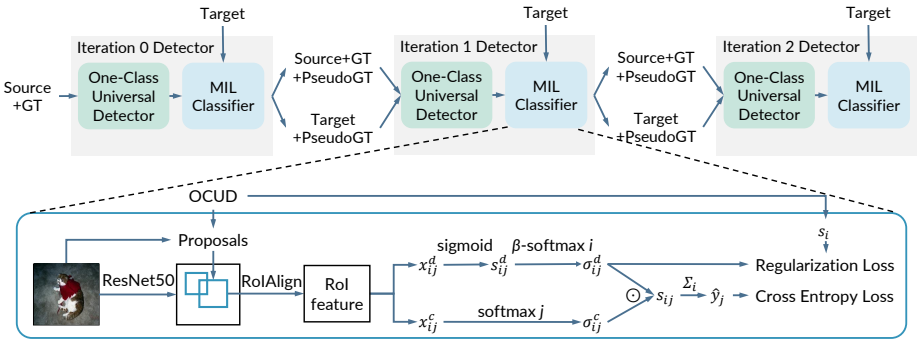


Fig. 1. An illustration of the proposed progressive knowledge transfer framework. One-class universal detector (OCUD) is initially trained with fully annotated source data and iteratively refined on source and target data with pseudo ground truths (GT). OCUD acts as the proposal generator during the subsequent training of target domain MIL classifiers. OCUD and MIL classifier together form the target domain detector.

special case of our algorithm with a single-step transfer and a different instantiation of network and MIL method. Comparatively, we differentiate our approach from them by progressively exploiting the knowledge in the source dataset in a multi-step way, such that the accuracy can improve gradually. Empirically, we observed non-trivial performance gain with progressive knowledge transfer.

3 Proposed Approach

Given source dataset $\mathcal{S}$ with bounding box annotations and target dataset $\mathcal{T}$ with only image-level labels, the goal is to train an object detector for object categories in $\mathcal{T}$. The categories of $\mathcal{S}$ and $\mathcal{T}$ can be non-overlapping, which differentiates our setting from a typical semi-supervised setting.

The proposed training framework and workflow are outlined in Fig. 1 and Alg. 1. The basic flow is to first train a target domain detector as a seed based on the existing labels, and then mine the pseudo ground truth boxes, which are then used to refine the detector. The process is repeated to improve the target domain detector gradually since more target domain boxes can be found in both $\mathcal{S}$ and $\mathcal{T}$ through the process. The architecture design of the detector is versatile. Here we present a simple solution consisting of a one-class universal detector (OCUD) and a MIL classifier.

3.1 One-Class Universal Detector (OCUD)

The one-class universal detector, which we refer to as OCUD for convenience, treats all categories as a single generic category. While we employ Faster RCNN [21] with ResNet50 [9] backbone, any modern object detector can be used.

Algorithm 1: WSOD with Progressive Knowledge Transfer.

Input: Max number of refinements N, source dataset $\mathcal{S}$, target dataset $\mathcal{T}$;
Train the one-class universal detector (OCUD) on the source dataset $\mathcal{S}$;
Train the MIL classifier based on the OCUD and the target dataset $\mathcal{T}$;
for $K = 1, 2, \dots N$ **do**
    Mine pseudo ground truths in $\mathcal{S}$ and $\mathcal{T}$ with OCUD and the MIL classifier;
    Refine the OCUD with the mined boxes and original source annotations;
    Refine the MIL classifier based on the OCUD and the target dataset $\mathcal{T}$;
return The OCUD and MIL classifier as the target domain detector;

Initially, the OCUD is trained on source data only, which is similar to [30]. Although the categories can be non-overlapping between the source domain and the target domain, the objects may be visually similar to some extent, which gives the detector certain capability to detect the target domain objects. For example, a detector trained on *cat* might be able to detect *dog*.

3.2 MIL Classifier

With the OCUD, we extract multiple proposals in the target dataset image and perform multiple instance learning (MIL) with the proposals. Our MIL classifier is based on WSDDN [2], but adapted to incorporate knowledge from the OCUD.

The MIL classifier has a two-stage Faster-RCNN-like architecture sketched in Fig. 1. Assume that the OCUD gives R proposals in a target dataset image: $\{b_i, s_i\}_{i=1}^{R}$. We run RoIAlign [8] to extract a feature map for each proposal, and feed the feature into two branches as in [2]: the detection branch and the classification branch. Each branch consists of 2 linear layers with ReLU. The last linear layer's output has the same dimension as the number of target domain categories. Let $x_{ij}^{\mathrm{d}} \in \mathbb{R}$ and $x_{ij}^{\mathrm{c}} \in \mathbb{R}$ be the output for the i-th proposal and the j-th category from the detection branch and the classification branch, respectively. The predicted score s_{ij} is calculated as follows,

$$\begin{aligned} s_{ij}^{\mathrm{d}} &= \mathrm{sigmoid}(x_{ij}^{\mathrm{d}}), \quad & \sigma_{ij}^{\mathrm{d}} &= \mathrm{softmax}_i(\beta s_{ij}^{\mathrm{d}}), \\ \sigma_{ij}^{\mathrm{c}} &= \mathrm{softmax}_j(x_{ij}^{\mathrm{c}}), \quad & s_{ij} &= \sigma_{ij}^{\mathrm{d}} \sigma_{ij}^{\mathrm{c}}. \end{aligned} \tag{1}$$

The softmax is computed along the i and j dimensions respectively. Different from [2], we squash the detection scores s_{ij}^{d} to $(0, 1)$ by sigmoid. This has two benefits: (1) It allows multiple proposals to belong to the same category: s_{ij}^{d} represents how likely each proposal individually belongs to category j, and σ_{ij}^{d} is a normalization; (2) It makes it easier to enforce the objectness regularization as we shall see below. To make σ_{ij}^{d} amenable to train, we introduce a scaling factor β to adjust the input range from $(0, 1)$ to $(0, \beta)$. With a larger β, the range of the scaled softmax is wider, and the value is easier to be spiked.

Let $\{y_j\}_{j=1}^{C} \in \{0, 1\}^C$ be the image-level label, and C be the number of categories. Given the scores of all proposals, an image-level classification prediction

$\hat{y}_j$ is calculated and used in an image-level binary classification loss L_{wsddn},

$$\hat{y}_j = \sum_{i=1}^{R} s_{ij}, \quad L_{\text{wsddn}} = -\frac{1}{C}\sum_{j=1}^{C} y_j \log \hat{y}_j + (1 - y_j)\log(1 - \hat{y}_j). \tag{2}$$

To further exploit the knowledge present in OCUD, we introduce the following L_2 regularization loss on the detection branch scores s_{ij}^{d}. The intuition behind is that the objectness score s_i predicted by the OCUD could guide MIL by promoting the object candidates' confidence. It should match the target domain detector's objectness of region i defined as the max over classes. The overall training loss for each image is the weighted sum with coefficient λ as in Eq. 4.

$$L_{\text{guide}} = \frac{1}{R}\sum_{i=1}^{R} \big(\max_{1 \le j \le C} s_{ij}^{\text{d}} - s_i \big)^2. \tag{3}$$

$$\mathcal{L} = L_{\text{wsddn}} + \lambda L_{\text{guide}}. \tag{4}$$

During inference, the final detection score is the linear interpolation of s_i from the OCUD and s_{ij} from the MIL classifier. This scheme is shown to be robust [30]. Specifically, with a coefficient $\eta \in [0, 1]$, we compute the final score by Eq. 5. The model trusts the MIL classifier more with a larger η.

$$s_{ij}^{\text{final}} = \eta s_{ij} + (1 - \eta) s_i. \tag{5}$$

3.3 Pseudo Ground Truth Mining

Given the OCUD and the MIL classifier, we mine the pseudo ground truth on both the source and the target dataset based on the latest target domain detector (OCUD + MIL classifier). Following [27,33,13], we adopt the simple heuristic to pick the most confident predictions, as summarized in Alg. 2.

In the source dataset, the predictions with high confidence (thresholded by τ) and low overlap ratio (thresholded by o) with the nearest ground truth bounding box are taken as a pseudo ground truth. Here we use the intersection over the predicted bounding box area as the overlap ratio, to conservatively mine the box in the source data and avoid mining object parts. Empirically, this simple scheme is effective to locate target domain objects in the source dataset.

In the target dataset, the image-level labels are used to filter the predictions in addition to the confidence scores. For each positive class, we select as pseudo ground truth the top one box and any detection result with a confidence score higher than the threshold τ. In this way, any misclassified bounding box is filtered out, and each positive class is guaranteed to have at least one box.

3.4 Refinement of OCUD and MIL Classifier

Pseudo ground truth augmented source and target datasets are used to refine the OCUD. The fine-tuning is the same as the initial OCUD training, except

Algorithm 2: Pseudo Ground Truth Mining.

Input: Detector $D_{\mathcal{T}}$, source $\mathcal{S}$, target $\mathcal{T}$, score threshold τ, overlap threshold o

$\mathcal{S}^+ \leftarrow \emptyset$, $\mathcal{T}^+ \leftarrow \emptyset$;
for (image I, boxes B) in $\mathcal{S}$ **do**
  predictions $P \leftarrow D_{\mathcal{T}}(I)$; annotations $anno \leftarrow B$;
  for predicted box p in P **do**
    if $p.score > \tau$ **then**
      $overlaps \leftarrow \text{overlap}(p.box, B)/\,\text{area}(p.box)$;
      if $\max overlaps < o$ **then** add p to $anno$;
  add $(I, anno)$ to $\mathcal{S}^+$;
for (image I, image label Y) in $\mathcal{T}$ **do**
  predictions $P \leftarrow D_{\mathcal{T}}(I)$; annotations $anno \leftarrow \emptyset$;
  for category y in Y **do**
    find subset predictions $P_y \leftarrow \{p \in P : p.category = y\}$;
    for box p in P_y **do**
      if $p.score > \tau$ or $p.score = \max P_y.scores$ **then** add p to $anno$;
  add $(I, anno)$ to $\mathcal{T}^+$;
return $\mathcal{S}^+$, $\mathcal{T}^+$;

that the two domain images are now mixed together, and the model is initialized from the last OCUD. More advanced techniques can be leveraged, e.g., assigning different weights for the pseudo ground truth in the target dataset, the source dataset, and the original source annotations. We leave it as future work.

In the experiments, we find this simple refinement approach is effective. Through the last target domain detector, the mined pseudo ground truth boxes are better aligned towards the target domain categories. In the target dataset, the objects could be correctly localized, and the boxes become the pseudo ground truths to improve the OCUD. In the source dataset, the pseudo ground truth can improve the recall rate, especially when the image content contains the target category (not in the source domain category). Without refinement, those regions will be treated as the background, which is detrimental.

With the improved OCUD, the MIL classifier is also fine-tuned by the improved object proposals detected by the OCUD. Before the refinements, the OCUD contains little information on the target domain categories, and the proposals are generated by solely relying on the similarity of the categories across domains (e.g., being able to detect *horse* might help detect *sheep*). Afterwards, the OCUD is improved to incorporate more information about the target domain, and the proposals will also likely be aligned to improve the MIL classifier.

4 Experiments

4.1 Experiment Settings

Target Dataset. Following [16,35], we use Pascal VOC 2007 dataset [6] as the target dataset, which has 2501 training images with 6301 box-level annotations,

2510 validation images with 6,307 annotations and 4,952 testing images with 12,032 annotations. As in [16,35,2,26,33], we combine the training and validation sets into one trainval set for training, and evaluate the accuracy on the test set. The bounding boxes are removed in the trainval set, and only the image-level labels are kept for the weakly supervised training. There are 20 categories.

Source Dataset. Similar to [16], we use COCO [18] 2017 detection dataset as the source dataset, which contains 118,287 training images with 860,001 box-level annotations and 5,000 validation images with 36,781 annotations. The number of categories is 80, and all the 20 categories of VOC are covered. As in [16], we remove all the images that have overlapped categories with VOC, resulting in a train set of 21,987 images with 70,549 annotations, and a validation set of 921 images with 2,914 annotations. The resulting train and validation sets are merged as the source dataset, which we denote as COCO-60. We aim to transfer the knowledge through the one-class universal detector from the COCO-60 dataset to the weakly labeled VOC dataset with no overlapping classes.

Another source dataset we investigate is ILSVRC 2013 detection dataset, which contains 395,909 train images (345,854 box annotations) and 20,121 validation images (55,502 box annotations) of 200 classes. After removing images of the 21[3] categories overlapping with VOC, we arrive at 143,095 train images and 6,229 val images of 179 classes. The train and validation images are combined as the source dataset, denoted as ILSVRC-179 in the ablation study. Without an explicit description, we use COCO-60 as the source dataset.

Evaluation Metrics. We adopt two evaluation metrics frequently used in weakly supervised detection literature, namely mean average precision (mAP) and Correct Localization (CorLoc). Average precision (AP) is the area under the precision/recall curve for each category, and mAP averages the APs of all categories. CorLoc [5] measures the localization accuracy on the training dataset. It is defined as the percentage of images of a certain class that the top one prediction of the algorithm correctly localizes one object. A prediction is correct if the intersection-over-union (IoU) with ground truth is larger than 0.5.

Network. We use the Faster RCNN as the one-class universal detector (OCUD), where the RPN network is based on the first 4 conv stages of ResNet, and the RoI CNN is based on the 5th conv stage. It is worth noting that any detector can be used here. Up to 100 detected boxes are fed from the OCUD to the MIL classifier. ResNet50 is used as the backbone for both OCUD and the MIL classifier. Our implementation is based on maskrcnn-benchmark [20].

Training and Inference. The training is distributed over 4 GPUs, with a batch size of 8 images. The OCUD is initialized with the ImageNet pre-trained model and trained with 17,500 steps. Afterwards, the OCUD is fine-tuned with 5000 steps in the refinements. The MIL classifier is trained for 5000 steps initially and then fine-tuned similarly for 2000 steps in each following refinement. The base learning rate is set to 0.008 for all experiments and all models and is reduced by 0.1 after finishing roughly 70% of the training progress.

[3] ILSVRC has 2 classes *water bottle* and *wine bottle* while COCO and VOC have *bottle*.

Table 1. mAP performance on VOC 2007 test set. 'Ours' are trained with COCO-60 as source. Superscript '+' indicates multi-scale testing. 'Distill' means to re-train a Faster RCNN based on the mined boxes. 'Ens' indicates ensemble methods.

Method	aero	bike	bird	boat	bottl	bus	car	cat	chair	cow	table	dog	horse	mbik	pers.	plant	sheep	sofa	train	tv	mAP
Pure WSOD:																					
WSDDN-Ens [2]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
OICR-Ens+FR [27]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
PCL-Ens+FR [26]	63.2	69.9	47.9	22.6	27.3	71.0	69.1	49.6	12.0	60.1	51.5	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
WSOD2$^+$ [33]	65.1	64.8	57.2	39.2	24.3	69.8	66.2	61.0	29.8	64.6	42.5	60.1	71.2	70.7	21.9	28.1	58.6	59.7	52.2	64.8	53.6
With transfer:																					
MSD-Ens$^+$ [35]	70.5	69.2	53.3	43.7	25.4	68.9	68.7	56.9	18.4	64.2	15.3	72.0	74.4	65.2	15.4	25.1	53.6	54.4	45.6	61.4	51.08
OICR+UBBR [16]	59.7	44.8	54.0	36.1	29.3	72.1	67.4	70.7	23.5	63.8	31.5	61.5	63.7	61.9	37.9	15.4	55.1	57.4	69.9	63.6	52.0
Ours:																					
Ours(single scale)	64.4	45.0	62.1	42.8	42.4	73.1	73.2	76.0	28.2	78.6	28.5	75.1	74.6	67.7	57.5	11.6	65.6	55.4	72.2	61.3	57.77
Ours$^+$	64.8	50.7	65.5	45.3	46.4	75.7	74.0	80.1	31.3	77.0	26.2	79.3	74.8	66.5	57.9	11.5	68.2	59.0	74.7	65.5	59.71
Ours(distill,vgg16)$^+$	62.6	56.1	64.5	40.9	44.5	74.4	76.8	80.5	30.6	75.4	25.5	80.9	73.4	71.0	59.1	16.7	64.1	59.5	72.4	68.0	59.84
Ours(distill)$^+$	65.5	57.7	65.1	41.3	43.0	73.6	75.7	80.4	33.4	72.2	33.8	81.3	79.6	63.0	59.4	10.9	65.1	64.2	72.7	67.2	60.24
Upper bounds:																					
Fully Supervised	75.9	83.0	74.4	60.8	56.5	79.0	83.8	83.6	54.9	81.6	66.8	85.3	84.3	77.4	82.6	47.3	74.0	72.2	78.0	74.8	73.82
Ideal OCUD	70.0	72.4	72.6	51.7	57.5	76.1	80.7	86.8	45.8	81.3	50.6	81.6	78.4	72.5	74.4	45.4	70.1	61.5	76.0	72.9	68.92

It is worth noting that the overhead training time of K refinements is less than K times the usual training time, due to the shortened training schedule for the refinements. For example, in the COCO-60-to-VOC experiments, the initial OCUD and MIL training cost 190min and 23min, but the OCUD and MIL refinements only took 50min and 8min. The testing time of the final distilled detector is similar to the usual detector. The details are in the supplementary.

Without explicit description, the parameter β in Eq. 1 is 5, the λ in Eq. 4 is 0.2, the η in Eq. 5 is 0.5, and the number of refinements is 5. In the phase of pseudo ground truth mining, the confidence threshold τ is 0.8, and the IoU threshold o is 0.1. We also studied the sensitivity of these parameters.

The training images are resized to have a short edge of 640 pixels. During testing, we study both the single-scale no-augmentation configuration, and the multi-scale (two: 320, 640 pixels) setting with horizontal flipping as adopted in prior work [33,35,26]. The non-maximum suppression IoU is 0.4 during testing.

4.2 Comparison with SOTA

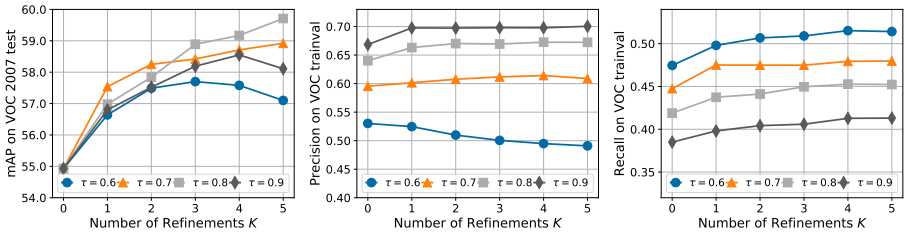
Table 1 and Table 2 compare our approach with previous state-of-the-art approaches in terms of mAP and CorLoc, respectively. We compare to pure WSOD methods: (1) WSDDN-Ens [2], the ensemble of 3 Weakly Supervised Detection Networks. Our two branch MIL is modified from WSDDN. (2) OICR-Ens+FR [27], a Fast RCNN [7] retrained from a VGG ensemble of the Online Instance Classifier Refinement models. (3) PCL-Ens+FR [26], an improvement over OICR [27] which leverages proposal clusters to refine classifiers. (4) WSOD2$^+$ [33], one of the best-performing WSODs on VOC which combines bottom-up and top-down objectness cues. We also compare with two WSOD-with-transfer methods: (1) MSD-Ens$^+$ [35] which transfers the objectness learned from source, (2) OICR+UBBR [16] which learns a universal box regressor on source data.

From the tables, we have the following observations.

Table 2. CorLoc performance on VOC 2007 trainval set. 'Ours' are trained with COCO-60 as source. Superscript '+' indicates multi-scale testing. 'Distill' means to re-train a Faster RCNN based on the mined boxes. 'Ens' indicates ensemble methods.

Method	aero	bike	bird	boat	bottl	bus	car	cat	chair	cow	table	dog	horse	mbik	pers.	plant	sheep	sofa	train	tv	Cor.
Pure WSOD:																					
WSDDN-Ens [2]	68.9	68.7	65.2	42.5	40.6	72.6	75.2	53.7	29.7	68.1	33.5	45.6	65.9	86.1	27.5	44.9	76.0	62.4	66.3	66.8	58.0
OICR-Ens+FR [27]	85.8	82.7	62.8	45.2	43.5	84.8	87.0	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	78.1	64.3
PCL-Ens+FR [26]	83.8	85.1	65.5	43.1	50.8	83.2	85.3	59.3	28.5	82.2	57.4	50.7	85.0	92.0	27.9	54.2	72.2	65.9	77.6	82.1	66.6
WSOD2^{+} [33]	87.1	80.0	74.8	60.1	36.6	79.2	83.8	70.6	43.5	88.4	46.0	74.7	87.4	90.8	44.2	52.4	81.4	61.8	67.7	79.9	69.5
With transfer:																					
WSLAT-Ens [22]	78.6	63.4	66.4	56.4	19.7	82.3	74.8	69.1	22.5	72.3	31.0	63.0	74.9	78.4	48.6	29.4	64.6	36.2	75.9	69.5	58.8
MSD-Ens^{+} [35]	89.2	75.7	75.1	66.5	58.8	78.2	88.9	66.9	28.2	86.3	29.7	83.5	83.3	92.8	23.7	40.3	85.6	48.9	70.3	68.1	66.8
OICR+UBBR [16]	47.9	18.9	63.1	39.7	10.2	62.3	69.3	61.0	27.0	79.0	24.5	67.9	79.1	49.7	28.6	12.8	79.4	40.6	61.6	28.4	47.6
Ours:																					
Ours(single scale)	86.7	62.4	87.1	70.2	66.4	85.3	87.6	88.1	42.3	94.5	32.3	87.7	91.2	88.8	71.2	20.5	93.8	51.6	87.5	76.7	73.6
Ours^{+}	87.5	64.7	87.4	69.7	67.9	86.3	88.8	88.1	44.4	93.8	31.9	89.1	92.9	86.3	71.5	22.7	94.8	56.5	88.2	76.3	74.4
Ours(distill,vgg16)$^{+}$	87.9	66.7	87.7	67.6	70.2	85.8	89.9	89.2	47.9	94.5	30.8	91.6	91.8	87.6	72.2	23.8	91.8	67.2	88.6	81.7	75.7
Ours(distill)$^{+}$	85.8	67.5	87.1	68.6	68.3	85.8	90.4	88.7	43.5	95.2	31.6	90.9	94.2	88.8	72.4	23.8	88.7	66.1	89.7	76.7	75.2
Upper bounds:																					
Fully Supervised	99.6	96.1	99.1	95.7	91.6	94.9	94.7	98.3	78.7	98.6	85.6	98.4	98.3	98.8	96.6	90.1	99.0	80.1	99.6	93.2	94.3
Ideal OCUD	97.5	85.1	96.7	83.5	84.4	91.9	92.5	94.5	65.4	95.2	70.0	94.2	94.6	91.6	90.6	81.3	96.9	61.3	96.6	88.2	87.6

1. Our approach without multi-scale testing and model retraining (distilling) achieves significantly higher accuracy than any pure WSOD. In terms of mAP, the gain is more than 4 points from 53.6% [33] to 57.77% (ours). For CorLoc, it is from 59.5% [33] to 73.6%. This demonstrates the superior advantage of leveraging existing detection source dataset to help the novel or unseen weakly supervised training task.
2. Compared with the approaches using external data, our approach performs consistently higher than the top related approach both in mAP and CorLoc. The best previous mAP is 52.0% [16], and the best CorLoc is 66.8% [35]. Both numbers are behind the best pure WSOD approach, and the reason might be the insufficient utilization of the external data. Instead, we utilize the external data more thoroughly with multiple progressive refinements, which significantly boosts the final accuracy.
3. For our approach, the multi-scale testing gives around 2 points' gain in mAP and 1 point's gain in CorLoc.
4. Similar to [17,13,26], we retrain a Faster RCNN detector on the VOC trainval images with the pseudo box annotations from our OCUD and MIL classifier. With VGG16 as the backbone, the accuracy (shown with distill) is 59.84% in mAP and 75.7% in CorLoc. With a more powerful backbone of ResNet50, mAP is 60.24% and CorLoc is 75.2%. Though the backbones are notably different, we observe the accuracy does not change accordingly, and the bottleneck may still be the quality of the mined pseudo ground truth. With 60.24% mAP, our approach surpasses the Faster RCNN fully supervised detector with the ZF network backbone (59.9% mAP) [21].
5. Two numbers are reported as upper bounds in Table 1. The first one is a fully supervised Faster RCNN (ResNet50) based on the true box annotations, which achieves 73.82% mAP. The other upper bound is estimated based on our training pipeline but with the fully annotated VOC as the source dataset.



(a) mAP (b) Precision (c) Recall

Fig. 2. Accuracy with different pseudo ground truth mining thresholds. (a) mAP on VOC test; (b)/(c): precision/recall of the mined pseudo ground truth on VOC trainval.

That is, the true ground truth bounding boxes of VOC trainval are used to train the OCUD, which yields 68.92% mAP. Thus, the gap from 60.24% (our best result) to 68.92% mAP may mainly come from data disparity between the source and the target, signifying room for further improvement. Investigating a more advanced pseudo ground truth mining approach and resorting to more source data could help close the gap in the future.

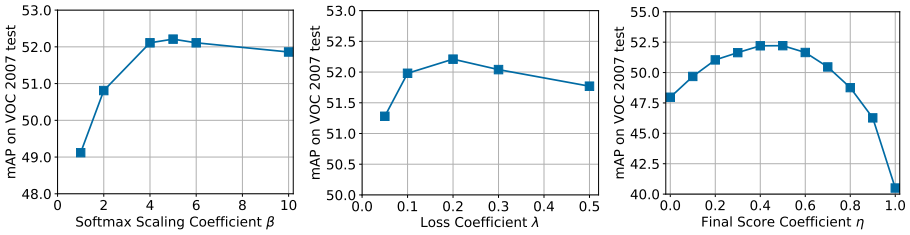
4.3 Ablation Study

***τ* and *K*.** Fig 2(a) shows the mAP with multi-scale testing under different thresholds of τ (0.6,0.7,0.8,0.9) and the number of refinements K. Fig 2(b) and (c) shows the corresponding precision and recall of the pseudo ground truth on the target dataset. The threshold τ is used in the pseudo ground truth mining in Alg. 2. From (b) and (c), we can see a higher threshold leads to higher precision but lower recall and vice versa. The threshold of 0.8 achieves the best trade-off mAP with $K \geq 3$. When $K \leq 2$, a smaller threshold is better. This is reasonable because more boxes can be leveraged.

Along the dimension of K, the precision and recall improve in general, except for $\tau = 0.6$ where the precision deteriorates when $K \geq 3$. For $\tau = 0.8$, the accuracy improves significantly from 55.0% to 59.7% when the number of refinements is increased from 0 to 5. The gradual accuracy improvement indicates that one-step knowledge transfer is sub-optimal, and the final accuracy benefits a lot from more iterations of knowledge transfer.

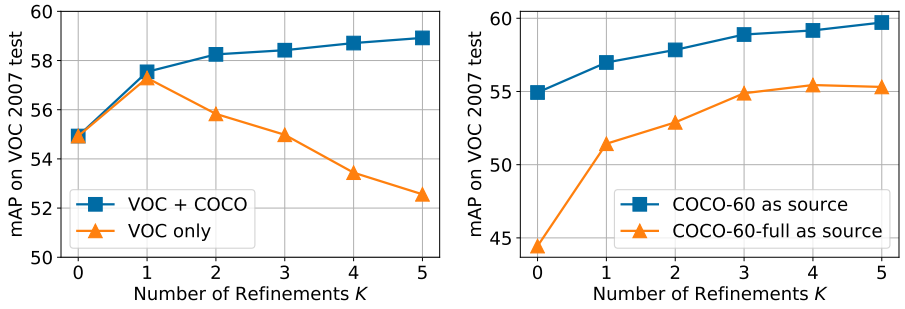
***β*.** The β parameter in Eq. 3 scales the detection score $s_{ij}^d \in (0, 1)$ before softmax. When $\beta = 0$, it is equivalent to remove the detection branch. When $\beta \to +\infty$, all the non-maximum values are zero after softmax, which reduces the importance of the classification branch. The best accuracy locates at $\beta = 5$ in Fig. 3(a).

***λ*.** Coefficient λ balances the image classification loss L_{wsddn} and detection score regularization L_{guide} in Eq. 4. A larger λ means stronger regularization. The result is shown in Fig. 3(b), and $\lambda = 0.2$ delivers the best performance. A non-zero λ performing well suggests that the OCUD can provide valuable information to guide the MIL classifier learning, which is overlooked in previous work [30].



(a) (b) (c)

Fig. 3. Ablation study of the scaling factor β in Eq. 1, λ in Eq. 4 and η in Eq. 5. The accuracy is based on the initial OCUD and MIL classifier with single-scale inference.



(a) (b)

Fig. 4. Accuracy with different configurations of the source datasets: (a) VOC+COCO vs VOC, (b) COCO-60 vs COCO-60-full. The accuracy is based on multi-scale testing.

***η*.** Linear coefficient η in Eq. 5 balances the score from the MIL classifier and the OCUD during inference. As illustrated in Fig. 3(c), the accuracy is worse if we rely on either MIL classifier ($\eta = 1$) or the OCUD ($\eta = 0$) alone. The best accuracy is located at $\eta = 0.4 \sim 0.5$.

VOC vs VOC+COCO. After we have the initial OCUD, one alternative is to remove the source dataset afterwards. Fig. 4(a) shows the experiment results with $\tau = 0.7$. As we can see, without the source dataset, the performance drops dramatically after one refinement. The reason might be that the error of the mined box annotation can be accumulated and the OCUD becomes unstable without the guidance of the manually-labeled boxes. This ablation is similar to pure WSOD methods such as OICR [27], where the detector is refined only on the target data. The inferior result suggests that transferring knowledge from the source is indeed critical in the success of our method.

COCO-60 vs COCO-60-full. Following [16,35], we removed all images in the source dataset (COCO) which has overlapping categories with the target VOC dataset. Instead of removing the images, we also conduct the experiments by keeping the images but removing the annotations of overlapping categories, and

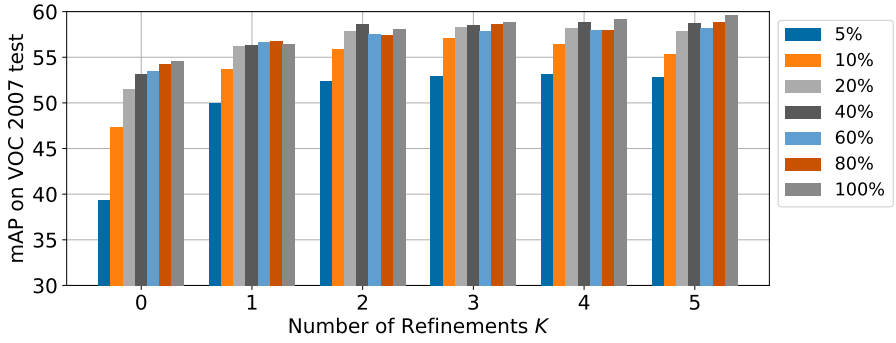


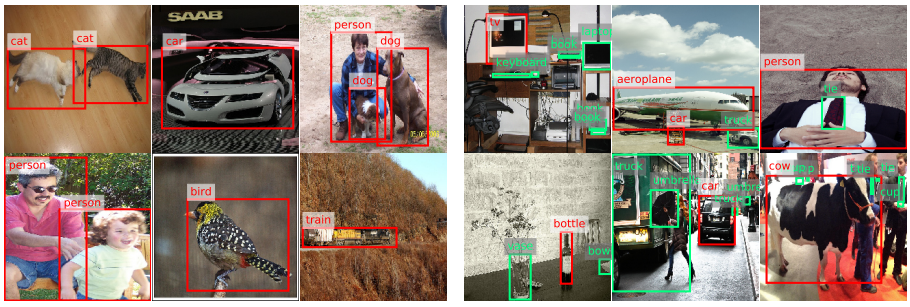
Fig. 5. Ablation study on the size of the source dataset (subsets of COCO-60).

denote this source set by COCO-60-full. Fig. 4(b) shows the experiment results. Obviously, the accuracy with COCO-60 is higher than that with COCO-60-full. The reason is that the regions with the annotation removed are treated as background in OCUD, which will reduce the recall rate for COCO-60-full. Another observation is that even with this challenging source dataset, we can still boost the accuracy from less than 45% to more than 55%, with a gain of more than 10 points with our progressive transfer learning. Comparatively, the gain on COCO-60 is much less at around 5 points. The reason is that the propagation on the COCO-60-full can provide more positive pseudo ground truth boxes.

Fig. 6(a) and 6(b) visualize the mined pseudo ground truth boxes (in red) of a few example images in the VOC trainval data and COCO-60-full after 2 refinements. From the results, we can see that some missing box-level annotations in VOC are successfully recovered, which helps the OCUD align with the target domain. The mined boxes in the COCO-60-full can also reduce the impact of the missing labels and improve the recall.

Size of the Source Data. We study the effect of varying the source dataset's size to explore the boundary of the amount of data necessary for a successful transfer. Specifically, we randomly sample 5%, 10%, ..., 80% of the COCO-60 as the source dataset. The smaller percentage subset is subsequently included in the larger percentage subset. Fig. 5 shows the experiment results. We can observe that as few as 20% of COCO-60 (4396 train + 194 val images) brings accuracy to more than 58% mAP on VOC.

COCO vs ILSVRC. We replace COCO-60 by ILSVRC-179 and run our algorithm for 4 refinements with the same hyper-parameters as in the COCO-60 experiment. The OCUD is trained with 4 times more gradient steps, because of the larger data size. The final accuracy is 56.46%, which is higher than COCO-60-full, but worse than COCO-60. Compared with COCO-60-full, the superiority might come from the larger dataset. Compared with COCO-60, we believe the inferiority is from the data quality and consistency with the target dataset. Although ILSVRC-179 contains more images than COCO-60, the quality is not



(a) (b)

Fig. 6. (a) Mined pseudo ground truth boxes (in red) in VOC trainval. (b) Original ground truth (in green) and pseudo ground truth boxes (in red) in COCO-60-full.

as good, and we observed more images with missing labels. Visual images are shown in the supplementary materials. This introduces more regions that are target domain objects but are taken as negative regions for OCUD.

4.4 ILSVRC Transfer Setting

Following the setting in [10,29,30,35], we also conduct experiments with the 200 classes in the ILSVRC 2013 detection dataset [4]. The setting uses the first 100 classes sorted in alphabetic order as the source classes, and the last 100 classes as the target weak classes. We were able to achieve 37.0% mAP on the weak 100 categories of val2 set with our algorithm and the ResNet50 backbone, which is comparable to the 36.9% mAP reported in [30] with the stronger Inception-ResNet and is much better than any earlier results under the same setting [10,29,35]. Note that our method without any refinement is 33.5%, and iterative knowledge transfer boosts the performance by 3.5 points after two refinements. This again confirms our argument that the multi-step transfer is more effective than one-step. The detail is provided in the supplementary material.

5 Conclusion

We have studied the weakly supervised object detection problem by transfer learning from a fully annotated source dataset. A simple yet effective progressive knowledge transfer algorithm is developed to learn a one-class universal detector and a MIL classifier iteratively. As such, the source dataset's knowledge can be thoroughly exploited and leveraged, leading to a new state-of-the-art on VOC 2007 with COCO-60 as the source dataset. The results suggest that knowledge transfer from an existing well-annotated dataset could be a fruitful future direction towards mitigating the annotation effort problem for novel domains.

References

1. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 384–400 (2018) 3
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2846–2854 (2016) 2, 3, 5, 8, 9, 10
3. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: A low-shot transfer detector for object detection. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 2
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009) 2, 14
5. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. International journal of computer vision **100**(3), 275–293 (2012) 2, 3, 8
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 2, 7
7. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015) 1, 9
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 5
9. 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) 1, 4
10. Hoffman, J., Guadarrama, S., Tzeng, E.S., Hu, R., Donahue, J., Girshick, R., Darrell, T., Saenko, K.: Lsda: Large scale detection through adaptation. In: Advances in Neural Information Processing Systems. pp. 3536–3544 (2014) 2, 3, 14
11. Hoffman, J., Pathak, D., Darrell, T., Saenko, K.: Detector discovery in the wild: Joint multiple instance and representation learning. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 2883–2891 (2015) 2
12. Hoffman, J., Pathak, D., Tzeng, E., Long, J., Guadarrama, S., Darrell, T., Saenko, K.: Large scale visual recognition through adaptation using joint representation and multiple instance learning. The Journal of Machine Learning Research **17**(1), 4954–4984 (2016) 2
13. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1377–1385 (2017) 6, 10
14. Kuen, J., Perazzi, F., Lin, Z., Zhang, J., Tan, Y.P.: Scaling object detection by transferring classification weights. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6044–6053 (2019) 2, 3
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018) 2
16. Lee, S., Kwak, S., Cho, M.: Universal bounding box regression and its applications. In: Asian Conference on Computer Vision. pp. 373–387. Springer (2018) 2, 3, 7, 8, 9, 10, 12

17. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3512–3520 (2016) 2, 3, 10
18. 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) 2, 8
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016) 3
20. Massa, F., Girshick, R.: maskrcnn-benchmark: Fast, modular reference implementation of instance segmentation and object detection algorithms in pytorch (2018) 8
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 1, 2, 3, 4, 10
22. Rochan, M., Wang, Y.: Weakly supervised localization of novel objects using appearance transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4315–4324 (2015) 3, 10
23. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8430–8439 (2019) 2
24. Shi, M., Caesar, H., Ferrari, V.: Weakly supervised object localization using things and stuff transfer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3381–3390 (2017) 2, 3
25. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. In: International Conference on Machine Learning. pp. 1611–1619 (2014) 2, 3
26. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.L.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence (2018) 2, 3, 8, 9, 10
27. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2843–2851 (2017) 2, 3, 6, 9, 10, 12
28. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 352–368 (2018) 3
29. Tang, Y., Wang, J., Wang, X., Gao, B., Dellandréa, E., Gaizauskas, R., Chen, L.: Visual and semantic knowledge transfer for large scale semi-supervised object detection. IEEE transactions on pattern analysis and machine intelligence **40**(12), 3045–3058 (2017) 2, 3, 14
30. Uijlings, J., Popov, S., Ferrari, V.: Revisiting knowledge transfer for training object class detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1101–1110 (2018) 2, 3, 5, 6, 11, 14
31. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013) 2, 3
32. Wu, Z., Bodla, N., Singh, B., Najibi, M., Chellappa, R., Davis, L.S.: Soft sampling for robust object detection. In: BMVC. p. 225. BMVA Press (2019) 2

33. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8292–8300 (2019) 2, 3, 6, 8, 9, 10
34. Zhang, D., Han, J., Zhao, L., Meng, D.: Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. International Journal of Computer Vision **127**(4), 363–380 (2019) 2
35. Zhang, J., Huang, K., Zhang, J., et al.: Mixed supervised object detection with robust objectness transfer. IEEE transactions on pattern analysis and machine intelligence **41**(3), 639–653 (2018) 2, 3, 7, 8, 9, 10, 12, 14
36. Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B.: W2f: A weakly-supervised to fully-supervised framework for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 928–936 (2018) 3
37. Zhong, Y., Wang, J., Peng, J., Zhang, L.: Anchor box optimization for object detection. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 1286–1294 (2020) 1
38. Zhu, Y., Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Soft proposal networks for weakly supervised object localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1841–1850 (2017) 3
39. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014) 3