Weak-shot Object Detection through Mutual Knowledge Transfer

Xuanyi Du[*], Weitao Wan[*][†], Chong Sun, Chen Li
WeChat, Tencent
{duxuanyi93, wanweitao1}@gmail.com, {waynecsun, chaselli}@tencent.com

Abstract

Weak-shot Object Detection methods exploit a fully-annotated source dataset to facilitate the detection performance on the target dataset which only contains image-level labels for novel categories. To bridge the gap between these two datasets, we aim to transfer the object knowledge between the source (S) and target (T) datasets in a bi-directional manner. We propose a novel Knowledge Transfer (KT) loss which simultaneously distills the knowledge of objectness and class entropy from a proposal generator trained on the S dataset to optimize a multiple instance learning module on the T dataset. By jointly optimizing the classification loss and the proposed KT loss, the multiple instance learning module effectively learns to classify object proposals into novel categories in the T dataset with the transferred knowledge from base categories in the S dataset. Noticing the predicted boxes on the T dataset can be regarded as an extension for the original annotations on the S dataset to refine the proposal generator in return, we further propose a novel Consistency Filtering (CF) method to reliably remove inaccurate pseudo labels by evaluating the stability of the multiple instance learning module upon noise injections. Via mutually transferring knowledge between the S and T datasets in an iterative manner, the detection performance on the target dataset is significantly improved. Extensive experiments on public benchmarks validate that the proposed method performs favourably against the state-of-the-art methods without increasing the model parameters or inference computational complexity.

1. Introduction

Recent rapid development of supervised object detection models [17, 20, 22, 23] largely relies on massive human-annotated bounding boxes and category labels. Since obtaining these annotations, especially the bounding boxes, are expensive and time-consuming on large-scale datasets, it motivates the researches of alternative algorithms with less annotation cost. Weakly Supervised Object Detection (WSOD) methods [1,8,12,13,24,26,34] only require image-level object category labels to train an object detector on a target dataset. Though the annotation cost is considerably reduced, a prominent performance gap exists between WSOD and full-supervised models.

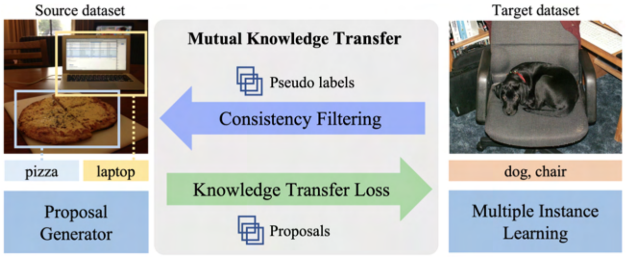

Figure 1. Overview of the proposed Mutual Knowledge Transfer scheme for the weak-shot object detection task.

While noticing class-invariant visual evidence can be transferred from base categories to unseen ones [14, 30], researches [3, 15, 18, 32, 36] show that the WSOD performance can be further improved by utilizing an additional source dataset with fully annotated data. This learning paradigm is referred to as the Weak-shot Object Detection (WSHOD) [21], for which a widely adopted model architecture is the combination of a proposal generator (PG) trained on the source (S) dataset and a multiple instance learning (MIL) module trained on the target (T) dataset. The S dataset contains both object category and bounding box annotations, while the T dataset has only image-level category labels and the object categories are not overlapped with those in the S dataset.

Although a well-trained PG on a full-annotated S dataset can assist the training of the MIL module on the T dataset, it is still essential to bridge the gap between these two datasets for handling non-overlapping categories. Previous efforts to address this issue mainly focus on transferring the knowledge about base categories from the S dataset to the T dataset by post-processing the predicted boxes [15] or designing various transferring scores [18, 32]. Zhong *et al*. [36] constrain the training of the MIL module by an ob-

[*] Equal contribution.
[†] Corresponding author.

jectness regularization loss. Unfortunately, this loss tends to exacerbate the classification ambiguity of novel categories since it enlarges multiple class probabilities for the same proposal (see details in Sec. 3.2). Some researches also adopt the predicted boxes on the T dataset as pseudo labels to refine the training of the PG module. The predicted boxes can be directly used as pseudo labels with confidence thresholding [36], adjusted by confidence maps [3], or softly weighted [21]. However, these practices are limited in discriminating inaccurate pseudo labels in the T dataset. For example, the intra-class feature variance can be significant, especially for the novel categories, which makes weighting these pseudo labels upon feature similarity [21] fail in discriminating inliers from outliers. Moreover, it is not exploited in previous works to incorporate the MIL module into discriminating inaccurate pseudo labels, which enables the knowledge transfer from the T dataset to S dataset.

To address the aforementioned issues in narrowing the gap between the S and T datasets, we design the Mutual Knowledge Transfer scheme for the WSHOD task, as illustrated in Fig. 1. Within this scheme, a novel Knowledge Transfer (KT) loss performs knowledge transfer from the S dataset to T dataset by constraining the training of the MIL module. In contrast to the regularization loss in [36], our KT loss enforces the predicted objectness score and class entropy of the MIL module to be consistent with the predictions of the PG, which helps to transfer the knowledge from S dataset to facilitate the training of the MIL module. Through mathematical analysis, we reveal that the formulation of KT loss intrinsically alleviates the class ambiguity issue of the regularization loss in [36].

Furthermore, we propose a novel and statistically robust Consistency Filtering (CF) method to improve the quality of the pseudo labels and boost knowledge transfer from the T dataset to S dataset. The intuition is that, by injecting noises into random regions in the feature maps of the predicted boxes[1], inaccurate boxes tend to be less stable in maintaining the original predictions than accurate ones. Inaccurate boxes usually only cover the most discriminative object fragment, which is a commonly addressed challenge in previous works [18,21], so the corresponding probability distribution of novel categories probably becomes uncertain when the designed noises are injected into the features of the MIL module. In contrast, accurate boxes usually contain the entire object and tend to be more stable against the injected feature noises. A detailed statistical verification for this intuition can be found in Tab. 7. We thus discover the inaccurate pseudo labels by evaluating the stability of the MIL outputs when varying noises are injected. The proposed CF method essentially takes advantage of the object knowledge of the MIL module regarding the discrimination of novel categories and transfers it to the PG module through refinement training with filtered pseudo labels.

By using the mutual knowledge transfer scheme iteratively, the detection performance on the T dataset with novel categories can be greatly improved. Through theoretical analysis and extensive experiments, we demonstrate that the proposed method significantly outperforms previous state-of-the-art WSHOD methods without increasing the model parameters or inference computational complexity.

2. Related works

2.1. Weakly supervised object detection

The WSOD task aims at training an object detector using only image-level category labels. Under the most commonly utilized multiple instance learning (MIL) structure, proposals are generated using unsupervised methods like Selective Search (SS) [33], Sliding Windows (SW), or Edge Boxes (EB) [38]. WSDDN [1] is a typical MIL-based model, which first proposes an end-to-end architecture to perform region selection and classification. Further improvements are made in OICR [29] and PCL [28] methods by adding instance refinement classifiers that facilitate the iterative refinement of candidate boxes. To help detectors focus on the entire objects rather than only the discriminative parts, ICMWSD [24] employs instance-aware self-training with bounding box regression. SDCN [16] takes advantage of the complementary collaboration of the weakly supervised detection and segmentation tasks. Besides, CASD [11] enforces consistent object detection across different transformations of the same images by computing comprehensive attention and conducting self-distillation on the WSOD networks. The aforementioned models frequently adopt the multi-stage strategies like self-training and self-distillation algorithms.

However, there remain some issues with the MIL-based methods. Conventional algorithms commonly used for generating proposals, including the SS, SW, and EB methods, are time-consuming. To tackle this problem, the class activation map (CAM) [37] method can be employed to generate proposals efficiently. However, CAM-based methods like TP-WSL [13], ACoL [35], and WCCN [5] can hardly localize multiple instances of the same class in an image [27], which restricts its generalization to real-world object detection tasks. To address these issues, researchers employ the Weak-shot Object Detection methods with the merits of both fast proposal generation and the ability to detect object instances of the same class.

2.2. Weak-shot object detection (WSHOD)

WSHOD methods take advantage of an existing fully-annotated source dataset and improve the detection perfor-

[1] "Feature maps of the predicted boxes" refers to the features produced by the *RoIAlign* layer given the predicted boxes.

mance on the target dataset. WSHOD first emerges from the method LSDA [10], and various methods [2, 15, 31, 32] follow. The MSD method [18] adversarially learns domain-invariant objectness to enable the MIL module to discriminate inaccurate proposals. Dong *et al.* [6] leverage the bounding box regression knowledge from a well-annotated auxiliary dataset to explore a series of learnable bounding box adjusters (LBBAs). Uijlings *et al.* [32] formulate various knowledge scores based on the hierarchy of categories and transfer the knowledge from the S to T dataset to improve the predictions of the MIL module, concluding that the objectness score is more favourable than the class-specific scores in knowledge transfer. Zhong *et al.* [36] improve the one-time knowledge transfer from the S dataset to T dataset by employing iterative refinement training for the PG and the MIL modules. Based on this work, Liu *et al.* [21] unify a mask generator with the object detection network to provide mask prior information and the SimNet which predicts semantic similarity to assign weights to pseudo labels. The SCM [3] method also aims to improve the pseudo labels by training an extra box regressor based on the score heatmaps of the original boxes with high confidence. However, the pseudo labels can be further improved by leveraging the MIL module which learns the knowledge of the novel categories on the T dataset. As such, we propose a pseudo label filtering method in this paper to reduce the inaccurate pseudo labels by exploiting the MIL module, thus transferring the knowledge from the T dataset to S dataset in return.

3. Method

Our network architecture consists of a proposal generator (PG) and a multiple instance learning (MIL) module, both of which are based on Faster-RCNN [23], as shown in Fig. 2. By employing the proposed Mutual Knowledge Transfer scheme, object knowledge is transferred between the PG and MIL modules to improve both.

3.1. Proposal generator and MIL modules

The detection head of the PG module contains a box regressor and an objectness predictor. By thresholding the objectness scores, the positive proposals are employed for training the MIL module. We utilize the S dataset to train the PG module, and all the object categories are regarded as a one-class foreground category.

The head of the MIL module consists of a classification branch and a detection branch. Each branch is composed of two fully connected layers. We train this module on the T dataset containing totally C categories in a weakly supervised manner. Given R proposals generated by the PG module and the feature maps produced by the backbone, R region feature maps are obtained through an *RoIAlign* layer. The region feature maps are fed into the classification and detection branches, obtaining matrices $M^c, M^d \in \mathbb{R}^{R\times C}$, respectively. The row- and column-wise softmax operations are performed for M^c and M^d respectively, which output two score matrices S^c and S^d with the (i, j) entry given by

$$S^c_{ij} = \frac{e^{M^c_{ij}}}{\sum_{n=1}^{C} e^{M^c_{in}}}, \quad S^d_{ij} = \frac{e^{M^d_{ij}}}{\sum_{m=1}^{R} e^{M^d_{mj}}}, \tag{1}$$

where S^c_{ij} is the probability of proposal i belonging to category j and S^d_{ij} denotes contribution of the i-th proposal to category j. Then we obtain the predicted score matrix S, whose (i, j) entry can be computed as

$$S_{ij} = S^d_{ij} S^c_{ij}. \tag{2}$$

We perform the column-wise sum operation on S, obtaining the image-level class probability for the j-th class as

$$\hat{y}_j = \sum_i^R S_{ij} = \sum_i^R S^d_{ij} S^c_{ij}. \tag{3}$$

A classification loss $\mathcal{L}_{cls}$ and the proposed Knowledge Transfer loss $\mathcal{L}_{kt}$ (see Sec. 3.2) are employed in training the MIL module. We exploit the binary cross entropy loss in Eq.(4), where $y_j \in \{0, 1\}$ is the image-level label.

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_j^C [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)], \tag{4}$$

During inference, the final class probability for each proposal is the element-wise product of S^c and $\dot{S}^d$, in which $\dot{S}^d = \mathrm{sigmoid}(M^d)$. Different from S^d, $\dot{S}^d$ denotes the confidence of a certain category existing in a proposal.

3.2. Knowledge Transfer loss

The lack of bounding box annotations is a crucial issue for the WSHOD task. We thus propose the Knowledge Transfer (KT) loss to transfer the knowledge from the PG module to constrain the training of the MIL module, through which the rich annotation information of bounding boxes on the S dataset is hopefully transferred to the MIL module, leading to potentially better detection performance on the T dataset.

Intuitively, the objectness score and the entropy of class distribution predicted by the MIL module should be consistent with those from the PG module trained with fully-annotated data. Although the categories in the S dataset and the T dataset have no overlap in our setting, this intuition holds as many low-level class-agnostic image features can be shared across object categories [18, 32].

It is worth noting that the PG module itself does not directly produce class distribution entropy for each proposal. Nonetheless, the objectness score conveys whether an object exists in a proposal, and the existence of an object

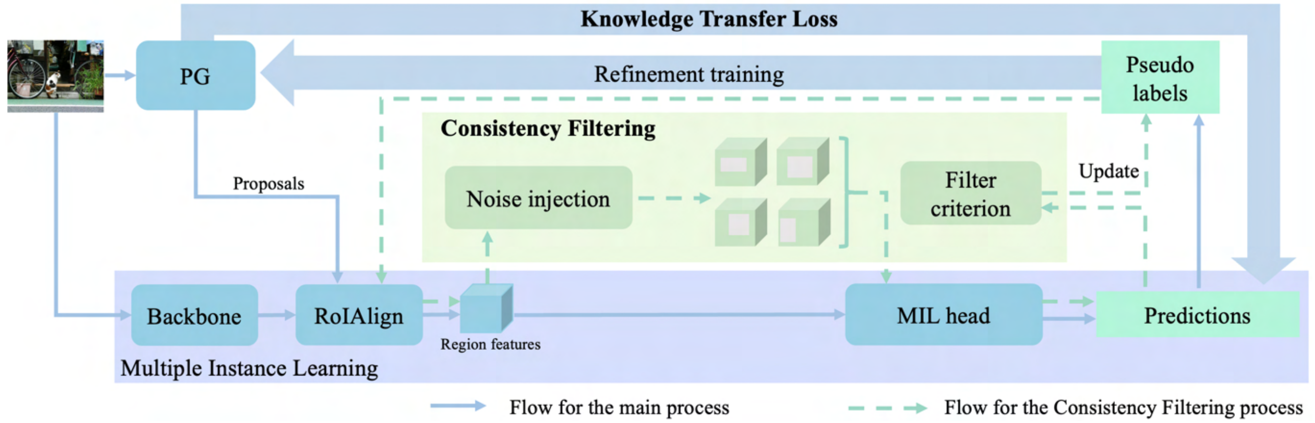

Figure 2. Overview of our model architecture. In the main process, the proposals generated by PG module are fed into the MIL module to produce detection predictions which are used to compute the Knowledge Transfer Loss $\mathcal{L}_{kt}$ with the predictions of the PG module. In the refinement training of the PG module, the Consistency Filtering method is conducted to update the pseudo labels. As such, the knowledge transfer is performed mutually between the source and target datasets.

causes large certainty of the class distribution, leading to small entropy. As such, the class distribution entropy of proposal i can be roughly estimated by its negative correlation with the objectness score $O_i \in (0, 1)$ predicted by the PG module. Therefore, the optimization target for the class distribution entropy of proposal i can be defined as

$$\mathcal{H}_i^t = (1 - O_i) \log C. \tag{5}$$

The class distribution entropy predicted by the MIL module for proposal i is defined by Eq. (6), where $\tilde{S}_i$ is the i-th row of $\tilde{S}^d$ which is obtained by performing row-wise *softmax* normalization on M^d in Eq. (7).

$$\mathcal{H}(\tilde{S}_i^d) = \sum_{j=1}^{C} (-\tilde{S}_{ij}^d) \log \tilde{S}_{ij}^d, \tag{6}$$

$$\tilde{S}_{ij}^d = \frac{e^{M_{ij}^d}}{\sum_{n=1}^{C} e^{M_{in}^d}}. \tag{7}$$

As such, the proposed Entropy Transfer loss $\mathcal{L}_{ent}$ is computed by

$$\mathcal{L}_{ent} = \frac{1}{R} \sum_{i=1}^{R} [\mathcal{H}(\tilde{S}_i^d) - (1 - O_i) \log C]^2. \tag{8}$$

The objectness transfer loss [36] is defined by $\mathcal{L}_{obj} = \frac{1}{R} \sum_{i=1}^{R} (\max_j \dot{S}_{ij}^d - O_i)^2$, where $\max_j \dot{S}_{ij}^d$ is the objectness score of the MIL module for the i-th proposal. Then the proposed KT loss $\mathcal{L}_{kt}$ is defined as

$$\begin{aligned} \mathcal{L}_{kt} &= \mathcal{L}_{ent} + \mathcal{L}_{obj} \\ &= \frac{1}{R} \sum_{i=1}^{R} \{[\mathcal{H}(\tilde{S}_i^d) - (1 - O_i) \log C]^2 + \\ &\qquad (\max_j \dot{S}_{ij}^d - O_i)^2\}. \end{aligned} \tag{9}$$

Finally, the overall loss $\mathcal{L}$ for training the MIL module is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{kt}, \tag{10}$$

where λ is a trade-off hyper-parameter.

Discussion. Prior to our work, Zhong [36] *et al.* also consider objectness transfer during model training but only exploit the objectness loss $\mathcal{L}_{obj}$. We find that using the objectness transfer loss $\mathcal{L}_{obj}$ alone (with the classification loss $\mathcal{L}_{cls}$) leads to unreasonable class distributions. For simplicity, we consider an image labeled with two categories g and h on the T dataset. Since no further constraints are enforced on the score matrix $\dot{S}^d$, a positive proposal i may simultaneously have peak responses for both categories, *i.e.*, large response values $\dot{S}_{ig}^d$ and $\dot{S}_{ih}^d$ in the score matrix. It can be theoretically proved that such score values of $\dot{S}^d$ will lead to an unreasonable score matrix S^c. With Eq.(2)-(4) and Eq.(9)-(10), we obtain the gradient of the overall loss with respect to S_{ig}^c as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial S_{ig}^c} &= \frac{\partial \mathcal{L}}{\partial \hat{y}_g} \frac{\partial \hat{y}_g}{\partial S_{ig}^c} = -\frac{1}{C \hat{y}_g} S_{ig}^d \\ &= -\frac{1}{C(S_{ig}^c + \frac{\sum_{r \neq i} S_{rg}^d S_{rg}^c}{S_{ig}^d})}. \end{aligned} \tag{11}$$

Note that the lower bound of its absolute value is S_{ig}^d / C, which reveals that the absolute value of the gradient with respect to S_{ig}^c is positively related to S_{ig}^d. As a result, large responses in S_{ig}^d and S_{ih}^d lead to high responses in both S_{ig}^c and S_{ih}^c through the back-propagation during training. This indicates that the proposal i is simultaneously recognized as categories g and h, which is unreasonable for the MIL module.

Fortunately, this phenomenon can be alleviated by using the proposed Knowledge Transfer loss $\mathcal{L}_{kt}$ instead. The optimization of $\mathcal{L}_{ent}$ leads to entropy reduction for positive proposals. As such, one element's spiking in $\tilde{S}^d$ leads to generally low values in other classes, ensuring a dominant contribution to only one class from each positive proposal in M^d, as is the case in S^d. Furthermore, according to the analysis of Eq.(11), we conclude that the optimization in our loss $\mathcal{L}$ can maintain unique peak response among all categories for each proposal in S^c. This learning paradigm is more reasonable in the knowledge transfer and MIL training process compared to using $\mathcal{L}_{obj}$ alone.

3.3. Consistency Filtering

With the trained PG and MIL modules, we can obtain the object detection results on both the source and target datasets for the novel categories, which is called the process of pseudo ground truth mining. The mined boxes are used as pseudo labels for the next iteration of refinement training, which is elaborated in Sec. 3.4. The motivation for the Consistency Filtering method is two-fold. First, the quality of pseudo labels can be further improved by the proposed approaches in addition to the naive way of abandoning the mined boxes whose classes are not contained in the image-level class labels. Second, object knowledge regarding the novel categories in the T dataset can boost the training of the PG module when transferred through label filtering.

It is a common challenge that the predicted boxes usually only cover the most discriminative part of an object instead of enclosing the entire object. The image features of the object fragment contain less semantic information and thus are susceptible to noise injections compared to those of the entire object. Motivated by this, we apply designed noises to random regions of feature maps corresponding to the predicted boxes. Specifically, the *RoIAlign* features (with spatial size 7 × 7) of the pseudo boxes are obtained, in which random regions of feature maps are replaced with noises. With the processed feature maps, the MIL module predicts classification probabilities and detection scores for different boxes. After repeatedly injecting random noises to N random regions, the evaluated pseudo box will be removed if all the N predictions meet the proposed filter criterion.

Three key factors of the CF method are the noise region selection, the noise formulation, and the filter criterion. First, we consider two kinds of noise regions, *i.e.*, *continuous* and *scattered* regions, respectively. The continuous noise region selection mechanism randomly locates a rectangle region on the feature maps, while the scatted noise region is selected by randomly sampling pixels on the feature maps. For the formulation of noises, we consider the truncated Gaussian noise post-processed by a *ReLU* activation layer. The *ReLU* layer is exploited to ensure consistent responses for noises and original features. In addition, we also consider using zero values as an alternative noise formulation. We propose two types of filter criterions for the CF method. The first one measures the mined box quality considering the predicted class probability and the detection-branch confidence of the MIL module. Two thresholds t_d and t_c are predefined. The boxes are abandoned if the max detection score and classification probability are lower than t_d and t_c, respectively. The second criterion considers label consistency for box selection. Boxes with inconsistent predicted class between the noisy and original input are regarded as inaccurate boxes. We refer to the CF method using these two filter criterions as CF-generative (denoted by CF-g) and CF-discriminative (denoted by CF-d), respectively. Ablation studies are provided in Sec. 4.4 to compare the performances of different variants of the CF method.

The CF method can effectively detect inaccurate pseudo labels. Visualizations and analysis of the removed boxes are illustrated in Fig. 3 and Sec. 4.5. Since false discarding of boxes harms the refinement training, the CF method concentrates more on its precision than recall.

3.4. Iterative training strategy

The whole network, including the PG and the MIL modules, is trained iteratively, following the training scheme in previous works [21, 36]. The initial PG and MIL modules are trained with the source and target datasets, respectively. Then candidate pseudo labels are generated on both datasets with the trained models. By applying the proposed CF method, we remove inaccurate predicted boxes and employ the remaining ones as pseudo labels. They are merged with the original annotations from the source dataset for the next refinement training of the PG module. As such, the refinement training of the PG and MIL modules can be conducted iteratively and progressively improves the detection performance on the target dataset.

4. Experiments

4.1. Datasets and evaluation metrics

For fair comparison, our experimental setting follows previous WSHOD works [21, 36]. *Pascal VOC 2007* [7] is adopted as the target dataset. Either *MS COCO 2017* [19] or *ILSVRC 2013* [25] detection dataset is adopted as the source dataset. *Pascal VOC 2007* consists of 20 categories and is split into a train-val set with 5,011 images and a test set with 4,952 images. The bounding box annotations are left unused during training. *MS COCO 2017* contains 80 categories in total which covers the 20 categories in *Pascal VOC 2007*. We remove all the images which contain the target categories. The remaining source dataset, referred to as *COCO-60*, consists of a training set with 21,987 images and a validation set with 921 images. *ILSVRC 2013* detection dataset originally has 200 categories. After removing

Table 1. mAP comparisons with state-of-the-art methods on VOC 2007 test set. '*Single scale*' denotes single-scale training and testing, and '+' means multi-scale. '*Distill*' indicates re-training a Faster-RCNN model based on the pseudo labels. 'Ens' indicates ensemble methods and 'FR' means distilling with Fast RCNN [9] model. The backbone is ResNet50 if not specified.

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 [1]	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 [29]	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 [28]	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
ICMWSD+FR [24]	66.4	69.1	58.9	32.5	27.6	71.5	73.1	66.2	32.8	75.4	47.4	53.7	63.3	71.7	34.8	28.5	57.4	54.7	62.5	67.1	55.7
CASD+FR [11]	66.6	81.3	58.4	33.5	31.6	75.7	55.2	68.3	36.8	59.5	61.0	52.9	65.4	72.0	29.1	29.4	65.7	54.2	74.5	70.7	57.1
WSHOD:																					
MSD [18]	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	51.1
OICR+UBBR [15]	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
Zhong et al. (single scale) [36]	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	57.8
Zhong et al.+ [36]	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	59.7
Zhong et al. (distill,vgg16)+ [36]	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	59.8
Zhong et al. (distill)+ [36]	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	60.2
TraMaS (single scale) [21]	65.6	53.7	67.4	47.2	46.9	76.3	76.6	81.7	33.0	76.9	29.3	80.9	76.8	66.2	61.1	12.6	65.8	58.9	74.4	66.7	60.9
TraMaS+ [21]	66.5	58.7	68.3	47.7	47.0	76.3	78.0	81.1	33.9	77.8	30.9	80.1	78.0	66.2	63.0	15.1	69.2	60.2	76.1	68.1	62.1
TraMaS (distill,vgg16)+ [21]	67.8	59.9	67.9	48.9	47.5	75.4	78.2	79.3	33.1	76.4	32.1	78.8	77.4	68.3	63.1	18.4	70.0	59.9	76.2	69.3	62.4
TraMaS (distill)+ [21]	68.6	61.1	69.6	48.1	49.9	76.3	77.8	80.9	34.9	77.0	31.1	80.9	78.5	66.3	64.0	19.1	69.1	62.3	74.4	69.1	62.9
Ours (single scale)	64.8	56.2	67.8	48.8	52.0	76.5	78.1	82.0	33.4	77.9	24.7	82.6	73.3	74.0	69.0	15.1	70.7	65.3	78.6	66.6	**62.9**
Ours+	68.5	57.6	68.5	47.3	50.9	79.2	78.4	81.8	34.7	77.5	23.1	81.8	74.3	73.0	69.6	15.9	70.8	62.3	78.2	69.1	**63.1**
Ours (distill,vgg16)+	64.9	64.6	69.4	44.9	48.3	72.0	81.4	80.9	38.7	74.5	26.4	79.3	75.3	74.2	72.1	20.2	65.5	62.3	76.4	69.6	**63.0**
Ours (distill)+	68.3	64.6	71.7	48.5	50.6	77.1	80.9	80.6	39.7	81.0	28.0	81.0	76.2	72.4	72.0	21.9	70.9	66.0	79.3	68.8	**65.0**

Table 2. CorLoc comparisons with state-of-the-art methods on VOC 2007 trainval set. See the definition of each notation in Tab. 1.

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 [1]	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 [29]	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 [28]	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
ICMWSD+FR [24]	86.2	55.8	78.8	44.7	15.9	68.8	81.8	62.2	32.2	78.3	26.3	54.7	58.0	76.9	28.6	32.9	76.1	36.5	77.2	59.6	56.6
CASD+FR [11]	83.4	79.7	75.1	46.9	42.7	76.5	72.5	53.6	75.4	46.2	37.7	32.0	44.9	86.7	27.5	46.2	74.3	70.8	79.4	65.1	60.8
WSHOD:																					
MSD-Ens [18]	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 [15]	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
Zhong et al. (single scale) [36]	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
Zhong et al.+ [36]	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
Zhong et al. (distill,vgg16)+ [36]	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
Zhong et al. (distill)+ [36]	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
TraMaS (single scale) [21]	88.9	66.5	87.3	69.2	70.6	86.2	90.3	90.6	49.5	95.5	31.6	93.7	93.5	87.4	73.6	24.9	93.5	67.3	89.6	82.7	76.6
TraMaS+ [21]	88.3	67.9	89.8	68.0	70.8	88.6	90.6	91.8	50.3	96.6	31.8	93.5	92.2	88.2	72.8	25.2	94.2	67.4	90.3	84.4	77.1
TraMaS (distill,vgg16)+ [21]	89.7	69.4	90.9	68.5	71.1	86.9	91.5	91.0	50.1	96.4	33.2	92.4	92.7	90.1	75.3	24.8	93.3	69.8	90.6	83.1	77.5
TraMaS (distill)+ [21]	90.6	67.4	89.7	70.5	72.8	86.6	91.7	89.8	51.0	96.1	34.0	93.7	94.8	90.3	73.0	26.5	95.2	68.2	89.8	83.1	77.7
Ours (single-scale)	88.3	75.1	87.3	77.4	76.3	90.1	93.9	88.9	54.5	97.9	30.0	90.8	94.2	91.5	81.6	34.4	95.9	69.8	92.4	82.1	**79.3**
Ours+	90.4	74.8	88.6	75.0	76.1	89.6	94.0	88.8	54.7	97.2	28.8	89.9	93.8	91.9	82.2	33.6	95.9	66.9	92.8	82.2	**79.4**
Ours (distill,vgg16)+	88.7	74.0	90.0	76.9	79.8	86.2	94.4	92.1	58.4	95.2	33.1	90.9	92.8	91.1	83.2	29.9	96.9	71.2	93.2	82.7	**80.0**
Ours (distill)+	91.7	79.2	89.7	76.3	76.0	88.6	94.1	89.4	59.2	96.6	37.1	91.1	94.8	93.1	83.8	37.3	93.8	71.6	90.5	82.3	**80.8**

the images which contain the categories in the *VOC* dataset, we obtain 143,095 train images and 6,229 validation images for 179 categories, which is denoted as the *ILSVRC-179* dataset. Two commonly used evaluation metrics are adopted, namely mean average precision (mAP) and correct localization (CorLoc) [4]. Average precision (AP) is the weighted mean of precision at each threshold and can be calculated as the area under the precision-recall curve. mAP is the average of AP for all categories. For a certain category, CorLoc is the percentage of correctly localized test images by the top-1 prediction of an algorithm.

4.2. Implementation details

We use 4 GPUs for training and set the batch size to 8. The architecture of the PG module is based on Faster-RCNN with a ResNet50 or VGG16 backbone which is initialized by ImageNet-pretrained weights. Following [21], the base learning rate is set to 8×10^{-3} and reduced to 8×10^{-4} after running 70% of the total training iterations. We train the PG module for 20,000 iterations and the MIL module for 5,000 iterations before the refinement training. Then the training iterations are reduced to 10,000 and 2,000, respectively. During pseudo label generation, the final score of the i-th bounding box is defined as $s_i^{final} = (\max_j S_{ij} + O_i)/2$. The predicted boxes with $s^{final} < 0.8$ are removed before applying the CF method. For the iterative training, we conduct 4 iterations of refinement training following [21,36]. For multi-scale setting, the scales of 240, 320, 480, 640, and 800 are chosen for both training and testing. The loss weight λ for the KT loss is set to 0.2.

In the Consistency Filtering process, the noise injection time N is set to 4. We use p to denote the area proportion of the noise region to the entire feature maps and r to denote the aspect ratio of the noise region. The value p and r are set as random values ranging in $(0.1, 0.33)$ and $(0.3, 3.3)$, respectively. For continuous noise regions, a rectangle region is selected based on p and r for noise injection. For scattered noise regions, the number of noisy pixels n_p is determined by the randomly selected p, and the positions of noisy pixels are randomly selected across the feature maps. The truncated Gaussian noises are created with a mean value of 0 and a variance of $\frac{M_f^2}{n_p}$, where M_f is the maximum value of the feature maps. For the filer criterion, t_d and t_c is set to 0.3 and 0.6, respectively. Ablation studies for t_d and t_c are presented in Sec. 4.4. The KT loss and the CF method introduce no extra model parameters. Besides, the proposed methods are conducted only at the training stage. As such, the inference computational complexity of the detection model is not increased.

4.3. Comparisons with SOTA

We compare the proposed approach with previous state-of-the-arts in this section. To present comprehensive comparisons with the competitive baselines in WSHOD, we also perform multi-scale testing or distillation training following experiments in [21, 36]. Table 1 and 2 show the mAP and CorLoc performance on VOC 2007 test set, respectively, with COCO-60 as the source dataset. In the WSHOD works, both [36] and [21] adopt the Faster-RCNN-based framework. Generally, the testing performance of the best WSHOD methods surpasses that of WSOD methods by leveraging knowledge in the source dataset. In terms of mAP, our method with single scale and no distillation setting gains 2.0% (62.9% vs. 60.9%) compared to the previous best method [21], which demonstrates the superiority of the proposed method. Besides, under the multi-scale setting, the mAP can be increased to 63.1%. When distillation is considered, our performance reaches 65.0%, which surpasses TraMaS by 2.1% [21]. In terms of the CorLoc metric, the performance in single-scale testing is 79.3%, obtaining a significant improvement of 2.7% compared to TraMaS. Considerable improvements can also be observed in multi-scale and distillation settings on both VGG16 and ResNet50 backbones.

Experiments on ILSVRC-179 dataset. Results of leveraging ILSVRC-179 as the source dataset are shown in Tab. 3. We use the reported results of state-of-the-art methods for comparison. In terms of mAP, our method gains 2.1% (60.4% vs. 58.3%) compared to [21], which indicates that our method is robust to different source datasets.

Table 3. Comparison with state-of-the-art methods when the source dataset is ILSVRC-179. The backbone is ResNet50 unless specified.

Methods	mAP	CorLoc
MSD (vgg16) [18]	47.5	65.3
Zhong et al. [36]	56.5	/
TraMaS (vgg16) [21]	57.8	74.1
TraMaS [21]	58.3	74.8
Ours (vgg16)	**58.9**	**75.7**
Ours	**60.4**	**77.5**

Table 4. Ablation studies on the proposed Knowledge Transfer loss and Consistency Filtering method. "+" and "-" means the method is used or unused, respectively.

KT loss	CF	mAP (%)
-	-	60.9
+	-	61.6
-	+	62.1
+	+	62.9

Table 5. Ablation studies on different kinds of filter criterion. The mAP (%) is evaluated for the final detection model, while the precision (Prec.) (%) is evaluated for the filtering operation, *i.e.*, the ratio of correctly removed boxes to all removed boxes.

metrics	CF-d	CF-g
mAP (%)	60.5	62.9
Prec. (box only) (%)	57.9	75.7
Prec. (class + box) (%)	74.4	77.9

Table 6. Ablation studies on the noise regions and noise formulations. The mAP (%) of the final detection evaluation on the T dataset is presented.

Noise formulation	scattered	continuous
zeros	61.6	62.9
truncated Gaussian	61.9	62.4

4.4. Ablation studies

Components of the proposed method. In this paper, we design a novel KT loss and propose the CF method to improve the performance of the WSHOD model. Table 4 demonstrates how the two proposed components contribute to the overall performance. Without our methods, a Faster-RCNN-based framework which uses the iterative training strategy and adopts the objectness transfer loss as in [36] obtains a mAP of 60.9%. The performance can be improved to 61.6% and 62.1% by solely adopting the KT loss or the

Table 7. Ablation studies on thresholds t_d and t_c for the CF method. "†" means directly filtering pseudo labels based on the two thresholds without injecting noises. "Prec." and "Prec.*" are the precisions of "box only" and "class + box", respectively. FPR is the ratio of falsely removed boxes to all correct boxes.

t_d t_c	0.4 0.7	0.4 0.6	0.5 0.5	0.3 0.7	0.3 0.6	0.3† 0.6†
mAP (%)	62.2	62.0	62.2	62.7	62.9	61.9
FPR (%)	0.9	0.7	0.8	0.5	0.4	/
Recall (%)	9.9	8.7	10.0	6.2	5.9	/
Prec. (%)	66.8	69.8	68.2	72.1	75.7	/
Prec.* (%)	70.6	72.7	72.0	75.5	77.9	/

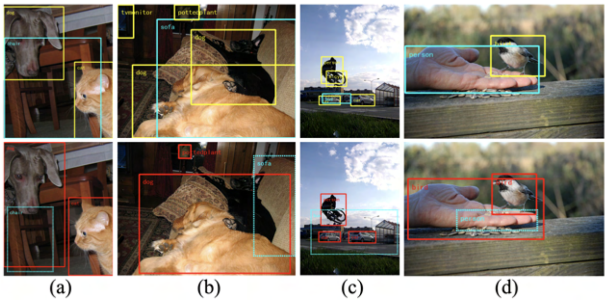

Figure 3. Visualizations of pseudo boxes removed by the Consistency Filtering method. The first row is the ground truth, in which the cyan boxes are the corresponding ground-truth for the categories of the removed boxes, and the yellow ones are ground-truth for other categories. The second row shows the removed boxes in cyan dashed boxes and the kept pseudo labels in red solid boxes.

CF method, respectively. By jointly using the KT loss and CF method, the mAP can be further increased to 62.9%. Both parts of our design have contributions to the performance gain and that they are complementary because the combination of both can obtain further improvement.

Variants of the CF method. We study the three key factors of the CF method, *i.e.*, the noise region selection, the noise formulation, and the filter criterion, using COCO-60 as the source dataset. We first compare the two filter criterions in terms of mAP of the ultimate detection results, which is shown in Tab. 5. In addition, to statistically analyze the effect of the filtering operation itself, we also compute the precision of the CF method, which is defined as the ratio of the correctly removed boxes to all the removed boxes. Regardless of the GT class, the removal is correct if the IoU of the removed box with the Ground-Truth on the T dataset is less than 0.7. Such a metric is referred to as "box only". We also compute the precision when the GT class is also considered (*i.e.*, the removal is correct if both the class and box predictions are wrong), which is denoted as "class + box" in Tab. 5. In terms of the "class + box" metric, CF-g is slightly better than CF-d (77.9% vs. 74.4%). However, CF-g surpasses CF-d by a large margin (75.7% vs. 57.9%) when the "box only" metric is used. As such, CF-g is chosen over CF-d as the method in comparison with SOTA. Besides, we demonstrate the effects of noise formulations and noise regions in Tab. 6. Generally, continuous noise regions are more favourable than scattered ones. As such, the zero noises on continuous regions are adopted as our best setting. Table 7 studies the thresholds t_d and t_h for CF-g method. The thresholds $t_d = 0.3$ and $t_c = 0.6$ are finally adopted. The result denoted with "†" is an important baseline which simply filters out all boxes by the filter criterion of CF-g method, without injecting any noises. A large performance gap is observed in the last two columns, which validates that random noise injection is crucial for the effectiveness of the CF method.

4.5. Visualizations

We visualize samples of the CF method of our best setting in Fig. 3. In most cases, the predicted boxes are correctly removed because they cover only part of an object, which validates our analysis. Occlusion is a common reason for the PG module to generate boxes covering object parts. As in Fig. 3(b), the dog lying on the sofa is a distraction for detecting the sofa. In addition, a number of predicted boxes only cover the most discriminative part of an object. In Fig. 3(d), fingers of the person are recognized while the other parts are ignored. Nonetheless, these inaccurate predictions can be identified by the CF method. An overly large predicted box is shown in Fig. 3(c). Features of the bus in the box can be severely influenced by the injected noises, leading to successful filtering of the inaccurate box.

5. Conclusions

We develop the Mutual Knowledge Transfer scheme for the Weak-shot Object Detection task. For mitigating the gap between the source and target datasets, we propose a novel Knowledge Transfer loss to constrain the training of the multiple instance learning module. Moreover, the statistically robust Consistency Filtering method is proposed to refine the proposal generator module with accurate pseudo bounding boxes annotations. We conduct mathematical analysis and statistical verification to demonstrate the advantages and effectiveness of the proposed Mutual Knowledge Transfer scheme. Extensive experiments demonstrate that the detection performance on the target dataset is significantly improved against the state-of-the-art WSHOD approaches without increasing the model parameters or inference computational complexity.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China under Grant No. 62106149.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 1, 2, 6

[2] Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, and Yan-Feng Wang. Cat: Weakly supervised object detection with category transfer. In *ICCV*, pages 3070–3079, 2021. 3

[3] Zitian Chen, Zhiqiang Shen, Jiahui Yu, and Erik Learned-Miller. Cross-supervised object detection. *arXiv preprint arXiv:2006.15056*, 2020. 1, 2, 3

[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012. 6

[5] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *CVPR*, pages 914–922, 2017. 2

[6] Bowen Dong, Zitong Huang, Yuelin Guo, Qilong Wang, Zhenxing Niu, and Wangmeng Zuo. Boosting weakly supervised object detection via learning bounding box adjusters. In *ICCV*, pages 2876–2885, 2021. 3

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[8] Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han. Weakly supervised rotation-invariant aerial object detection network. In *CVPR*, pages 14146–14155, 2022. 1

[9] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 6

[10] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. *NeurIPS*, 27, 2014. 3

[11] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *NeurIPS*, 33:16797–16807, 2020. 2, 6

[12] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 1

[13] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *ICCV*, pages 3534–3543, 2017. 1, 2

[14] Jason Kuen, Federico Perazzi, Zhe Lin, Jianming Zhang, and Yap-Peng Tan. Scaling object detection by transferring classification weights. In *ICCV*, pages 6044–6053, 2019. 1

[15] Seungkwan Lee, Suha Kwak, and Minsu Cho. Universal bounding box regression and its applications. In *ACCV*, pages 373–387. Springer, 2018. 1, 3, 6

[16] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *ICCV*, pages 9735–9744, 2019. 2

[17] Yali Li and Shengjin Wang. R (det) 2: Randomized decision routing for object detection. In *CVPR*, pages 4825–4834, 2022. 1

[18] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE TPAMI*, 41(3):639–653, 2018. 1, 2, 3, 6, 7

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 1

[21] Yan Liu, Zhijie Zhang, Li Niu, Junjie Chen, and Liqing Zhang. Mixed supervised object detection by transferring mask prior and semantic similarity. *NeurIPS*, 34:3978–3990, 2021. 1, 2, 3, 5, 6, 7

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 3

[24] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, pages 10598–10607, 2020. 1, 2, 6

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[26] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *ECCV*, pages 312–329. Springer, 2022. 1

[27] Feifei Shao, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao. Deep learning for weakly-supervised object detection and localization: A survey. *Neurocomputing*, 2022. 2

[28] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, 42(1):176–191, 2018. 2, 6

[29] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, pages 2843–2851, 2017. 2, 6

[30] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *CVPR*, pages 2119–2128, 2016. 1

[31] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE TPAMI*, 40(12):3045–3058, 2017. 3

[32] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *CVPR*, pages 1101–1110, 2018. 1, 3

[33] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 2

[34] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE TPAMI*, 44(9):5866–5885, 2021. 1

[35] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018. 2

[36] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *ECCV*, pages 615–631. Springer, 2020. 1, 2, 3, 4, 5, 6, 7

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2

[38] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014. 2