Mixed Supervised Object Detection with Robust Objectness Transfer

Li, Yan; Zhang, Junge ; Zhang, Jianguo; Huang, Kaiqi

Published in:
IEEE Transactions on Pattern Analysis and Machine Intelligence

DOI:
10.1109/TPAMI.2018.2810288

Publication date:
2018

Citation for published version (APA):
Li, Y., Zhang, J., Zhang, J., & Huang, K. (2018). Mixed Supervised Object Detection with Robust Objectness Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 639-953. https://doi.org/10.1109/TPAMI.2018.2810288

Mixed Supervised Object Detection with Robust Objectness Transfer

Li, Yan; Zhang, Junge ; Zhang, Jianguo; Huang, Kaiqi

Published in:
IEEE Transactions on Pattern Analysis and Machine Intelligence

Publication date:
2018

Document Version
Peer reviewed version

Citation for published version (APA):
Li, Y., Zhang, J., Zhang, J., & Huang, K. (2018). Mixed Supervised Object Detection with Robust Objectness Transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence. 10.1109/TPAMI.2018.2810288

Mixed Supervised Object Detection with Robust Objectness Transfer

Yan Li, *Student Member, IEEE,* Junge Zhang, *Member, IEEE,* Jianguo Zhang, *Senior Member, IEEE*
Kaiqi Huang*, *Senior Member, IEEE,*

Abstract—In this paper, we consider the problem of leveraging existing fully labeled categories to improve the weakly supervised detection (WSD) of new object categories, which we refer to as mixed supervised detection (MSD). Different from previous MSD methods that directly transfer the pre-trained object detectors from existing categories to new categories, we propose a more reasonable and robust objectness transfer approach for MSD. In our framework, we first learn domain-invariant objectness knowledge from the existing fully labeled categories. The knowledge is modeled based on invariant features that are robust to the distribution discrepancy between the existing categories and new categories; therefore the resulting knowledge would generalize well to new categories and could assist detection models to reject distractors (*e.g.*, object parts) in weakly labeled images of new categories. Under the guidance of learned objectness knowledge, we utilize multiple instance learning (MIL) to model the concepts of both objects and distractors and to further improve the ability of rejecting distractors in weakly labeled images. Our robust objectness transfer approach outperforms the existing MSD methods, and achieves state-of-the-art results on the challenging ILSVRC2013 detection dataset and the PASCAL VOC datasets.

Index Terms—Weakly supervised detection, Mixed supervised detection, Robust objectness transfer.

1 Introduction

Recently, object detection has been improved drastically in performance and scale with the development of convolutional neural networks (CNNs) [1], [2], [3], [4], [5] and the introduction of benchmarking detection datasets (*e.g.*, PASCAL VOC [6], MS COCO [7] and ILSVRC2013 detection dataset [8]). The supervised training process of state-of-the-art object detectors requires a large number of fully labeled images with bounding box annotations. However, the bounding box annotations are very difficult to acquire; thus supervised training of a high-performance detector at such a scale is not feasible, when we need to deal with hundreds of thousands of new objects in real-world applications.

On the other hand, weakly labeled data (*i.e.*, object categories with image labels only) are relatively easier to obtain either manually or through search engine. The problem of learning a detector on images with such weak labels is called *weakly* supervised detection (WSD). It is often formulated as a multiple instance learning (MIL) problem [9], in which each image is used as a bag of regions. However, unfortunately, the clear definition of *instances* for object regions is missing in such formulation. As illustrated in Fig. 1, WSD tends to confuse the objects with co-occurring distractors (context or object part) [10], [11], [12], [13], [14], [15], [16]. Those co-occurring distractors are in principle perfectly *valid* positive instances considering the image class labels only, though not the actual object being sought, for example *boat vs. boat with water* and "*whole cat*" *vs.* "*cat face*". The key issue in weakly supervised setting is that WSD trains an object detector by optimizing objective functions for an image-level classification instead of the region-level detection. Considering a typical case when both "*whole cat*" region and "*cat face*" region present in many positive *cat* images but do not appear in negative images, WSD cannot distinguish the two regions and is also likely to select "*cat face*" region as the desired *cat* region, as either of them could distinguish *cat* from other categories. Such an incapability

Fig. 1. Failure examples of weakly supervised detection†. The weakly supervised detection tends to confuse the objects (*cat, boat*) with the co-occurring distractors (*cat face, water*).

• *This work is funded by the National Key Research and Development Program of China (Grant 2016YFB1001004 and Grant 2016YFB1001005), the National Natural Science Foundation of China (Grant 61673375, Grant 61721004 and Grant 61403383) and the Projects of Chinese Academy of Sciences (Grant QYZDB-SSW-JSC006 and Grant 173211KYSB20160008).*
• *Y. Li, J. Zhang and K. Huang are with the Center for Research on Intelligent Perception and Computing (CRIPAC) and National Laboratory of Pattern Recognition (NLPR) of Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the University of Chinese Academy of Sciences (UCAS), Beijing, 100049, China. E-mail: yan.li@cripac.ia.ac.cn; jgzhang@nlpr.ia.ac.cn; kqhuang@nlpr.ia.ac.cn (*Correspondence author).*
• *K. Huang is also with the CAS Center for Excellence in Brain Science and Intelligence Technology.*
• *Jianguo Zhang is with Computing, School of Science and Engineering, University of Dundee, UK. E-mail: j.n.zhang@dundee.ac.uk.*

†. The results are obtained using our implementation of WSDDN [10].

of distinguishing the objects from distractors leads to lots of false detections and limits the detection performance. This pitfall is inherent in current weakly supervised setting.

However we argue that the pitfall could be better addressed by building detectors over a mixed set of images with strong labels (*i.e.*, bounding box annotations) and weak labels (*i.e.*, image-level labels); we call such a problem *mixed* supervised detection (MSD). The mixed supervised learning offers two key advantages: 1) limiting the amount of annotations due to the use of weak labels; 2) leveraging fully labeled public datasets to assist training on weak labels. For simplicity and clarity, in the following we term the object categories with bounding box annotations as *strong* categories, while the categories with image-level annotations only are called *weak* categories. We would like to highlight that, different from *semi*-supervised detection [17], [18], [19], strong categories in mixed supervised detection have **NO** overlap with weak categories, *i.e.*, objects in the weak categories are novel categories w.r.t. the strong categories. The trained detector is expected to detect an object instance from one of those novel categories in an unseen image. Thus MSD is a more challenging problem. The existing MSD methods [20], [21], [22], [23] utilize a straightforward pipeline that directly transfers the object detectors learned on strong categories to weak categories following some hand-crafted strategies. We argue that a better approach to solve MSD should be capable of 1) learning strong domain-invariant knowledge from strong categories and 2) robustly transferring the learned knowledge to weak categories.

In this paper, we propose a robust objectness transfer approach for mixed supervised learning. In our approach, we first aim to learn domain-invariant objectness knowledge from strong categories with CNN models. The objectness knowledge learned from annotated boxes of strong categories could facilitate rejecting the distractors in weakly labeled images. Meanwhile, the unlabeled regions from weak categories are utilized to make the learned objectness generalize well to new categories. Concretely, we cast the learning of objectness as a domain adaptation problem, considering the strong categories as our source domain, and the weak categories as our target domain. The objectness model is trained with the embedding representations where both strong and weak categories are indistinguishable. Thus the learned objectness will be invariant to the change of domains (from strong to weak categories). The domain-invariant objectness has two important characteristics: 1) category independent, to generalize well to unseen categories and 2) object sensitive, to reliably reject distractors in weak categories. We believe that such knowledge is appropriate for mixed supervised detection scenarios. After that, the objectness knowledge is applied to separate the objects and distractors in weak categories and a simple way is to use the separated objects as pseudo ground truths to train object detectors on weak categories. However, we believe that the difference between objects and distractors in weak categories can be better modeled by a further learning process. Specifically, we consider the separated objects and distractors as "object bag" and "distractor bag" and aim to model the concepts of objects and distractors under a standard multiple instance learning (MIL) framework. Finally, with the improved object and distractor concepts, the detection model is capable of distinguishing the ground truth objects from distractors, and leads to much better performance.

In summary, our contributions are three-fold:

- A robust objectness transfer approach is proposed for MSD. Different from previous MSD methods that directly transfer pre-trained object detectors from strong to weak categories with hand-crafted strategies, our method automatically *learns* the domain-invariant knowledge by incorporating weak categories into the knowledge learning process.
- We design a MIL-based framework to further model the difference between objects and distractors and to improve the ability of reliably rejecting distractors in weakly labeled images.
- The proposed method outperforms both the state-of-the-art MSD methods and the baselines on the challenging ILSVRC2013 detection and PASCAL VOC 2007, 2010, 2012 datasets.

2 Related Work

2.1 Weakly Supervised Detection

To reduce the annotation cost in object detection, weakly supervised detection (WSD) methods [10], [11], [12], [13], [14], [15], [16], [24], [25], [26], [27] attempt to learn object detectors using only image category labels. In weakly supervised setting, the optimization of WSD methods is an image-level classification instead of the required region-level detection, thus the WSD methods tend to select distractors (local optima) and their performance strongly depends on the initialization. Song *et al*. [16] and Wang *et al*. [27] use clustering method to obtain better initializations. Cinbis *et al*. [13], [14] propose a multi-fold training strategy of MIL to avoid the local optima: the dataset is split into 10 subsets. When selecting high-score proposals from a subset, the detectors trained on other subsets are used. Bilen *et al*. [11] propose a smoothed version of MIL where soft labels are related to the region proposals instead of choosing the ones with highest confidence. WSDDN [10] utilizes a two-stream architecture to train the recognition model and to select the discriminative regions in parallel to avoid using the recognition model itself to select high confident regions, which is able to relieve the local optima phenomenon. Based on WSDDN, Kantorov *et al*. [25] propose to utilize the context information to reject distractors and obtain more reliable detections. While these approaches are promising, the local optima problem has not yet been solved. The performance of WSD methods is still far from acceptable.

2.2 Mixed Supervised Detection

To learn well-performing object detectors with image category labels, several methods aim to utilize fully labeled data of different categories (strong categories) to improve the detection performance on the weakly labeled categories (weak categories), which is referred to as mixed supervised detection (MSD). Shi *et al*. [28] propose to learn a rank model on strong categories based on the appearance similarity. Then the rank model is transferred to weak categories to select the top-ranked regions as objects. Guillaumin and

Ferrari [29] conduct the MSD on ImageNet [8]. By exploiting the semantic hierarchy of ImageNet, the key idea in [29] is to localize objects of a weak category by transferring knowledge from its *ancestor* and *sibling* strong categories. Hoffman *et al*. [20] propose a Large Scale Object Detection through Adaptation (LSDA) algorithm to address the MSD problem. In their method, the classifier and detector differences are learned on strong categories and then transferred from several "similar" strong categories to a weak category. The weak category applies the transferred differences to adjust its classifier to corresponding object detector. In [21], Hoffman *et al.* utilize the same strategy to adapt the intermediate representations from strong to weak categories and then solve a standard MIL problem on weak categories based on transferred representations. Recently, Tang *et al*. [23] propose a Large Scale Semi-supervised Object Detection (LSSOD) method to improve the LSDA. LSSOD follows the same approach in LSDA but selects the "similar" categories by considering more informed visual and semantic similarities. In LSDA-based methods, both classifiers and object detectors are trained with 8-layer AlexNet model [30], and the parameters of layers 1-7 are the same for all categories in the models. Thus LSDA has to assume as a prior that the differences learned on layers 1-7 are category-invariant (note they are not *learned* to achieve category-invariance as what we do in this paper), those differences learned on strong categories are directly applied to weak categories. However, when the distribution discrepancy between strong and weak categories becomes significant, this assumption is no longer valid and the detection performance will be weakened greatly. Rocha *et al*. [22] propose a method called Weakly Supervised Localization Using Appearance Transfer (WSLAT) to solve the MSD based on semantic knowledge. In WSLAT, the strong and weak categories are represented as fixed length vectors (called "word embedding") [31]. Then the object detectors can be transferred from one category to another based on their semantic relationship. However, the semantic information is still an indirect measurement of complex object categories. The transferred detectors cannot obtain good performance in their method. In contrast, our model learns more robust and transferable objectness to support the learning of WSD on weak categories, which is able to effectively relieve the impacts caused by distribution discrepancy between strong and weak categories. Recently, Shi *et al*. [32] propose a new mixed supervised learning setting, where the auxiliary fully labeled annotations correspond to the pixel-level segmentation annotations and the knowledge is learned from the segmentation models.

2.3 Objectness Knowledge

Learning objectness knowledge to improve the detection performance has been explored by many previous works [33], [34], [35], [36]. Most of them measure objectness with low-level cues such as saliency [33], contour information [36] and hierarchical superpixels [35]. These approaches are "class-agnostic" and can be applied to new categories. However, the ability of these objectness knowledge is limited due to their incapabilities of capturing high-level cues. In fact, some high-level cues are also beneficial to detecting objects. For example, many animal object categories might share the same high-level structures (limbs around the body). Detecting such structures could infer to the presence of objects, but it is difficult to learn such high-level structures with low-level information only. Based on this consideration, learning objectness from annotated images with deep CNN models is likely to perform better, and such CNN-based objectness knowledge has already been explored by DeepBox [34]. With a large number of annotated images, DeepBox aims to let CNN model itself figure out what low-, mid- and high-level object cues are most discriminative and it achieves promising results in fully supervised detection scenario. However, such CNN-based approach is not appropriate for our MSD task. Since the objectness in DeepBox is learned on strong categories and directly applied to different categories without considering the distribution discrepancy problem. It would decrease the performance of DeepBox on new categories, especially when the available strong categories is limited. Different from previous objectness methods, our domain-invariant deep objectness model is the first work that incorporates the domain invariance into the CNN-based objectness method, which makes it possible to obtain sufficient objectness from existing annotated categories and to transfer the knowledge to new categories well.

3 Task Definition

In the mixed supervised learning case, we assume that we already have a set of fully labeled categories, which is called "strong categories" and denoted as $\mathcal{S}$. Meanwhile we have some weakly labeled categories called "weak categories", denoted as $\mathcal{W}$. Both bounding box annotations and image-level labels are available for set $\mathcal{S}$; for set $\mathcal{W}$, we only have access to their image category labels. In our detection scenario, the strong categories and weak categories have **no** overlap. This is quite different from the "mixed supervised learning" explored in Cinis's work [14], which actually belongs to *semi*-supervised learning where typically a small amount of fully labeled data with a large amount of weakly labeled data are provided for the same category.

4 Method

Our robust objectness transfer framework is illustrated in Fig. 2. We first learn domain-invariant objectness knowledge to assist the weakly supervised learning on weak categories. During the learning of objectness, the annotated boxes from strong categories are used to train the objectness predictor. Meanwhile, the unlabeled boxes from weak categories are also applied to learn a domain classifier, and the gradients from the domain classifier are reversed to achieve the domain invariance. The learned objectness is firstly utilized to roughly distinguish the objects and distractors and then a MIL-based approach is used to further model the difference between the objects and distractors. Finally, the detection model is able to recognize the *distractor* category in addition to object categories and learn that these confused distractors are false detections.

4.1 Learning Domain-invariant Objectness

In our approach, we aim to model the objectness knowledge using CNN-based method, and the objectness model is directly trained on the bottom-up proposals that are generated

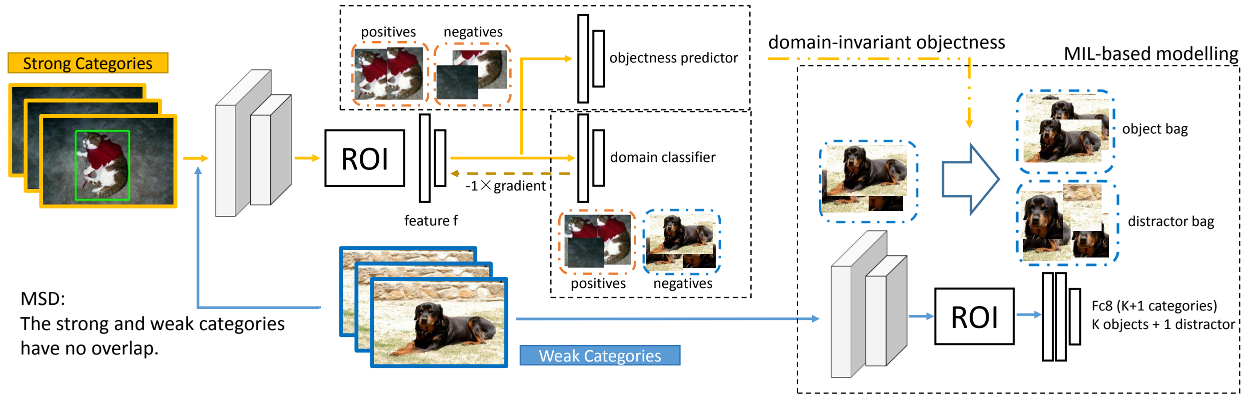

Fig. 2. The proposed robust objectness transfer approach for MSD. During the learning of objectness, the annotated boxes from fully labeled categories ("strong" categories, *e.g. cat*) are used to train the objectness predictor; meanwhile, the unlabeled regions from weakly labeled categories ("weak" categories, *e.g. dog*) are also applied to learn a domain classifier. During training, the gradients from the domain classifier are reversed to make the feature f invariant to the change of categories. The learned objectness is firstly utilized to roughly distinguish the objects and distractors and then a MIL-based approach is used to further model the difference between the objects and distractors.

by selective search [35]. By leveraging the bounding box annotations in strong categories, we cast the learning of objectness as a binary classification task: considering the regions that largely overlap with the ground truth boxes as the "objects" and the regions with smaller overlaps as "non-objects". We aim to let CNN models to automatically figure out beneficial cues for learning "objectness". During training, the images as well as a set of region proposals are fed to several convolutional layers, the RoI pooling layer [2] and fully-connected (fc) layers. Each region $\mathbf{r}_i$ is finally mapped to a 256-dimensional vector $\mathbf{f}_i \in \mathbb{R}^{256}$ (*i.e.*, **feature f** in Fig. 2), which can be considered as the internal representations for input regions. Then the $\mathbf{f}_i$ is connected to two branches.

The first branch is the objectness predictor, denoted as G_{obj}. G_{obj} consists of fc layers and predicts from $\mathbf{f}_i$ whether a region $\mathbf{r}_i$ is an object or not. As bounding box annotations are required to separate object regions and non-object regions in an image, only regions from set $\mathcal{S}$ are used to train the objectness predictor. The binary logistic loss can be used:

$$L_{obj}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i^{obj}\log(p_i) + (1-y_i^{obj})\log(1-p_i)] \quad (1)$$

where $p_i = \frac{1}{1+\exp(-G_{obj}(\mathbf{f_i}))}$ is the posterior probability that a region $\mathbf{r}_i$ belongs to "objects". The regions whose intersection-over-union (IoU) with any ground truth object is no less than 0.5 are considered as positive examples, *i.e.*, $y_i^{obj} = 1$; the regions that have a maximum IoU with ground truth in the interval $[0.1, 0.5)$ are negative examples, *i.e.*, $y_i^{obj} = 0$. Additionally, we balance the ratio of positive and negative samples in each image to 1:3 as the number of negative examples is far more than the positive ones.

Such objectness is trained on set $\mathcal{S}$ and cannot be extended to set $\mathcal{W}$ well, since the statistical distributions of categories in the two sets are different. Thus, we need to make the objectness learned on set $\mathcal{S}$ generalize well on set $\mathcal{W}$. Inspired by [37], [38], in our approach, we cast the learning of objectness as a domain adaptation problem where set $\mathcal{S}$ and $\mathcal{W}$ can be considered as source and target domain respectively, and the domain invariance is achieved by connecting a second branch to **feature f**, which we call domain classifier branch, as shown in Fig. 2 and denoted as G_{dom}.

Different from the objectness predictor, the domain classifier receives regions from both set $\mathcal{S}$ and set $\mathcal{W}$ to predict the origin of the input regions ($\mathcal{S}$ or $\mathcal{W}$). It is also a binary classification task and the optimization objective used is similar to the objectness predictor G_{obj}:

$$L_{dom}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i^{dom}\log(p_i) + (1-y_i^{dom})\log(1-p_i)] \quad (2)$$

where $p_i = \frac{1}{1+\exp(-G_{dom}(\mathbf{f_i}))}$ is the probability that a region $\mathbf{r}_i$ belongs to set $\mathcal{S}$. In this domain classification task, the regions sampled from set $\mathcal{S}$ are positives, *i.e.*, $y_i^{dom} = 1$; the regions sampled from set $\mathcal{W}$ are negatives, *i.e.*, $y_i^{dom} = 0$.

During the forward propagation, the domain classifier proceeds standardly and calculates L_{dom}. While in the backward propagation process, the gradients from the domain classifier are reversed (multiplied by -1) before passed to $\mathbf{f}$. With this *gradient reversal* operator, the network actually **maximizes** L_{dom} during the training process, which results in the **incapability** of modeling the discriminative information between two domains, *i.e.*, makes the internal representations $\mathbf{f}$ as indistinguishable as possible (**domain-invariant**) for both domains. In our objectness model, domain invariance is achieved by learning objectness $G_{obj}(\mathbf{f})$ with such domain-invariant representations $\mathbf{f}$, which thus could be well transferred to target domains with unseen objects. ‡

At each training iteration, for objectness predictor, 64 annotated regions are sampled from set $\mathcal{S}$ to learn the objectness and the ratio of positives and negatives are balanced to 1:3. Meanwhile, for domain classifier, 64 unlabeled regions are randomly sampled from set $\mathcal{W}$. It should be

‡. We refer the readers to the theoretical proof of the gradient reversal strategy in [38].

noticed that directly utilizing 64 "balanced" regions of set $\mathcal{S}$ and 64 "random" regions from set $\mathcal{W}$ to achieve domain invariance is inappropriate, since the data distribution of set $\mathcal{S}$ has already been changed with the balancing process. To address this issue, we randomly sample **another** 64 regions from set $\mathcal{S}$ to train the domain classifier together with the regions of set $\mathcal{W}$. Finally, each training mini-batch contains 192 samples, including 64 "balanced" regions sampled from set $\mathcal{S}$, another 64 "random" regions sampled from set $\mathcal{S}$ and 64 "random" regions sampled from set $\mathcal{W}$. Note that the image-level labels are not used in this process. The detailed training strategy is described in Step 1 of Algorithm 1.

Algorithm 1 Robust Objectness Transfer Approach for MSD

Input: strong set $\mathcal{S}$ and its region proposals $\mathcal{R}_s$
weak set $\mathcal{W}$ and its region proposals $\mathcal{R}_w$
Objectness model Obj with parameters $\mathcal{W}_o$
Objectness-aware detection model Det with parameters $\mathcal{W}_d$
max iteration *maxiter*$_{obj}$ and *maxiter*$_{det}$

Step 1: Domain-invariant Objectness Learning
Initialize $\mathcal{W}_o$
for *iter* $= 1$; *iter* $<$ *maxiter*$_{obj}$; ++*iter* **do**
  Sample 64 "balanced" regions from $\mathcal{R}_s$
  Calculate objectness prediction loss using (1)
  Propagate the *gradients*
  Sample 64 "random" regions from $\mathcal{R}_s$
  Sample 64 "random" regions from $\mathcal{R}_w$
  Calculate domain classification loss using (2)
  Propagate the ***reversed*** *gradients*
  Update $\mathcal{W}_o$
end for
Step 2: Objectness-aware Detection
Initialize $\mathcal{W}_d$
for *iter* $= 1$; *iter* $<$ *maxiter*$_{det}$; ++*iter* **do**
  Sample one image $\mathbf{x}$ from $\mathcal{W}$
  Scoring its region proposals $\mathbf{R}$ with Obj
  Object bag $\mathbf{B}_{obj} \leftarrow$ the top 15% regions of $\mathbf{R}$
  Distractor bag $\mathbf{B}_{dis} \leftarrow$ the last 85% regions of $\mathbf{R}$
  Calculating the Loss using (6)
  Update $\mathcal{W}_d$
end for

Output: Detection model's parameter $\mathcal{W}_d$

4.2 Objectness-aware Detection Model

After the objectness knowledge is learned, it is used to separate the objects and distractors in weakly labeled images of set $\mathcal{W}$. In our approach, for each weakly labeled image, its region proposals are firstly fed to the learned objectness model to get their "objectness" scores (*i.e.*, the outputs of objectness predictor in Fig. 2), and then sorted according to the scores. The top m% (m is set to 15 in our algorithm) proposals are selected as the "object regions", and the rest 1-m% ones are used as "distractor regions". A simple way to utilize the selected object regions is to consider these regions as pseudo ground truth and then train fully supervised detectors. But it is not a well-performing approach since such separation between objects and distractors is not prominent. To address this issue, we aim to further model the difference between objects and distractors based on a multiple instance learning (MIL) approach and propose the objectness-aware detection model.

In MIL framework, we first construct the "object bag" with "object regions" and the "distractor bag" with "distractor regions" for each weakly labeled image in set $\mathcal{W}$. The labels for these bags are denoted as $y \in \{-1, 1\}^{K+1}$. The "distractor bags" are tagged with $y_0 = 1$ while the "object bags" are labeled as their corresponding object categories ($y_k = 1, k > 0$).

Then we adopt the Fast R-CNN framework for the objectness-aware detection model. During training, a weakly labeled image $\mathbf{x}$ as well as its region proposals R that are generated by selective search are imported as the input of the network (each image $\mathbf{x}$ contains two bags: the object bag and the distractor bag). The network simultaneously computes features for each proposal and finally maps the features to K+1-dimensional vectors $s^R \in \mathbb{R}^{(K+1)\times|R|}$, which represent the classification scores for regions. These region-level scores are directly used to evaluate the detection performance at testing time.

During training, regions in the bag cannot be labeled since we do not have bounding box annotations. Thus the region-level scores s^R need to be aggregated to a bag-level classification score s^B to train the model. In traditional MIL settings, the highest region-level score is selected as the bag-level score:

$$s^B = \max_r (s_r^R). \tag{3}$$

This max operator utilizes only one region per bag as the positive sample. To relax this restriction, we use "*exp-sum-log*" operator proposed in [11] to serve as a soft approximation for the max operator:

$$s^B = \log\left(\sum_r^{|R|} \exp\left(s_r^R\right)\right). \tag{4}$$

After obtaining the bag-level scores s^B, we utilize the sigmoid function to compute the posterior probability that each bag $\mathbf{B_i}$ belongs to the k-th class:

$$p_{ki} = \frac{1}{1 + \exp(-s_{ki}^B)}. \tag{5}$$

Finally, the network can be trained end-to-end using cross-entropy loss:

$$L(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K+1}\left(\mathbf{1}\{y_{ki} = 1\}\log\left(p_{ki}\right)\right) \tag{6}$$

where $y_i \in \{-1, 1\}^{K+1}$ is the bag-level labels, and $\mathbf{1}\{\cdot\}$ is the indicator function. λ is the weight decay parameter on the weight $\mathbf{w}$ of CNNs used to improve the generalization of the model and is set to 0.0005 in all experiments. Using this MIL-based approach, the objectness-aware detection model is able to model the concepts of both distractors and K object categories. Finally, with the learned *distractor* concept, our method is able to reliably reject the distractors in images and significantly improve the detection performance on set $\mathcal{W}$. The overall approach is summarized in Algorithm 1.

TABLE 1
The summaries of Ours-MSD and three baselines. The contents in parentheses indicate the architectures of the models, *i.e.*, Alexet or VGG16.

Method	Learning the knowledge from set $\mathcal{S}$	Traning detectors on set $\mathcal{W}$
Ours-MSD	domain-invariant objectness model (VGG16)	objectness-aware detection model (AlexNet/VGG16)
B-WSD	-	standard MIL-based WSD model (AlexNet/VGG16)
B-MSD	fully supervised Fast RCNN detector (VGG16)	standard MIL-based WSD model (VGG16)
OOM-MSD	original objectness model (VGG16)	objectness-aware detection model (AlexNet)

5 EXPERIMENTS

The proposed MSD method is evaluated on both intra-dataset detection task (Section 5.2) and cross-dataset detection task (Section 5.3). We use two metrics for evaluation: mAP and CorLoc. Mean Average Precision (mAP) is the evaluation metric to test our model on the test set, which follows the standard PASCAL VOC protocol [6]. Correct localization (CorLoc) is to test the proposed model on the training set measuring the localization accuracy [39]. CorLoc is the percentage of images for which the most confident detected bounding box overlaps (IoU $\geq$ 0.5) with a ground truth box.

5.1 Baselines

In this section, we first introduce three baseline methods to compare with our robust objectness transfer MSD approach. For simplicity, we term our robust objectness transfer MSD approach as **Ours-MSD** in the following sections.

B-WSD (the **B**aseline **W**eakly **S**upervised **D**etection method). This baseline is a standard MIL-based WSD method, which is trained on set $\mathcal{W}$ only and does not utilize the objectness knowledge. Specially, in B-WSD, **all** regions in an image construct an image bag tagged with its object category label. During training, B-WSD only aims to model the concepts of K object categories and the loss function in (6) is adapted to sum over K categories.

B-MSD (the **B**aseline **M**ixed **S**upervised **D**etection method). In Ours-MSD, the objectness knowledge is first modelled on set $\mathcal{S}$ and then transferred to set $\mathcal{W}$ to learn the objectness-aware detection model. In B-MSD, we utilize a straightforward *fine-tuning* approach to transfer the knowledge of *detection task* from set $\mathcal{S}$ to set $\mathcal{W}$. We first train a fully supervised Fast RCNN detector on set $\mathcal{S}$. Then we fine tune the obtained detector on set $\mathcal{W}$ under a MIL-based WSD framework. That is, we train a B-WSD model on set $\mathcal{W}$ initialized using the fully supervised detector.

OOM-MSD (the **O**riginal **O**bjectness **M**odel for **M**ixed **S**upervised **D**etection). Similar to Ours-MSD, the OOM-MSD also utilizes objectness knowledge to separate the objects and distractors and then trains objectness-aware detection model. The difference lies in the architecture of the objectness model. In OOM-MSD, the domain adaptation component of the objectness model (*i.e.*, the domain classifier branch in Fig. 2) is removed and the *orginial* objectness knowledge is learned from set $\mathcal{S}$ only.

5.2 Intra-dataset Detection

5.2.1 Benchmark Data

In the intra-dataset detection case, we evaluate our method on the PASCAL VOC 2007 [6] dataset and the ILSVRC2013 detection [8] dataset.

The PASCAL VOC 2007 dataset contains 20 common object classes, 2,501 training images, 2,510 validation images and 5,011 test images. To train the objectness models and objectness-aware WSD models, we split the *trainval* set (5,011 images in total) into two sets: images belonging to the first 10 categories of PASCAL constitute set $\mathcal{S}$ and images of the last 10 categories construct the set of weak categories (set $\mathcal{W}$). The strong set $\mathcal{S}$ includes 3,002 images. We have access to their bounding box annotations and model objectness knowledge from these annotated boxes. The weak set $\mathcal{W}$ contains 3,340 weakly labeled images, which are used to train objectness-aware detection models. The detection models are evaluated on *test* set and the mAP is computed over the last 10 categories.

The ILSVRC2013 detection dataset contains 200 basic level object cateogories, 395,909 images for training, 20,121 images for validation, and 40,152 images for testing. The validation set is split in half: *val1* and *val2*, as in R-CNN [3]. Then we collect images with bounding box annotations from both *train* and *val1* to construct our training set, *trainval1* (107,452 images in total). Similar to PASCAL VOC, we also split the *trainval1* set into two sets: the first 100 and the last 100 categories (in alphabetical order) correspond to the strong categories (54,735 images) and weak categories (57,584 images) respectively. Finally, the detection models are evaluated on *val2* set (9,917 images) and the mAP is calculated over the last 100 categories.

It is noted that the training images in both datasets are possible to contain more than one object class. Thus a portion of training images would be included in both strong and weak sets. For example, an image containing both *dog* (strong category) and *person* (weak category) will be used in both sets. In this case, our method considers the image as a fully labeled *dog* image in set $\mathcal{S}$ and a weakly labeled *person* image in set $\mathcal{W}$.

5.2.2 Implementation Details

In this section, we introduce the detailed implementation settings for Ours-MSD and three baselines. Table 1 summarizes their statistics.

Learning the knowledge from set $\mathcal{S}$. In OOM-MSD baseline and Ours-MSD, we train the original objectness model and the domain-invariant objectness model from set $\mathcal{S}$ respectively. Both the two objectness models start from VGG16 models pre-trained on ImageNet classification [40]. The last 1000-way fc layer of VGG16 is changed to a new 256-way fc layer and its output, the 256-dimensional vector, serves as the **feature f** introduced in Section 4.1. The **feature f** is then used for both objectness prediction and domain classification. For objectness prediction, $\mathbf{f}$ is directly connected to a 2-way fc layer; for domain classification, $\mathbf{f}$ is connected to 3 fc layers (1024-1024-2). Note that this domain classifier branch is removed in OOM-MSD.

During training, the objectness models are trained for 10 epochs using stochastic gradient descent (SGD). The initial learning rate is set to 0.001 for the first 5 epochs and decreased to 0.0001 for the last 5 epochs. A momentum 0.9 and a weights decay of 0.0005 are used. For domain-invariant model, the reversed gradients collected from domain classifier start from $-0.0 \times$ *gradient* and gradually decrease to $-0.1 \times$ *gradient* in the first 8,000 training iterations to make the training process stable in the early iterations.

In B-MSD baseline, we train a supervised Fast RCNN detector [2] on set $\mathcal{S}$. The detector is initialized using VGG16 pre-trained on ImageNet classification, which is the same as the objectness models used in OOM-MSD and Ours-MSD. During training, the Fast RCNN detector is trained for 20 epochs, and the learning rates are set to 0.001 and 0.0001 for the first and the last 10 epochs respectively.

Training object detectors on set $\mathcal{W}$. For OOM-MSD and Ours-MSD, the objectness knowledge is utilized to train the objectness-aware detection models; for B-WSD and B-MSD, standard MIL-based WSD models are learned. As shown in Table 1, the implementations of the four detection models (Ours-MSD and three baselines) are different. The details are as follows:

In previous MSD method (*i.e.*, LSSOD [23]), the weakly supervised detectors are based on AlexNet [30]. Thus, for a fair comparison, the detection models used in B-WSD, OOM-MSD and Ours-MSD are initialized using the same AlexNet model pre-trained on ImageNet. The only exception is B-MSD. B-MSD proposes to fine tune VGG16 Fast RCNN detectors to the VGG16 WSD models. To compare with the B-MSD baseline, we additionally train detection models based on VGG16 for B-WSD baseline and Ours-MSD.

During training, the detection experiments run for 20 epochs and the learning rates are set to 5×10^{-5} and 5×10^{-6} respectively for the first and last 10 epochs. Similar to Fast RCNN [2], the aspect ratios of the input images are retained fixed, and the shorter sides of images are resized to 600. Only horizontal flipping is applied as a form of data augmentation. At testing time, non-maximum suppression (NMS) is used to ignore redundant, overlapped boxes and the threshold of NMS is set to 0.3.

5.2.3 Evaluation on Large-scale Dataset

In this section, the proposed method is evaluated on ILSVRC2013 detection dataset. The experiment results (mAP %) on set $\mathcal{W}$ (the last 100 categories of ILSVRC2013 detection) are shown in Table 2. We first focus on the WSD methods (the first compartment of Table 2). With the image-level labels only, B-WSD achieves a relatively low performance (B-WSD-AlexNet, 13.78%). When utilizing deeper networks, the performance of B-WSD is even lower (B-WSD-VGG16, 11.82%). The goal of B-WSD is to distinguish between different object categories. Thus, with more capable VGG16 networks, B-WSD might tend to search more discriminative object parts, rather than the whole objects. It will lead to inferior detection results. Regarding the B-WSD baseline, an interesting comparison is that the performance of B-WSD outperforms the state-of-the-art WSD result [26] by a large margin (13.78% *vs.* 6.25%). While in PASCAL VOC dataset (as will be shown in Table 4), [26] could easily surpass B-WSD (31.0% *vs.* 23.87%). In the algorithm of [26], several hyperparameters, *e.g.*, the number of mined proposals, need to be set. In small dataset, such as PASCAL VOC, these hyperparameters could be decided by using cross-validation. However, in large-scale dataset, searching proper hyperparameters for hundreds of object categories is much more difficult and it finally results in unsatisfactory performance in large-scale scenarios. We believe that the robust algorithms with few hyperparameters are more appropriate, especially in large-scale datasets, to obtain high-performance detectors.

TABLE 2
Object detection performance (mAP %) on ILSVRC2013 *val2* set. The mAP is computed over the last 100 categories. * The results of [26] are obtained by averaging the reported numbers (in the supplementary material of [26]) over the last 100 categories

Method	mAP on set $\mathcal{W}$ (100 categories)
weakly supervised:	
B-WSD-AlexNet	13.78
B-WSD-VGG16	11.82
(OM+MIL)+FRCN-AlexNet [26] *	6.25
(OM+MIL)+FRCN-VGG16 [26] *	9.10
mixed supervised:	
LSSOD (visual) [23]	19.02
LSSOD (semantic) [23]	19.04
ens-LSSOD [23]	20.03
B-MSD-VGG16	16.44
OOM-MSD-AlexNet	18.54
Ours-MSD-AlexNet	22.28
Ours-MSD-VGG16	**25.26**
fully supervised:	
FRCN-AlexNet:	26.40
FRCN-VGG16:	30.82

Different from the WSD methods, the OOM-MSD and Ours-MSD that are based on objectness transfer approach obtain much better results. Regarding the OOM-MSD baseline, as we have sufficient strong categories (the first 100 categories) in large-scale dataset, the original objectness model already learns, to some extent, "general" objectness knowledge. Thus, the subsequent objectness-aware detection models could obtain remarkable improvement on set $\mathcal{W}$ (18.54% *vs.* 13.78%). The performance of Ours-MSD is improved to 22.28% by applying domain-invariant model, where the learned objectness knowledge is more robust to the change of categories. Ours-MSD could obtain further improvements by utilizing deeper detection networks (Ours-MSD-VGG16, 25.26%). As the object parts are more likely recognized as *distractors* and rejected with deeper networks in Ours-MSD. By fine tuning the supervised detectors to WSD models, the B-MSD baseline can improve the B-WSD baseline (16.44% *vs.* 11.82%). But the performance is still obviously lower than Ours-MSD (16.44 % *vs.* 25.26%). In some way, with the fine-tuning approach, B-MSD aims to reduce the gap between the *detection task* and the *classification task*. In contrast, our

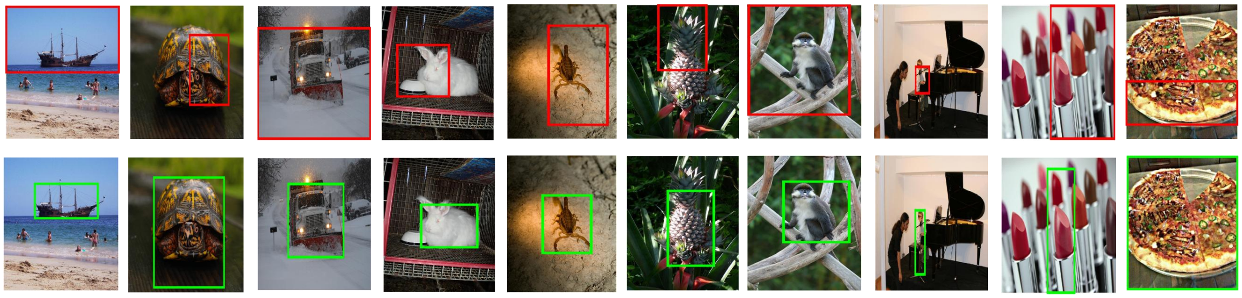

Fig. 3. Sample detection results on ILSVRC2013 *val2* set of the last 100 categories. The first row shows the results of the baseline WSD model (B-WSD-AlexNet) and the second row lists the results of the same images predicted by the objectness-aware model (Ours-MSD-AlexNet).

TABLE 3
Object detection performance (mAP %) on PASCAL VOC 2007 *test* set. The mAP is computed over the last 10 categories.

Method	mAP on set $\mathcal{W}$ (10 categories)
B-WSD-AlexNet	22.63
LSSOD (visual)	5.76
OOM-MSD-AlexNet	14.00
Ours-MSD-AlexNet	**32.21**

MSD method transfers the *objectness* knowledge to weak categories, which is crucial to detect objects. Moreover, our method also aims to reduce the discrepancy between different categories. The last compartment of Table 2 shows the results of fully supervised Fast RCNN detectors trained on the last 100 categories. Compared with these oracle detectors, our MSD method attains comparable performance (*e.g.*, 22.28% *vs*. 26.40%).

In order to illustrate our improvements more clearly, we draw the detection results of the B-WSD and Ours-MSD for the same image, as shown in Fig. 3. We can see that, in Ours-MSD, the objectness-aware model is endowed with the ability to reject the distractors (contexts or object parts) by incorporating the objectness knowledge. Such ability is essential for object detection and finally results in the remarkable improvements.

Our method is then compared with the state-of-the-art MSD method LSSOD [23], which performs intra-dataset detection in ILSVRC2013 (our experimental setup is slightly different from LSSOD, and we will clarify this in supplementary materials). LSSOD transfers the classifier and detector differences from strong to weak categories based on their visual and semantic similarities, which correspond to the visual transfer model and semantic transfer model respectively. The best performance of the two models are 19.02% (visual) and 19.04% (semantic). The results are slightly higher than the OOM-MSD baseline (18.54%), but they are lower than that of our method (22.28%). The ensemble of the two models (20.03%) still cannot compete with our method. These results demonstrate the effectiveness of our method on large-scale datasets and prove that making use of transferable objectness knowledge to improve WSD is reasonable and successful.

5.2.4 Evaluation on Smaller Dataset

We also compare these methods on a much smaller dataset, PASCAL VOC 2007, where only a small quantity of fully labeled categories are available. The results are shown in Table 3. In the smaller dataset, the distribution discrepancy between strong and weak categories becomes very large. The original objectness model trained with only 10 strong categories will have strong bias on these categories, and the subsequent objectness-aware detection model only achieves 14.00% on weak categories, which is even far below the baseline method (22.63%). LSSOD cannot deal with the large distribution discrepancy as well. During the knowledge transfer process, LSSOD assumes that the differences learned on layers 1-7 in AlexNet are category-*invariant*, and those differences are shared between strong and weak categories. However, with large distribution discrepancy, layers 1-7 actually learn category-*specific* representations, and this assumption is no longer hold. As a result, LSSOD completely fails (5.76%) in this scenario.

On the contrary, we can relieve these impacts by training domain-invariant objectness model. With the domain-invariant knowledge, our method significantly improves the detection performance to 32.21%. It proves the ability of our domain-invariant method to cope with the large distribution discrepancy. Even if we only have access to limited fully labeled categories, we still be able to use the proposed method to localize new objects.

5.3 Cross-dataset Detection

5.3.1 Benchmark Data

In this section, we evaluate our MSD framework on a cross-dataset detection task. The task is conducted between PASCAL VOC 2007 dataset and ILSVRC2013 detection dataset. In the cross-dataset scenario, all the images in VOC 2007 *trainval* constitute set $\mathcal{W}$ (20 categories, 5,011 images); the images whose categories do not overlap with set $\mathcal{W}$ are selected from the ILSVRC2013 *trainval1* to form the set $\mathcal{S}$ (180 categories, 89,391 images). The mAP is used to evaluated the performance of the detection models on VOC 2007 *test* set over all the 20 categories. The CorLoc [39] is applied to measure the localization accuracy of models on VOC 2007 *trainval* set. We also conduct the cross-dataset detection on PASCAL VOC 2010 and PASCAL VOC 2012. The object categories of VOC 2010 and VOC 2012 are the same as VOC 2007. The size of VOC 2010 and VOC 2012 is approximately

TABLE 4
Object Detection performance (mAP %) on PASCAL VOC 2007 *test* set. "Ours-MSD-Ens" is the ensemble model obtained by averaging the outputs of AlexNet and VGG16 objectness-aware models. "Ours-MSD-Ens + FRCN-VGG16" indicates the method that utilizes our ensemble model to select top-scoring regions as pseudo ground truths and then trains a supervised Fast RCNN detector [2] using VGG16 model.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
weakly supervised:																					
B-WSD-AlexNet	40.5	35.3	19.5	5.8	7.7	38.9	39.9	23.3	1.6	25.0	11.1	25.2	29.9	49.5	21.3	16.4	24.4	16.8	35.1	10.5	23.87
B-WSD-VGG16	43.1	32.5	18.5	8.7	13.1	33.5	37.3	18.0	11.4	18.3	21.8	28.5	23.3	46.4	8.5	16.3	22.6	27.8	43.3	27.9	25.03
Wang *et al*. [27]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
WSDDN-VGG_F [10]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
WSDDN-VGG_M [10]	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	37.8	46.9	53.4	47.9	34.9
WSDDN-VGG16 [10]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
WSDDN-Ens [10]	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
(OM+MIL)+FRCN-AlexNet [26]	49.7	33.6	30.8	19.9	13.0	40.5	54.3	37.4	14.8	39.8	9.4	28.8	38.1	49.8	14.5	24.0	27.1	12.1	42.3	39.7	31.0
(OM+MIL)+FRCN-VGG16 [26]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	**21.8**	57.1	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5
ContextLocNet [25]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	**49.2**	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
WCCN-AlexNet [24]	43.9	57.6	34.9	21.3	14.7	64.7	52.8	34.2	6.5	41.2	20.5	33.8	47.6	56.8	12.7	18.8	39.6	46.9	52.9	45.1	37.3
WCCN-VGG16 [24]	49.5	60.6	38.6	29.2	16.2	**70.8**	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	**60.8**	54.5	42.8
mixed supervised:																					
B-MSD-VGG16	40.1	49.2	31.2	13.9	23.9	48.7	52.9	42.6	4.9	40.5	26.1	38.5	36.5	61.3	**22.3**	14.6	39.1	22.4	37.4	24.7	33.53
OOM-MSD-AlexNet	52.5	49.1	37.9	35.1	15.5	49.9	51.9	53.2	8.9	40.1	17.9	47.6	50.9	59.3	11.6	16.8	43.7	26.3	43.6	40.9	37.65
Ours-MSD-AlexNet	55.8	56.6	41.1	35.1	22.8	60.1	58.5	55.0	10.3	48.5	22.2	50.5	55.8	61.6	12.8	21.7	44.4	26.1	46.8	49.4	41.77
Ours-MSD-VGG16	61.1	66.3	**54.8**	12.5	**36.5**	63.1	61.9	**66.9**	17.5	66.1	14.3	69.3	65.4	69.6	2.4	20.5	54.6	34.3	58.3	54.6	47.50
Ours-MSD-Ens	65.6	67.2	54.3	31.2	35.8	68.1	65.0	63.3	17.3	**66.8**	18.3	70.6	66.7	**69.8**	3.7	24.7	**55.0**	37.4	58.3	57.3	49.82
Ours-MSD-Ens + FRCN-VGG16	**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**
fully supervised:																					
FRCN-AlexNet	61.5	64.6	46.3	34.3	20.4	65.0	65.9	63.8	27.0	57.2	55.0	54.3	62.0	66.8	48.0	23.7	49.0	49.0	62.7	58.6	51.8
FRCN-VGG16	71.7	72.0	60.8	45.1	32.3	73.6	76.6	78.7	35.8	72.3	62.8	75.1	71.1	73.0	61.2	32.0	62.3	65.6	70.6	65.5	62.9

twice larger than VOC 2007 for both *trainval* and *test* sets. For VOC 2010, the *trainval* images construct set $\mathcal{W}$ and the detection performance is evaluated on VOC 2010 *test* set over all the 20 categories. The same cross-dataset setting is adopted for VOC 2012 dataset.

5.3.2 Implementation Details

In cross-dataset detection task, the most of the experimental settings (learning rates, nms threshold, *etc*.) are as same as the ones in intra-dataset detection case (Section 5.2.2). The only difference is that, when the detection models in Ours-MSD and the three baselines are learned, we use five image scales $\{480, 576, 688, 864, 1200\}$ for both training and testing as an additional form of data augmentation. This multi-scale training/testing strategy is widely-used in recent fully/weakly supervised detection methods [5], [10], [26] and has proven effective on PASCAL VOC dataset.

5.3.3 Evaluation on PASCAL VOC 2007

Our results for each class on PASCAL VOC 2007 are reported in Table 4 (mAP) and Tabel 5 (CorLoc). The first compartment in both tables shows the results obtained by state-of-the-art WSD methods [10], [26] and B-WSD baseline, which are trained on the weak categories (20 categories in PASCAL VOC) only. The second compartment reports the results of the MSD methods that leverage extra strong categories (180 categories in ILSVRC2013) for training. Additionally, in Table 4, the performance of fully supervised Fast RCNN detectors is listed in the third compartment. The Fast RCNN detectors are trained without bounding box regression and the results are cited from the *experiment logs* released by Fast RCNN. §.

As shown in Table 4 and Table 5, using a single AlexNet model, Ours-MSD achieves huge improvements in mAP (41.17% *vs*. 23.87%) and in CorLoc (61.00% *vs*. 41.35%) compared with the B-WSD baseline. This performance also significantly exceeds the state-of-the-art WSD results [24], [25] (41.77% *vs*. 37.3%/36.3%). When compared with OOM-MSD, our method with robust objectness knowledge also shows superiority (41.77% *vs*. 37.65%). When Ours-MSD is trained with the deeper VGG16 detectors, the result is improved to 47.50%, which also largely outperforms the B-WSD baseline (25.03%), the B-MSD baseline (33.53%) and previous WSD results [24], [26] (42.8%/39.3%) that also adopt VGG16 detectors. Similar to WSDDN [10], our results can be further improved by combing multiple models. The ensemble model used in our method (Ours-MSD-Ens) is obtained by simply summing up the scores of AlexNet detector (Ours-MSD-AlexNet) and VGG16 detector (Ours-MSD-VGG16) and it finally achieves 49.82%, which outperforms the ensemble results in WSDDN by a large margin (49.82% *vs*. 39.3%). Also, similar to [26], we use the obtained ensemble model to select top-scoring regions (select one highest-score region from each image of each category) as pseudo ground truth boxes to train a supervised VGG16 Fast RCNN detector [2] with no bounding box regression. Further improvements can be obtained with this process (51.08% *vs*. 49.82%). Finally, as shown in Table 4, even when compared with fully supervised Fast RCNN detectors, our MSD methods can achieve comparable detection results.

Then we compare the proposed method with the state-of-the-art MSD method, WSLAT [22], which also leverages the strong categories of ILSVRC2013 to support the WSD learning on PASCAL VOC 2007. The proposed method is compared with WSLAT in terms of CorLoc (%) in Table 5, since WSLAT does not report their mAP results on VOC 2007 *test* set. It is noted that WSLAT has *three* variants in [22]: 1) a WSD model directly trained on weak categories (denoted as WSLAT$_{weak}$); 2) a transfer model (denoted as WSLAT$_{trans}$) built on both strong and weak categories leveraging their semantic relationships, and 3) ensemble model (denoted as WSLAT-Ens) by combining the above two. Our method outperforms *all* the three variants. In

§. https://dl.dropboxusercontent.com/s/q4i9v66xq9vhskl/fast_rcnn_experiments.tgz?dl=0

TABLE 5
Object Detection performance (CorLoc %) on PASCAL VOC 2007 *trainval* set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
weakly supervised:																					
B-WSD-AlexNet	67.50	54.51	36.04	20.21	14.12	54.31	66.10	49.13	9.27	54.79	11.79	41.63	45.58	75.10	38.04	40.00	48.45	30.65	50.95	21.86	41.35
B-WSD-VGG16	76.67	58.82	39.04	26.06	34.73	58.88	72.40	48.55	26.57	43.15	42.21	43.95	48.98	78.71	28.07	38.46	50.52	43.01	59.32	41.94	48.00
Wang *et al*. [27]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
WSDDN-VGG_F [10]	68.5	67.5	56.7	34.3	32.8	69.9	75.0	45.7	17.1	68.1	30.5	40.6	67.2	82.9	28.8	43.7	71.9	62.0	62.8	58.2	54.2
WSDDN-VGG_M [10]	65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
WSDDN-VGG16 [10]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	**34.8**	64.5	**60.5**	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
WSDDN-Ens [10]	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
(OM+MIL)+FRCN-AlexNet [26]	77.3	62.6	53.3	41.4	28.7	58.6	76.2	61.1	24.5	59.6	18.0	49.9	56.8	71.4	20.9	44.5	59.4	22.3	60.9	48.8	49.8
(OM+MIL)+FRCN-VGG16 [26]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
ContextLocNet [25]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
WCCN-AlexNet [24]	79.7	68.1	60.4	38.9	36.8	61.1	78.6	56.7	27.8	67.7	20.3	48.1	63.9	75.1	21.5	46.9	64.8	23.4	60.2	52.4	52.6
WCCN-VGG16 [24]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	**49.1**	68.1	27.9	64.5	55.7	56.7
WSLAT$_{weak}$ [22]	77.31	55.55	62.76	40.88	21.31	77.96	72.1	54.9	14.83	68.79	29.50	56.29	70.38	74.69	43.18	27.35	47.91	26.20	70.88	67.19	53.00
mixed supervised:																					
WSLAT$_{trans}$ [22]	48.32	48.97	17.58	55.25	6.15	32.26	15.85	40.36	28.54	70.92	4.50	15.91	43.55	34.69	13.75	3.26	51.04	28.38	46.74	19.92	31.3
WSLAT-Ens [22]	78.57	63.37	66.36	56.35	19.67	**82.26**	74.75	**69.13**	22.47	72.34	31.00	62.95	74.91	78.37	**48.61**	29.39	64.58	36.24	**75.86**	**69.53**	58.84
B-MSD-VGG16	60.93	68.24	52.25	34.04	38.55	69.04	75.03	58.43	12.41	74.66	36.12	58.37	55.78	83.13	40.81	37.00	73.20	41.40	47.91	42.29	52.97
OOM-MSD-AlexNet	86.25	66.67	66.37	57.45	29.01	66.50	71.48	66.86	20.80	73.97	30.04	67.44	72.11	87.15	29.40	31.87	78.35	47.85	65.40	56.63	58.58
Ours-MSD-AlexNet	85.00	72.16	66.37	66.49	38.17	73.60	78.71	66.28	22.90	79.45	27.38	72.09	78.91	87.55	15.75	39.56	81.44	41.94	63.50	62.72	61.00
Ours-MSD-VGG16	85.83	**77.25**	73.57	49.47	**66.41**	76.14	84.36	74.13	32.87	**86.99**	19.01	**84.19**	**85.03**	90.36	13.37	34.43	**88.66**	44.09	71.86	67.38	65.27
Ours-MSD-Ens	**89.17**	75.69	**75.08**	**66.49**	58.78	78.17	**88.89**	66.86	28.15	86.30	29.66	83.49	83.33	**92.77**	23.68	40.29	85.57	**48.92**	70.34	68.10	**66.79**

TABLE 6
Object Detection performance (mAP %) on PASCAL VOC 2010 *test* set and VOC 2012 *test* set.

Method	VOC 2010	VOC 2012
weakly supervised		
B-WSD-AlexNet	25.25	24.65
B-WSD-VGG16	21.51	21.44
(OM+MIL)+FRCN-AlexNet [26]	21.4	22.4
(OM+MIL)+FRCN-VGG16 [26]	30.7	29.1
ContextLocNet [25]	-	35.3
WCCN-AlexNet [24]	28.8	28.4
WCCN-VGG16 [24]	39.5	37.9
mixed supervised		
B-MSD-VGG16	33.92	33.93
OOM-MSD-AlexNet	35.42	35.43
Ours-MSD-AlexNet	37.98	38.12
Ours-MSD-VGG16	**42.87**	**43.42**

particular, WSLAT$_{trans}$ is mostly close to us. However, its reported CorLoc value is 31.3% [22], far below that of our method (61.00%). Only by combining a high-performance WSD model (WSLAT$_{weak}$, 53.00%), does the WSLAT obtain an acceptable result (WSLAT-Ens, 58.84%).

5.3.4 Evaluation on PASCAL VOC 2010 and VOC 2012

In this section, our MSD method (Ours-MSD) is compared with the baseline methods and other state-of-the-art WSD methods on VOC 2010 and VOC 2012. The detection results (mAP) are reported in Table 6. In general, the detection performance of VOC 2010 and VOC 2012 is lower than that of VOC 2007. When compared with WSD methods, Ous-MSD significantly outperforms B-WSD baseline and the state-of-the-art results [24], [25], [26] with both AlexNet detectors and VGG16 detectors. When the two objectness transfer methods are compared, Ous-MSD also surpasses OOM-MSD. Finally, Ours-MSD outperforms B-MSD baseline by a large margin on both VOC 2010 and VOC 2012 datasets. It conforms the superiority of the robust objectness transfer approach over the straightforward fine-tuning approach.

TABLE 7
Object detection performance (mAP) on PASCAL VOC 2007. "MSD-no-distractor" indicates the alternative baseline that only learns 20 object categories from the selected object regions (top 15%).

Method	mAP
B-WSD	23.87
MSD-no-distractor	25.95
Ours-MSD	41.77

5.4 Ablation Studies

In this section, we conduct some ablation experiments to illustrate the effectiveness of our robust objectness transfer MSD approach. Without loss of generality, the comparisons are performed on cross-dataset detection task and trained with the AlexNet model. All the experiments follow the same settings mentioned in Section 5.3.2 (learning rates, nms threshold, multi-scale strategies, *etc*.).

5.4.1 Is Learning the Concept of Distractors Necessary for WSD?

To explore the necessity of modelling distractors in WSD, we propose an alternative baseline, MSD-no-distractor and compare it with the proposed method (Ours-MSD). MSD-no-distractor also utilizes the objectness knowledge but aims to learn object categories only. Specially, MSD-no-distractor first uses the obtained objectness model to score the regions in each weakly labeled image. Then it selects the top 15% of regions as the "object regions", which is similar to Ours-MSD. The difference is that MSD-no-distractor only utilizes the selected top 15% "object regions" to train a 20-class (20 categories of PASCAL VOC) WSD model; while Ours-MSD utilizes both the top 15% "object regions" and the last 85% "non-object regions" to train a 20+1-class objectness-aware detection model.

The performance comparison of the three methods (B-WSD, Ours-MSD and MSD-no-distractor) is shown in Table 7. We can see that the performance of MSD-no-distractor is only slightly better than that of B-WSD (25.95% *vs*. 23.87%). Compared with B-WSD, MSD-no-distractor has already

TABLE 8
Object detection performance (mAP) on PASCAL VOC 2007 *test* set. The mAP is computed over 8 *natural object* categories.

Method	bird	cat	cow	dog	horse	person	plant	sheep	mAP
B-WSD	21.5	20.8	25.9	26.7	23.5	18.3	12.8	19.9	21.17
Ours-MSD	40.7	37.0	37.6	23.1	45.1	18.6	17.9	28.6	31.08

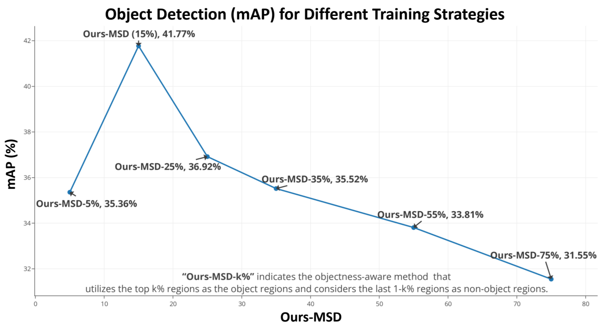

Fig. 4. Object Detection performance (mAP) on PASCAL VOC 2007. "Ours-MSD-k%" indicates the objectness-aware method that utilizes the top k% regions as the object regions and considers the last 1-k% regions as non-object regions.

largely reduced the search space for object categories (from 100% regions used in B-WSD to the selected 15% regions used in MSD-no-distractor), but it still cannot distinguish the objects from distractors. When the trained detectors saw a distractor, *e.g.*, a *cat face*, the MSD-no-distractor cannot recognize it as a false detection due to the missing of distractor concept and the obtained improvements are quite small. Only when we learn the distractor concept together with the object categories (Ours-MSD), does the detector correctly distinguish between objects and distractors and achieve remarkable improvements (41.77% *vs*. 23.87%).

5.4.2 The Effect of the Quantity of Selected Object Regions

In Ours-MSD, we select the top 15% of regions as object regions to train the objectness-aware detection models. In this ablation experiment, we analyse the influence of selected object regions's quantity. We apply the same domain-invariant objectness model to re-rank the regions in weakly labeled images. Then we use top 5%, top 25%, top 35%, top 55%, top 75% of regions as object regions respectively meanwhile utilize the remaining regions as non-object regions to train the objectness-aware detection models. The results are shown in Fig. 4. It can be seen that when the number of selected object regions are small (*e.g.*, top 5%), the selected object bags do not contain enough positive regions (IoU$\geq$0.5) in images (The detailed recall numbers of positive regions can be found in Fig. 5). Thus the performance improvement of subsequent detection models is limited (Ours-MSD-5%, 35.36%). With more regions selected as objects (from top 5% to top 15%), the recall of positive regions in selected object bags gradually increases, and the performance of detection models also improves (from 35.36% to 41.77%). But when more regions are chosen as objects, the improvements of the recall are relatively small. Moreover, the large number of selected regions would bring in lots of false positives and decrease the performance of detectors.

5.4.3 Discrepant Domains Transfer (Natural Objects vs. Man-made Objects)

In this section, we aim to apply our objectness transfer approach between more discrepant domains. We select 8 *natural object* categories from PASCAL VOC 2007 and construct set $\mathcal{W}$ with images of natural objects (1,929 images). Then all the *man-made* objects are selected from ILSVRC2013 detection (139 categories in total) and their images construct set $\mathcal{S}$. The proposed method is compared with the B-WSD method and the results are shown in Table 8. It can be seen that the performance of Ours-MSD significantly outperforms the B-WSD (31.08% *vs*. 21.17%). Even when the strong categories and weak categories come from more discrepant domains, *i.e.*, *man-made vs. natural*, our objectness transfer approach is still effective to improve the WSD performance.

5.4.4 Comparisons with Other Objectness Detectors

Further experiments are conducted to compare our domain-invariant objectness model with other objectness/proposal methods for *objectness* and *object instance* detection on PASCAL VOC 2007. Four models are considered: Objectness [33], EdgeBox [36] ¶, original objectness model (OOM-MSD, Section 5.1) and our domain-invariant objectness model (Ours-MSD, Section 4.1). For each model, we re-rank the selective search windows based on their *objectness* scores, and compute the *recall* for different percentage of the proposals (*i.e.*, percentage of windows considered containing an object instance) when IoU=0.7. The results are shown in Fig. 5. It can be seen that our domain-invariant objectness model outperforms existing objectness models (*i.e.*, Objectness & EdgeBox) and the original objectness model in all cases. This confirms that our domain-invariant objectness model is a better objectness *detector*, which accounts for the better performance of the proposed domain-invariant objectness.

To test the effect of different objectness models on *object instance* detection, for each of them, we select the top 15% of the re-ranked selective search windows as "objects" to train the objectness-aware detection model. The detection results are shown in Table 9. We can see that when we use existing objectness models, the subsequent detection models (Objectness-MSD & EdgeBox-MSD) obtain significantly lower performance than the ones using the CNN-based objectness models, OOM-MSD & Ours-MSD (10.99%&19.78% *vs*. 37.65%&41.77%). When the two objectness models are compared, Ours-MSD outperforms OOM-MSD.

It is quite a surprising result that the performance of EdgeBox-MSD is much lower than that of OOM-MSD & Ours-MSD. As shown in Fig. 5, the recall numbers at 15% for

¶. The official code of EdgeBox [36] uses the edge-driven *objectness* measure to score sliding windows and finally outputs the selected windows with their *objectness* scores. To calculate the *objectness* score for a specific given box, we use its modified code provided by [14].

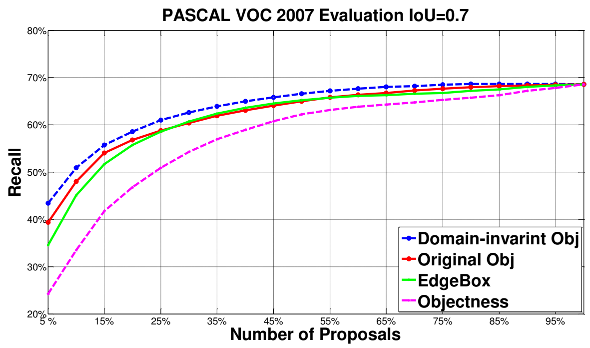

Fig. 5. The recall rates on PASCAL VOC 2007 *trainval* set. We compute the recall using the code provided by [36]. The four curves join at 100% region proposals because we re-rank the existing selective search windows.

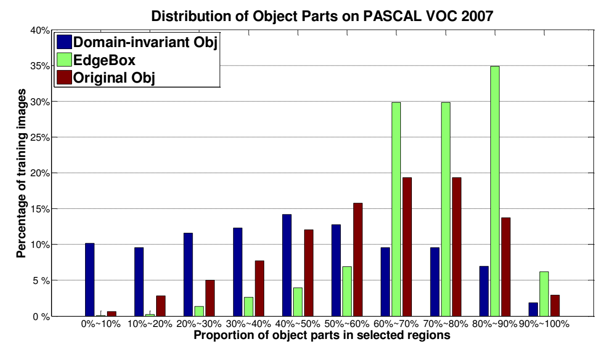

Fig. 6. The distribution of *object parts* for three methods evaluated on PASCAL VOC 2007 *trainval* set. The *object parts* are the regions whose IoUs with ground truths are in the interval $(0.0, 0.5)$. The *object parts* are difficult to reject in weakly supervised setting.

TABLE 9
Object detection performance (mAP) on PASCAL VOC 2007. "Objectness-MSD" & "EdgeBox-MSD" indicate the methods that utilize objectness detector (Objectness & EdgeBox) to score selective search windows and then train objectness-aware detection models.

Method	mAP
Objectness-MSD [33]	10.99
EdgeBox-MSD [36]	19.78
OOM-MSD	37.65
Ours-MSD	41.77

three methods (EdgeBox & original objectness & domain-invariant objectness) are very close, while the detection performance of subsequent detection models differs greatly (19.78% & 37.65% & 41.77%). So what makes such a contradiction?

We explore the reasons from the pitfall in weakly supervised detection (WSD) as we mentioned in Section 1. The regions in an image can be divided into three types according to their IoUs with ground truths: *positive objects* (IoU$\geq$0.5), *object parts* (0<IoU<0.5) and *backgrounds* (IoU=0). For WSD, the *positive objects* and *backgrounds* are easily to be distinguished in most cases and the main difficulty is how to separate *positive objects* from *object parts*. In selected object regions (*i.e.*, the top 15% of regions in our experimental setting), few *positive objects* or excess *object parts* would both hurt the WSD performance. The recall rates in Fig. 5 only show the number of *positive objects* in selected regions and do not consider the *object parts*. Thus it cannot roundly reflect the effectiveness of these objectness detectors in WSD. To address this issue, we conduct another experiment to visualize the distribution of *object parts* for three methods. The results are shown in Fig. 6.

In Fig. 6, the x-coordinate stands for the proportion of *object parts* in selected regions and the y-coordinate stands for the percentage of images having the corresponding proportion range of object parts in all training images. For example, the x-coordinate of the first blue bar is (0%$\sim$10%), indicating that "for a weakly labeled image with n (n=2000 for instance) region proposals, we select the top 15% regions as objects ($2000 \times 0.15 = 300$ regions); in the selected 300 regions, the 0%$\sim$10% of these regions (0$\sim$30 regions) are actually *object parts*". Meanwhile, the y-coordinate of the first blue bar, 10.18%, means that "in all 5011 training images, there are 10.18% of the images (around 500 images), in each of which 0%$\sim$10% of the selected regions are *object parts*." As shown in Fig. 6, the mode of distribution curve of *object parts* for EdgeBox (green bars) is obviously on the **right** to the ones of our objectness models (red & blue bars), which means there are more *object parts* in selected regions for EdgeBox. It is noted that EdgeBox only utilizes low-level cues (*i.e.*, contour information) to measure objectness, which has limited capability of rejecting *object parts* in images. Considering a typical case that many *object parts*, such as *cat face*, also have closed contours and can be recognized as *objects* with EdgeBox. In contrast, our objectness model *learns* to capture the concept of "complete objectness" from lots of annotated data, which results in less *object parts*.

We can also observe that the *object parts* in domain-invariant objectness models are also less than the original objectness models. The reason is that the domain-invariant models would include more *backgrounds* in selected regions than the original objectness models. One overlooked fact in PASCAL VOC is that the images in VOC contain lots of "non-target objects" that do not belong to VOC 20 categories [41]. That is, the *backgrounds* regions also contain a lot of "complete objects" belonging to non-target categories. Moreover, when we train domain-invariant objectness models, all regions in images of target domain (PASCAL VOC) are randomly sampled to learn domain-invariant features. Thus, the learned domain-invariant objectness would be robust to not only VOC 20 categories but also the non-target categories. Finally, the "non-target objects" in *backgrounds* are more likely to be recognized as *objects* in domain-invariant objectness models, which results in higher objectness scores for *backgrounds* and leads to fewer *object parts* in selected top 15% regions. Considering both *positive objects* (Fig. 5) and *object parts* (Fig. 6), our experiments clearly confirm the superiority of our domain-invariant objectness model for both *objectness* and *object instance* detection, especially in weakly supervised settings.

5.5 Error Analysis

Though our method achieves outstanding performance for many categories, its performance is still poor for classes

Fig. 7. Sample failure detection results on PASCAL VOC 2007 *test* set for class *chair*, *table* and *person*. Green rectangles are ground truth boxes, and red rectangles indicate failure detections (IoU<0.5).

such as "chair", "table" and "person". For analysis, we show failure detection results on VOC 2007 *test* in Fig. 7. We can see that, for "chair" images, multiple *chairs* often get close together and the *chairs* typically co-appear with a *table*. In this case, it is difficult to figure out single complete *chair* from images. The detectors would prefer to select the whole *table*, as the *table* is the most likely "object". The similar situation also exists in "table" images. The *tables* often appear together with other categories, such as *bottle*, *plate* and *person*. The exact closed contours for such tables cannot be clearly and easily confirmed, and the complete *bottles*, *plates*, or even *pizzas* would be more easily recognized as objects in such table images. The main kind of failure detections of *person* images are caused by the multiple category setting. In PASCAL VOC, some images contain more than one categories, and for example in many person images, the person appears together with *bicycle*, *horse*, or *motorbike*. In such a situation, the selected object regions are considered as positive for both *person* and *horse* categories, and the detectors cannot distinguish between the *person* object and the *horse* object. This issue caused by multiple category setting is, to some extent, intrinsic in weakly supervised settings where only image category labels are available.

6 Conclusion

In this paper, we consider mixed supervised detection (MSD), which aims to leverage the existing fully labeled categories to localize objects of new categories with weak labels only. The weakly supervised detection of new objects does not require expensive bounding box annotations, and satisfactory detection solutions can be obtained by exploiting the existing fully labeled categories. These characteristics make MSD be a practically important problem.

In MSD, the existing fully labeled categories have no overlap with new categories. Thus, the key issue to be solved is how to learn the transferable and robust knowledge from the existing categories to assist the detection on new categories. Previous MSD works [20], [21], [22], [23] transfer the learned object detectors from the existing categories to new categories following some hand-crafted strategies. In contrast, the proposed robust objectness transfer approach automatically *learns* the domain-invariant knowledge, and the proposed objectness-aware detection model further utilizes the learned objectness knowledge to distinguish the objects from distractors. The state-of-the-art object detection performance has been achieved on benchmarking datasets, which confirms the superiority of our proposed method.

References

[1] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *ICCV*, 2015, pp. 1134–1142.

[2] R. Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[6] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "Pascal visual object classes challenge results," *Available from www. pascal-network. org*, 2005.

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

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.
[10] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.
[11] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *BMVC*, 2014.
[12] ——, "Weakly supervised object detection with convex clustering," in *CVPR*, 2015, pp. 1081–1089.
[13] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in *CVPR*, 2014, pp. 2409–2416.
[14] ——, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.
[15] A. Kolesnikov and C. H. Lampert, "Improving weakly-supervised object localization by micro-annotation," in *BMVC*, 2016.
[16] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *ICML*, vol. 32, 2014.
[17] I. Misra, A. Shrivastava, and M. Hebert, "Watch and learn: Semi-supervised learning for object detectors from video," in *CVPR*, 2015, pp. 3593–3602.
[18] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," 2005.
[19] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *CVPR*, 2013, pp. 1650–1657.
[20] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko, "Lsda: Large scale detection through adaptation," in *NIPS*, 2014, pp. 3536–3544.
[21] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko, "Detector discovery in the wild: Joint multiple instance and representation learning," in *CVPR*, 2015, pp. 2883–2891.
[22] M. Rochan and Y. Wang, "Weakly supervised localization of novel objects using appearance transfer," in *CVPR*, 2015, pp. 4315–4324.
[23] Y. Tang, J. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, and L. Chen, "Large scale semi-supervised object detection using visual and semantic knowledge transfer," in *CVPR*, 2016, pp. 2119–2128.
[24] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *CVPR*, 2017.
[25] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *ECCV*, 2016, pp. 350–365.
[26] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *CVPR*, 2016, pp. 3512–3520.
[27] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *ECCV*, 2014, pp. 431–445.
[28] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *BMVC*, 2012.
[29] M. Guillaumin and V. Ferrari, "Large-scale knowledge transfer for object localization in imagenet," in *CVPR*, 2012, pp. 3202–3209.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.
[32] M. Shi, H. Caesar, and V. Ferrari, "Weakly supervised object localization using things and stuff transfer," in *ICCV*, 2017.
[33] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.
[34] W. Kuo, B. Hariharan, and J. Malik, "Deepbox: Learning objectness with convolutional networks," in *CVPR*, 2015, pp. 2479–2487.
[35] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.
[36] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014, pp. 391–405.
[37] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.
[38] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
[39] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *International journal of computer vision*, vol. 100, no. 3, pp. 275–293, 2012.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[41] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra, "Object-proposal evaluation protocol is' gameable'," in *CVPR*, 2016, pp. 835–844.

Yan Li received the B.Eng degree in Automation from Tsinghua University, Beijing, China, in 2014. He is currently working toward the PhD degree in the Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include computer vision, object detection, weakly supervised learning and few shot learning.

Junge Zhang received his B.S. degree from China University of Geosciences (CUG) in 2008. He received his PhD degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2013. In July 2013, Dr. Zhang joined the Center for Research on Intelligent Perception and Computing (CRIPAC). Now he serves as a Special-appointed Professor. His major research interests include computer vision, pattern recognition, deep learning and general artificial intelligence.

Jianguo Zhang is currently a Reader of Visual Computation at University of Dundee, UK. He received a PhD in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2002. His research interests include visual surveillance, object recognition, image processing, medical image analysis and machine learning.

Kaiqi Huang received his BSc and MSc from Nanjing University of Science Technology, China and obtained his PhD degree from Southeast University. He has worked in National Lab. of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science, China and now he is a professor in NLPR. His current research interests include visual surveillance, digital image processing, pattern recognition and biological based vision and so on.