CaT: Weakly Supervised Object Detection with Category Transfer

Tianyue Cao[1]   Lianyu Du[1]   Xiaoyun Zhang[1*]   Siheng Chen[1,2]   Ya Zhang[1,2]   Yan-Feng Wang[1,2]

Cooperative Medianet Innovation Center, Shanghai Jiao Tong University[1]   Shanghai AI Laboratory[2]

{vanessa_, dulianyu, xiaoyun.zhang, sihengc, ya_zhang, wangyanfeng}@sjtu.edu.cn

Abstract

A large gap exists between fully-supervised object detection and weakly-supervised object detection. To narrow this gap, some methods consider knowledge transfer from additional fully-supervised dataset. But these methods do not fully exploit discriminative category information in the fully-supervised dataset, thus causing low mAP. To solve this issue, we propose a novel category transfer framework for weakly supervised object detection. The intuition is to fully leverage both visually-discriminative and semantically-correlated category information in the fully-supervised dataset to enhance the object-classification ability of a weakly-supervised detector. To handle overlapping category transfer, we propose a double-supervision mean teacher to gather common category information and bridge the domain gap between two datasets. To handle non-overlapping category transfer, we propose a semantic graph convolutional network to promote the aggregation of semantic features between correlated categories. Experiments are conducted with Pascal VOC 2007 as the target weakly-supervised dataset and COCO as the source fully-supervised dataset. Our category transfer framework achieves 63.5% mAP and 80.3% CorLoc with 5 overlapping categories between two datasets, which outperforms the state-of-the-art methods. Codes are avaiable at `https://github.com/MediaBrain-SJTU/CaT`.

1. Introduction

Object detection is one of the most fundamental tasks in computer vision [44]. In the past decade, based on deep neural networks, many methods [16, 26, 27] have achieved great success. However, most methods follow the fully supervised setting, which requires a huge number of high-quality annotations, including the precise bounding boxes of objects and their corresponding category labels. This set-
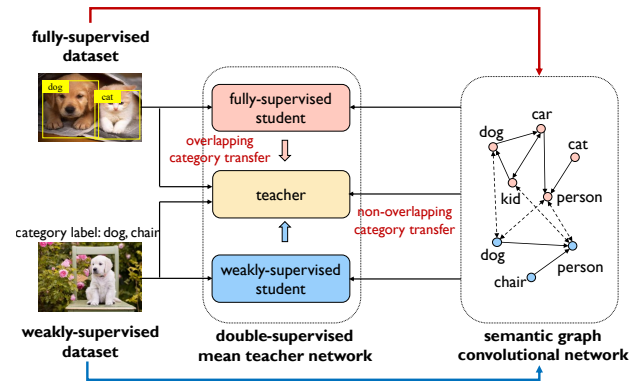
Figure 1. The sketch map of our category transfer framework. We use the double-supervised mean teacher network to transfer overlapping category and objectness knowledge; and use the semantic graph convolutional network to transfer non-overlapping category knowledge. The final prediction is generated by the teacher.

ting usually costs extensive amount of time and resources to acquire such annotations. To reduce the annotation cost, weakly supervised object detection (WSOD) are proposed [2, 32, 33] to train detection models with only image-level category labels. However, the lack of bounding-box-level supervision leads to significant issues, such as instance ambiguity and low-quality proposals. Therefore, a large performance gap still exists between fully-supervised object detection (89.1% mAP, SOTA [10]) and weakly-supervised object detection (56.8% mAP, SOTA [13]).

To narrow this gap, some previous methods consider knowledge transfer from additional data. There are two main approaches: the objectness transfer approach and the semi-supervised approach. For example, [18, 29, 43] train a generic object detector on source data and apply it to target data; however, this objectness transfer methods ignores the category information in the source dataset, causing the decline of classification. [11, 37, 35, 15] follow the semi-supervised setting with partial fully annotated data, and transform an image classifier to an object detector. Such a semi-supervised method leverages both box and category information, but it usually can not solve the domain gap between datasets, especially the category inconsistency be-

tween the source and target datasets. Moreover, the correlations between categories have not been exploited. Therefore, the lack of exploiting category information still limit the empirical performance of the target dataset.

To solve those issues, this work specifically considers category transfer; that is, leverage both visually-discriminative and semantically-correlated category information in the fully-supervised dataset to enhance the discriminative ability of a weakly-supervised detector. Based on whether the categories are shared in both the fully-supervised and the weakly-supervised datasets, the category transfer includes overlapping category transfer, where the fully-supervised and the weakly-supervised datasets share the same categories, and non-overlapping category transfer, where two datasets have different, yet correlated categories.

To realize overlapping category transfer, we propose a double-supervision mean teacher network. The double-supervision mean teacher network adopts the similar structure of mean teacher method [36], but works with two students, each of which is supervised by either fully-supervised or weakly-supervised datasets. Our teacher thus can gather overlapping category information from both fully-supervised and weakly-supervised students, leading to better discriminative ability. Moreover, the mean teacher structure could bridge the domain gap between two datasets, including differences in image and category distributions.

To achieve non-overlapping category transfer, we propose a semantic graph to model the correlations among all the categories in both fully-supervised and weakly-supervised datasets. The intuition is that even two categories are different, they may be highly-correlated and their corresponding category information can be strategically transferred according to the correlation strength. Figure 2 illustrates a toy example of a semantic graph. The category *kid* in the weakly-supervised dataset does not exactly match with the category *person* in the fully-supervised dataset, but they are semantically related. Based on this semantic graph, we use graph convolutional networks to exploit non-overlapping category information and provide semantic guidance for object classification.

Overall, we propose an end-to-end framework; see Figure 1. It includes a Faster R-CNN [27] as the backbone network, a double-supervised mean teacher network for overlapping category transfer, and a semantic graph convolutional network for non-overlapping category transfer. This framework can fill the domain gap by using a mean teacher structure, and fully exploit category information by aggregating semantic features over a semantic graph. Compared with objectness transfer approaches, we transfer category information from a fully-supervised dataset to improve the classifier. Compared with approaches under the semi-supervised setting [11, 37, 35, 15], our approach applies a double supervised mean teacher to solve domain
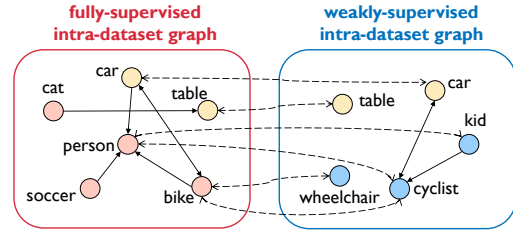

Figure 2. An example of category label graph. Red, blue and yellow nodes represent fully-supervised, weakly-supervised and common categories separately. The inter-dataset graph edges are built according to common categories (*"car(fully)"* ↔ *"car(weakly)"*) and category relationships. The relationships include subclass relationship (*"person(fully)"* ↔ *"kid(weakly)"*), including relationship (*"cyclist(weakly)"* ↔ *"person(fully)"*), and similarity relationship (*"bike(fully)"* ↔ *"wheelchair(weakly)"*). Each intra-dataset graph is a digraph, and the inter graph is a bigraph with bi-directional edges. Best viewed in color.

gap between datasets, and uses a semantic graph convolutional network to fully exploit the correlations between categories, leading to semantic transfer of non-overlapping categories. Our method outperforms the state-of-the-art methods in WSOD, and achieves results competitive to FSOD baseline.

We summarize our main contributions as:

• We propose a novel category transfer framework for WSOD, which specifically handles the issues on category transfer, including the domain gap problem in overlapping category transfer and information aggregation issue in non-overlapping category transfer.

• We propose a novel double-supervised mean teacher network to handle overlapping category transfer. This network gathers common category information and bridges the domain gap between two datasets.

• We propose a novel semantic graph convolutional network to tackle non-overlapping category transfer. This network promotes the aggregation of semantic features between correlated categories.

• We conduct extensive experiments and show that the proposed method outperforms the state-of-the-art weakly-supervised object detection methods and is competitive to fully-supervised object detection baseline on benchmarks.

2. Related Works

Weakly-Supervised Object Detection. Recent works usually formulate WSOD as a MIL problem. WSDDN[2] is the first end-to-end MIL deep neural network in WSOD. Inspired by WSDDN, many methods are proposed. [34] integrates WSDDN and a multi-stage online instance classifier refinement (OICR) algorithm into a simple deep network. [32] use an iterative process, Proposal cluster learning (PCL), to learn refined instance classifiers based on

OICR. [33] introduces continuation multiple instance learning (C-MIL) by smoothing the original loss function to solve the part domination problem. [42] uses top-down confidence and bottom-up evidence with an adaptive training mechanism to distill box boundary knowledge. Follow-up works further improve the performance by self-training [28], self-distillation [42] [13], etc.

Weakly-Supervised Object Detection with Knowledge Transfer. Transfer learning is typically used in domain adaptation (DA) [38]. It can bridge different datasets[36], categories[3], or even tasks[11, 12], which reduces the cost of training and leverages the knowledge from another dataset/domain/task. In this paper, we focus on knowledge transfer in weakly-supervised object detection. Additional data are usually used to provide auxiliary prior information for detection. [35, 1] use the word embedding of category label text to represent the semantic relationship between classes, and the dependency of categories is used to assist object detection. [30, 29, 19] transfer the model learned in source domain to target domain. Also, some works [15, 11, 35] use weight prediction to effectively transform an image classifier into an object detector. Recent studies [6, 37, 18, 23] raise the idea of sharing the general knowledge learned in source domain. The knowledge can be object predictor [6, 23], object candidate region [37], or general boundary box regression [18, 43]. However, categories in two datasets are usually not the same, most of the existing methods do not work under this situation. In this paper, we combine WSOD with mean teacher framework to fully leverage knowledge from both a public fully-supervised dataset and a weakly-supervised dataset, which can also solve the category mismatch problem.

Knowledge-Guided Graph Reasoning. Graph reasoning is proved to be effective in many tasks, including image classification [5, 4], object detection [40, 39], human skeleton-based action recognition [21] or motion prediction [22], etc. These methods model domain knowledge as a graph to transfer knowledge based on category dependency, object spatial relationship, or object semantic relationship. Some classification models [5, 4] build a category dependency graph based on the dataset statistical information. [40] uses knowledge graph to discover most relative categories for feature evolving. Our method not only uses a semantic graph for both fully-supervised and weakly-supervised datasets to reason category dependencies but also transfers category similarity or dependency knowledge between two datasets.

3. Category Transfer Framework

Problem formulation. Mathematically, given a weakly-supervised dataset $\mathcal{D}_w$, each image $\mathbf{x}_w$ in the dataset has an image-level category label $\mathbf{y}_w$ with $\mathbf{y}_w \in \mathbb{R}^{C_w}$, where C_w is the number of weakly-supervised dataset categories.

Similarly, for a fully-supervised dataset $\mathcal{D}_f$, each image $\mathbf{x}_f$ has the instance-level annotations, including the bounding boxes $B_f = \{\mathbf{b}_f^{(i)}\}_{i=1}^r$ and their corresponding category labels $Y_f = \{\mathbf{y}_f^{(i)}\}_{i=1}^r$, where r is the number of instances in an image and $\mathbf{y}_f^{(i)} \in \mathbb{R}^{C_f}$ is the category label of the ith instance with C_f, the number of fully-supervised dataset categories. We aim to train a model $\mathcal{M}$ on $\mathcal{D}_w$ with knowledge transfer from $\mathcal{D}_f$. For each testing image $\mathbf{x}$, the model can output instance-level detection, including estimated bounding boxes and category estimations; that is $\widehat{B}, \widehat{Y} = \mathcal{M}(\mathbf{x})$.

Overview. To design such a model, we specifically consider category transfer; that is, leverage the category information for object recognition in a fully-supervised dataset to enhance the discriminative ability of a detector, which is trained on a weakly-supervised dataset. Category transfer enables the discriminative ability transfer from a fully-supervised dataset to a weakly-supervised detector, leading to better detection performance.

As shown in Figure 3, the proposed category transfer framework includes three parts: the backbone network, the double-supervision mean teacher network (Section 3.1) and the semantic graph convolutional network (Section 3.2). For an input image, our backbone network follows the backbone of Faster R-CNN [27], extracts visual features and generates the regions of proposals. The double-supervision mean teacher network takes the pooled visual features in each region and estimates the bounding boxes and classes by leveraging the bounding box and overlapping category information from both fully-supervised and weakly-supervised datasets. To further enable non-overlapping category transfer, the semantic graph convolutional network exploits the semantic correlations among categories and outputs the semantic features for each category. To classify each region, we aggregate the semantic features produced by the semantic graph convolutional network and the visual features from the R-CNN output of the double-supervision mean teacher network to generate final category features, boosting the recognition ability.

3.1. Double-Supervision Mean Teacher Network

The proposed double-supervision mean teacher network follows the traditional mean teacher, yet with two different student branches. This novel architecture can bridge the domain gap and estimate the bounding boxes and classes via overlapping category transfer.

Network architecture. This network consists of a fully-supervised student branch, a weakly-supervised student branch, and a teacher branch. Each of the three branches takes the pooled visual features for proposal regions produced by the backbone network as the input and updates the visual feature through a few convolutional layers; that is, given the pooled visual feature F^p, the visual features after
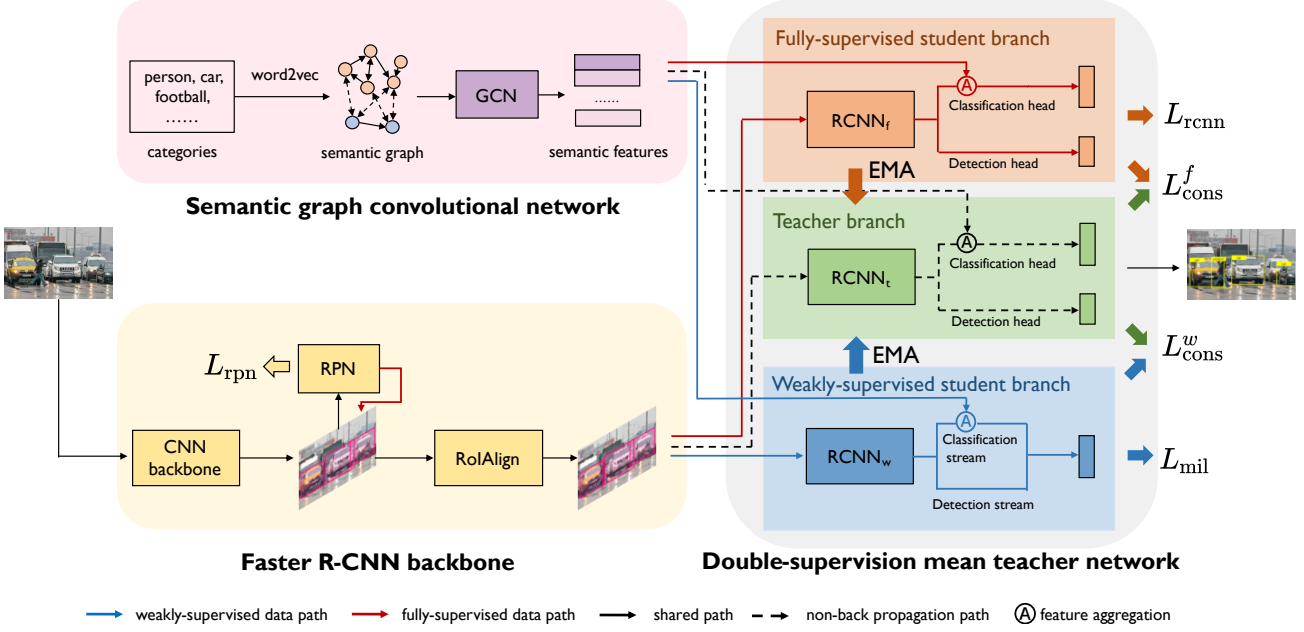
Figure 3. Overall architecture of our category transfer framework. Image region features are forward to the double-supervision mean teacher, which has three branches: the fully-supervised student branch, the weakly-supervised student branch, and the teacher branch. The semantic features are updated by the semantic graph convolutional network and fused to visual features for classfication. The training process in one iteration contains two forward steps (solid lines), one backpropagation step (dotted lines), and one exponential moving average (EMA) step.

convolutional layers in Branch $*$ is $F_*^o = \mathrm{RCNN}_*(F^p) \in \mathbb{R}^{r \times d}$, where r is the number of region proposals, d is the dimension of features, $\mathrm{RCNN}_*(\cdot)$ is the regional convolutional neural network and $*$ is the branch index with f for the fully-supervised branch, w for the weakly-supervised branch and t for the teacher branch.

Taken fully-supervised region features F_f^o as input, the fully-supervised student branch uses a regression head and a classification head to output the estimated bounding boxes $\widehat{B}_f = \{\widehat{\mathbf{b}}_f^{(i)}\}_{i=1}^r$ and the corresponding categories $\widehat{Y}_f = \{\widehat{\mathbf{y}}_f^{(i)}\}_{i=1}^r$, respectively. The outputs are supervised by the ground-truth bounding boxes and the the corresponding category labels in the fully-supervised dataset. Note that the estimated label is $\widehat{\mathbf{y}}_f^{(i)} \in \mathbb{R}^{C_f}$, where C_f is the number of categories in the fully-supervised dataset. The architecture of this branch is the same as class-agnostic Faster R-CNN [27] after RoIAlign layer.

The weakly-supervised student branch trains a multiple-instance-learning (MIL)-based network inspired by [2]. Taken weakly-supervised region features F_w^o as input, this branch outputs the image-level category prediction $\widehat{\mathbf{y}}_w \in \mathbb{R}^{C_w}$, which is supervised by the image-level category labels in the weakly-supervised dataset. Figure 4 illustrates its architecture. The output feature of the RCNN is input to a classification stream and a detection stream. The classification stream uses a fully-connected layer ϕ_c to map $F_w^o \in \mathbb{R}^{r \times d}$ to $U \in \mathbb{R}^{r \times C_w}$, where r is the number of region
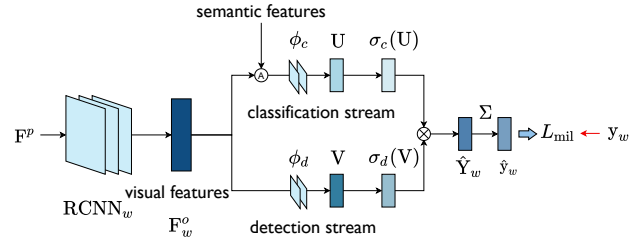

Figure 4. Architecture of the weakly-supervised student branch.

proposals, C_w is the number of categories in the weakly-supervised dataset. We then apply the softmax operation along the category (second) dimension of U and obtain the category prediction scores $\sigma_c(U) \in \mathbb{R}^{r \times C_w}$ for r proposals. Meanwhile, the detection stream uses another fully-connected layer ϕ_d to map F_w^o to $V \in \mathbb{R}^{r \times C_w}$ and applies the softmax operation along the proposal (first) dimension of V to obtain the detection score $\sigma_d(V) \in \mathbb{R}^{r \times C_w}$ for r proposals. Finally, we apply the element-wise multiplication between the results of two streams to get the image-level category prediction.

The teacher branch accumulates knowledge from two student branches by exponential moving average (EMA) [36] without training. It takes both fully-supervised and weakly-supervised region features as input, and outputs the final box predictions $\widehat{B}$ and category predictions $\widehat{Y}$. The teacher branch has a similar architecture as the

fully-supervised student branch and also includes the regression head and the classification head. The difference is that the number of output categories of the teacher branch follows the categories in the weakly-supervised dataset. To transfer the regression ability, we take the exponential moving average of the weights in the regression head of the fully-supervised student branch as the weights in the regression head of the teacher branch. To transfer the classification ability, we take the exponential moving average of the weights in the classification stream of the weakly-supervised student branch as well as the weights of overlapping categories in the fully-supervised student branch to be the weights in the classification head of the teacher branch.

Loss function. The overall loss function includes three parts: the weakly-supervised MIL loss L_{mil}, which is an image-level cross-entropy classification loss [2]; the fully-supervised loss L_{full}, which is same as Faster R-CNN [27], including RPN loss L_{rpn} and R-CNN loss L_{rcnn}; and the consistency loss to promote consistency between the teacher and two student branches; that is, $L_{\mathrm{cons}} = L_{\mathrm{cons}}^f + L_{\mathrm{cons}}^w$, where the fully-supervised and weakly-supervised consistency losses are $L_{\mathrm{cons}}^f = L_{\mathrm{box}}(\widehat{\mathbf{b}}_f, \widehat{\mathbf{b}}) + L_{\mathrm{cls}}(\widetilde{\mathbf{y}}_f, \widetilde{\mathbf{y}}), L_{\mathrm{cons}}^w = L_{\mathrm{cls}}(\widehat{\mathbf{y}}_w, \widehat{\mathbf{y}})$, respectively, with $\widehat{\mathbf{b}}$ the estimated bounding box of the teacher branch, $\widetilde{\mathbf{y}}$ the classification score of the teacher branch for the overlapping categories, $\widehat{\mathbf{y}}$ the classification score of the teacher branch for all the categories, L_{cls} is the smooth-ℓ_1 loss for classification consistency and L_{box} is the mean square error for box regression consistency.

The overall loss function is $L = L_{\mathrm{mil}} + \lambda_{\mathrm{full}}L_{\mathrm{full}} + \lambda_{\mathrm{cons}}L_{\mathrm{cons}}$, where λ_{full} and λ_{cons} are the hyperparameters to balance the weakly-supervised, fully-supervised, and consistency losses.

3.2. Semantic Graph Convolutional Network

To leverage non-overlapping categories in the fully-supervised dataset, we propose a novel semantic graph convolutional network. A semantic graph reflects the correlations between categories. Based on the semantic graph, graph convolutional networks update the semantic features of all categories. The optimized semantic features are finally aggregated with visual features to guide classification.

Semantic graph construction. We build a semantic graph based on the correlations between all the categories. The corresponding objects of the highly correlated categories often appear jointly, which provides a hint for object detection. Therefore, appropriately leveraging the correlations could lead to better classification and detection results.

For each of the fully-supervised and the weakly-supervised datasets, we build a intra-dataset graph by the co-occurrence of the categories to model the dependences between categories. Here each node models a unique category and each edge reflects the dependence between two
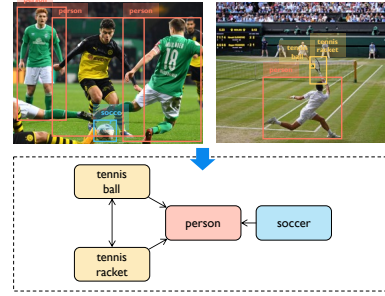

Figure 5. An example of category relationships. The related label graph is a digraph, arrows represent the asymmetric dependencies between categories. "*soccer → person*" means that when *soccer* appears, *person* appears with high probability.

categories. Note that the relationship between categories is not bidirectional, such as the *soccer* and the *person* in Figure 5. Therefore, the dependence is asymmetric and the intra-dataset graph is a digraph. To define the connectivities for each graph, we first compute a graph transition matrix based on the co-occurrence of the categories and then apply thresholding to obtain the graph adjacent matrix. For example, the graph transition matrix for the fully-supervised dataset is $\mathrm{P}_f \in \mathbb{R}^{C_f \times C_f}$. The (i,j)th element of P is the co-occurrence probability between the ith and the jth categories; that is, $(\mathrm{P}_f)_{ij} = M_{ij}/M_i$, where M_i is the number of images with the ith category, and M_{ij} is the number of images with both the ith and the jth categories. Since the graph transition matrix P_f is mostly a full matrix, which could be noisy and cause expensive computation for the subsequent procedures, we introduce a threshold τ to obtain a binary graph adjacent matrix $\mathrm{A}_f \in \mathbb{R}^{C_f \times C_f}$. The (i,j)th element of the graph adjacent matrix, $(\mathrm{A}_f)_{ij} = 0$ when $(\mathrm{P}_f)_{ij} < \tau$ and 1, otherwise. Similarly, we can construct the graph adjacent matrix $\mathrm{A}_w \in \mathbb{R}^{C_w \times C_w}$ for the weakly-supervised dataset.

To further capture semantic relations between categories across two datasets, we introduce inter-dataset edges to connect two intra-dataset graphs, forming a bipartite graph $\mathrm{B} \in \mathbb{R}^{C_f \times C_w}$. Each edge weight can be obtained by the cosine similarity as well as the hand-crafted design. For the similarity-based inter-dataset edges The edge weight between the ith node in the fully-supervised dataset and the jth node in the weakly-supervised dataset is

$$\mathrm{B}_{ij} = \frac{\exp(\mathrm{sim}(\mathbf{v}_i^f, \mathbf{v}_j^w))}{\sum_j \exp(\mathrm{sim}(\mathbf{v}_i^f, \mathbf{v}_j^w))},$$

where $\mathrm{sim}(\cdot, \cdot)$ is the cosine similarity, $\mathbf{v}_i^f$ and $\mathbf{v}_j^w$ are the semantic features of the ith node in the fully-supervised graph and the jth node in the weakly-supervised graph respectively. For the hand-crafted inter-dataset edges, we consider binary edge weights to capture the logical relation-
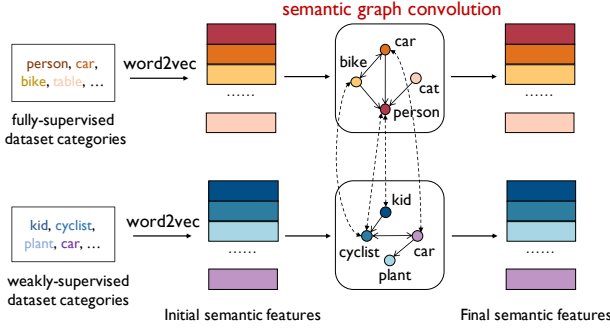
semantic graph convolution

Figure 6. Details of the semantic graph convolutional network. The fully-supervised label graph and the weakly-supervised label graph share the same GNN weights and transmit related semantics between two graphs by inter-dataset graph connections.

ships between categories. When two nodes in the respective graphs have the subclass or inclusion relation, the corresponding edge weight is set to be 1, else 0. For example, *pedestrain* is a subclass of *person*, so the edge weight between *pedestrain* and *person* is 1.

Network architecture. Figure 6 illustrates the detailed architecture of the semantic graph convolutional network. We first use word2vec pretrained on wiki corpus to initialize the semantic feature for each category, incorporating semantic information as well as linguistic knowledge. Let $H_f^{(0)} \in \mathbb{R}^{C_f \times k}$ and $H_w^{(0)} \in \mathbb{R}^{C_w \times k}$ be the semantic features of all the categories for the fully-supervised and weakly-supervised datasets, respectively, where C_f is the number of categories of the fully-supervised dataset, C_w the number of categories of the weakly-supervised dataset, and k the dimension of word2vec. We then use two graph convolution layers [14] to update the semantic features based on the intra-dataset graphs. For the fully-supervised dataset, the ith graph convolution layer takes the semantic feature from the previous layer as the input and outputs the latent feature, $Z_f^{(i+1)} = \text{GCN}(A_f, H_f^{(i)})$, Similarly, we can obtain the latent feature $Z_w^{(i+1)}$ for the weakly-supervised dataset. Next, we update the semantic features of the categories in both datasets via the inter-dataset graph:

$$H_f^{(i+1)} = Z_f^{(i+1)} + \sigma(B\, Z_w^{(i+1)}\, W_f^{(i)}),$$
$$H_w^{(i+1)} = Z_w^{(i+1)} + \sigma(B^T\, Z_f^{(i+1)}\, W_w^{(i)}),$$

where $B \in \mathbb{R}^{C_f \times C_w}$ reflects the inter-dataset graph, $W_*^{(i)}$ is a trainable weight matrix for each dataset, and σ is the activation function. The semantic graph convolutional network outputs the final semantic features for both datasets, $H_f \in \mathbb{R}^{C_f \times d}$ and $H_w \in \mathbb{R}^{C_w \times d}$, where d is the feature dimension, which is the same as the feature dimension of the visual features F_*^o in each of two student branches of the double-supervision mean teacher network.

Fusion with double-supervision mean teacher network. Both semantic features are consumed by the classification heads of two student branches. For example, in the fully-supervised student branch, we aggregate the semantic features and the visual features to obtain the final visual features for classification; that is, $\check{F}_f^o = g(F_f^o\, H_f^T) + F_f^o \in \mathbb{R}^{r \times d}$, where r is the number of proposal regions, $g(\cdot)$ is a 1×1 convolution. The fully-supervised student branch follows the same procedure. Experimental results show that the fused feature $\check{F}_f^o$ is more discriminative than the visual feature F_f^o; see Table 3.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on Pascal VOC 2007 [8] and use COCO 2014 [25] as the supplementary public fully-supervised dataset. Pascal VOC dataset contains 20 categories with 24,640 objects in 9,963 images. COCO dataset contains 80 common object categories with 2.5 million labeled instances in 328k images. The categories in Pascal VOC are totally included by COCO dataset. To better simulate different category overlapping conditions, we process COCO dataset to be COCO-$(80 - n)^*$ by deleting the annotations of the randomly picked n overlapping categories, where $(80 - n)$ is the number of the remaining categories in COCO dataset. Unlike [43], we only remove the *annotations* but not the *images* of the overlapping categories as objects of the non-overlapping categories may appear in the source fully-supervised dataset. We use COCO-$(80 - n)^*$ to denote COCO dataset removing all *images* of the overlapping categories.

We further validate on KITTI dataset [9]. It contains 7 categories related to self-driving excluding *misc* and *dontcare*, with only *car* and *truck* in the label set of COCO dataset. The categories have some logical relationships: *pedestrain* and *person_sitting* in KITTI are sub-categories of *person* category in COCO; *cyclist* is the combination of *person* and *bicycle*. We use these relationships to build hand-crafted inter-dataset edges between two datasets.

Evaluation. We use mean average precision (mAP) to evaluate the detection performance over categories, and CorLoc [7] to measure the localization accuracy.

Implementation details. We use Faster R-CNN[27] as our base model and follow the same settings. We use vgg16[31] pretrained on ImageNet as the feature extraction backbone. For Pascal VOC dataset, we use the similarity inter-dataset edges to build the inter-dataset graph. The dimension of the input semantic features is 300. The semantic graph convolutional network applies two graph convolution layers with the hidden feature dimensions 2048 and 4096.

During training, we do image augmentation by horizontally flipping, randomly cropping, and randomly resizing the image between 0.5 and 2. The initial learning rate is

Method	Source Dataset	mAP(%)	CorLoc(%)
pure WSOD:			
WSDDN Ensemble [2]	-	39.3	58.0
OICR-Ens+FR [34]	-	47.0	64.3
PCL-Ens+FR [32]	-	48.8	66.6
WSOD2 [42]	-	53.6	69.5
CASD [13]	-	56.8	70.4
WSOD with transfer:			
MSD-Ens [24]	ILSVRC2013-180	51.1	66.8
OICR+UBBR [17]	COCO-60*	52.0	47.6
Boosting [43]	COCO-60	55.2	72.4
Boosting* [43]	COCO-60*	57.8	73.6
Ours:			
CaT_0	COCO-60	58.0	73.8
CaT_0^*	COCO-60*	59.2	75.9
CaT_5	COCO-65	**63.5**	**80.3**
FSOD:			
Faster R-CNN [27]	-	69.6	94.3

Table 1. Comparison of our method on Pascal VOC 2007 test set to FSOD method and the state-of-the-art WSOD methods in terms of mAP (%) and CorLoc (%). Our method achieves better performance than the previous WSOD methods.

Method	Sim	HC	mAP(%)	CorLoc(%)
WSDDN	×	×	64.5	26.7
CaT	✓	×	83.8	56.8
CaT	×	✓	80.6	51.5
CaT	✓	✓	79.5	58.9

Table 2. KITTI test 2D detection mAP (%) and CorLoc (%) with different inter-dataset edges, with 'Sim' for the similarity inter-dataset edges, and 'HC' for the hand-crafted inter-dataset edges. Our model has good domain adaptation ability and the similarity inter-dataset edges can lead to better performance.

0.002. We adopt an SGD optimizer with 0.9 momentum and 0.0001 weight decay. The teacher branch weights do not participate in back propagation, and are updated by EMA with EMA decay $\alpha = 0.999$. The weights to balance losses are $\lambda_{full} = 0.5$, $\lambda_{cons} = 1.0$. Our method is implemented based on Pytorch Faster R-CNN framework[41]. All the models are trained on 4 NVIDIA 1080 Ti GPU. The batch size is 8.

4.2. Comparison with State-of-the-Art Methods

Results on Pascal VOC 2007. We compare our method with the state-of-the-art approaches on Pascal VOC 2007 dataset, including (1) *pure WSOD methods*: WSDDN [2], OICR [34], PCL [32], WSOD2 [42], CASD [13]. These approaches are trained without any supplementary data. (2) *WSOD methods with knowledge transfer*: MSD [24], OICR+UBBR [17], Boosting[43]. These methods transfer knowledge from an additional source dataset with bounding box annotations. As for the source dataset, the COCO-$(80 - n)$ dataset removes the *annotations* and the COCO-$(80 - n)^*$ dataset removes the *images* of the n overlapping categories in the original COCO dataset; ILSVRC2013-

180* is the ILSVRC2013 dataset removing all the images of the 20 categories contained by the Pascal VOC dataset. (3) fully-supervised method: we also compare our method with Faster R-CNN [27] trained with fully annotations on the weakly-supervised dataset, as in [43, 20, 32].

Table 1 compares our approach with the previous state-of-the-art approaches on the Pascal VOC test set in terms of mAP and CorLoc. We use CaT_n to denote our method trained with the fully-supervised dataset COCO-$(80 - n)$, and use CaT_n^* to denote our method trained with the fully-supervised dataset COCO-$(80 - n)^*$.

We first evaluate our method using COCO-60 or COCO-60* as the source fully-supervised dataset. For pure WSOD approaches, CaT_0 outperforms the state-of-the-art method CASD by 1.2% mAP and 3.4% CorLoc; and CaT_0 improves the WSDDN baseline by 18.7% mAP and 15.8% CorLoc. For WSOD approaches with transfer, CaT_0 outperforms the state-of-the-art method Boosting [43] by 2.8% mAP and 1.4% CorLoc with COCO-60 as source; and CaT_0^* outperforms Boosting* [43] by 1.4% mAP and 2.3% CorLoc with COCO-60* as source. CaT_0^* using COCO-60* as source achieves better performance than CaT_0 using COCO-60 as source, as the regions with annotation removed in COCO-60 are treated as background while training, which will reduce the recall rate of CaT_0.

We then evaluate our method using COCO-65 as the source fully-supervised dataset. With only 5 overlapping categories between COCO-65 dataset and Pascal VOC dataset, CaT_5 gets an increase of 6.7% mAP and 9.9% CorLoc over the state-of-the-art pure WSOD method CASD, and significantly reduces the mAP gap between weakly supervised object detection method and the fully-supervised method, like Faster R-CNN![1]

Results on KITTI. We also validate the domain adaptation ability of our method. We use a domain general dataset, COCO, as the source fully-supervised dataset and do weakly-supervised object detection on a domain specific dataset, KITTI. For the semantic graph construction, we can build similarity inter-dataset edges as well as hand-crafted inter-dataset edges according to the logical relationships between the categories in KITTI and COCO datasets. Table 2 compares the performance of our methods using different inter-dataset edges and a WSDDN baseline. We see that i) our method significantly outperforms our baseline, WSDDN, validating the domain adaptation ability of our method; and ii) our method achieves the best mAP 80.6% using the similarity inter-dataset edges and the best CorLoc 58.9% using both the similarity and the hand-crafted inter-dataset edges, which suggests the similarity inter-dataset edges can lead to better category transfer.

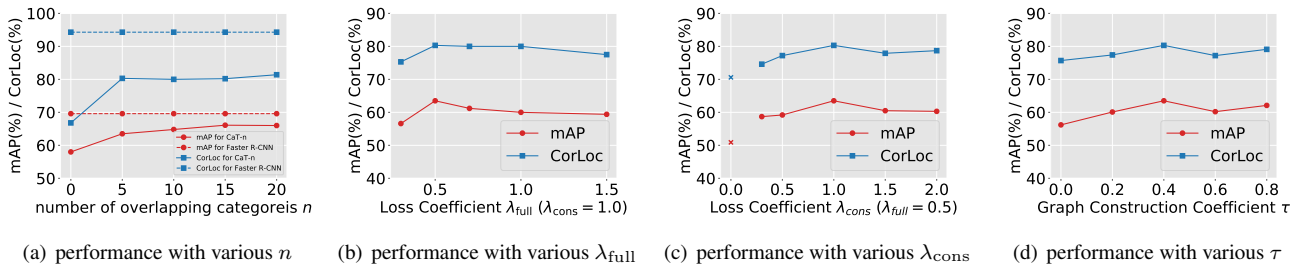[1]The qualitative results are shown in the supplementary.

(a) performance with various n | (b) performance with various λ_{full} | (c) performance with various λ_{cons} | (d) performance with various τ

Figure 7. Ablation study of the number of overlapping category n, the balance weight for the fully-supervised loss λ_{full}, the balance weight for the consistency loss λ_{cons}, and the graph construction threshold τ. The performance of our method improves with the increase of n; and our method is robust to λ_{full}, λ_{cons}, and τ.

	DSMT	SGCN	aero	bike	bird	boat	bottle	bus	car	**cat**	chair	cow	**table**	**dog**	horse	**mbike**	person	**plant**	sheep	sofa	train	tv	mAP(%) / CorLoc (%)
(A)	×	×	29.1	49.3	31.0	25.5	25.2	40.4	63.2	34.8	26.5	39.0	0.2	29.7	41.1	53.7	33.9	27.5	36.5	33.5	32.5	51.3	35.2
			58.5	73.2	51.6	44.3	42.9	61.8	78.3	48.8	40.0	70.1	0.0	51.5	62.4	72.5	57.3	51.6	71.4	51.3	52.1	70.98	55.5
(B)	✓	×	72.6	49.2	54.9	40.2	49.6	79.0	80.4	69.1	43.3	75.5	30.6	65.9	71.1	59.3	67.1	31.5	68.8	59.7	70.8	66.9	60.3
			87.8	69.6	78.2	64.2	71.3	90.2	92.3	82.2	60.2	92.9	58.7	81.8	84.6	77.3	79.6	58.3	91.8	73.8	80.3	82.0	77.8
(C)	✓	✓	74.0	70.7	60.0	31.1	50.0	75.9	82.0	70.7	32.8	74.3	69.5	70.2	69.5	77.0	37.5	45.8	67.0	61.1	72.4	68.0	63.0
			87.3	84.4	80.3	59.1	71.3	89.1	91.7	80.7	52.5	92.9	86.2	84.5	85.0	92.3	62.6	70.1	89.8	73.5	81.9	84.7	80.0

Table 3. The effectiveness of the double-supervision means teacher network (DSMT) and the semantic graph convolutional network (SGCN) in terms of mAP (%) and CorLoc (%). The overlapping categories between COCO-65 and Pascal VOC are bold. Both DSMT and SGCN are effective for detection performance and localization accuracy.

4.3. Ablation Study

Number of overlapping categories. Figure 7 (a) shows the effects of the number of the overlapping categories n between the fully-supervised and the weakly-supervised datasets. We see that i) the performance improves with the increase of n. The reason is that more bounding box and category training data for the specific overlapping categories improve the detector; and 2) our method outperforms Faster R-CNN trained on only COCO-$(80-n)$ dataset with n overlapping categories, validating the domain adaptation ability of our method. For the following ablation studies, we fix the number of overlapping categories as 5.

Network Components. To validate the effectiveness of the proposed double-supervised mean teacher network (DSMT) and the proposed semantic graph convolutional network (SGCN), we compare three network settings: (A) a pure weakly-supervised object detector using a traditional mean teacher structure without using any proposed network, which consists of the Faster R-CNN backbone, a WSDDN like weakly-supervised student branch, and a teacher branch which has the same architecture with the student branch; (B) the baseline network with double-supervision mean teacher network (DSMT); and (C) the baseline network with double-supervision mean teacher network (DSMT) and the semantic graph convolutional network (SGCN). Table 3 compares the detection performances of these three settings. We see that i) Setting (B) outperforms (A) by over 25% and 20% in terms of the mAP and CorLoc, respectively, indicating the huge benefits

brought by DSMT; and ii) Setting (C) further improves (B), validating the effectiveness of the SGCN.

Hyperparameters. Figure 7 (b) - (c) shows the effects of hyperparameters in the lost function, λ_{full}, λ_{cons} and λ_{cons}, respectively. We see that i) too small or too large λ_{full} can both causes the performance drop, and our method achieves the best performance with $\lambda_{\text{full}} = 0.5$; and ii) A larger λ_{cons} means a stronger regularization on L_{cons}. The method achieves the best performance with $\lambda_{\text{full}} = 1.0$. Plot (d) shows the effect of the intra-dataset graph construction threshold, τ. We see that a larger τ leads to a sparser intra-datset graph and $\tau = 0.4$ models a proper density of the graph to achieve the best performance.

5. Conclusion

This paper studies the weakly-supervised object detection with fully-supervised knowledge transfer. We specifically focus on category transfer; that is, leveraging the category information in a fully-supervised dataset to improve the discriminative ability of the detector. For overlapping category transfer, we propose a double-supervision mean teacher network to gather common category information and bridge the domain gap between two datasets. For non-overlapping category transfer, we propose a semantic graph convolutional network to gather semantic features by propagating semantic information between correlated categories. The advantages of the proposed networks lead to a new state-of-the-art on PASCAL VOC 2007.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, pages 397–414. Springer, 2018. 3

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854. IEEE Computer Society, 2016. 1, 2, 4, 5, 7

[3] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763. IEEE Computer Society, 2017. 3

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, pages 522–531. IEEE, 2019. 3

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5177–5186. Computer Vision Foundation / IEEE, 2019. 3

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

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

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 6

[9] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. 6

[10] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *CoRR*, abs/2012.07177, 2020. 1

[11] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. LSDA: large scale detection through adaptation. In *NIPS*, pages 3536–3544, 2014. 1, 2, 3

[12] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross B. Girshick. Learning to segment every thing. In *CVPR*, pages 4233–4241. IEEE Computer Society, 2018. 3

[13] Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *NeurIPS*, 2020. 1, 3, 7

[14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 6

[15] Jason Kuen, Federico Perazzi, Zhe L. Lin, Jianming Zhang, and Yap-Peng Tan. Scaling object detection by transferring classification weights. In *ICCV*, pages 6043–6052. IEEE, 2019. 1, 2, 3

[16] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 765–781. Springer, 2018. 1

[17] Seungkwan Lee, Suha Kwak, and Minsu Cho. Universal bounding box regression and its applications. In *ACCV (6)*, volume 11366 of *Lecture Notes in Computer Science*, pages 373–387. Springer, 2018. 7

[18] Seungkwan Lee, Suha Kwak, and Minsu Cho. Universal bounding box regression and its applications. *CoRR*, abs/1904.06805, 2019. 1, 3

[19] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520. IEEE Computer Society, 2016. 3

[20] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520. IEEE Computer Society, 2016. 7

[21] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3595–3603, 2019. 3

[22] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2020. 3

[23] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):639–653, 2019. 3

[24] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):639–653, 2019. 7

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 6

[26] Zheng-Chun Lin, Dian Li, Yun-Zhi Jiang, Jing Wang, and Chao Zhang. Yolov3: Face detection in complex environments. *Int. J. Comput. Intell. Syst.*, 13(1):1153–1160, 2020. 1

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 2, 3, 4, 5, 6, 7

[28] 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 10595–10604. IEEE, 2020. 3

[29] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *CVPR*, pages 4315–4324. IEEE Computer Society, 2015. 1, 3

[30] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *ICCV*, pages 3401–3410. IEEE Computer Society, 2017. 3

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[32] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan L. Yuille. PCL: proposal cluster learning for weakly supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(1):176–191, 2020. 1, 2, 7

[33] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, pages 3059–3067. IEEE Computer Society, 2017. 1, 3

[34] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3059–3067, 2017. 2, 7

[35] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert J. Gaizauskas, and Liming Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):3045–3058, 2018. 1, 2, 3

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, pages 1195–1204, 2017. 2, 3, 4

[37] Jasper R. R. Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *CVPR*, pages 1101–1110. IEEE Computer Society, 2018. 1, 2, 3

[38] Ying Wei, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5072–5081. PMLR, 2018. 3

[39] Hang Xu, Linpu Fang, Xiaodan Liang, Wenxiong Kang, and Zhenguo Li. Universal-rcnn: Universal object detector via transferable graph R-CNN. In *AAAI*, pages 12492–12499. AAAI Press, 2020. 3

[40] Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *CVPR*, pages 6419–6428. Computer Vision Foundation / IEEE, 2019. 3

[41] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017. 7

[42] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *ICCV*, pages 8291–8299. IEEE, 2019. 3, 7

[43] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *ECCV (26)*, volume 12371

of *Lecture Notes in Computer Science*, pages 615–631. Springer, 2020. 1, 3, 6, 7

[44] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *CoRR*, abs/1905.05055, 2019. 1