Missing Labels in Object Detection

Mengmeng Xu[1] mengmeng.xu@kaust.edu.sa

Yancheng Bai[2] baiyancheng20@gmail.com

Bernard Ghanem[1] bernard.ghanem@kaust.edu.sa

[1] Visual Computing Center, King Abdullah University of Science and Technology (KAUST)
[2] Institute of Software, Chinese Academy of Sciences (CAS)

Abstract

Object detection is a fundamental problem in computer vision. Impressive results have been achieved on large-scale detection benchmarks by fully-supervised object detection (FSOD) methods. However, FSOD performance is highly affected by the quality of annotations available in training. Furthermore, FSOD approaches require tremendous instance-level annotations, which are time-consuming to collect. In contrast, weakly supervised object detection (WSOD) exploits easily-collected image-level labels while it suffers from relatively inferior detection performance. In this paper, we study the effect of missing annotations on FSOD methods and analyze approaches to train an object detector from a hybrid dataset, where both instance-level and image-level labels are employed. Extensive experiments on the challenging PASCAL VOC 2007 and 2012 benchmarks strongly demonstrate the effectiveness of our method, which gives a trade-off between collecting fewer annotations and building a more accurate object detector. Our method is also a strong baseline bridging the wide gap between FSOD and WSOD performances.

1. Introduction

Object detection is a fundamental and essential problem yet to be deciphered in computer vision. Impressive results have been achieved on large-scale detection benchmarks by fully-supervised object detection (FSOD) methods, especially with the convenience of deep convolutional neural networks (CNNs) [21, 15], whose success mainly benefits from the flexibility of deep learning models and an abundance of instance-level annotations in extensive datasets [32, 25]. However, annotating such large-scale datasets is expensive and time-consuming. More importantly, the performance of FSOD is profoundly affected by the quality of these annotations. For instance, imperfect bounding box annotations or missing annotations of objects in training images can have a drastic impact on FSOD performance. This will be the noteworthy focus of our paper.

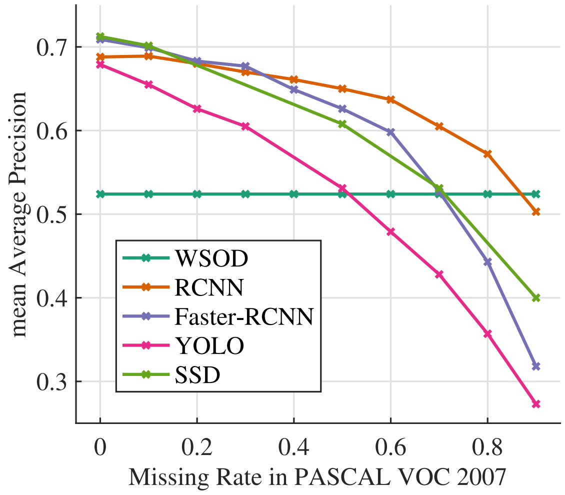

Figure 1: **The Mean Average Precision (mAP) for one WSOD detector and four classical FSOD detectors training under different instance-level missing label rates** ($M_r = 0 : 0.1 : 0.9$) **of the training dataset.** The performance of FSOD detectors is largely affected by the amount of missing annotations. These models are trained on VOC2007 train-val set and evaluated on the VOC2007 test dataset.

When collecting large-scale object detection datasets, the missing label problem (*i.e.* some instance-level bounding box annotations are missing in some images) does arise and it becomes more prevalent when the dataset grows in size (both in the number of training images and object classes). Fig. 1 exhibits the detection performance of a standard FSOD object detector at different instance-level **missing label rates** (M_r). This value shows the proportion of discarded annotations over all the annotations in the original dataset. Ranging from 0 to 0.4, FSOD performance decreases slightly. However, the performance drops significantly when M_r is larger than 0.5. Nevertheless, efforts to identify the effects of this problem on detector performance are not sufficient to produce considerable outcomes. As such, it becomes worthwhile to develop object detectors that can handle the missing label problem.

In this paper, we firstly investigate the robustness of current CNN based detectors trained on datasets with different missing label rates. After analyzing the limitations of these existing FSOD detectors and realizing that state-of-the-art WSOD methods register much lower detection performance, we propose a hybrid supervised learning framework for missing label object detection. This framework firstly uses a WSOD detector[41] as a teacher model to generate pseudo labels. Then these labels are merged with existing annotations to train a novel object detector. We compare the object prediction with the images label and introduce loss functions generally used in weakly supervised object detection. Also, we show that repeated refinement is an efficient way to improve the model performance.

To sum up, we make the following three *contributions* for the missing label object detection problem.

(1) We evaluate the robustness of mainstream FSOD methods to varying rates of missing labels, varying object categories (Sec. 4.2) and different object sizes (Sec. 4.3). We conclude that the performance of all these FSOD methods drops more significantly as the missing rate increases, thus, indicating that current FSOD detectors are less robust to missing annotations in the training dataset.

(2) We compares different setups in a novel teacher-student framework that combines image-level and instance-level information to train a robust end-to-end detection model. This framework inherits the advantages of both weakly- and fully-supervised detection methods while avoiding their drawbacks. Both the compared FSOD and WSOD methods carry out no better performance than our proposed detector on VOC2007 and VOC2012 with ten different missing rates (in Sec. 4.5).

(3) Our framework gives a trade-off between collecting large-scale fully-annotated dataset and training a better object detection model. Experiments in Sec. 4.4 reveals the highly practical value of our work to do object detection.

2. Related Work

2.1. Fully-Supervised Object Detection

With the development of deep learning, many CNN based methods have been proposed to solve the FSOD problem, such as Fast RCNN [13], Faster RCNN [38], SSD [26], YOLO2 [29], and many of their variants [1, 10, 14, 24, 5, 43, 2]. Faster RCNN [38] is a typical proposal based detection CNN, which balances both detection performance and computational efficiency. This method has become the *de facto* framework for fully-supervised object detection due to its plasticity and flexibility. YOLO2 [29], on the other hand, achieves real-time detection by predicting bounding boxes in a dense manner, specifically for each predefined region in the image. Fully-supervised methods have achieved impressive results in object detection. However, training them required the collection and curation of large-scale instance-level bounding-box annotations, which is expensive and time-consuming. As we pointed out in Fig. 1, the performance of FSOD detectors is largely affected by the amount of missing annotations. In our hybrid learning framework, we also take Faster-RCNN as our basic model, but any state-of-the-art detection model such as [20, 7, 27] can be used and compared in our study.

2.2. Weakly Supervised Object Detection

If there are no instance-level labels available in training, we can resort to training a weakly-supervised object detector. Most classical approaches treat Weakly Supervised Object Detection (WSOD) as a Multiple Instance Learning problem [33, 22, 16, 31, 3, 18, 40, 39, 28, 42]. For example, Bilen *et al*. [4] present a weakly supervised deep detection network (WSDDN), which selects positive samples by multiplying the score of recognition and detection, and updates the scores by comparing the predicted positive samples and image level annotation. Others focus on improving the optimization strategy in training. Tang *et al*. [35, 34] design an online instance classifier refinement (OICR) algorithm to refine the predicted object positions and alleviate the local-optimum problem that plagues WSDDN.

The different set-ups in out hybrid learning framework are inspired by weakly supervised learning methods. We combine WSDDN [4] with a fully-supervised detection network to generate the positive samples and consider instance level annotations in OICR algorithms [35] to enhance our missing label object detector.

2.3. Hybrid Supervision and Pseudo Labels

Hybrid supervised learning aims to use different level of supervision to train a detection model. This topic is drawing more attention in segmentation problem [23, 17], which requires expensive pixel level annotation. In object detection domain, people use high scoring predicted bounding boxes generated by a weakly supervised detector as pseudo labels [36, 11, 41, 9, 37, 8]. Pseudo label method has been recently used in a fully-supervised setup to compensate for the absence of all instance-level box annotations. The cascade detector in Diba's work [11] and OICR [35] both use pseudo object labels to train Fast-RCNN and achieve eminent WSOD performance. Mining pseudo labels can also increase the success of fully-supervised detectors. Zhang *et al*. [41] determine the most accurate bounding box using pseudo ground-truth excavation (PGE) and pseudo ground-truth adaption (PGA) algorithm from predictions.

Apply this method on the hybrid supervised learning domain, our innovation is to generate pseudo labels from different levels of annotations and update the generator in every training cycle.

Algorithm 1 Frequency based instance-level annotation sampling. We randomly drop M_r of the annotations of each category based on the number of instance level annotations and missing rate M_r.

```
Input: missing rate $M_r$, number of categories $M$, number
    of images $N$;
for each annotation $j$ in each image $i$ do
    find annotation category $k$;
    append annotations to $objLabels[k][i]$;
end for
for each category $k$ do
    sample $M_r$ of annotations from $objLabels[k][:]$
    (without replacement )
    remove the sampled data from $objLabels[k][:]$;
end for
save labeled image index as $ml\_ImageList$;
save $objLabels[k][:]$ in PASCAL VOC standard as
    $annotation\_m_r$;
Output: $annotation\_m_r$, $ml\_ImageList$.
```

3. Approach

In this section, we give a comprehensive description of our hybrid supervised learning framework for the missing instance-level label object detection problem. To our best knowledge, no attention has been paid on this problem, and there is no standard instance-level missing object detection training dataset. Therefore, we firstly describe how to modify the standard detection benchmarks into such kind of dataset under different instance-level missing rates. Then, we present each component of our proposed detector in detail. A brief overview of our framework is shown in Fig. 2.

3.1. Missing Label Datasets

Alg. 1 presents how to constructs the instance-level missing label dataset from any fully supervised dataset for a given missing rate m_r. Firstly, we collect all instance-level labels for each category in all images in the dataset. And then we randomly drop the instance-level labels for each category with the ratio m_r. Meanwhile, we also record images without any instance-level labels after dropping, which will not be sampled when training the FSOD models. The reason is that no positive training examples exist in these images, which will make the detector bias to the background and degrade its performance. However, our detector can mine valuable information from these images to boost performance. Note that, in this paper, the missing label object detection dataset is based on PASCAL VOC2007/2012.

3.2. Basic Hybrid Supervised Architecture

The missing label dataset is hybrid-supervised by both instance- and image-level annotations. We use a **teacher-student learning** architecture to solve the hybrid supervised learning problem. Our teacher detector is a decent object detection model that forward-passes an image and gives pseudo label for object categories and localization predictions, and the student network can be an off-the-shell object detector such as[30, 29], or an adapted one for the missing label background (Sec. 3.3).

In the learning process, the teacher detector is fixed and the student detection model is trained from both instance level labels and post-processed pseudo labels. We stack confident predictions from the teacher detector and the ground truth object labels, and apply Non-Maximum Suppression (NMS) for the object categories that appear in image level supervision. This process removes three type of predictions. The first type is the false alarms whose category contradicts with images labels. Also, the predictions which have a small positive confidence score is discarded to collect a high precision training data. The NMS operation also removes the unnecessary predictions which are similar to any instance labels on both classification side and localization side.

We also compared two different setups for the teacher model. (1) For the simple setup, this model is a decent object detector and is learned from only weakly annotations. It performs well at classification tasks for the most distinguishable part of the object. The student model can inherit the abilities to find the correct object classes from the teacher model, and also get improved on localization accuracy from the remained instance level annotations. (2) Another way to design the teacher model is to use the hybridly learned object detector. Each time we get a high-performance object detection model, it takes the place of teacher model and providing more accurate pseudo labels. When we update the teacher model, we reset the student model to the original weights.[1]

3.3. More Adaptions to Missing Label Training

We also adapt two main modules in WSOD field into our hybrid supervised learning framework. The first module is a Multiple Instance Detection (MID) network [4] which introduces loss function for the images-level labels. And the second Instance Classifier Refinement (ICR) network [35] further improve the localization accuracy from more reliable regression targets.

In the following discussion, we assume an image has both image-level labels $\boldsymbol{y}{=}[y_1, \ldots, y_C] \in \{0,1\}^C$ and instance-level labels $\boldsymbol{P}{=}[\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L] \in \mathbb{R}^{L\times 5}$. Here C denotes the number of object categories while L is the number of labelled objects in the image. In addition, it also has L' pseudo labels from the teacher model, denoted as $\boldsymbol{P}'{=}[\boldsymbol{p}'_1, \ldots, \boldsymbol{p}'_{L'}] \in \mathbb{R}^{L'\times 5}$.

[1] The VGG16 weights are pretrained from ImageNet, and the remained are initialized randomly.

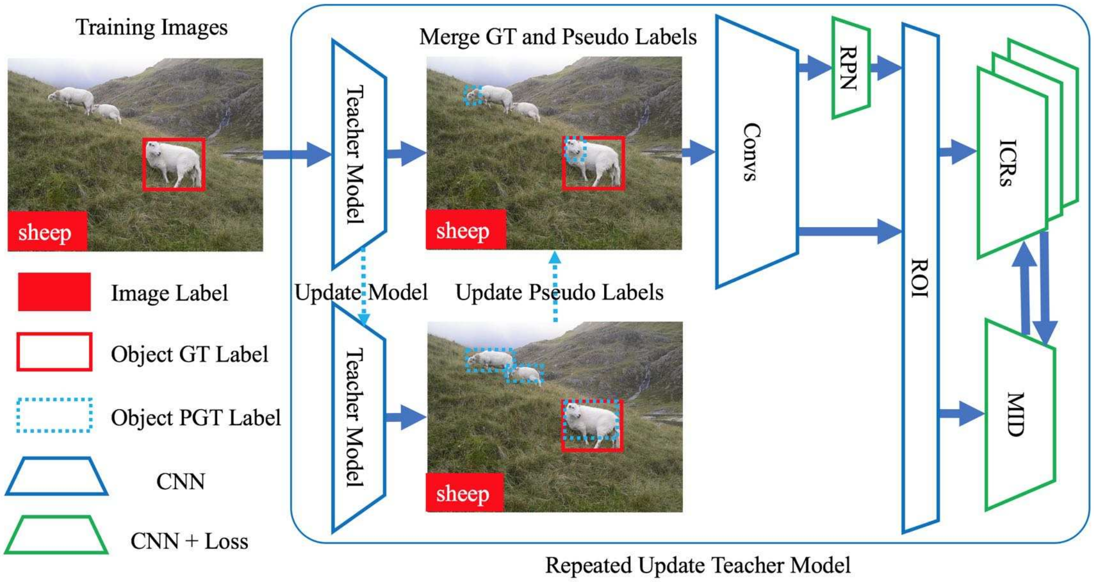

Figure 2: **Illustration of the proposed Hybrid Supervised Architecture with Adaptions.** Given an image collection with image-level labels and partial instance-level labels, we firstly use W2F [41] to generate pseudo label information (*e.g.* blue rectangles in the top image.) Then we combine the ground-truth object bounding boxes (*e.g.* red rectangles in the top image), pseudo information and image-level labels to train an end-to-end object detection network. The detection network can have three modules: RPN, MID and ICRs. RPN provides proposals as in most two stage methods. MID predict region detection results, and it is supervised by both level of annotations. ICRs further refine the learning target for each proposals. When the detection network is ready, it takes the place of the teacher models and generates more accurate pseudo label (*e.g.* blue bounding boxes in the bottom image) , the updated instance-level pseudo bounding boxes are utilized to retrain the model.

Learn from Image- and Instance-Level Labels. We modify the training process of the detection model such that it can learn from both image level and instance level labels.

In weakly supervised learning, n object proposals are used to extract n feature vectors. These feature vectors pass through classification and localization sub-networks and get two $C \times |\boldsymbol{R}|$ matrices $X^{\mathcal{C}}$ and $X^{\mathcal{D}}$, respectively. Inspired by [4], the scores of each proposal (subject to C classes) can be presented as the element-wise production of the softmax of the two matrices. $X^{\mathcal{R}} = \sigma_{cls}(X^{\mathcal{C}}) \odot \sigma_{det}(X^{\mathcal{D}})$. Given an image, the c-th class prediction score $[\boldsymbol{p}]_c \in \mathbb{R}$ can be obtained by summation of $X^{\mathcal{R}}$ for all proposals, as Eq. 1 [2]. Cross entropy loss is used to regress the module.

$$[\boldsymbol{p}]_c = \sum_{r=1}^{|R|} [\sigma_{cls}(X^{\mathcal{C}}) \odot \sigma_{det}(X^{\mathcal{D}})]_{c,r} \tag{1}$$

We also changed the training process on the proposal classification and object boundary regression to learn from ground-truth and pseudo instance labels differently. A positive proposal as the regions which have a high IOU with any (ground truth or pseudo) instance label, while a negative proposal must has a small IOU with either instance labels. Moreover, we only regress the proposal boundaries if it has a high IOU with ground truth instance label. In another words, we only use ground truth bounding boxes to regress locations, while avoiding pseudo label giving inaccurate bounding boxes. During training, the modified MID loss function can be formulated in Eq. 2:

$$\begin{aligned} Loss_{MID} = & \sum_{c=1}^{C} L_{lab}(y_c, p_c) + \\ & \frac{\alpha}{N_{cls}} \sum_{\boldsymbol{p^*} \in \hat{\boldsymbol{P}}} L_{cls}(\boldsymbol{p}, \boldsymbol{p^*}) + \\ & \frac{\beta}{N_{reg}} \sum_{\boldsymbol{p^*} \in \boldsymbol{P}} L_{reg}(\boldsymbol{p}, \boldsymbol{p^*}). \end{aligned} \tag{2}$$

Here, $\hat{\boldsymbol{P}} = \boldsymbol{P} \cup \boldsymbol{P'}$ is the union of true and pseudo instance labels. Both L_{lab} and L_{cls} are cross-entropy losses and L_{reg} is a smooth l_1 loss. N_{cls} and N_{reg} are the number of proposals used in classification and regression, respectively.

Refinement Layers The key idea of ICR is to integrate the basic detection network and the multi-stage instance-level classifier into a single network. We set-up this module almost the same as in Tang's work[35]. For the k-th refining subbranches in ICR, each of them classifies r-th proposal as $\boldsymbol{x}_r^{(k)} \in \mathbb{R}^{C+1}$. The label $\boldsymbol{y}_r^{(k)}$ for r-th proposal in k-th subbranch is the most confident predictions from $k-1$-th subbranch. Differently, if the instance label exist, it is also the training target for all of the k branches.

For each subbranch in ICR, the weighted cross-entropy

[2] the subscripts are used to show the element of a vector or a matrix.

loss (classification loss) is used as in Eq.3:

$$Loss_{ICR} = \frac{1}{|R|} \sum_{k=1}^{K} \sum_{r=1}^{|R|} L_{cls}(\boldsymbol{x}_r^{(k)}, \boldsymbol{y}_r^{(k)}, \boldsymbol{w}_r^{(k)}), \qquad (3)$$

where $\boldsymbol{w}_r^{(k)}$ denotes the confidence vector, $\boldsymbol{y}_r^{(k)}$ is training target, and K is the number of refinement layers.

Finally, we train our end-to-end student model by combining the loss functions from the above two modified modules and Region Proposal Network (RPN), as in Eq.(4).

$$Loss_{Total} = Loss_{RPN} + Loss_{MID} + Loss_{ICR}. \qquad (4)$$

$Loss_{RPN}$ is a regular RPN loss as in [30].

4. Experiments and Analysis

In this section, first, we experimentally test the classical FSOD detectors under different instance-level label missing rates, which demonstrates the weaknesses of current FSOD methods. Then, we verify the effectiveness of each component of our proposed detector, which is designed to deal with the instance-level missing problem. Finally, we compare the proposed method with other typical detectors on the public detection benchmark (PASCAL VOC datasets) and present some qualitative results.

4.1. Experimental Setup

Datasets and Evaluation Metrics We construct the missing label datasets based on COCO [25] and PASCAL VOC datasets [12]. For the COCO[25] experiments, we only consider four typical classes over 80 classes of object labels: dog, table, book and cow. To investigate the robustness of current detectors, we generate the missing label training set with missing rates ranging from 0.0 to 0.9 with step size 0.1 on those four categories, and also collect the object labels corresponding to the other 76 classes. After training a detector with the ability to detect 80 classes of objects, the Average Precision is reported for the typical four classes to show model performance. Also, we use PASCAL VOC datasets [12] to do more analysis on the missing label problem and investigate the possible solutions. PASCAL VOC 2007 and 2012 comprise of 9963 images and 22531 images respectively, which both include 20 categories of objects. To generate dataset with missing labels, we process the data for all the 20 categories as in COCO. We use VOC2007 and VOC2012 train-val sets to train our models, and evaluation is on the corresponding test datasets. Afterwards, the mean average precision (mAP) is utilized to evaluate the performance of the detectors.

Implementation Details Our framework employs the VGG16 model pre-trained on ImageNet [32] as the backbone network. In our training setting, the total number of iterations is set to $70k$ for VOC2007 and $80k$ for VOC2012, and the learning rate is 0.001 for the first $40k$ iterations and 0.0001 in the remaining iterations. Grounded on Chen's work[6] in Tensorflow, we design the hybrid learning part that will be publicly accessible later.

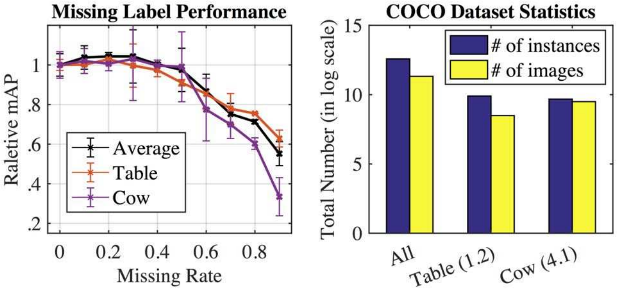

Figure 3: **Effect of the missing rate of instance-level labels on detection performance.** (*left*) shows the average precision for four COCO[25] classes with different missing label rates. The results are given by Faster-RCNN[6] with Mr from 0.0 to 0.9 by step 0.1. (*right*) plots statistics the instance number and image number of the four classes. Number are shown in *log* scale. Parentheses after each class gives the number of instances per image.

More specifically, we set three layers on the ICR network (*i.e.* K=3). The negative example of RCNN is the predictions which have Intersection Over Union (IOU) of ground-truth or pseudo instance-level label between 0.1 and 0.5.

4.2. The Effect of Missing Instance-Level Labels

To verify the robustness of current FSOD methods, we re-train and evaluate four typical FSOD models on the missing label dataset under different missing ratio: RCNN, Faster-RCNN, YOLO and SSD. Fig. 1 displays their mean Average Precision (mAP) on the PASCAL VOC2007 test sets. The performance of all FSOD methods drops significantly as the missing rate increases, which demonstrates that the performance of FSOD techniques is considerably affected by the quality of the training set. More surprisingly, the performances of Faster-RCNN, YOLO and SSD is inferior to the weakly-supervised method[41] when the missing rate M_r is higher than 0.7. From the experiments, we can conclude that current FSOD detectors are very sensitive to the quality of the training dataset.

As discussed above, different models are tested on our dataset, and we also run Faster-RCNN in another COCO dataset. Fig. 3 shows the same AP pattern when M_r is increasing. We acquire more interesting discoveries in the larger dataset investigation. (1) A small M_r can hardly affect model performance. In COCO dataset, we do not observe a clear decrease when $M_r < 0.4$. (2) Object class that appears as a crowd is more sensitive to M_r. AP for *cow* class falls faster than *table* class, and there are always multiple cows in an image but with only one table. We expect more emphasis laid on this phenomena by researchers.

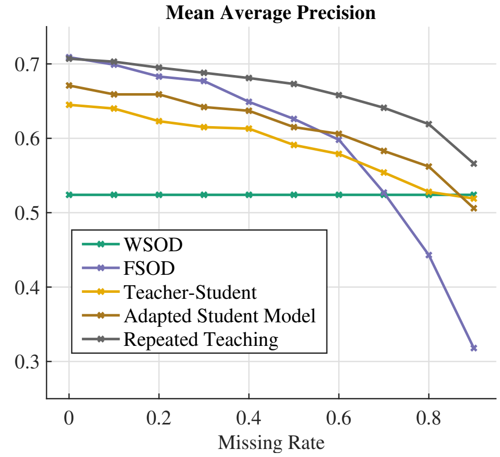

Figure 4: **Model performance when we use Teacher-Student Model, Student Model Adaption and Repeated Teaching among all the M_rs.** The final mAPs with full modules are higher than both FSOD and WSOD methods.

Cls.	FSOD	WSOD	T-S	MID*	ICR*	Repeated
aero	49.8	**63.5**	58.0	59.2	60.7	62.7
bike	66.4	70.1	75.9	74.6	72.2	**76.0**
bird	56.3	50.5	52.1	50.9	52.1	**58.9**
boat	37.7	31.9	37.6	38.8	41.0	**47.0**
bottle	33.0	14.4	32.7	31.9	33.7	**42.4**
bus	60.6	**72.0**	69.5	68.3	68.4	70.2
car	66.5	67.8	74.5	74.2	74.2	**75.8**
cat	72.4	73.7	76.1	75.3	75.6	**80.7**
chair	31.8	23.3	31.7	31.9	31.7	**42.0**
cow	46.5	53.4	65.9	66.5	68.0	**71.0**
table	46.2	49.4	47.0	43.8	47.5	**62.9**
dog	63.8	65.9	68.7	70.2	70.1	**76.8**
horse	71.1	57.2	73.5	74.5	75.7	**77.4**
mbike	64.0	67.2	67.9	70.4	73.7	**72.5**
person	62.3	27.6	56.2	55.2	56.7	**67.7**
plant	22.8	23.8	23.3	27.0	27.3	**31.7**
sheep	50.2	51.8	61.0	60.5	**62.5**	62.3
sofa	44.5	58.7	55.2	53.8	57.0	**62.9**
train	56.3	64.0	61.6	64.5	69.1	**73.4**
tv	51.1	62.3	62.4	65.7	65.0	**70.7**
mAP	52.7	52.4	57.5	57.9	59.1	**64.2**

Table 1: **Model performance with different modules when M_r is fixed to** 0.7. When we sequentially add modified MID, add modified ICR and use Repeated "Teaching" on the basic hybrid supervised architecture (S-T), we get increasing AP on most of the classes. The final APs with full modules are higher than both FSOD and WSOD methods.

4.3. Normal Object *vs*. Small Object

Since small objects are often missed, therefore, we investigate the effect of missing small instance-level labels on the detector's performance. To build this kind of dataset, (1) we compute the mean area (width × height) of all instances in each category in PASCAL VOC 2007; (2) for an instance whose area is smaller than the average, we discard its bounding box and only keep its image label. By doing this, around 30% percent of bounding boxes are removed and only large object instances remain.

Table 2: **The affect of missing small scale instance-level objects.** We compare both regular FSOD model and our proposed method in three dataset: fully-labeled dataset (Fully), missing label dataset at $M_r = 0.3$ and missing small label dataset (Small). The last two datasets have similar amount of annotations.

	Fully	M_r@0.3	M_r@0.3	Small	Small
Model	FSOD	FSOD	Hybrid	FSOD	Hybrid
mAP	71.3%	68.3%	68.8%	45.3%	57.9%

Table 2 shows the performance of standard Faster-RCNN and our hybrid learning method on both normal missing label dataset ($M_r = 0.3$) and the small missing label object dataset (Small). From the table, we can see that the Faster-RCNN trained on small scale missing dataset shows significantly performance drop (45.3% *vs*. 71.3%) compared to the model trained with fully-annotated dataset. This demonstrates that Faster-RCNN are very sensitive to the small scale install-level object missing problem. Compared to Faster-RCNN, our proposed method registers nearly 12.6% improvement (57.9% *vs*. 45.3%), from which we can conclude that our method is more robust to the small scale object missing problem.

4.4. More Comparison to Other Methods

The proposed hybrid supervised method strives a balance between labelling images and obtaining a more accurate detection model. We compare this method with the mainstream FSOD and WSOD methods in Table 3 and 4.

Comparison at Different M_r Table 3 shows AP performance on the VOC 2007 test set. The central block of the table shows our results in five different missing rate from 0.1 to 0.9. Our method achieves outstanding performance from 61.9% to 70.7% on different missing rate, which is between our compared fully-supervised object detection methods and weakly supervised models. Our method has better robustness when missing rate is small. For example, if the missing rate varies from 0.1 to 0.3, model performance decreases 1.2 percent. However, if the M_r decreases from 0.9 to 0.7, our mAP increases 3.9 percentage. Bounding boxes information is efficiently used with high M_r.

Table 4 lists our performance in mAP on the PASCAL VOC 2012 test set. These models are trained on PASCAL VOC 2012 train-val set only. On the left side of the table, we compare our method with OICR [35] and W2F [41]. Since our approach is based on Faster-RCNN, we compare it with Fast-RCNN [13] and Faster-RCNN [30] on the right side. Our baseline method successfully bridges between the gap in performance between WSOD and FSOD. Due to the flexibility of our method, the student detector can be taken to be any state-of-the-art FSOD detector.

Comparison with WSOD The first block in Table 3 compares our model to WSOD methods. The reason for the

Table 3: **Average Precision of FSOD, Missing Label Object Detection and WSOD.** All of them are trained/tested on VOC2007 *train-val/test* dataset.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Jie, 2017 [18]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	03.0	53.6	24.7	43.6	48.4	65.8	06.6	18.8	**51.9**	43.6	53.6	62.4	41.7
Krishna, 2016 [19]	53.9	-	37.7	13.7	-	-	56.6	51.3	-	24.0	-	38.5	47.9	47.0	-	-	-	-	48.4	-	41.9
Tang, 2017 [35]	**65.5**	67.2	47.2	21.6	**22.1**	68.0	**68.5**	35.9	5.7	**63.1**	**49.5**	30.3	**64.7**	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
Zhang, 2017 [41]	63.5	**70.1**	**50.5**	**31.9**	14.4	**72.0**	67.8	**73.7**	**23.3**	53.4	49.4	**65.9**	57.2	**67.2**	**27.6**	**23.8**	51.8	58.7	**64.0**	62.3	**52.4**
$M_r = 0.9$	56.4	71.9	46.0	32.6	34.5	70.9	69.1	73.3	32.6	65.2	46.4	69.7	74.0	67.0	59.3	24.6	55.2	49.9	66.4	67.2	56.6
$M_r = 0.7$	65.9	73.4	59.0	51.7	42.1	70.3	74.2	78.6	40.9	72.1	57.9	76.6	78.8	72.5	70.0	30.9	63.6	61.8	73.3	68.2	64.1
$M_r = 0.5$	68.1	78.0	61.7	51.7	50.4	74.8	78.4	82.1	46.9	72.6	63.3	78.2	80.9	72.7	74.9	36.4	64.4	67.9	71.7	70.1	67.3
$M_r = 0.3$	67.1	78.4	66.2	53.2	54.2	76.7	80.0	81.9	47.7	75.5	63.9	81.2	82.3	75.2	76.7	38.8	69.5	61.8	73.4	71.9	68.8
$M_r = 0.1$	69.1	78.3	70.3	54.6	56.1	78.5	81.2	82.3	52.9	73.3	66.9	80.7	83.3	74.6	77.4	43.0	71.0	66.1	74.2	72.6	70.3
Liu, 2016 [26]	75.4	**82.3**	67.4	**61.6**	41.7	**80.9**	82.2	80.3	49.2	71.9	**68.6**	82.1	83.5	**80.4**	75.9	**46.6**	**69.6**	**73.4**	**82.0**	**70.1**	**71.2**
Chen, 2017 [6]	67.6	78.9	67.6	55.2	**56.9**	78.8	**85.2**	**83.9**	**49.8**	**81.9**	65.5	80.1	**84.4**	75.7	**77.6**	45.3	70.8	66.9	78.2	72.9	71.2
Redmon, 2016 [29]	73.4	77.6	65.2	55.0	42.4	76.9	77.3	80.5	45.4	69.4	72.6	76.5	80.1	77.0	72.3	42.9	63.3	64.8	78.7	66.6	67.9
Girshick, 2015 [13]	**77.4**	78.3	**68.6**	59.7	37.5	80.0	78.3	83.8	43.8	74	67.8	**82.9**	80.0	76.6	67.9	35.7	69.4	69.8	77.7	67.5	68.8

improvement is that we also append accurate instance-level labels in the training set and regress the object location. In fact, these weakly supervised detectors can give a reliable classification probability, but not a precise localization. WSOD only highlights the discriminative parts of objects (*e.g.* face from human, nose from dogs, etc.), since it does not have object boundary priors. When $M_r = 0.9$, only very few bounding boxes are seen in training, and yet our method improves upon W2F [41] by 4.2% in mAP. Clearly, training the regression part of the model using ground truth improves the localization accuracy.

Compare to FSOD Compared to the methods in the last block in Table 3, our performance boost mainly comes from two contributions. (1) We use image-level labels to predict objects from the training set in two steps. Firstly we predict instance-level labels from a well-trained detection model before feeding the image to the network. Secondly, the image labels are also applied to evaluate the current model prediction. (2) The missing instance-level labels confuse the model. A missing positive bounding box can be taken to be a negative sample in training. Our method marks these areas as positive samples and reduces the related loss.

4.5. Ablation Study

In this part, we first compares different methods discussed in Sec. 3.3 under different missing rates. Then, we study the effects of the adapted modules with $M_r = 0.7$ to validate the contribution of each modification. All models are trained on PASCAL VOC 2007 train-val set and tested on PASCAL VOC 2007 test set.

Basic Hybrid Supervised Architecture The light yellow curve in the plot of Fig. 4 shows the model performance from the teacher-student training framework with basic setup. Its accuracy is strongly depressed at low M_r because of the imprecise pseudo labels. Also, it is approximately the same as WSOD at 0.8 and 0.9 missing rates. The combination of ground truth and pseudo labels gives more information to train the object detector, but a regular Faster-RCNN could not learn from these priors appropriately.

More Adaptions to Missing Label Training To properly use ground truth, pseudo labels and image labels, we adapted MID and ICR modules into our student model. The table in Fig. 4 gives details when we cumulatively adding this two modules to the RCNN model with $M_r = 0.7$. Comparing the result given by Teacher-Student (T-S) training and MID branch, we can see that the performance does not increase too much. However, when we introduce the modified ICR module, the performance goes from 57.9% to 59.1%. The feature map is encouraged to generate similar predictions from strong overlapped proposals. The dark yellow curve of Repeated Teaching in the plot of Fig. 4 indicates the overall improvement using both branches. Compared to RCNN, the MID part less relies on pseudo labels, and uses image labels to produce more reliable predictions; it gives better reaction at most missing rates.

Repeated "Teaching" As described in Sec. 3.2, repeatedly upgrading teacher model further improves the overall performance. Our grey curve in the plot of Fig. 4 reaches the original Faster-RCNN record when no bounding box is missed and surpasses 4.2 percentage from W2F model at $M_r = 0.9$. The bottom table also shows that this method makes average precision for each class higher than before. The improvement contributes to both the maintained ground truth labels and the updated pseudo labels because unreliable pseudo object labels are frequently replaced by more accurate ones.

4.6. Qualitative Results

In Fig. 5, we illustrate some detection results generated by our framework and compare them to those from FSOD or WSOD models. The dataset missing rate is set to 0.7. Faster-RCNN trained on the missing label dataset will miss some objects, while W2F tends to highlight only parts of objects. Our network combines labels from different sources and makes full use of them to produce bounding boxes that are tight and accurately classified.

Table 4: **Mean Average Precision of FSOD, our method and WSOD.** Our baseline method successfully bridges between the gap in performance between WSOD and FSOD. All of them are train/test on VOC2012 *train-val/test* dataset.

WSOD		Missing Label Object Detection (M_r)									FSOD	
Tang[35]	Zhang[41]	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	Ren[30]	Girshick[13]
42.5	**47.8**	52.3	56.8	59.9	62.0	63.6	64.6	65.3	66.1	66.9	**67.0**	65.7

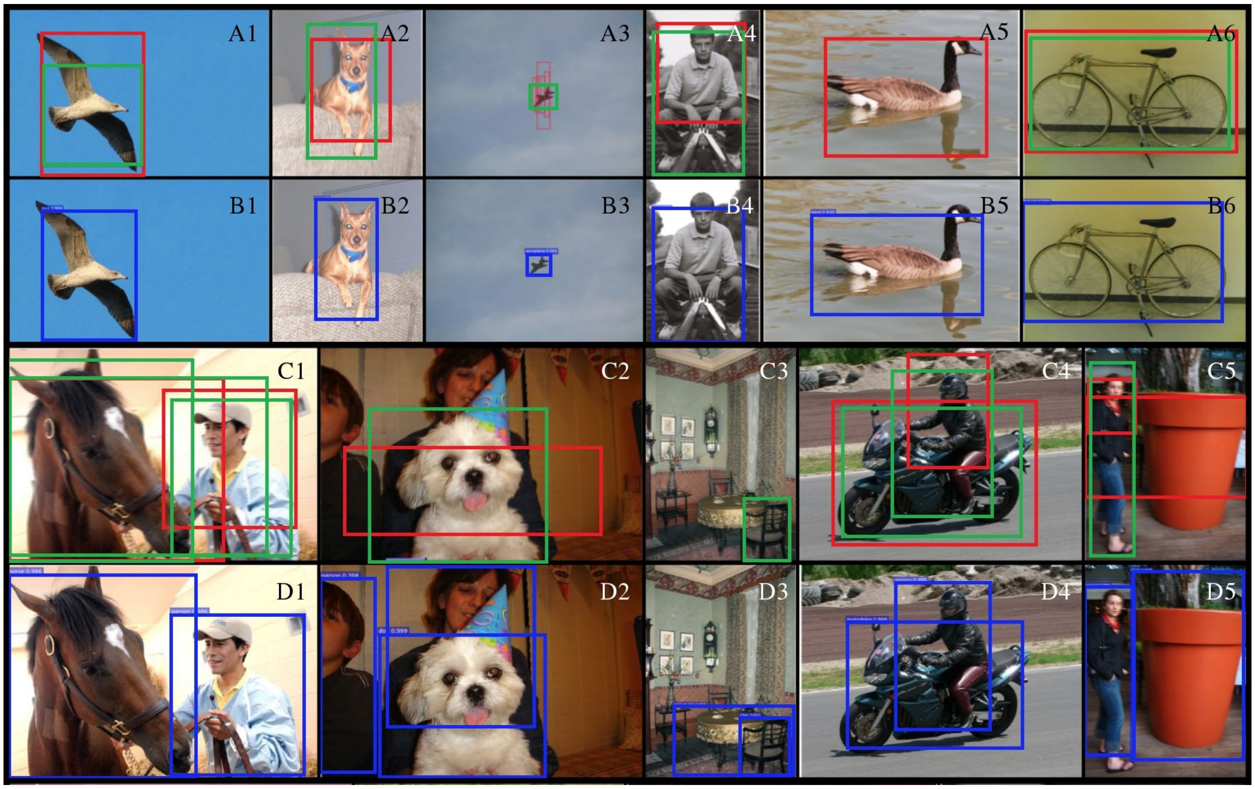

Figure 5: **Qualitative detection results of our method and two references (Faster-RCNN and W2F).** Blue bounding boxes indicate objects detected by our method, while red and green ones correspond to those detected by Faster-RCNN and W2F respectively. $M_r = 0.7$. Faster-RCNN trained on the missing label dataset will miss some objects, while W2F tends to highlight only parts of objects. Our network combines labels from different sources and makes full use of them to produce bounding boxes that are tight and accurately classified.

In Fig. 5 $A4$, FSOD predicts a tight green bounding box around the man, while WSOD only detects the upper body in the red bounding box. Our method in $B4$ is very similar to the FSOD result. In $C5$, FSOD failed to find the large flower pot, but WSOD can give a rough prediction. Our model locates the plant with a tight blue bounding box in $D5$. Line $3^{rd}, 4^{th}$ in Fig. 5 show images which have multiple objects from different classes. Combining both instance-level labels and image-level labels gives much better results than using each label independently. Moreover, we visualize some failed detection in $B5$ and $D5$, which indicate that there is still much room for improvement.

5. Conclusion

In this paper, we study the missing instance-level label problem in object detection and present a novel framework for this task. Our pipeline combines the advantages of fully-supervised and weakly-supervised learning. We first generate pseudo ground truth instance-level labels using weakly supervised object detection method, and then train an end-to-end missing label object detector. The pseudo ground truth object labels are upgraded once the detector reaches a better performance. Extensive experiments on PASCAL VOC 2007 and 2012 compared the improvement between fully and weakly supervised methods, and show that our method stands out among all of them at different levels of missing instance-level labels.

Acknowledgments This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research and by Natural Science Foundation of China, Grant No. 61603372.

References

[1] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. In *CVPR*, pages 21–30, 2018. 2

[2] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *ECCV*, 2018. 2

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, pages 1081–1089, 2015. 2

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 2, 3, 4

[5] Z. Cai and N. Vasconcelos. Cascade R-CNN: delving into high quality object detection. *CoRR*, abs/1712.00726, 2017. 2

[6] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017. 5, 7

[7] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017. 2

[9] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *CoRR*, abs/1503.00949, 2015. 2

[10] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016. 2

[11] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. 2

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 5

[13] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2, 6, 7, 8

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[16] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, pages 2883–2891, 2015. 2

[17] R. Hu, P. Dollár, K. He, T. Darrell, and R. B. Girshick. Learning to segment every thing. *CoRR*, abs/1711.10370, 2017. 2

[18] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. *arXiv preprint arXiv:1704.05188*, 2017. 2, 7

[19] K. Kumar Singh, F. Xiao, and Y. Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *CVPR*, pages 3548–3556, 2016. 7

[20] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. *CoRR*, abs/1808.01244, 2018. 2

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998. 1

[22] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520, 2016. 2

[23] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[24] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Featuref pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 2

[25] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1, 5

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2, 7

[27] J. Ma, A. Ming, Z. Huang, X. Wang, and Y. Zhou. Object-level proposals. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 2

[29] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 2, 3, 7

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 3, 5, 6, 8

[31] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *CVPR*, pages 4315–4324, 2015. 2

[32] 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. *IJCV*, 115(3):211–252, 2015. 1, 5

[33] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, pages 1637–1645, 2014. 2

[34] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille. PCL: proposal cluster learning for weakly supervised object detection. *CoRR*, abs/1807.03342, 2018. 2

[35] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. *arXiv preprint arXiv:1704.00138*, 2017. 2, 3, 4, 6, 7, 8

[36] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[37] Z. Wu, N. Bodla, B. Singh, M. Najibi, R. Chellappa, and L. S. Davis. Soft sampling for robust object detection. *CoRR*, abs/1806.06986, 2018. 2

[38] J. Yang, J. Lu, D. Batra, and D. Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017. 2

[39] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. *CoRR*, abs/1804.06962, 2018. 2

[40] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. *CoRR*, abs/1807.08902, 2018. 2

[41] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *IEEE CVPR*, 2018. 2, 4, 5, 6, 7, 8

[42] Y. Zhang, M. Ding, Y. Bai, M. Xu, and B. Ghanem. Beyond weakly-supervised: Pseudo ground truths mining for missing bounding-boxes object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2

[43] C. Zhu, Y. He, and M. Savvides. Feature selective anchor-free module for single-shot object detection. *CoRR*, abs/1903.00621, 2019. 2