EHSOD: CAM-Guided End-to-End Hybrid-Supervised Object Detection with Cascade Refinement

Linpu Fang,[1*] Hang Xu,[2*†] Zhili Liu,[2]
Sarah Parisot,[2] Zhenguo Li[2]
[1]South China University of Technology
[2]Huawei Noah's Ark Lab

Abstract

Object detectors trained on fully-annotated data currently yield state of the art performance but require expensive manual annotations. On the other hand, weakly-supervised detectors have much lower performance and cannot be used reliably in a realistic setting. In this paper, we study the hybrid-supervised object detection problem, aiming to train a high quality detector with only a limited amount of fully-annotated data and fully exploiting cheap data with image-level labels. State of the art methods typically propose an iterative approach, alternating between generating pseudo-labels and updating a detector. This paradigm requires careful manual hyper-parameter tuning for mining good pseudo labels at each round and is quite time-consuming. To address these issues, we present EHSOD, an end-to-end hybrid-supervised object detection system which can be trained in one shot on both fully and weakly-annotated data. Specifically, based on a two-stage detector, we proposed two modules to fully utilize the information from both kinds of labels: 1) CAM-RPN module aims at finding foreground proposals guided by a class activation heat-map; 2) hybrid-supervised cascade module further refines the bounding-box position and classification with the help of an auxiliary head compatible with image-level data. Extensive experiments demonstrate the effectiveness of the proposed method and it achieves comparable results on multiple object detection benchmarks with only 30% fully-annotated data, e.g. 37.5% mAP on COCO. We will release the code and the trained models.

Introduction

Recent advances of object detectors trained on large-scale datasets with instance-level annotations have shown promising results which predict both the class labels and the locations of objects in an image (Lin et al.; Redmon and Farhadi; Li et al.; Wang et al. 2017a; 2017; 2018; 2019). Those detectors are typically trained under full supervision which requires huge manual annotations of the objects' locations and categories for a large number of training images. However, it is expensive and time-consuming to recruit annotators for labeling the images. This becomes more severe when the number of categories is large. Thus, training a customized good object detector with a limited budget becomes a crucial problem in community. On the other hand, image-level labels that indicate the presence of an object can be acquired cheaply even in large amounts as such labels can be collected easily using an Internet crawler in an image search engine. Unfortunately, training solely with weakly-supervised methods yields models of subpar performance that cannot be reliably used in real life scenarios. As a result, we study the Hybrid Supervised Object Detection (HSOD) problem focusing on training a good detector with some fully-annotated data with bounding-box labels while fully utilizing weakly-annotated data with image-level labels. Note that this task is different from semi-supervised detection settings which usually focus on training with some existing categories with annotated labels and infering on some new categories.

The current state-of-the-art weakly-supervised/few-shot

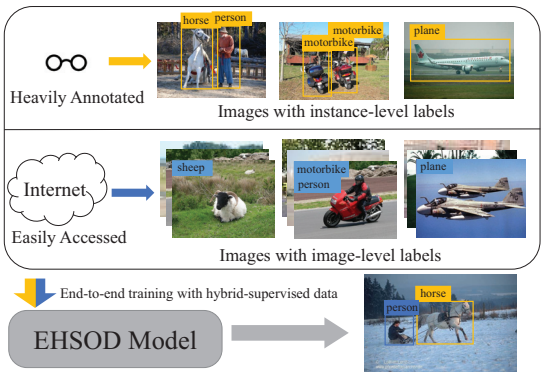

Figure 1: An illustration of our hybrid-supervised object detection task. Hybrid-supervised object detection problem focuses on training a good detector with a) limited fully-annotated data with bounding-box labels; b) fully utilizing cheap data with image-level labels. The conventional weakly-supervised methods usually require alternating iterative strategy while we aim to design an end-to-end hybrid-supervised object detection which can fully utilize both kinds of data.

*Both authors contributed equally to this work.
†Corresponding Author: xbjxh@live.com

object detection methods usually require generating pseudo labels and updating the detector iteratively. Most of the previous methods follow the Multiple Instance Learning (MIL) pipelines (Cinbis, Verbeek, and Schmid; Li et al.; Jie et al. 2016; 2016; 2017): images are decomposed into object proposals and the learning process iteratively alternates between re-localizing objects given the current detector and re-training the detector given the current object locations. During re-localization, different kinds of scoring systems are adopted to select the best proposal for an object class in each image. The recently proposed end-to-end PCL method requires mining instance labels for online instance classifier refinement (Tang et al.; Tang et al. 2017; 2018). Although some promising results have been obtained for Weakly Supervised Object Detector (WSOD) (Zhang et al.; Wan et al.; Wei et al.; Wan et al.; Li et al.; Kosugi, Yamasaki, and Aizawa 2018b; 2018; 2018; 2019; 2019; 2019), they are not comparable to those of fully supervised ones to meet the standard of deployment on the product.

Moreover, this kind of iterative training paradigm can easily get stuck in a local minimum, and are therefore unstable, which means it requires careful hyper-parameter tuning in mining good pseudo labels for each round of training. The weakly-supervised detector will easily fail when bad proposals are adding into training. When using a new dataset, many human efforts are further required to find a new set of good hyper-parameters. This situation is more severe when the number of categories and the dataset is large which hinders their usage in industry. To solve these issues, we seek to find a useful end-to-end object detection system which can be trained only once on both kinds of data, which nearly increase no extra hyper-parameters compared to the fully-supervised counterpart.

In this work, we present EHSOD for designing an end-to-end hybrid-supervised object detection network. The proposed method incorporates end-to-end joint training with both kinds of data and fully utilizes relevant information of the image-level labeled data to reach better performance. Specifically, based on a standard two-stage detector (Ren et al. 2015a), we proposed two modules to upgrade the existing system and manage to learn from both kinds of data. Stacked on an ImageNet pretrained ResNet, a CAM-RPN is proposed to localize the foreground proposals guided by a class activation heat-map (CAM) (Zhou et al. 2016). The CAM is jointly trained by the image-level labels and bounding-box annotations within the feature hierarchy of FPN (Lin et al. 2017a) and provides further information for Region Proposal Network (RPN) to select proposals. Furthermore, we design a hybrid-supervised cascade module to progressively refine the bounding-box position and improve classification accuracy. This module is a sequence of cascaded hybrid-supervised heads that contains a regular RCNN head as in Lin et al. and an auxiliary Multiple Instance Detection (MID) head as in Bilen and Vedaldi. The hybrid-supervised head is compatible with both kinds of data and can use image-level data to enhance the learning of classification.

The training of the resulting detection network follows a standard procedure of two-stage detection (He, Girshick, and Dollár 2018). The final detection system is greatly enhanced by abundant relevant image-level information and the performance is then boosted by sharing and distilling essential information across weakly/fully-supervised data.

Extensive experiments are conducted on the widely used detection benchmarks, including Pascal VOC (Everingham et al. 2015) and MS-COCO (Lin et al. 2014). The proposed method outperforms current state-of-the-art HSOD and WSOD methods, e.g. PCL (Tang et al. 2018) and BAOD (Pardo et al. 2019). We observe consistent performance gains on the base detection network FPN (Lin et al. 2017a) training with fully-annotated data. In particular, our method achieves comparable results with fully-supervised methods on multiple object detection benchmarks, e.g. 37.5% mAP on MS-COCO using only 30% fully-annotated data and 40.0% mAP with 50% MS-COCO fully-annotated data (compared to 37.2% with FPN trained with whole data).

To sum up, we make the following contributions:

- We are among the first to investigate the hybrid-supervised object detection problem focusing on training on a limited amount of fully-annotated images and a large amount of weakly labeled data.
- By fully exploiting the potential of both kinds information flow with different kinds of label, we develop EHSOD, a CAM-guided end-to-end hybrid-supervised object detection system with cascade refinement which can be trained in an one shot fashion.
- Extensive experiments demonstrate the effectiveness of the proposed method and achieve reliable results on multiple object detection benchmarks by only 30% fully-annotated data.

Related Work

Fully Supervised Object Detection (FSOD). Object detection is a core problem in computer vision. Significant progress has been made in recent years on FSOD task using CNN. Modern CNN based FSOD methods may be categorized in two groups: one-stage detection methods such as SSD and YOLO (Liu et al.; Redmon et al. 2016a; 2016) and two-stage detection methods such as Faster R-CNN and R-FCN (Dai et al.; Ren et al.; Xu et al.; Xu et al. 2016; 2015a; 2019b; 2019a). Although these methods have achieved satisfactoring detection results, the requirement of large-scale bounding-box annotations may hinder their usage in some budget-aware scenarios.

Weakly Supervised Object Detection (WSOD). WSOD aims at training a detector with only image-level labels, and received extensive attention from both academia and industry. Some classical methods formulate WSOD as a MIL problem (Cinbis, Verbeek, and Schmid; Li et al.; Jie et al.; Zhang et al. 2016; 2016; 2017; 2018a), which treats each training image as a bag of candidate instances and work in an iterative way to find the positive proposals and train a detector. In contrast to those interative MIL methods, some works try to construct end-to-end MIL models for WSOD (Tang et al.; Tang et al.; Bilen and Vedaldi; Diba et al. 2017; 2018; 2016; 2017). Another kind of methods mine pseudo labels from the location information obtained by the WSOD approaches to learn a supervised detector (Zhang

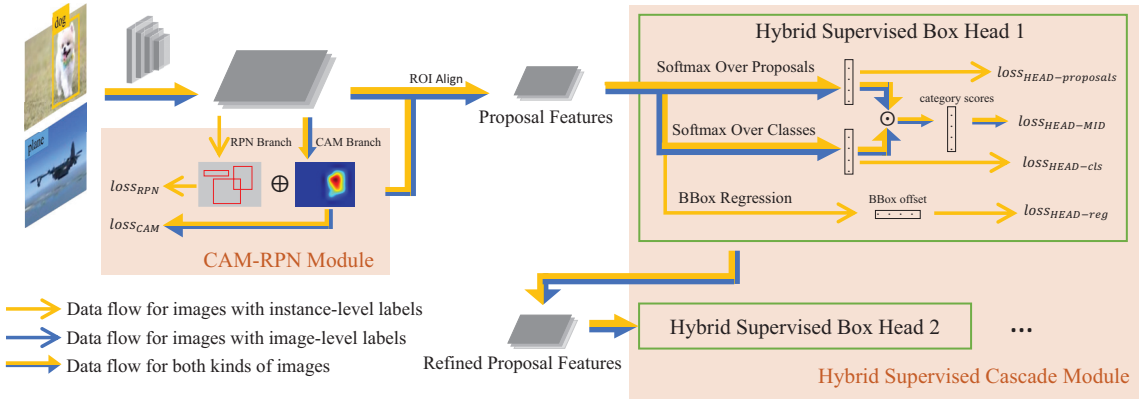

Figure 2: An overview of our EHSOD: a hybrid-supervised object detection framework trained by both image-level labels and instance-level labels. Stacked on an ImageNet pretrained backbone, we first proposed a CAM-RPN module to generate the foreground proposals guided by a class activation heat-map (CAM). The CAM branch is trained by both kinds of data jointly and provides enhanced objectness score for RPN to select proposals. Then a sequence of cascaded hybridsupervised heads further refines the bounding box position and improve classification accuracy. Within each head, image-level information is effectively incorporated to enhance the learning of classifiers. Both the bounding-box labels and the image-level labels are used to improve the performance of both classification and localization in an end-to-end manner.

et al.; Zhang et al.; Shen et al. 2018b; 2018a; 2018). These approaches easily fail without heavy hyper-parameter tuning and hardly achieve the performance requirement of most real-life applications.

Hybrid Supervised Object Detection (HSOD). HSOD aims at using a small number of fully-supervised data and a large number of weakly-supervised data to train a high-performan detector. Yan et al. developed an Expectation-Maximization (EM) based method for both WSOD and HSOD. Pardo et al. used teacher-student learning method to solve the HSOD problem. These methods work in an multi-step way, which heavily relys on hyperparameter tuning for mining high-quality pseudo object labels. We aims to solve this issue by construcing an end-to-end HSOD model that adds some weakly-supervised branches to the standard FSOD models without changing its structure. The whole model can be trained in an one shot fashion jointly on both kinds of data nearly without extra hyper-parameter tuning. It should be noted that our work differents from previous semi-supervised detection works that transfer knowlege from fully-supervised categories to weakly-supervised categories (Tang et al.; Uijlings, Popov, and Ferrari 2016; 2018). Our work focus on the setting where both instance-level labels and image-level labels are in the same categories.

The Proposed Approach

Overview. In this paper, we introduce EHSOD: a hybrid-supervised object detection framework to develop a general detection model to incorporate both image-level and bounding-box information. An overview of our EHSOD can be found in Figure 2. The proposed EHSOD is stacked on an ImageNet pretrained backbone to extract feature. A CAM-RPN is used to propose the foreground proposals guided by a class activation heat-map (CAM). The CAM is generated by two convolutional layers trained by the image-level labels and provide further information for RPN to select proposals. To generate classification score and location offset for each proposal, a sequential of cascaded hybrid-supervised heads progressively refines the bounding box position and improves classification accuracy. Within each hybrid-supervised head, image-level information is incorporated to enhance the learning of classifiers. Both the bounding-box labels and the image-level labels are used to improve the performance of both classification and localization in an end-to-end manner.

CAM-RPN Module

Conventional RPN in two-stage detection framework generates proposals based on predefined anchors. It has one classification layer to outputs the foreground scores and a regression layer to produce bounding-box offsets which is effective on fully-annotated data. We further improve the RPN module to utilize the image-level information.

Inspired from some works on WSOD and segmentation (Zhou et al.; Diba et al. 2016; 2017), class activation heat-map (CAM) presents the activated area in the feature map of an CNN and can help to locate the object. To enable end-to-end training and improve the usage of both kinds of data, we make following modifications on CAM. Let $\mathbf{f} = \{f_l\}_{l=1}^{4}, f_l \in \mathbb{R}^{W_l \times H_l \times D}$ be the four different feature maps with different resolutions in the FPN hierarchy extracted from the output of each stage of the ResNet backbone network. We use the same design of RPN (3x3 conv and 1x1 conv) to transform the f_l to a class activation heat-map A_l with C channels (C is the number of the categories). By adopting global average pooling over A_l and then performing the softmax operation, we obtain a vector $\boldsymbol{y}_{CAM} = [y_1, ..., y_C]$ presenting the predicted probability of each category in an image. Thus, we can construct a multi-

label classification loss over the ground truth of image-level information:

$$\mathcal{L}_{CAM-cls} = -\sum_{c} \{y_c^* log y_c + (1-y_c^*) log(1-y_c)\},$$

where $\mathcal{L}_{CAM-cls}$ is the binary cross-entropy loss and $y_c^* \in \{0,1\}$ is the label of the presence or absence of class c.

Note that the conventional CAM is purely learned from image-level information. To better reflect the position of objects of the generated CAM, we fully make use of the ground truth bounding-box. We first generate a ground truth heat-map for each A_l from the annotated bounding-boxes. Given an object instance of category c and its bounding-box coordinates $b = (x, y, w, h)$, we map it to the corresponding feature map scale s for generating the ground truth map of $A_s(c)$ (the cth channel of A_s). The mapped box coordinates is denoted as $b_m = (x^s, y^s, w^s, h^s)$. For the ground truth map of $A_s(c)$, we define the positive region $b_m^p = (x^s, y^s, \sigma w^s, \sigma h^s)$ as the proportional region of b_m by a constant scale factor σ, the remaining region of b_m as the ignoring region $b_m^i = b_m \backslash b_m^p$, and the whole map excluding the b_m as the negative region. According to the ground truth heat-map $A^* = \{A_l^*\}_{l=1}^4$, we add a pixel-level segmentation loss between the generated CAM:

$$\begin{aligned}
&\mathcal{L}_{CAM-seg} \\
&= -\alpha \sum_{l} \sum_{c} \sum_{i,j} \{A_l^*(c,i,j)(1-A_l(c,i,j))^\gamma log A_l(c,i,j) \\
&+ (1-A_l^*(c,i,j))(A_l(c,i,j))^\gamma log(1-A_l(c,i,j))\},
\end{aligned}$$

where $\mathcal{L}_{CAM-seg}$ is the pixel-level focal loss, $A_l(c,i,j)$ is the predicted probability on pixel (i,j) of cth channel of $A_l$obtained by performing an element-wise sigmoid function on the CAM, and $A_l^*(c,i,j)$ is the corresponding ground truth on pixel (i,j). For each proposal, we can calculate a objectness score from the CAM by performing a softmax operation over channels on the matched A_l and calculate the mean value over the proposal's region on the obtained single-channel CAM. The summation of the CAM objectness score and the confidence score from the RPN classification layer is used to perform non-maximum suppression (NMS) to select the best proposals for the next stage of hybrid-supervised cascade module.

To fully incorporate both kinds of data, the loss function of CAM-RPN is formulated as the weighted summation of the following four loss items:

$$\begin{aligned}
\mathcal{L}_{CAM-RPN} =& \alpha_1 \mathcal{L}_{CAM-cls} + \alpha_2 \mathcal{L}_{CAM-seg} \\
&+ \alpha_3 \mathcal{L}_{RPN-cls} + \alpha_4 \mathcal{L}_{RPN-reg},
\end{aligned}$$

Where the$\mathcal{L}_{RPN-cls}$ and $\mathcal{L}_{RPN-reg}$ are the regular RPN losses for predicting foreground confidence score and bounding-box offets as in (Ren et al. 2015b). During training, the $\mathcal{L}_{CAM-cls}$ and $\mathcal{L}_{CAM-seg}$ are calculated from both kinds of data, while the $\mathcal{L}_{RPN-cls}$ and $\mathcal{L}_{RPN-reg}$ are only gererated by fully-supervised data. The detailed computational flowchart is shown in Figure 3.

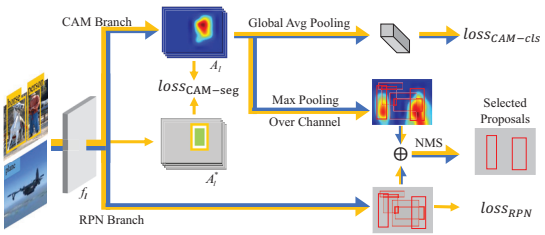

Figure 3: Detailed flowchart of the CAM-RPN Module. The whole module includes: a CAM branch trained by both kinds of data; a RPN branch trained only by fully-supervised data. The CAM branch is supervised by a image-level classification loss and a pixel-level bounding-box segmentation loss. The RPN branch generates a set of proposals and their confidence scores. Then another enhanced objectness score of each proposal is calculated by averaging the corresponding region on the CAM. The summation of these two scores is used as the final foreground score for each proposal, which is used to select the best proposals for the next stage.

Hybrid Supervised Cascade Module

In the conventional two-stage detection, RCNN head performs fine-grained classification and bounding-box refinement. The main purpose of this module is to exploit the weakly-labeled data to enhance the performance of the classifier and refine bounding-boxes.

Given the ROI feature $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_n\}$ and $\boldsymbol{x}_i \in \mathbb{R}^D$ from the n proposals by the previous CAM-RPN module, like conventional RCNN head for fully-supervised detector, we use one regression branch to predict the offset of bounding-box and one classification branch to predict the classification score s_{ic} for proposal i and categories c. Inspired by the network proposed in WSDNN (Bilen and Vedaldi 2016), we further add another new proposal-confidence branch to calculate the confidence score for each proposal. A fully connected layer takes X as input and outputs confidence score p_{ic} for proposal i on category c. Different from the classification branch, the softmax operation is taken over the proposals. Thus $\boldsymbol{p}_{:c}$ is a term that ranks all the proposals with the probability of containing category c while $\boldsymbol{s}_i$ are the probability of proposal i belongs to each category. Thus, for each image, the probability of containing an object with category c can be calculated as:

$$g_c = \sum_i p_{ic} s_{ic},$$

which can be also written as an elementary-wise matrix product $\boldsymbol{p} \odot \boldsymbol{s}$ and sum on all the proposals. Softmax is not performed at this step as images are allowed to contain more than one object class. Note that the s_{ic} is only calculated by softmax operation over the logits of foreground object classes at this step.

From the image-level labels (both from weakly-supervised data and fully-supervised data), we can train the classification branch and proposal-confidence branch jointly by a multiple instance detection (MID) loss:

Figure 4: Qualitative examples of our EHSOD trained on 30% fully-supervised COCO data. EHSOD can detect tiny and occusion objects due to the help of image-level information.

$$\mathcal{L}_{HEAD-MID} = -\sum_{c} \{y_c log g_c + (1-y_c) log(1-g_c)\},$$

where $\mathcal{L}_{HEAD-MID}$ is the binary cross-entropy loss and g_c is the predicted score of c category. Furthermore, for each image with bounding-box annotations, we can learn the proposal-confidence branch with supervision. After assigning category labels to all proposals, we set the ground truthp^*_{ij} as 0 for proposal i assigned with background class and set the ground truth p^*_{ij} as $1/N_j$ for proposal i assigned with category j (N_j is the total number of proposals assigned with category j). The resulting $\boldsymbol{p}^*_{:j}$ is set as the ground truth of $\boldsymbol{p}_{:j}$. Thus, we can train the proposal-confidence branch by a cross-entropy loss:

$$\mathcal{L}_{HEAD-proposals} = -\frac{1}{R}\sum_{j}\sum_{i} p^*_{ij} log p_{ij},$$

Where R is total number of region proposals. To fully utilize both kinds of data, the loss function of a hybrid-supervised head is formulated as the weighted summation of the following four loss items:

$$\begin{aligned}\mathcal{L}_{HS-head} =& \beta_1 \mathcal{L}_{HEAD-MID} + \beta_2 \mathcal{L}_{HEAD-proposals} \\ &+ \beta_3 \mathcal{L}_{HEAD-cls} + \beta_4 \mathcal{L}_{HEAD-reg},\end{aligned}$$

Where the $\mathcal{L}_{HEAD-cls}$ and $\mathcal{L}_{HEAD-reg}$ are the regular bounding-box losses as in (Ren et al. 2015b). During traing, the $\mathcal{L}_{HEAD-MID}$ is calculated from both kinds of data, while the $\mathcal{L}_{HEAD-proposals}$, $\mathcal{L}_{HEAD-cls}$, and $\mathcal{L}_{HEAD-reg}$ are only gererated by fully-supervised data. The detail computational flowchart is shown in Figure 2.

Note that the performance of the bounding-box regression branch is very weak since we have little data to train it. To further refine the bounding-box position, we adopt the cascade-structure with increasing IoU threshold. Empirically, we found that a sequence of three cascade works very well and can boost the localization performance of the detector.

Training EHSOD

Having discussed the EHSOD architecture in the previous section, here we explain how the model is trained. The prposed EHSOD framework is optimized in an end-to-end fashion using a multi-task loss. Apart from the conventional loss of cascade detection network (Cai and Vasconcelos 2018), we also introduce two losses $\mathcal{L}_{CAM-cls}$ and $\mathcal{L}_{CAM-seg}$ for CAM learning and two losses $\mathcal{L}_{HEAD-MID}$ and $\mathcal{L}_{HEAD-proposals}$ for hybrid-supervised head learning. They are jointly optimized by the following weighted summation of all losses:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{CAM-rpn} + \sum_{i} \lambda_i \mathcal{L}_{HS-head-i}.$$

Where $\mathcal{L}_{HS-head-i}$is the loss for the ith head. In practice, we found that the convergence of the model is very fast and we can train the model with the default setting of a two-stage detection network such as FPN (Lin et al. 2017b).

Experiments

Datasets and Evaluations. We evaluate the performance of our proposed EHSOD method on two common detection benchmarks: the PASCAL VOC 2007 (Everingham et al. 2015), and the MS-COCO 2017 dataset (Lin et al. 2014). The PASCAL VOC 2007 has 9,962 images with 20 categories. For PASCAL VOC 2007 , we choose trainval set (5,011 images) for training and choose the test set (4,952 images) for testing. The MS-COCO dataset has 80 object classes, which is divided into train set (118K images), val set (5K images) and test set (20K unannotated images). We train our model on the MS-COCO train set and test our model on the val set.

For the hybrid-supervised setting, we select a proportion of training images randomly as the fully-supervised training data, the remaining training images are used as weakly-supervised training data. We employ the standard mean Average Precision (mAP) metric with IoU=0.5 to evaluate our method on the PASCAL VOC dataset and employ mAP@[.5, .95] on the MS-COCO dataset.

Implementation Details. We use the popular FPN (Lin et al. 2017b) as our baseline detector and implement the

Method	Backbone	AP	AP_{50}
SSD (Liu et al. 2016b)	VGG16	26.8	46.5
RetinaNet (Lin et al. 2017c)	Res50	35.6	54.7
Faster R-CNN (Ren et al. 2015b)	Res50	32.6	53.1
FPN (Lin et al. 2017b)	Res50	35.9	56.9
FSAF (Zhu, He, and Savvides 2019)	Res50	37.2	57.2
AlignDet (Chen et al. 2019)	Res50	**37.9**	**57.7**
EHSOD w **30%** Fully supervised img	Res50	35.3	54.2
EHSOD w **50%** Fully supervised img	Res50	37.8	56.5
SSD (Liu et al. 2016b)	Res101	31.2	50.4
RetinaNet (Lin et al. 2017c)	Res101	37.7	57.2
Faster R-CNN (Ren et al. 2015b)	Res101	34.9	55.7
FPN (Lin et al. 2017b)	Res101	37.2	59.1
FSAF (Zhu, He, and Savvides 2019)	Res101	39.3	59.2
AlignDet (Chen et al. 2019)	Res101	**39.8**	**60.0**
EHSOD w **30%** Fully supervised img	Res101	37.5	56.8
EHSOD w **50%** Fully supervised img	Res101	**40.0**	59.4

Table 1: Comparison with Fully Supervised Object Detection methods on MS-COCO. The competing methods are trained with full data (1x schedule and no multi-scale training/testing). EHSOD trained with only 30%/50% fully-supervised images can reach comparable performance with fully-supervised object detection methods.

EHSOD network based on it. ImageNet pretained backbone is used as the backbone network. We use a sequence of three cascaded heads with increasing IoU threshold in the hybrid-supervised cascade module. Thus, our EHSOD network have four stages in total, one CAM-RPN for generating proposals and three heads for detection with IoU threshold $\{0.5, 0.6, 0.7\}$. We set the loss weights $\alpha_1$and α_2 in $\mathcal{L}_{CAM-RPN}$ to 0.1 and 0.2 respectively, set the loss weights λ_1, λ_2 and λ_3 for three hybrid-supervised heads to 1, 0.5. 0.25 respectively, and set all the other loss weights to 1. The scale factor σ for generating the positive region of the ground truth CAM is set to 0.8. The hyper-parameters α and γ for focal loss in the $\mathcal{L}_{CAM-seg}$ are set to 0.25 and 2 respectively. No data augmentation was used except standard horizontal image flipping.

During both training and testing, we resize the input image such that the shorter side has 600 pixels and 800 pixels for the PASCAL VOC dataset and the MS-COCO dataset respectively. All experiments are conducted on a single server with 8 Tesla V100 GPUs by using the Pytorch framework. For training, SGD with weight decay of 0.0001 and momentum of 0.9 is adopted to optimize all models. For the PASCAL VOC dataset, the batch size is set to be 8 with 4 images on each GPU, the initial learning rate is 0.005, reduce by 0.1 at epoch 9 during the training process. For the MS-COCO dataset, the batch size is set to be 16 with 2 images on each GPU, the initial learning rate is 0.01, reduce by 0.1 at epoch 8 and 11 during the training process. We only train **12 epochs for all models** in an end-to-end manner. **Multi-scale training/testing is not used** for all the models.

Comparison with Fully Supervised Object Detection methods. To show the effectiveness our method in using low-cost annotating (e.g., weakly-supervised) data to boost the detection performance, we compare the overall performance of our EHSOD method with its fully-supervised object detection counterpart. Specifically, we train the same proposed model without the data-flow of the weakly-supervised data. Note that the image-level information of the fully-supervised data will still be used in the model to train the CAM module and the classification branch/proposal-confidence branch in Hybrid Supervised Cascade Module. The backbone is ResNet-50. Figure 5 shows comparison between EHSOD (Red) and its fully-supervised version (Blue) on different portions of data. It can be found that our model can boost the performance mostly in 10% fully-supervised data. On Pascal VOC07, EHSOD can significantly increase the performance of mAP from 35% to 55% under only 10% of fully-supervised data.

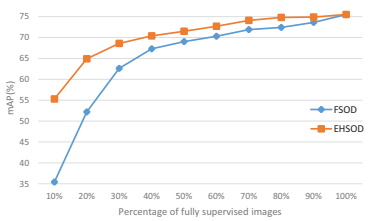

Figure 5: Comparison between our EHSOD (Orange) and its fully-supervised version (FSOD: Blue) on different portions of data of PASCAL VOC07. The blue line is its fully-supervised object detection counterpart. Note that the image-level information of the fully-supervised data is still used in the blue line model. The backbone is ResNet-50.

In Table 1, we further compared our method with fully object detection training with 100% data on MS-COCO. The competing methods are trained with full data. The reported results use 1x schedule (He, Girshick, and Dollár 2018) and no multi-scale training/testing, which is under the same setting with us. It can be found that our method trained with 30% data has comparable performance with fully-trained detector such as Faster-RCNN (Ren et al. 2015a), FPN (Lin et al. 2017a), RetinaNet (Lin et al. 2017c) and SSD (Liu et al. 2016a). Note that our model with backbone Resnet-101 can reached mAP of 40% with only 50% fully-supervised data. Figure 4 further shows some quantitative results for our EHSOD trained with 30% fully-supervised data on COCO. Our method has a very high accuracy and can detect very small items such as birds and traffic lights.

Comparison with Hybrid Supervised Object Detection method. Despite of the limited research works on hybrid-supervised detection network, we can compare the performance with BAOD (Pardo et al. 2019) with same setting of experiments. BAOD considered an iterative training scheme by an optimal image/annotation selection and retraining the detector. Table 2 compares the mAP_{50} under different setting of fully supervised data proportion from 10% to 100% on Pascal VOC07. Both methods are under same setting of experiments. Our method is trained once with 12 epochs while BAOD has several rounds of training.

From Table 2, it can be found that our method consistently

data	Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
10%	BAOD (Pardo et al. 2019)	51.6	50.7	52.6	41.7	36.0	52.9	63.7	69.7	34.4	65.4	22.1	66.1	63.9	53.5	59.8	24.5	60.2	43.3	59.7	46.0	50.9
	Our EHSOD	60.6	65.2	55.0	35.4	32.8	66.1	71.3	75.3	38.4	54.1	26.5	71.7	65.0	67.8	63.0	27.7	52.6	48.6	70.9	57.3	**55.3**
20%	BAOD (Pardo et al. 2019)	57.0	62.2	60.0	46.6	46.7	60.0	70.8	74.4	40.5	71.9	30.2	72.7	73.8	64.7	69.8	37.2	62.9	48.4	64.1	59.1	58.6
	Our EHSOD	65.5	72.3	66.7	45.6	50.8	72.2	77.8	82.2	44.3	73.1	44.8	79.3	76.0	73.0	73.8	35.5	63.0	62.1	74.0	65.5	**64.9**
40%	BAOD (Pardo et al. 2019)	68.6	71.3	66.6	52.5	53.1	69.6	77.7	77.2	45.7	72.7	54.0	74.4	74.6	74.7	74.4	42.4	66.2	56.8	71.7	65.4	65.5
	Our EHSOD	75.8	78.4	72.9	56.7	55.2	76.1	81.3	83.9	51.2	76.2	60.0	83.3	81.5	77.9	79.3	41.2	68.0	64.4	75.2	68.8	**70.4**
50%	BAOD (Pardo et al. 2019)	70.1	73.1	70.4	52.0	57.0	73.1	79.4	77.1	47.4	77.5	54.0	76.6	73.5	74.6	77.1	43.8	68.5	61.3	73.7	69.1	67.5
	Our EHSOD	73.4	77.8	72.9	57.5	57.2	79.5	81.6	83.5	53.7	79.0	60.4	83.5	81.9	76.6	79.7	45.4	69.1	67.3	77.9	71.9	**71.5**
60%	BAOD (Pardo et al. 2019)	73.5	75.3	72.4	52.5	53.3	76.5	81.1	81.0	51.0	76.7	57.9	76.8	79.2	77.0	79.0	45.4	69.3	63.0	75.3	67.2	69.2
	Our EHSOD	74.4	81.3	72.7	58.1	58.9	82.3	83.9	82.9	54.2	77.6	63.5	82.6	82.1	79.6	80.5	46.8	71.2	71.8	80.0	69.9	**72.7**
80%	BAOD (Pardo et al. 2019)	76.7	76.4	74.0	56.8	62.0	81.4	82.1	84.8	57.3	78.2	61.2	81.9	79.3	78.1	80.6	46.8	73.0	67.6	76.9	71.7	72.3
	Our EHSOD	83.1	82.9	77.0	60.6	63.4	81.5	85.2	86.1	56.2	80.5	65.9	84.2	83.1	79.1	82.5	47.8	73.8	71.7	79.7	72.4	**74.8**
100%	Our EHSOD	82.5	82.7	75.4	63.3	63.2	82.1	85.8	86.3	57.6	79.5	67.5	84.1	82.6	80.2	82.8	51.5	73.6	73.5	82.7	74.3	75.5

Table 2: Comparison of Hybrid Supervised Object Detection methods on VOC07. Both methods are trained with same settings of hybrid-supervised data. Our method is trained once with 12 epochs while BAOD has several rounds of training (each with 10 epochs). It can be found that our method consistently outperforms the BAOD especially under lower proportions of fully-supervised data.

VOC 2007	AP_{50}	*MS-COCO*	AP_{50}
PCL	48.8	PCL	19.6
Our EHSOD	**55.3**$^{+6.5}$	Our EHSOD	**46.8**$^{+27.2}$

Table 3: Comparison of the weakly-supervised method PCL and our method on VOC07 and MS-COCO. The EHSOD method is trained with 10% of fully-supervised data. In MS-COCO, our method outperforms PCL method by a large margin of 27.2% in terms of AP_{50}.

outperforms the competitor BAOD. Note that our method is significantly better than BAOD by around 5% of mAP under 10%-40% settings. This demonstrates the effectiveness of our method in utilizing weakly-supervised data. By observing the performance gain compared to the baseline method, it can be found that our method can successfully detect difficult categories such as aero, bike, bird, bottle and plant. These categories usually suffer from the problems of tiny-size and occlusion. Our method can alleviate these problems by the refinement of the hybrid-supervised heads.

Comparison with Weakly Supervised Object Detection methods. We further compare our approach with weakly-supervised object detection methods. Current WSOD methods mainly focus on easy detection datasets such as Pascal VOC, and only PCL (Tang et al. 2018) is tested on a much harder dataset: MSCOCO. Table 3 shows the comparison between PCL and our method on VOC07 and MS-COCO. The EHSOD method is trained with 10% of supervised data. Although PCL performs well in Pascal VOC, its mAP_{50} in MS-COCO is only 19.4. Our method significantly outperforms PCL by 27.2%, which implies that our method is superior in harder tasks.

Ablative Analysis. We conduct ablation analysis of the proposed method EHSOD, including the influence of adding CAM branch in RPN, using hybrid-supervised branches in the head and the effect of adding $L_{HEAD-proposals}$. For the hybrid-supervised head, it can be found that it can boost the performance by 9.1% mAP which demonstrates the importance of utilizing the image-level labels in heads. Adding $L_{HEAD-proposals}$ can further improve the mAP by 6.3%.

Training with 10% fully-supervised data	Cascade Heads	CAM Branch	Head adds $\mathcal{L}_{HEAD-MID}$	Head adds $\mathcal{L}_{HEAD-proposals}$	mAP_{50}
✓	✓				35.4
✓	✓		✓		44.5$^{+9.1}$
✓	✓		✓	✓	50.8$^{+6.3}$
✓	✓	✓	✓	✓	**55.3**$^{+4.5}$

Table 4: Ablative Analysis of EHSOD on VOC07. We compare the influence of adding CAM branch in RPN, using hybrid-supervised branches and the effect of adding $L_{HEAD-proposals}$ in the head.

# Cascaded modules	mAP_{50}	Speed/*fps*
1	67.8	22.5
2	70.3$^{+2.5}$	20.1
3	71.5$^{+3.7}$	18.6
4	71.6$^{+3.8}$	17.1

Table 5: Results on the different number of proposed Hybrid Supervised Cascade Modules. All models are trained with 50% of fully-supervised data and 50% of weakly-supervised data. All the inference time is tested on a single V100 GPU. "3" is the default setting of our model.

Our CAM branch in RPN achieves 4.5% improvements.

Impact of Different Number of Cascaded Modules. We evaluate the performance of the different number of cascaded modules of the proposed Hybrid Supervised Cascade Module. The comparison results are shown in Table 5. "Three cascaded modules" is the default setting of our model. It can be seen that our hybrid supervised cascade module with three cascaded modules can significantly improve the performance by 3.7% of mAP while only having a runtime overhead with 3.9*fps* out of 22.5*fps* comparing to the single module. It can be also found that adding too many cascaded modules (say more than 3) will not help much.

Conclusion

We study the hybrid-supervised object detection problem and present EHSOD, an end-to-end hybrid-supervised object detection system which can be trained jointly on both

fully-annotated data and image-level data. The performance of the proposed method is comparable to fully-supervised detection models with only a limited amount of fully annotated-samples, e.g. 37.5 mAP on COCO with 30% of fully-annotated data.

References

Bilen, H., and Vedaldi, A. 2016. Weakly supervised deep detection networks. In *CVPR*, 2846–2854.

Cai, Z., and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*.

Chen, Y.; Han, C.; Wang, N.; and Zhang, Z. 2019. Revisiting feature alignment for one-stage object detection. *arXiv preprint arXiv:1908.01570*.

Cinbis, R. G.; Verbeek, J.; and Schmid, C. 2016. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* 39(1):189–203.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*.

Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; and Van Gool, L. 2017. Weakly supervised cascaded convolutional networks. In *CVPR*, 914–922.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111(1):98–136.

He, K.; Girshick, R.; and Dollár, P. 2018. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*.

Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; and Liu, W. 2017. Deep self-taught learning for weakly supervised object localization. In *CVPR*, 1377–1385.

Kosugi, S.; Yamasaki, T.; and Aizawa, K. 2019. Object-aware instance labeling for weakly supervised object detection. *arXiv preprint arXiv:1908.03792*.

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

Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; and Sun, J. 2018. Detnet: A backbone network for object detection. In *ECCV*.

Li, X.; Kan, M.; Shan, S.; and Chen, X. 2019. Weakly supervised object detection with segmentation collaboration. *arXiv preprint arXiv:1904.00551*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *CVPR*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017b. Feature pyramid networks for object detection. In *CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017c. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016a. Ssd: Single shot multibox detector. In *ECCV*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016b. Ssd: Single shot multibox detector. In *ECCV*.

Pardo, A.; Xu, M.; Thabet, A.; Arbelaez, P.; and Ghanem, B. 2019. Baod: Budget-aware object detection. *arXiv preprint arXiv:1904.05443*.

Redmon, J., and Farhadi, A. 2017. Yolo9000: better, faster, stronger. In *CVPR*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015a. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015b. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Shen, Y.; Ji, R.; Zhang, S.; Zuo, W.; and Wang, Y. 2018. Generative adversarial learning towards fast weakly supervised detection. In *CVPR*, 5764–5773.

Tang, Y.; Wang, J.; Gao, B.; Dellandréa, E.; Gaizauskas, R.; and Chen, L. 2016. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *CVPR*, 2119–2128.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2843–2851.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. L. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*.

Uijlings, J.; Popov, S.; and Ferrari, V. 2018. Revisiting knowledge transfer for training object class detectors. In *CVPR*, 1101–1110.

Wan, F.; Wei, P.; Jiao, J.; Han, Z.; and Ye, Q. 2018. Min-entropy latent model for weakly supervised object detection. In *CVPR*, 1297–1306.

Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; and Ye, Q. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*, 2199–2208.

Wang, J.; Chen, K.; Yang, S.; Loy, C. C.; and Lin, D. 2019. Region proposal by guided anchoring. In *CVPR*, 2965–2974.

Wei, Y.; Shen, Z.; Cheng, B.; Shi, H.; Xiong, J.; Feng, J.; and Huang, T. 2018. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV*, 434–450.

Xu, H.; Jiang, C.; Liang, X.; and Li, Z. 2019a. Spatial-aware graph relation network for large-scale object detection. In *CVPR*.

Xu, H.; Jiang, C.; Liang, X.; Lin, L.; and Li, Z. 2019b. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *CVPR*.

Yan, Z.; Liang, J.; Pan, W.; Li, J.; and Zhang, C. 2017. Weakly-and semi-supervised object detection with expectation-maximization algorithm. *arXiv preprint arXiv:1702.08740*.

Zhang, X.; Feng, J.; Xiong, H.; and Tian, Q. 2018a. Zigzag learning for weakly supervised object detection. In *CVPR*, 4262–4270.

Zhang, Y.; Bai, Y.; Ding, M.; Li, Y.; and Ghanem, B. 2018b. W2f: A weakly-supervised to fully-supervised framework for object detection. In *CVPR*, 928–936.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhu, C.; He, Y.; and Savvides, M. 2019. Feature selective anchor-free module for single-shot object detection. *arXiv preprint arXiv:1903.00621*.