Object Detection Using Multiple Level Annotations

Thesis by
Mengmeng Xu

In Partial Fulfillment of the Requirements

For the Degree of

Masters of Science

King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia

April, 2019
The thesis of Mengmeng Xu is approved by the examination committee

Committee Chairperson: Bernard Ghanem
Committee Co-Chair: Tareq Alnaffouri
Committee Members: Ali Thabet
Object Detection Using Multiple Level Annotations
Mengmeng Xu

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 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.

This thesis studies hybrid learning methods on the object detection problems. We intend to train an object detector from a 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.

Based on the hybrid learning framework, we further study the problem of object detection from a novel perspective in which the annotation budget constraints are taken into consideration. When provided with a fixed budget, we propose a strategy for building a diverse and informative dataset that can be used to optimally train a robust detector. We investigate both optimization and learning-based methods to sample which images to annotate and which level of annotations (strongly or weakly supervised) to annotate them with.

By combining an optimal image/annotation selection scheme with the hybrid supervised learning, we show that one can achieve the performance of a strongly super-
vised detector on PASCAL-VOC 2007 while saving 12.8% of its original annotation budget. Furthermore, when 100% of the budget is used, it surpasses this performance by 2.0 mAP percentage points.
This thesis owns a considerable debt of gratitude to the many people who contributed its creation and accomplishment. I would first like to thank my thesis advisor Prof. Bernard Ghanem, Associate Professor of the Visual Computing Center at King Abdullah University of Science and Technology. Prof. Ghanem always gave instructive and insightful suggestions to me through slack, email or personal meetings whenever I ran into a trouble spot. He is also very patient to review my research reports and paper drafts, and provides valuable ideas on how to write a high-quality academic article. I do benefit a lot from his supervision.

I would also like to thank my colleagues who guide me to walk into this domain and shared their advice and suggestion in many of the discussions: Prof. Yancheng Bai, Dr. Ali Thabet. Without their passionate instructions and assistance, this work could not have been successfully conducted.

I would also like to acknowledge Professor Tareq Alnaffouri of the Computer, Electrical and Mathematical Science and Engineering Division at King Abdullah University of Science and Technology as the Committee Co-Chair of this thesis, and I am gratefully indebted to him for his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents: Mr. Liang Xu, Ms. Xuexia Ge and special thanks to my lover: Ms. Sisi Sally Qu for providing me with unfailing support and continuous encouragement throughout my master years of study and through the process of doing researches and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

Author

Mengmeng Frost Xu
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Committee Page</td>
<td>2</td>
</tr>
<tr>
<td>Copyright</td>
<td>3</td>
</tr>
<tr>
<td>Abstract</td>
<td>4</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>10</td>
</tr>
<tr>
<td>List of Figures</td>
<td>10</td>
</tr>
<tr>
<td>List of Tables</td>
<td>14</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>16</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>16</td>
</tr>
<tr>
<td>1.2 Objectives and Contributions</td>
<td>20</td>
</tr>
<tr>
<td>2 Review of Related Literature</td>
<td>22</td>
</tr>
<tr>
<td>2.1 Fully-Supervised Object Detection</td>
<td>22</td>
</tr>
<tr>
<td>2.2 Weakly Supervised Object Detection</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Hybrid Supervised Learning</td>
<td>24</td>
</tr>
<tr>
<td>2.4 Pseudo Object Labels</td>
<td>24</td>
</tr>
<tr>
<td>2.5 Active Learning</td>
<td>25</td>
</tr>
<tr>
<td>3 Design and Methodology – Hybrid Learning</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Construction of Hybrid Supervised Datasets</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Learning from Multiple Level of Annotations</td>
<td>28</td>
</tr>
<tr>
<td>3.2.1 Teacher-Student Framework</td>
<td>29</td>
</tr>
<tr>
<td>3.2.2 Post-processing</td>
<td>30</td>
</tr>
<tr>
<td>3.3 More Adaptations in Hybrid Learning</td>
<td>31</td>
</tr>
<tr>
<td>3.3.1 MID Modification</td>
<td>31</td>
</tr>
<tr>
<td>3.3.2 ICR Modification</td>
<td>33</td>
</tr>
</tbody>
</table>
3.4 Budget-Aware Hybrid Dataset

3.4.1 Random Sampling and Uncertainty Sampling

3.4.2 Optimization Based Selection using US

3.4.3 Optimization Based Selection using LAL

4 Experiments and Analysis of Missing Label Problem

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics

4.1.2 Implementation Details

4.2 The Effect of Missing Instance-Level Labels

4.3 Normal Object v.s. Small Object

4.4 Ablation Study

4.4.1 Teacher-Student Training

4.4.2 MID and ICR Adaption

4.5 Comparing to Other Methods

4.6 Qualitative Results

5 Experiment and Analysis of BAOD

5.1 Experimental Setup

5.2 Uncertainty Sampling and Hybrid Training

5.2.1 Uncertainty Sampling

5.2.2 Hybrid Training

5.3 Optimization-Based Active Selection

5.4 Effect of Per-Image Annotation Cost

5.5 Improving Detection using Fixed Budget

5.6 Easy Images and Weak Annotation First

5.7 Qualitative Results of the Active Selection

6 Conclusion and Future Work

References

Appendices

A.1 Experiments on COCO

A.2 Experiments of Hybrid Learning

A.3 Experimental Data of BOAD

A.4 Definition of Easy, Medium and Hard Classes

B.1 Image Sampling
B.2 Post Processing
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_r$</td>
<td>(instance-level) Missing Rates</td>
</tr>
<tr>
<td>BAOD</td>
<td>Budget-Aware Object Detection</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>FSOD</td>
<td>Fully Supervised Object Detection</td>
</tr>
<tr>
<td>ICR</td>
<td>Instance Classifier Refinement</td>
</tr>
<tr>
<td>IOU</td>
<td>Intersection over Union</td>
</tr>
<tr>
<td>LAL</td>
<td>Learning Active Learning</td>
</tr>
<tr>
<td>mAP</td>
<td>Mean Average Precision</td>
</tr>
<tr>
<td>MID</td>
<td>Multiple Instance Detection</td>
</tr>
<tr>
<td>NMS</td>
<td>Non-Maximum Suppression</td>
</tr>
<tr>
<td>RS</td>
<td>Random Sampling</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
<tr>
<td>US</td>
<td>Uncertainty Sampling</td>
</tr>
<tr>
<td>VOC07</td>
<td>PASCAL VOC 2007 trainval</td>
</tr>
<tr>
<td>VOC12</td>
<td>PASCAL VOC 2012 trainval</td>
</tr>
<tr>
<td>WSOD</td>
<td>Weakly Supervised Object Detection</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1.1 The Mean Average Precision (mAP) for one WSOD detector and four classical FSOD detectors training under different instance-level missing rates ($M_r = 0 : 0.1 : 0.9$) of the training dataset. These models are trained on VOC2007 train-val set and evaluated on the VOC2007 test dataset. .......................... 17

1.2 Budget-Aware performance of detectors with different levels of supervision. The models are trained on PASCAL VOC 2007 trainval (VOC07), PASCAL VOC 2012 trainval (VOC12) as specified in the legend. Our proposed Budget aware object detection (BAOD) (green -∇- curve) has a higher mAP than FSOD (yellow -| - curve) and WSOD (orange - - curve) methods at most budgets. Given a larger unlabeled image pool (Blue -♦- curve, VOC07+VOC12), our BAOD can reach a higher mAP using the same budget needed to annotate VOC07 with instance-level labels. Since the dataset is finite, WSOD cannot increase its performance with more budget. .......................... 18

3.1 Illustration of the Hybrid Learning framework. Given an image collection with hybrid labels, we firstly use a warm-up detection model to generate pseudo instance labels (e.g. black solid rectangles in the inference phase.) After cleaning noise, the second detection model learns from both the ground truth and pseudo instance labels. .......................... 29
### 3.2 More adaptions in the detailed teacher-student learning framework

Given an image collection with image-level labels and partial instance-level labels, we firstly use W2F [1] 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 to ROI layer. MID gives region detection results. And ICRs further refine the learning target. 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. We take the replacement repeatedly until convergence.

### 3.3 Overview of active learning pipeline to construct a hybrid labeled dataset

For any weakly labeled or unlabeled image in the image pool (circular shapes), the selection method (bottom blue rectangle) decides which type of action to apply on the image based on the sample function and image status: weakly label ($x_1$) or strong label ($x_2, x_3$). Then such image is appended into the hybrid dataset and we train an object detection model with the hybrid supervision.

### 4.1 Effect of the missing rate of instance-level labels on detection performance

(left) shows the average precision for four COCO[2] classes with different missing label rates. The results are given by Faster-RCNN[3] 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.

### 4.2 Ablation plot for different framework setups

It compares the mAP when we use Teacher-Student Model, Student Model Adaption and Repeated Teaching among all the $M_r$s.
4.3 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. The missing rate of training set is 0.7.

5.1 Comparison of the cost and image number on every active selection step. Left: Budget usage distribution to learn different difficulty categories. More budget is used to annotate Easy images (green area) at beginning. The cost spent on Hard images (red area) grows up when the active model is mature. Right: Selected images distribution in different difficulty categories and different annotation type. The selection agent gives more weak annotations (light color) at the first steps. Given more budget, the proportion of strong annotations (dark color) increase. We run out of unlabeled images after 9-th step. The mapping is motivated and shown in Appendix A.

5.2 Visualization of the selected images in each step. Left: Two examples in the warm-up set which is fully annotated by 10% budget. Up-Right: Strongly annotated images per step. They are hard examples including occlusion, multiple instance or tiny scale. Bottom-Right: Weakly annotated images per step. They are simple in the beginning but the difficulty increases when the detector is mature.
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The affect of missing small scale instance-level objects. We compare both regular FSOD model and our proposed Hybrid training model 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.</td>
</tr>
<tr>
<td>4.2</td>
<td>Model performance with different modules. This table shows the Average Precision when we add different modules when ( M_r ) is fixed to 0.7.</td>
</tr>
<tr>
<td>4.3</td>
<td>Average Precision of FSOD, Missing Label Object Detection and WSOD. All of them are trained/tested on VOC2007 ( train-val/test ) dataset.</td>
</tr>
<tr>
<td>4.4</td>
<td>Mean Average Precision of FSOD, our method and WSOD. All of them are train/test on VOC2012 ( train-val/test ) dataset.</td>
</tr>
<tr>
<td>5.1</td>
<td>Budget-Average mAP using fully and hybrid training pipelines with random and uncertainty selection. Uncertainty sampling is always better than random sampling selection, and hybrid training is always better than FSOD.</td>
</tr>
<tr>
<td>5.2</td>
<td>Budget-Average mAP using simple hybrid training and optimization methods. US based optimization is slightly better than LAL one. The optimization methods perform better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.</td>
</tr>
<tr>
<td>5.3</td>
<td>Budget-Average mAP using a lower cost for strong annotations. If we assume the weak and strong annotation costs are more close (7 seconds and 1.5 seconds), US based optimization (BAOD) still performs better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.</td>
</tr>
</tbody>
</table>
5.4 Simulation of a larger unlabeled image pool. With 87.2% budget, BAOD achieves the same performance as a detector trained on fully annotated VOC07. If the budget equals to the total budget of VOC07, BAOD achieves 2% mAP improvement over FSOD with the same budget. Further annotating all the images can only improve 3.4% mAP.

A.1 COCO Experiments: Missing labels for each single categories. Because of the noise, some missing label performances are better than fully annotated one.

A.2 Hybrid Learning Experiments. They are trained on VOC 2007/2012 train-val set, tested on 2007/2012 test set. Missing rate ranges from 0.1 to 0.9.

A.3 Table Representation for Raw Experiment Data

A.4 The mapping from twenty categories in PASCAL VOC to Easy, Medium and Hard classes. This mapping is based on the final mAP of each category.
Chapter 1

Introduction

1.1 Background

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) [4, 5], whose success mainly benefits from the flexibility of deep learning models and an abundance of instance-level annotations in extensive datasets [6, 7]. 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, more seriously, missing annotations of objects in training images can have a drastic impact on FSOD performance. In this case, we focus on two problems of great interest and importance in object detection area in the thesis: (1) how to design an object detection model that can learn from not only instance level annotations but also the image level annotations such that it is less affected by the missing labels. (2) Based on the hybrid learning method, we study how to build a dataset with multiple level annotations on which we can train the detector. The dataset should take less amount of budget since it is not necessarily strongly annotated for every image, but the detector can still obtain a good performance.

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
Figure 1.1: The Mean Average Precision (mAP) for one WSOD detector and four classical FSOD detectors training under different instance-level missing rates ($M_r = 0 : 0.1 : 0.9$) of the training dataset. These models are trained on VOC2007 train-val set and evaluated on the VOC2007 test dataset.

arise and it becomes more prevalent when the dataset grows in size (both in the number of training images and object classes). Fig. 1.1 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 in all the annotations in the original dataset. Ranging from 0 to 0.4, its performance decreases slightly while 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.

On the other hand, building a successful object detector encompasses three main dimensions: (1) the image dataset to be annotated for training the detector. A larger dataset allows for a more accurate detector, but the number of training images is limited by the annotation budget; (2) the annotation scheme used to label
Figure 1.2: **Budget-Aware performance of detectors with different levels of supervision.** The models are trained on PASCAL VOC 2007 trainval (VOC07), PASCAL VOC 2012 trainval (VOC12) as specified in the legend. Our proposed Budget aware object detection (BAOD) (green -∇- curve) has a higher mAP than FSOD (yellow -| curve) and WSOD (orange - - curve) methods at most budgets. Given a larger unlabeled image pool (Blue -♦- curve, VOC07+VOC12), our BAOD can reach a higher mAP using the same budget needed to annotate VOC07 with instance-level labels. Since the dataset is finite, WSOD cannot increase its performance with more budget.

the training images. One could annotate either image-level labels (the categories of the objects are known but their locations are unknown, denoted weakly supervised annotation) or instance-level labels (both categories and locations are known, denoted strongly supervised annotation) \[8, 2, 9, 10\]; (3) **the detection model.** Most works on object detection fix the first two dimensions and only explore the third. In fact, they tend to focus on optimizing the detection model based on one or more training datasets that typically provide the same kind of annotations. With a hybrid learning detection model, we can further investigate solutions in the first two dimensions of the problem.

A large group of object detectors fall under the umbrella of Fully-Supervised Object Detection (FSOD) \[11, 12, 13, 14, 15, 16, 17, 18\]. It has been shown in recent years that these techniques can reach high detection performance, especially with the introduction of large datasets with strong annotations \[19, 20, 2\]. This requirement makes
FSOD methods expensive and time-consuming. In contrast, Weakly-Supervised Object Detection (WSOD) \cite{21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31} aims at building object detectors from cheaper but less informative image-level or weak annotations.

Training an object detector using multiple level annotation is a trade-off between dataset annotation cost and model precision in order to combine both weak and strong annotations and train an object detector with multiple level annotations. We put all the detectors on the same footing when different annotation schemes are available. Ideally, detectors should only be compared when they are trained using image datasets that offer the same amount of information (not necessarily the same number of images). Since this notion is difficult to define quantitatively, we take the training budget of a detector as a unifying surrogate measure. Here, we define budget as the effort, or cost, to annotate a dataset, thus combining the first two dimensions of the detection problem: dataset scale and annotation scheme. In fact, with a fixed budget and a set of unlabeled images, the number of images we can label depends highly on the annotation cost for each image. This cost varies significantly between image- and instance-level annotations. Typically, annotating a bounding box around an object in an image is significantly more expensive than simply annotating its category \cite{8, 19}. Therefore, an FSOD method with the same budget as a WSOD method contains fewer images in its training set.

We explore strategies to build better object detector models when constrained with a training budget, a novel problem we coin budget-aware object detection (BAOD). As such, we focus on choosing the best images for training and how to annotate them. For this, we survey several selection methods to sequentially choose both the image and type of annotation following an active learning paradigm. Figure 1.2 shows that actively selecting images and their annotations to sequentially train a hybrid supervised detector outperforms its FSOD and WSOD counterparts, when the budget is larger than 20\% of the budget needed to annotate VOC2007 trainval at the
instance-level. Moreover, the yellow curve shows that we can reach an even higher mAP if we use the same budget of VOC07 to annotate images from both datasets VOC07 and VOC12.

1.2 Objectives and Contributions

This thesis aims to study the problem of training an object detector with multiple level supervisions and its application. To sum up, we make the following five contributions.

(1) We evaluate the robustness of mainstream FSOD methods to varying rates of missing labels, varying object categories (section 4.2) and different object sizes (section 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 not robust enough to missing annotations in the training dataset.

(2) We compare 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 state-of-the-art FSOD and WSOD methods carry out no better performance when compared with our proposed detector on VOC2007 and VOC2012 with 10 different missing rates (in section 4.4). To the best of our knowledge, we are the first to use hybrid training (i.e. with both instance- and image-level annotations) in the context of object detection.

(3) Our framework gives a trade-off between collecting large-scale fully-annotated dataset and training a better object detection model. Explicit experiments in section 4.5 reveals the highly practical value of our work to do object detection.

(4) We propose the BAOD problem and present a new evaluation criterion Budget-Average mAP for object detection algorithms. This criterion takes into account both detection performance and budget. Experiments in Section 5.3 show that in many cases we can reach 95% of state-of-the-art performance with only half
of the total budget.

(5) Following a BAOD approach (i.e. combining intelligent image and annotation scheme selection with hybrid supervised learning), we show that the mAP test performance on VOC07 can be improved by 2 percentage points for the same budget used to annotate the training set of VOC07 (by combining it with VOC12). We also show the opposite, i.e. that state-of-the-art test performance on VOC07 can be achieved, while saving 12.8% of the budget used in strongly annotating its training set.

In the thesis, we firstly review some of related work in Chapter 2. Then we explain the proposed hybrid supervised dataset construction and hybrid learning method in the followed Chapter 3. Specially, we attach the hybrid dataset with its contraction cost to study the benefit the hybrid learning method, in the last section of chapter three. In the experiment Chapter 4, we evaluate the performance of our hybrid supervised learning framework on the missing label problem, each module of which is validated by an ablation study. Our method is a strong baseline bridging the wide gap between FSOD and WSOD performances. Next, it is discussed in Chapter 5 that the multiple level supervision can be applied to solve the budget-aware object detection problem using hybrid learning on a dataset constructed by an intelligent image selection method. We can either save budget while keep the same performance of the detection, or build a better object detection model with the same amount of budget. Final Chapter 6 comes the conclusion of the whole thesis.

To be clearer on the terms in this paper, when talking about an image, strong annotation denote the instance-level annotation (a set of bounding boxes and their classes) of the image, and weak annotation means the image-level annotation (a list of objects in the images). Furthermore, when we say the dataset is Weakly Supervised, all of the images have weakly annotation. When it is Fully Supervised, all of the images are strongly annotated. If some of the images are weakly annotated, and others are strongly annotated, the dataset is Hybrid Supervised.
Chapter 2

Review of Related Literature

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 [32], Faster RCNN [33], SSD [34], YOLO2 [12], and their variants [35, 36, 37]. Faster RCNN [33] 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 [12] 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 requires the collection and curation of large-scale instance-level bounding-box annotations, which is expensive and time-consuming. Su et. al. report that it takes around 26 seconds to draw one bounding-box without quality control (asking multiple people to do annotation on the same images), and 42 seconds with it [38]. There has been recent progress in developing tools to further reduce annotation time (most recently to 7 seconds in average) [11, 8, 9]. As we pointed out in Fig. 1.1, the performance of FSOD detectors is largely affected by the amount of missing annotations.

However, if we train the object detection model using multiple level annotations, then datasets do not need to be fully annotated, which can save the annotating time and the data storage effectively. We can constrain the annotation budget and focus
on how to reach the best detection performance within this budget by intelligently selecting between both weak and strong annotations. Our work investigates a complementary aspect of annotation to emphasize that some images do not need to be strongly annotated to achieve excellent performance. While our discussion of budget aware object detection only considers the labor cost for image annotation, we note that sequentially training detection models also consume computational resources. However, given the continual development of hardware acceleration and re-organization, which leads to steady decrease in their cost, manual annotation remains the more expensive component in the detection problem.

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 (MIL) problem [39, 27, 28, 29, 21, 31]. Li et. al. [27] collect class-specific object proposals and then exploits them to fine-tune the network. It also uses confident object candidates to optimize the representation in the target domain. Bilen et. al. [24] present a weakly supervised deep detection network (WSDDN), which selects positive samples by multiplying the score of recognition and detection. Others focus on improving the optimization strategy in training. In [31], instead of using the static absolute CNN score during training, the model relies on a relative improvement in output CNN scores. Tang et. al. [40] improve WSDDN and design an online instance classifier refinement (OICR) algorithm to alleviate the local-optimum problem that plagues WSDDN.

We use weak supervision to improve pseudo object labels. These labels are generated from a previously trained detector, and are post-processed through a pseudo label mining process. The image-level labels help remove false predictions that cor-
respond to a wrong object category.

2.3 Hybrid Supervised Learning

In the thesis, we study hybrid supervised learning for object detection. This type of learning exploits multiple types of supervision during training. Semi-supervised learning is a special case of this family, since it learns a model from a set of labeled and unlabeled data. Several works [41, 42, 43, 44, 45] try to solve the generic semi-supervised learning problem through *teacher-student learning*. They first train a model from the strongly annotated subset as a *teacher*, then the predictions obtained from the teacher model on the unlabeled data are used to regress a second model that is called the student model. More generally, Hu *et. al.* [46] propose a novel weight transfer function to solve the instance segmentation problem from a hybrid dataset of instance- and pixel-level annotations. Chéron *et. al.* [47] also use several kinds of annotations to train an activity recognition model, and they revealed that strongly annotating every training sample is not necessary to achieve noteworthy localization results in the video domain.

In our case, since images can be labeled either weakly or strongly, a hybrid supervised dataset is always considered. Inspired by Rosenberg’s work [45], we train a teacher-student model to use the hybrid dataset. Our teacher detector is learned from strong annotations only, while the student detector is learned from both ground-truth and/or processed pseudo object labels.

2.4 Pseudo Object Labels

Pseudo object labels are confident bounding boxes predicted by a decent detector, especially by a weakly supervised detector. They have been recently used in a fully-
supervised setup to compensate for the absence of all instance-level box annotations. A framework [48] is proposed to exploit tracked object bounding boxes from videos to serve as Pseudo label to train an object detector. The cascade detector in Diba’s work [49] and OICR [40] both use pseudo object labels to train Faster-RCNN and achieve eminent WSOD performance. Mining pseudo labels can also increase the success of fully-supervised detectors. Zhang et. al. [1] determine the most accurate bounding box using pseudo ground-truth excavation (PGE) and pseudo ground-truth adaption (PGA) algorithm from predictions.

These object label generators have fixed output. Once produced, these methods cannot utilize other annotations to update pseudo bounding boxes. Our innovation here is to generate pseudo labels from different levels of annotations and update them in every training cycle.

2.5 Active Learning

In general, active learning is a sequential decision making process. It iteratively selects the most useful examples that an oracle should annotate and add to the labeled training set. It aims at training more accurate models with the minimum data required. This field has been widely studied in the context of image and video classification [50, 51, 52, 53, 54, 55], object and action localization [56, 9, 57, 10, 58], human pose estimation [59] and visual question answering [60]. A commonly used approach to select new training images is by means of their entropy score [61]. The intuition is that higher entropy examples attribute to more learning information. More recent research directly predicts the improvements of adding a new sample to the training set, and uses this measurement as a selection criteria [52]. One commonly used method is to calculate an entropy score of the training example [61] expecting that this score be directly related to the knowledge one can get from it. Another recent work proposes to predict the improvement if the sample is added to the training
set\textsuperscript{[52]}. As specified in previous work, active learning aims at maximizing performance while minimizing the human cost in labeling the training samples \textsuperscript{[50, 59, 56]}.

The BAOD problem can also be formulated as a sequential decision making process. We study a few well-known selection techniques in our new active learning pipeline, which contains two active processes: (1) select the next batch of training images to annotate and (2) decide the type of annotation for each selected image. Thus, we focus on the annotation sequence that can provide the most useful information to the detector. To the best of our knowledge, we are the first to use active learning and hybrid training to study object detection.
Chapter 3

Design and Methodology – Hybrid Learning

In this section, we will give a comprehensive study of our hybrid supervised learning framework for the multiple supervision problem in three aspects. To our best knowledge, no attention has been paid to this problem, and there is no standard multiple level supervised object detection training datasets. 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. 1.2 and Fig. 3.2. Finally, we discuss how to design a budget-aware object detection scheme based on this framework.

3.1 Construction of Hybrid Supervised Datasets

Our hybrid supervised datasets are constructed from two level of annotations: the image level and instance level. For example, the image level annotation of a dog image is to know there is a dog, but the instance level one need to further know where the dog is. Most public large scale object detection datasets have both annotations, or we can retrieve the image annotation from the instance one. In our case, the dataset is mixed of weakly labeled images and strongly labeled images with a certain ratio. As mentioned in the first section, we denote $M_r$, missing rate as the ratio from only weakly labeled image numbers to the number of strongly labeled images.

Alg. 1 presents how to constructs the instance-level missing label dataset from any
Algorithm 1 Instance-Level Label Sampling

Require: missing rate $m_r$, number of categories $M$, number of images $N$;
1: for each annotation $j$ in each image $i$ do
2: find annotation category $k$;
3: append annotations to $objLabels[k][i]$;
4: end for
5: for each category $k$ do
6: sample $m_r$ of annotations from $objLabels[k][:]$ (without replacement)
7: remove the sampled data from $objLabels[k][:]$;
8: end for
9: save labeled image index as $ml_{ImageList}$;
10: save $objLabels[k][:]$ in PASCAL VOC standard as $annotation_{m_r}$;
Ensure: $annotation_{m_r}$, $ml_{ImageList}$.

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. 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 thesis, the missing label
object detection dataset is based on PASCAL VOC2007/2012.

3.2 Learning from Multiple Level of Annotations

We use a teacher-student learning framework followed by a post-processing step to
solve the hybrid supervised learning problem, illustrated by Fig.3.1. 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 net-
work can be any off-the-shell object detector, or an adapted one for the missing label
background (section 3.3). On the other side, the post processing is a Non-Maximum
Suppression (NMS) for the object categories that appear in image level supervision.
Figure 3.1: **Illustration of the Hybrid Learning framework.** Given an image collection with hybrid labels, we firstly use a warm-up detection model to generate pseudo instance labels (e.g. black solid rectangles in the inference phase.) After cleaning noise, the second detection model learns from both the ground truth and pseudo instance labels.

### 3.2.1 Teacher-Student Framework

In this framework, we firstly train an decent object detector with only one level of annotation as a teacher model. There are two options to make the teacher model accurate. Firstly, if most of the images in the dataset have weakly annotations, we can use the Weakly Supervised Object Detection method to train first detection. On the other hand, if we assume there are more information concluded in the strongly annotated images, we can use FSOD method on this partition of the hybrid dataset. Training on one level of annotation can only learn a decent detector, but it can also transfer knowledge to the images which are annotated differently. This detector works as a teacher that predicts objects in every unlearned image in the dataset. If these predictions are post-processed properly, they can be viewed as pseudo labels, which we merge with the ground truth instance-level labels to train a fully supervised student detector. In order to overcome undesirable local minima, the student detector is also pretrained from Imagenet [19]. Since it has both strong and weak annotations from training samples in the whole dataset, we expect the student detector to perform better than its teacher.

The student detector can also be used to replace the teacher model to further improve the hybrid 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.

3.2.2 Post-processing

The predictions from the teacher model have many redundant or erroneous pseudo labels. We use minimal knowledge to post-process them so as to focus our study on the active selection methods and the benefit of hybrid training. More concretely, given an image $I_k$ with a weak annotation $\omega \in \{0, 1\}^c$, where $c$ is the number of categories, we assume that the teacher model gives $M$ positive predictions with localization $P \in \mathbb{R}^{M \times 4}$ (4 components defining a bounding box), classification $A \in \{0, 1\}^{M \times c}$, and the confidence score for each of the $M$ positive predictions $q \in [0, 1]^M$. $P$, $A$ and $q$ are obtained from the teacher model. Among these $M$ predictions, we seek to mine the pseudo labels in the form of a sparse $M$-dimensional binary vector $y$ that solves the following constrained optimization:

$$
\max_{y \in \{0, 1\}^M} y^\top q \\
\text{s.t.} \quad \begin{cases} 
    y^\top A(1 - \omega) = 0 \\
    \text{IoU}(P_i, P_j) \leq \alpha & \forall \ y_i, y_j = 1 \\
    1 \leq \|y\|_0 \leq \beta
\end{cases}
$$

(3.1)

Here the IoU function evaluate the intertersection over union of the two predected bounding boxes $P_i, P_j$. The intuition behind solving this problem is that we seek to maximize the confidence score across all pseudo labels indexed by $y$. The first constraint enforces image-level consistency of the pseudo labels with the ground truth weak annotations. The second constraint removes predictions that are highly over-

\footnote{The VGG16 weights are pretrained from ImageNet, and the remained are initialized randomly.}
lapping (similar to an NMS post-processing step). The third constraint enforces a sparsity condition on $y$, thus limiting the number of possible pseudo labeled bounding boxes (potential objects) to be between 1 and $\beta$. In our experiments, we finally take $(\alpha = 0.3, \beta = 3)$ using hyper-parameter search. Details on this optimization are in the Appendix.

As a result, 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.

## 3.3 More Adapations in Hybrid Learning

We also adapt two main modules in WSOD field into our hybrid supervised learning framework to take fully use of the weakly annotation when we are training the student model. These two modification usually lead to a more accurate detector but takes more storage space and computation time. The first module is a Multiple Instance Detection (MID) network [24] which introduces loss function for the images labels. And the second Instance Classifier Refinement (ICR) network [40] further improves the localization accuracy from more reliable regression targets. The framework with these adaptions is shown in Fig. 3.2.

In the following discussion, we assume an image has both image-level labels $y=[y_1, \ldots, y_C] \in \{0, 1\}^C$ and instance-level labels $P=[p_1, \ldots, p_L] \in \mathbb{R}^{L \times 5}$. 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 $P'=[p'_1, \ldots, p'_{L'}] \in \mathbb{R}^{L' \times 5}$. 
Given an image collection with image-level labels and partial instance-level labels, we firstly use W2F \cite{1} 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 to ROI layer. MID gives region detection results. And ICRs further refine the learning target. 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. We take the replacement repeatedly until convergence.

### 3.3.1 MID Modification

In weakly supervised learning, \( n \) object proposals are used to crop the images, then the image patch is represented by feature vectors, which passes through classification and localization sub-networks and get two \( C \times |R| \) matrices \( \sigma_{cls}(X^c) \) and \( \sigma_{det}(X^d) \). In original MID, The score of each proposal is pixel-wise production \( X^R = \sigma_{cls}(X^c) \odot \sigma_{det}(X^P) \). Finally, the image-level class prediction score \( p \in \mathbb{R}^C \) can be obtained by summation of \( X^R \) for all proposals, as Eq. 3.2:

\[
[p]_c = \sum_{r=1}^{|R|} [\sigma_{cls}(X^C) \odot \sigma_{det}(X^P)]_{c,r} \tag{3.2}
\]
A sub-network is developed in this module to employ ground-truth and pseudo instance labels. This branch has similar structure as a RCNN[62], but is learned differently. These proposals are divided into three groups and used for different targets: (1) Positive proposals are the regions which have a high IOU with any ground truth instance label. They are both the classification and regression target. (2) Semi-positive proposals are the regions that own a highest IOU with pseudo labels. They are only used for classification loss. (3) Negative samples are regions which have a small IOU with either instance labels or pseudo labels.

In short, we only use ground truth bounding boxes to regress locations, while avoiding pseudo label giving inaccurate bounding boxes. During training, the modified loss function can be formulated in Eq. 3.3:

\[
\text{Loss}_{MID} = \sum_{c=1}^{C} L_{lab}(y_c, p_c) + \frac{\alpha}{N_{cls}} \sum_{p^* \in \tilde{P}} L_{cls}(p, p^*) + \frac{\beta}{N_{reg}} \sum_{p^* \in P} L_{reg}(p, p^*). \tag{3.3}
\]

Here, \( \tilde{P} = P \cup P' \) 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.

### 3.3.2 ICR Modification

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[40]. For the \( k \)-th refining subbranches in ICR, each of them classifies \( r \)-th proposal as \( z_r^{(k)} \in \mathbb{R}^{C+1} \). The label \( y_r^{(k)} \) for \( r \)-th proposal in \( k \)-th subbranch is the most confident predictions from \( k - 1 \)-th subbranch. The instance label is also the training target for all of the \( k \) branches if it exists.

For each subbranch in ICR, the weighted cross-entropy loss (classification loss) is
used as in Eq.3.4:

\[
\text{Loss}_{ICR} = \frac{1}{|R|} \sum_{k=1}^{K} \sum_{r=1}^{|R|} L_{cls}(x_r^{(k)}, y_r^{(k)}, w_r^{(k)}),
\]

(3.4)

where \(w_r^{(k)}\) denotes the confidence vector, \(y_r^{(k)}\) represents the ground truths, and \(K\) means the number of classification subbranches.

Finally, we train our end-to-end student model by combining the loss functions from the above three modules, as in Eq. (3.5).

\[
\text{Loss} = \text{Loss}_{RPN} + \text{Loss}_{MID} + \text{Loss}_{ICR}.
\]

(3.5)

\(\text{Loss}_{RPN}\) is a regular RPN loss as in [16].

### 3.4 Budget-Aware Hybrid Dataset

In this section, we study methods to sequentially create an object detection dataset with a fixed budget, which includes both image-level and instance-level annotations. A complete overview of the whole active learning procedure is shown in Figure 3.3.

Our hybrid supervised dataset should consider which images to include in training and how to label them. To tackle this problem, we propose an active hybrid learning framework which acts as an active learning agent to simulate the annotation process. At every active step \(t\), we have a budget constraint \(d\) to spend on annotations. Each image in this set belongs to one of three pools: unlabeled \(U_t\), weakly labeled \(W_t\), or strongly labeled \(S_t\). At each active step, the hybrid supervised dataset \(W_t \cup S_t\) is used to train a hybrid detector, which is described in detail in Section 3.2. This detector is used to find a selection function that takes an action on \(U_t\) or \(W_t\) pools. As shown in Figure 3.3, we have three possible actions with their associated cost: (1) annotate strongly \(x_1\) \((i.e.\) send the image from \(U_t\) to \(S_t)\), (2) annotate weakly \(x_2\) \((i.e.\) send the
Figure 3.3: Overview of active learning pipeline to construct a hybrid labeled dataset. For any weakly labeled or unlabeled image in the image pool (circular shapes), the selection method (bottom blue rectangle) decides which type of action to apply on the image based on the sample function and image status: weakly label ($x_1$) or strong label ($x_2, x_3$). Then such image is appended into the hybrid dataset and we train an object detection model with the hybrid supervision.

image from $U_t$ to $W_t$), or (3) strongly annotate a weakly annotated image $x_3$ (i.e. send the image from $W_t$ to $S_t$). Once the actions have been made, the image sets $U_t$, $W_t$, and $S_t$ are updated. We proceed iteratively until we either run out of images in $P_t = U_t \cup W_t$ or run out of budget $d$.

To this end, we study three active learning sampling functions (Random, Uncertainty, and Learning Active Learning [52]) within four action selection methods: Random Sampling (RS), Uncertainty Sampling (US), Optimization based on US and Optimization based on Learning Active Learning (LAL) in the followed three subsections.

### 3.4.1 Random Sampling and Uncertainty Sampling

For RS, we randomly choose an active batch of images at each active step to include into $W_t$ or $S_t$. For US, images are sorted in descending order of uncertainty (measured
by entropy) to choose high uncertainty images to train on with full supervision based on an uncertainty score denoted as $s_k$, and the top images are included into $W_t$ or $S_t$. We adhere to the following annotation policy for RS and US when a budget constraint is enforced. For both RS and US, we prioritize weak labels first. In other words, so long as the budget constraint is not exceeded, (1) the image batch that is selected in each active step for these two methods will only contain images from $U_t$ that will be weakly labeled or, (2) if $U_t$ becomes empty, images from $W_t$ will be selected and get a strongly label.

There are several ways to evaluate $s_k$ for an image $I_k$ [61][8]. We follow convention and model this score in Eq. 3.6

$$s_k = \frac{1}{M} \sum_{i=1}^{M} \sum_{p \in p_i} -p \log(p) \quad (3.6)$$

This is an entropy measure for each of the $M$ bounding boxes in an image using the classification score predicted by the current detection model. Collecting $M$ predictions from an image $I_k$, each prediction has a probability score vector $p_i \in [0, 1]^c$ for the $c$ object categories.

### 3.4.2 Optimization Based Selection using US

At the $t$-th active step, let $P_t = \{I_k\}_{k=1}^N$ be the $N$ images that can be further annotated, and $s \in \mathbb{R}^N$ be their corresponding uncertainty scores (as defined above). We have three possible action vectors $x_1, x_2, x_3 \in \{0, 1\}^N$, as defined earlier. If the $k$-th element of $x_i$ is 1, we annotate the image $I_k$ using option $i$. Assume the next active batch of annotations has a linear impact on the model performance increment $\delta_t$.

$$\delta_t = f_1(P_t)^T x_1 + f_2(P_t)^T x_2 + f_3(P_t)^T x_3 \quad (3.7)$$

---

2We also conducted an experiment in which images in US were sorted in ascending order and low uncertainty images were chosen in every step. We discuss this in greater detail in the Appendix.
To quantitatively show the contribution of new images, we also assume that the uncertainty score is a complete statistic of an unlabeled or weakly labeled set of images. Then, we can simply approximate \((f_1, f_2, f_3)\) as linear functions. We observe that many active learning studies \([63, 56]\) have experimentally shown that incorporating images with a higher uncertainty score in training may further improve the detector, but at the risk of having more difficulty in producing true positive predictions. Therefore, we model \((f_1, f_2, f_3)\) as linear functions that tend to favor actions \(x_1, x_3\) over action \(x_2\) for images with high enough uncertainty score \(i.e.\), higher than the median). As such, and by combining the intuitions above, the expected increment in performance is modeled as:

\[
\delta_t \approx s^\top(x_1 + x_3) + (\mu_1 - s)^\top x_2
\] (3.8)

For the active selection method based on optimization using US, we seek to maximize \(\delta_t\), while staying within the budget constraint. To model this latter constraint, we define \(a\) as the cost of strongly annotating an unlabeled image, \(b\) as the cost of weakly annotating an unlabeled image, and \(c\) as the cost of strongly annotating an already weakly labeled image. Therefore, this selection method seeks to solve the following binary optimization problem for \((x_1, x_2, x_3)\):

\[
\max_{x_1, x_2, x_3 \in \{0, 1\}^N} s^\top(x_1 + x_3) + (\mu_1 - s)^\top x_2 \\
\text{s.t.} \begin{cases} 
    x_3 \leq \psi \\
    x_1 + x_2 \leq 1 - \psi \\
    1^\top(ax_1 + bx_2 + cx_3) \leq d \\
    1^\top(ax_1 + bx_2 + cx_3) \geq d - a 
\end{cases}
\] (3.9)

Here, we define a vector \(\psi\) as the indicator vector for images that have already
been weakly annotated, \( i.e. \) the \( k \)-th component of \( \psi \) is 1 if \( I_k \in W_t \) and 0 otherwise. Using this indicator, the first two constraints in Eq (3.9) enforce that only one action is performed on each image, \( i.e. \) among all the \( k \)-th components of \( (x_1, x_2, x_3) \), only one can be 1. The third and fourth constraints enforce that the budget be used as much as possible in each active step.

The linear binary problem in Eq (3.9) is NP-hard in general, so exact solvers (\( e.g. \) the Branch and Bound Algorithm) tend to have long run-times especially when \( N \) is large. As a tradeoff between optimization accuracy and per active step runtime, we employ a conventional linear relaxation of the original problem to form an approximate linear program (LP), which can be efficiently solved at large-scale with off-the-shelf LP solvers. More details about this problem are shown and proven in the Appendix.

### 3.4.3 Optimization Based Selection using LAL

The uncertainty score \( s \) is not the only selection measure that has been studied in the active learning literature. In the Learning Active Learning (LAL) method proposed by Konyushkova \textit{et. al.} \cite{52}, the increment in performance function \( \delta_t \) is learned to be a function of the current model state as well. More concretely, given images \( I_k \in P_t \) and the current detection model, we build a feature vector \( v = [O_t, s] \) that concatenates both the current model state \( O_t \), represented as the average precision curves under five different Intersection over Union (IoU) thresholds, and the uncertainty scores \( s \). Following the LAL method in \cite{52}, we train a Support Vector Regression (SVR) model to regress the actual increment in mAP performance from \( v \) at each active step for both weak and strong annotation actions. Obviously, these SVRs are trained on a separate detection dataset than the one used to evaluate BAOD. At each active step and by denoting the output predictions of these SVRs as \( h_w \) and \( h_s \), we formulate the
same constrained optimization problem in Eq (3.9) but with the *learned* objective:

\[ \delta_t \approx h_w^\top x_2 + h_s^\top (x_1 + x_3) \] (3.10)
Chapter 4

Experiments and Analysis of Missing Label Problem

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 state-of-the-art detectors on the public detection benchmark (PASCAL VOC2007 and VOC2012 datasets) and present some qualitative results.

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics

We construct the missing label datasets based on PASCAL VOC datasets [64]. PASCAL VOC 2007 and 2012 comprise of 9963 images and 22531 images respectively, which include 20 categories of objects. We use VOC2007 and VOC2012 train-val sets to train our models, which are evaluated on the corresponding test datasets. To investigate the robustness of our detectors, we generate the missing label training set with missing rates ranging from 0.0 to 0.9 with step size 0.1. For the COCO[2] experiments, we only discard four classes of object label: dog, table, book, and cow, and process the data as in PASCAL VOC. Afterwards, the mean average precision (mAP) is utilized to evaluate the performance of the detectors.
4.1.2 Implementation Details

Our framework employs the VGG16 model pre-trained on ImageNet [6] as the backbone network. We set three subbranches on the classification coherence supervisor \((i.e. K=3)\). In our training setting, the total number of iterations is set to 70\(k\) for VOC2007 and 80\(k\) for VOC2012, and the learning rate is 0.001 for the first 40\(k\) iterations and 0.0001 in the remaining iterations. The mini-batch size is one for each iteration. We also use 0.9 momentum and 0.0005 weight decay. Grounded on Chen’s work[3] in Tensorflow, we design the hybrid learning part that will be publicly accessible later.

During training process, we keep the original aspect ratio of images and resize the shortest side to one of these five scales \(\{480, 576, 688, 864, 1200\}\), and ensure the longest side is not larger than 2,000 simultaneously. The negative example of RCNN is the predictions which have Intersection Over Union (IOU) of any object label between 0.1 and 0.5. Furthermore, we randomly flip images in the horizontal direction with a probability of 0.5 during training.

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 methods on the missing label dataset under different missing ratio: RCNN, Faster-RCNN, YOLO and SSD. Fig. 1.1 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[1] when the missing rate \(M_r\) is 0.7. From the experiments, we can conclude that current FSOD detectors are very sensitive to the quality of the training dataset.
Figure 4.1: **Effect of the missing rate of instance-level labels on detection performance.** *(left)* shows the average precision for four COCO\[2\] classes with different missing label rates. The results are given by Faster-RCNN\[3\] 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.

As discussed above, different models are tested on our dataset, and we also run Faster-RCNN in another COCO dataset. Fig. 4.1 shows the same mAP pattern when $Mr$ is increasing. We conclude some more interesting discoveries in the larger dataset investigation. *(1)* A small $Mr$ can hardly affect model performance. In COCO dataset, we do not observe a clear decrease when $Mr < 0.4$. *(2)* Object class that appears as a cloud is more sensitive to $Mr$. AP for *cow* class falls faster than *table* class, and there are always multiple cows in an image but with only one table. We report data for more categories in the Appendix A and expect to see more emphasis could be laid on this phenomena by researchers.

### 4.3 Normal Object v.s. 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 $\times$ height) of all instances in each category in PASCAL VOC 2007; *(2)* for an instance whose area is smaller than the
Table 4.1: **The affect of missing small scale instance-level objects.** We compare both regular FSOD model and our proposed Hybrid training model 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fully</th>
<th>$M_r = 0.3$</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>71.3%</td>
<td>68.3%</td>
<td>68.8%</td>
</tr>
</tbody>
</table>

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 4.1 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% v.s. 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% v.s. 45.3%), from which we can conclude that our method is more robust to the small scale object missing problem.

### 4.4 Ablation Study

In this part, we first compares different methods discussed in section 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.
Table 4.2: Model performance with different modules. This table shows the Average Precision when we add different modules when $M_r$ is fixed to 0.7.

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

4.4.1 Teacher-Student Training

The light yellow curve in the plot of Fig. 4.2 shows the model performance from the teacher-student training framework with simplest setup. Its accuracy is strongly depressed at low $M_r$ because of the imprecise pseudo labels. On the other side, 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.

4.4.2 MID and ICR Adaption

To properly use ground truth, pseudo labels and image labels, we adapted MID and ICR modules into our student model. The table in Fig. 4.2 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 can generate similar predictions by different branches and by strong overlapped proposals. The dark yellow curve of Repeated Teaching in the plot of Fig. 4.2 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 Repeatedly upgrading teacher model further improves the overall performance. Our grey curve in the plot of Fig. 4.2 reaches the original Faster-RCNN record when no bounding box is missed and surpasses 4.2 percentage points.
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.

### 4.5 Comparing to Other Methods

Hybrid supervised method strives a balance between labelling images and obtaining a more accurate detection model. We also compare our model with the mainstream FSOD [34] and WSOD [40, 31, 48] methods in Table 4.3 and 4.4.

**Comparison at Different $M_r$**

Table 4.3 shows AP performance on the VOC 2007 test set. The central block of the table shows our results in different missing rate, i.e., $M_r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

Our method achieves outstanding performance from 61.9% to 70.7% on different missing rate, which is between state-of-the-art 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 4.4 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 [40] and W2F [1]. Since our approach is based on Faster-RCNN, we compare it with Fast-RCNN [32] and Faster-RCNN [16] On the right side. Our baseline method successfully bridges between the gap in performance between WSOD and FSOD. Except for Faster-RCNN, the fully-supervised detector can be taken to be any general off-the-shelf detector.

**Comparison with WSOD**

The first block in Table 4.3 compares our model

\footnote{The other five sets of results are in the Appendix.
Table 4.3: 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 et. al. 2017 [31]  | 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 |     | 41.7 |
| Krishna et. al. 2016 [48]| 53.9 | -    | 37.7 | 13.7 | -      | -   | 56.6  | 51.3  | -     | 24.0  | 2017  | 51.9  | 43.6  | 53.6  | 62.4   | 51.9  | 43.6  | 53.6  |     |     |     |
| Tang et. al. 2017 [10]  | 65.5 | 67.2 | 47.2 | 21.6 | 22.1   | 68.0| 68.5  | 35.9  | 5.7   | 63.1  | 51.9  | 43.6  | 53.6  | 62.4  | 41.7   |       |       |       |     |     |     |
| Zhang et. al. 2017 [1]  | 63.5 | 70.1 | 50.5 | 31.9 | 62.4   | 72.0| 73.7  | 23.3  | 53.4  | 52.6  | 2016  | 51.9  | 43.6  | 53.6  | 62.4   | 51.9  | 43.6  | 53.6  |     |     |     |
| Mr = 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 |
| Mr = 0.7                | 65.9 | 73.4 | 59.0 | 51.7 | 42.1   | 70.3| 74.2  | 40.9  | 72.1  | 64.1  | 57.9  | 76.6  | 78.8  | 72.5   | 70.0  | 30.9  | 63.6  | 61.8 | 73.3 | 68.2 |
| Mr = 0.5                | 68.1 | 78.0 | 61.7 | 51.7 | 50.4   | 74.8| 78.4  | 82.1  | 76.6  | 64.3  | 63.3  | 78.2  | 80.9  | 72.7   | 74.9  | 36.4  | 64.4  | 67.9 | 71.7 | 70.1 |
| Mr = 0.3                | 67.1 | 78.4 | 66.2 | 53.2 | 54.2   | 76.7| 80.0  | 81.9  | 75.5  | 63.9  | 81.2  | 82.3  | 75.2  | 76.7   | 38.8  | 69.5  | 61.8  | 73.4 | 71.9 | 68.8 |
| Mr = 0.1                | 69.1 | 78.3 | 70.3 | 54.6 | 56.1   | 78.5| 81.2  | 82.3  | 52.9  | 66.9  | 80.7  | 83.3  | 74.6  | 77.4   | 43.0  | 71.0  | 66.1  | 74.2 | 72.6 | 70.3 |
| Liu et. al. 2016 [34]   | 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  | 46.3  | 69.6  | 73.4 | 82.0 | 71.2 |
| Chen et. al. 2017 [3]   | 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 |
| Redmon et. al. 2016 [12] | 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 | 67.9 |
| Girshick et. al. 2015 [65]| 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 |

to WSOD methods. The reason for the 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 Mr = 0.9, only very few bounding boxes are seen in training, and yet our method improves upon W2F [1] by 4.2% in mAP. Clearly, training the regression part of the model using ground truth improves the localization accuracy.
Table 4.4: Mean Average Precision of FSOD, our method and WSOD. All of them are train/test on VOC2012 train-val/test dataset.

<table>
<thead>
<tr>
<th>WSOD</th>
<th>Missing Label Object Detection (M_r)</th>
<th>FSOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tang[40] Zhang[1]</td>
<td>0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1</td>
<td>Ren[16] Girshick[65]</td>
</tr>
<tr>
<td>42.5</td>
<td>47.8</td>
<td>67.0</td>
</tr>
</tbody>
</table>

**Compare to FSOD** Compared to the methods in the last block in Table 4.3, 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 Classification Coherent Supervision module marks these areas as positive samples and reduces the related loss.

4.6 Qualitative Results

In Fig. 4.3, 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.

In image A, 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 B is very similar to the FSOD result. In C, 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 D. The E, F line shows images which have multiple objects from more than one class. Our combines both instance-level labels and image-level labels
Figure 4.3: 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. The missing rate of training set is 0.7.

Moreover, we visualize some failed detection. Our model detects the reflection as a duck in image B5. In D5, the large potted plant is wrongly classified as a sofa, and the black cow in F2 is incorrectly identified as a dog. In summary, there is still much room for improvement.
Chapter 5

Experiment and Analysis of BAOD

5.1 Experimental Setup

Datasets. Similar to the previous chapter, we use the PASCAL VOC 2007 (VOC07) or 2012 (VOC12) datasets \cite{20} to perform most of the experiments. Given the active learning pipeline of our method, we emulate the active learning procedure using VOC07-12 annotations to selectively annotate 20 categories in 5011 images in VOC07 or 16551 images in the union of VOC07 and VOC12 (VOC0712). All detection models are evaluated on the VOC07 test dataset. For each annotation type (weak or strong), we assume that each image has a fixed annotating cost/time, which is not necessarily true in practice but it simplifies the analysis. In most of the experiments, we set \((a = 34.5, b = 1.6, c = a - b)\), in unit seconds, according to the annotation procedure of \cite{19}. In Section 5.4, we vary these cost value to \((a = 7, b = 1.6, c = a - b)\) following the more efficient annotation procedure of \cite{8}. The cost of fully annotating VOC07 trainval is denoted as 100% (or total) budget for every experiment.

Evaluation Metrics. To evaluate the active selection methods with the hybrid detector, we compute a budget-performance curve at various budget limits. The budget axis varies the percentage of the total budget, and the model detection performance is taken to be mAP. In doing so, we propose a new budget-aware metric denoted as Budget-Average mAP, measured as the normalized area under the budget-mAP curve for a certain budget range. We take three ranges \([10\%, 30\%]\), \([30\%, 50\%]\), and \([50\%, 100\%]\) to evaluate our experiment such that three to five data
points are located in each range and the metric is less affected by noise. The first range starts from 10% since we need a small fully labeled warm up set to initialize the fully supervised detector, and start our pipeline. This warm up set is randomly selected and fixed in all experiments. The last budget range is wider because the performance curve saturates at high budgets, and we observe more subtle changes in performance.

Implementation Details. In every active step, we choose Faster-RCNN as the object detection model (teacher and student), which is trained in the hybrid supervised way described in Section 3.2. Both teacher and student models use VGG16 as the backbone network pre-trained on ImageNet [19]. We follow the default setup in [66] to train Faster-RCNN. During training, the total number of epochs is set to 10, the learning rate is 0.01 for the first 8 epochs and 0.001 for the remaining. The batchsize is set to 16 on four-GPU cluster nodes equipped with Titan Xp. The student model is cloned from the ImageNet pre-trained VGG16 in every active step and has the same training schedule as the teacher model. For LAL experiment, we collect 10 categories from the MS-COCO dataset [2] to train the SVRs. Since the LAL method is dataset agnostic [52], we take the SVR training categories to be different from the 20 categories in PASCAL VOC.

5.2 Uncertainty Sampling and Hybrid Training

5.2.1 Uncertainty Sampling

In order to explore the influence of the uncertainty score on model performance, we emulate the annotation process of the oracle using VOC07 and incorporate it into the active learning pipeline with only strong annotations and use a FSOD method to train our model. We use two selection methods (RS and US) to build the dataset for FSOD. From the first two columns in Table 5.1 that compare these two sampling

---

1The source code of the framework will be made publicly available.
Table 5.1: **Budget-Average mAP using fully and hybrid training pipelines with random and uncertainty selection.** Uncertainty sampling is always better than random sampling selection, and hybrid training is always better than FSOD.

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>FSOD</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Function</td>
<td>RS</td>
<td>US</td>
</tr>
<tr>
<td><strong>Low Budget Range</strong></td>
<td>52.5</td>
<td>53.1</td>
</tr>
<tr>
<td><strong>Mid Budget Range</strong></td>
<td>62.5</td>
<td>63.6</td>
</tr>
<tr>
<td><strong>High Budget Range</strong></td>
<td>67.9</td>
<td>68.7</td>
</tr>
</tbody>
</table>

methods, we see that collecting images with large uncertainty score is more effective and preferable than random sampling in the three budget ranges.

### 5.2.2 Hybrid Training

Table 5.1 also shows the influence of the active selection scheme on model performance but when the hybrid supervised learning process is used to combine weak and strong annotations at each active step. We again use RS and US to select the images in each step for hybrid training. It is clear that both hybrid methods outperform their corresponding FSOD counterpart at every budget range, especially when the budget range is low (under 50%).

### 5.3 Optimization-Based Active Selection

Here, we evaluate the optimization-based selection methods (using US and LAL) described in Section 3.4 when they are combined with hybrid supervised training. Table 5.2 compares these methods by measuring their average mAP at each budget range and compares them to the RS and US selection methods for references. These results indicate that the optimization methods are the best ways to combine and use the hybrid annotations at every budget, where the optimization-based US method is slightly better than its LAL counterpart. As such, we denote the former method as the BAOD approach, which was mentioned and highlighted in Figure 1.2. Interestingly,
Table 5.2: **Budget-Average mAP using simple hybrid training and optimization methods.** US based optimization is slightly better than LAL one. The optimization methods perform better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Hybrid</th>
<th>Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Function</td>
<td>RS US</td>
<td>LAL US (BAOD)</td>
</tr>
<tr>
<td>Low Budget Range</td>
<td>56.4 55.1</td>
<td>56.3 57.1</td>
</tr>
<tr>
<td>Mid Budget Range</td>
<td>64.5 65.8</td>
<td>65.9 66.0</td>
</tr>
<tr>
<td>High Budget Range</td>
<td>68.7 69.3</td>
<td>69.3 69.5</td>
</tr>
</tbody>
</table>

we observe that the RS method requires 62% of the total VOC07 budget (all images are strongly annotated) to achieve 95% of the detection performance at that budget (i.e. 67.4% mAP). In comparison, the BAOD method requires only 48.5% of the total budget to reach the same performance. This performance gap attests to the effectiveness of this method. We include more results in Appendix A.

### 5.4 Effect of Per-Image Annotation Cost

The cost of strong annotations can vary due the annotation strategy that is used. For instance, Papadopoulos et. al. created a method that reduces the time needed to draw bounding boxes to 7 seconds per image on average. Here, we use this cost ($a = 7$) to study the behavior of our BAOD method given a smaller gap between strong and weak annotation. We report these results in Table 5.3 which depicts the best performing FSOD method and the RS hybrid baseline for reference. We observe the same relative behavior between the methods as in the case when $a$ was several times higher. However, the performance gap between these methods decreases in this case because the number of images that can be weakly annotated for the cost of one strong annotation is much smaller.
Table 5.3: **Budget-Average mAP using a lower cost for strong annotations.** If we assume the weak and strong annotation costs are more close (7 seconds and 1.5 seconds), US based optimization (BAOD) still performs better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>FSOD</th>
<th>Hybrid</th>
<th>Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Function</td>
<td>RS</td>
<td>US</td>
<td>US (BAOD)</td>
</tr>
<tr>
<td>Low Budget Range</td>
<td>53.1</td>
<td>52.3</td>
<td>54.2</td>
</tr>
<tr>
<td>Mid Budget Range</td>
<td>62.8</td>
<td>63.1</td>
<td>63.6</td>
</tr>
<tr>
<td>High Budget Range</td>
<td>68.3</td>
<td>68.6</td>
<td>69.3</td>
</tr>
</tbody>
</table>

Table 5.4: **Simulation of a larger unlabeled image pool.** With 87.2% budget, BAOD achieves the same performance as a detector trained on fully annotated VOC07. If the budget equals to the total budget of VOC07, BAOD achieves 2% mAP improvement over FSOD with the same budget. Further annotating all the images can only improve 3.4% mAP.

<table>
<thead>
<tr>
<th>Train Set Pool</th>
<th>VOC07</th>
<th>VOC0712</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Budget*</td>
<td>100%</td>
<td>87.2%</td>
</tr>
<tr>
<td>Final mAP</td>
<td>0.710</td>
<td>0.710</td>
</tr>
</tbody>
</table>

* The values are normalized by VOC07 trainval set cost.

### 5.5 Improving Detection using Fixed Budget

This experiment simulates a real-world application of our method. We combine VOC07 and VOC12 data to simulate a larger pool of unlabeled images (VOC0712) to choose from. As reported in Table 5.4, when the budget is set at 87.2% of the total budget (annotation budget for VOC07), we learn an object detector whose performance is the same as an FSOD detector trained on the entire strongly labeled VOC07 training set. This is a budget saving of 12.8%. Now, if we choose to use this total budget on VOC0712, our method achieves a 73.0% mAP, which is 2% improvement over the aforementioned FSOD detector on the same VOC07 test set. If all the 16551 images in the union of VOC07 and VOC12 set are strongly labeled, we use an extra 230% of budget and only improve 3.4% mAP.
5.6 Easy Images and Weak Annotation First

We analyze the cost and number of images selected at every active selection step to investigate which type of training examples are more helpful in the sequential training process. Based on the final mAP of each category in VOC07 [14], we divide the twenty categories into three groups: Easy, Medium, and Hard. The left plot of Figure 5.1 illustrates that BAOD spends more budget to annotate Easy categories (shown in the green area) in the first several steps, while the cost of Hard categories (red area) increases when the detector becomes more accurate. These results align with concepts from curriculum learning, in which a larger number of easy samples can be trained on first to bootstrap the model and then hard samples are introduced progressively.

The right plot of Figure 5.1 measures the percentage of selected images per active step that belong to various subsets of the data (combinations of annotation type chosen and difficulty). Interestingly, the BAOD model tends to select more weak
Figure 5.2: **Visualization of the selected images in each step.** *Left:* Two examples in the warm-up set which is fully annotated by 10% budget. *Up-Right:* Strongly annotated images per step. They are hard examples including occlusion, multiple instance or tiny scale. *Bottom-Right:* Weakly annotated images per step. They are simple in the beginning but the difficulty increases when the detector is mature.

5.7 **Qualitative Results of the Active Selection**

Figure 5.2 shows some selected strongly and weakly labeled images based on the BAOD experiment in section 5.5. Each row of the images is from the dog category in the VOC07 or VOC12 trainval set, and each column of images is selected at the same active step.

We show that in the first five active steps, strongly annotated images contain dog instances that are difficult to detect due to occlusion, multiple close instances or small-scale. In contrast, the selected weakly annotated images during the same steps are relatively easier to locate. As the model gets mature the difficulty for both levels of annotations increases. For example, the image chosen in step 4 contains
three small dogs, and the dogs appeared in step 5’s images are small and black which makes them barely visible.
Chapter 6

Conclusion and Future Work

In this thesis, we present a novel framework for multiple level supervised object detection. Our pipeline combines the advantages of fully-supervised and weakly-supervised learning. We first generate pseudo ground truth instance-level labels using multiple level supervised object detection method, and then train an end-to-end 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 supervised and weakly supervised methods, and show that our method stands out among all of them at different proportion of instance-level labels.

In addition, we introduce a novel budget-aware perspective to study the unexplored dimensions of the object detection problem using the hybrid learning method we developed. With a fixed budget, we compare both optimization and learn based sample methods to build diverse hybrid supervised object detection datasets which consist of both image level supervision and instance level supervision. With the optimal set-up, our proposed budget-aware approach can achieve the performance of a strongly supervised detector on PASCAL-VOC 2007 while saving 12.8% of its original annotation budget. Furthermore, when 100% of the budget is used, our approach surpasses this performance by 2 percentage points of mAP.

In the future work, we are looking for the application of this work on industrial side. There are over 1000 categories in dataset imagenet \[19\], but only 80 classes of them are strongly annotated. Using the combination of hybrid training and hybrid
dataset can expand the object detection tasks to a more general field. On the other hand, we can also apply this framework on the action detection problem in the video. In this case, the weak annotation is to classify what actions are included, while the strong annotation need to further point out in which frame range that the activity happens. This work is also promising because annotating an video is much more expensive than drawing an bounding box in the image.
REFERENCES


## APPENDICES

### A Raw Data and Extra Experiments

#### A.1 Experiments on COCO

Table A.1: **COCO Experiments**: Missing labels for each single categories. Because of the noise, some missing label performances are better than fully annotated one.

<table>
<thead>
<tr>
<th>$M_r$</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1</td>
<td>1.03</td>
<td>1.04</td>
<td>1.04</td>
<td>1.00</td>
<td>0.97</td>
<td>0.86</td>
<td>0.75</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>1.084</td>
<td>1.095</td>
<td>1.100</td>
<td>1.059</td>
<td>1.057</td>
<td>0.990</td>
<td>0.816</td>
<td>0.822</td>
<td>0.712</td>
</tr>
<tr>
<td>cow</td>
<td>1</td>
<td>1.019</td>
<td>1.006</td>
<td>1.030</td>
<td>0.996</td>
<td>0.990</td>
<td>0.774</td>
<td>0.699</td>
<td>0.602</td>
<td>0.334</td>
</tr>
<tr>
<td>table</td>
<td>1</td>
<td>1.001</td>
<td>1.028</td>
<td>0.996</td>
<td>0.974</td>
<td>0.910</td>
<td>0.854</td>
<td>0.778</td>
<td>0.754</td>
<td>0.628</td>
</tr>
<tr>
<td>bool</td>
<td>1</td>
<td>0.975</td>
<td>0.900</td>
<td>0.933</td>
<td>0.800</td>
<td>0.633</td>
<td>0.591</td>
<td>0.475</td>
<td>0.325</td>
<td>0.200</td>
</tr>
</tbody>
</table>
A.2 Experiments of Hybrid Learning

Table A.2: Hybrid Learning Experiments. They are trained on VOC 2007/2012 train-val set, tested on 2007/2012 test set. Missing rate ranges from 0.1 to 0.9.

<table>
<thead>
<tr>
<th>Data/missing Rate</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>motorcycle</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>69.1</td>
<td>78.3</td>
<td>70.3</td>
<td>54.6</td>
<td>56.1</td>
<td>78.5</td>
<td>81.2</td>
<td>82.3</td>
<td>52.9</td>
<td>73.3</td>
<td>66.9</td>
<td>80.7</td>
<td>83.3</td>
<td>74.6</td>
<td>77.4</td>
<td>43.0</td>
<td>71.0</td>
<td>66.1</td>
<td>74.2</td>
<td>72.6</td>
<td>70.3</td>
</tr>
<tr>
<td>0.2</td>
<td>67.3</td>
<td>78.1</td>
<td>69.8</td>
<td>55.1</td>
<td>53.3</td>
<td>75.2</td>
<td>79.5</td>
<td>82.8</td>
<td>49.6</td>
<td>78.5</td>
<td>65.3</td>
<td>83.0</td>
<td>74.7</td>
<td>72.2</td>
<td>65.0</td>
<td>69.1</td>
<td>66.0</td>
<td>73.7</td>
<td>70.4</td>
<td>69.5</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>67.1</td>
<td>78.4</td>
<td>66.2</td>
<td>53.2</td>
<td>54.2</td>
<td>76.7</td>
<td>80.0</td>
<td>81.9</td>
<td>47.7</td>
<td>75.5</td>
<td>63.9</td>
<td>81.2</td>
<td>82.3</td>
<td>75.2</td>
<td>76.7</td>
<td>38.8</td>
<td>69.5</td>
<td>61.8</td>
<td>73.4</td>
<td>71.9</td>
<td>68.8</td>
</tr>
<tr>
<td>0.4</td>
<td>67.3</td>
<td>77.1</td>
<td>63.4</td>
<td>55.8</td>
<td>51.2</td>
<td>74.8</td>
<td>80.2</td>
<td>80.8</td>
<td>47.1</td>
<td>73.8</td>
<td>63.9</td>
<td>77.5</td>
<td>81.7</td>
<td>74.5</td>
<td>75.6</td>
<td>37.7</td>
<td>68.5</td>
<td>64.2</td>
<td>75.7</td>
<td>70.3</td>
<td>68.1</td>
</tr>
<tr>
<td>0.5</td>
<td>68.1</td>
<td>78.0</td>
<td>61.7</td>
<td>51.7</td>
<td>50.4</td>
<td>74.8</td>
<td>78.4</td>
<td>82.1</td>
<td>46.9</td>
<td>72.6</td>
<td>63.3</td>
<td>78.2</td>
<td>80.9</td>
<td>72.7</td>
<td>74.9</td>
<td>36.4</td>
<td>64.4</td>
<td>67.9</td>
<td>71.7</td>
<td>70.1</td>
<td>67.3</td>
</tr>
<tr>
<td>0.6</td>
<td>64.4</td>
<td>76.9</td>
<td>61.2</td>
<td>47.9</td>
<td>47.1</td>
<td>73.2</td>
<td>78.1</td>
<td>81.8</td>
<td>43.0</td>
<td>72.4</td>
<td>60.8</td>
<td>78.5</td>
<td>79.0</td>
<td>73.3</td>
<td>72.3</td>
<td>32.7</td>
<td>64.5</td>
<td>65.6</td>
<td>73.8</td>
<td>68.8</td>
<td>65.8</td>
</tr>
<tr>
<td>0.7</td>
<td>65.9</td>
<td>73.4</td>
<td>59.0</td>
<td>51.7</td>
<td>42.1</td>
<td>70.3</td>
<td>74.2</td>
<td>78.6</td>
<td>40.9</td>
<td>72.1</td>
<td>57.9</td>
<td>76.6</td>
<td>78.8</td>
<td>72.5</td>
<td>70.0</td>
<td>30.9</td>
<td>63.6</td>
<td>61.8</td>
<td>73.3</td>
<td>68.2</td>
<td>64.1</td>
</tr>
<tr>
<td>0.8</td>
<td>59.5</td>
<td>74.7</td>
<td>57.5</td>
<td>41.2</td>
<td>41.7</td>
<td>72.0</td>
<td>73.2</td>
<td>77.7</td>
<td>37.2</td>
<td>66.9</td>
<td>59.3</td>
<td>73.5</td>
<td>78.2</td>
<td>68.3</td>
<td>64.0</td>
<td>27.4</td>
<td>61.7</td>
<td>61.9</td>
<td>72.5</td>
<td>68.9</td>
<td>61.9</td>
</tr>
<tr>
<td>0.9</td>
<td>56.4</td>
<td>71.9</td>
<td>46.0</td>
<td>32.6</td>
<td>34.5</td>
<td>70.9</td>
<td>69.1</td>
<td>73.3</td>
<td>32.6</td>
<td>65.2</td>
<td>46.4</td>
<td>69.7</td>
<td>74.0</td>
<td>67.0</td>
<td>59.3</td>
<td>24.6</td>
<td>55.2</td>
<td>49.9</td>
<td>66.4</td>
<td>67.2</td>
<td>56.6</td>
</tr>
<tr>
<td>0.1</td>
<td>81.0</td>
<td>73.8</td>
<td>70.6</td>
<td>49.9</td>
<td>50.5</td>
<td>72.7</td>
<td>74.3</td>
<td>86.9</td>
<td>44.5</td>
<td>69.7</td>
<td>42.5</td>
<td>85.4</td>
<td>76.7</td>
<td>79.0</td>
<td>78.5</td>
<td>43.3</td>
<td>70.3</td>
<td>52.9</td>
<td>72.8</td>
<td>63.4</td>
<td>66.9</td>
</tr>
<tr>
<td>0.2</td>
<td>80.6</td>
<td>74.1</td>
<td>66.0</td>
<td>48.0</td>
<td>51.3</td>
<td>72.3</td>
<td>75.4</td>
<td>87.0</td>
<td>43.7</td>
<td>69.9</td>
<td>35.9</td>
<td>85.0</td>
<td>77.4</td>
<td>79.9</td>
<td>76.8</td>
<td>41.5</td>
<td>71.2</td>
<td>52.6</td>
<td>70.8</td>
<td>62.8</td>
<td>66.1</td>
</tr>
<tr>
<td>0.3</td>
<td>79.7</td>
<td>74.1</td>
<td>66.8</td>
<td>47.8</td>
<td>49.8</td>
<td>72.6</td>
<td>72.5</td>
<td>85.8</td>
<td>43.3</td>
<td>70.3</td>
<td>35.2</td>
<td>83.6</td>
<td>75.3</td>
<td>77.9</td>
<td>75.6</td>
<td>41.7</td>
<td>68.8</td>
<td>51.9</td>
<td>71.4</td>
<td>61.9</td>
<td>65.3</td>
</tr>
<tr>
<td>0.4</td>
<td>79.0</td>
<td>74.7</td>
<td>67.1</td>
<td>45.2</td>
<td>47.7</td>
<td>71.4</td>
<td>71.7</td>
<td>85.0</td>
<td>41.8</td>
<td>67.5</td>
<td>38.3</td>
<td>83.8</td>
<td>73.7</td>
<td>78.4</td>
<td>27.2</td>
<td>68.9</td>
<td>54.2</td>
<td>70.7</td>
<td>62.6</td>
<td>64.6</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>78.3</td>
<td>73.6</td>
<td>67.2</td>
<td>43.8</td>
<td>47.7</td>
<td>71.3</td>
<td>71.7</td>
<td>85.8</td>
<td>41.2</td>
<td>67.9</td>
<td>31.3</td>
<td>82.3</td>
<td>75.4</td>
<td>78.8</td>
<td>69.5</td>
<td>40.0</td>
<td>66.6</td>
<td>51.8</td>
<td>66.3</td>
<td>63.6</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>77.9</td>
<td>71.9</td>
<td>64.1</td>
<td>43.2</td>
<td>47.2</td>
<td>67.9</td>
<td>69.9</td>
<td>85.1</td>
<td>37.6</td>
<td>66.0</td>
<td>27.7</td>
<td>82.8</td>
<td>73.9</td>
<td>78.3</td>
<td>67.7</td>
<td>36.5</td>
<td>65.7</td>
<td>50.2</td>
<td>66.0</td>
<td>61.6</td>
<td>62.1</td>
</tr>
<tr>
<td>0.7</td>
<td>76.5</td>
<td>69.6</td>
<td>60.7</td>
<td>38.6</td>
<td>44.5</td>
<td>65.7</td>
<td>68.5</td>
<td>83.7</td>
<td>35.1</td>
<td>65.9</td>
<td>25.7</td>
<td>82.0</td>
<td>73.7</td>
<td>76.0</td>
<td>60.9</td>
<td>34.3</td>
<td>60.4</td>
<td>48.4</td>
<td>67.7</td>
<td>60.1</td>
<td>59.9</td>
</tr>
<tr>
<td>0.8</td>
<td>74.6</td>
<td>69.1</td>
<td>56.9</td>
<td>36.3</td>
<td>41.7</td>
<td>63.8</td>
<td>64.5</td>
<td>82.4</td>
<td>32.1</td>
<td>61.6</td>
<td>20.5</td>
<td>80.3</td>
<td>71.0</td>
<td>75.5</td>
<td>55.2</td>
<td>29.0</td>
<td>59.2</td>
<td>48.3</td>
<td>55.1</td>
<td>58.8</td>
<td>56.8</td>
</tr>
<tr>
<td>0.9</td>
<td>68.0</td>
<td>66.0</td>
<td>49.2</td>
<td>27.5</td>
<td>37.4</td>
<td>57.8</td>
<td>61.0</td>
<td>81.7</td>
<td>25.4</td>
<td>59.2</td>
<td>13.2</td>
<td>78.7</td>
<td>65.3</td>
<td>72.9</td>
<td>47.3</td>
<td>28.8</td>
<td>54.9</td>
<td>41.8</td>
<td>53.8</td>
<td>57.7</td>
<td>52.4</td>
</tr>
</tbody>
</table>
### Table A.3: Table Representation for Raw Experiment Data

|         | H   | CR | Budget Percentages | Full | Budget | 10.8 | 15.8 | 20.8 | 25.8 | 30.8 | 35.7 | 40.8 | 45.7 | 50.7 | 55.7 | 60.7 | 65.8 | 80.8 | 92.8 | 96.8 | 99.8 |
|---------|-----|----|--------------------|------|--------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| RS  ✓ H | 0.414 | 0.508 | 0.547 | 0.572 | 0.597 | 0.622 | 0.639 | 0.643 | 0.652 | -    | -    | -    | 0.692 | 0.71 |    |     |
| US  ✓ H | 0.450 | 0.506 | 0.519 | 0.552 | 0.590 | 0.621 | 0.623 | 0.651 | 0.657 | -    | -    | -    | 0.695 | 0.71 |    |     |
| SUS  ✓ H | 0.397 | 0.502 | 0.546 | 0.575 | 0.602 | 0.623 | 0.633 | 0.636 | 0.648 | -    | -    | -    | 0.682 | 0.71 |    |     |
| LAL  ✓ H | 0.429 | 0.495 | 0.545 | 0.569 | 0.58  | 0.606 | 0.634 | 0.64  | 0.644 | -    | -    | -    | 0.692 | 0.71 |    |     |
| RS  ✓ H | 0.408 | 0.562 | 0.576 | 0.608 | 0.612 | 0.637 | 0.651 | 0.655 | 0.663 | -    | -    | -    | 0.71 |    |     |
| US  ✓ H | 0.433 | 0.462 | 0.586 | 0.62  | 0.638 | 0.647 | 0.658 | 0.672 | 0.676 | -    | -    | -    | 0.71 |    |     |
| OPT-US ✓ H | 0.433 | 0.529 | 0.594 | 0.624 | 0.643 | 0.646 | 0.665 | 0.668 | 0.679 | 0.689 | 0.687 | 0.691 | 0.71 |    |     |
| OPT-LAL ✓ H | 0.389 | 0.529 | 0.59 | 0.62  | 0.638 | 0.645 | 0.661 | 0.674 | 0.677 | -    | -    | -    | 0.71 |    |     |
| US  ✓ L | 0.412 | 0.498 | 0.509 | 0.527 | 0.537 | 0.613 | 0.626 | 0.653 | 0.664 | 0.675 | 0.673 | -    | 0.71 |    |     |
| RS  ✓ L | 0.412 | 0.492 | 0.519 | 0.517 | 0.539 | 0.587 | 0.614 | 0.633 | 0.641 | 0.662 | 0.664 | -    | 0.71 |    |     |
| OPT  ✓ L | 0.42 | 0.518 | 0.548 | 0.586 | 0.609 | 0.62  | 0.648 | 0.665 | 0.676 | 0.67 | 0.683 | 0.689 | 0.71 |    |     |
| RS  ✓ M | 0.443 | 0.554 | 0.576 | 0.5  | 0.62  | 0.623 | 0.652 | 0.655 | 0.672 | 0.675 | 0.689 | -    | 0.71 |    |     |
| OPT  ✓ M | 0.433 | 0.544 | 0.582 | 0.608 | 0.632 | 0.643 | 0.652 | 0.664 | 0.677 | 0.684 | 0.687 | 0.693 | 0.71 |    |     |
| Budget | -    | -   | 10.8  | 20.8  | 30.8  | 40.8  | 50.8  | 60.8  | 70.8  | 80.8  | 90.8  | -    | -    | -    | -    |
| VOC0712 ✓ M | 0.356 | 0.506 | 0.511 | 0.641 | 0.674 | 0.672 | 0.696 | 0.701 | 0.715 | 0.717 | 0.72 | -    | -    | 0.74 |    |     |
| Budget | -    | -   | 11.09 | 13.62 | 18.1 | 22.58 | 27.06 | 31.53 | 36.15 | 40.49 | 44.97 | 49.45 | 53.93 | -    | -    | -    |
| RL  ✓ H | 0.433 | 0.525 | 0.561 | 0.579 | 0.603 | 0.615 | 0.634 | 0.645 | 0.648 | 0.657 | 0.664 | -    | -    | 0.71 |    |     |

* H for Hybrid Learning  
† CR for cost ratio of strong and weak annotations
A.4 Definition of Easy, Medium and Hard Classes

We analyze the cost and number of images selected at every active selection step to investigate which type of training examples is more helpful in the sequential training process. Based on the final mAP of each category in VOC07, we divide the twenty categories into three groups: Easy, Medium and Hard. The mapping is motivated and shown in Table 2.

Table A.4: The mapping from twenty categories in PASCAL VOC to Easy, Medium and Hard classes. This mapping is based on the final mAP of each category.

<table>
<thead>
<tr>
<th>Class</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>88.9</td>
<td>88.4</td>
<td>88.1</td>
</tr>
<tr>
<td>Category</td>
<td>cat</td>
<td>car</td>
<td>bus</td>
</tr>
<tr>
<td>mAP</td>
<td>82.3</td>
<td>82.0</td>
<td>80.4</td>
</tr>
<tr>
<td>Category</td>
<td>person</td>
<td>bike</td>
<td>sheep</td>
</tr>
<tr>
<td>mAP</td>
<td>75.8</td>
<td>65.7</td>
<td>68.9</td>
</tr>
<tr>
<td>Category</td>
<td>sofa</td>
<td>table</td>
<td>boat</td>
</tr>
</tbody>
</table>

* data is from the last setup of Table 6 in the report [66].

B Solutions to Optimization Functions

B.1 Image Sampling

In section 3.1.1 and 3.1.2 of the submission, we propose two linear functions to approximate the mAP increment. Then we maximize it with several constraints. However, the proposed solution is an integer programming problem, and it cannot be solved in polynomial time. For example, the default solver uses Branch and Bound Algorithm (B&B) and it takes more than 24 hours to find a global maximum. To solve this problem more efficiently, we take the relaxation of the original integer program shown as Eq. B.1. Noted that the solution space of $x_1, x_2, x_3$ becomes an interval $[0, 1]$, but not
a discrete set \{0, 1\}, the relaxation Eq. B.1 can be solved in linear time.

\[
\hat{x}_1, \hat{x}_2, \hat{x}_3 = \arg \max_{x_1, x_2, x_3 \in [0, 1]^N} s^\top(x_1 + x_3) + (\mu - s)^\top x_2 \\
\text{s.t.} \quad x_3 \leq \psi \\
\quad x_1 + x_2 \leq 1 - \psi \\
1^\top(ax_1 + bx_2 + cx_3) \leq d \\
1^\top(ax_1 + bx_2 + cx_3) \geq d - a
\] (B.1)

To project \(x_1, x_2, x_3\) back to the discrete set \{0, 1\}, we take three float numbers \(\epsilon_1, \epsilon_2, \epsilon_3\) as the thresholds of \(\hat{x}_1, \hat{x}_2, \hat{x}_3\), respectively. In another word, every element in \(x_k\) larger than \(\epsilon_k\) is set as 1, otherwise it is 0.

\[
\begin{align*}
x_1^* &= 1\{\hat{x}_1 > \epsilon_1\} \\
x_2^* &= 1\{\hat{x}_2 > \epsilon_2\} \\
x_3^* &= 1\{\hat{x}_3 > \epsilon_3\}
\end{align*}
\] (B.2)

The solution needs to be feasible after thresholding. More concretely, the first two constraints must be satisfied because we cannot give an invalid action (*e.g.* It’s impossible to annotate an image both weakly and strongly, but it might appear from the solution).

We set \(\epsilon_1 = \epsilon, \epsilon_2 = 1 - \epsilon, \epsilon_3 = \epsilon\), where \(\epsilon\) exhaustively goes from 0 to 1 to satisfy the last two budget constraints. In this case, it can be proven that the solution of Eq. B.2 is also feasible from the original constraints.

*Proof.* (1) If \(\psi(k) = 0\), where \(k\) can be the index of any element of vector \(\hat{x}_1\). Since \(\hat{x}_1, \hat{x}_2, \hat{x}_3\) is feasible, the first inequality in Eq. B.1 shows

\[
\hat{x}_3(k) \leq 0 \implies \hat{x}_3(k) = 0.
\] (B.3)
The second inequality in Eq. B.1 gives the inequality for \( \hat{x}_1 \) and \( \hat{x}_2 \).

\[
\hat{x}_1(k) + \hat{x}_2(k) \leq 1 \implies 1 - \hat{x}_2(k) \geq \hat{x}_1(k) \implies 1\{1 - \hat{x}_2(k) < \epsilon\} \leq 1\{\hat{x}_1(k) < \epsilon\}.
\] (B.4)

Here \( 1\{\} \) is the indicator function. When we apply Eq. B.2 to check the two restrictions, the discrete solutions \( x_1^*, x_2^*, x_3^* \) are still feasible.

\[
x_3^*(k) = 1\{\hat{x}_3(k) > \epsilon\} = 1\{0 > \epsilon_3\} = 0 = \psi(k)
\]

\[
x_1^*(k) + x_2^*(k) = 1\{\hat{x}_1(k) > \epsilon\} + 1\{\hat{x}_2(k) > 1 - \epsilon\} = 1\{\hat{x}_1(k) > \epsilon\} + 1\{1 - \hat{x}_2(k) < \hat{x}_2(k)\} \leq 1\{\hat{x}_1(k) > \epsilon\} + 1\{\hat{x}_1(k) < \hat{x}_2(k)\} \leq 1 = 1 - \psi(k).
\] (B.5)

(2) If \( \psi(k) = 1 \), we have

\[
\hat{x}_3(k) \leq 1 \implies \hat{x}_3(k) = 1
\]

\[
\hat{x}_1(k) + \hat{x}_2(k) \leq 0 \implies \hat{x}_1(k), \hat{x}_2(k) = 0
\] (B.6)

Similar to previous discussion, we can have \( x_3^*(k) \leq 1 \) and \( x_1^*(k) = 0, x_2^*(k) = 0 \) to fit the two constraints.

\[\square\]

**B.2 Post Processing**

In the article, we give another optimization function to do post processing. This procedure checks the consistency between the image and instance-level annotations, and it removes abundant pseudo labels.

Given an image with the weakly annotation \( \omega \in \{0, 1\}^K \), where \( K \) is the number of categories. We assume the first detection model (teacher model) gives \( M \) positive predictions for the localization \( P \in \mathbb{R}^{M \times 4} \), classification \( A \in \{0, 1\}^{M \times K} \), and the positive class confidence vector \( q \in [0, 1]^M \). Eventually, the post processing returns a
Algorithm 2 Post-Processing for Noise Cleaning

Require: Weak annotation $\omega \in \{0,1\}^K$, localization $P \in \mathbb{R}^{M \times 4}$, classification $A \in \{0,1\}^{M \times K}$, the positive class confidence for $M$ predictions $q \in [0,1]^M$, $\alpha$, $\beta$;

1: for $i=1:M$ do
2: if not $A(i,:) < \omega$
3: $y(i) = 0$;
4: end for
5: $y' = \text{NMS index for } [P,A,q]$, do it by class with threshold $\alpha$;
6: $y = y' \cdot y'$;
7: if $\text{sum}(y) > \beta$
8: assign the top $\beta$ remained predictions to $y$;

Ensure: a binary vector $y \in \{0,1\}^M$

sparse $M$-dimensional binary vector $y$ as Eq. B.7, where $\alpha = 0.3$, $\beta = 3$.

$$
\min_{y \in \{0,1\}^M} -y^\top q \\
\text{s.t. } y^\top A(1 - \omega) = 0 \\
\forall_{y_i,y_j=1} \text{IoU}(P_i, P_j) \leq \alpha \\
|y|_0 \in [1,\beta]
$$

(B.7)

In this problem, our objective function takes high confidence predictions by choosing $y \in \{0,1\}^M$. In the first constraint, $y^\top A$ accumulates the selected predictions by their categories. The following inner product with $(1 - \omega)$ returns zero only when all the predicted classes are in the weak annotation. On the other hand, the second constraint removes the heavily overlapped predictions while the third one makes the binary vector $y$ sparse. We develop Alg. 2 to solve the above optimization.
C Papers Submitted and Under Preparation
