Many-shot from Low-shot: Learning to Annotate using Mixed Supervision for Object Detection

Carlo Biffi[1], Steven McDonagh[1], Philip Torr[3], Aleš Leonardis[1], and Sarah Parisot[1,2]

[1] Huawei Noah's Ark Lab
[2] Mila Montréal
[3] University of Oxford

Abstract. Object detection has witnessed significant progress by relying on large, manually annotated datasets. Annotating such datasets is highly time consuming and expensive, which motivates the development of weakly supervised and few-shot object detection methods. However, these methods largely underperform with respect to their strongly supervised counterpart, as weak training signals *often* result in partial or oversized detections. Towards solving this problem we introduce, for the first time, an online annotation module (OAM) that learns to generate a many-shot set of *reliable* annotations from a larger volume of weakly labelled images. Our OAM can be jointly trained with any fully supervised two-stage object detection method, providing additional training annotations on the fly. This results in a fully end-to-end strategy that only requires a low-shot set of fully annotated images. The integration of the OAM with Fast(er) R-CNN improves their performance by 17% mAP, 9% AP50 on PASCAL VOC 2007 and MS-COCO benchmarks, and significantly outperforms competing methods using mixed supervision.

1 Introduction

Object detection is an essential building block of many computer vision systems [26]. State-of-the-art (SOTA) methods mainly rely on large scale datasets with manually annotated bounding boxes to train fully supervised CNN-based models [8, 17, 16, 12, 3]. However, the prohibitive cost and time requirements associated with data annotation reduce the applicability of SOTA detection models in real life scenarios. This has motivated research on object detection strategies with reduced data annotation requirements. Amongst the most popular low data regimes, we distinguish Weakly Supervised Object Detection (WSOD), which aims to train object detectors using only image-level annotations [2, 18, 21, 1, 25], and Few-Shot or Low-Shot Object Detection (FSOD/LSOD), training supervised models with only a handful of training examples on all (LSOD) or only a subset of novel test classes (FSOD) [10, 24, 5]. FSOD and in particular WSOD have been the focus of a large body of work with innovative strategies obtaining promising performance. Nonetheless, these models typically fall far short of their strongly supervised counterparts. Numerical performance gaps are attributed to

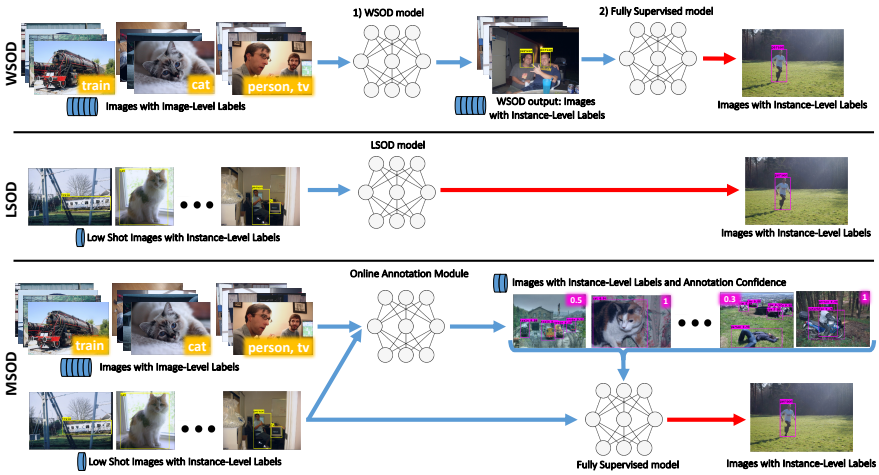

Fig. 1. Weak and low data detection strategies and our proposed mixed supervision-based setting. First row: Weakly Supervised Object Detection (WSOD) models learn to annotate images with image-level annotations which are then used to train fully supervised models. WSOD annotations are often partial or oversized, resulting in poor detector performance. Second row: Few-shot or low-shot object detection (LSOD) trains models on only a handful of training examples. Research mainly focuses on situations where only a specific subset of novel classes have limited training data. Bottom row: Mixed Supervision for Object Detection (MSOD) combines a low shot set of images containing object annotations with a large volume of images comprising only image-level annotations. We train an online annotation module to generate a many-shots set which, at the same time, is used to train a fully supervised model.

the low quality of bounding-box annotations produced, *e.g.* by WSOD methods, that often manifest as partial or oversized boxes. Such results are not reliable enough for use in real-world scenarios and can be observed to cause deterioration of detection performance when used in fully supervised models training. This can be attributed to weak training signals requiring very large and curated datasets (WSOD) or very representative and carefully selected annotated examples (FSOD).

To address the aforementioned challenges, we focus on a recent training paradigm relying on Mixed Supervision for Object Detection (MSOD) [14, 7]. The distinction between this protocol and the previously introduced weak and low data settings is illustrated in Fig. 1. The objective of MSOD is to exploit and combine the complementary advantages provided by WSOD and LSOD; weak (image-level) supervision affords the construction of large databases with minimal effort, while low-shot supervision provides information rich, fully annotated ground truth examples. The MSOD paradigm has, only very recently, been initially investigated in two related works. Fang *et al.* [7] propose a cascaded architecture yielding performance competitive with fully supervised counterparts yet using a significant fraction of the full training data to achieve comparable

performance. Pan *et al.* [14] use low-shot examples to refine bounding box annotations obtained from a pre-trained WSOD model [19], resulting in a method intrinsically linked to the performance and drawbacks of WSOD techniques.

In this work, we approach the MSOD scenario from a different angle. Due to the sparsity of rich training information provided, we expect a MSOD model to output annotations of variable quality, especially for images containing crowded scenes or objects with appearance substantially dissimilar to the training data. In contrast to existing MSOD models we introduce an Online Annotation Module (OAM), trained with mixed supervision, that can be used in conjunction with any two-stage fully supervised object detection method to improve its performance (*e.g.* Fast(er) R-CNN family [8, 17]). Our OAM generates, on the fly, additional reliable automated annotations obtained from a larger set of weakly annotated images (containing only image-level class labels). Furthermore, we exploit prediction stability to reason about annotation reliability resulting in associated confidence scores. Generated annotations are used to train, concurrently to the OAM, a fully supervised detector that shares the same encoding features. This produces an intrinsic training curriculum for the standard detector model; only simple images, labelled with high confidence will be presented to the model at the outset. Compared to previous MSOD work, our OAM strategy provides increased robustness against mislabeled crowded and ambiguous training images as only confident MSOD annotations are exploited for fully supervised training. Furthermore, our joint MSOD and fully supervised training provides intrinsic regularisation for both tasks, allowing the learning of higher quality and more discriminative feature extractors.

Experiments show that our strategy allows effective training of standard detection algorithms with only minimal annotation requirements and significantly outperforms WSOD and competitive MSOD approaches on PASCAL VOC 2007 and MS-COCO benchmarks. Additionally, we report competitive performance in comparison to fully supervised alternatives, illustrating the ability of our OAM to annotate a many-shot set of (weakly labelled) images that can be leveraged to improve the fully supervised model performance.

In summary, we propose a new direction using Mixed Supervision for Object Detection (MSOD). Our main contributions are the following:

- We introduce a novel Online Annotation Module (OAM), trained using mixed supervision. This module allows expansion of the low-shot training set of fully annotated images by generating *reliable* annotations from a larger volume of weakly labelled images.
- Training our OAM concurrently with any two-stage object detection model introduces a strategy for object detection performance improvement due to the generated annotation. We report on the benefits of intrinsic regularisation afforded to both tasks when common encoding features are shared.
- The integration of the OAM with Fast(er) R-CNN improves their performance by 17% mAP, 9% AP50 on PASCAL VOC 2007 and MS-COCO benchmarks, and significantly outperforms MSOD approaches.

2 Related work

Weakly Supervised Object Detection. A large body of recent work, considering WSOD, couples CNN feature extractors with Multiple Instance Learning (MIL) frameworks, thus casting weakly supervised object detection as a multi-label classification problem. Each image is typically represented as a bag of pre-computed proposals (*e.g.* Selective Search [20], Edge Boxes [27], etc.) and the objective is to identify proposals that are most relevant for bag classification [2, 18, 21, 25]. Being framed as a classification task, MIL WSOD models typically focus on proposals that comprise of either the most discriminative object parts or image regions that define the presence of an object category. They therefore struggle to detect full object extent (*e.g.* human faces in contrast to an entire human body) or group multiple object instances of the same object within a single bounding box [25, 14]. In order to address this issue, recent work has focused on bounding-box refinement strategies using cascaded refinements of MIL classifications [19, 18], using saliency maps [23, 25], adopting continuation strategies [21, 22] and modelling uncertainty [1]. However, the ill-posed nature of the WSOD problem and insufficient statistics provided by the PASCAL VOC dataset (on which these approaches are usually evaluated) has lead to the development of ad-hoc training strategies and parameter sensitive methods to cope with the weak training signal, which substantially reduce generalisability across datasets. In this paper, we argue that including a handful of labelled samples yields accuracy and stability model improvements at only minimal annotation cost. Usually, all the images annotated by MIL WSOD methods are used, in a second step, to train fully supervised models [18, 21, 25]. Further previous work has also focused on alternating between the pseudo-labelling of images and, in conjunction, training a fully supervised model [9, 5]. In this work, we generate bounding box annotations on the fly from mixed supervision and we concurrently train a fully supervised detector *only* on the images annotated with high confidence.

Few-Shot and Mixed Supervision Object Detection. Few-Shot Object Detection (FSOD) considers a fully supervised training set, and aims to achieve strong performance on a set of novel classes comprising of only K annotated training images per class. To date only a handful of works have focused on FSOD [10, 24, 11, 4]. Such approaches typically adapt few-shot classification techniques to the object detection setting, exploring metric learning [11] or meta-learning [24] strategies. Mixed Supervision for Object Detection (MSOD) enhances a WSOD training set containing only image-level labels with a small set of fully annotated (strong) images (*e.g.* K images per class, analogous to an FSOD scenario) and aims to achieve strong performance on *all* the training classes. Pan *et al.* recently propose BCNet [14], which learns to refine the output of a pre-trained WSOD model using a small set of strong images. The definition of small set explored in their work ranges from 10 shots to 20% of the entire dataset ($\sim$1000 images in PASCAL VOC 2007 training set). This approach provides a strong performance increase with respect to WSOD methods, however remains highly dependent on the original WSOD model detections as input. If

detections are originally missed by the pre-trained model, the approach cannot recover. Moreover, BCNet requires the training of two independent models which makes the adaption of WSOD parameters, *i.e.* training for new datasets, challenging. In this work, we instead propose a one-stage approach relying on an adaptive pool of annotations, updated dynamically as training progresses. EHSOD [7] and BAOD [15] focus on larger data regimes (*e.g.* 10% to 90%) and aim to reduce the data required to reach fully supervised performance using a cascaded MIL model and a student-teacher setup trained on weak and strong annotations, respectively. In contrast to all outlined methods, we propose instead to learn, and annotate on the fly, only a subset of weak images that can be labelled with high confidence. These additional samples are then used together with strong images to train an object detector and thus improve performance.

3 Method

Let $\mathcal{I}$ be a set of training images annotated with image-level supervision. Under our mixed supervision paradigm, a subset of these images, $\mathcal{S} \subset \mathcal{I}$ with $|\mathcal{S}| \ll |\mathcal{I}|$, is further annotated with bounding box annotations. We refer to the images contained in $\mathcal{S}$ as *strong* training images, while the images in $\mathcal{W} = \mathcal{I} \setminus \mathcal{S}$, that have only image-level annotations, are referred to as *weak* training images. An overview of our proposed method is reported in Fig. 2. Our model comprises two branches with shared encoder backbone, and employs an ROI pooling layer to compute a fixed-length feature representation for each image bounding box proposal. The first branch of our model employs both weak and strong training images to learn an Online Annotation Module (OAM) for weak training images. The OAM generates bounding box annotations, with associated confidence scores, on the fly, for every weak training image. Annotated weak images are added to a third set of images, $\mathcal{P} \subset \mathcal{W}$, if they have been annotated with high confidence, and can be subsequently removed if their annotation confidence drops during training. Images contained in $\mathcal{P}$ are referred to as *semi-strong* training images throughout the paper. The second branch of our model is designed as a standard fully supervised component and trained, in parallel, in an end-to-end manner using strong and semi-strong images. At testing, only the fully supervised model is used for object detection.

Given an input image, we first compute a set of B candidate proposals $\{b_r\}_{r=1}^{B}$, using either an unsupervised method (*e.g.* Selective Search [20] or Edge Boxes [27]) or a Region Proposal Network (RPN) [17], and their associated feature vectors $\{\boldsymbol{\xi_r}\}_{r=1}^{B}$. These feature vectors $\{\boldsymbol{\xi_r}\}_{r=1}^{B}$ are obtained using a standard CNN backbone and ROI Pooling layer and provide a common input to both of our model branches: the OAM and the fully supervised branch.

3.1 Online Annotation Module

Our OAM is designed to jointly exploit weak and strong supervision in an efficient manner. It comprises three main components: 1) a joint detection module

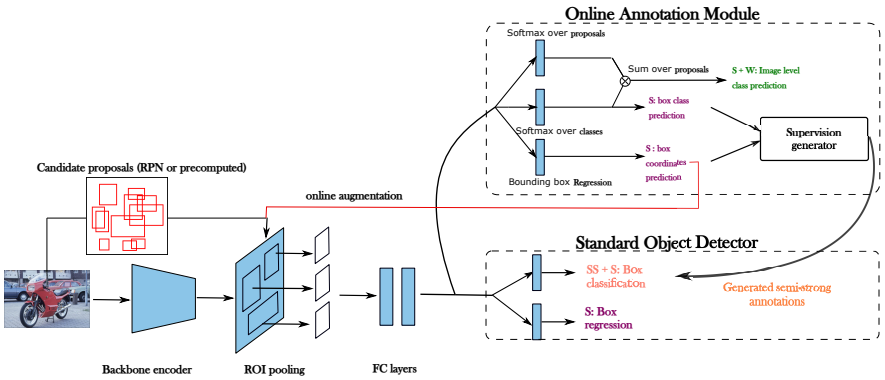

Fig. 2. Architecture of the proposed approach. Our model comprises two branches with a shared encoder. Our Online Annotation Module (OAM) is trained on weakly ($\mathcal{W}$) and strongly ($\mathcal{S}$) annotated images to generate, on the fly, confident annotations for $\mathcal{W}$ images which are added to a pool of semi-strong (SS) images. The model's second branch uses SS and $\mathcal{S}$ images to train a standard fully supervised detection model.

exploiting weak and strong labels in a single, common architecture to predict bounding boxes and their classes, 2) an online bounding box augmentation step that generates refined bounding box proposals, 3) a supervision generator, identifying confident annotations to be used as supervision. We next describe all three components in detail.

Joint detection module. Similarly to the strategy proposed in [7], we combine a multiple instance learning (MIL) type image-level classification task with a fully supervised joint classification and regression task. Our joint detection module hence comprises three parallel, fully connected layers focusing on three different subtasks: proposal scoring, classification and regression (Fig. 2, online annotation module block). Proposal scoring $\boldsymbol{\gamma}_C(c,l)$ and classification $\boldsymbol{\gamma}_R(c,l) \in \mathbb{R}^{C\times B}$ are obtained by applying the softmax function to the output of their layers along both dimensions, independently, (classes for $\boldsymbol{\gamma}_C$, proposals for $\boldsymbol{\gamma}_R$). After this operation, $\boldsymbol{\gamma}_C(c,l)$ represents the probability that the l-th proposal belongs to class c, while $\boldsymbol{\gamma}_R(c,l)$ represents the proportional contribution that proposal l provides to the image being classified as class c. Following [2], these layers are trained by exploiting the image-level supervision of both strong and weak images. In particular, a proposal score $\boldsymbol{\phi}_P = \boldsymbol{\gamma}_C \odot \boldsymbol{\gamma}_R$, per class, is obtained by combining them where $\odot$ is a Hadamard product. Then, summing these scores over proposals, $\alpha_c = \sum_{r=1}^{R} \boldsymbol{\phi}_P$, enables the use of a binary cross-entropy loss as image-level loss function:

$$L_{gc}(\alpha_c, y_c) = -\sum_{c=1}^{C}[(1-y_c)log(1-\alpha_c) + y_c log(\alpha_c)] \tag{1}$$

where y_c is the label indicating the presence or absence of class c in an image.

Similar to traditional object detectors, we use strong images to compute bounding box regression and classification via the corresponding fully connected layers. We therefore combine weak and strong supervision by providing direct supervision to proposal-level class prediction $\boldsymbol{\gamma}_c$. For regression, each bounding box b is parametrised as a four-tuple (x, y, h, w) that specifies its center coordinate (x, y) and its height and width (h, w). For each proposal classified as foreground in a strong image, this regression branch predicts the offset of these coordinates $t^k = (t_x, t_y, t_h, t_w)$. Hence, for every strong image, the following additional loss is computed on a batch of M proposals:

$$L_p(\gamma, u, t, v) = L_{cls}(\gamma, u) + 1[u \geq 1] L_{reg}(t, v) \tag{2}$$

where:

$$L_{cls} = -\frac{1}{M} \sum_{r=1}^{M} \sum_{c=1}^{C+1} u_{cr} \log(\gamma_{cr}), \quad L_{reg}(t, v) = \sum_{i \in (x,y,h,w)} \text{smooth}_{L1}(t_i - v_i) \tag{3}$$

Parameters γ and u constitute the predicted and target proposal classes respectively, t and v the predicted and target bounding box offsets respectively and smooth_{L1} is a smooth $L1$ loss function [8].

The loss function of the joint detection module is hence $L_{I_s} = L_p + L_{gc}$ on strong images, while the loss function on weak images is $L_{I_w} = L_{gc}$. Enforcing synergy between the two types of supervision regularises the low-shot task thanks to the statistical information provided by weak images. Moreover, due to the instance-level annotations provided by strong images, this also constrains the MIL task and encoder to learn stronger discriminative features between full and partial-extent object proposals.

Online Bounding Box Augmentation Strategy. Learning to update and improve bounding box spatial regions via low-shot regression is highly challenging. When initial inference and ground-truth box overlap is small, large corrections (spatial offsets) are required. Previous work (BCNet [14]) actively elects to exclude such challenging samples, further reducing already highly limited data. We alternatively fully exploit available annotations and push our regression branch output through a second forward pass of our OAM (red arrow in Fig. 2).

More specifically; after the first forward pass, we select the M top scoring proposals, per class, corresponding to image-level ground-truth. M is defined as half the size of the proposal batch used to train the strongly supervised component. This accounts for the presence of irrelevant background proposals and allows us to fix this hyperparameter. Once regression branch offsets have been applied, our ROI pooling layer ingests the proposals and yields a new set of bounding box features. Loss functions are evaluated using the updated boxes features and combined with the first pass loss. The overall loss function of our OAM branch is then: $L_{1B} = L^I_{I_s} + L^I_{I_w} + L^{II}_{I_s} + L^{II}_{I_w}$, where superscripts I and II indicate the first and second pass, respectively. At every iteration, a batch with the same number of weak and strong images is used.

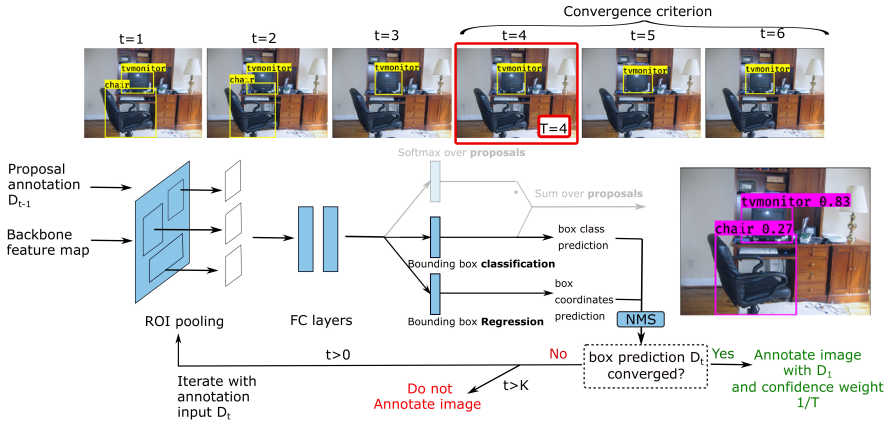

Fig. 3. Proposed online pseudo-supervision generation strategy. At each iteration, a new set of bounding boxes D_t is computed via classification, regression and NMS of the features from previous set D_{t-1}. If bounding box predictions converge at iteration $T < K$, and all proposal classes agree with the image-level label, the weak image is annotated.

Motivation for our second pass is two-fold. Firstly augmentation is intrinsically provided as new sets of proposal candidates are generated for regression and classification task training. In contrast pre-computed proposals (predominant in WSOD), that lack additional external augmentation strategies, provide only static input, reducing sample variability during training. Secondly, our regression task is regularised as any weak proposals receiving modifications that hinder correct image-level classification are penalised.

Online Pseudo-Supervision Generation. The key objective of our OAM is to generate reliable annotations on a large set of weakly labelled images in order to guide the training of a fully supervised second branch. As the OAM is trained concurrently with the second branch, it is critical to identify and add only reliable annotations to the pool of training images. Our rationale is that only these images should be used to train the final supervised detection network, while images that the joint detection module struggles to annotate with high confidence should not be used for model training, as they may hurt the training process and deteriorate detector performance.

During early stages of the training process, uncertainty regarding both the class of bounding box proposals and the related regression refinement of box coordinates will be high. As training progresses and model predictive quality improves, confidence, accuracy and stability will increase. This results in an increasingly difficult set of images being accurately annotated. We propose to exploit this behaviour by introducing a supervision generator that is able to reliably identify annotated images, creating a set we refer to as *semi-strong* images $\mathcal{P} \subset \mathcal{W}$, that are used to train the fully supervised branch. Intuitively, $\mathcal{P}$ will comprise "easy" images in early stages of training (*e.g.* single instances,

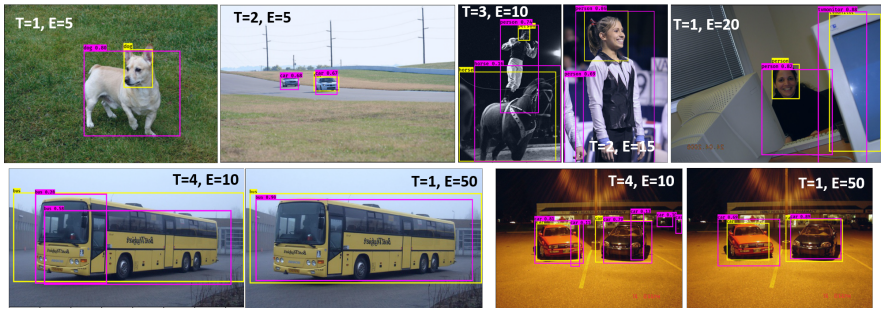

Fig. 4. Examples of semi-strong images. First row: annotated semi-strong images at epoch E, with T iterations required for convergence (see text for details). Second row: examples of semi-strong annotation at pairs of early and late epochs. Magenta color: OAM annotation (class, bounding box score). Yellow: OICR [19] annotation. The results are obtained from a model trained on PASCAL VOC 2007 with 10 shot strong supervision.

uniform colour backgrounds) and sample diversity will progressively increase as the model becomes more accurate (examples of images annotated by our OAM at different training epochs are reported in Fig. 4).

In order to build a set of semi-strong images $\mathcal{P}$, with bounding boxes and associated annotation confidence scores, we propose the following mechanism. Given a weak image I, we obtain a set of N_1 bounding boxes $D_1 = \{c_r, p_r\}_{r=1}^{N_1}$ after Non-Maximum Supression (NMS) is performed on the output of the joint detection module, where c_r and p_r correspond to the class label and coordinates of box r respectively. $D_t = \{c_r, p_r\}_{r=1}^{N_t}$ at every iteration $t > 1$, using D_{t-1} as input candidate proposals. More specifically, the bounding boxes D_{t-1} obtained at the previous iteration are fed again to the RoI Pooling Layer, providing a new set of image features allowing to compute new proposal coordinates. The process iterates until bounding box prediction stabilises and is stopped when $D_t = D_{t-1}$ for three consecutive iterations, i.e. for each bounding box $b_t \in D_t$, there exists a corresponding box $b_{t-1} \in D_{t-1}$ such that b_t and b_{t-1} have intersection-over-union (IoU) ≥ 0.5 and possess matching class predictions (*i.e.* a standard criterion for characterising object equivalence in detection methods). We assign a global confidence weight $1/T$, per image, where T is defined as the first of three iterations in which $D_t = D_{t-1}$. Pseudo-code for the OAM algorithm is found in Supplementary Materials A.

The set of proposals D_1 obtained at iteration 1 constitute the final bounding box annotations. Each box is weighted (box level confidence) by its average IoU with the best matching box at all subsequent iterations. Boxes absent at a given iteration (IoU < 0.5) are, by definition, down weighted due to being assigned an overlap of 0 at that iteration (Fig. 3 shows an example). Images that do not reach convergence by K iterations, or that fail to find any foreground proposals at iteration t, are considered to be annotated with low confidence and are not added

to the semi-strong pool. We set the maximum number of updates $K = 30$, to prevent large sets of iterations and observe that large T (*e.g.* $T > K$) would only occur during early stage training in practice. Finally, the image is only added to the semi-strong pool if the set of obtained annotations contains *all* classes pertaining to the image-level label. We highlight that images requiring large iteration count T for convergence are assigned low confidence scores by design and therefore have limited influence on the training procedure of the second branch. As weak images get annotated by the proposed OAM during training; the semi-strong set expands, while at the same time refining annotations and confidence as the model improves. At a given training step, a weak image that is not successfully annotated, and yet was present in the pool of semi-strong images, will be removed. In this way, the set of semi-strong images has the ability to both expand and contract during training.

3.2 Fully Supervised Branch

Concurrently to OAM training, the obtained strong and semi-strong sets of images are used to train a fully supervised second branch, that comprises both bounding box classification and regression modules on the proposal features ξ_{rf} in a similar fashion to Fast(er) R-CNN [8] style methods. In particular, at every training iteration a batch with the same number of strong and semi-strong images is used. The loss function for this branch is:

$$L_{2B}(p, u, t, v) = L_{cls}(p, u) + L_{reg}(t, v), \tag{4}$$

where p is the ROI class predictions, t is the predicted offset between ROIs and targets, u is the class label and v is the target offset. Only ROIs with foreground labels contribute to the regression loss, L_{reg}. The L_{cls} loss constitutes a weighted cross-entropy for each image:

$$L_{cls}(p, u) = -\frac{1}{T}\sum_{i} \omega_i p_i log(u_i) \tag{5}$$

where the proposals in each batch, contributing to the loss, are indexed by i, the confidence for GT proposal u_i is denoted ω_i and the image-level annotation confidence score is denoted $\frac{1}{T}$. Strong images are assigned image and proposal-level weights of 1. In the early stages of the training process, the semi-strong annotations present some localisation inaccuracies, but are nonetheless highly informative to learn foreground vs background proposals. As training progresses, our OAM improves annotation quality with tighter object coverage and these additional high accuracy annotations will more often contain proposals of exactly full object extent. Such annotations reinforce and strengthen a base signal, provided by strong images alone, towards better bounding-box classification. We also explored utilising semi-strong images to improve bounding-box regression, analogously. In practice, however, this produced slightly worse results. We hypothesise that the discrete problem, associated with the bounded classification loss, affords more robustness to (early-stage) imperfect semi-strong annotations

method	backbone	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP(%)
10% images																						
Fast R-CNN	VGG	47.9	62.9	45.5	34.2	23.0	54.6	70.8	65.5	27.2	61.1	39.8	60.6	70.0	63.3	64.2	14.7	52.9	43.0	55.7	49.5	50.3
BAOD	VGG	51.6	50.7	52.6	41.7	36.0	52.9	63.7	69.7	34.4	65.4	22.1	66.1	63.9	53.5	59.8	24.5	60.2	43.3	59.7	46.0	50.9
BCNet	VGG	64.7	**73.1**	55.2	37.0	39.1	**73.3**	74.0	75.4	**35.9**	69.8	56.3	**74.7**	77.6	71.6	66.9	25.4	61.0	**61.4**	**73.8**	**69.3**	61.8
Ours	VGG	**65.6**	**73.1**	**59.0**	**49.4**	**42.5**	72.5	**78.3**	**76.4**	35.4	**72.3**	**57.6**	73.6	**80.0**	**72.5**	**71.1**	**28.3**	**64.6**	55.3	71.4	66.2	**63.3**
EHSOD	ResNet	60.6	65.2	55.0	35.4	32.8	66.1	71.3	75.3	38.4	54.1	26.5	71.7	65.0	67.8	63.0	27.7	52.6	48.6	70.9	57.3	55.3
BCNet	ResNet	**68.3**	72.0	61.2	48.1	40.8	73.3	73.4	77.8	37.0	69.7	58.3	78.2	80.0	67.5	70.5	**27.4**	62.9	**63.6**	**73.4**	63.6	63.4
Ours	ResNet	62.3	**73.2**	**61.8**	**56.2**	**44.3**	**75.4**	**76.7**	**80.5**	**39.5**	**73.7**	**61.7**	**78.8**	**82.8**	**71.5**	**74.3**	27.0	**67.4**	62.7	71.2	**64.4**	**65.3**
10 shots																						
BCNet	VGG	59.7	69.1	44.6	29.4	40.1	69.2	73.2	**72.9**	32.9	58.1	53.3	66.7	71.3	66.0	61.7	**24.6**	53.0	**62.0**	**67.2**	**67.4**	57.1
Ours	VGG	**60.2**	**71.6**	**51.5**	**45.6**	**43.5**	**71.1**	**75.8**	72.2	**33.8**	**62.9**	**54.0**	**70.0**	**72.9**	**67.5**	**67.4**	23.6	**61.5**	59.1	63.6	66.7	**59.7**
BCNet	ResNet	**63.4**	69.4	54.7	39.5	35.9	70.6	71.8	71.8	33.5	64.6	50.0	65.3	72.7	62.5	61.6	**29.2**	54.5	**63.3**	66.7	**69.4**	58.5
Ours	ResNet	61.7	**72.3**	**56.5**	**52.0**	**37.2**	**71.3**	**74.6**	**77.8**	**36.0**	**67.1**	**58.3**	**78.1**	**77.6**	**68.0**	**71.8**	25.5	**63.6**	62.4	**72.7**	61.2	**62.3**

Table 1. Detailed detection performance (%) on VOC07 dataset. In all the setting, the same BCNet data splits were employed [14].

and therefore compute bounding box regression on only strong images in our final model. To conclude, collecting the introduced components results in the complete loss function for our model: $L_{tot} = L_{1B} + L_{2B}$. At testing, only this fully supervised model is deployed.

4 Results

4.1 Datasets and Implementation Details

We evaluate the performance of our proposed method on two common detection benchmarks: the PASCAL VOC 2007 [6] and the MS-COCO 14 dataset [13], referred to as VOC07 and COCO14. VOC07 has 5011 training and 4952 testing images across 20 categories. COCO14 has $82k$ training and $5k$ testing images across 80 categories. Following evaluation strategies used in the literature, we evaluate detection accuracy on VOC07 using mean Average Precision (mAP), while we employ the COCO metrics, AP_{50} and $AP_{50:95}$, on the COCO dataset. In the reported experiments, reference to 10% of labelled images dictates that 10% of all images have bounding box annotations while the remaining 90% have image-level labels. This corresponds to 500 images in VOC07, $8.2k$ images in COCO14. With reference to our "N-shot" experimental setup, we define each class to have access to N images possessing bounding box annotations. All the experiments on VOC07 use the same data splits provided by BCNet [14], experiments on COCO14 use random selection.

We employ popular network backbones VGG16 and ResNet101 in our experiments to retain consistency with recent approaches. We combine our OAM with Fast R-CNN [8] (using Edge Boxes [27]) and Faster R-CNN using a trainable RPN [17]. Optimisation of all models is performed using SGD with weight decay 0.0001 and momentum 0.9. For experiments concerning the VOC07 dataset, models are trained for 60 epochs. The initial learning rate is 0.001 (first 40 epochs) and reduced to 0.0001 for the final 20 epochs. Analogously for MS COCO experiments; models are trained for 12 epochs, with learning rate 0.001 in the

Method type	Method	10-shots/WSOD		10% images	
		AP (%) person class	mAP (%)	AP (%) person class	mAP (%)
fully supervised	Fast R-CNN	58.0	42.1	64.2	50.3
fully supervised	Faster R-CNN	54.3	37.7	55.7	46.7
WSOD	PCL	17.8	43.5	-	-
WSOD	PCL + Fast R-CNN	15.8	44.2	-	-
WSOD	WSOD^2	21.9	53.6	-	-
MSOD	BAOD	-	-	59.8	50.9
MSOD	EHSOD (ResNet + FPN)	-	-	63.0	55.3
MSOD	BCNet	61.7	57.1	66.9	61.8
MSOD	Ours	**67.4**	**59.7**	**71.1**	**63.3**
MSOD	Ours + RPN	64.3	54.6	68.9	60.5
fully supervised	Fast-RCNN 100 % images (Ours upper bound)	76.8 (person), 71.6			
fully supervised	Faster-RCNN 100 % images (Ours + RPN upper bound)	75.6 (person), 67.0			

Table 2. Comparison to SOTA on VOC07 dataset. A VGG backbone is used unless specified. Gray rows correspond to methods learning an RPN (vs methods using precomputed proposals).

first 9 epochs and then reduced to 0.0001 for the final 3 epochs. Remaining model hyper-parameters follow the values reported in [14]. For data augmentation, we apply the same augmentation strategy as BCNet [14] for fair comparison, *i.e.* we bilinearly resize images to induce a minimum side length $\in \{400, 600, 750\}$ and, for fully supervised training, uniformly crop image regions with a fixed 600×600 window. All experiments are implemented in PyTorch using a single GeForce GTX 1080 GPU.

4.2 Comparisons with State-of-the-art

Baselines: We evaluate our model with respect to two SOTA WSOD methods, PCL [18] and WSOD^2 [25], that were evaluated on both VOC07 and COCO14. We further compare to three MSOD approaches: the two level approach of BCNet [14], end-to-end methods BAOD [15] and EHSOD [7]. To the best of our knowledge, these are the only three methods adopting mixed supervision. All three methods were evaluated on VOC07. Results for BCNet, the best performing baseline on VOC07, were not available for the COCO dataset. The approach requires training two models (OICR and BCNet) with two separate sets of parameters that need to be adapted to the new dataset, making it highly challenging and time consuming to provide a fair comparison, hence we were not able to provide it. Similarly, EHSOD was evaluated only on the COCO 2017 database with a much larger set of annotated training images (approx. $12k$), making results not directly comparable to our experiments and different from the low-shot setting studied in this work. Finally, we compare our results with respect to Fast R-CNN and Faster R-CNN trained with full supervision (our upper bounds) and low-shot supervision (*i.e.* 10% and 10-shot training data), using the same augmentation strategy as all previous models.

PASCAL VOC 2007: We report detailed per-class results, compared to competing MSOD approaches in Tab. 1 using 10% annotated training images, and 10 shots. We consistently outperform all competing methods in terms of mAP,

Method type	Method	AP@.50	AP@[.50,.95]
fully supervised	Fast R-CNN - 10 shots	22.1	10.0
fully supervised	Faster R-CNN - 10 shots	16.1	6.7
WSOD	PCL	19.4	8.5
WSOD	PCL+ Fast R-CNN	19.6	9.2
WSOD	WSODˆ2	22.7	10.8
MSOD	Ours - 10 shots	**31.2**	**14.9**
MSOD	Ours + RPN - 10 shots	**24.9**	**10.2**
fully supervised	Fast R-CNN - 100% data	49.9	29.0
fully supervised	Faster R-CNN - 100% data	42.1	20.5

Table 3. Comparison with the SOTA on MS-COCO14 with 10-shot training examples (VGG backbone). Gray rows correspond to methods learning an RPN (vs methods using precomputed proposals).

with an improvement of up to 4% with respect to BCNet in the 10 shot scenario (ResNet), and 10% with respect to EHSOD in the 10% images scenario. We further highlight that BCNet constitutes a two-level WSOD dependent method. The influence of the chosen WSOD component is clearly visible; object classes where their method excels, and surpasses our per-class performance, are the same classes for which their adopted WSOD component (OICR) provides best initial bounding box estimations [19]. In Tab. 2, we provide more comparisons in the 10 shots and 10% images scenarios using precomputed proposals (white rows) and an RPN [17] (grey rows). We highlight that we use an off-the-shelf RPN without parameter optimisation, and expect performance to be worse, and not directly comparable to strategies relying on pre-computed proposals. We further compare with top performing WSOD methods and Fast(er)-RCNN approaches and highlight our performance on the "person" class, often reported as one of the most challenging classes for WSOD methods due to the large intra-class variability in terms of appearance [14, 25]. We significantly outperform all SOTA methods, and substantially improve with respect to WSOD methods, in particular for the person class, with only minor additional labelling cost. Comparing to Fast(er)-RCNN methods, we highlight that our OAM improves upon models trained on 10% data and 10 shots by a large margin (13% and 17% respectively), reaching performance close to the fully supervised upper bound.

MS-COCO14: We provide further comparison to additional benchmark datasets in order to highlight model generalisability. We note that contemporary WSOD methods mainly focus on detection datasets of modest size such as VOC07. COCO14 is significantly larger, and constitutes a more challenging dataset due to both the increased size and variability expressed in image content. Tab. 3 reports comparisons between our method (precomputed and RPN proposals) and WSOD approaches PCL and WSOD^2 on COCO14 using 10 shots labelled images. As we compare solely to WSOD methods, we limit our experiments to the 10 shots setting, as 10% annotated examples provide a very significant advantage compared to WSOD methods. We additionally provide comparison to Fast(er) R-CNN methods trained on 10 shots as well as their fully supervised equivalent on 100% images. We highlight that our method maintains robust performance and significantly outperforms the WSOD methods and 10 shots Fast(er)-RCNN

10 shots					AP (%)																		
SE	BBA	OAM	1B	2B	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP(%)
	✓		✓		42.0	57.1	40.2	34.2	30.3	62.6	69.0	62.5	23.2	63.8	33.0	58.5	72.2	63.3	62.9	20.8	54.9	44.2	54.3	55.2	50.2
				✓	30.9	53.2	35.8	27.8	19.9	51.6	65.8	54.7	19.3	48.3	27.8	46.3	57.7	54.3	58.0	14.9	49.1	37.5	43.8	44.7	42.1
✓	✓		✓		44.3	60.2	40.4	37.8	28.1	67.0	72.8	64.1	24.2	64.6	40.9	60.5	70.5	61.6	63.5	16.1	55.0	46.2	57.5	58.0	51.7
✓	✓			✓	47.3	62.1	42.4	35.2	28.2	67.0	72.8	65.1	21.7	65.3	43.4	61.4	70.6	63.5	63.0	16.5	57.6	45.8	58.7	54.7	52.1
✓		✓	✓		50.3	67.3	49.8	44.1	35.9	64.3	72.7	70.3	32.6	57.7	44.5	66.3	65.6	68.3	62.8	25.2	60.0	48.8	62.6	64.5	55.7
✓		✓		✓	61.4	71.0	48.5	42.9	37.8	69.8	75.6	72.8	34.0	63.2	47.6	71.9	71.1	71.1	64.6	25.7	63.4	55.6	61.9	65.8	58.8
✓	✓	✓	✓		57.9	71.4	48.2	42.7	38.0	71.4	75.5	75.5	34.0	67.1	54.0	71.4	74.3	69.4	65.7	23.7	61.6	56.1	61.0	65.0	59.2
✓	✓	✓		✓	60.2	71.6	51.5	45.6	43.5	71.1	75.8	72.2	33.8	62.9	54.0	70.0	72.9	67.5	67.4	23.6	61.5	59.1	63.6	66.7	59.7

Table 4. Ablative analysis of our method on VOC07 for the 10 shot scenario. SE: shared encoder, OAM: second branch training also on OAM generated semi-strong images, BBA: bounding box augmentation strategy. 1B: first branch output, 2B: second branch output.

models (9%). This provides evidence in support of our claim that the strategy of providing mixed supervision significantly improves generalisation ability in settings that entail more difficult tasks with higher variability.

4.3 Ablation Studies

We conduct experiments to understand the different contributions and assignment of credit for our OAM components using the VOC07 dataset and a VGG backbone. Tab. 4 shows ablative results for the 10 shots scenario while additional results for the 10% images scenario are reported in supplementary materials. Studied components are: *SE*: shared encoder (*i.e.* no SE entails independent branch training); *OAM:* fully supervised branch is also trained on semi-strong images generated by the OAM; *BBA:* online bounding box augmentation strategy. For each configuration, we report mAP with respect to the output of the OAM (first branch; 1B) as well as the output of the fully supervised branch (second branch; 2B). We experimentally verify the importance of each component; performance consistently improves as new components are integrated. We note that the shared encoder strongly improves the fully supervised branch, while the OAM, and communication between branches, affords mutual branch improvement. Both performance gains can be attributed to the more discriminative full *vs*. partial object proposal features learned by the shared encoder.

5 Conclusion

We have introduced a novel online annotation module (OAM), trained using mixed supervision, that learns to generate annotations on the fly and thus affords concurrent training for fully supervised object detection. The OAM can be combined with any two-stage object detector and provides an intrinsic curriculum to improve the training procedure. Extensive experiments on two popular benchmarks show SOTA performance in the mixed supervision scenario, and significant improvement of two-stage detection methods in low-shot settings. Moreover, our method has the potential to increase performance on rare, long tail classes that typically only possess a handful of annotated examples.

References

1. Arun, A., Jawahar, C., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9432–9441 (2019)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2846–2854 (2016)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: A low-shot transfer detector for object detection. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. Dong, X., Zheng, L., Ma, F., Yang, Y., Meng, D.: Few-example object detection with model communication. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1641–1654 (2018)
6. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
7. Fang, L., Xu, H., Liu, Z., Parisot, S., Li, Z.: EHSOD: CAM-Guided End-to-End Hybrid-Supervised Object Detection with cascade refinement. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence. pp. xxx–yyy. AAAI Press (2020)
8. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
9. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1377–1385 (2017)
10. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8420–8429 (2019)
11. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2019)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Pan, T., Wang, B., Ding, G., Han, J., Yong, J.: Low shot box correction for weakly supervised object detection. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 890–896. AAAI Press (2019)
15. Pardo, A., Xu, M., Thabet, A., Arbelaez, P., Ghanem, B.: Baod: Budget-aware object detection. arXiv preprint arXiv:1904.05443 (2019)
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)

17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
18. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.L.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence (2018)
19. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2843–2851 (2017)
20. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
21. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2199–2208 (2019)
22. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-entropy latent model for weakly supervised object detection. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
23. Wei, Y., Shen, Z., Cheng, B., Shi, H., Xiong, J., Feng, J., Huang, T.: Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 434–450 (2018)
24. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9577–9586 (2019)
25. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8292–8300 (2019)
26. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
27. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014)