Supplementary Materials

Paper ID 438

We provide additional material to supplement our work.

In **Appendix A**, we report a pseudocode description of the proposed online pseudo-supervision generation algorithm in the Online Annotation Module (OAM).

In Sec. 4.3 of the main paper, we presented an ablation study to confirm the influence of each component of our method. This was carried out using the 10 shot scenario, with the PASCAL VOC 07 dataset (VOC07) and a VGG16 backbone [1]. Here in **Appendix B** we present an extended analysis using, alternatively, 10% of VOC07 training images for strong supervision. Additionally, to further explore method sensitivity, **Appendix C** investigates variance caused by the selection process of the fully annotated image set; we report a five-fold experiment, under the 10 shot scenario again employing VOC07 with a VGG16 backbone.

The EHSOD paper [2] reports detection results for the MS-COCO 17 (COCO17) dataset corresponding to the 10% training data scenario. In that setting, $\sim$12000 fully annotated images are available to the model, which strays from the *low-shot* scenario studied in our work. Nonetheless, for completeness, we report comparison between our method (considering both pre-computed and RPN proposal instances) and EHSOD [2], and provide additional qualifying discussion in **Appendix D**.

In **Appendix E** we report additional detailed per-class detection results for both 20% and 20 shot annotation scenarios on VOC07, with comparisons to alternative Mixed Supervision Object Detection (MSOD) approaches. Per-class detection results aim to further reader understanding and offer deeper insight into competing methods' performance and individual per-class traits.

In **Appendix F** additional visual results are provided; **Appendix F.1** and **Appendix F.2** show (1) examples of images annotated by our OAM and (2) test time detection performance (for VOC07, COCO14) respectively. Finally, in **Appendix G**, we highlight some common failure cases of our method.

A Online Pseudo-Supervision Generation algorithm

Algorithm 1 Online Pseudo-Supervision Generation algorithm

Input: Initial set of N detections $D_0 = \{c_r, p_r\}_{r=1}^{N}$, stopping criterion K, image feature vector $f(\mathbf{x})$, OAM layers parameters θ.
Output: M output detections $D_1 = \{c_r, p_r, w_r\}_{r=1}^{M}$ with confidence weights w_r, number of iterations required for convergence T.
Initialise variables: $D \leftarrow D_0$, $counter \leftarrow 0$
For t = 1 **to** K :
    $\{\boldsymbol{\xi_r}\}_{r=1}^{N} \leftarrow$ RoIPooling$(D, f(\mathbf{x}))$
    $D_t = forward_\theta(\{\boldsymbol{\xi_r}\}_{r=1}^{N})$
    if D_t is empty : ▷ No detections
        break
    if $D_t == D$ **:** ▷ $\forall\ b_t \in D_t$, $\exists\ b \in D$ where $IOU(b_t, b) \leq 0.5$ and $class(b_t) = class(b)$.
        $counter++$
        if $counter == 3$ **:**
            $T \leftarrow t + 1 - counter$ ▷ First of three iterations where $D_t == D$
            $w_r \leftarrow$ averageOverlap$(\{D_t\}_{t=1}^{T})$
            break
    else:
        $counter \leftarrow 0$
    $D \leftarrow D_t$

B Ablation study: 10% data scenario

In Tab. 1, we report ablation study results for the proposed model (VGG16 backbone) where 10% of images from VOC07 provide strong supervision. Results for the analogous 10 shot scenario were reported in the main paper, Sec. 4.3. Considered ablation components are *SE*: presence of shared encoder (*i.e.* no *SE* entails independent branch training); *OAM:* the fully supervised branch is additionally 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*).

As was also observed for the 10 shot scenario (reported in Sec. 4.3 of the main paper), the performance increases as additional components are added, providing further evidence for component validity and contribution. The performance gaps between differing ablations are smaller than our analogous main paper experiment due to the increased strong supervision available in the current case. Congruent with the results reported in Sec. 4.3, this ablation highlights that the shared encoder strongly improves the fully supervised branch, while the OAM and communication between branches, afford mutual branch improvement.

10 %					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(%)
	✓		✓		56.7	69.9	52.5	42.7	36.7	72.9	76.4	70.6	31.8	72.6	48.2	66.9	77.7	68.9	67.1	22.9	59.9	55.5	62.8	63.2	58.8
				✓	47.9	62.9	45.5	34.2	23.	54.6	70.8	65.5	27.2	61.1	39.8	60.6	70.	63.3	64.2	14.7	52.9	43.	55.7	49.5	50.3
✓	✓		✓		57.3	67.4	51.4	42.	37.2	72.2	77.2	72.5	31.7	69.5	52.8	71.1	76.5	67.8	67.4	21.8	57.7	54.6	64.5	62.3	58.7
✓	✓			✓	57.5	68.2	53.	41.8	37.4	70.1	77.2	73.2	33.	69.3	54.8	71.8	78.4	69.	67.7	22.2	59.4	54.3	66.1	62.3	59.3
✓		✓	✓		64.3	69.7	56.1	48.3	39.8	71.4	78.1	76.5	37.8	71.1	56.4	76.5	76.5	70.9	68.4	25.7	62.1	55.7	70.2	65.8	62.1
✓		✓		✓	67.1	70.3	56.2	48.4	42.1	71.7	76.9	76.7	39.2	71.5	60.1	74.1	79.6	71.3	70.9	26.3	61.6	56.4	71.1	66.1	62.9
✓	✓	✓	✓		66.4	71.8	57.3	50.3	41.5	72.6	78.5	77.3	38.4	71.6	59.8	74.3	79.4	71.5	71.4	26.1	61.8	57.6	72.3	66.5	63.3
✓	✓	✓		✓	65.6	73.1	59.	49.4	42.5	72.5	78.3	76.4	35.4	72.3	57.6	73.6	80.	72.5	71.1	28.3	64.6	55.3	71.4	66.2	63.3

Table 1: Ablative analysis of our method using VOC07 in the 10% scenario. *SE*: shared encoder, *OAM*: second branch trained also using OAM generated semi-strong images, *BBA*: bounding box augmentation strategy. *1B*: first branch output, *2B*: second branch output.

C Sensitivity to the selected annotation

In order to test the sensitivity of our method, with respect to annotated image-subset selection variance, we perform a five-fold experiment, under the 10 shot scenario. We test using VOC07 and a standard VGG16 backbone architecture. This scenario represents the setting most susceptible and sensitive to image subset selection as the pool of strong images is the smallest among all considered scenarios (including MS-COCO experiments). It can be observed in Table 2 that image selection variance is small. Varying the selected image subset has only minor effect on final mAP, providing evidence towards the robustness of our proposed approach. This variance intuitively reduces further in cases where the model is trained using a larger number of fully annotated images.

SPLIT	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP(%)
1	64.1	73.7	53.0	49.2	46.8	73.4	75.1	70.5	33.1	73.4	46.9	75.1	72.4	69.8	63.8	31.0	62.6	52.2	69.2	62.5	60.9
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
3	62.5	73.9	60.1	42.0	40.0	74.1	74.7	75.2	33.7	74.5	51.4	71.4	79.9	71.9	64.6	30.3	63.6	55.8	64.8	66.8	61.6
4	62.2	75.1	56.1	42.7	38.9	73.4	75.3	75.0	32.1	68.1	46.3	69.6	75.3	71.1	62.5	26.4	59.3	54.3	69.4	63.4	59.8
5	64.0	73.5	60.1	50.6	38.9	72.6	75.6	70.3	32.7	70.1	55.4	73.9	75.1	70.2	64.3	25.6	62.6	49.2	67.9	65.3	60.8
mean	62.6	73.6	56.2	46.0	41.6	72.9	75.3	72.6	33.1	69.8	50.8	72.0	75.1	70.1	64.5	27.4	61.9	54.1	67.0	64.9	60.6
std	1.6	1.3	3.9	3.8	3.4	1.1	0.4	2.4	0.7	4.6	4.1	2.4	2.3	1.7	1.8	3.1	1.6	3.7	2.6	1.9	0.8

Table 2: Five-fold experiment for the 10 shot scenario using VOC07 and a standard VGG16 backbone [1]. Fold mean and standard deviation statistics are reported in the final rows. The second split is the split used in [3], and the split used for all our remaining experiments.

D MS-COCO 2017 comparisons

The EHSOD [2] method reported results using the COCO17 dataset, corresponding to a 10% training data scenario. We thus report here comparison between our method (considering both pre-computed and RPN [4] proposal setups) and the EHSOD mixed supervision approach. We also provide additional comparison to both Fast and Faster-RCNN methods, trained using the same 10% of COCO17 images, as well as their fully supervised equivalent; using 100% of the training images. Results are found in Tab. 3. We note this setting corresponds to approximately ∼12000 fully annotated images, a much larger set than the ones used in all other experiments.

It can be observed that, in this setting, our model performs on-par with EHSOD when using RPN proposals, while significantly outperforms their approach when pre-computed (Edge Boxes) proposals are employed. Moreover, we observe that our method also performs on-par with the Fast(er)-RCNN baselines in the 10% images scenario. Interestingly we note only a reasonably modest gap between Fast(er)-RCNN performance with regard to the considered 10% and 100% baselines. This suggests that the gap between the 10% and 100% setting can be closed by providing the network with images containing object class appearance outliers or by images containing difficult, crowded scenes. As a consequence, the problem, in this setting, can be considered to have a greater affinity with a fully supervised task than with a low-shot setting. This observation provides some explanation towards why our method provides limited improvement in this setting. Images required to improve the detector performance (high information content) may not be annotated with high confidence and therefore not considered for object detector training. As highlighted in our future work discussion (main paper; Sec. 5), we believe active learning strategies may prove fruitful in such cases.

Method type	Method	AP@.50	AP@[.50,.95]
fully supervised	Fast RCNN - 10% data	53.7	31.6
fully supervised	Faster RCNN - 10% data	46.3	25.6
MSOD	EHSOD - 10% data	46.8	-
MSOD	Ours - 10% data	54.2	31.6
MSOD	Ours + RPN - 10% data	46.0	25.4
fully supervised	Fast RCNN - 100% data	61.6	48.0
fully supervised	Faster RCNN - 100% data	51.1	28.8

Table 3: Comparison with state of the art on COCO17. All the models were trained with a ResNet101 backbone [5], while EHSOD uses FPN [6]. Gray rows correspond to methods learning an RPN [4] (*vs.* methods using precomputed proposals).

E Additional PASCAL VOC 07 results

We report here detailed *per-class* detection results and compare competing MSOD approaches using both 16% annotated training images and 20 shot scenarios. Results are found in Tab. 4. We consistently outperform all competing methods in terms of mAP, with an improvement of up to 4% with respect to BCNet in the 20 shot scenario (ResNet101 [5] backbone). We highlight that in the 16% training image scenario, we report both EHSOD and BAOD results, trained using 20% of training images as only these results were available. This highlights the ability of our method to outperform these competing models even in the case where we have access to 200 fewer training examples.

method	backbone	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP(%)
20 shot																						
BCNet	ResNet101	**66.5**	67.6	56.7	40.5	40.4	72.8	71.3	76.6	**39.4**	65.0	54.1	71.4	72.9	66.6	66.0	26.1	59.0	65.5	**67.7**	**67.6**	60.7
Ours	ResNet101	66.2	**73.3**	**57.0**	**53.2**	**42.8**	**76.0**	**76.0**	**79.1**	38.6	**74.6**	**61.1**	**79.9**	**77.4**	**70.2**	**73.1**	**26.7**	**64.3**	**65.7**	67.6	64.5	**64.4**
16% images																						
BCNet	VGG16	63.7	**77.2**	**62.9**	48.0	39.7	73.3	76.0	**78.0**	39.4	72.9	56.1	75.4	79.9	69.5	70.2	31.0	60.6	62.2	**75.0**	68.6	64.0
Ours	VGG16	**66.5**	76.2	59.1	**53.0**	**49.2**	**77.1**	**79.4**	76.9	**41.4**	**75.4**	**63.7**	**80.2**	**80.9**	**71.6**	**73.0**	**35.7**	**67.5**	**64.0**	73.5	**68.9**	**66.7**
BAOD*	ResNet101	57.0	62.2	60.0	46.6	46.7	60.0	70.8	74.4	40.5	71.9	30.2	72.7	73.8	64.7	69.8	**37.2**	62.9	48.4	64.1	59.1	58.6
BCNet	ResNet101	**67.3**	74.2	65.2	51.7	40.8	**74.1**	72.7	77.2	39.2	70.3	59.9	77.2	78.5	69.9	68.6	30.6	60.0	**68.2**	**75.9**	**66.8**	64.4
EHSOD*	ResNet101	65.5	72.3	**66.7**	45.6	**50.8**	72.2	77.8	82.2	**44.3**	73.1	44.8	79.3	76.0	**73.0**	73.8	35.5	63.0	62.1	74.0	65.5	64.9
Ours	ResNet101	65.8	**78.8**	63.7	**55.3**	49.7	73.0	**79.6**	**84.5**	42.7	**75.0**	**61.6**	**84.7**	**83.3**	71.8	**75.1**	33.9	**64.6**	64.9	73.3	66.2	**67.4**

Table 4: Detailed *per-class* detection performance (%) on VOC07. For each instance of our model, identical data splits, from the BCNet paper [3] were consistently used. Method rows marked * indicate models trained using 20% of images, due to the availability of comparable results, *c.f.* only 16%.

F Additional visual results

F.1 Annotated Semi-Strong Images

In Fig. 1 we provide additional examples of images annotated by our OAM, named semi-strong images, during progressive training epochs E. These online annotations are obtained by our model using VOC07 data with 10 shot strong supervision (other examples of semi-strong images are reported in the main manuscript, Fig. 4). We observe that typically uncomplicated and simple images are labelled with high confidence when training begins (for example at epoch rows $E = \{5, 10\}$). During later training stages (here $E > 10$), more complex images with increased appearance diversity and also with multiple, overlapping object instances are added to the pool by our OAM. In general, T ranged from 1-10 (first 5 epochs) to 1-3 (end of training); and the semi-strong set contained approx. 10% (first epochs) to 45-60% (end of training) of annotated weak images

Furthermore, we compare the annotations obtained by our method (magenta) with annotations generated by a popular Weakly Supervised Object Detection (WSOD) approach; OICR [7] (yellow detections). We highlight that, from early epochs, our method is providing better, more reliable annotations that are then employed for concurrent object detector training. Moreover, our annotations cover the full extent of the object of interest. This can be explained due to the high quality information being distilled from the low-shot fully annotated images (strong images), while the WSOD method annotations exhibit the well understood problem of tending to focus on object parts and on (only) the most discriminative object in the image.

Figure 1: Examples of semi-strong images at epoch E with iterations required for OAM convergence T (definition in the main paper, Sec. 3). Magenta: our OAM annotation (class, bounding box score). Yellow: OICR [7] (WSOD) annotations. Results are obtained using model trained on VOC07 with 10 shot strong supervision.

F.2 Examples of Detections

Further exemplar test-time detections, obtained by our method with 10 shot strong supervision, are shown in Fig. 2 and Fig. 3 for VOC07 and COCO14 test images respectively. Due to the low-shot set of fully annotated images, that are leveraged by our model, we observe that obtained detections cover full object extent, even for classes typically difficult for WSOD (*e.g. person*). In comparison with WSOD approaches, our method avoids enclosing only the most discriminative object parts. Moreover, multiple instances of the same class within a single image can now be captured. This is usually problematic when training a model by relying only on image-level supervision, as in WSOD.

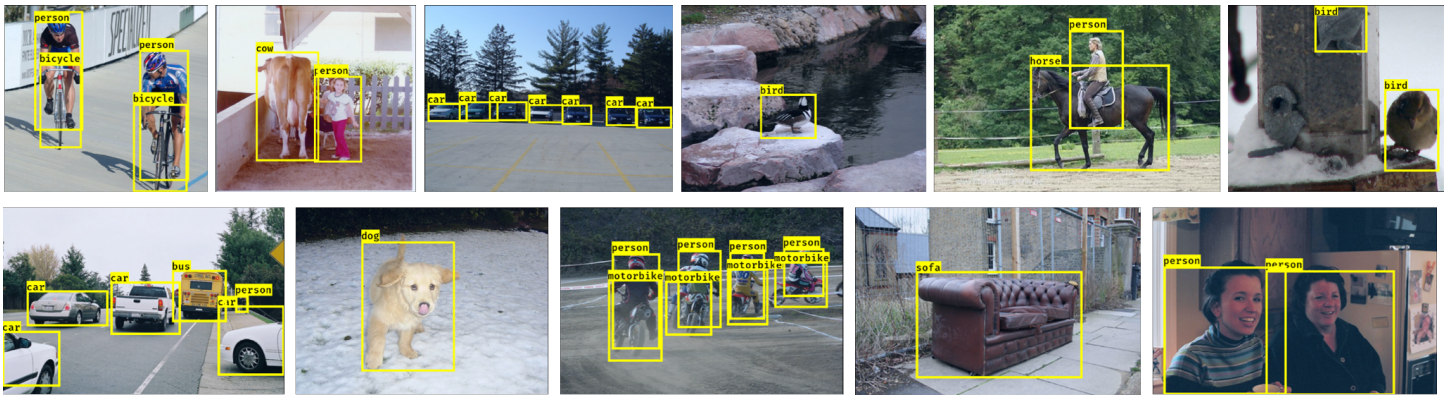

Figure 2: Detection results on VOC07 test. Results are obtained from a model trained on VOC07 with 10 shot strong supervision, VGG16 backbone.

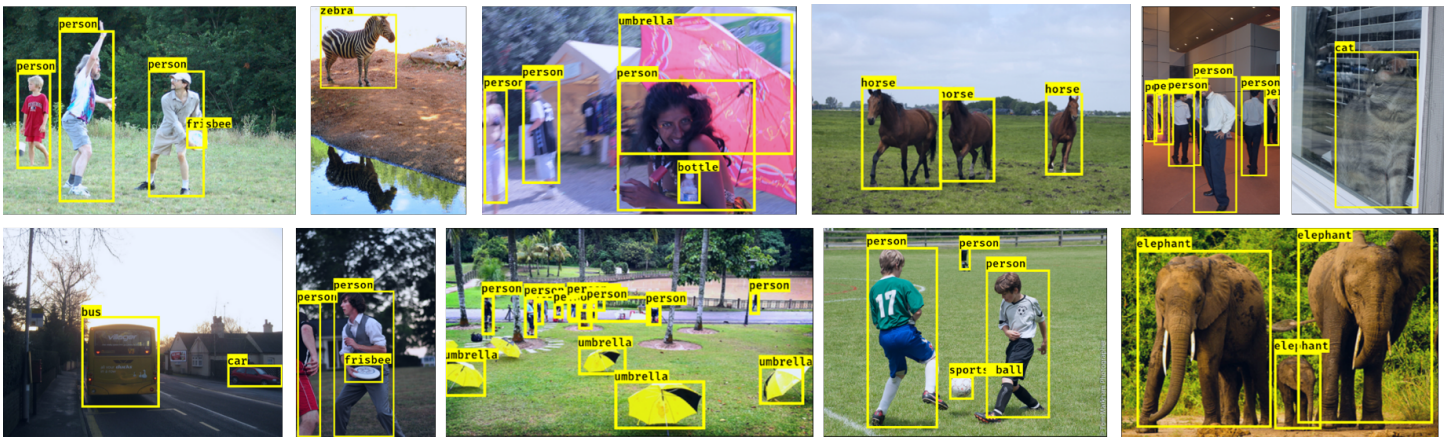

Figure 3: Detection results on COCO14 test. Results are obtained from a model trained on COCO14 with 10 shot strong supervision, VGG16 backbone.

G Common Modes of Failure

We conducted additional investigation to identify instances of detection failures for our model trained with 10 shot supervision. For both datasets (VOC07, COCO14) considered in our work, the most common mode of failure is represented by multiple detection for an object of interest. Given that the model is only trained with 10 shot, we partially attribute such failures to the (weakly-learned) bounding box regressor. In corroboration with competing work [3, 2] we note bounding box regression is an intrinsically difficult task, especially in cases when limited training data is available or where substantial background pixels need be included to provide an optimal object bounding box, such as for objects with elongated or articulated shapes. As discussed in the main paper (Sec. 5), additional future work may explore strengthening of regression task performance.

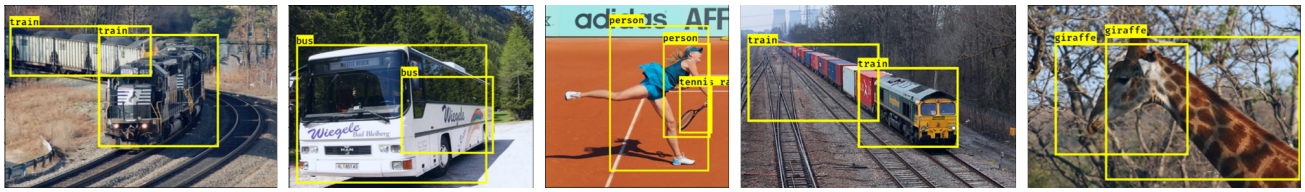

Figure 4: Example detection failures obtained from our proposed model. Images are obtained from a model trained on VOC07 (left-most two images) and on COCO14 (right-most three images) with 10 shot supervision.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] L. Fang, H. Xu, Z. Liu, S. Parisot, and Z. Li, "EHSOD: CAM-Guided End-to-End Hybrid-Supervised Object Detection with cascade refinement," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, AAAI Press, 2020.

[3] T. Pan, B. Wang, G. Ding, J. Han, and J. Yong, "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.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[7] P. Tang, X. Wang, X. Bai, and W. Liu, "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.