Active Learning Strategies for Weakly-supervised Object Detection

—Supplemental Material—

Huy V. Vo[1,2] Oriane Siméoni[2] Spyros Gidaris[2] Andrei Bursuc[2]
Patrick Pérez[2] Jean Ponce[1,3]

[1]Inria and DI/ENS (ENS-PSL, CNRS, Inria)
[2]Valeo.ai [3]Center for Data Science, New York University

{van-huy.vo, jean.ponce}@inria.fr
{oriane.simeoni,spyros.gidaris,andrei.bursuc,patrick.perez}@valeo.com

Table of Contents

1 Additional Qualitative Results

We provide in this section visualizations to get insights into the benefits of our method BiB.

1.1 Prediction Examples

We show in Figure 5 predictions obtained with the weakly-supervised detector MIST (top row) and the detector after the first cycle of BiB (bottom row) with $B = 50$ on VOC07 and $B = 160$ on COCO. We observe that the failures modes of MIST are corrected by our BiB detector: objects and parts are not confused (3^{rd} and 4^{th} images), objects are covered (1^{st}) and better separated (2^{nd}).

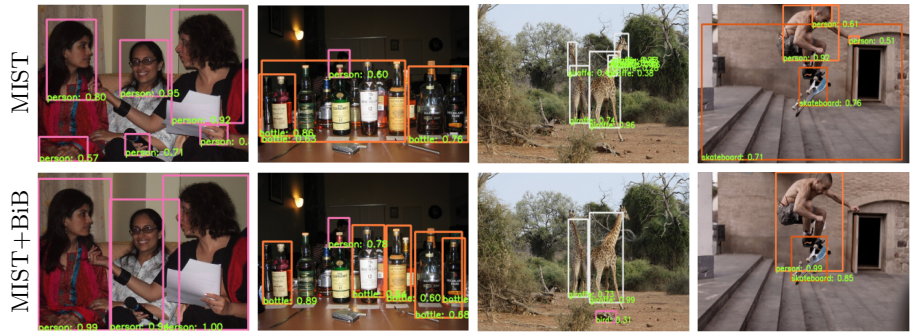

Fig. 5: Examples of predictions on the VOC07 and COCO test sets, by MIST [4] (first row) and BiB after the first cycle (second row). Fine-tuning MIST with images selected by BiB significantly remedies its limitations.

Table 4: Comparison of active learning strategies on VOC07. For each experiment, we conducted 5 cycles with a budget of 50 images per cycle. We repeated the experiment six times for each strategy and report the average and standard deviation of their performance. BiB yields significantly better performance than the others. *loss* performs well in the first cycle but fares worse than BiB in subsequent cycles. Additionally, it performs much worse, even than random, on COCO (see Table 5).

Method	Number of fully-annotated images				
	50	100	150	200	250
u-random	56.5 ± 0.4	58.4 ± 0.4	59.3 ± 0.7	60.2 ± 0.4	61.1 ± 0.5
b-random	56.7 ± 0.7	58.4 ± 0.7	59.7 ± 0.8	60.4 ± 0.5	61.2 ± 0.4
core-set	55.5 ± 0.6	57.7 ± 0.6	58.7 ± 0.5	59.5 ± 0.4	60.1 ± 0.2
core-set-ent	55.5 ± 0.4	57.6 ± 0.4	59.0 ± 0.4	60.0 ± 0.2	60.5 ± 0.2
entropy-max	57.0 ± 0.4	58.7 ± 0.2	59.6 ± 0.4	60.6 ± 0.2	60.9 ± 0.2
entropy-sum	56.5 ± 1.0	58.6 ± 0.4	59.8 ± 0.3	60.5 ± 0.5	61.2 ± 0.8
loss	**59.7 ± 0.2**	60.5 ± 0.5	61.3 ± 0.7	62.0 ± 0.5	62.5 ± 0.3
BiB	58.5 ± 0.8	**60.8 ± 0.5**	**61.9 ± 0.4**	**62.9 ± 0.5**	**63.5 ± 0.4**

1.2 More Visualization of BiB Pairs

Our selection method relies on the discovery of *box-in-box* patterns. We provide in Figure 6 more visualization of BiB pairs on images of VOC07 and COCO.

2 Additional Quantitative Results

2.1 Detailed Results of Active Learning Strategies

For experiments with active learning strategies, we have run each strategy six times on VOC07 and three times on COCO and reported the average performance in the main paper. For completeness, we provide in Table 4 and Table 5 both the average and the standard deviation of the detector's performance in these experiments.

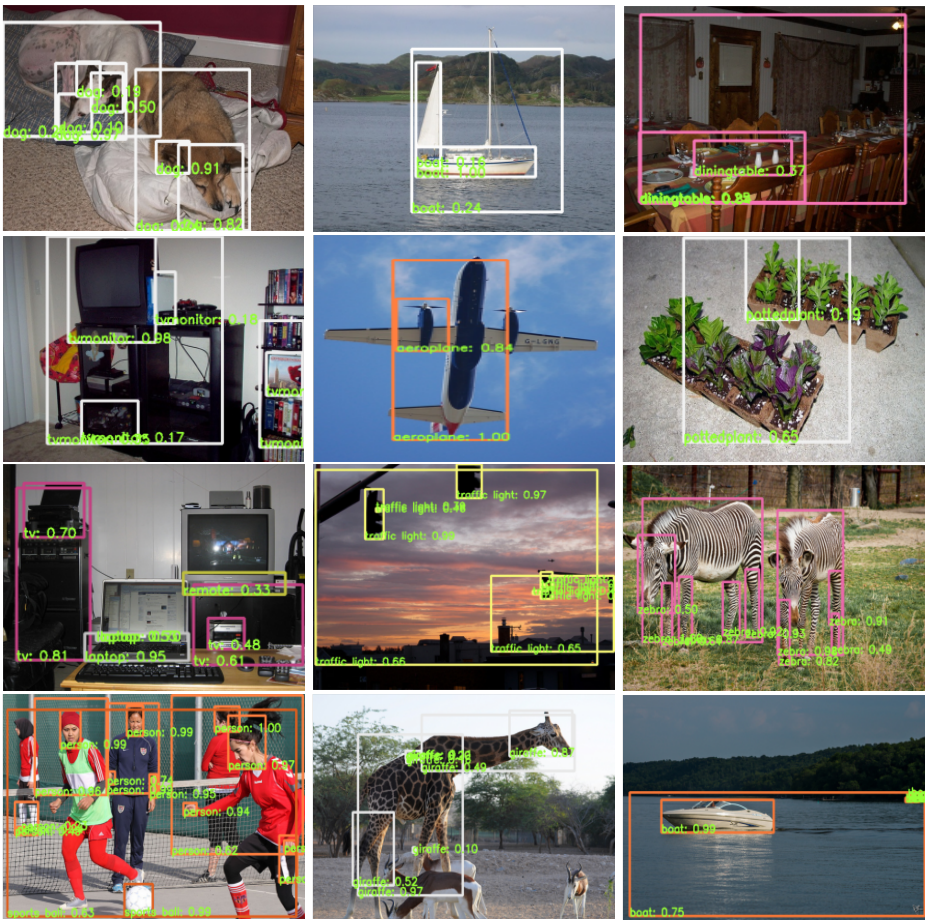

Fig. 6: Examples of *box-in-box* (BiB) pairs on VOC07 (first two rows) and COCO (last two rows) extracted using the MIST [4] detector.

2.2 Different Variants of *loss*

MIST [4] is trained with a combination of losses coming from different heads. The Multiple Instance Learner produces $\mathcal{L}^{\text{MIL}}$ using the ground-truth class information while each refinement head $k \in \{1, 2, 3\}$ produces the refinement loss $\mathcal{L}_w^{(k)}$ using pseudo objects generated from the previous head. We have tested each of these losses and the combination of the three refinement losses $\sum_{k=1}^{3} \mathcal{L}_w^{(k)}$ in our experiments with *loss* strategy. We present a summary of the results in Table 6. For each experiment, we have conducted 5 cycles with a budget of 50 images per cycle on VOC07. On average, $\mathcal{L}_w^{(3)}$ yields the best results on this dataset and we use it for all experiments with the loss strategy in our submission.

Table 5: Comparison of active learning strategies on COCO. For each experiment, we conducted 5 cycles with a budget of 160 images per cycle. We repeated the experiment three times for each strategy and report the average and standard deviation of their performance. BiB significantly outperforms all other methods.

Method	AP					AP50				
	160	320	480	640	800	160	320	480	640	800
u-random	14.1 ± 0.1	15.1 ± 0.2	15.7 ± 0.2	16.1 ± 0.4	16.5 ± 0.3	29.1 ± 0.4	30.8 ± 0.3	31.7 ± 0.4	32.4 ± 0.4	33.0 ± 0.3
b-random	14.4 ± 0.4	15.2 ± 0.3	15.9 ± 0.1	16.2 ± 0.2	16.8 ± 0.2	29.5 ± 0.6	30.8 ± 0.4	31.8 ± 0.2	32.3 ± 0.1	33.3 ± 0.2
entropy-sum	12.3 ± 0.3	12.8 ± 0.2	13.3 ± 0.3	13.6 ± 0.4	13.7 ± 0.3	25.6 ± 0.4	26.5 ± 0.1	27.2 ± 0.2	27.7 ± 0.5	27.8 ± 0.1
entropy-max	12.7 ± 0.2	13.9 ± 0.1	14.5 ± 0.5	14.9 ± 0.3	15.2 ± 0.2	26.9 ± 0.2	28.9 ± 0.1	29.7 ± 0.5	30.4 ± 0.3	30.8 ± 0.3
loss	13.5 ± 0.1	14.1 ± 0.2	14.5 ± 0.2	14.7 ± 0.3	14.9 ± 0.3	27.8 ± 0.1	29.1 ± 0.1	29.7 ± 0.1	30.1 ± 0.3	30.4 ± 0.3
core-set	12.9 ± 0.2	14.5 ± 0.3	15.3 ± 0.2	15.9 ± 0.1	16.4 ± 0.3	26.9 ± 0.3	29.6 ± 0.5	30.9 ± 0.2	31.7 ± 0.2	32.5 ± 0.4
core-set-ent	13.1 ± 0.0	14.2 ± 0.1	15.1 ± 0.2	15.5 ± 0.3	16.0 ± 0.2	27.3 ± 0.2	29.2 ± 0.1	30.7 ± 0.2	31.3 ± 0.4	32.1 ± 0.2
BiB	**14.8** ± **0.3**	**15.9** ± **0.2**	**16.5** ± **0.1**	**16.9** ± **0.2**	**17.2** ± **0.2**	**30.6** ± **0.1**	**32.4** ± **0.3**	**33.1** ± **0.2**	**33.8** ± **0.1**	**34.1** ± **0.1**

Table 6: Performance of the loss strategy with different choices of the detector's loss on VOC07. For each experiment, we perform 5 cycles with a budget of 50 images per cycle. We have repeated the experiment six times for each strategy and report the average and standard deviation of their performance.

AL method	Number of fully-annotated images				
	50	100	150	200	250
$\mathcal{L}^{\text{MIL}}$	57.1 ± 0.3	57.9 ± 0.2	58.4 ± 0.5	59.4 ± 0.2	60.0 ± 0.3
$\mathcal{L}_w^{(1)}$	58.2 ± 0.4	58.5 ± 0.4	59.6 ± 0.7	60.3 ± 0.8	61.1 ± 0.5
$\mathcal{L}_w^{(2)}$	59.4 ± 0.3	**60.7** ± **0.2**	**61.4** ± **0.3**	61.8 ± 0.3	62.4 ± 0.1
$\mathcal{L}_w^{(3)}$	59.7 ± 0.2	60.5 ± 0.5	61.3 ± 0.7	**62.0** ± **0.5**	**62.5** ± **0.3**
$\sum_{k=1,2,3} \mathcal{L}_w^{(k)}$	**59.9** ± **0.4**	60.6 ± 0.5	60.9 ± 0.5	61.6 ± 0.3	62.2 ± 0.6

2.3 Ablation study on COCO.

We have provided an ablation study on different components of BiB on VOC07 dataset in the main paper. For completeness, we report in Table 7 the averaged AP50 scores (over 3 repetitions) of the ablation study on COCO. The results are similar to those obtained on VOC, except for the difficulty-aware sampling, which helps with the u-random strategy but not always with BiB.

2.4 Are diverse samples important?

We propose in BiB to find diverse images on which the weakly-supervised detector fails. We investigate the importance of sample diversity in BiB by comparing it to two variants. In the first variant, we randomly select images containing BiB pairs ('U(BiB)'), and in the second variant, we use a mix, with half selected with BiB and the other half with randomly uniform sampling ('U+BiB'), to be fully annotated. We show the results in Table 8. The fact that U(BiB) is worse than BiB and U+BiB outperforms U(BiB) in general shows that diversity sampling is important once BiB patterns have been discovered.

Table 7: **Ablation study on COCO**. Results in AP50 on COCO with 5 cycles and a budget $B = 160$. We provide averages and standard deviation results over several runs. *DifS* stands for the difficulty-aware region sampling module. Images are selected by applying k-means++ init. (*K selection*) on image-level features (*im.*), confident predictions' features (*reg.*) or BiB pairs.

DifS	K selection			AP50				
	im.	reg.	BiB	160	320	480	640	800
				29.0	30.6	31.4	32.3	32.8
✓				29.1	30.8	31.7	32.4	33.0
✓	✓			29.2	30.7	31.6	32.3	32.9
✓		✓		30.5	31.6	32.6	33.5	34.1
			✓	30.7	32.3	33.2	33.7	34.2
✓			✓	30.6	32.4	33.1	33.8	34.1

Table 8: A comparison between BiB, u-rand and two other variants that combine them. BiB outperforms the variants, showing that diversity sampling is important to the effectiveness of BiB.

Method	Dataset	AL cycles				
		1	2	3	4	5
u-rand.	VOC	56.5	58.4	59.3	60.2	61.1
U(BiB)		57.6	59.2	60.1	61.2	61.8
U+BiB		57.9	59.4	60.7	61.6	62.4
BiB		**58.5**	**60.8**	**61.9**	**62.9**	**63.5**
u-rand	COCO	29.1	30.8	31.7	32.4	33.0
U(BiB)		30.0	31.4	32.3	33.1	33.5
U+BiB		29.7	31.4	32.4	33.2	33.7
BiB		**30.6**	**32.4**	**33.1**	**33.8**	**34.1**

2.5 Verification of BiB pairs

We propose in our paper the use of BiB pairs as an indicator of a detector's confusion on images. With its design, we argue that at least one box in the pair is likely a wrong prediction. We verify this assumption on MIST's predictions on VOC07 and COCO. Among 8,758 BiB pairs on VOC, there are 8,633 pairs (98.6%) with at least one wrong prediction while 99.6% of the 854,004 BiB pairs have at least one wrong box on COCO.

2.6 Number of BiB examples reduced with active learning cycles.

We use BiB pairs as an indicator of the model's confusion on images. Intuitively, as the model becomes more accurate with more active learning cycles, fewer BiB pairs will be found. We computed the number of BiB pairs during active learning cycles on VOC07 and COCO datasets to verify this assumption. As expected,

our results show that it decreases with iterations. On VOC, it drops from 8801 in cycle 1 to 5170 in cycle 5 with budget $B = 50$. On COCO, it decreases from 854k in cycle 1 to 152k in cycle 5 with budget $B = 160$.

2.7 Influence of Hyper-Parameters

We use two intuitive hyper-parameters in BiB design: the area ratio μ between two boxes in a BiB pair and the ratio δ of the overlap over the smallest box. By design, the latter should be close to 1 so that the small box is "contained" in the large box and it is set to 0.8 in our experiments. For the former, we test BiB on VOC07 when its value varies in $\{2, 3, 4\}$ and report results in Table 9. It can be seen that the performance is relatively insensitive to μ. We use $\mu = 3$ in our experiments.

Table 9: Performance of BiB on VOC07 with different values of the area ratio μ in BiB design. We conducted 5 cycles with a budget of 50 images per cycle, repeated the experiment six times for each value of μ and report the average and standard deviation of their performance.

	Number of fully-annotated images				
μ	50	100	150	200	250
$\mu = 2$	58.5 ± 0.5	60.4 ± 0.3	61.6 ± 0.4	62.4 ± 0.3	63.1 ± 0.2
$\mu = 3$	58.5 ± 0.8	60.8 ± 0.5	61.9 ± 0.4	62.9 ± 0.5	63.5 ± 0.4
$\mu = 4$	58.3 ± 0.5	60.6 ± 0.3	61.7 ± 0.3	62.5 ± 0.4	63.3 ± 0.2

3 More Details

3.1 MIST Architecture

We use MIST [4] as our base weakly-supervised object detector and briefly describe it in the main submission. MIST follows OICR [6] and consists of a Multiple Instance Learner (MIL) trained to produce coarse detections which are then refined with several refinement heads using automatically-generated pseudo-boxes. We have given details about the refinement heads in the main paper and provide here a description of the MIL head as well as the procedure to generate the pseudo-boxes. We consider an image $\mathbf{I}$, its class labels $\mathbf{q} \in \{0, 1\}^C$ and the set of pre-computed region proposals $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$. Please note that we drop here the image index in order to ease understanding.

Multiple Instance Learner. MIL receives $\mathbf{I}$ and $\mathcal{R}$ as input and yields a class probability vector $\phi \in \mathbb{R}^C$. It is trained to classify the image with the Binary Cross Entropy (BCE) loss $\mathcal{L}_{\text{MIL}}$ on ϕ:

$$\mathcal{L}_{\text{MIL}} = -\frac{1}{C}\sum_{c=1}^{C} q(c)\log(\phi(c)) + (1 - q(c))\log(1 - \phi(c)). \tag{5}$$

In MIST, class probabilities ϕ are obtained by aggregating scores in a region score matrix $\mathbf{s} \in \mathbb{R}^{R \times C}$ with $c \in \{1, .., C\}$:

$$\phi(c) = \sum_{i=1}^{R} \mathbf{s}(i, c), \tag{6}$$

where $\mathbf{s} = \mathbf{s}_c \odot \mathbf{s}_d$ is the point-wise product of a classification score matrix $\mathbf{s}_c \in \mathbb{R}^{R \times C}$ and a detection score matrix $\mathbf{s}_d \in \mathbb{R}^{R \times C}$. Matrices $\mathbf{s}_c$ and $\mathbf{s}_d$ are built by concatenating projected regions features extracted with the backbone network for each of the regions in $\mathcal{R}$. Matrix $\mathbf{s}_c$ is normalized row-wise with the softmax operation and models the class probabilities of the region proposals. Matrix $\mathbf{s}_d$, which is normalized column-wise, represents the relative objectness of the proposals with respect to the corresponding classes. Given those interpretations, $\mathbf{s}(i, c)$ expresses the likelihood that region i is an object of class c.

Pseudo-boxes generation. MIST [4] introduces a heuristic to generate the pseudo-boxes $\mathbf{D}^{(k-1)}$ that are used to train the refinement heads k. Such boxes are generated either from the region score matrix $\mathbf{s}$ of the MIL (giving $\mathbf{D}^{(0)}$) or the region classification score matrices $\mathbf{s}^{(k)}$ $(k = 1, 2, 3)$ of the refinement heads (giving $\mathbf{D}^{(k)}$). In particular, for each ground-truth class c in image $\mathbf{I}$, the corresponding column scores $[\mathbf{s}(1, c), \ldots, \mathbf{s}(R, c)]$ in $\mathbf{s}$ (or $\mathbf{s}^{(k)}$) are sorted in descending order. Then, given the top-15% region proposals with the highest scores, we select all boxes that do not have an IoU$\geq$ 0.3 with a higher-ranked region. Selected boxes for all classes are aggregated to construct the final set of pseudo-boxes.

3.2 Active Learning Strategies

We compare in the main text our proposed BiB to different active learning strategies. We detail here all considered methods. As described in the Algorithm 1 of the submission, a set of images A^t of B images is selected at each cycle t. The selection is performed with an active learning method within the set of images W^{t-1}, possibly using the detector M^{t-1} trained at the end of the previous cycle and the set of selected images S^{t-1}.

Random. We implement two variants of a random sampling: *u-random* and *b-random*. In *u-random*, B images are selected uniformly at random from W^{t-1}; *b-random* seeks to have a balance sampling among the classes. Images are iteratively selected until the budget B is reached. At each iteration, an image containing at least an object of the class that is the least represented[1] in $S^{t-1} \cup A^t$ is randomly chosen and added to A^t.

[1] In case of draw, a class is randomly selected.

Diversity-based strategies. The core-set [5] approach attempts to select a representative subset of a dataset. We employ the greedy version of *core-set* in our experiments. In particular, at cycle t, let $\psi_{t-1}(\mathbf{I}_i)$ be the features of image $\mathbf{I}_i$ extracted from detector M^{t-1}, *core-set* iteratively selects the image i^* to be added in A^t by solving the optimization problem:

$$i^* = \underset{i \in W^{t-1} \setminus A^t}{\operatorname{argmax}} \min_{j \in S \cup A^t} \|\psi_{t-1}(\mathbf{I}_i) - \psi_{t-1}(\mathbf{I}_j)\|. \tag{7}$$

In the first cycle, the very first image is randomly selected.

Selection using model uncertainty. The concept of *informativeness* has been widely exploited in the literature [8,7,1,2]. For a classification task, the uncertainty can be computed by measuring the *entropy* over the class predictions of an image. Here, we first compute the entropy over the class predictions of each predicted box in an image, and then the box entropy-scores of an image are aggregated using the *sum* and *max* pooling, resulting in two strategies, *entropy-sum* and *entropy-max*. Concretely, let $p_{i,j} \in \mathbb{R}^{C+1}$ be the predicted class scores of the predicted box j for image $\mathbf{I}_i$ given by M^{t-1}, and $\mathbf{D}_i$ be the set of all predictions in $\mathbf{I}_i$, we compute the uncertainty score u_i of image $\mathbf{I}_i$ as

$$\max_{1 \le j \le |\mathbf{D}_i|} \sum_{c=1}^{C+1} -p_{i,j}^c \log(p_{i,j}^c) \tag{8}$$

for *entropy-max* and

$$\sum_{1 \le j \le |\mathbf{D}_i|} \sum_{c=1}^{C+1} -p_{i,j}^c \log(p_{i,j}^c) \tag{9}$$

for *entropy-sum*. Then, the B images with the highest scores in $\mathbf{u}$ are selected.

Combining diversity and uncertainty. Following [3], we consider a selection strategy function that incorporates the uncertainty information into *core-set* by multiplying the distances between image features with the uncertainty score $\mathbf{u}$ defined above. Specifically we combine *core-set* and *entropy-max*, in a new active learning method *core-set-ent* which iteratively selects an image i^* following:

$$i^* = \underset{i \in W^{t-1} \setminus A^t}{\operatorname{argmax}} \min_{j \in S \cup A^t} u_i \times \|\psi_{t-1}(\mathbf{I}_i) - \psi_{t-1}(\mathbf{I}_j)\|. \tag{10}$$

Selection using losses. In [7], the authors propose to learn – through an auxiliary module – an object detection loss predictor which later allows choosing samples that produce the highest losses. Conveniently, the refinement heads of MIST produce refinement losses ($\mathcal{L}_w^{(k)}$ with $k \in \{1, 2, 3\}$) that are *detection* losses computed using pseudo-boxes. We therefore propose the active learning method *loss* which selects the B images with the highest loss $\mathcal{L}_w^{(3)}$. We have discussed in Section 2.2 results obtained when considering different losses of MIST.

References

1. Brust, C.A., Kading, C., Denzler, J.: Active learning for deep object detection. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) (2019) 8
2. Choi, J., Elezi, I., Lee, H.J., Farabet, C., Alvarez, J.M.: Active learning for deep object detection via probabilistic modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 8
3. Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C., Alvarez, J.M.: Scalable active learning for object detection. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV) (2020) 8
4. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 3, 6, 7
5. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018) 8
6. 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 (CVPR) (2017) 6
7. Yoo, D., Kweon, I.S.: Learning loss for active learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 8
8. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 8