

## A Details of the Experimental Setup

### A.1 Datasets

Two large-scale image datasets are adopted in our experiments, including Pascal VOC 2007 [9] and MS-COCO 2017 [20]. Pascal VOC 2007 (VOC) is a standard dataset for object detection, consisting of 9,963 images with 24,640 box annotations. MS-COCO 2017 (COCO) is also a popular object detection benchmark, containing 328,000 images of generic objects. Following [22], clean annotations are perturbed to simulate noisy bounding box annotations in our experiments, which is performed once for each dataset. Specifically, let  $c_x, c_y, w, h$  represent the central x-axis coordinate, central y-axis coordinate, width, and height of a clean bounding box, respectively. We simulate a noisy bounding box by randomly shifting and scaling a clean one, which can be formulated as

$$\begin{cases} \hat{c}_x = c_x + \Delta_x \cdot w, & \hat{c}_y = c_y + \Delta_y \cdot h, \\ \hat{w} = (1 + \Delta_w) \cdot w, & \hat{h} = (1 + \Delta_h) \cdot h, \end{cases} \quad (15)$$

where  $\Delta_x, \Delta_y, \Delta_w,$  and  $\Delta_h$  obey the uniform distribution  $U(-n, n)$  and  $n$  is the noise level. For example, when  $n$  is set to 40%,  $\Delta_x, \Delta_y, \Delta_w,$  and  $\Delta_h$  would ranges from  $-0.4$  to  $0.4$ . Note that Equation 15 is conducted on each bounding box of the training set. Such a noise simulation can guarantee access to real ground-truths for analyzing training behaviors and evaluating the performance of box refinement. Noise levels are set to  $\{10\%, 20\%, 30\%, 40\%\}$  for VOC and  $\{20\%, 40\%\}$  for COCO.

### A.2 Implementation Details

Following [22], we implement our method on FasterRCNN [28] with ResNet-50 [12] as the backbone. The idea of DISCO can be easily generalized to other frameworks and we choose to perform our experiments with FasterRCNN as it is widely adopted [17]. As a common practice, the model is trained with the “ $1\times$ ” schedule [10]. Notably, all other training configurations are aligned with [22] to ensure fairness. As commonly done, mean average precision (mAP@.5) and mAP@[.5, 95] are used for VOC and COCO respectively. Specifically, we report AP<sub>50</sub> for VOC and  $\{\text{AP}, \text{AP}_{50}, \text{AP}_{75}, \text{AP}_S, \text{AP}_M, \text{AP}_L\}$  for COCO.

### A.3 Hyperparameter Selections

There are six hyperparameters in DISCO, including the temperature coefficient  $T$ , the augmented proposal number  $N'$ , two box fusion hyperparameters  $\alpha$  and  $\beta$ , and two loss weights  $\gamma$  and  $\lambda$ . As there are no additional validation sets available, we tuned these hyperparameters based on the performance on the training set with clean annotations, which can also avoid the leakage of test data. For the sake of simplicity, we empirically fix  $N'$  and  $\gamma$  to 10 and 0.3, and then tuning  $T \in [0.01, 0.2]$ ,  $\alpha \in [3, 10]$ ,  $\beta \in [0.7, 0.9]$ , and  $\lambda \in [0.01, 0.2]$ . To ensure reproducibility, the selected hyperparameters for all settings are reported in Table 4. Notably, we have just roughly tuned these hyperparameters by selecting some regular values, thus the performance of our method in Table 1 has the potential to be better.

Dataset	Noise Level	Hyperparameter					
		$T$	$N'$	$\alpha$	$\beta$	$\gamma$	$\lambda$
VOC	10%	0.05	10	10	0.7	0.3	0.05
	20%	0.05	10	10	0.7	0.3	0.05
	30%	0.1	10	10	0.8	0.3	0.1
	40%	0.1	10	5	0.8	0.3	0.1
COCO	20%	0.01	10	10	0.7	0.3	0.01
	40%	0.1	10	5	0.8	0.3	0.1

Table 4: **Hyperparameter selections.** We report the hyperparameters for all settings to ensure reproducibility.

Hyper.	Value	AP <sub>50</sub>	Hyper.	Value	AP <sub>50</sub>	Hyper.	Value	AP <sub>50</sub>
$T$	0.01	68.6	$N'$	5	68.5	$\alpha$	3	68.3
	0.1	<b>68.7</b>		10	<b>68.7</b>		5	<b>68.7</b>
	0.2	67.8		20	68.6		7	68.2
Hyper.	Value	AP <sub>50</sub>	Hyper.	Value	AP <sub>50</sub>	Hyper.	Value	AP <sub>50</sub>
$\beta$	0.7	68.1	$\gamma$	0.1	68.3	$\lambda$	0.05	67.9
	0.8	<b>68.7</b>		0.3	<b>68.7</b>		0.1	<b>68.7</b>
	0.9	67.3		0.5	68.4		0.15	67.4

Table 5: **Ablation studies of hyperparameter sensitivity.** DISCO can still achieve relatively stable performance when these hyperparameters vary within a moderate range.

## B More Ablation Studies

In this section, we conduct more ablation studies to further verify the effectiveness of the proposed DISCO. These ablation studies contain hyperparameter sensitivity, backbone compatibility, and the execution number of DISCO. Unless otherwise specified, the following experiments are all based on VOC at the 40% noise level.

### B.1 Hyperparameter Sensitivity

Here we evaluate the sensitivity of the hyperparameters used in DISCO. Note that we choose some moderate values rather than extreme ones to reasonably evaluate the sensitivity of each hyperparameter. The experimental results are reported in Table 5. As we can see, the temperature coefficient  $T$  is relatively robust when set to 0.01 or 0.2. Tuning  $T$  to a proper value can contribute to better performance. Besides, it can be observed that the augmented proposal number  $N'$  is insensitive when varying from 5 to 20. This is the reason why we empirically fix  $N'$  to 10 for all settings. Moreover, the hyperparameters regulate the fusion of two bounding boxes (*i.e.*,  $\alpha$  and  $\beta$ ) is also insensitive when varying within a moderate range, showing the effectiveness of our method. Furthermore, two loss weights  $\gamma, \lambda$  also remain insensitive while  $\lambda$  is relatively crucial. This is because it controls the strength of an extra classification loss term, directly affecting classification accuracy.

Method	Backbone	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	Execution Num.	
								AP <sub>50</sub>	
OA-MIL [22]	ResNet-50	18.6	42.6	12.9	9.2	19.0	26.5	1	68.1
DISCO (Ours)		<b>21.2</b>	<b>45.7</b>	<b>16.9</b>	<b>11.4</b>	<b>24.7</b>	<b>27.8</b>		
OA-MIL [22]	ResNet-101	19.3	44.1	13.1	9.3	20.8	27.8	3	67.9
DISCO (Ours)		<b>22.7</b>	<b>47.6</b>	<b>18.4</b>	<b>12.9</b>	<b>26.6</b>	<b>29.8</b>		

Table 6: Left: **Ablation studies of backbone compatibility.** The experiment is conducted on COCO at 40% noise level with ResNet-50 and ResNet-101. DISCO can still outperform OA-MIL when equipped with different backbones. Right: **Ablation studies of the execution number of DISCO.** Our execution strategy can achieve superior performance.

## B.2 Backbone Compatibility

Following [22], the benchmark experiments are performed with ResNet-50 [12] as the backbone. To further demonstrate the superior performance of our method, we conduct an additional experiment based on different backbones. Specifically, in this experiment, DISCO [22] is compared to OA-MIL on COCO at the 40% noise level with the backbone set to ResNet-101 [12], and other experiment setups remain the same. In this way, we aim to evaluate the performance of our DISCO for a large-scale dataset when it is equipped with an advanced backbone. The experimental results are reported in Table 6. It can be observed that DISCO can further improve performance and still achieve SOTA results.

## B.3 Execution Number of DISCO

In this work, DISCO is performed twice in a training iteration, where the first time is for proposal re-assignment and the second time is for obtaining better supervision. We compare such an execution strategy with two other options: 1) The execution number of DISCO is set to 1: proposal re-assignment is removed and the only one time of DISCO is for obtaining better supervision; 2) The execution number of DISCO is set to 3: the first two times are for proposal re-assignment and the third time is for obtaining better supervision. As shown in Table 6, more execution numbers of DISCO do not contribute to better detection performance. This is because such an improper strategy could result in excessive box refinement and thus influence the learning stability of detectors. Moreover, it also can be observed that our execution strategy can achieve superior performance.

# C More Qualitative Results

## C.1 Box Refinement

As an extension to Figure 5, we present more qualitative results of box refinement in DISCO (see Figure 6), which shows that DISCO can attain tighter bounding boxes than noisy ground-truths. As shown in Figure 6, it is worth noting that DISCO can achieve consistent refinement of bounding boxes for different objects varying in size.

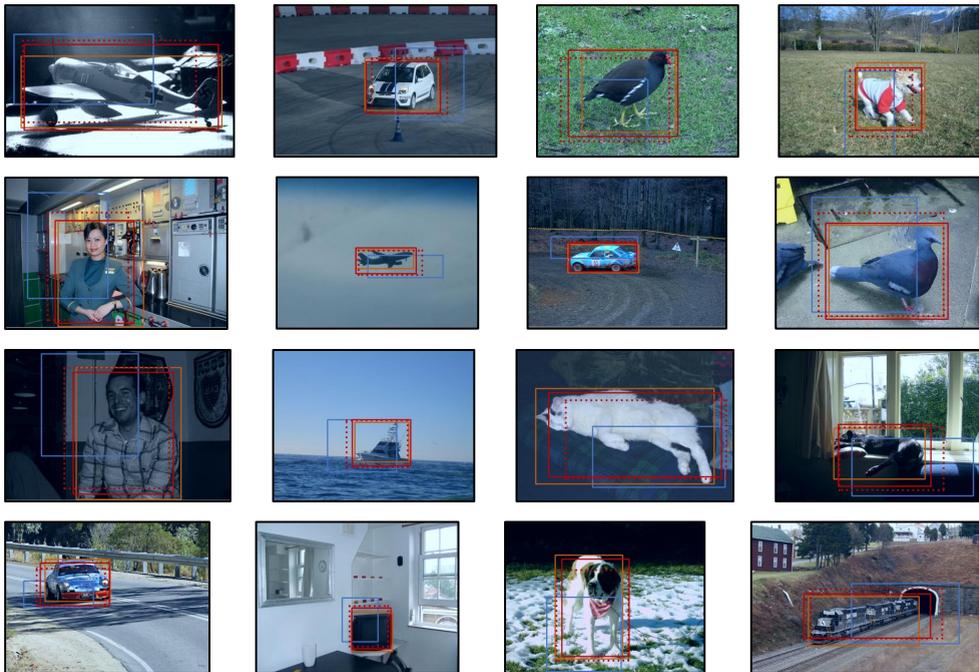


Figure 6: **Qualitative results of box refinement in DISCO.** Real ground-truths and noisy ground-truths are marked in *orange* and *blue*. Refined bounding boxes produced by the first-/second-time DISCO are indicated in *dotted/solid red*. The first-time refined boxes can cover the objects more tightly than noisy ground-truths, and the second-time refinement can further contribute to more precise ones.

## C.2 Interpretability

In Figure 7, more qualitative results of interpretability in DISCO are provided to demonstrate such a characteristic of our method. As shown in Figure 7, when trained with DISCO, the detector can output a reasonable variance as the confidence for each border of predicted bounding boxes, which shows that the detector is capable of realizing which border may be inaccurately predicted.

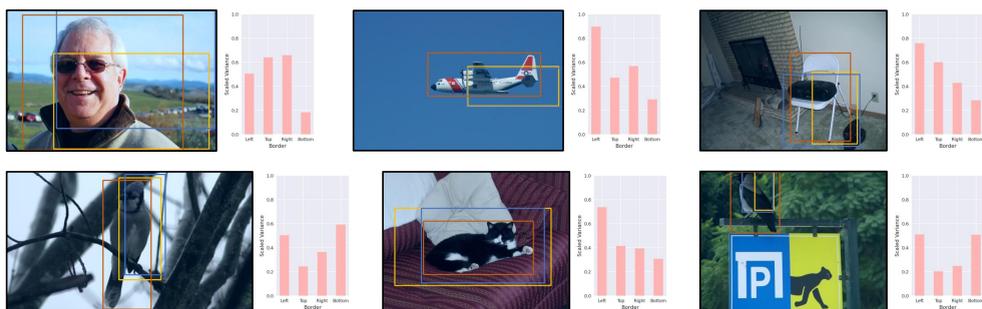


Figure 7: **Qualitative results of interpretability in DISCO.** We randomly choose an assigned proposal (yellow) per image to report its estimated variances. Real ground-truths and noisy ground-truths are marked in orange and blue. Note that the variance is scaled by the width and height for clarity. With the proposed DA-Est, DISCO can estimate reasonable variances for each border of box prediction.