Appendix

Table 1. Comparison of our method on PASCAL VOC 2007 test set to state-of-the-art WSOD methods in terms of mAP (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	AP
Pure WSOD:																					
WSDDN [2]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
OICR$^+$ [26]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
PCL$^+$ [25]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
Yang *et al.*$^+$ [29]	57.6	70.8	50.7	28.3	27.2	72.5	69.1	65.0	26.9	64.5	47.4	47.7	53.5	66.9	13.7	29.3	56.0	54.9	63.4	65.2	51.5
C-MIDN$^+$ [28]	53.3	71.5	49.8	26.1	20.3	70.3	69.9	68.3	28.7	65.3	45.1	64.6	58.0	71.2	20.0	27.5	54.9	54.9	69.4	63.5	52.6
Arun *et al.* [1]	66.7	69.5	52.8	31.4	24.7	74.5	74.1	67.3	14.6	53.0	46.1	52.9	69.9	70.8	18.5	28.4	54.6	60.7	67.1	60.4	52.9
WSOD2$^+$ [31]	65.1	64.8	57.2	39.2	24.3	69.8	66.2	61.0	29.8	64.6	42.5	60.1	71.2	70.7	21.9	28.1	58.6	59.7	52.2	64.8	53.6
GradingNet-C-MIL [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.3
MIST-Full [20]	68.8	77.7	57.0	27.7	28.9	69.1	74.5	67.0	32.1	73.2	48.1	45.2	54.4	73.7	35.0	29.3	64.1	53.8	65.3	65.2	54.9
IM-CFB$^+$ [30]	63.3	77.5	48.3	36.0	32.6	70.8	71.9	73.1	29.1	68.7	47.1	69.4	56.6	70.9	22.8	24.8	56.0	59.8	73.2	64.6	55.8
CASD [9]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.8
SoS [24]	72.9	79.4	59.6	20.4	49.8	81.2	82.9	84.0	31.5	76.6	57.4	60.7	74.7	75.1	33.0	34.3	66.3	61.1	80.6	71.8	62.7
SoS$^+$ [24]	77.9	81.2	58.9	26.7	54.3	82.5	84.0	83.5	36.3	76.5	57.5	58.4	78.5	78.6	33.8	37.4	64.0	63.4	81.5	74.0	64.4
OICR+REG (reproduce)	54.0	61.9	43.9	22.6	31.7	73.8	65.1	60.6	14.4	68.0	17.0	48.8	58.3	69.9	12.8	22.0	53.9	53.6	69.7	60.4	48.3
CASD (reproduce)	68.8	67.2	53.9	38.2	21.5	70.4	69.7	68.9	23.6	66.3	48.8	62.3	56.4	70.6	17.3	24.9	55.9	58.9	66.0	69.1	54.0
OICR+REG+W2N (Ours)	71.0	74.2	60.8	28.8	44.6	78.0	72.6	80.3	16.7	74.3	24.3	58.2	64.6	75.1	13.3	29.9	60.3	65.3	80.1	67.6	**57.0(+8.7)**
CASD+W2N (Ours)	74.0	81.7	71.2	48.9	51.0	78.6	82.3	83.5	29.1	76.9	51.5	82.1	76.9	79.1	28.5	34.3	65.0	64.2	75.2	74.8	**65.4(+11.4)**
WSOD with transfer learning:																					
MSD-Ens$^+$ [14]	70.5	69.2	53.3	43.7	25.4	68.9	68.7	56.9	18.4	64.2	15.3	72.0	74.4	65.2	15.4	25.1	53.6	54.4	45.6	61.4	51.1
OICR+UBBR [11]	59.7	44.8	54.0	36.1	29.3	72.1	67.4	70.7	23.5	63.8	31.5	61.5	63.7	61.9	37.9	15.4	55.1	57.4	69.9	63.6	52.0
LBBA$^+$ [6]	70.3	72.3	48.7	38.7	30.4	74.3	76.6	69.1	33.4	68.2	50.5	67.0	49.0	73.6	24.5	27.4	63.1	58.9	66.0	69.2	56.6
Zhong *et al.* (R50-C4)$^+$ [33]	64.8	50.7	65.5	45.3	46.4	75.7	74.0	80.1	31.3	77.0	26.2	79.3	74.8	66.5	57.9	11.5	68.2	59.0	74.7	65.5	59.7
TraMaS$^+$ [17]	68.6	61.1	69.6	48.1	49.9	76.3	77.8	80.9	34.9	77.0	31.1	80.9	78.5	66.3	64.0	19.1	69.1	62.3	74.4	69.1	62.9
CaT_5 [4]	74.0	70.7	60.0	31.1	50.0	75.9	82.0	70.7	32.8	74.3	69.5	70.2	69.5	77.0	37.5	45.8	67.0	61.1	72.4	68.0	63.0
LBBA (reproduce)	70.2	75.5	49.2	41.9	30.5	80.5	78.2	72.8	36.4	73.8	52.3	67.0	46.4	76.2	34.6	29.4	67.9	66.6	68.3	74.1	59.1
LBBA+W2N (Ours)	71.8	83.0	69.9	50.3	54.5	79.0	83.9	83.9	39.4	79.2	52.9	82.2	83.6	79.2	62.6	32.7	68.5	66.1	75.8	74.5	**68.6(+9.5)**
Upper bounds:																					
Faster R-CNN (Res50+FPN) [19]	82.8	84.2	75.2	62.4	67.0	81.4	87.1	82.6	57.3	82.5	64.9	83.0	84.0	82.7	83.7	54.0	76.1	73.4	81.8	76.1	**76.1**

A Experiments

A.1 Implementation Details

In this subsection we show some implementation details of experiments.

Overall. All programs are conducted based on PyTorch toolkit and run on NVIDIA GTX1080Ti GPU ×8.

Training of Weakly Supervised Detector. In this work we select three WSOD baseline methods to play the role of generators: *OICR+REG*, *CASD* and *LBBA*. For OICR+REG, we adopt the code by [22], and for CASD and LBBA, we adopt code with official implementations. To make a long story short, the is no any trick introduced in training phase and all the training configuration and hyper-parameter are the same as the default version of the code base, except for that we select the COCO-60-clean [6,33] as the auxiliary dataset to train the LBBA [6] network.

Noisy Label Generation. We follow the Pseudo Ground-truth Excavation (PGE) algorithm [32] to implement the noisy label generation on training set, where the threshold T_{nms} for NMS is set to 0.3, while T_{score} and T_{fusion} are set to 0.2 and 0.4 respectively.

Table 2. Comparison of our method on PASCAL VOC 2007 trainval set to state-of-the-art WSOD methods in terms of CorLoc (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	CorLoc
Pure WSOD:																					
WSDDN [2]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
OICR$^+$ [26]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
PCL$^+$ [25]	79.6	85.5	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	68.5	75.7	78.9	62.7
Li$^+$ [13]	85.0	83.9	58.9	59.6	43.1	79.7	85.2	77.9	31.3	78.1	50.6	75.6	76.2	88.4	49.7	56.4	73.2	62.6	77.2	79.9	68.6
C-MIL$^+$ [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
Yang *et al.*$^+$ [29]	80.0	83.9	74.2	53.2	48.5	82.7	86.2	69.5	39.3	82.9	53.6	61.4	72.4	91.2	22.4	57.5	83.5	64.8	75.7	77.1	68.0
MIST (Full)$^+$ [20]	87.5	82.4	76.0	58.0	44.7	82.2	87.5	71.2	49.1	81.5	51.7	53.3	71.4	92.8	38.2	52.8	79.4	61.0	78.3	76.0	68.8
WSOD2$^+$ [31]	87.1	80.0	74.8	60.1	36.6	79.2	83.8	70.6	43.5	88.4	46.0	74.7	87.4	90.8	44.2	52.4	81.4	61.8	67.7	79.9	69.5
Arun *et al.*[1]	88.6	86.3	71.8	53.4	51.2	87.6	89.0	65.3	33.2	86.6	58.8	65.9	87.7	93.3	30.9	58.9	83.4	67.8	78.7	80.2	70.9
GradingNet-C-MIL [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72.1
IM-CFB$^+$ [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72.2
OICR+REG (reproduce)	91.6	78.3	62.6	46.0	44.8	86.4	87.7	80.3	34.4	87.1	30.1	69.4	81.1	90.8	31.3	44.8	76.0	76.1	83.1	60.5	67.4
CASD (reproduce)	68.8	67.2	53.9	38.2	21.5	70.4	69.7	68.9	23.6	66.3	48.8	62.3	56.4	70.6	17.3	24.9	55.9	58.9	66.0	69.1	68.5
OICR+REG+W2N (Ours)	87.4	86.0	69.7	50.8	59.8	89.8	88.4	86.9	37.5	86.5	26.0	69.8	84.0	95.1	31.6	57.6	78.12	75.6	85.8	77.3	**71.2(+3.8)**
CASD+W2N (Ours)	92.0	90.5	82.4	71.3	73.0	85.5	94.7	89.0	46.3	89.4	63.5	87.9	92.7	96.7	47.1	70.2	84.4	75.1	82.4	87.5	**80.1(+12.6)**
WSOD with transfer learning:																					
OICR+UBBR [11]	47.9	18.9	63.1	39.7	10.2	62.3	69.3	61.0	27.0	79.0	24.5	67.9	79.1	49.7	28.6	12.8	79.4	40.6	61.6	28.4	47.6
WSLAT-Ens [21]	78.6	63.4	66.4	56.4	19.7	82.3	74.8	69.1	22.5	72.3	31.0	63.0	74.9	78.4	48.6	29.4	64.6	36.2	75.9	69.5	58.8
MSD-Ens$^+$ [14]	89.2	75.7	75.1	66.5	58.8	78.2	88.9	66.9	28.2	86.3	29.7	83.5	83.3	92.8	23.7	40.3	85.6	48.9	70.3	68.1	66.8
Zhong *et al.* (R50-C4)$^+$ [33]	87.5	64.7	87.4	69.7	67.9	86.3	88.8	88.1	44.4	93.8	31.9	89.1	92.9	86.3	71.5	22.7	94.8	56.5	88.2	76.3	74.4
LBBA$^+$ [6]	93.3	90.6	71.8	69.2	59.5	90.9	94.4	78.5	55.4	96.6	51.0	82.3	72.5	93.2	48.5	52.8	100.0	66.7	78.3	87.5	76.7
TraMaS$^+$ [17]	90.6	67.4	89.7	70.5	72.8	86.6	91.7	89.8	51.0	96.1	34.0	93.7	94.8	90.3	73.0	26.5	95.2	68.2	89.8	83.1	77.7
CaT$_5$ [4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.3
LBBA (reproduce)	86.9	84.5	74.6	65.6	55.1	85.4	86.8	84.4	42.5	88.0	45.0	83.3	82.3	88.6	47.6	49.1	88.3	50.8	81.1	84.3	72.7
LBBA+W2N (Ours)	89.5	93.4	83.9	70.2	73.4	87.1	94.5	92.0	58.9	95.7	64.0	91.0	94.8	93.5	80.7	64.1	91.7	78.2	84.3	89.1	**83.5(+10.8)**
Upper bounds:																					
Faster R-CNN (Res50+FPN)[19]	91.7	93.7	92.6	75.0	84.0	95.4	95.3	93.2	76.5	94.5	86.9	92.3	96.0	93.2	93.0	76.8	94.9	89.2	85.7	90.4	**89.5**

Learning Detector with Noisy Annotations For localization adaptation stage, we adopt Faster R-CNN [19] with backbone of ResNet-50 [8] combined with FPN [15] as the supervised object object detector f .During training, the f is optimized by stochastic gradient descent (SGD) [3] with the batch size of 16, initialized learning rate of 0.02, momentum of 0.9 and weight decay of 1×10^{-4}. The number steps of training is set to 5,000, 10,000, 90,000, 250,000 and the learning rate is decayed by 0.1 after 3,500, 7,000, 60,000, 180,000 steps for PASCAL VOC 2007, PASCAL VOC 2012, MS-COCO and ILSVRC respectively. τ_{score}, τ_{assign}, λ_{re}, α and β are set to 0.1, 0.5, 1.0, 0.05 and 0.8 respectively. After training, we use f to refine the noisy labels set $\mathbb{X}_p$ by the same procedure of the noisy label generation. For semi-supervised learning stage, we choose the *two tasks instance-level* data split method and the proportion p of clean data is set to 60%. We implement the semi-supervised object detection (SSOD) framework by Unbiased Teacher [18] with its official code. During SSOD training, batch size of labeled set and unlabeled set are set to 8, and the learning rate is set to 0.01, the number of training step is set to 30,000, 50,000, 100,000 and 200,000 for PASCAL VOC 2007, PASCAL VOC 2012, MS-COCO and ILSVRC respectively and learning rate decay is not adopted. The iteration time T for the whole iterative manner is set to 2 for CASD and LBBA while set to 1 for OICR+REG and the further ablation study has shown in *submission file*. The weight of unsupervised loss λ_u is set to 2 for all settings. Other hyper-parameters are adopted the default configuration of [18].

Table 3. Comparison of our method on PASCAL VOC 2012 test set to state-of-the-art WSOD methods in terms of mAP (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	AP
Pure WSOD:																					
OICR$^+$ [26]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
PCL$^+$ [25]	58.2	66.0	41.8	24.8	27.2	55.7	55.2	28.5	16.6	51.0	17.5	28.6	49.7	70.5	7.1	25.7	47.5	36.6	44.1	59.2	40.6
Yang *et al.*$^+$	64.7	66.3	46.8	28.5	28.4	59.8	58.6	70.9	13.8	55.0	15.7	60.5	63.9	69.2	8.7	23.8	44.7	52.7	41.5	62.6	46.8
WSOD2$^+$ [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.2
Arun *et al.* [1]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48.4
IM-CFB [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.4
C-MIDN$^+$ [28]	72.9	68.9	53.9	25.3	29.7	60.9	56.0	78.3	23.0	57.8	25.7	73.0	63.5	73.7	13.1	28.7	51.5	35.0	56.1	57.5	50.2
GradingNet-C-MIL [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.5
MIST (Full)$^+$ [20]	78.3	73.9	56.5	30.4	37.4	64.2	59.3	60.3	26.6	66.8	25.0	55.0	61.8	79.3	14.5	30.3	61.5	40.7	56.4	63.5	52.1
CASD [9]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	53.6
SoS [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59.6
SoS$^+$ [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.9
OICR+REG (reproduce)	57.1	60.7	45.7	24.9	30.6	62.7	52.4	65.9	24.4	52.5	19.1	50.0	55.6	69.5	10.3	22.9	51.0	42.3	61.8	54.6	45.7
CASD (reproduce)	65.9	69.3	55.3	27.9	40.2	61.6	61.6	75.3	32.4	55.2	22.7	51.6	60.0	74.7	10.0	27.1	55.1	48.2	68.3	61.1	51.2
OICR+REG+W2N (Ours)	75.2	76.7	63.8	32.9	48.7	70.3	70.3	81.1	38.5	63.9	23.8	57.5	69.1	78.6	9.8	36.4	65.6	54.7	77.8	64.0	**57.9(+12.2)**
CASD+W2N (Ours)	81.8	78.9	69.8	33.5	48.0	75.0	73.9	84.9	33.2	71.1	16.2	84.9	78.1	78.4	11.2	38.9	71.7	45.7	77.5	64.5	**60.8(+9.6)**
WSOD with transfer learning:																					
MSD-Ens$^+$ [14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.4
LBBA	77.0	71.0	62.0	40.0	37.5	67.4	62.5	68.3	23.6	71.4	25.6	78.4	71.9	74.3	6.7	29.2	62.8	50.6	47.8	62.1	**54.5**
LBBA (reproduce)	76.8	71.4	61.4	40.0	38.1	66.6	63.8	69.9	22.6	65.7	23.9	77.5	72.8	74.4	6.4	29.3	59.3	51.4	47.2	62.8	54.0
LBBA+W2N (Ours)	81.4	80.7	72.6	39.5	52.7	78.2	76.7	82.3	34.9	77.4	20.9	83.6	79.4	81.9	11.1	37.7	75.3	49.4	74.2	65.4	**62.7(+8.7)**

Table 4. Comparison of our method on PASCAL VOC 2012 trainval set to state-of-the-art WSOD methods in terms of CorLoc (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	CorLoc
Pure WSOD:																					
OICR$^+$ [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.1
PCL$^+$ [25]	77.2	83.0	62.1	55.0	49.3	83.0	75.8	37.7	43.2	81.6	46.8	42.9	73.3	90.3	21.4	56.7	84.4	55.0	62.9	82.5	63.2
Shen [23]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.5
Li$^+$ [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.9
C-MIL$^+$ [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
Yang *et al.*$^+$ [29]	82.4	83.7	72.4	57.9	52.9	86.5	78.2	78.6	40.1	86.4	37.9	67.9	87.6	90.5	25.6	53.9	85.0	71.9	66.2	84.7	69.5
Arun *et al.*[1]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.5
WSOD2$^+$ [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9
MIST (Full)$^+$ [20]	91.7	85.6	71.7	56.6	55.6	88.6	77.3	63.4	53.6	90.0	51.6	62.6	79.3	94.2	32.7	58.8	90.5	57.7	70.9	85.7	70.9
IM-CFB$^+$ [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.6
GradingNet-C-MIL [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9
OICR+REG (reproduce)	78.8	80.3	67.1	49.4	52.2	88.7	73.9	74.0	50.9	81.8	37.3	59.8	77.1	86.9	21.6	46.6	74.0	70.6	72.3	79.0	66.1
CASD (reproduce)	71.0	83.1	75.1	49.6	61.7	91.0	79.0	81.1	56.8	75.4	38.5	61.2	80.0	87.1	19.0	56.0	87.8	67.0	75.9	81.0	69.4
OICR+REG+W2N (Ours)	87.4	86.0	69.7	50.8	59.8	89.8	88.4	86.9	37.5	86.5	26.0	69.8	84.0	95.1	31.6	57.6	78.12	75.6	85.8	77.3	**71.2(+5.1)**
CASD+W2N (Ours)	89.7	88.4	82.1	53.5	71.8	93.4	88.2	86.8	59.9	94.1	57.4	85.3	94.6	92.8	32.9	64.7	94.5	64.1	84.2	87.3	**78.3(+8.9)**
WSOD with transfer learning:																					
LBBA$^+$ [6]	92.3	90.0	85.0	68.0	63.2	92.8	82.4	66.8	57.2	96.8	54.1	80.4	92.1	94.4	16.8	66.4	94.5	70.8	71.8	91.3	76.4
LBBA (reproduce)	86.9	84.5	74.6	65.6	55.1	85.4	86.8	84.4	42.5	88.0	45.0	83.3	82.3	88,6	47.6	49.1	88.3	50.8	81.1	84.3	72.7
LBBA+W2N (Ours)	93.1	91.5	85.0	68.1	76.1	96.0	90.2	86.8	63.3	95.7	57.3	86.3	94.8	94.3	27.6	66.4	93.2	67.1	83.6	87.1	**80.2(+7.5)**

A.2 Comparison with State-of-the-arts

First we compare our method with several state-of-the-art WSOD approaches in terms of mAP and CorLoc on PASCAL VOC 2007 and 2012 [7] reported by Table 1, 3, 2 and 4. Our all results are obtained with single-scale testing approach. Based on these results, we obtain the following observations: First, our W2N framework outperforms all WSOD baselines in terms of both mAP and CorLoc. Specifically, on PASCAL VOC 2007 dataset, it outperforms OICR+REG by 8.7% mAP and 3.8% CorLoc, outperforms CASD by 11.4% mAP and 12.6% CorLoc, and outperforms LBBA by 9.5% mAP and 10.7% CorLoc. Performance on PASCAL VOC 2012 also demonstrates favorable performance improvement. Second, our W2N outperforms all of the state-of-the-art WSOD methods as well as transfer learning based methods. Specifically, OICR+REG+W2N achieve 57.0% mAP on PASCAL VOC 2007 test set, outperforming CASD [9] by 0.2% mAP which is the state-of-the-arts of pure WSOD method. CASD+W2N achieves

Table 5. Results of our method on MS-COCO 2017 validation set.

Methods	mAP	AP50	AP75	AP_S	AP_M	AP_L
OICR+REG [26] (reproduce)	9.8	20.8	7.9	1.4	9.2	17.7
OICR+REG+W2N	**15.3**	**30.0**	**13.9**	**4.9**	**18.5**	**24.6**
CASD [9] (reproduce)	10.5	24.1	8.3	2.7	12.2	18.3
CASD+W2N	**15.9**	**33.3**	**13.4**	**5.6**	**18.4**	**27.2**

Table 6. Results of our method on ILSVRC 2013 detection validation set.

Methods	AP50
OICR+REG [26]	17.4
OICR+REG+W2N	**22.6**
CASD [9]	18.4
CASD+W2N	**27.9**

65.4% mAP on PASCAL VOC 2007 test set, outperforming CaT_5 by 2.4% mAP which is the state-of-the-arts of transfer learning based WSOD method. Moreover, LBBA+W2N obtains 68.6% mAP and 83.4% CorLoc, which achieves a new state-of-the-arts for WSOD problem and catches up with the performance of fully supervised methods Faster R-CNN.

Fig. 1 shows the visualization results of our method on PASCAL VOC 2007 test set. The top row is the detection results of LBBA (reproduce) , the medium row is the detection results of LBBA+W2N and the bottom row is ground-truth annotations. Obviously, with the help of W2N framework, the bounding box predicted by model have a better location performance . And the phenomenon that predict bounding boxes covering at the discriminative part have been eased.

A.3 Results on More Datasets

We also deploy more experiments on MS-COCO 2017 dataset [16] and ILSVRC 2013 detection dataset [5] to prove the effectiveness of our method. MS-COCO dataset includes 80 categories with 118,287 training images and 5,000 validation images. ILSVRC 2013 detection dataset include 200 categories with 395,909 train images and 20,121 validation images. We choose *OICR+REG* to implement the weakly supervised detector and conduct our proposed on these two datasets and the results are shown in Table 5 and Table 6. With the help of W2N training process, the OICR+REG outperform 10.8% and 5.5% in terms of AP50 and mAP on MS-COCO, 5.2% AP50 on ILSVRC 2013, which demonstrates the generalization ability of our W2N framework.

A.4 More Ablation Study

Effect of two modules. Table 7 shows the ablation study of each module on different WSOD baselines. Simply re-training Faster R-CNN(FRCNN*) with pseudo GT only outperforms *OICR+REG* by 0.7% mAP, *CASD* by 1.0% mAP

Table 7. Effect of two modules on VOC 2007.

WSOD	FRCNN*	LA	SSL	ITER	mAP
OICR+REG[26]					48.3
	✓				49.0
		✓			49.9
			✓		56.1
		✓	✓		56.8
		✓	✓	✓	**57.0**
CASD[9]					54.0
	✓				55.0
		✓			55.6
			✓		62.0
		✓	✓		62.7
		✓	✓	✓	**65.4**
LBBA[6]					59.1
	✓				59.4
		✓			60.3
			✓		66.1
		✓	✓		67.0
		✓	✓	✓	**68.6**

Table 8. Training and Inference time comparison.

Second phase methods	Training time(h/stage)	Inference time(s/img)
Faster R-CNN	0.5	0.06
SoS	7.2	0.06
W2N(Ours)	7.8	0.06

Table 9. Effect of different backbone of W2N on VOC 2007.

Methods	mAP
LBBA(VGG16)	59.1
LBBA(VGG16)+W2N(VGG16)	66.7
LBBA(VGG16)+W2N(Res50+FPN)	**68.6**

and *LBBA* by 0.7% mAP respectively. By introducing localization adaption and semi-supervised learning separately, these improvements respectively outperform the *OICR+REG* by 1.6% and 7.8% mAP; outperform the *CASD* by 1.6% and 8.0% mAP; and outperform the *LBBA* by 1.2% and 7.0% mAP; Furthermore, our full method combining these two modules can further improve the detection performance to 56.8%, 62.6% and 67.0% mAP respectively.

Time cost of W2N: We show the training and inference time of W2N in Table 8. Our framework increase slight training time in comparison to SoS [24], while having the same inference time as Faster R-CNN and SoS, which illustrates the high efficiency of our method.

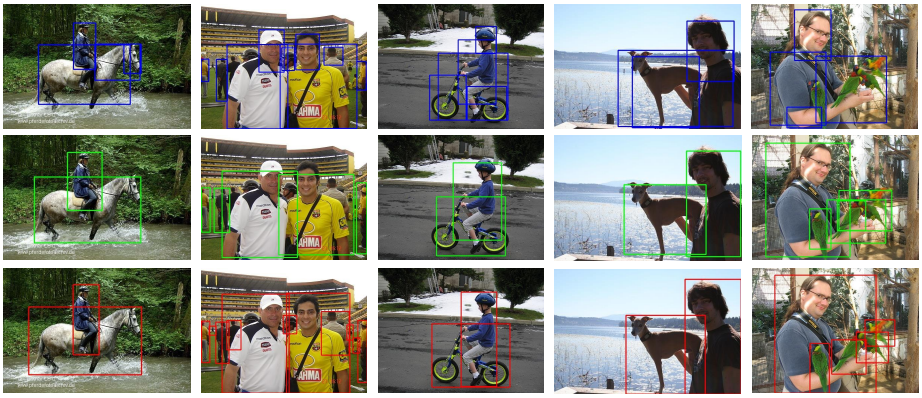


Fig. 1. Visualization results of our method on PASCAL VOC 2007 test set. **Top row**: detection results from LBBA. **Medium row**: detection results from LBBA with W2N. **Bottom row**: ground-truth annotations.

Effect of different backbones: For a fair comparison with [33,24], we use VGG16 as the backbone of WSOD baselines and use Res50 with FPN as the backbone of target detector. Furthermore, we additionally conduct experiments by using Faster R-CNN with VGG-16 backbone as target detector to explore the performance that our method adapted on other backbones. Table 9 shows the effect of using different backbones as target detectors. W2N leads to better mAP on better backbone.

Table 10. Comparison of different dataset split method on Pascal VOC 2007 for different WSOD baseline at iteration 0.

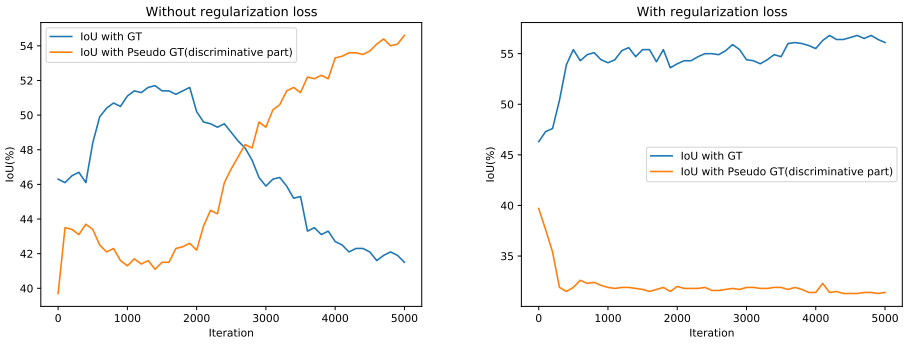
Methods	image-level split	instance-level split	two tasks instance-level split	ideal split	mAP
OICR+REG+W2N	✓				55.4
OICR+REG+W2N		✓			56.8
OICR+REG+W2N			✓		**56.8**
OICR+REG+W2N				✓	**61.4**
CASD+W2N	✓				61.9
CASD+W2N		✓			62.6
CASD+W2N			✓		**62.7**
CASD+W2N				✓	**66.5**
LBBA+W2N	✓				65.4
LBBA+W2N		✓			66.8
LBBA+W2N			✓		**67.0**
LBBA+W2N				✓	**72.8**

B Discussion

B.1 Ideal Data Split

In this work we propose a *hybrid-level dataset split* method, which aim to keep more pseudo label with high quality for subsequent SSOD training process. Theoretically, if we can keep more label with high quality, then the SSOD will perform better, which may be a important research direction for future work. To prove that the quality of labeled set will affect the performance of detector, we propose an simple *ideal data split* with the help of ground-truth dataset $\mathbb{X}_{gt}$. $\mathbb{X}_{gt} = \{\mathbf{I}, \{\mathbf{S}_{gt}\}\}$, $\mathbf{S}_{gt} = (\mathbf{b}_{gt}, c_{gt})$, where $\mathbf{S}_{gt}$ denotes a ground-truth box annotation with box coordinate $\mathbf{b}_{gt}$ and category c_{gt}. Specifically, given the i th pseudo bounding box label $\mathbf{S}^{(i)} = (\mathbf{b}^{(i)}, c^{(i)})$ of a training image $\mathbf{I}$ with location $\mathbf{b}^{(i)}$ and category $c^{(i)}$ predicted by weakly-supervised detector, calculate the Intersection over Union (IoU) between the $\mathbf{S}^{(i)}$ and every ground-truth of $\{\mathbf{S}_{gt}\}$ whose category c_{gt} as same as c. Then we keep the i th pseudo box label in labled set if there is at least one ground-truth box label has IoU with $\mathbf{S}^{(i)}$ higher than 0.5. And other pseudo label will be discard. Note that introducing the instance-level ground-truth annotation is only for illustrating the effect of clean label and it is not allowed in WSOD task.

We also deploy experiments for ideal split method for every WSOD baseline and Table 10 shows the result of ideal data split method. Obviously, comparing with other data split method which we proposed, ideal split further improves performance a lot. For example, for LBBA+W2N, ideal split outperforms 5.8% mAP than two tasks instance-level split. Hence we believe that it is worth to explore how to design a more effective data split method in future work.



B.2 Quantitative analyze of "outer proposals" observation:

We propose the localization adaptation module and the regularization loss based on the early stage learning characteristic of bounding box regression. To prove

this observation and investigate the effectiveness of regularization loss, we conduct an quantitative experiment by plotting IoU curves between proposals and GT/pseudo GT. In each iteration, we randomly sample N (e.g., 30) outer proposals around each pseudo GT with "discriminative part problem", obtain the current bbox regression w/ and w/o regularization loss, and calculate the IoU metric between estimated box and GT/pseudo GT (Fig B.1). Without regularization loss, the outer proposals first regress to the GT at early learning stage, but finally regress to pseudo GT (the left sub-figure). In comparison, with regularization loss, the outer proposals can be regressed consistently towards GT along with the training iterations(the right sub-figure). Thus, the above result can provide an empirical support to our observation as well as the effectiveness of regularization loss.

B.3 Discussion of Sui *et al.*

Sui *et al.*[24]. proposed a novel WSOD framework named SoS, which is first to adopt SSOD in WSOD task. There are several differences between this work and ours.

First, in overall, SoS only adopt SSOD method to enhance the performance of WSOD, while our work formulates the multi-phase weakly supervised object detection problem as a noisy-label object detection problem. Learning with noisy labels has been widely studied for image classification, where the noise is image-level and is on classification labels. In contrast, noisy label remains less investigated for object detection due to that: (i) noise of pseudo label on localization is also inevitable, and (ii) both classification and localization label noise are instance-level instead of image level. Here we present *Location Adaption module* and *Hybrid-Level Dataset Split* for handling these two issues, which are novel for incorporating noisy label learning with WSOD. In addition, we believe that as the noisy-label learning develops, we can absorb more idea from it and design more better performing model for WSOD task.

Second, SoS applies [12] for object detection task directly, while we think more about the characteristic of (weakly-supervised) object detection task in terms of the noise distribution. For example, In *Location Adaption Module*, we focus on the "discriminative part" issue of WSOD and analyze the change of regression of bound boxes during training, which inspires us to propose the regularization loss to make the location performance better. In *Semi-supervised Learning Module*, we propose that both classification and localization label noise are instance-level instead of image level. So, we calculate classification and localization loss respectively for every instance, which can screen more high-quality samples for training.

Third, our W2N outperforms 2.7% mAP and 1.2% mAP than SoS with the same WSOD baseline(CASD) at single scale testing on PASCAL VOC 2007 and 2012 dataset respectively. Note that we don't adopt any training tricks e.g.multi input training strategy and PGF proposed in [24].

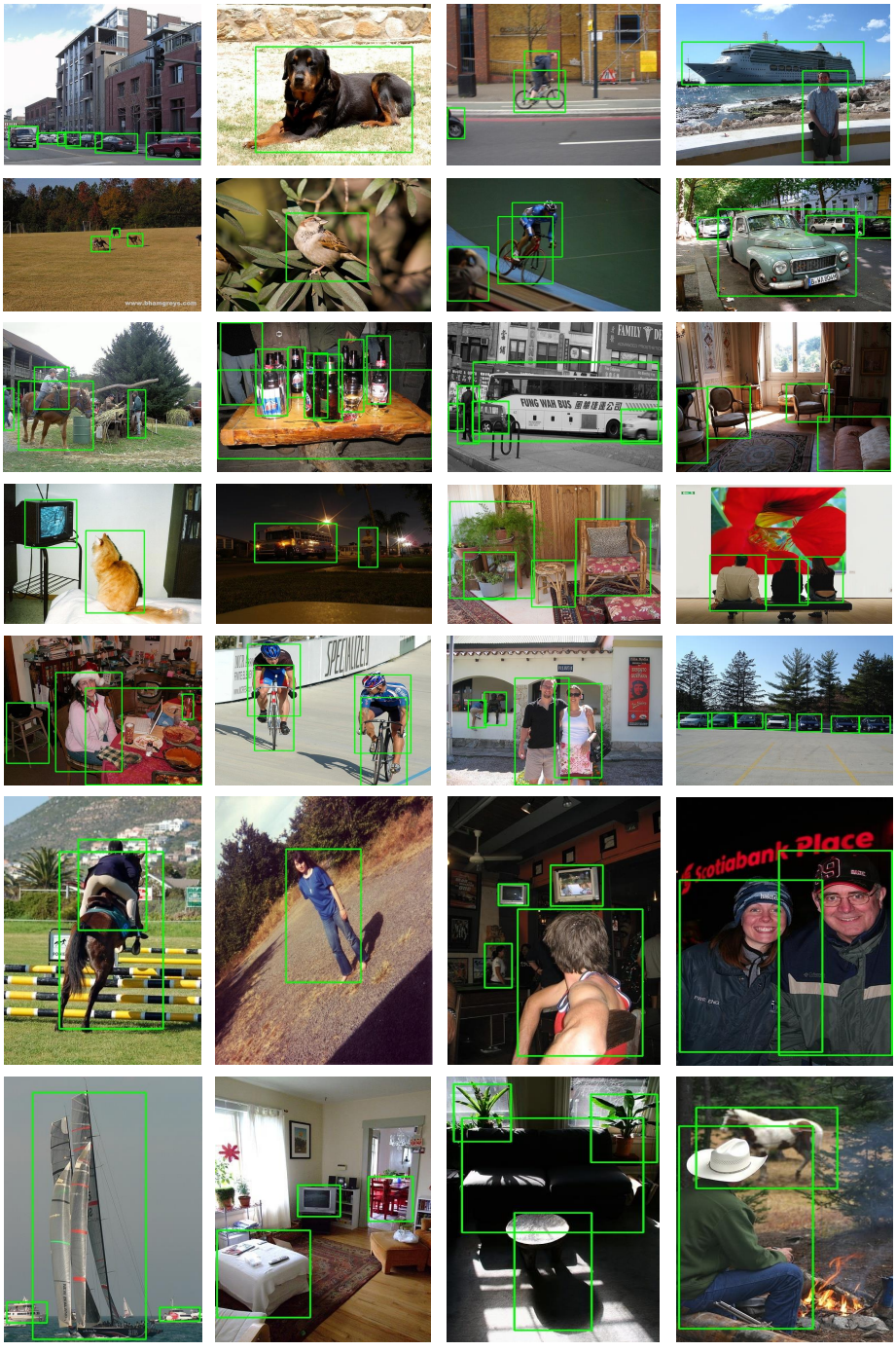


Fig. 2. More visualization results of LBBA+W2N on PASCAL VOC 2007 test set.

References

1. Arun, A., Jawahar, C., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019). https://doi.org/10.1109/cvpr.2019.00966
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. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)
4. Cao, T., Du, L., Zhang, X., Chen, S., Zhang, Y., Wang, Y.F.: Cat: Weakly supervised object detection with category transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3070–3079 (October 2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2876–2885 (2021)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
9. Huang, Z., Zou, Y., Bhagavatula, V., Huang, D.: Comprehensive attention self-distillationfor weakly-supervised object detection. In: NeurIPS (2020)
10. Jia, Q., Wei, S., Ruan, T., Zhao, Y., Zhao, Y.: Gradingnet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. Proceedings of the AAAI Conference on Artificial Intelligence **35**(2), 1682–1690 (May 2021), `https://ojs.aaai.org/index.php/AAAI/article/view/16261`
11. Lee, S., Kwak, S., Cho, M.: Universal bounding box regression and its applications. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. pp. 373–387. Springer International Publishing, Cham (2019)
12. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations (2019)
13. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
14. Li, Y., Zhang, J., Zhang, J., Huang, K.: Mixed supervised object detection with robust objectness transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (02 2018). https://doi.org/10.1109/TPAMI.2018.2810288
15. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
16. 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)

17. Liu, Y., Zhang, Z., Niu, L., Chen, J., Zhang, L.: Mixed Supervised Object Detection by Transferring Mask Prior and Semantic Similarity. arXiv e-prints arXiv:2110.14191 (Oct 2021)
18. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: International Conference on Learning Representations (2020)
19. 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 (NIPS) (2015)
20. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
21. Rochan, M., Wang, Y.: Weakly supervised localization of novel objects using appearance transfer. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4315–4324 (2015). https://doi.org/10.1109/CVPR.2015.7299060
22. Shen, Y., Ji, R., Wang, Y., Chen, Z., Zheng, F., Huang, F., Wu, Y.: Enabling deep residual networks for weakly supervised object detection. In: European Conference on Computer Vision (ECCV) (2020)
23. Shen, Y., Ji, R., Wang, Y., Wu, Y., Cao, L.: Cyclic guidance for weakly supervised joint detection and segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
24. Sui, L., Zhang, C.L., Wu, J.: Salvage of supervision in weakly supervised detection (2021)
25. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence **42**(1), 176–191 (2018)
26. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR (2017)
27. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
28. Yan, G., Liu, B., Guo, N., Ye, X., Wan, F., You, H., Fan, D.: C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9833–9842 (2019). https://doi.org/10.1109/ICCV.2019.00993
29. Yang, K., Li, D., Dou, Y.: Towards precise end-to-end weakly supervised object detection network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8372–8381 (2019)
30. Yin, Y., Deng, J., Zhou, W., Li, H.: Instance mining with class feature banks for weakly supervised object detection. Proceedings of the AAAI Conference on Artificial Intelligence **35**(4), 3190–3198 (May 2021), `https://ojs.aaai.org/index.php/AAAI/article/view/16429`
31. 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/CVF International Conference on Computer Vision (ICCV) (October 2019)
32. Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B.: W2f: A weakly-supervised to fully-supervised framework for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 928–936 (2018)

33. Zhong, Y., Wang, J., Peng, J., Zhang, L.: Boosting weakly supervised object detection with progressive knowledge transfer. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 615–631. Springer International Publishing, Cham (2020)