[Supplementary Material] Object Discovery via Contrastive Learning for Weakly Supervised Object Detection

Jinhwan Seo[1], Wonho Bae[2], Danica J. Sutherland[2,3], Junhyug Noh[4⋆], and Daijin Kim[1⋆]

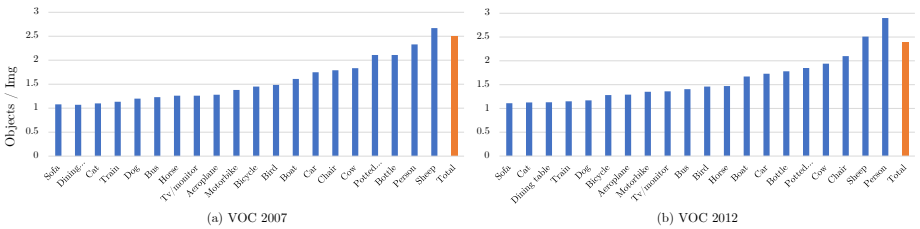
[1] Pohang University of Science and Technology
[2] University of British Columbia
[3] Alberta Machine Intelligence Institute
[4] Lawrence Livermore National Laboratory
tohoaa@gmail.com, whbae@cs.ubc.ca, dsuth@cs.ubc.ca, noh1@llnl.gov, dkim@postech.ac.kr
https://github.com/jinhseo/OD-WSCL

In this supplementary material, we provide further results, both quantitative and qualitative, in the following order.

- Section A reports per-class Average Precision and Correct Localization results on PASCAL VOC datasets.
- Section B compares different proposal generation methods.
- Section C demonstrates the robustness of proposed method using similarity threshold guided by WSCL.
- Section D provides overall pipeline of Object Discovery.
- Section E provides additional qualitative results on PASCAL VOC and MS-COCO datasets.

A Detailed Performance on PASCAL VOC

In Tables 4 and 5, we provide additional performance of per-class average precision (AP) using Selective Search (SS) [13] with VGG16 on VOC07 and VOC12 [4]. Our method achieves the second-highest performance on VOC07 and the highest performance on VOC12. The proposed method successfully addresses the issue of missing objects with high performance for the classes with a large number of objects per image, such as *cow*, *person* and *sheep* in Fig. 7.



(a) VOC 2007 (b) VOC 2012

Fig. 7: Analysis of a number of objects per image on PASCAL VOC datasets.

In Tables 6 and 7, we report the results of per-class Correct Localization (CorLoc) scores using SS with VGG16 on VOC07 and VOC12. CorLoc is an additional evaluation metric commonly reported in WSOD literature to measure localization accuracy, equivalent to precision ($= \frac{\text{true positives}}{\text{true positives+false positives}}$). More specifically, it measures the percentage of correct localization predictions where a prediction is treated as "correct" if the IoU between the prediction and corresponding ground truth is greater than or equal to 0.5. Our method achieves the third-best result in CorLoc on both VOC07 and VOC12. Our slightly worse performance in CorLoc than in mAP is because, as a multiple instance labeling method, our approach captures more proposals than argmax-based methods: this significantly increases recall, but may slightly decrease precision (the only thing measured by CorLoc).

Table 4: Per-class AP results on VOC07

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mAP
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	56.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
OICR[11]	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
C-WSL[5]	62.9	64.8	39.8	28.1	16.4	69.5	68.2	47.0	27.9	55.8	43.7	31.2	43.8	65.0	10.9	26.1	52.7	55.3	60.2	66.6	46.8
WSRPN[12]	60.3	66.2	45.0	19.6	26.6	68.1	68.4	49.4	8.0	56.9	**55.0**	33.6	62.5	68.2	20.6	29.0	49.0	54.1	58.8	58.4	47.9
C-MIL[14]	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	52.9	53.5	57.4	68.9	8.4	24.6	51.8	58.7	66.7	63.5	50.5
C-MIDN[6]	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
WSOD2[15]	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
OIM[10]	55.6	67.0	45.8	27.9	21.1	69.0	68.3	70.5	21.3	60.2	40.3	54.5	56.5	70.1	12.5	25.0	52.9	55.2	65.0	63.7	50.1
SLV[3]	65.6	71.4	49.0	**37.1**	24.6	69.6	70.3	70.6	30.8	63.1	36.0	61.4	65.3	68.4	12.4	**29.9**	52.4	**60.0**	67.6	64.5	53.5
MIST[9]	**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
CASD[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**56.8**
Ours	65.8	**79.5**	**58.1**	23.7	28.6	**71.2**	**75.0**	**71.7**	31.7	69.8	45.2	55.7	57.2	**75.7**	29.6	24.3	61.0	55.3	**71.7**	**72.0**	56.1

Table 5: Per-class AP results on VOC12

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mAP
OICR[11]	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
C-WSL[5]	74.0	67.3	45.6	29.2	26.8	62.5	54.8	21.5	22.6	50.6	24.7	25.6	57.4	71.0	2.4	22.8	44.5	44.2	45.2	**66.9**	43.0
WSRPN[12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.4
C-MIL[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
C-MIDN[6]	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
WSOD2[15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.2
OIM[10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.3
SLV[3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.2
MIST[9]	**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[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	53.6
Ours	73.8	**74.7**	**61.3**	**32.9**	**40.0**	**64.6**	**59.8**	68.1	26.3	**67.5**	23.0	67.1	62.8	**80.6**	**17.3**	**34.1**	**63.4**	**44.4**	**66.2**	64.9	**54.6**

Table 6: Per-class CorLoc results on VOC07

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	CorLoc
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[11]	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
C-WSL[5]	85.8	81.2	64.9	50.5	32.1	84.3	85.9	54.7	43.4	80.1	42.2	42.6	60.5	90.4	13.7	57.5	**82.5**	61.8	74.1	82.4	63.5
WSRPN[12]	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	87.7	**68.4**	52.1	84.4	91.6	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
C-MIL[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
C-MIDN[6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.7
WSOD2[15]	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
OIM[10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.2
SLV[3]	84.6	84.3	73.3	58.5	**49.2**	80.2	87.0	79.4	46.8	83.6	41.8	**79.3**	**88.8**	90.4	19.5	59.7	79.4	**67.7**	**82.9**	**83.2**	**71.0**
MIST[9]	**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
CASD[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.4
Ours	86.3	**87.8**	74.5	47.3	43.9	**85.8**	84.6	**78.2**	**49.1**	83.6	49.4	61.6	74.5	92.4	42.2	46.9	80.4	62.1	**82.9**	82.8	69.8

Table 7: Per-class CorLoc results on VOC12

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	CorLoc
OICR[11]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.1
C-WSL[5]	90.9	81.1	64.9	57.6	50.6	84.9	78.1	29.8	49.7	83.9	50.9	42.6	78.6	87.6	10.4	58.1	85.4	**61.0**	64.7	**86.6**	64.9
WSRPN[12]	85.5	60.8	62.5	36.6	53.8	82.1	80.1	48.2	14.9	87.7	**68.5**	60.7	**85.7**	89.2	**62.9**	**62.1**	87.1	54.0	45.1	70.6	64.9
C-MIL[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
C-MIDN[6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.2
WSOD2[15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9
OIM[10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.1
SLV[3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.2
MIST[9]	**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
CASD[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**72.3**
Ours	88.2	**88.3**	**75.0**	**59.7**	**58.9**	**89.3**	73.2	57.8	53.4	88.0	48.7	**67.5**	78.3	94.0	34.8	61.6	**91.7**	59.4	**70.9**	84.4	71.2

B Comparison of Proposal Generation Method

Current WSOD models rely on pre-computed proposal methods such as Selective Search (SS) [13] and Edge Boxes (EB) [16]. Although the choice of proposal generation methods has a significant impact on localization performance, most previous studies still exploit SS for PASCAL VOC and MCG for MS-COCO datasets. To better understand the effect of using different proposal methods, we compare our algorithm's performance to that of several state-of-the-art algorithms with different proposal methods (SS [13], MCG [1], and COB [8]) on VOC07 (Table 8) and MS-COCO (Table 9) datasets. Note that COB generally captures the groundtruths the best among the three proposal generation methods whereas SS performs the worst.

In general, the better the proposals are, the higher the performance of detection is regardless of model. In Table 8, Ours performs the best with COB and then with MCG (COB: 61.8%, MCG: 58.7%, and SS: 56.1%), which is the same for CASD and MIST. Similarly, COB outperforms MCG with a large margin as observed on MS-COCO datasets as shown in Table 9. We chose to report only the performance of SS and MCG in the main paper because additional boundary information is required to train COB, which violates the definition of image-level supervision. Based on this experiment, we believe MCG should be the default proposal generation method for both PASCAL VOC and MS-COCO datasets unlike the previous convention in WSOD.

Table 8: Per-class AP results with different proposal generation methods on VOC07

Method	Proposal	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mAP
MIST[9]	SS	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
CASD[7]	SS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.8
Ours	SS	65.8	79.5	58.1	23.7	28.6	71.2	75.0	71.7	31.7	69.8	45.2	55.7	57.2	75.7	29.6	24.3	61.0	55.3	71.7	72.0	56.1
MIST[9]	MCG	65.7	78.9	55.5	25.1	31.3	74.5	76.8	67.5	16.1	68.7	50.3	36.0	73.4	76.7	31.7	30.7	61.6	64.5	74.9	70.0	56.5
CASD[7]	MCG	65.1	70.5	55.6	42.8	31.3	72.4	71.7	75.5	16.0	64.1	**60.2**	68.4	71.5	70.7	39.6	27.5	58.3	53.9	63.6	69.2	57.4
Ours	MCG	**69.2**	**81.5**	56.4	28.5	30.5	77.6	79.1	71.6	13.0	70.8	48.8	56.9	74.9	**78.4**	34.9	27.6	61.4	**65.4**	74.4	**73.4**	58.7
MIST[9]	COB	65.1	74.8	57.5	34.0	45.0	77.8	**80.6**	56.1	20.5	**71.2**	50.0	51.9	58.0	78.2	27.2	**32.6**	**62.2**	63.4	72.9	69.8	57.4
CASD[7]	COB	69.1	71.1	**63.2**	**48.5**	40.0	76.4	74.2	**77.1**	17.6	67.4	59.9	**76.1**	74.4	70.4	20.8	30.2	59.4	58.3	67.2	68.1	59.4
Ours	COB	68.6	78.4	62.2	36.6	**49.8**	**79.2**	80.9	77.0	**29.4**	71.0	38.1	62.7	**80.6**	78.0	**40.8**	31.6	61.7	62.8	**75.7**	69.8	**61.8**

Table 9: Performance with different proposal generation methods on MS-COCO

Dataset	Backbone	Method	Proposal	AP	AP^{50}	AP^{75}	AP^s	AP^m	AP^l	AR^1	AR^{10}	AR^{100}	AR^s	AR^m	AR^l
COCO14	VGG16	MIST [9]	MCG	11.4	24.3	9.4	3.6	12.2	17.6	13.5	22.6	23.9	8.5	25.4	38.3
		CASD [7]	MCG	12.8	26.4	-	-	-	-	-	-	-	-	-	-
		Ours	MCG	13.7	27.7	11.9	4.4	14.5	21.2	14.7	24.8	26.9	8.8	27.8	44.0
		Ours	COB	**15.1**	**29.3**	**13.8**	**4.5**	**15.9**	**23.4**	**16.0**	**26.5**	**28.2**	**8.9**	**29.5**	**46.5**
	ResNet50	MIST [9]	MCG	12.6	26.1	10.8	3.7	13.3	19.9	14.8	23.7	24.7	8.4	25.1	41.8
		CASD [7]	MCG	13.9	27.8	-	-	-	-	-	-	-	-	-	-
		Ours	MCG	13.9	29.1	11.8	**4.9**	16.8	22.3	15.5	26.1	28.0	9.0	31.8	46.6
		Ours	COB	**15.4**	**30.4**	**14.0**	4.8	**18.0**	**24.6**	**16.9**	**29.2**	**31.4**	**9.4**	**35.1**	**53.1**
	ResNet101	MIST [9]	MCG	13.0	26.1	10.8	3.7	13.3	19.9	14.8	23.7	24.7	8.4	25.1	41.8
		Ours	MCG	14.4	29.0	12.4	4.8	17.3	23.8	15.8	27.0	30.0	9.2	33.6	51.0
		Ours	COB	**16.2**	**31.6**	**14.8**	**5.0**	**18.7**	**26.4**	**17.5**	**29.6**	**31.9**	**10.0**	**35.4**	**53.5**
COCO17	VGG16	MIST [9]	MCG	12.4	25.8	10.5	3.9	13.8	19.9	14.3	23.3	24.6	9.7	26.6	39.6
		Ours	MCG	13.6	27.4	12.2	4.9	15.5	21.6	14.6	24.8	26.8	9.2	28.7	43.8
		Ours	COB	**15.6**	**29.9**	**14.3**	**5.1**	**17.2**	**25.1**	**16.4**	**27.1**	**28.7**	**9.8**	**30.5**	**47.8**
	ResNet50	Ours	MCG	13.8	27.8	12.1	**5.7**	17.7	23.8	15.1	26.6	29.7	10.1	33.7	50.7
		Ours	COB	**16.0**	**30.5**	**14.9**	5.4	**19.0**	**27.2**	**17.0**	**29.1**	**31.4**	**10.4**	**35.2**	**53.3**
	ResNet101	Ours	MCG	14.4	28.7	12.6	5.4	17.9	25.5	15.4	26.8	29.6	10.0	33.3	50.6
		Ours	COB	**16.5**	**31.6**	**15.2**	**5.7**	**19.6**	**28.2**	**17.4**	**29.7**	**31.9**	**11.3**	**35.5**	**54.2**

C Different Criterion for Object Discovery

In Section 4.3, we claimed that the similarity of two proposals in the embedding space can be large even though they are not similar in classification score. To justify the necessity of using additional similarity scores for the object discovery module, Table 10 compares object discovery based on classification score, with various threshold values, to similarity score. Recall that the "adaptive" threshold we used for similarity score is determined by the average value of similarity between the argmax and its augmented samples: $\tau_c^{n,k} = \frac{1}{|\mathcal{S}_c|}\sum_{i=1}^{|\mathcal{S}_c|} sim(z^n_{\tilde{m}_c^{n,k}}, \mathcal{S}_{c,i})$. For object discovery based on classification score, we not only try fixed thresholds but also adaptive threshold defined as $\tau_c^{n,k} = \frac{1}{|\mathcal{S'}_c|}\sum_{i=1}^{|\mathcal{S'}_c|} \mathcal{S'}_{c,i}$ where $\mathcal{S'}$ is the collection of classification scores that are calculated using the augmented features (same features for $\mathcal{S}$).

In Table 10, the performance of the object discovery based on classification score is significantly worse than similarity score. Moreover, the best-performing threshold $\tau_c^{n,k} = 0.4$ (57.4%) is dramatically better than a similar threshold value $\tau_c^{n,k} = 0.2$. Thus, unlike similarity score (as shown in Section 5.3), performance is also very sensitive to the choice of threshold. Note that we train the model with the same hyperparameters ($\tau_{nms} = 0.1$, $\lambda = 0.03$) for fair comparison.

Table 10: The results of different criteria for object discovery

Criterion	Threshold ($\tau_c^{n,k}$)	mAP
Classification Score	0.2	50.3
	0.3	56.2
	0.4	57.4
	0.5	56.2
	0.6	56.1
	0.7	55.6
	0.8	54.4
	Adaptive	53.2
Similarity Score	Adaptive	**58.7**

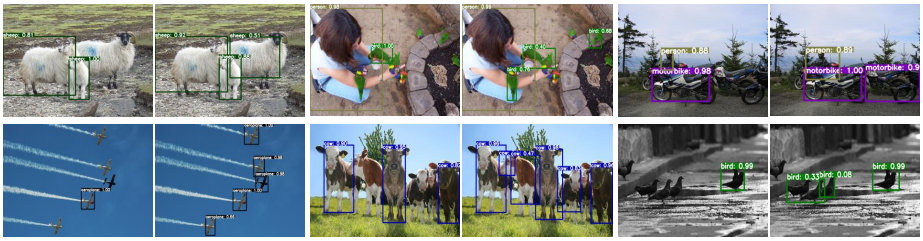


Fig. 8: Comparison of pseudo groundtruths generated by classification score *vs.* similarity score. The left and right images of each pair correspond to pseudo groundtruths based on classification and similarity scores, respectively.

D Pseudo Code: Sampling and Object Discovery

Along with Fig.2 in the main paper, we provide the detailed procedure of sampling steps and object discovery. The main purpose of object discovery is to obtain more pseudo groundtruths in addition to the top-scoring proposals.

Algorithm 1 Sampling steps and object discovery

Network: RoI feature extractor $\eta(\cdot)$, similarity head $\varphi(\cdot)$
Input: Proposal Scores: $x_{c,m}^{n,k}$, Embedding Vectors: z_m^n, Proposals: R^n, Image Labels: Y^n, Proposal Labels: $Y_{c,m}^{n,k}$
Output: Updated S_c, $Y_{c,m}^{n,k}$

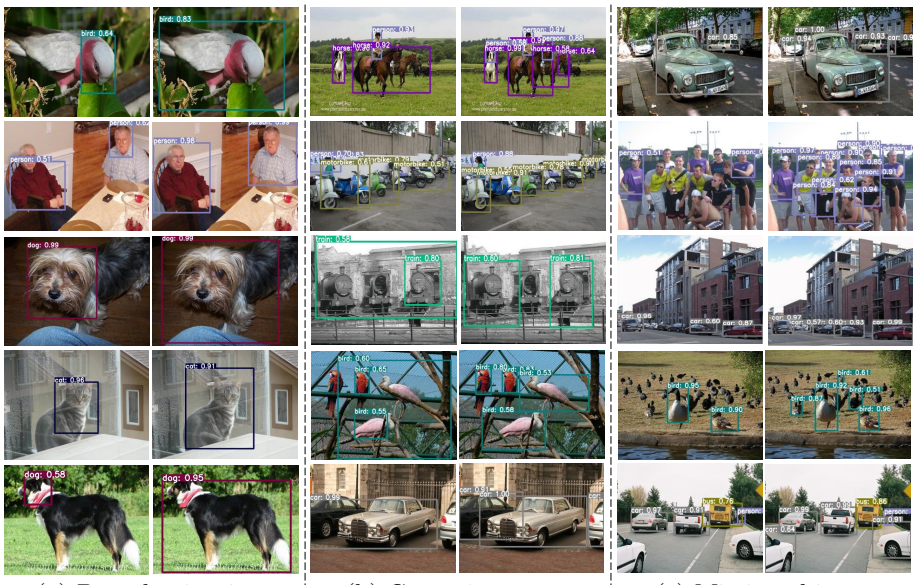
$S_c \leftarrow \emptyset, y_{c,m}^{n,k} = 0, y_{(c+1),m}^{n,k} = 1$
for n = 1 **to** N **do**
  for k = 0 **to** K-1 **do**
    if y_c^n==1 **then**
      $\bar{m}_c^{n,k} = \text{argmax}_m \, x_{c,m}^{n,(k-1)}$
      if $IoU(r_m, r_{\bar{m}_c^{n,k}}) > \tau_{IoU}, \forall m \in M^n$ **then**
        $\mathcal{M}_c^{n,k} \leftarrow m$
  $\mathcal{Z}_{IoU}^{n,c} = \{\varphi(\eta(f_m^n)) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\}$
  $D : D_{i,j} \sim U(0,1) \in \mathbb{R}^{H \times W}$
  $D_{drop} = \begin{cases} 0 & \text{if } D < \tau_{drop} \\ 1 & \text{otherwise} \end{cases}$
  $\mathcal{Z}_{mask}^{n,c} = \{\varphi(\eta(f_m^n \odot D_{drop})) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\}$
  $D_{noise} : D_{i,j} \sim N(0,1) \in \mathbb{R}^{H \times W}$
  $\mathcal{Z}_{noise}^{n,c} = \{\varphi(\eta(f_m^n + f_m^n \odot D_{noise})) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\}$
$\mathcal{S}_c = \bigcup_{n=1}^{N} (\mathcal{Z}_{IoU}^{n,c} \cup \mathcal{Z}_{mask}^{n,c} \cup \mathcal{Z}_{noise}^{n,c})$
for n = 1 **to** N **do**
  for k = 0 **to** K-1 **do**
    if y_c^n==1 **then**
      $\bar{m}_c^{n,k} = \text{argmax}_m \, x_{c,m}^{n,(k-1)}$
      $\tau_c^n = Avg(sim(z_{\bar{m}_c^{n,k}}^n, S_c))$
      if $sim(z_{\bar{m}_c^{n,k}}^n, z_m^n) > \tau_c^n, \forall m \in M^n$ **then**
        $S_c \leftarrow z_m^n$
        if $IoU(r_m, r_{\bar{m}_c^{n,k}}) > 0.5, \forall m \in M^n$ **then**
          $y_{c,m}^{n,k} = 1$

E More Qualitative Results

In Fig. 9, we provide more qualitative results for the three challenges of WSOD on VOC07. Columns on the left and right of each pair correspond to qualitative results from OICR [11] and our model, respectively.

In Fig. 10, we compare prediction results of OICR [11] on the left and Ours on the right. Our model shows much better results for COCO, which contains more instances per image. Although the issue of grouped instances is observed in some cases, our model correctly captures multiple objects and classifies them correctly, despite extremely complex backgrounds.

Fig. 11 shows failure cases of the proposed method. Our model misclassfies background objects that looks like a target class, for example human-like statues or dolls. In addition, the predicted boxes are separated in some cases, even though the object its full extent is captured.



(a) Part domination (b) Group instances (c) Missing objects

Fig. 9: More qualitative results for the three challenges of WSOD on VOC07.

Fig. 10: Qualitative results on COCO14.

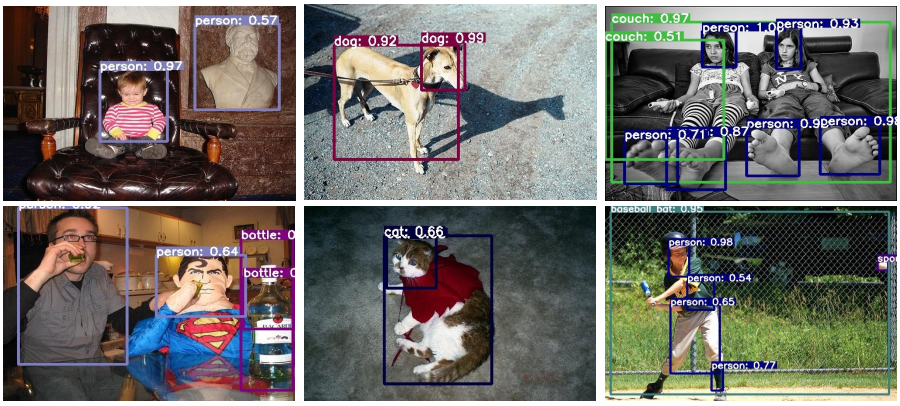


Fig. 11: Failure cases of the proposed method.

References

1. Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 328–335 (2014)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Chen, Z., Fu, Z., Jiang, R., Chen, Y., Hua, X.S.: SLV: Spatial likelihood voting for weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12995–13004 (2020)
4. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (Jan 2015)
5. Gao, M., Li, A., Yu, R., Morariu, V.I., Davis, L.S.: C-wsl: Count-guided weakly supervised localization. In: The European Conference on Computer Vision (ECCV) (September 2018)
6. Gao, Y., 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9834–9843 (2019)
7. Huang, Z., Zou, Y., Bhagavatula, V., Huang, D.: Comprehensive attention self-distillation for weakly-supervised object detection. arXiv preprint arXiv:2010.12023 (2020)
8. Maninis, K.K., Pont-Tuset, J., Arbeláez, P., Van Gool, L.: Convolutional oriented boundaries. In: European conference on computer vision. pp. 580–596. Springer (2016)
9. 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. pp. 10598–10607 (2020)

10. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: European conference on computer vision. pp. 347–365. Springer (2020)
11. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
12. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: The European Conference on Computer Vision (ECCV) (September 2018)
13. 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)
14. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
15. 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. pp. 8292–8300 (2019)
16. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014)