End-to-End Weakly Supervised Object Detection with Sparse Proposal Evolution –Supplementary Material–

Mingxiang Liao[1], Fang Wan[1] *, Yuan Yao[1], Zhenjun Han[1], Jialing Zou[1], Yuze Wang[2], Bailan Feng[2], Peng Yuan[2], and Qixiang Ye[1]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Huawei Noah's Ark Lab
{liaomingxiang20,yaoyuan17}@mails.ucas.ac.cn,
{wanfang,hanzhj,qxye}@ucas.ac.cn, zjl7223009@163.com,
{wangyuze1,fengbailan,yuanpeng126}@huawei.com

1 Details of the Attention Maps

In this section, we first analyze the strategies of generating attention map A. We then visualize the attention maps including A and $\{A_j, j = 1, ..., J\}$ from the multiple head of the class-attention layer.

Method	CorLoc	mAP
Vanilla	57.0	33.3
Top 1	58.8	35.0
Min-max norm	**61.0**	**37.2**

Table 1: Performance of SPE with respect to the strategies of generating the final attention map A on PASCAL VOC 2007 *test* set.

1.1 Generation of Attention Map A

Table 1 shows the performance of SPE with respect to strategies of generating A on PASCAL VOC 2007 *test* set. For attention maps $\{A_j, j = 1, ..., J\}$ from the J heads of the class-attention layer, the final attention map A is generated as

$$A = \sum_{j}^{J} w_j A_j \tag{1}$$

where w_j is weight of A_j, and can be defined as:
(1) Vanilla, $w_j = 1/J$,
(2) Top 1, $w_j = \mathbb{1}(\sigma_j = \max_i \sigma_i)$,
(3) Min-max Norm: $\sigma_j = \frac{\sigma_j - \min_i \sigma_i}{\max_i \sigma_i - \min_i \sigma_i}$.

* Corresponding author.

σ_j is the standard deviation of attention map A_j. It can be seen that "Top 1" outperforms "Vanilla" by 1.8% and 1.7% on CorLoc accuracy and mAP respectively, indicating the attention maps with small standard deviation contain non-negligible noise. "Min-max Norm" further use the attention maps with large standard deviation while suppressing those with small standard deviation, which significantly outperforms the Vanilla method by 4.0% and 3.9% on CorLoc accuracy and mAP respectively.

1.2 Visualization of the Attention Maps

Fig. 1 shows the attention maps $\{A_j, j = 1, ..., J\}$ generated by the multiple heads of the class-attention layer. It can be seen that activation maps $\{A_j, j = 1, ..., J\}$ various among the attention heads, where each of them activates full object(s), object parts or backgrounds. It is observed that the attention maps which activate a large area of backgrounds have small standard deviation, while those actives foreground areas have large standard deviations.

Fig. 2 shows the attention map A generated by three weighting strategies defined in the last subsection. It shows that "Min-Max Norm" can significantly reduce the activation of backgrounds and thus improve the quality of generated seed proposals.

2 Additional Visualization Results

In this section, we present additional visualization results, including the localization results of SPR with/without proposal augmentation. We also give more visualization about seed proposals and matched proposals.

2.1 SPR w(/o) Proposal Augmentation

Fig. 3 shows the matched proposals in SPR with/without seed proposal augmentation. Without seed proposal augmentation, the matched proposals contain localization noise caused by the seed proposals. When introducing seed proposal augmentation, the sparse proposals are able to enjoy the "teacher ensemble" to suppress noise proposals and thereby achieve more accurate localization.

2.2 Seed Proposals and Matched Proposals

Fig. 4 shows the additional visualization of seed proposals generated by SPG and matched proposals learned by SPR. It can be seen that SPG is able to generate sparse yet high-quality seed proposals. When introducing SPR, the matched proposal is able to reduce the error in the seed proposals and achieves preciser object localization. These results again validate the effectiveness of the "teacher-student" learning mechanism of SPE, where the seed proposals and sparse proposals evolve towards true object locations.

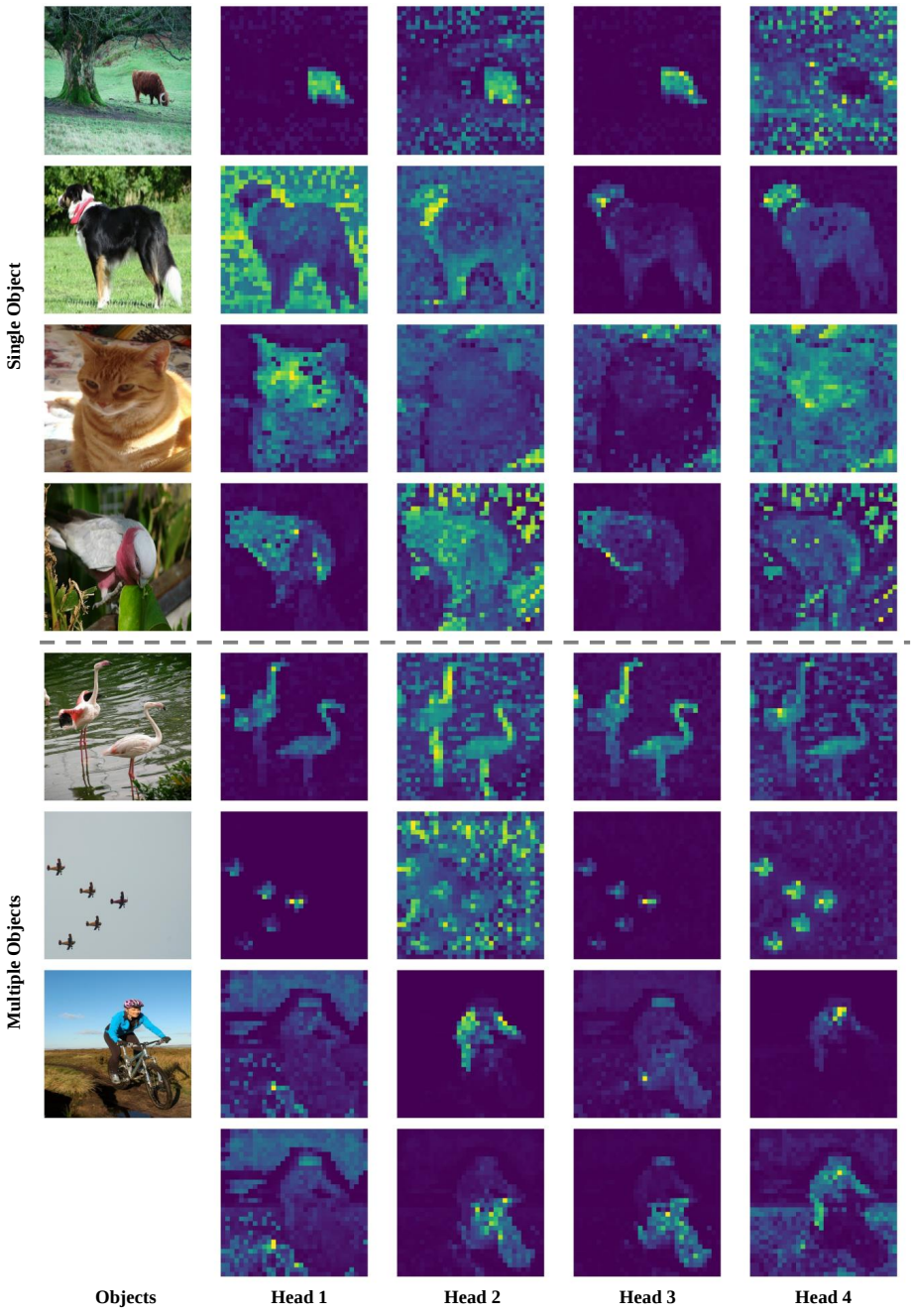

Fig. 1: Attention maps generated by each heads of class-attention layer

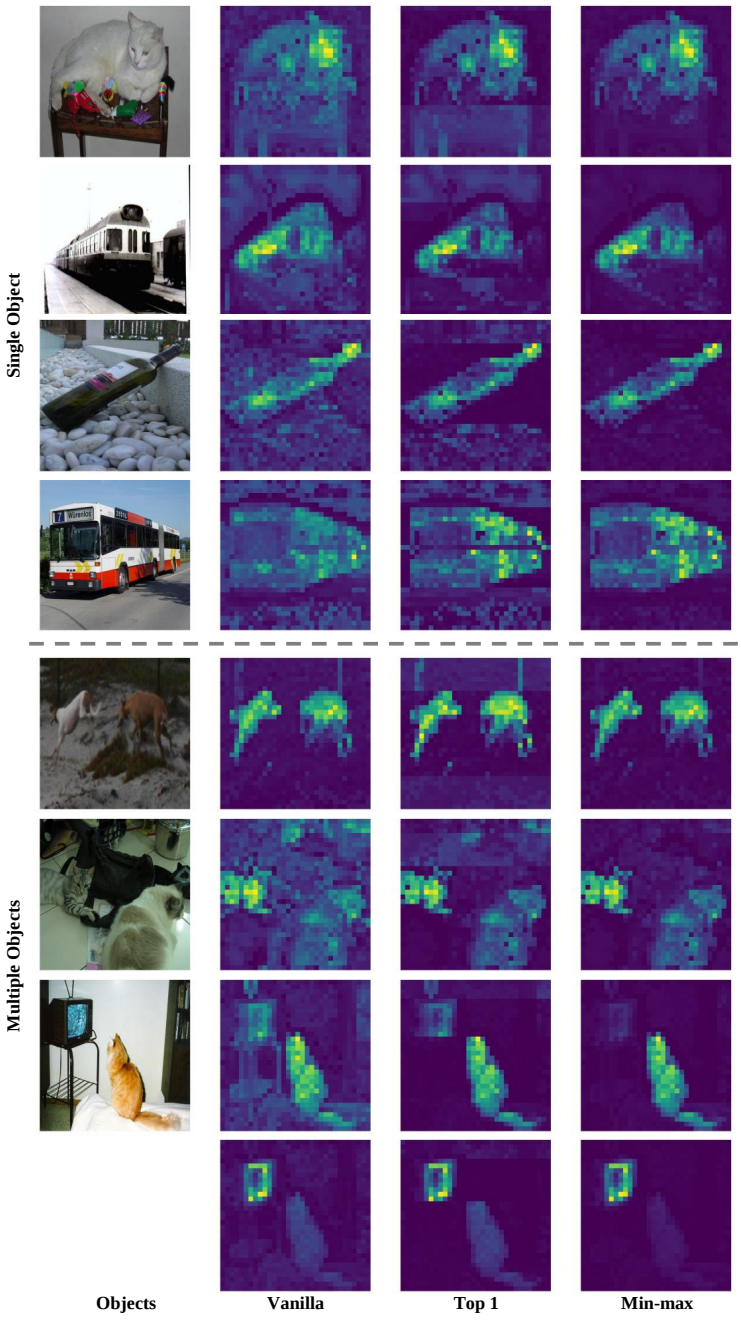

Fig. 2: Activation map generated by different normalization methods

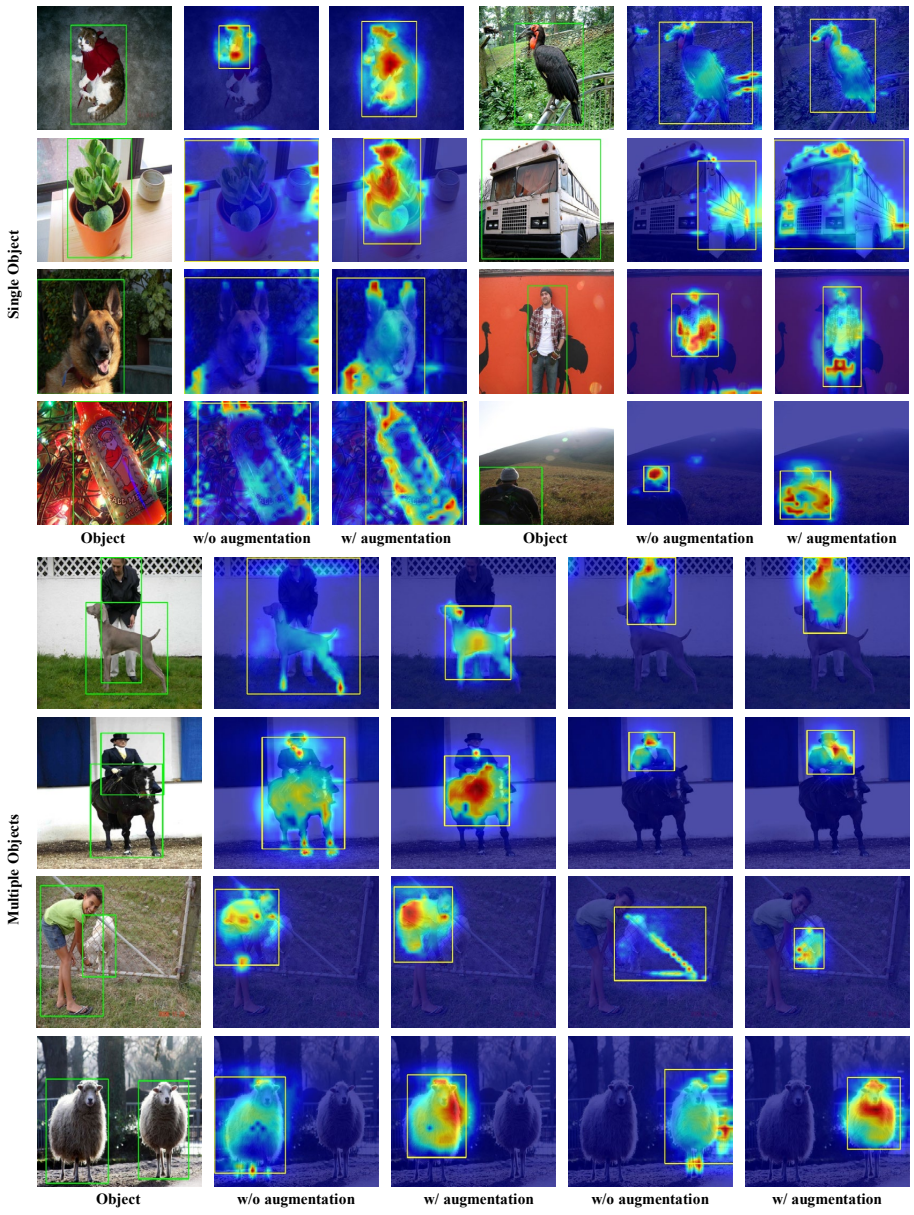

Fig. 3: Matched proposals (yellow bounding boxes) learned by SPR w(/o) seed proposal augmentation. Heatmaps show the cross-attention maps of the matched proposals.

2.3 Proposal Evolution

Fig. 5 shows more results of the evolution of seed proposals and matched proposals and their corresponding attention maps (heatmaps) generated by SPG and SPR. It is also observed that proposals generated by SPG suffer from covering background areas or only part of objects at early training epochs. SPR then refine proposals that match to seed proposals to more accurate object locations, which demonstrates the effectiveness of the SPR module with the proposal augmentation strategy. As training goes on, seed proposals can be gradually refined by and matched with the sparse proposals, and finally evolve to full object extent.

3 Per-Class Performance on PASCAL VOC

Table 2 and Table 3 show the per-class performance of SPE and state-of-the-art methods. For categories of "bird", "cat", "dog", "horse" and "person", which are known to be dominated by the part localization problem in WSOD, SPE achieves creditable performance, *i.e.*, significantly outperforming OICR [10] by 13.4%∼43.1%. It reflects that SPE is able to take advantage of the self-attention mechanism of transformer and activate full object extent for accurate localization.

Network	Method	Set	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGG16	"Enumerate-and-Select" Methods (Two-Stage)																						
	WSDDN [2]	07	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 [10]	07	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
	SLV [3]	07	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
	DC-WSOD [1]	07	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
	TS^2C [12]	07	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
	SDCN [6]	07	59.8	67.1	32.0	34.7	22.8	67.1	63.8	67.9	22.5	48.9	47.8	60.5	51.7	65.2	11.8	20.6	42.1	54.7	60.8	64.3	48.3
	C-MIL [4]	07	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
	PCL [9]	07	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
	ICM [7]	07	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
	WSDDN† [2]	0712	58.0	45.1	34.1	20.6	13.1	73.6	36.9	33.1	14.2	40.9	37.2	35.2	29.5	56.9	11.1	14.1	34.9	48.6	50.0	49.2	36.9
	OICR† [10]	0712	66.7	70.2	40.0	23.1	19.8	68.5	67.1	25.8	25.3	46.8	38.4	25.8	39.8	69.4	4.8	21.3	47.8	44.7	59.6	66.8	43.6
	"Enumerate-and-Select" Methods (End-to-End)																						
	WeakRPN [11]	07	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	57.3	35.2	64.2	68.6	32.8	28.6	50.8	49.5	41.1	30.0	45.3
	UWSOD [8]	07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44.0
CaiT	"Seed-and-Refine" (End-to-End)																						
	TS-CAM [5]	0712	23.9	11.9	10.7	10.8	3.5	24.6	19.2	26.6	0.8	17.0	18.2	21.8	30.1	36.1	6.0	2.0	8.8	6.9	39.0	10.9	16.4
	SPE(Ours)	0712	65.6	64.7	63.5	29.9	12.6	64.4	47.2	82.4	15.0	42.6	49.8	78.7	66.6	57.2	31.4	24.2	45.7	59.0	77.0	43.5	51.0

Table 2: Detection Performance(%) on the PASCAL VOC 2007 *test* set. Comparison of SPE to the state-of-the-arts. "07" in "Set" column denotes the *trainval* set of VOC 2007, and "0712" denotes *trainval* set of VOC 2007 and 2012 datasets. † refers to our implementation.

References

1. Arun, A., Jawahar, C.V., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. In: IEEE CVPR. pp. 9432–9441 (2019) 6, 9
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE CVPR. pp. 2846–2854 (2016) 6, 9

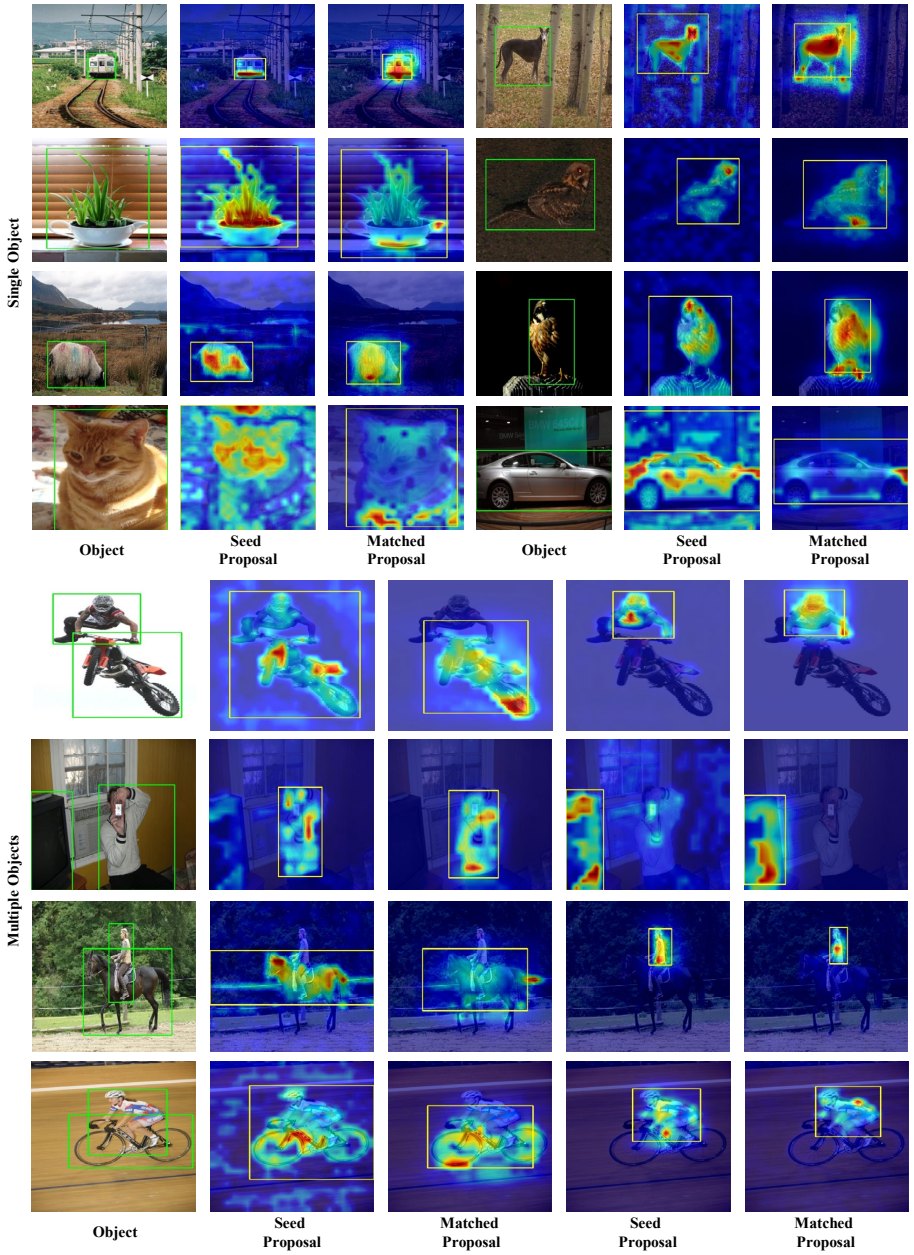

Fig. 4: Seed proposals generated by SPG and matched sparse proposals learned by SPR (yellow bounding boxes). Heatmaps in "seed proposal" column show the semantic-aware attention maps for object classes, while heatmaps in "matched proposal" column show the cross-attention maps of the matched sparse proposals.

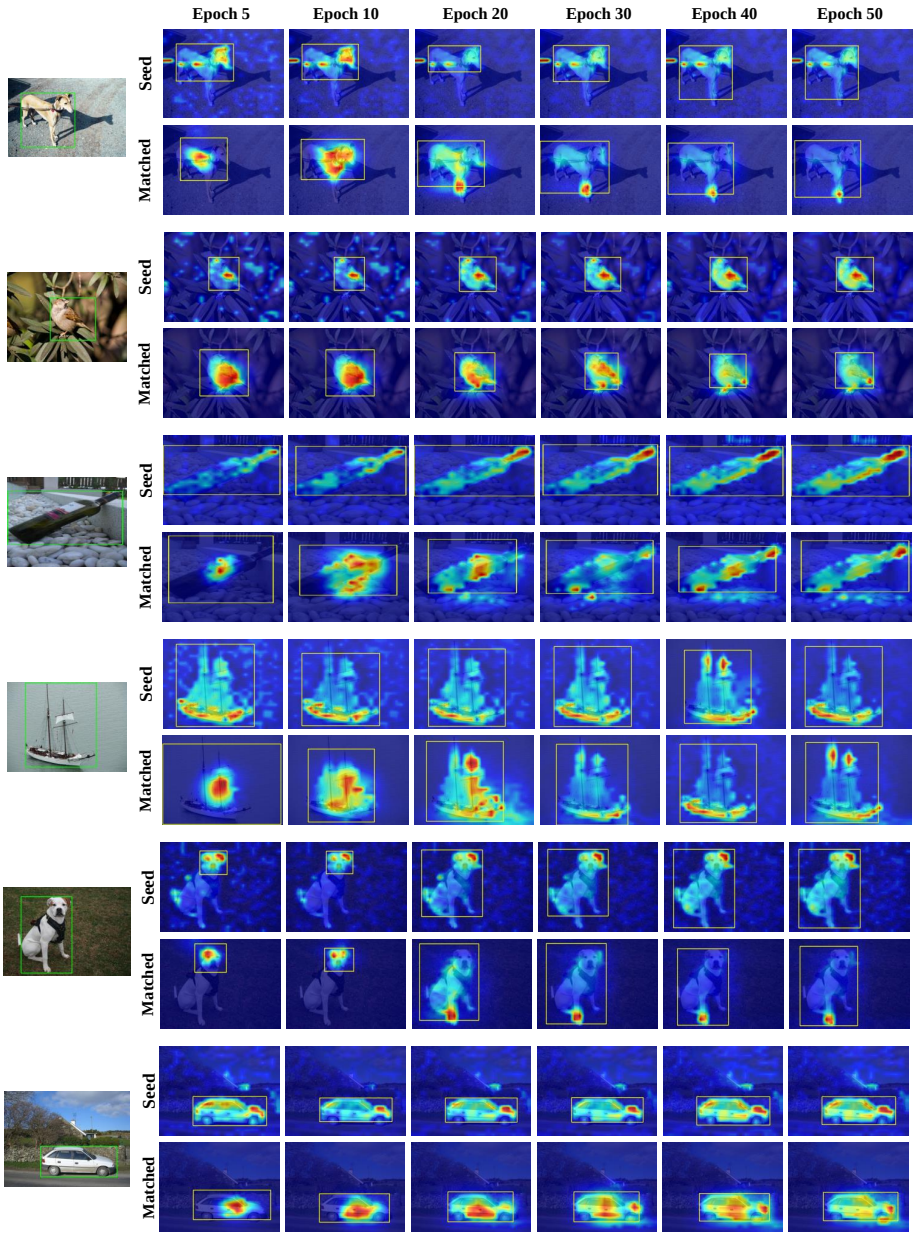

Fig. 5: Evolution of seed proposals and matched sparse proposals (yellow bounding boxes) during training. Heatmaps in the "Seed" column show the semantic-aware attention maps, while heatmaps in "Matched" column show the cross-attention maps of the matched sparse proposals.

Network	Method	Set	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
	"Enumerate-and-Select" Methods (Two-Stage)																						
VGG16	WSDDN [2]	07	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 [10]	07	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
	SLV [3]	07	84.6	84.3	73.3	85.5	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
	DC-WSOD [1]	07	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
	TS^2C [12]	07	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
	SDCN [6]	07	85.8	83.1	56.2	58.5	44.7	80.2	85.0	77.9	29.6	78.8	53.6	74.2	73.1	88.4	18.2	57.5	74.2	60.8	76.1	79.2	66.8
	C-MIL [4]	07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
	PCL [9]	07	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
	ICM [7]	07	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
	WSDDN† [2]	0712	83.8	78.4	56.5	41.0	35.5	74.1	71.9	44.2	45.5	56.2	47.2	37.0	57.5	85.9	11.0	51.3	73.3	40.9	71.9	74.6	56.8
	OICR† [10]	0712	84.9	82.3	66.2	49.6	47.3	83.9	79.9	33.0	48.4	73.9	51.0	39.7	63.6	89.6	12.2	54.7	74.5	44.6	71.0	84.1	61.7
	"Enumerate-and-Select" Methods (End-to-End)																						
	WeakRPN [11]	07	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
	UWSOD [8]	07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.0
	"Seed-and-Refine" (End-to-End)																						
CaiT	TS-CAM [5]	0712	56.6	36.7	43.7	34.9	13.4	54.6	35.4	47.8	11.8	53.2	28.2	45.3	52.0	62.6	26.5	11.8	36.9	32.4	58.7	16.6	37.9
	SPE(Ours)	0712	86.3	80.8	82.0	54.6	32.5	81.4	72.5	88.9	44.0	72.4	65.2	85.9	82.1	79.8	59.2	54.3	77.6	62.5	83.8	62.8	70.4

Table 3: Localization Performance(%) on the PASCAL VOC. Comparison of SPE to the state-of-the-arts. "07" in "Set" column denotes the *trainval* set of VOC 2007, and "0712" denotes *trainval* set of VOC 2007 and 2012 datasets. † refers to our implementation.

3. Chen, Z., Fu, Z., Jiang, R., Chen, Y., Hua, X.: SLV: spatial likelihood voting for weakly supervised object detection. In: IEEE CVPR. pp. 12992–13001 (2020) 6, 9
4. Fang, W., Chang, L., Wei, K., Xiangyang, J., Jianbin, J., Qixiang, Y.: Cmil: Continuation multiple instance learning for weakly supervised object detection. In: IEEE CVPR (2019) 6, 9
5. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: TS-CAM: token semantic coupled attention map for weakly supervised object localization. CoRR **abs/2103.14862** (2021) 6, 9
6. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: IEEE ICCV (2019) 6, 9
7. Ren, Z., Yu, Z., Yang, X., Liu, M., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: IEEE CVPR. pp. 10595–10604 (2020) 6, 9
8. Shen, Y., Ji, R., Chen, Z., Wu, Y., Huang, F.: UWSOD: toward fully-supervised-level capacity weakly supervised object detection. In: NeurIPS (2020) 6, 9
9. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.L.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE TPAMI **42**(1), 176 – 191 (2020) 6, 9
10. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE CVPR. pp. 3059–3067 (2017) 6, 9
11. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: ECCV. pp. 352–368 (2018) 6, 9
12. Wei, Y., Shen, Z., Cheng, B., Shi, H., Xiong, J., Feng, J., Huang, T.: Ts2c:tight box mining with surrounding segmentation context for weakly supervised object detection. In: ECCV. pp. 434–450 (2018) 6, 9