

Supplementary Materials: WeakSAM

Anonymous Authors

1 MORE DETAILS OF WEAKSAM PROPOSALS

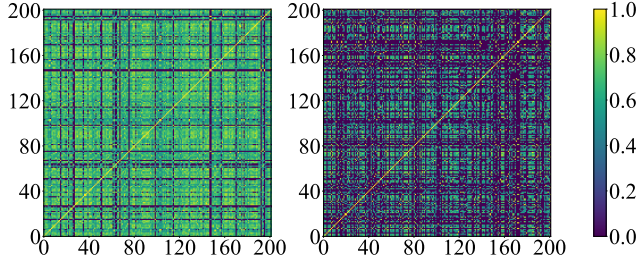


Figure 1: The cosine similarity among the features of proposals, i.e., Left for Selective Search proposals and Right for WeakSAM proposals. For a single image from PASCAL VOC 2007, we randomly sampled 200 proposal features to calculate their similarity.

We further analyze the proposal similarity in the different weakly-supervised object detection (WSOD) proposals, as shown in Fig. 1. We randomly sample 200 proposal features each from Selective Search [15] proposals and WeakSAM proposals, and then compute their cosine similarity, respectively. Please note that all the features are output by the RoI pooling layer. It can be seen that the features from Selective Search tend to have higher similarity with other ones. In contrast, the features from WeakSAM proposals show lower similarity, which usually means it has less overlap and redundancy.

2 MORE DETAILS OF ROI DROP REGULARIZATION

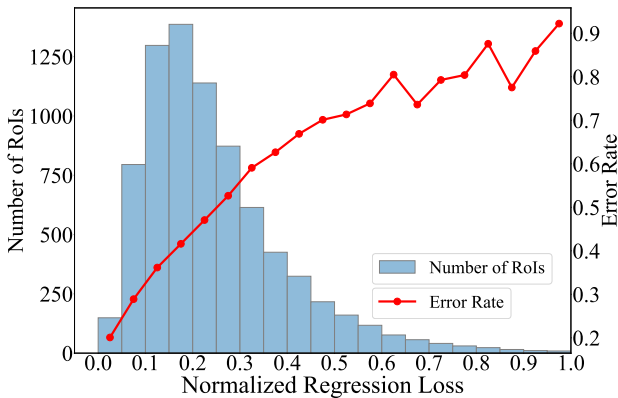


Figure 2: The relationship between the normalized regression loss, the corresponding number of RoIs, and the corresponding error rate. The results are obtained from training the Faster-RCNN [10] using PASCAL VOC 2007 pseudo ground truth (PGT) in the preliminary training stage.

We further present the relationship between the normalized regression loss, the corresponding number of RoIs, and the corresponding error rate in Fig. 2. It shows that the regression losses of RoIs have different number distributions compared to the classification losses. However, they exhibit similar error rate curves. This observation further demonstrates the necessity of RoI drop regularization with a regression threshold τ_{reg} .

3 MORE DETAILS OF QUERY DROP REGULARIZATION

Because DINO [21] employs Focal loss [7] as the classification loss, queries associated with background classes tend to have higher predicted probabilities and lower losses. This results in the inadvertent omission of most foreground category queries when directly dropping queries. To mitigate this issue, our first step involves normalizing the unweighted Focal loss, which is essentially the binary cross-entropy loss, for both foreground and background queries within each training batch. Normalizing at the batch level broadens the sampling scope from a single image to the size of the batch. In the second step, queries are dropped based on their loss ranking post-normalization. This approach avoids making the model converge slowly due to the dropping of the most foreground queries.

3.1 Ablation Studies for RoI Drop Regularization

Table 1: Ablation study for the regression threshold and classification threshold in RoI drop regularization in terms of AP₅₀ on the PASCAL VOC 2007 test set.

(a) Ablation study for the regression threshold.

τ_{reg}	0.8	1.0	1.2
AP ₅₀	71.0	71.8	71.3

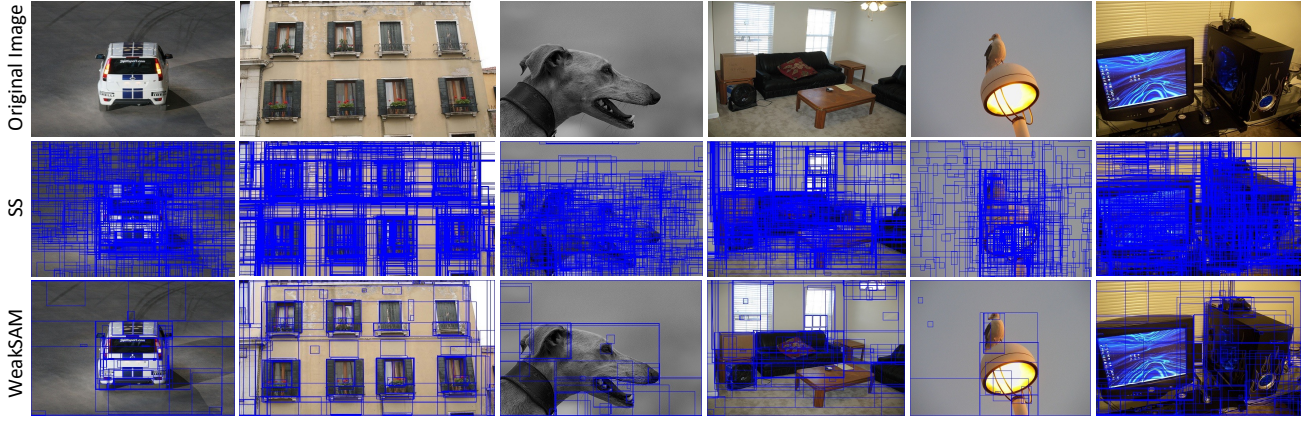
(b) Ablation study for the classification threshold.

τ_{cls}	3.0	4.0	5.0
AP ₅₀	71.2	71.8	71.1

To further analyze the impact of the regression threshold and classification threshold in RoI drop regularization, we conduct experiments as shown in Table 1. It is observed that the best regression threshold τ_{reg} and classification threshold τ_{cls} for RoI drop regularization is 1.0 and 4.0, respectively.

3.2 Ablation Studies for Query Drop Regularization

To analyze the impact of dropping queries corresponding to foreground categories Query_{things} and background categories Query_{bkg}, we conduct ablations as shown in Table 2. Experimental results indicate that dropping only the foreground queries (Query_{things}) leads to the best performance, whereas dropping both types of queries results in a slight decrease in performance. We maintain

Figure 3: Visualization of the proposals boxes on the PASCAL VOC 2007 *trainval* set.Figure 4: Visualization of the pseudo ground truth boxes and pseudo instance labels on the PASCAL VOC 2012 *trainaug* set.Table 2: Ablation studies for query drop regularization on the PASCAL VOC 2007 *test* set.

Baseline	Query _{things}	Query _{bkg}	AP ₅₀
✓			72.8
	✓		73.4+0.6
	✓	✓	73.3+0.5

the viewpoint that dropping more background queries may also lead to slower convergence. Consequently, we choose to drop only Query_{things} to achieve better performance.

We further analyze the impact of classification threshold τ in query drop regularization, as shown in Table 3. Quantitative results demonstrate that 90 is the best percentile classification threshold.

4 ADDITIONAL QUANTITATIVE RESULTS

We present the comparison on PASCAL VOC 2007 *trainval* set in terms of CorLoc, as shown in Table 4. It can be seen that the WeakSAM achieves the 13.9% and 14.1% CorLoc improvements on OICR and MIST, respectively. The WeakSAM (OICR) outperforms

Table 3: Ablation study for the classification threshold in query drop regularization in terms of AP₅₀ on the PASCAL VOC 2007 *test* set.

τ (%)	100	90	80
AP ₅₀	72.8	73.4	71.8

the WSOD-CBL [19] by 11.1% CorLoc. The results demonstrate the significant performance improvement brought by our WeakSAM.

5 ADDITIONAL ABLATION STUDIES

5.1 Analysis on Classification Methods

To further analyze the impact of methods that generate classification clues, we replaced WeakTr in WeakSAM with MCTformer and CLIP-ES. As indicated in Table 5, WeakSAM (MCTformer) achieves a 1.6% higher Recall (IoU=90) than WeakSAM (WeakTr). Furthermore, WeakSAM (CLIP-ES) records increases of 0.6% and 2.4% in Recall over WeakSAM (WeakTr) at IoU thresholds of 75 and 90, respectively. These results demonstrate the versatility of the WeakSAM proposal-generating method across different classification methods. Please note that all classification methods employed in

Table 4: Comparison on PASCAL VOC 2007 trainval set in terms of CorLoc(%) with multi-scale testing.

Methods	Sup.	Proposal	CorLoc
WSDDN [1]	\mathcal{I}	EB [23]	53.5
Yang et al. [18]		SS [15]	68.0
C-MIL [16]		SS	65.0
C-MIDN [3]		SS	53.5
WSOD2 [20]		SS	69.5
CASD [5]		SS	70.4
OD-WSCL [12]		SS	69.8
WSOD-CBL [19]		SS	71.8
WSOVOD [6]	\mathcal{I}	LO-WSRPN+SAM	77.2
WSOVOD ‡		LO-WSRPN+SAM	80.1
OICR [14]	\mathcal{I}	SS	60.6
WeakSAM (OICR)		WeakSAM	74.5+13.9
MIST [11]	\mathcal{I}	SS	68.8
WeakSAM (MIST)		WeakSAM	82.9+14.1

Table 5: Ablation studies for classification methods that generate WeakSAM queries on PASCAL VOC 2007 trainval set. We evaluate the average number of proposals and Recall. We present the results of Selective Search [15] at the first row as a baseline.

CLS Methods	Num.	Recall		
		IoU=0.50	IoU=0.75	IoU=0.90
None	2001	92.6	57.7	19.2
WeakTr [22]	213	95.6	75.0	42.1
MCTformer [17]	173	93.2	74.8	43.7
CLIP-ES [8]	205	93.8	75.6	44.5

this study are CAM networks from weakly-supervised semantic segmentation (WSSS) methods. Since these networks are typically well-tuned on specific datasets, such as PASCAL VOC 2012 and COCO 2014, they are adept at providing rich classification clues.

6 ADDITIONAL VISUALIZATION RESULTS

Fig.3 compares the Selective Search proposals with those generated by WeakSAM. The WeakSAM proposals exhibit less redundancy than Selective Search proposals. Fig.4 contrasts the Top-1 PGT with adaptive PGT, demonstrating that adaptive PGT generation captures a greater number of target objects, which might be missed by the Top-1 approach. Additionally, adaptive PGT can be seamlessly integrated to generate pseudo instance labels.

7 MORE IMPLEMENTATION DETAILS

For Algorithm. 1 in paper, we set the kernel size k to 128 and activation threshold τ to 0.9 following default parameters from WeakTr [22]. And for Algorithm. 2, we follow the default manners similar to SoS-WSOD [13], in which score threshold τ_s is set to 0.3, and overlap threshold τ_o is set to 0.85.

For Faster R-CNN [10] retraining, we adopt the same training strategy and hyper-parameters as the fully-supervised ones. For

DINO [21] retraining, we use a learning rate of $9e-5$ with the AdamW [9] optimizer, and a max epoch of 14. Moreover, we apply multi-scale augmentation and horizontal flips in both training and testing.

For implementations of Mask R-CNN [4] and Mask2Former [2], we follow their default hyper-parameters.

REFERENCES

- [1] Hakan Bilen and Andrea Vedaldi. 2016. Weakly supervised deep detection networks. In *CVPR*. 2846–2854.
- [2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*. 1290–1299.
- [3] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. 2019. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *ICCV*. 9834–9843.
- [4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*. 2961–2969.
- [5] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. 2020. Comprehensive attention self-distillation for weakly-supervised object detection. *NeurIPS* 33 (2020), 16797–16807.
- [6] Jianghang Lin, Yunhang Shen, Bingquan Wang, Shaohui Lin, Ke Li, and Liujuan Cao. 2024. Weakly Supervised Open-Vocabulary Object Detection. In *AAAI*.
- [7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*. 2980–2988.
- [8] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*. 15305–15314.
- [9] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS* 28 (2015).
- [11] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*. 10598–10607.
- [12] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. 2022. Object discovery via contrastive learning for weakly supervised object detection. In *ECCV*. Springer, 312–329.
- [13] Lin Sui, Chen-Lin Zhang, and Jianxin Wu. 2022. Salvage of supervision in weakly supervised object detection. In *CVPR*. 14227–14236.
- [14] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple instance detection network with online instance classifier refinement. In *CVPR*. 2843–2851.
- [15] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. 2013. Selective Search for Object Recognition. *IJCV* 104 (2013), 154–171.
- [16] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*. 2199–2208.
- [17] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*. 4310–4319.
- [18] Ke Yang, Dongsheng Li, and Yong Dou. 2019. Towards precise end-to-end weakly supervised object detection network. In *ICCV*. 8372–8381.
- [19] Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. 2023. Cyclic-Bootstrap Labeling for Weakly Supervised Object Detection. In *ICCV*. 7008–7018.
- [20] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. 2019. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *ICCV*. 8292–8300.
- [21] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv:2203.03605* [cs.CV].
- [22] Lianghui Zhu, Yingyue Li, Jieming Fang, Yan Liu, Hao Xin, Wenyu Liu, and Xinggang Wang. 2023. WeakTr: Exploring Plain Vision Transformer for Weakly-supervised Semantic Segmentation. *arXiv preprint arXiv:2304.01184* (2023).
- [23] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*. Springer, 391–405.