

# Supplementary Materials: CBNet: Cooperation-Based Weakly Supervised Polyp Detection

\*\*\*\*\*

## 1 RELATED WORK

### 1.1 Region Proposals Generation

Among various studies Selective search [5] and Edge Boxes [3] are the standard methods to create candidate proposals. However, later studies believe that most of them are negative cases, which not only affect the detection effect but also reduce the prediction speed [10]. With that in mind, they designed the region proposal network (RPN) [7]. Indeed, it shows a significant speed improvement compared to traditional methods. Unfortunately, this strategy calls for a large number of priori parameters such as scales and ratios. Even worse, the effectiveness of these parameters heavily controls the quality of the proposals and thereby impacts the detection results. We statistically correlated information from two publicly available along with a private polyp dataset.

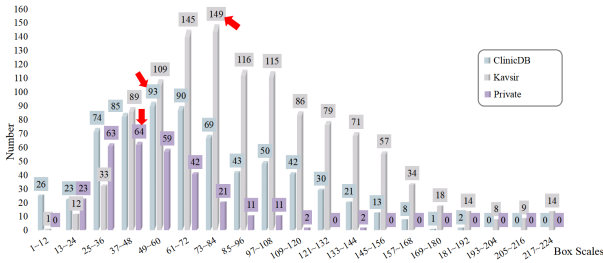


Figure 1: The number of ground truth box scales for different datasets (CVC-ClinicDB, Kvasir, Private). The red arrow indicates the most frequent scale in the dataset and its number of occurrences.

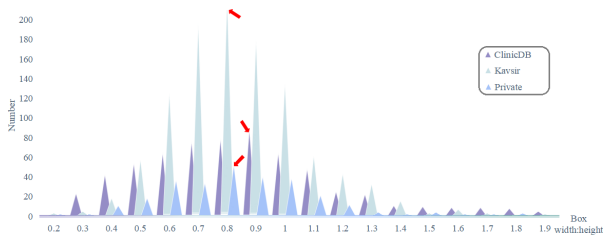


Figure 2: The number of ground truth box aspect ratio (width: height) for different datasets (CVC-ClinicDB, Kvasir, Private). The red arrow indicates the most frequent ratio in the dataset and its number of occurrences.

As Figure 1 & 2 shows the range of polyp scales and box ratios in different datasets is extremely variable. These variations in differentiation pose a severe challenge to parameter setting, as setting only the high-frequency parameter will cause a lower recall rate resulting in missed detections, while resulting in a large number of

redundant candidate proposals and more complex calculations if all the parameters are set. In addition, these numbers are constantly

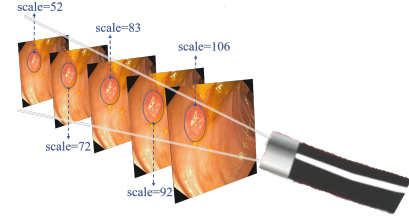


Figure 3: An example of scale change for the same polyp at different distances with colonoscopy.

dynamic because they are closely linked to the distance between the colonoscope and polyps, as shown in Figure 3. Furthermore, to ensure the high performance of RPN, instance-level annotations are required to train the network, which not only deviates from the requirement of weak supervision (WS) but also fails to deliver on weakly supervised polyp detection (WSPD). Considering these, we argue that rather than RPN probably traditional methods are more adequate for WSPD.

## 2 EXPERIMENTS

### 2.1 Quantitative Results

**Main results:** To further demonstrate the average precision (AP) performance of different comparison methods in each category. We show the class AP of all comparison methods on CVC-ClinicDB, Kvasir, and internal dataset, as shown in Table 1, Table 2, Table 3. Among weakly supervised methods, our method outperforms others (except for the AP of pedicle polyps on CVC-ClinicDB when the IOU is set to 10%) under different IOU thresholds. In particular, our method performs much better than the state-of-the-art (both weakly sup. and fully sup.) on flat polyps, as our approach has stronger classification ability and fusion feature maps, which can maintain localization ability, though in most cases instances of these categories are extremely obvious or camouflaged. Besides, the performance of our weakly supervised method is even comparable with the fully supervised methods in some aspects (e.g. the AP of edge on Kvasir and private with 50% iou threshold), illustrating the effectiveness of the proposed CBNet.

### 2.2 Ablation Study

**Impact of CBRPN:** Weakly supervised networks lack the guidance of the instance boundary box and rely only on the boundary box's suggestion generator, which often results in more negative examples than positive ones. This makes it difficult for the network to learn positive features, which affects the accuracy of polyp detection. Before training the network, pre-filtering is necessary to address this imbalance. To evaluate the effectiveness of CBRPN, we

Methods	Supervision	IOU@10			IOU@30			IOU@50		
		Flat	Pedicle	Edge	Flat	Pedicle	Edge	Flat	Pedicle	Edge
Faster Rcnm [7]	Fully Sup.	–	0.78	0.46	–	0.78	0.46	–	0.76	0.46
Yolo [6]	Fully Sup.	–	0.82	0.39	–	0.82	0.38	–	0.82	0.35
DiffusionDet50 [2]	Fully Sup.	–	0.98	0.92	–	0.98	0.92	–	0.96	0.82
DiffusionDet500 [2]	Fully Sup.	–	0.94	0.91	–	0.94	0.91	–	0.94	0.82
WSDDN [4]	Weakly Sup.	0.14	0.23	0.13	0.07	0.16	0.05	0.06	0.11	0.03
OICR [9]	Weakly Sup.	–	0.05	0.08	–	0.01	–	–	–	–
WSOD2 [11]	Weakly Sup.	0.27	0.52	0.34	0.05	0.39	0.34	0.03	0.39	0.34
Grad-CAM [8]	Weakly Sup.	0.32	0.74	0.20	0.04	0.35	–	–	0.12	–
Grad-CAM++ [1]	Weakly Sup.	0.27	0.73	0.46	0.04	0.30	0.05	–	0.11	0.02
CBNet(SAM&SSW)	Weakly Sup.	0.86	0.69	0.76	0.82	0.63	0.43	0.77	0.43	0.36
CBNet(SAM(filter)&SSW)	Weakly Sup.	0.73	0.47	0.38	0.73	0.45	0.38	0.64	0.45	0.38

**Table 1: Average precision for different methods on CVC-ClinicDB test set. The top two results of weakly supervised are marked in red and blue, respectively.**

Methods	Supervision	IOU@10			IOU@30			IOU@50		
		Flat	Pedicle	Edge	Flat	Pedicle	Edge	Flat	Pedicle	Edge
Faster Rcnm [7]	Fully Sup.	–	0.80	–	–	0.79	–	–	0.77	–
Yolo [6]	Fully Sup.	–	0.76	0.07	–	0.75	0.07	–	0.71	0.07
DiffusionDet50 [2]	Fully Sup.	–	0.74	0.12	–	0.72	0.12	–	0.70	0.09
DiffusionDet500 [2]	Fully Sup.	–	0.71	0.11	–	0.70	0.11	–	0.67	0.04
WSDDN [4]	Weakly Sup.	0.05	0.17	0.01	0.03	0.14	–	0.03	0.11	–
OICR [9]	Weakly Sup.	–	–	–	–	–	–	–	–	–
WSOD2 [11]	Weakly Sup.	–	0.05	–	–	0.05	–	–	0.05	–
Grad-CAM [8]	Weakly Sup.	0.01	0.37	–	–	0.11	–	–	0.02	–
Grad-CAM++ [1]	Weakly Sup.	–	0.37	–	–	0.11	–	–	0.03	–
CBNet(SAM&SSW)	Weakly Sup.	0.10	0.5	–	0.09	0.39	–	0.07	0.26	–
CBNet(SAM(filter)&SSW)	Weakly Sup.	0.16	0.42	0.08	0.16	0.31	0.08	0.14	0.18	0.08

**Table 2: Average precision for different methods on Kvasir test set. The top two results of weakly supervised are marked in red and blue, respectively.**

Methods	Supervision	IOU@10			IOU@30			IOU@50		
		Flat	Pedicle	Edge	Flat	Pedicle	Edge	Flat	Pedicle	Edge
Faster Rcnm [7]	Fully Sup.	–	0.83	–	–	0.83	–	–	0.83	–
Yolo [6]	Fully Sup.	–	0.74	0.17	–	0.74	0.15	–	0.74	0.12
DiffusionDet50 [2]	Fully Sup.	–	0.80	0.33	–	0.80	0.33	–	0.79	0.32
DiffusionDet500 [2]	Fully Sup.	–	0.84	0.36	–	0.84	0.34	–	0.81	0.21
WSDDN [4]	Weakly Sup.	0.08	0.30	0.06	0.05	0.27	0.03	0.03	0.26	0.03
OICR [9]	Weakly Sup.	0.04	–	–	0.04	–	–	0.04	–	–
WSOD2 [11]	Weakly Sup.	–	0.03	–	–	0.03	–	–	0.03	–
Grad-CAM [8]	Weakly Sup.	–	0.69	0.03	–	0.40	–	–	0.31	–
Grad-CAM++ [1]	Weakly Sup.	–	0.62	0.03	–	0.39	–	–	0.30	–
CBNet(SAM&SSW)	Weakly Sup.	0.54	0.6	0.4	0.47	0.6	0.4	0.47	0.55	0.4
CBNet(SAM(filter)&SSW)	Weakly Sup.	0.53	0.58	0.4	0.45	0.58	0.4	0.44	0.58	0.4

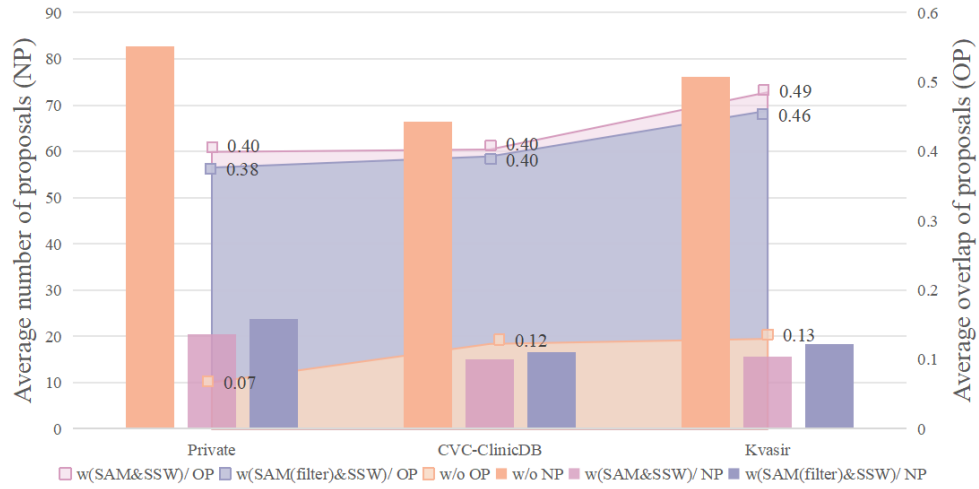
**Table 3: Average precision for different methods on Private test set. The top two results of weakly supervised are marked in red and blue, respectively.**

compared the average number of proposals (NP) with or without CRM, as well as their average overlap (OP, intersection over union) with ground truth. As shown in Figure 4, when we add ARFM, the number of generated proposals decreased from 83 to 21 or 24 on the private dataset, while the overlap increased by  $\delta + 4.71$  or  $\delta + 4.43$ . In addition, we also compared CVC-ClinicDB and Kvasir, NP decreased from 66 to 15 or 17 (CVC-ClinicDB), from 76 to 16 or 18 (Kvasir), while OP increased by  $\delta + 2.33$  and  $\delta + 2.77$  or  $\delta + 2.54$ , respectively. In addition, we also find that the strategy that directly uses SAM pseudo-mask to filter SSW is better than SAM self-filtering first. For example, on the Kvasir dataset, NP of SAM & SSW is 16 v.s. SAM (filter) & SSW is 18, while OP of SAM & SSW is 0.49 v.s. SAM

(filter) & SSW is 0.46. Therefore, to obtain better performance, we believe that CBRPN should choose the strategy of direct filtering.

## REFERENCES

- [1] Chattopadhyay Aditya, Sarkar Anirban, Howlader Prantik, and Balasubramanian Vineeth N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *IEEE Winter Conference on Applications of Computer Vision*. 839–847.
- [2] Shoufa Chen, Pei Sun, Yibing Song, and Ping Luo. 2022. DiffusionDet: Diffusion Model for Object Detection. *arXiv preprint arXiv:2211.09788* (2022).
- [3] Zitnick C.L and Doll ’ar P. 2014. Edge boxes: Locating object proposals from edges. In *Proceedings of the European conference on computer vision (ECCV)* (2014), 391–405.



**Figure 4: Visualization qualitative results of ablation study in CBRPN.**

- [4] Bilen Hakan and Vedaldi Andrea. 2016. Weakly Supervised Deep Detection Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2846–2854.
- [5] Uijlings J.R, van de Sande K.E, Gevers T., and Smeulders A.W. 2013. Selective search for object recognition. *International Joint Conference on Artificial Intelligence, - IJCAI* 104, 2 (2013), 154–171.
- [6] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767* (2018).
- [7] Ren S., He K., Girshick R., and Sun J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, 6 (2017), 1137–1149.
- [8] Ramprasaath R. Selvaraju, Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision*. 618–626.
- [9] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3059–3067.
- [10] Chen Z Y, Wang Z D, and Gong C. 2023. Image-level labeled weakly supervised object detection: a survey. *Journal of Image and Graphics* 28, 9 (2023), 2644–2660.
- [11] 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 *IEEE International Conference on Computer Vision*. 8291–8299.