Rebuttal

Thanks for all valuable reviews and our answers as follows:

Reviewer #1:

Q1: What does hetero-centric segmentation network, tandem feeding and L_{BCE} mean?

A1: This network means that we train on A hospital data and test on other hospital data, different datasets are called hetero center[1]. Tandem feeding means that the output of the previous one serves as input for the next one. L_{BCE} is binary cross entropy loss function.

Q2: In formula 9, can N be zero?

A2: N will not be zero. It is the number of proposals generated by CRM. When it is zero, we will randomly select some proposals from the initial ones to ensure that N is not zero.

Q3: The text is poor with some errors.

A3: We will carefully check and correct all errors (especially some examples which you list) in the revised version.

Reviewer #2:

Q1: How to eliminate significantly different proposals if segmentation results are incorrect?

A1: We set low scores for significantly different proposals through SCM (classify polyps) and MIB (locate polyps) and filter them by DPS. Through these operations, MDNet can reduce the number of false positives and improve the precision.

Q2: What is the rationale behind the correlation between ROI coordinates and polyp categories?

A2: We first locate the position of polyps and then further determine their categories. The coordinates of the ROI can provide position information of the polyps, so there is an indirect connection between coordinates and polyp categories.

Q3: The multiple instance branch (MIB) should be evaluated in the ablation study.

A3: As mentioned in your comments, MIB is crucial in training weakly supervised detection networks. Thus, we will conduct ablation study to evaluate its effectiveness based on your suggestions and add the results to our revised version.

Q4: Should include more recent weakly-supervised detection methods.

A4: We will compare with more recent weakly-supervised methods (e.g., AME-CAM 2023, SPA 2021) in the revised paper to further evaluate the learning ability of MDNet.

Q5: The paper contains several errors.

A5: Thank you for pointing out our mistakes. According to your suggestions, we will carefully check the entire paper and correct these errors in the revised version.

Reviewer #3:

Q1: What is the evidence that pre-trained model can learned generic features of polyps?

A1: In the generalization experiment[2], the model was trained with Kavsir and CVC-ClinicDB, and evaluated on CVC-ColonDB, ETIS and test set (CVC-T) of EndoScene. The slight performance difference (0.18, 0.27, 0.028) shows that the model achieves considerable robustness, which indicates that the pre-trained model can learn generic features.

Q2: Could directly use the pre-trained or a better model to entirely solve weakly supervised polyp detection?

A2: As reviewer #2 Q1 said, results of the pre-trained model may be wrong. If directly use the model, MDNet will lose opportunity to modify final results, because there is no other proposals can be selected. Better models can well solve the problem, but they need to be further designed for specific tasks[3]. Inspired by your idea, we will further improve our model from this direction to achieve better performance.

Q3: What is the performance if directly use pre-trained model? Could provide independent validation of the pre-trained model on the polyp detection task?

A3: These are very good suggestions. According to your suggestions, we have conducted some elementary validation. The results show that when we set the iou threshold to 0.1, 0.3 and 0.5, respectively, the mean Average Precision (mAP) on the private will decrease$\downarrow$ (4.93%, 9.88%, 5.84%), and the mAP on the Kavsir-SEG will increase$\uparrow$ (1.27%, 6.13%, 9.62%) but the edge AP will all decrease$\downarrow$ to 0. More detailed results and analysis will be described in the revised paper.

Reviewer #4:

Q1: Model mostly shows less competitive accuracy than fully supervised models.

A1: In the detection task, results with image-level annotations are indeed less competitive with instance-level annotations. Specifically, **image-level annotations** only provide category information and lack location guidance, so this detection **(MDNet)** is more difficult and less competitive. Moreover, this situation also exists in natural images, for example, the best mAP of full supervision is 89.3% but only 58.1% of weak supervision. In particular, the best AP is only 39.2% when detecting variable shape categories (similar to polyps) (e.g., boat, bottle, chair, person, plant)[4].

Q2: Do morphological category labels increase physicians annotation compared with frame-level pathological ones?

A2: Thank you very much. They do not increase the annotation effort. In our method, we do not need pathological labels and morphological labels are also frame-level.

Q3: Should give Cross Reference module (CRM) and Spatial Category module (SCM) better explanations.

A3: CRM can eliminate negative proposals through pseudo-labels before training to reduce noise and let the network learn more accurate polyp information. SCM can carefully learn category differences through high-low level features to better judge the polyp morphology. For some explanations, please see mentary material Section 1. And we will add more explanations in the revised paper.

Q4: Why does YOLO or Faster RCNN appear less competitive in polyp detection?

A4: These methods are designed for natural images and do not consider the special polyp detection challenges of variable size and blurred edges, so they appear less competitive.

[1] [1] Ali et al., A multi-centre polyp detection and segmentation dataset for generalisability assessment[J]. Sci Data 2023

[2] [2] Fan et al., Pranet: Parallel reverse attention network for polyp segmentation[C]. MICCAI 2020

[3] [3] Mei et al, A Survey on Deep Learning for Polyp Segmentation: Techniques, Challenges and Future Trends. arXiv:2311.18373

[4] [4] Chen et al., Image-level labeled weakly supervised object detection: a survey[J]. Journal of Image and Graphics 2023