# WSPD: Morphology-Driven Weakly Supervised Polyp Detection on Colon Image

**Xiuquan Du, Jiajia Chen, Xuejun Zhang, Hancheng Wu** and **Lou Sun**

School of Computer Science and Technology, Anhui University, Hefei, China

dxqllp@ahu. edu. cn

# 1. Introduction

Colorectal cancer (CRC) is the third most prevalent type of malignancy throughout the world [10]. Polyps are anomalous tissue and the precursor of colon cancer. Therefore, an effective way to reduce morbidity and mortality from colon cancer is to remove them before they become cancerous. Colonoscopy represent the most widely utilized tool of colon polyps examination globally and hold immense significance in the early screening, including provides information about the location and appearance of polyps, making it a gold standard in clinical diagnosis. A number of studies have shown that early colonoscopy has contributed to a 30% decline in the incidence of CRC [1]. Whereas, in the process of colonoscopy, there are different degrees of missing detection, 8.65% by experienced physicians and 57.14% by unskilled[9]. Missed polyps can lead to survival rate as low as 10%[8]. **Hence, an accurate and automatic polyp detection approach capable of finding all possible polyps at an early stage is of great meaning[11]**.

With the development of computer technology in medical image analysis, various methods have been tried in polyp detection.**Traditional** methods rely on classifiers and combine morphological hand-crafted features based on colour, texture and shape,such as Haar, colour wavelet analysis, local binary pattern (LBP) , histogram of oriented gradients(HoG), edge detection ,etc to distinguish polyps from the normal mucosa[2-7]. **Nonetheless, these methods commonly yield inferior detection results and struggle with limited generalization ability.**The main reason is that the representation capability of hand-crafted features is exceedingly limited when it comes to dealing with polyps and backgrounds that lack strong contrast.To address this issue,[abundant](https://jfy.xiaowazi.com/jfc004z.html" \o "abundant的近义词、同义词) **deep learning** based methods have been developed for polyp detection.Yuan et al.[14]brought rotation invariance and image similarity to polyps recognition. Fang et al.[12] designed a double-branch structure to detect polyps and used edge information to improve the detection rate. Jiang et al.[13]proposed cross-frame contrastive learning module to solve rapid video jitters problem in video polyp detection.Although these methods have greatly improved accuracy and generalization ability compared to traditional methods, **they are all fully supervised methods and the clinical application of them is still flawed**, as shown in Figure. 1(a), for several reasons: (1)Hard Annotation. During the data collection process, boxes with (xmin,ymin) and (xmax,ymax) are not given at the same time[16], which need clinicians spend extra time to annotate them. As a result, this greatly leads to the inapplicability of fully supervised polyp detection; (2)Imperfect prediction. The category、class confidence and location of polyps, with different information, are equally important and expected to be given simultaneously. However, existing methods have diverse prediction content missing. (3)Inadequate category evaluation. Current models’ performance are mostly evaluated by mAP of pathology (Adenoma and Hyperplasia), but studies have shown that the missed rate of polyps depends on their morphology (Flat, Pedicle, Edge)[17]. Therefore, evaluation should be based on the mAP of morphology rather than pathology.

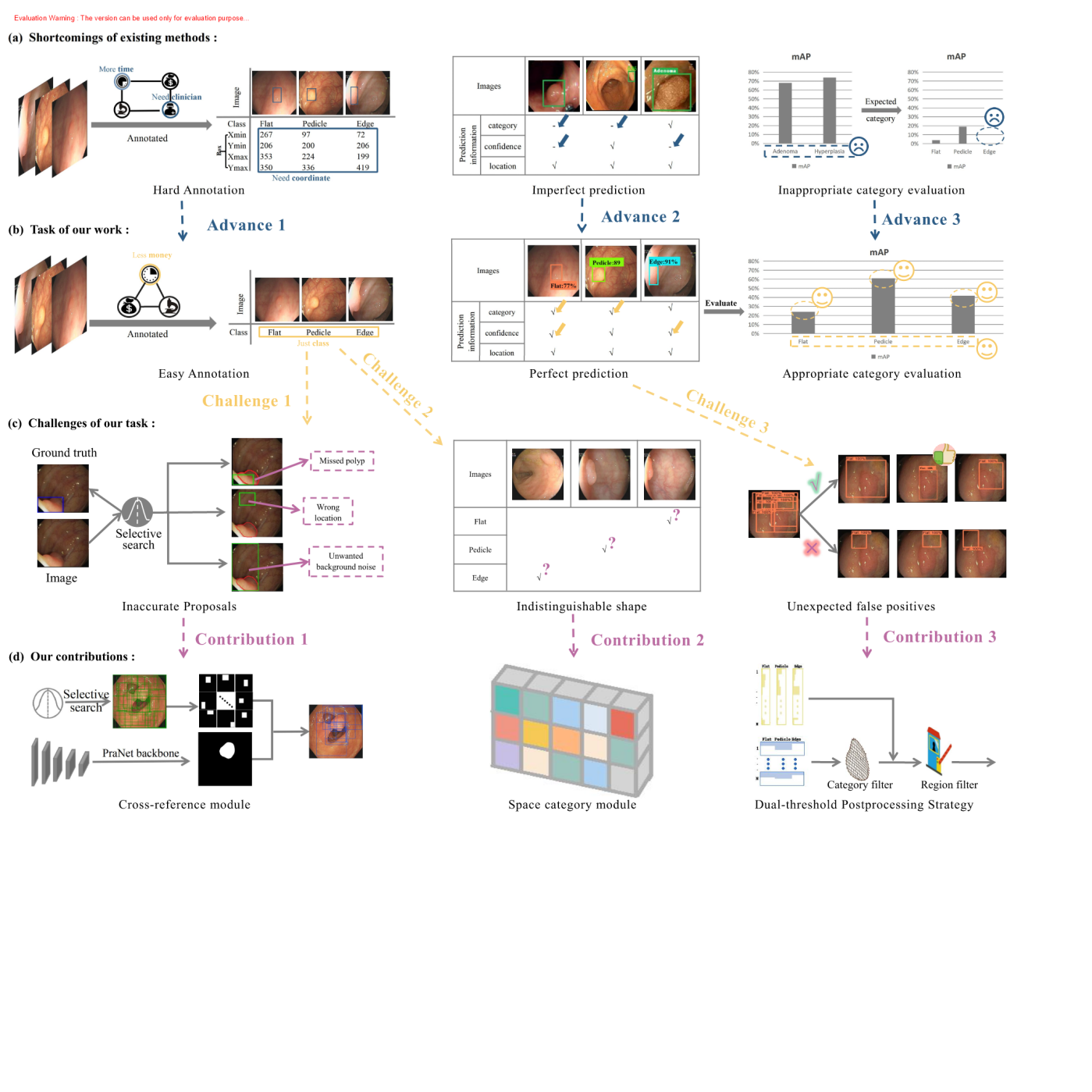


Figure 1: The motivation of our framework to handle the polyps detection via using the colonoscopic image. From (a) to (d)，they are shortcomings of existing methods, Task of our work, Challenges of our task and our contributions, respectively.

In the face of existing methods’ shortcomings, inspired by **multiple instance learning(MIL)[23]**, **we design the weakly supervised polyp detection work based on** , as shown in Figure. 1(b), which aims to detect instances with only image-level labels (e.g. category of instances in the whole images). Specifically, given an image, we first annotate the class. Then feed the image and label into our model for training and evaluation(by morphological mAP). Finally, select the best one to test, and output the prediction with location, confidence and category classified by morphology.Unfortunately, with the low contrast between polyps and backgrounds, initial proposals are very inaccurate(e.g. missing polyps, fault location, background noise), as shown in Figure. 1(c) left. These inaccurate proposals not only waste computational resources during training, which also cause the false positives of detection as shown in Figure. 1(c) right. In addition, the shape of polyps are very similar, with low discriminability, providing them with powerful camouflage properties, and making them difficult to classify. Figure 1(c) middle shows polyps of different types.

Given the above challenges, we propose an weakly supervised polyp detection network(WSPD). As shown in Figure 1(d), our approach mainly includes three main parts. The first one is cross-domain reference module(CRM) for eliminating inaccurate suggestions before training, through pre-training a network to generate pseudo-labels, and uses them as the reference to filter original proposals. In this way, the network will have the ability to ensure the quality of candidate boxes, and minimize the impact of imprecise regions. The second part proposes a spatial category module (SCM) to put more learning pressure on the feature extraction stage and take full account of high-level classification features and low-level compositions such as position, shape and size so that identify polyps of similar shapes. Lastly, we further develop a dual-threshold post-processing strategy(DPS)to better choose the predictions, which capture ahead high confidence boxes via the category threshold and then drop low scores areas from them controlled by region threshold.

Our contributions can be briefly summarized as follows:

* We propose the cross-domain reference module to assemble and ensure the quality of proposals needed for polyp detection, meanwhile reducing the possibility of false positives.
* We design the spatial category module to further learn the global features of spatiality,which can help to classify polyps with similar shapes.
* We present the dual-threshold post-processing strategy, consists of both fixed class and dynamic regional thresholds to refine the detection results.
* We are the first work on weakly supervised detection guided by morphological categories, with significant results on three datasets.

# **2. Related Work**

## 2.1 Multiple Instance Learning (MIL)

MIL[23] is a classical concept in weakly supervised machine learning, and was first proposed in [20] for drug activity division. Formally, given a training image I, Selective search Windows (SSW) [19] or Edge Boxes (EB) [21] generates instance boxes, all of which in some form constitute a "bag" with the image category as the label. According to whether there is at least one positive instance of the class in the package, it will be divided into positive or negative, when it is positive, the positive instance should be found from the positive bag, so as to achieve the combination of MIL and weakly supervised detection.

## 2.2 Pathology-Driven polyp detection

Box-level pathology labels have been utilized to train detectors or to improve weakly supervised polyp segmentation. In practice, different computer vision techniques are adapted to perform polyp detection, which can be broadly classified into two main categories:two-stage detectors and one-stage detectors.

**Two-stage detectors** needs a region proposal stage. Among them, Faster RCNN[31] is the most classic representative, which replaces slow selective search algorithm with a region proposal network, resulting in a faster detection rate. With the development of diffusion models, [Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen,+S) et al. proposed DiffusionDet[32] based on diffusion that formulates detection as a denoising diffusion process from noisy boxes to object boxes.

**One-stage detectors** get rid of the region proposal stage and directly predict bounding boxes by densely sampling the entire image. For instances, YOLO[33] splits the image into grids. Each cell is in charge of predicting boxes coordinates, region confidence scores and category probability with centers located inside the cell, which is supervised through the regression loss function.

Although the two-stage detector has a region proposal stage, which takes more processing time than one-stage, it produces higher accuracy, so we select two-stage method to detect polyp.

## 2.3 Weakly supervised location

Due to the scarcity of bounding box annotations, the community has started to tackle object using weakly supervised detection to address the issue, which can be summarized as two solutions, one used Class Activation Mapping (CAM)[22](see the blue path in Figure 2) and the other via MIL[23](see the green path in Figure 2). For example, Ahn et al. [24]used CAM from the classification network for the location of areas. Li et al.[25]suggested a guided attention inference network that iteratively tunes the model by erasing the CAM area of interest. Further, Kwon et al. [26]interpreted the histological classifier through Grad-CAM[27] to locate polyps. In the MIL-based approach, WSDDN[28] proposed a composing of two branches acting as a proposal selector and a proposal classifier, respectively. However, due to the absence of label location information, it suffers from the discriminative region problem. To alleviate the problem, OICR[29] adds three instance classifier to refine procedures after the baseline. Unlike the above method, WSOD2[30]simultaneously calculates the region scores of high and low level features to solve the challenge.

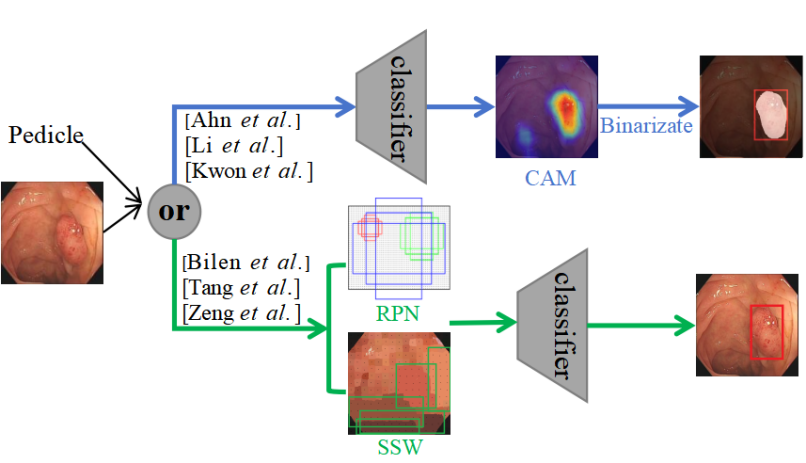


Figure 2: Overview of weakly supervised detection.The blue and green arrows indicate the workflow of class activation mapping (CAM) and multiple instance learning(MIL), respectively.

# **3. Method**

In this section, we present the structure of the weakly supervised polyp detection network, which consists of a cross-domain reference module(CRM), a spatial category module (SCM), and a multiple instance branch for train and the dual-threshold post-processing strategy(DPS) for test.

The overall architecture is shown in Figure 3. Given an image, features are extracted by the pre-trained ConvNet on ImageNet [42] and proposals are generated by cross-domain reference module. Then, send features to the ROI pooling layer for region mapping, whose results are fed to the spatial category module and the multiple instance branch to obtain location and category score, respectively. After that, the scores are dotted product, clamped column sums to form image category. The regression training of the network is guided by the categorical labels y = [y1, y2, ···, yC] ∈{0, 1}C, where yc = 1 or 0 denotes the presence or absence of the class c. The rest of this section discusses these three modules in detail, as well as the dual-thresholding post-processing strategy.

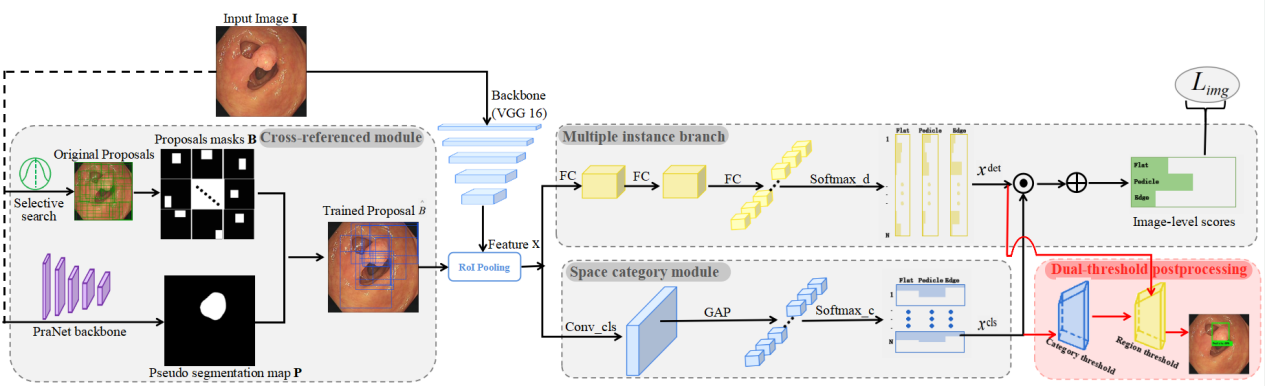


Figure 3: Overview of the proposed WSPD. The WSPD includes three parts for train with cross-reference module (CRM), multiple instance branch, space category module (SCM) and an extra dual-threshold post-processing for test. The solid ellipses denote the image-level loss functions . The black and red arrows are denoted as train data streams and post-processing operations, respectively.

## 3.1. Cross-domain reference module

We use SSW generate proposals to locate polyps. However, they have a large number of erroneous, missed or noisy items, which will bring false positives to the prediction. To solve the problem, we propose a cross-domain reference module(CRM) to select proposals with less interference. Specifically, we pre-trained a heterocentric segmentation network based on public dataset, after which we fed our data into the network to produce pseudo results (P). At the same time, the SSW algorithm is used to generate initial proposals and then transformed into binary masks (B). Intuitively, if B differs significantly from P, then B may be a hard or mislabelled sample, as the pre-trained segmentation model must have learned some generic features of the polyp. In this case, this bounding box will be filtered out from using. Given image I, we send it to segmentation network S, load the saved weighting parameters and get the prediction result P with the following formulation:



where I denotes the input image and  is the weight of the pre-trained segmentation network S. P is a binary segmentation map with the size of h × w.

At the same time, the SSW algorithm is applied to the image and the binary mask is converted to obtain the coarse candidate box B. The Jaccard Coefficient jc=  is chosen to measure the difference between B and P, and only  the boxes are selected, and the final output is the refined candidate boxes ={b1, b2, . . ., bN}(where N is the number of proposals), the formula is as follows:





wheredenote filter threshold and SSW is the selective search windows, details are referred to[19]. For each image, is a shortlist of object proposals. Each element B(i, :) indicates the coordinate of the proposals with the form (xmin, ymin)、(xmax, ymax).

After optimization according to Eq. (3), potential interference can be minimized, and fine boxes are fed into the network for subsequent training.

## **3.2. Spatial Category Module**

Generally, the pattern of multi-classification network is to extract features, flatten them into one-dimensional vectors by fully connected layer, and give classification by softmax. However, direct flattening will loss a lot of spatial information, since larger region with more variation are more likely to have high category consistency, which will cause the indivisibility of categories. To ensure the accuracy of the classification, we design the spatial category module(SCM). Formally, the spatial category module contains a convolutional layer, a global average pooling layer and a flatten layer as follows:



where  is the flatten operation and  is the global average pooling (GAP) layer designed for network.  and b are the parameters of the module, which is a 3 × 3 convolution. Then, it is delivered to the softmax operator. The classification scores can be calculated by:



The classification scores is a matrix with the same shape as the detection scores, which is described in detail in the next section. In this case, we expect it to learn more foreground/background information to help categorize objects.

## **3.3. Mutiple Instance** Branch

The standard object detector requires instance-level supervision. However, we only have image level labels, so we design a multi instance branch to achieve category supervised detection,which performs detection by scoring regions correlation to object. Given a region feature map graph, the branch takes it as input and outputs region scores matrix 。Specifically, the feature of proposals is tandem fed into three fully connected layers FC1~3 to obtain stretching vector, and then passed to the softmax operator, this time defined as follows:





Here,  has  channels, which differents from the input feature map channels, and the former is the result of the fully connected layer processing. Each elementindicates the probability of the ith proposal bi belonging to the jth category.

In the case, in fact, the softmax operator compares, for each region independently, class scores。In this method, the module can evaluate which region holds a better information . After that, the scores of all proposals are generated by element-wise product⊙. During the training stage, the loss function can be formulated as follows:



## **3.4. Dual-threshold Postprocessing Strategy**

To achieve accurate prediction that efficiently captures object parts while minimally covering background, we propose a dual-threshold post-processing strategy(DPS). In this section, we will make a detail introduction to the setting up, which are specifically designed for the given task. The workflow is shown in Figure 4.

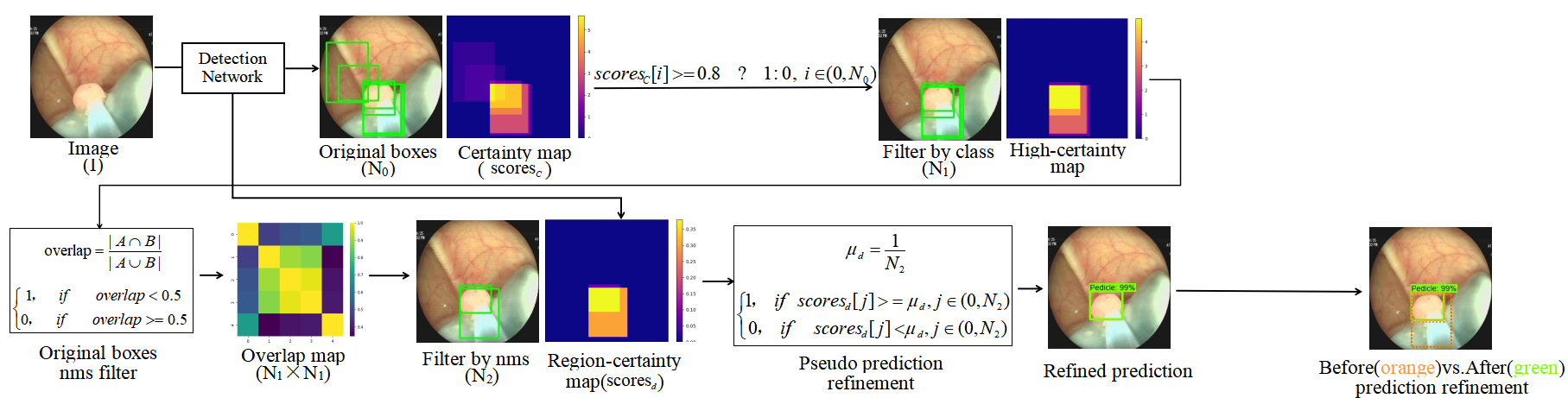


Figure 4:The workflow of dual-threshold post-processing strategy(DPS).

**Category filtering strategy.** Although categories may not be able to distinguish region with high spatial coverage in terms of categories, they are effective in distinguishing the target from the background. To filter out untrustworthy candidates, we set a category threshold, and only the item with greater category confidence than the threshold are retained. After that, non-maxima suppression (with the threshold) is applied to the proposals. Finally, proposals with scores higher thanand overlap lighter than are held for the further filtering process.

**Region filtering strategy.** Filter only on categories will lead to a large number of false positive (FP) predictions. This is due to the fact that classification is unable to select the most valuable regions, and it prefers boxes that contain a lot of background as predictions because they have more categorical information. In contrast to that, region confidence is more concerned with whether bounding boxes tightly surround the object. So naturally, we further set the region threshold, which dynamically changes unlike the fixed category thresholds. Consider the worst situation, where the proposals have high spatial overlap that tightly frame different parts of the polyp. In this case, the region scores of the proposals should be averaged. Therefore, we set the region threshold with the following formula under the worst-case :



where N denotes the number of proposals in an image, and the result is between 0 and 1.

# 4. Experiments

In this section, we introduce the evaluation datasets and the implementation details. Then we perform thorough experiments to analyse WSPD and its components.

## 4.1. Experimental Setup

**Datasets.** Our evaluation is conducted on three polyp datasets: publicly available CVC-ClinicDB[34], KvasirSEG[35] and our private dataset. CVC-ClinicDB contains 612 annotated frames extracted from 29 different colonoscopy sequences. KvasirSEG has 1000 images with polyp regions, manually annotated by an experienced doctor released in 2020. Private dataset consists of 290 images ,which comes from The Second Hospital of Anhui Medical University (Furong Road, Hefei Economic Development Zone, Hefei City) contributed by different patients.

**Implementation.** For the backbone we use VGG16[38] to extra feature, which pre-trained on ImageNet [37] and has some conv layers with max-pooling layer. We replace the last max-pooling layer of the model by ROI pooling. To reduce errors and noise of candidate regions, cross-domain reference module is added and the PraNet[36]is selected for the segmentation network, which details are formulated in Section 3.1. To accurately classify polyps, we develop a spatial category module consisting three operations of convolution, global average pooling and softmax introduced in Section 3.2. To locate polyps, we apply three different fc layers and the last fc layer, softmax operation and image-level loss are described in Section 3.3.

We follow a two-step training strategy: 1) the segmentation network is trained with fixed learning rate 10-4 for 20 epochs. 2) the entire architecture is trained following the end-to-end manner. The WSPD runs for 15 epochs with 10-5 following 10 epochs with learning rate 10−6. To achieve the filtration of initial proposals and detection post-processing, the value of the threshold are simply set, i.e. , , , respectively.

## 4.2. Ablation experiments

For the sake of evaluating the potence of our components in the proposed method, we conducted sufficient ablation studies on our private dataset with all the combinations of multi-pretext tasks of CRM, SCM, and DPS with consistency loss. There are four ablation types as below: 1) Original task: Baseline. 2) Single pretext task: CRM, SCM, DPS. 3) Dual pretext task: CRM+SCM, CRM+DPS, SCM+DPS 4) Triple pretext task: CRM+SCM+DPS. The results are shown in Table 1. The foundational detection method without any pretext is the same as the WSDDN[28], and we continually add modules as mentioned strategies above. When the implementation of aforementioned modules by degree, boosting mAP by 6.64%, 8.4%, 9.83%, 13.26%, 11.01%, 13.93% and 24.52%. These results proves the effectiveness and necessity of the dual-threshold post-processing strategy and other modules.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Types | Baseline | +CRM | +SCM | +DPS | mAP@0.5 |
| Original | √ |  |  |  | 6. 17 |
| Single | √ | √ |  |  | 12. 81 |
| √ |  | √ |  | 14. 57 |
| √ |  |  | √ | 16. 00 |
| Dual | √ | √ | √ |  | 19. 43 |
| √ | √ |  | √ | 17. 18 |
| √ |  | √ | √ | 20. 10 |
| Triple | √ | √ | √ | √ | **30. 69** |

Table 1: mAP (in %) of different weakly supervised strategies with the same backbone on our private dataset.

In order to give more intuitive results, we graphically contrast the contributions of CRM in Figure S1~S2，give some example on DPS in Figure S3 and visualize the class activation map for the SCM in Figure S4.

## **4.3. Comparison with other methods**

### Learning Ability

To validate our model’s learning ability, we evaluate comparison methods on ClinicDB and Kvasir-SEG datasets. We refer to the ratio eight to two to divide the trainval set and test set. We resize the images to 224×224 on both ClinicDB dataset and Kvasir-SEG dataset. As shown in Table 2, our proposed method achieves competitive results on all two datasets. On the ClinicDB dataset, our model achieves 87.93% Dice and 89.56% mean IoU, meanwhile achieving 92.35% Dice and 91.99% mean IoU on the Kvasir-SEG dataset. It demonstrates that our WSPD significantly improves the performance of polyp detection on average and has a strong learning ability. Additionally, in the Figure 5 we also give some visual prediction where our model successfully detected but failed to compare the models.

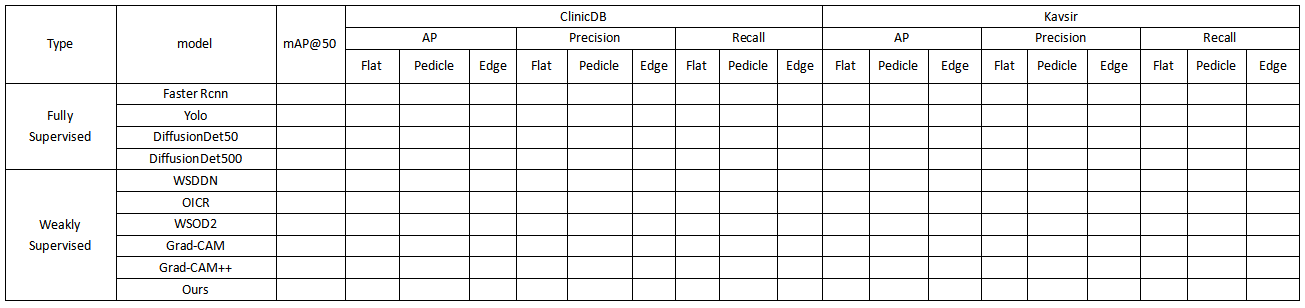


Table 2: Quantitative results of the test datasets ClinicDB and Kvasir-SEG. The best results are highlighted in **bold**.

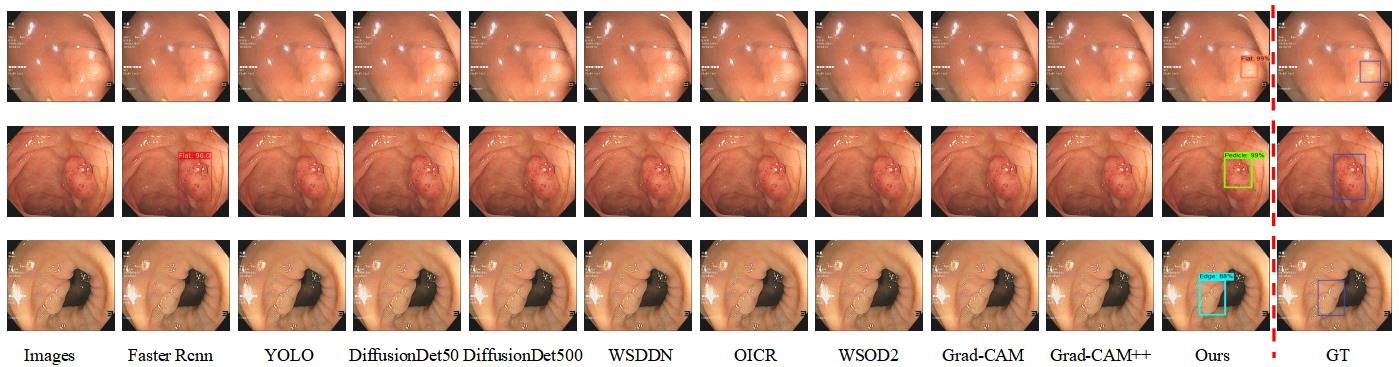


Figure 5: Qualitative results of different methods, where the orange, green, green proposals represent class flat, pedicle and edge, respectively. (Absence of boxes means no detected object given the detection threshold.)

Though our method significantly outperforms other methods, the performance is poor for class“Flat”. For analysis, we visualize some success and failure detection results on VOC 2007 trainval by WSPD Ens., as in Figure 6. We can observe that, our method is robust to the size and aspect of polyp. The main failures for the class are always due to overlarge boxes that not only contain objects, but also include their adjacent similar objects. For flat polyps , they are always with great deformation, while there is less deformation of their most representative parts (like head), so our detector is still inclined to find these parts. In addition, class imbalance may also the reason of failed detection. An ideal solution is yet wanted because there is still room for improvement.



Figure 6: Some detection results for class flat. Green rectangle indicates success cases(IoU>0.5), and red rectangle indicates failure cases (IoU<0.5).

### Generalization Capability

In order to adapt different shapes of polyps in clinical practices, we conduct one experiment on the private dataset to test the model’ s generalizability. EndoScene is a combination of CVC612 and CVC300. We use it to serve as the generalized dataset with 80% as training set and the remaining 20% as a testing set. All the images are set to 224x224.

To evaluate the generalization capability of WSPD, we present mAP, class ap, class precision, and class recall results in Table 3. As can be seen, our methods achieves the highest values on Dice (74.97%) and mIoU (82.64%) metrics ,which indicates that our detection has a high degree of coincidence with ground truth. In the meantime, we also present the radar plot in Figure 7 to investigate the performance from different dimensions, simultaneously. Specific qualitative result in Figure S5.

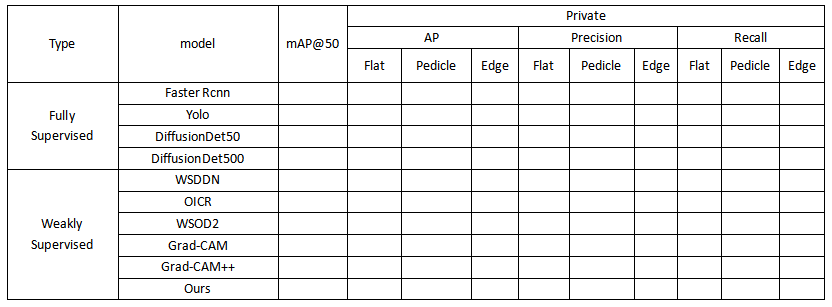


Table 3: Quantitative results of the test datasets private dataset. The best results are highlighted in **bold**.

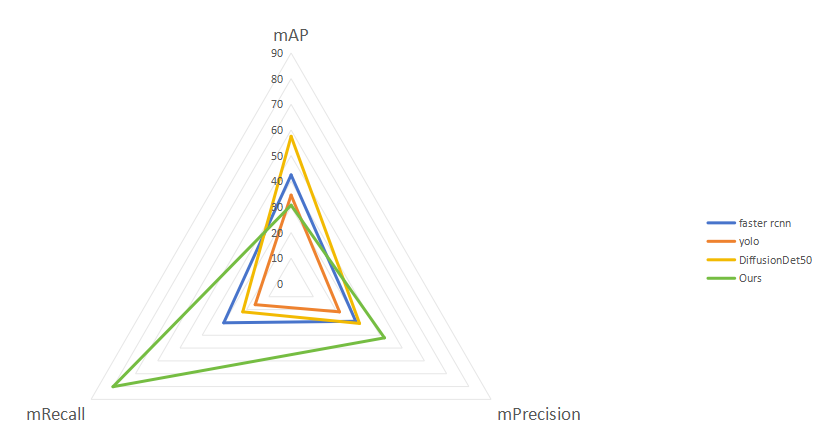
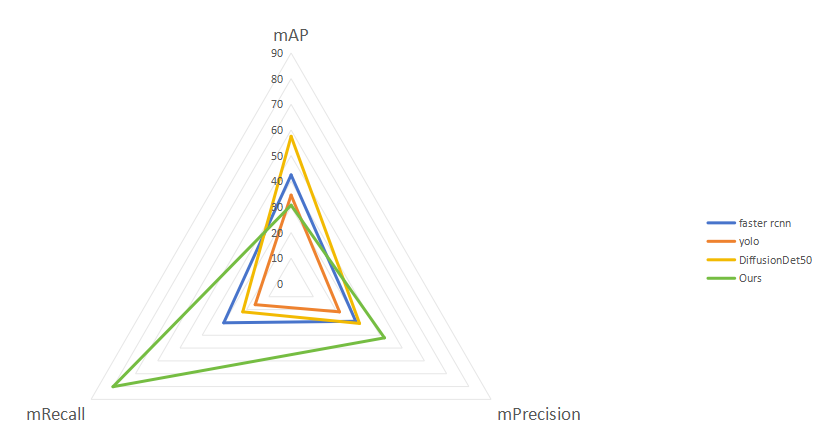


Figure 7: The radar plot of different methods.

# Conclusion

In this paper, we present a novel weakly supervised network for automatic polyp detection from colonoscopic images. The cross-domain reference module integrates initial proposals to polyp around to avoid erroneous, missing and noisy location. The spatial category module combines local-global spatiality features of each categories to help identify less-diverse forms polyp. And The dual-threshold post-processing strategy focus on prominent region with fewer background to reduce false positives. Extensive experiments and ablation studies safely reach the superiority of the proposed method and achieve new state-of-the-art performance.

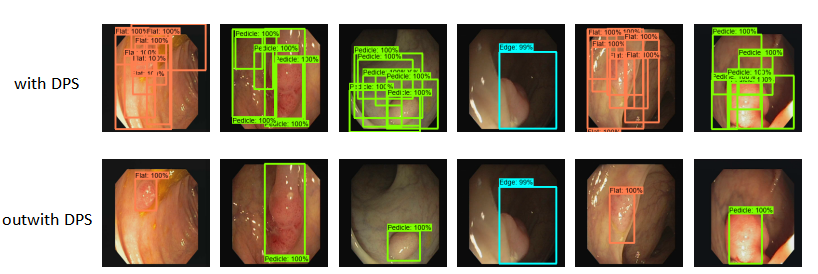
**Acknowledgments**

This work was supported in part by the Provincial Natural Science Research Program of Higher Education Institutions of Anhui province under Grant KJ2020A0035. And then, the authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

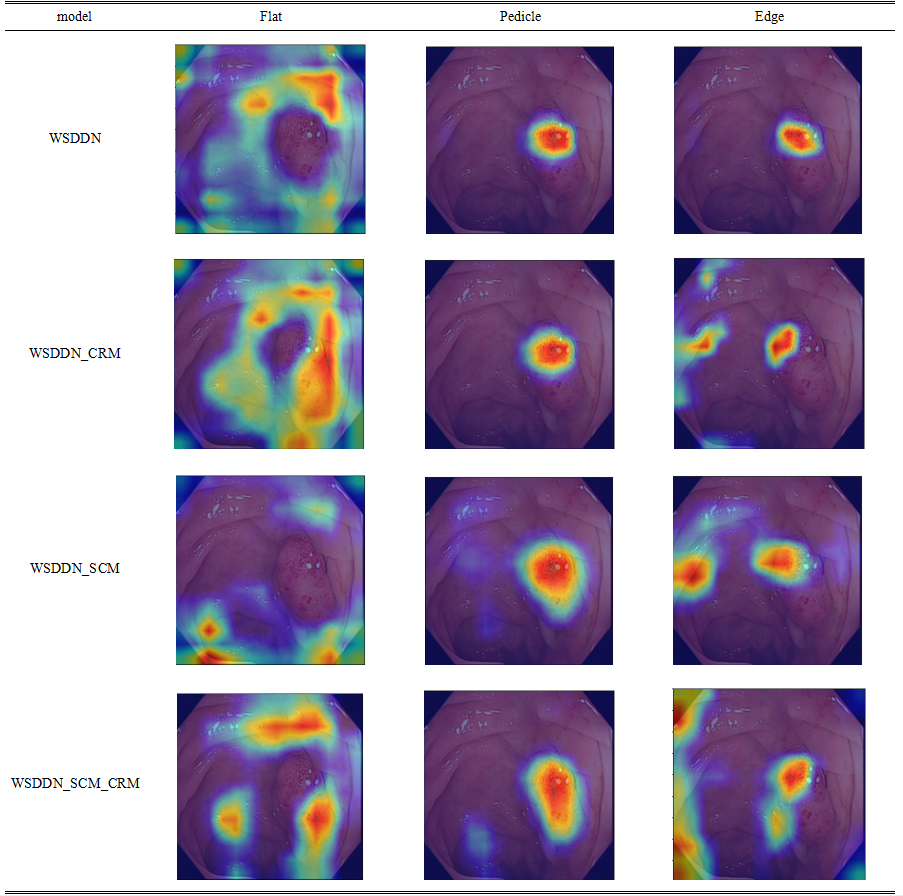
**References**

1. Haggar, F. A., Boushey, R. P. : Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics in colon and rectal surgery 22(04), 191–197 (2009)
2. S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, “Computer-aided tumor detection in endoscopic video using color wavelet features, ” IEEE Trans. Inf. Technol. Biomed., vol. 7, no. 3, pp. 141–152, Sep. 2003.
3. S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, “Texture based polyp detection in colonoscopy, ” in Proc. Bildverarbeitung fur die Medizin, 2009, pp. 346–350.
4. S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. de Groen, “Polyp detection in colonoscopy video using elliptical shape feature,” in Proc. IEEE Int. Conf. Image Process., vol. 2, pp. II-465-468, 2007.
5. L. A. Alexandre, N. Nobre, and J. Casteleiro, “Color and position versus texture features for endoscopic polyp detection, ” in Proc. Int. Conf. BioMed. Eng. Informat., 2008, vol. 2, pp. 38–42.
6. J. Bernal, J. Sanchez, and F. Vilarino, “Towards automatic polyp detection with a polyp appearance model, ” Pattern Recognit., vol. 45, no. 9, pp. 3166–3182, 2012.
7. S. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim, “A colon video analysis framework for polyp detection, ” IEEE Trans. Biomed. Eng., vol. 59, no. 5, pp. 1408–1418, May 2012.
8. L. Rabeneck, J. Souchek, and H. B. El-Serag, “Survival of colorectal cancer patients hospitalized in the veterans affairs health care system, ” Am. J. Gastroenterology, vol. 98, no. 5, pp. 1186–1192, 2003.
9. Chen Y L. Analysis of missed diagnosis rate and related factors of colorectal polyps in colonoscopy [J] . 2020. (in Chinese)
10. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293 (2014)
11. Jia, X., Xing, X., Yuan, Y., Xing, L., Meng, M.Q.H.: Wireless capsule endoscopy:A new tool for cancer screening in the colon with deep-learning-based polyp recognition. Proceedings of the IEEE 108(1), 178–197 (2019)
12. ABC-Net: Area-Boundary Constraint Network With Dynamical Feature Selection for Colorectal Polyp Segmentation[J].IEEE Sensors Journal , 2021, 21(10):11799 - 11809. DOI:10.1109/JSEN.2020.3015831.
13. Jiang, Yuncheng et al. “YONA: You Only Need One Adjacent Reference-frame for Accurate and Fast Video Polyp Detection.” International Conference on Medical Image Computing and Computer-Assisted Intervention (2023).
14. Yixuan Yuan, Wenjian Qin, Bulat Ibragimov, Bin Han, Lei Xing. RIIS-DenseNet: Rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection,International Conference on Medical Image Computing and Computer-Assisted Intervention,11071:620-628,2018.
15. Shin Y , Qadir H A , Aabakken L ,et al.Automatic Colon Polyp Detection using Region based Deep CNN and Post Learning Approaches.DOI:10.1109/ACCESS.2018.2856402
16. A N R, B M W S, A J G, et al. Automated Detection of Acute Respiratory Distress Syndrome from Chest X-Rays Using Directionality Measure and Deep Learning Features [J] . Computers in Biology and Medicine, 2021. DOI:10. 1016/j. compbiomed. 2021. 104463.
17. Chen Y L. Analysis of missed diagnosis rate and related factors of colorectal polyps in colonoscopy [J] . 2020. (in Chinese)
18. H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In CVPR, pages 2846–2854, 2016. 1, 2, 3, 4, 7, 8
19. K. van de Sande, J. Uijlings, T. Gevers, and A. Smeulders. Segmentation as selective search for object recognition. In Proc. ICCV, 2011.
20. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 89(1):31–71, 1997.
21. C. L. Zitnick and P. Dollar. Edge boxes: Locating object ´proposals from edges. In Proc. ECCV, pages 391–405, 2014.
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016). <https://doi.> org/10.1109/CVPR.2016.319
23. Thomas G. Dietterich, Richard H. Lathrop, and Tomas´ Lozano-Perez. Solving the multiple instance problem with ´axis-parallel rectangles. Artificial Intelligence., 89(1-2):31–71, 1997.
24. J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.
25. K. Li, Z. Wu, K.-C. Peng, J. Ernst, Y. Fu, Guided attention inference network, IEEE Trans. Pattern Anal. Mach. Intell. 42 (12) (2019) 2996–3010.
26. Kwon J, Choi K. Weakly Supervised Attention Map Training for Histological Localization of Colonoscopy Images. Annu Int Conf IEEE Eng Med Biol Soc. 2021 Nov;2021:3725-3728. doi: 10.1109/EMBC46164.2021.9629608. PMID: 34892046.
27. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
28. Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2846–2854, 2016.
29. P. Tang, X. Wang, X. Bai, and W. Liu, “Multiple instance detection network with online instance classifier refinement,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
30. Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, “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, 2019.
31. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems; 2015. p. 91–99.
32. Shoufa Chen., Peize Sun., Yibing Song., Ping Luo. Diffusiondet: Diffusion model for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2023: 2211. 09788.
33. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In:Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 779–788.
34. Bernal, J., S´anchez, F. J., Fern´andez-Esparrach, G., Gil, D., Rodr´ıguez, C., Vilari˜no, F. : Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. CMIG 43, 99–111 (2015)
35. Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pal Halvorsen, Thomas de Lange, Dag Johansen, ˚and Havard D Johansen. Kvasir-seg: A segmented polyp ˚ dataset. In MMM, pages 451–462, 2020.
36. Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. MICCAI, pages 263–273, 2020.
37. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. IJCV, 2015
38. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

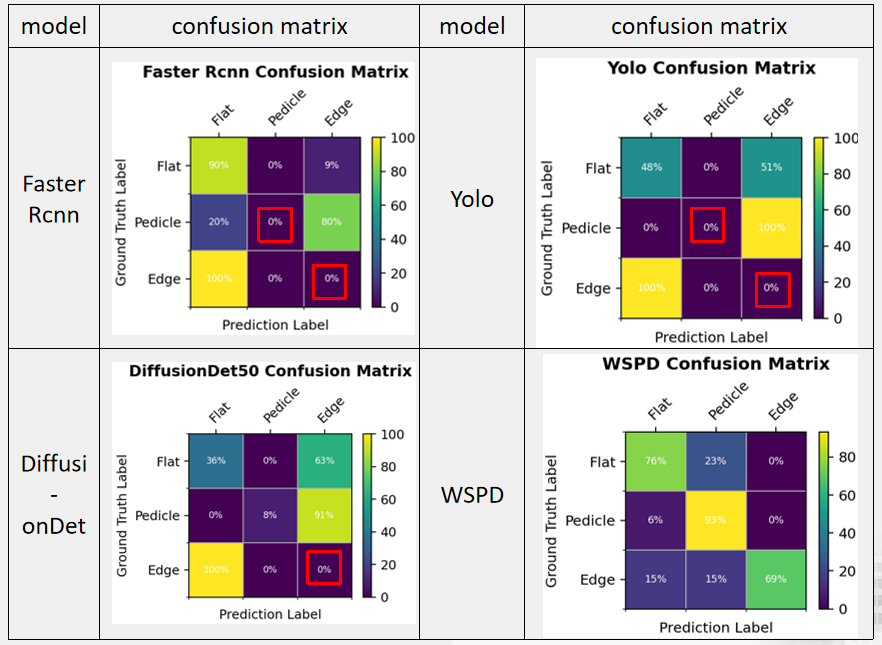
S1: CRM前后proposal的数量图

S2: 与GT的overlap图

S3：DPS作用前后图像检测结果图



S4: SCM类激活图



S5: 混淆矩阵