

**MDNet: Morphology-Driven Weakly Supervised Polyp Detection**

**\*\*\*\*\*\***

**Abstract**

Polyps – indicators of lethal colorectal cancer and

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

Figure 1: (a) The clinical classification of missed polyps. (b) Avisu- alization, representing the number of detected polyps under different supervision.

With the development of computer technology in medi- cal image analysis, numerous methods have been tried in polyp detection. The traditional models extract hand-crafted

[features and are based on various saliency assumptions [Jia](#bookmark5) [*et al.*, 2019;](#bookmark5) [Borji *et al.*, 2015;](#bookmark6) [Borji *et al.*, 2014]](#bookmark7). How-

ever, these methods commonly yield inferior detection re- sults 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 lack strong contrast between polyps and backgrounds. To solve the above limitation, abundant deep learning based methods [have been developed for polyp detection [Jiang *et al.*, 2023b;](#bookmark8) [Reamaroon *et al.*, 2021;](#bookmark9) [Fang *et al.*, 2021]](#bookmark10). Unfortunately, most of them are driven by pathology, demand laborious and lavish collection of instance-level annotations and miss prediction information to some extent, which caused clini- cal infeasibility. As shown in Figure [2](#bookmark11) (A): (1) **Hard An- notation**. During the data collection process, this is typi- cally not possible for bounding boxes with (xmin,ymin) and [(xmax,ymax) [Reamaroon *et al.*, 2021], which limits the](#bookmark9) availability of large datasets for pathology detection. Addi- tionally, manually annotating pathology bounding boxes is a time-consuming task, further exacerbating the inapplica- bility; (2) **Imperfect prediction.** The category, class con- fidence and location of polyps, with different information,

which

41

42

43

44

45

46

47

48

the focus of early screening and prevention, are

missed to varying degrees in clinic due to morpho-

logical difference. Existing methods either need

annotated bounding boxes, neglect the reality of

unprepared polyp proposals or lack complete pre-

dictions. Even worse, they only focus on the de-

tection rate of polyps under pathological classifica-

tion. To overcome these issues, we creatively pro-

pose the ***M****orphology-****D****riven network* (MDNet),

which detects polyps with only image-level super-

vision. Specifcally, by thinking the generic fea-

ture between detection and segmentation, the cross-

domain reference module (CRM) is devised to de-

crease the negative effect of the uncertain propos-

als. Based on spatial differences in polyp mor-

phologies, the spatial category module (SCM) is

designed, who enhances the ability to discriminate

similar polyp of different morphology. In addition,

class and region scores are integrated into the dual-

threshold post-processing strategy (DPS) to im-

prove detection accuracy. We carry out the experi-

ments on three datasets (one internal and two pub-

lic) and experimental results indicate that MDNet

has better robustness and performance. All code is

available at [https://github.com/dxqllp/MDNet.](https://github.com/dxqllp/MDNet)

49

50

51

**1 Introduction**

27

52

Colorectal cancer (CRC) is the third most prevalent type of

28

29

30

31

32

33

34

35

36

37

38

39

40

53

54

55

56

57

58

59

60

61

62

63

64

65

[malignancy throughout the world [Siegel *et al.*, 2023]](#bookmark2). Polyps

are abnormal tissue growth and the precursor to colon cancer.

Whereas, in the process of colonoscopy, there are different

morphological missing detection, which can be roughly di-

vided into three categories, as shown in Figure [1](#bookmark1) (a): flat,

pedicle, edge. Among them, the miss rate of flat is 23.52%

[(the hightest) instead the pedicle 8.95% (the lowest) [Li,](#bookmark3)

[2020]](#bookmark3). Missed polyps may lead to survival rate as low as

10% [[Haggar and Boushey, 2009]](#bookmark4). Hence, an accurate and

automatic polyp detection approach capable of finding all

possible polyps at an early stage is of great meaning in clini-

cal practice.

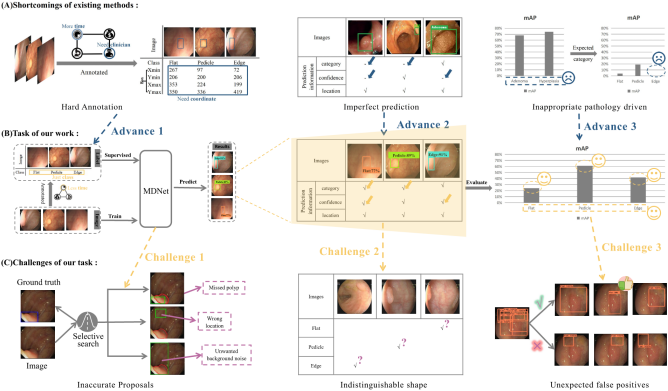


Figure 2: The motivation of our framework to handle the polyps detection via using the colonoscopic image. From (a) to (d), they are shortcomings of existing methods, task of our work, challenges of our task and our contributions, respectively. 

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

are equally important and expected to be given simultane- ously. However, existing methods have different prediction content missing; **(3) Inappropriate pathology driven**. In clinical practice, pathological features are the main factor de- termining prognosis, morphology is the main basis for en- doscopists to recognize polyps and the main task of detec- tion models is to identify various morphologies polyps. But current models primarily focus on the mean average preci- sion (mAP) driven by pathology (adenoma and hyperpla- sia). For instance, on the ClinicDB dataset, Faster R-CNN achieves an AP of 84.18% [[Debesh *et al.*, 2021]when driven](#bookmark12) by pathology, but it significantly declines while switch to the morphology-driven, especially in the detection of flat polyps. [Unfortunately, the study [Jiang *et al.*, 2023a] has shown that](#bookmark13) flat polyps have the highest odds ratio (OR),which highlights the inappropriateness of pathological driving.

In the face of existing methods’ shortcomings, we cre- atively propose a novel end-to-end weakly supervised mor- phology driven network (MDNet), as shown in Figure [2](#bookmark11) (B) , for more meaning polyp detection(in Figure [1](#bookmark1) (b)), which solely uses image-level morphology annotation. Un- fortunately, with the low contrast between polyps and back- grounds, initial proposals are very inaccurate (e.g., missing polyps, fault location, background noise) , as shown in Fig- ure [2](#bookmark11) (C) left. These inaccurate proposals not only waste computational resources during training but also cause the false positives of detection as shown in Figure [2](#bookmark11) (C) right. In addition, the shape of polyps shown as Figure [2](#bookmark11) (C) mid-

dle are very similar, with low discriminability, providing them with powerful camouflage properties, and making them diffi- cult to be classified.

Given the above challenges, our approach makes three ma- jor improvements via morphology-driven technology. The first one is cross-domain reference module (CRM) forelim- inating inaccurate suggestions before training with the help of pseudo-labels. Through this module, MDNet will have the ability to ensure the quality of candidate boxes, and minimize the impact of imprecise regions. The second promotion pro- poses 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 composi- tions so that identify polyps of similar shapes. Lastly, we fur- ther 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:

• To the best of our knowledge, we are the first to draw polyp detection under morphological guidance, where AP on different morphologies is used to comprehen- sively evaluate the performance.

• We propose MDNet based on the morphological cate- gories of polyps that consists of the cross-domain refer- ence module (CRM), the spatial category module (SCM) and the dual-threshold post-processing strategy (DPS).

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

Using CRM, the unaccurate proposals can be filted to reduce the possibility of false positives. With the aid of SCM, the high-low level features can serve more effec- tively to distinguish polyps with comparable morphol- ogy. Additionally, the last module transforms the tra- ditional category-based post-processing into simultane- ous category and region-based post-processing to better choose the predictions.

• Our method outperforms weakly supervised methods and competes with fully supervised detection on three datasets (one internal and two public) in terms of evalu- ation metrics.

**2 Related Work**

**2.1 Pathology-driven polyp detection**

**Two-stage detectors** The main idea of these detec-

[](#bookmark12)tors [[Jiang *et al.*, 2024;](#bookmark20) [Yang *et al.*, 2023;](#bookmark21) [Debesh *et al.*,](#bookmark12)

[2021] looks over region proposals globally in the image and](#bookmark12)

regresses the bounding box in each of the proposals. Among

[them, [Mo *et al.*, 2018]](#bookmark22) fine-tuned the typical Faster R-

CNN [[Ren *et al.*, 2017], and obtained highly competitive per](#bookmark23)-

formance on dataset CVC-ClinicDB compared to the state-

of-the-art approaches. With diffusion models have achieved

[great success in many generation tasks, DiffusionDet [Chen](#bookmark14)

[*et al.*, 2022]](#bookmark14) was proposaled who tackles the object detection

task with a diffusion model by casting detection as a genera-

tive task over the space of the positions (center coordinates)

and sizes (widths and heights) of bounding boxes in the im-

age.

**One-stage detectors** They directly predict bounding boxes

by densely sampling the entire image in a single network

pass, which gets rid of the time-consuming region proposal

[stage. Notably, the YOLO series [Ge *et al.*, 2021] have at](#bookmark24)-

tracted attention due to their desirable features of fast infer-

[ence and low computational burden. For instances, [Karaman](#bookmark15)

[*et al.*, 2023a] demonstrated the importance of combining op](#bookmark15)-

timization algorithms with a real-time detection framework,

such as scaled-YOLOv4. Furtherly, they improved the perfor-

mance on SUN and PICCOLO polyp datasets by integrating

YOLOv5 [[Karaman *et al.*, 2023b]](#bookmark25).

Despite one-stage detectors have obtained the remarkable

progress on speed, they suffer from inflexibility due to the

extensive use of anchor boxes, which leads to a limited range

of scales and aspect ratios. Even more unfortunately, since

the absence of region proposals one-stage detectors usually

have less chance to detect the polyps. Overall, considering

the importance of recall, we employs the two-stage training.

**2.2 Weakly supervised location**

[](#bookmark16)Caused by the scarcity of bounding box annotations, the community has started to tackle object using weakly super- vised detection to address the issue, which can be summa- rized as two solutions, one used Class Activation Mapping [(CAM) [Zhou *et al.*, 2015]](#bookmark26) and the other via MIL. For ex- [ample, [Jiwoon and Suha, 2018]](#bookmark27) used CAM from the clas- sification network for the location of areas. Li et al. [[Li *et*](#bookmark16)[*al.*, 2019] suggested a guided attention inference network](#bookmark16)

|  |
| --- |
| **Algorithm 1** Training and Testing MDNet |
| **Input:** training set with morphology category **T1** = {(**I**, **y**)}. 1: **procedure** TRAINING  2: forward MDNet fS (**I**) → **P**, f SSW (**I**) → **B**, fV GG (**I**) → **x**  3: generate variables with **P** and **B**  4: generate variables **xROI** with **x** and  5: forward **xROI** fSCM (**xROI** ) → **xc** , f MIB (**xROI** ) → **xd**  6: generate variables **xdet** and **xcls** with **xd** and **xc**  7: compute and backward Limg in Equation (8) for MDNet  8: continue until convergence  **Output:** the optimized MDNet for detection  **Input:** test set **T2** = {**I**}.  1: **procedure** TESTING  2: forward MDNet fS (**I**) → **P**, f SSW (**I**) → **B**, fV GG (**I**) → **x**  3: generate variables with **P** and **B**  4: generate variables **xROI** with **x** and  5: forward **xROI** fSCM (**xROI** ) → **xc** , f MIB (**xROI** ) → **xd**  6: generate variables **xdet** and **xcls** with **xd** and **xc**  7: DPS for detected bounding boxes with **xdet** and **xcls Output:** the detected bounding boxes with category and  confidence for **T2** |

[](#bookmark17)that iteratively tunes the model by erasing the CAM area [of interest. Further, [Kwon and Choi, 2021] interprete the](#bookmark28) histological classifier through Grad-CAM [[Selvaraju *et al.*,](#bookmark17) [2017] to locate polyps](#bookmark17). In the MIL-based approach, WS- DDN [[Hakan and Andrea, 2016] proposed a two branches](#bookmark29) network to select and class proposals, respectively. How- ever, due to the absence of label location information, it suf- fers from the discriminative region problem. To alleviate the [problem, OICR [Tang *et al.*, 2017] adds three instance classi](#bookmark30)- fier to refine procedures after the baseline. Unlike the above [method, WSOD2 [Zeng *et al.*, 2019]](#bookmark31) simultaneously calcu- latesthe region scores of high and low level features to solve the challenge.

[Nonetheless, they are all based on previous work [van de](#bookmark18) [Sande Koen E. A. *et al.*, 2011] who adopts segmentation as a](#bookmark18) selective search strategy for object recognition and pubilcize the proposal results on natural images. However, there is a lack of relevant research on polyp images, and, to the best of our knowledge, this field of polyps remains unexplored.

**3 Method**

说明CHW是什么

**3.1 Overview**

The overall architecture is shown in Figure [3.](#bookmark32) Given an image I ∈ RC×H×W, feature x is extracted by the pre- trained VGG [[Simonyan and Zisserman, 2015] and proposals](#bookmark33)

are generated by Cross-domain Reference Module (CRM).

Then, send x to the ROI pooling layer for region mapping

xROI ∈ RC × × , whose results are fed to the Spatial

′

Category Module (SCM) and the multiple instance branch

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

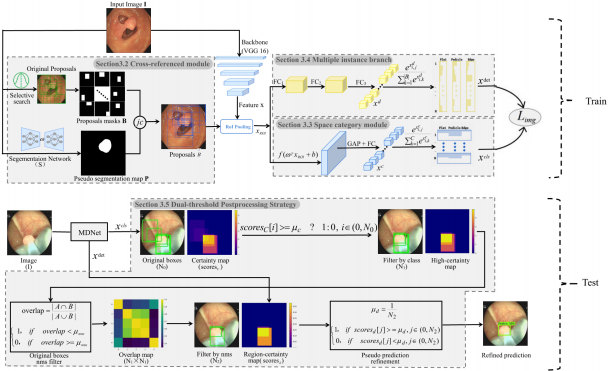
200

201

202

203



缺少后处理指向

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

Figure 3: Overview of the proposed MDNet. TheMDNet includes three parts for train with cross-reference module (CRM),multiple instance branch (MIB), space category module (SCM) and an extra dual-threshold post-processing (DPS) 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.



(MIB) to obtain location and category score, respectively. Af-

and get the prediction result P with the following formulation: 231

P = *f*S (I, w), P ∈ [0, 1]h×w ( [1)](#bookmark35)

where I denotes the input image and w is the weight of the 232 pre-trained segmentation network S. P is a binary segmenta- 233 tion map with the size of h × w. 234 At the same time, the SSW is applied to the image and 235 the binary mask is converted to obtain the coarse candidate 236 box B. Here, the Jaccard Coefficient jc =  237

chosen to measure the difference between B and P, and only 238 jc>τ of the boxes are selected, and the final output is the 239

refined candidate boxes = {b1 , b2 , . . . , bN } (where N is the 240 number of proposals), the formula is as follows: 241

ter that, the scores are dotted product, clamped column sums

to form image category. The training of the network is guided

by the morphological category 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 modules in

detail, as well as the Dual-thresholding Post-processing Strat-

egy (DPS).

For clarity, the training and the testing of our MDNet are

summarized in Algorithm [1.](#bookmark19)



**3.2 Cross-domain reference module**

[The selective search windows (SSW) [van de Sande Koen](#bookmark18)

parameters are missing useful to

compute the outcome

B = *f*SSW (I), B ∈ RN×4 (2)

[E. A. *et al.*, 2011] is used to generate proposals to prelimi](#bookmark18)-

narily locate polyps. However, it has a large number of erro-

242

{B}是一组候选框 ,而在公式3中， 它被描述为一个 函数

= { se jc>τ (3)

where τ denotes filter threshold. For each image, B() is 243 a shortlist of object proposals and B(i,:) indicates the coor- 244

dinate of the proposals with the form (xmin,ymin), (xmax, 245

nep 246

是什么？

什么意思

247

the network for subsequent training.   

248

**3.3 Spatial Category Module** 

249

Generally, the pattern of multi-classification network is to ex- 250

goal

tract features, flatten them into one-dimensional vectors xc 251

neous, 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 hetero-centric seg-

mentation network based on a public dataset, after which we

feed our data into the network to produce pseudo results (P).

At the same time, the SSW is used to generate initial propos-

als and then transformed into binary masks (B). Intuitively, if

B differs significantly from P, then B may be a hard or mis-

labelled 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. Given an image I, we

send it to segmentation networkS, load the weight parameters



252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

flattening

by fully connected layer, and give classification scores matrix

xcls. However, direct flatten will loss a lot of spatial informa-

tion, since larger region with more variation are more likely to

have high category consistency, which will cause the indivis-

ibility of categories. To ensure the accuracy of the classifica-

tion, 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:

xc = fFC (fGAP (f(wTxROI + b))), xc ∈ RN×D (4) where fFC is the flatten operation and fGAP is the global av- erage pooling (GAP) layer designed for network. wand b are the parameters of the module, which is a 3 × 3 convolution. Then, it is delivered to the softmax operator. The classifica- tion scores can be calculated by:

x =  , x ∈ [0, 1]N×C (5)

x shows the probability of the ith proposal bi belonging to

the jth category. The classification scores are 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.

**3.4 Mutiple Instance Branch**

The standard object detector requires instance-level supervi- sion. However, we only have image-level annotations, so we design a multi instance branch (MIB) to achieve category su- pervised detection, which performs detection by scoring re- gions correlation object. Given a region mapping xROI , the branch takes it as input and outputs region scores matrix xdet. Specifically, the feature of proposals is tandem fed into three fully connected layers FC1 3 to obtain stretching vector xd , and then passed to anthor softmax operator, this time defined as follows:

xd = fFC3 (fFC2 (fFC1 (xROI ))), xd ∈ RN×D (6)

不清楚 xt =  , xt ∈ [0, 1]N×C (7)

differs

Here, xd has D channels, which differents from the input fea- ture map channels, and the former is the result of the fully

connected layer processing. Each element xt indicates the

probability of the ith proposal bi including the jth category objects. In fact, the softmax operator compares, for each class independently, the scores of different regions, so the module can evaluate which region holds a better information. After that, the scores of all proposals are generated by element-wise product xB = xdet ⊙ xcls. During the training stage, the loss function can be formulated as follows:

Limg = ：LBCE (： xj , yj ) (8)

什么意思

**3.5 Dual-threshold Postprocessing Strategy**

To achieve accurate prediction that efficiently captures ob- ject parts while minimally covering background, we propose a dual-threshold post-processing strategy (DPS). In this sec- tion, we will make a detail introduction to the setting up, which are specifically designed for the given task.

**Category filtering strategy** Although categories may not be able to distinguish region with high spatial coverage, 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 sup- pression (with the threshold µnms ) is applied to the propos- als. Finally, proposals with scores higher than µc and overlap lighter than µnms 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 categori- cal information. In contrast to that, region confidence is more concerned with whether bounding boxes tightly surround the object. Naturally, we further set the region threshold, which dynamically changes unlike the fixed category thresholds. Considering the worst situation which that the proposals have high spatial overlap that tightly frame different parts of the polyp, the region scores of the proposals should be averaged. Therefore, we set the region threshold with the following for- mula under the worst-case:

µd =  , µd ∈ [0, 1] (9)

N能为0吗

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

**4 Experiments**

**4.1 Experimental Setup**

[](#bookmark38)**Datasets** Our evaluation is conducted on three polyp datasets: two publicly available CVC-[ClinicDB1](#bookmark40) [[Bernal *et*](#bookmark38)[*al.*, 2015], Kvasir-SEG2](#bookmark41) [[Debesh *et al.*, 2020] and one inter](#bookmark42)- nal dataset. CVC-ClinicDB contains 612 annotated frames extracted from 29 different colonoscopy sequences. Kvasir- SEG has 1000 images with polyp regions, manually an- notated by an experienced doctor released in 2020. In- ternal dataset consists of 290 static images extracted from OlympiusEurope colonoscopy videos, consisting of 177 pa- tients, annotated and validated by experienced endoscopists. Each frame of the image is accompanied by a morphological category label that is used to instruct the detection of polyps in the image and the ground truth bounding box for evalua- tion. This dataset consists of images with resolutions ranging from 564 × 480 to 600 × 530 pixels.

[](#bookmark39)**Implementation** For the backbone we use VGG16 to ex- [tract feature, which pre-trained on ImageNet [Olga *et al.*,](#bookmark39) [2014] and has some conv layers with max-pooling layer. We](#bookmark39) replace the last max-pooling layer of the model by ROI pool- ing. To reduce errors and noise of candidate regions, CRM is added and the PraNet [[Deng-Ping *et al.*, 2020] is selected](#bookmark43) for the segmentation network, which details are formulated in Section [3.2.](#bookmark34) To accurately classify polyps, we develop SCM including three operations of convolution, global av- erage pooling and softmax introduced in Section [3.3.](#bookmark37)

[1https://polyp.grand-challenge.org/CVCClinicDB/](https://polyp.grand-challenge.org/CVCClinicDB/) [2https://datasets.simula.no/downloads/kvasir-seg.zip](https://datasets.simula.no/downloads/kvasir-seg.zip)

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | Supervision | | CVC-ClinicDB | | | | | Kavsir-SEG | | | | | Private | | | | |
| IoU@10-50 IoU@10 IoU@30 | | | | | IoU@10-50 IoU@10 IoU@30 | | | | | IoU@10-50 IoU@10 IoU@30 | | | | |
| Box Class | | mAP | mAP | CorLoc | mAP | CorLoc | mAP | mAP | CorLoc | mAP | CorLoc | mAP | mAP | CorLoc | mAP | CorLoc |
| Faster R-CNN [[Ren *et al.*, 2017]](#bookmark23) | / | / | 41.08 | 41.22 | 94.38 | 41.22 | 91.01 | 26.16 | 26.56 | 95.56 | 26.22 | 94.58 | 27.59 | 27.59 | 97.14 | 27.59 | 97.14 |
| YOLOv3 [[Redmon and Farhadi, 2018]](#bookmark46) | / | / | 39.94 | 40.54 | 98.68 | 40.30 | 96.05 | 26.88 | 27.61 | 95.34 | 27.22 | 93.02 | 29.42 | 30.23 | 84.21 | 29.60 | 76.32 |
| DiffusionDet50 [[Chen *et al.*, 2022]](#bookmark14) | / | / | 62.08 | 63.41 | 97.10 | 63.41 | 97.10 | 27.66 | 28.59 | 97.41 | 27.94 | 95.69 | 37.40 | 37.82 | 100.00 | 37.59 | 100.00 |
| DiffusionDet500 [[Chen *et al.*, 2022]](#bookmark14) | / | / | 60.89 | 61.88 | 98.46 | 61.88 | 98.46 | 25.84 | 27.17 | 96.00 | 26.83 | 94.40 | 37.83 | 40.06 | 100.00 | 39.28 | 100.00 |
| WSDDN [[Hakan and Andrea, 2016]](#bookmark29) | – | / | 10.94 | 16.81 | 34.63 | 9.33 | 13.56 | 6.14 | 7.92 | 37.28 | 5.79 | 12.95 | 12.26 | 14.65 | 24.43 | 11.61 | 7.77 |
| OICR [[Tang *et al.*, 2017]](#bookmark30) | – | / | 1.68 | 4.30 | 9.40 | 0.37 | 1.88 | – | – | – | – | – | 1.24 | 1.24 | 5.24 | 1.24 | 4.04 |
| WSOD2 [[Zeng *et al.*, 2019]](#bookmark31) |  | / | 29.51 | 37.53 | 58.55 | 25.80 | 36.03 | 1.55 | 1.60 | 10.18 | 1.54 | 7.78 | 1.01 | 1.01 | 3.92 | 1.01 | 3.92 |
| Grad-CAM [[Selvaraju *et al.*, 2017]](#bookmark17) | – | / | 19.65 | 42.11 | 51.46 | 12.91 | 23.97 | 5.60 | 12.41 | 36.56 | 3.63 | 15.57 | 15.84 | 23.79 | 31.19 | 13.29 | 20.18 |
| Grad-CAM++ [[Aditya *et al.*, 2018]](#bookmark47) | – | / | 21.99 | 48.53 | 53.80 | 13.01 | 24.45 | 5.67 | 12.40 | 36.04 | 3.67 | 16.89 | 14.92 | 21.65 | 27.82 | 13.15 | 18.80 |
| MDNet(Ours) | – | / | ***54.28*** | ***74.01*** | ***96.97*** | ***54.83*** | ***73.94*** | ***18.31*** | ***25.92*** | ***92.40*** | ***18.26*** | ***69.20*** | ***37.16*** | ***45.37*** | ***88.75*** | ***36.12*** | ***68.75*** |

Table 1: Quantitative results of different methods on the polyp datasets from CVC-ClinicDB, Kvasir-SEG and internal. The best results of fully supervised and weakly supervised are marked in underline and **bold**, respectively.

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

**Training and Testing** We follow a two-step training strat- egy: 1) the segmentation network is trained with fixed learn- ing rate 10 −4 for 20 epochs. 2) the entire architecture is trained following the end-to-end manner. The MDNet runs for 15 epochs with 10 −5 following 10 epochs with learning rate 10−6. To achieve the filtration of initial proposals and de- tection post-processing, the value of the threshold are simply set,i.e. τ = 0.01, μc = 0.8 , μnms = 0.5, respectively.

**Train and test split** We divided data into two splits on each dataset: training, and test. The training split comprises about 80% of data; test about 20% each.

**4.2 Ablation Experiments**

For the sake of evaluating the potence of our components in the proposed method, we conducted sufficient ablation stud- ies on internal 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. that

The results are shown in Table [2, it demonstrates](#bookmark48) the each discriminator’s contribution to the final performance. When CRM, SCM or DPS is used separately to refine the predic- tion results, it achieve comparable improvement (CRM by ∆+1.08, SCM by ∆+1.36, DPS by ∆+1.59). The second type shows that when two out of the three modules are utilized simultaneously, that is, the mAP has increased by 6.62 (CRM + SCM compared with CRM), 4.86 (CRM + SCM compared with SCM), 4.37 (CRM + DPS compared with CRM), 1.18 (CRM + DPS compared with DPS), 5.53 (SCM + DPS com- pared with SCM), 4.10 (SCM + DPS compared with DPS). Only by using all three modules together the effect can be maximized, which demonstrates each module contribution to the final performance.

In order to give more intuitive results, we graphically con- trast the contributions of CRM in Figure S1[3](#bookmark49), the heat map (w/ or w/o SCM) of outputs in Figure S2[3](#bookmark49) and the prediction results with or without DPS in Figure S3[3](#bookmark49).

**4.3 Comparison With Other Methods**

We select some well-performed detection methods for com- parison. Among them, Faster R-CNN [[Ren *et al.*, 2017]](#bookmark23), [YOLOv](#bookmark14)3 [[Redmon and Farhadi, 2018], DiffusionDet [Chen](#bookmark14)

[3Figure S\* represents the Figure in the supplementary material.](#bookmark14)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Types | Baseline | +CRM | +SCM | +DPS | mAP@0.5 |
| Original | / |  |  |  | 6.17 |
| Single | /  /  / | / | / | / | 12.81↑ 6.64 14.57↑ 8.4 16.00↑ 9.83 |
| Dual | /  /  / | /  / | /  / | /  / | 19.43↑ 13.26 17.18↑ 11.01 20.10↑ 13.93 |
| Triple | / | / | / | / | **29.99**↑ **23.82** |

Table 2: mAP (in %) of different components strategies with the same baseline on internal dataset.

[](#bookmark29)[](#bookmark31)[](#bookmark44)[*et al.*, 2022] are fully supervised methods; WSDDN [Hakan](#bookmark29) [and Andrea, 2016], OICR [Tang *et al.*, 2017], WSOD2 [Zeng](#bookmark31) [*et al.*, 2019], Grad-CAM](#bookmark31) [[Selvaraju *et al.*, 2017]](#bookmark17) and Grad- CAM++ are weakly supervised methods. The results of Faster R-[CNN, YOLOv3, DiffusionDet are given from4](#bookmark50) [[Kai](#bookmark44) [*et al.*, 2019], OICR and WSOD2 are referred from codes5 ,](#bookmark51) the rest are implemented by ourselves with source codes.

**Learning Ability**

**Quantitative results** To validate our model’s learning abil- ity, we evaluate comparison methods on three datasets sepa- rately. We resize images to 224 × 224 pixels for WSDDN, Grad-CAM, Grad-CAM++ and MDNet while keeping the original size for other methods. As shown in Table [1, the pro](#bookmark45)- posed MDNet achieves the SOTA performance in terms of all the evaluation metrics compared with all other weakly super- vised methods on both benchmarks in mAP and CorLoc (cor- [rect localization) [Deselaers *et al.*, 2012]](#bookmark52). Although the fully supervised method perform richly for highly depending on instance annotation, our approach shows comparability with them within the range of 0.67% to 9.35% on average thresh- olds (IoU@10–50). Excitingly, on CVC-ClinicDB dataset, our method achieve the mAP of 74.01% and defeat all other methods on smallest threshold (IoU@10). For detailed results per category we refer to [Table S16 ,](#bookmark53) Table S2[6](#bookmark53) and Table S3[6](#bookmark53). **Qualitative results** To report an visualization results of polyp detection, we present the prediction of different meth- ods as shown in Figure [4. We analyze the results and reveal](#bookmark54)

a

[4https://github.com/open-mmlab/mmdetection](https://github.com/open-mmlab/mmdetection)

[5https://github.com/researchmm/WSOD2](https://github.com/researchmm/WSOD2)

6Table S\* represents the Table in the supplementary material.

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

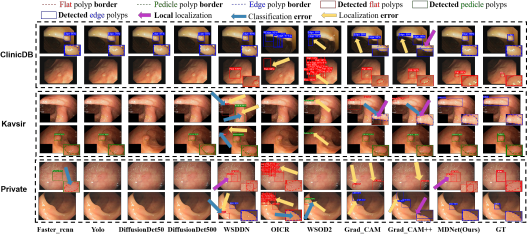


Figure 4: Result visualization of different detection methods on three datasets.

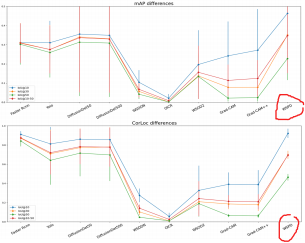


Figure 5: The mAP and CorLoc difference under different IoU thresholds for comparison methods and MDNet.

420 that there are mainly three types of errors in the results. 1) 421 ***Local location***. The predicted bounding box is insufficiently 422 precise,who only contains a portion of the target (purple ar- 423 row). 2) ***Classification error***. In this case, the bounding box 424 has the ability to accurately locate polyp position, yet cannot 425 correctly distinguish the morphology (blue arrow). 3) ***Loca-*** 426 ***tion error***. The failure in these results stems from the fact that 427 the box failed contains the object (yellow arrow).

428 **Generalization Capability**

429 Considering polyp images of different devices in clinical 430 practices, we conduct a cross center generalization experi- 431 ments to test the model’s generalizability. We use CVC- 432 ClinicDB to serve as the visible dataset, while internal dataset 433 as the seen images for testing.

434 To prove the generalization capability of MDNet, we 435 present two main metrics results under different IoU thresh- 436 olds in Table [3. As can be seen, our method have the highest](#bookmark54) 437 values both mAP and CorLoc at each threshold compared to

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | Supervision | | IoU@10-50 | | IoU@10 | | IoU@30 | |
| Box Class | | mAP CorLoc | | mAP CorLoc | | mAP CorLoc | |
| Faster R-CNN [[Ren *et al.*, 2017]](#bookmark23) | / | / | 20.54 | 83.21 | 21.07 | 87.79 | 20.88 | 83.21 |
| YOLOv3 [[Redmon and Farhadi, 2018]](#bookmark46) | / | / | 15.10 | 49.40 | 17.74 | 63.69 | 14.62 | 45.83 |
| DiffusionDet50 [[Chen *et al.*, 2022]](#bookmark14) | / | / | 5.21 | 60.00 | 7.64 | 75.00 | 4.70 | 57.50 |
| DiffusionDet500 [[Chen *et al.*, 2022]](#bookmark14) | / | / | 5.01 | 57.58 | 7.96 | 72.73 | 4.24 | 57.58 |
| WSDDN [[Hakan and Andrea, 2016]](#bookmark29) |  | / | 2.64 | 10.50 | 4.08 | 20.46 | 2.09 | 7.03 |
| OICR [[Tang *et al.*, 2017]](#bookmark30) |  | / |  | 0.61 |  | 1.82 |  |  |
| WSOD2 [[Zeng *et al.*, 2019]](#bookmark31) |  | / | 1.77 | 6.08 | 1.84 | 6.93 | 1.74 | 5.84 |
| Grad-CAM [[Selvaraju *et al.*, 2017]](#bookmark17) | – | / | 3.17 | 14.97 | 6.32 | 26.68 | 2.73 | 13.24 |
| Grad-CAM++ [[Aditya *et al.*, 2018]](#bookmark47) | – | / | 2.97 | 13.13 | 5.77 | 23.82 | 2.67 | 11.30 |
| MDNet(Ours) | – | / | ***15.17*** | ***67.62*** | ***18.65*** | ***87.39*** | ***15.29*** | ***65.62*** |

Table 3: Generalization study of different detection methods on the polyp data from internal dataset.

weakly supervised methods. Furthermore, we calculate the performance differences compared with the test set of CVC- ClinicDB for all methods as shown in Figure [5. The length of](#bookmark54) the error bar indicates the extent of performance change, and the model with shorter bars signify stronger generalization capabilities.

Although our method significantly outperforms other weakly supervised models in mAP, the performance of in- ter class is vast. For analysis, we visualize some success and failure detection results by MDNet, as shown in Figure S4[3](#bookmark49). We can observe that, our method is robust to the size and as- pect of polyps. The main failures are always due to overlarge boxes that not only contain objects, but also include their ad- jacent similar background.

**5 Conclusion**

In this paper, we present a novel MDNet using morphol- ogy category for automatic polyp detection. Specifcally, we design a CRM based on pseudo label to general features for avoiding inappropriate proposals, the SCM to combine low-high spatiality features for distinguishing targets and the DPS to focus on polyps compactly for reducing false posi- tives. Experiments on CVC-ClinicDB, Kvasir-SEG and in- ternal dataset validate the effectiveness of our network.

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

**References**

[Aditya *et al.*, 2018] Chattopadhay Aditya, Sarkar Anirban, Howlader Prantik, and Balasubramanian Vineeth N. Grad- cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Confer- ence on Applications of Computer Vision*, pages 839–847, 2018.

[Bernal *et al.*, 2015] Jorge Bernal, F Javier Snchez, Glo-

ria Fernndez-Esparrach, Debora Gil, Cristina Rodrguez,

and Fernando Vilario. Wm-dova maps for accurate polyp

highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and* *Graphics*, 43:99–111, July 2015.

[Borji *et al.*, 2014] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Com-* *putational Visual Media*, 5:117–150, 2014.

[Borji *et al.*, 2015] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706– 5722, 2015.

[Chen *et al.*, 2022] Shoufa Chen, Pei Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detec- tion. *arXiv preprint arXiv:2211.09788*, 2022.

[Debesh *et al.*, 2020] Ha Debesh, Pia H. Smedsrud, Michael A. Riegler, Halvorsen Pl, De Lange Thomas, Johansen Dag, and Johansen Hvard D. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling*, pages 451–462, 2020.

[Debesh *et al.*, 2021] Jha Debesh, Ali Sharib, Jo- hansen Hvard D., Johansen Dag D., and Halvorsen Pl. Real-time polyp detection, localization and segmen- tation in colonoscopy using deep learning. *IEEE Access*, (99):40496–40510, 2021.

[Deng-Ping *et al.*, 2020] Fan Deng-Ping, Ji Ge-Peng, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical im- age computing and computer assisted intervention*, pages 263–273, 2020.

[Deselaers *et al.*, 2012] Deselaers, Alexe, and Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):275–293, 2012.

[Fang *et al.*, 2021] Yuqi Fang, Delong Zhu, Jianhua Yao, Yixuan Yuan, and Kai-Yu Tong. Abc-net: Area-boundary constraint network with dynamical feature selection for colorectal polyp segmentation. *IEEE Sensors Journal*, 21(10):11799–11809, 2021.

[Ge *et al.*, 2021] Zheng Ge, Songtao Liu, Feng Wang, Zem- ing Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[Haggar and Boushey, 2009] Fatima Haggar and Robin P. Boushey. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 2009.

[Hakan and Andrea, 2016] Bilen Hakan and Vedaldi An- drea. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[Jia *et al.*, 2019] X. Jia, X. Xing, Y. Yuan, L. Xing, and M.Q.H. Meng. 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.

[Jiang *et al.*, 2023a] Wenxi Jiang, Linying Xin, Shefeng Zhu, Zhaoxue Liu, Jiali Wu, Feng Ji, Chaohui Yu, and Zhe Shen. Risk factors related to polyp miss rate of short-term repeated colonoscopy. *Digestive Diseases and Sciences*, 68(5):2040–2049, May 2023.

[Jiang *et al.*, 2023b] Yuncheng Jiang, Zixun Zhang, Ruimao Zhang, Guanbin Li, Shuguang Cui, and Z. Li. Yona: You only need one adjacent reference-frame for accurate and fast video polyp detection. In *International Conference on Medical Image Computing and Computer-Assisted Inter-* *vention*, 2023.

[Jiang *et al.*, 2024] Yuncheng Jiang, Zixun Zhang, Yiwen Hu, Guanbin Li, Xiang Wan, Song Wu, Shuguang Cui, Silin Huang, and Zhen Li. Ecc-polypdet: Enhanced cen- ternet with contrastive learning for automatic polyp detec- tion. *arXiv preprint arXiv:2401.04961*, 2024.

[Jiwoon and Suha, 2018] Ahn Jiwoon and Kwak Suha. Learning pixel-level semantic affinity with image-level su- pervision for weakly supervised semantic segmentation. In

*IEEE Conference on Computer Vision and Pattern Recog-* *nition*, pages 4981–4990, 2018.

[Kai *et al.*, 2019] Chen Kai, Wang Jiaqi, Pang Jiangmiao, Yuhang Cao, Xiong Yu, Li Xiaoxiao, Sun Shuyang, Feng Wansen, Liu Ziwei, Xu Jiarui, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Shi Jianping, Ouyang Wanli, Loy Chen Change, and Lin Dahua. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[Karaman *et al.*, 2023a] Ahmet Karaman, Dervis Karaboga,

Ishak Pacal, Bahriye Akay, Alper Bastrk, zkan Ufuk

Nalbantoglu, Seymanur Coskun, and Omur Sahin. Hyper-

parameter optimization of deep learning architectures us-

ing artificial bee colony (ABC) algorithm for high perfor-

mance real-time automatic colorectal cancer (CRC) polyp

detection. *Biomedical Signal Processing and Control*,

71(12):15603–15620, 2023.

[Karaman *et al.*, 2023b] Ahmet Karaman, Ishak Pacal, Alper

Bastrk, BahriyeAkay, zkan Ufuk Nalbantoglu, Seyma-

nur Coskun, Omur Sahin, and Dervis Karaboga. Robust

real-time polyp detection system design based on YOLO

algorithms by optimizing activation functions and hyper-

parameters with artificial bee colony (ABC). *Expert Sys-*

*tems With Applications*, 221(12):119741, 2023.

[Kwon and Choi, 2021] Jangho Kwon and Kihwan Choi. Weakly supervised attention map training for histological

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

localization of colonoscopy images. In *Annual Interna- tional Conference of the IEEE Engineering in Medicine* *and Biology Society*, pages 3725–3728, 2021.

[Li *et al.*, 2019] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Ernst Jan, and Fu Yun. Guided attention inference net- work. *IEEE transactions on pattern analysis and machine* *intelligence*, 42(12):2996–3010, 2019.

[Li, 2020] Yong Li. Analysis of missed diagnosis rate and related factors of colorectal polyps in colonoscopy. *Gems* *of Health*, (12):260–261, July 2020.

[Mo *et al.*, 2018] XiMo, Ke Tao, Quan Wang, and Guanghui Wang. An efficient approach for polyps detection in endo- scopic videos based on faster r-cnn. pages 3929–3934, 2018.

[Olga *et al.*, 2014] Russakovsky Olga, Jia Deng, Hao Su, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Zhiheng Huang, Karpathy Andrej, Khosla Aditya, and Bernstein Michael. Imagenet large scale visual recognition chal- lenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[Reamaroon *et al.*, 2021] Narathip Reamaroon, Michael W. Sjoding, Jonathan Gryak, Brian D. Athey, Kayvan Najar- ian, and Harm Derksen. Automated detection of acute res- piratory distress syndrome from chest x-rays using direc- tionality measure and deep learning features. *Computers* *in Biology and Medicine*, 134, July 2021.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv* *preprint arXiv:1804.02767*, 2018.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Gir- shick, and Jian Sun. Faster r-cnn: Towards real-time ob- ject detection with region proposal networks. *IEEE Trans- actions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[Siegel *et al.*, 2023] Rebecca L. Siegel, Nikita Sandeep Wa- gle, Andrea Cercek, RobertA. Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2023. *CA:A Cancer Journal* *for Clinicians*, 73(3):233–254, May–June 2023.

[Simonyan and Zisserman, 2015] Karen Simonyan and An- drew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Confer-* *ence on Learning Representations*, 2015.

[Tang *et al.*, 2017] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3059– 3067, 2017.

[van de Sande Koen E. A. *et al.*, 2011] van de Sande Koen E. A., Uijlings Jasper R. R., Gevers Theo, and Smeulders

Arnold W. M. Segmentation as selective search for object recognition. In *International Conference on Computer Vi-* *sion*, pages 1879–1886, 2011.

[Yang *et al.*, 2023] Xiao Yang, Enmin Song, Guangzhi Ma, Yunfeng Zhu, Dongming Yu, Bowen Ding, and Xianyuan Wang. Yolo-ob: An improved anchor-free real-time multi- scale colon polyp detector in colonoscopy. *arXiv preprint* *arXiv:arXiv:2312.08628*, 2023.

[Zeng *et al.*, 2019] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learn- ing bottom-up and top-down objectness distillation for weakly-supervised object detection. In *IEEE International Conference on Computer Vision*, pages 8291–8299, 2019.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, gata

Lapedriza, Aude Oliva, and Antonio Torralba. Learn- ing deep features for discriminative localization. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2015.

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643