基于合作的弱监督息肉检测

**CBNet:Cooperation Based Weakly Supervised Polyp Detection**

## Abstract

漏诊的息肉是导致结直肠癌的主要因素。为了减少误诊的风险，各种自动息肉检测方法已经被开发出来。然而，现有的方法要么依赖于费力的实例级注释，要么需要额外标点，要么缺乏过滤噪声建议和完整检测息肉的能力，这给该领域带来了严峻的挑战。在本文中，我们提出了一种新的基于合作网络(CBNet)，这是一个由类别标签监督的两阶段息肉检测框架，它通过分类与分割的协作来去除错误的建议，并通过聚合自适应多层次区域特征来获得更准确的检测器。具体来说，我们采用基于合作的区域建议网络(CBRPN)，通过删除没有息肉的建议来减少噪声的负面影响，使我们的网络能够捕获息肉特征。此外，为了提高息肉定位的完整性和分类的精度，我们主干网络分类层的指导下聚合了多层次感兴趣区域(ROI)特征，我们称之为自适应ROI融合模块(Adaptive ROI Fusion Module, ARFM)。在公共和私有数据集上进行的大量实验表明，我们的方法达到了弱监督方法的最先进性能，甚至在某些方面优于完全监督方法。所有代码可在https://github.com/dxqllp/CBNet上获得。

Missed polyps are the major factor for colorectal cancer. To minimize the risk of misdiagnosis, variously automatic polyp detection methods have been developed. However, existing methods either rely on laborious instance-level annotations, require labeling of prompt points, or lack the ability to filter noise proposals and detect polyps integrally, resulting in severe challenges in this area. In this paper, we propose a novel Cooperation Based network (CBNet), a two-stage polyp detection framework supervised by category labels that removes wrong proposals through classification in collaboration with segmentation and obtains a more accurate detector by aggregating adaptive multi-level regional features. Specifically, we conduct Cooperation Based Region Proposal Network (CBRPN) to reduce the negative impact of noises by deleting proposals without polyps, enabling our network to capture polyp features. Moreover, to enhance location integrity and classification precision of polyps, we aggregate multi-level region of interest (ROI) features under the guidance of the backbone classification layer, which we called Adaptive ROI Fusion Module (ARFM). Extensive experiments on the public and private datasets show that our method achieves the state-of-the-art performance for weakly supervised methods and even outperforms full supervision in some terms. All code is available at https://github.com/dxqllp/CBNet.

Missed polyps are the major factor for colorectal cancer. To minimize the risk of misdiagnosis, variously automatic polyp detection methods have been developed. However, existing methods either rely on laborious instance-level annotations, lack the ability to filter noise proposals and integrally detect polyps, or require some prompting, resulting in severe challenges in this area. In this paper, we propose the Weakly Supervised Polyp Detection network (CBIA-WSPD), a two-stage framework in collaboration with no-prompt segment anything (SAM) and supervised by image-level annotation that removes negative proposals and adaptively learns region features to obtain a remarkable performance. Specifically, we conduct Cooperation Based Region Proposal Network (CBRPN) to reduce the negative impact of noises by deleting proposals without polyps, enabling our network to capture polyp features. Moreover, to enhance the integrity of detection and improve the accuracy of flat polyps, we aggregate multi-level region of interest (ROI) features under the guidance of the CBRPN classification layer, which we called Adaptive ROI Fusion Module (ARFM). Extensive experiments on the public and private datasets show that our method achieves the state-of-the-art performance for weakly supervised methods and even outperforms full supervision in some terms. All code is available at https://github.com/dxqllp/CBIA\_WSPD.

## Introduction

结直肠癌(Colorectal cancer, CRC)是第三大最常见的癌症，其中80%以上起源于息肉[1]。尽管息肉在早期是良性的，但未经治疗可能会变成恶性的危及生命的癌症[2]。在临床实践中，结肠镜检查被认为是发现和切除息肉的黄金标准。然而，该方法高度依赖于医生的经验和熟练程度，由于内镜医师长期值班后操作不规范和疏忽，漏检率高(高达30%)[3]。**幸运的是，计算机辅助检测(CAD)已被证明可以帮助医生发现息肉并减少漏诊率**[4-5]。

Colorectal cancer (CRC) ranks the third most frequent cancer, which more than 80% originates from polyps[1]. Untreated polyps might become malignant and life-threatening cancer [2] although they are benign in the early stage. In clinical practice, colonoscopy procedure is regarded as the golden standard to detect and remove polyps. However, the method highly depends on the experience and proficiency of the medical practitioner that suffers a high miss rate (as much as 30%) due to the irregular operation and negligence of endoscopists after long duty[3]. Fortunately, computer-aided detection(CAD) has been proven to help doctors detect polyps and reduce the incidence of missed diagnoses[4-5].

尽管CAD在息肉检测方面表现出了令人印象深刻的能力[6-12]，但它们需要与具有实例级边界框注释的数据集相结合，由于结肠图像的复杂性和息肉的多样性，这是非常费力和具有挑战性的。为了减轻标注负担，研究人员希望使检测器以弱监督(WS)的方式工作，例如:学习一个只有类别标签而不是边界框的检测器。最近，一些有前景的作品将MIL与深度学习相结合的[17-19]极大地推动了自然图像的边界，成功地将WS应用于医学领域。不幸的是，准确可靠的息肉检测很容易被扁平息肉所欺骗，因为扁平息肉缺乏清晰可见的边界，并且与周围的结直肠表现出相似的方面，这使得网络很难准确地从背景中区分息肉。此外，这些研究大多直接使用标准方法生成的粗建议进行训练，但大多数是不包含目标的负例(噪声)，这影响了目标特征的学习，降低了检测的性能。更糟糕的是，它们严重依赖区域级分类器，这些分类器过度关注最具歧视性的局部区域，这不仅降低了分类精度，而且导致了过度拟合，如图1 (a)左所示。更近地，其他一些备受瞩目的研究希望通过迁移学习来完成这一任务，如借助Segment Anything Model (SAM)实现弱监督分割。但它本质上是一种基于提示的方法，如果没有额外的点提示，它将失去准确识别息肉的能力，如图1 (a)右所示。附加点不仅增加了标注的负担，而且偏离了仅使用类别标注的要求。因此，需要进一步的机制来进行更合理的推理。

Although CAD has demonstrated impressive capabilities in polyp detection [6-12], they need to be combined with datasets with instance -level bounding box annotations, which is very laborious and challenging due to the complexity of colonic images and the diversity of polyps. In order to reduce the labeling burden, researchers hope to make detectors work in a weakly supervised (WS) fashion, e.g. learning a detector with only category labels rather than bounding boxes. Recently, several promising works combining multiple instance learning (MIL) with deep learning [17-19] have greatly pushed the boundaries of natural images to successfully apply WS to the medical field. Unfortunately, accurate and reliable polyp detection can be easily fooled by flat polyps because they lack clearly visible borders and exhibits a similar aspect to the surrounding colorectum, which makes networks struggle to accurately distinguish polyps from backgrounds. In addition, most of these studies directly use coarse proposals generated by standard methods for training, yet most of them are negative cases (noise) that do not contain the target, which affects the learning of the target features while lowering the performance of the detection. Worse still, they severely rely on region-level classifiers that overly focus on the most discriminatory local regions, which not only reduces classification accuracy but also results in over-fitting, as shown in Figure 1 (a) left. More recently, other high-profile works hope to solve this task through transfer learning, such as weakly supervised polyp segmentation (WSPS) with the help of the Segment Anything Model (SAM). But it is essentially a promptable method, which will lose the ability to accurately identify polyps if there are no additional point annotations as shown in Figure 1 (a) right. Additional points not only increases the burden of annotation but also deviates from the requirements of only category labels. Therefore, a further mechanism is necessary for more reasonable inference.

通过深入研究，我们发现基于多例的方法和基于sam的方法具有互补的优点和缺点。具体而言，分类器倾向于关注对象与SAM中需求提示相匹配，因为它可以根据关注的区域自动生成点提示，从而消除了额外点注释的负担。同时，由点提示产生的高质量的息肉掩模不仅可以保证目标的完整性，避免过拟合，还可以过滤提案中的噪声，使网络更好地学习息肉特征，获得鼓舞人心的性能。总而言之，他们不应该单打独斗，而应该齐心协力克服各自的内在弱点。

Through deeper study we found that the multi-example based and SAM-aided approaches have complementary advantages and disadvantages. Specifically, the classifier tends to focus on parts of the object that matches the need for prompts in SAM, because it can automatically generate point prompt based on the focused region, eliminating the burden for additional point annotations.Meanwhile, the high-quality polyp mask produced from point prompts can not only ensure the integrity of the target to avoid over-fitting, but also filter the noise in the proposal so that the network can better learn polyp features and gain inspiring performance. In a word, instead of working separately, they should work together to overcome their internal weaknesses.

在这项工作中，我们提出了CBNet，如图1 (b)所示，这是一个弱监督息肉检测模型，旨在解决以下挑战:(1)不完整息肉检测。(2)无法过滤的噪声。(3)没有点标注的息肉掩模失败。(4)扁平息肉。该方法主要包括两个创新模块:基于合作的区域建议网络(CBRPN)、自适应ROI融合模块(ARFM)一个通用模块:多实例检测网络(MIDN)。具体地说，受互补优势和劣势的启发，我们设计了CBRPN来主要解决(2)~(3)的挑战同时初步解决挑战1，它包括提案生成器和SAM。前者通过基于图的分割和合并策略生成一系列粗略的建议。后者根据点提示(从主干的梯度矩阵派生)生成伪掩码。然后逐个计算掩模和提案的交并比(IOU)，留下相对完整包含息肉的提案，并去除不含息肉的噪声。此外，我们使用ARFM来解决挑战4并进一步解决挑战1，他由两部分构成：主干网络分类层，融合模块。其中分类层在训练该模块时参数是冻结的以保证主干网络能从图像的角度关注息肉进一步保证目标的完整性并从全局的角度对它的形态进行判断提高分类准确度。在融合部分，息肉和背景之间的差异将会被细化通过结合深-浅层的特征来细化提高检测扁平息肉的能力。该模块自适应调整第二阶段学习到的特征，以得到更全面的全局-局部特征。最后，我们引入MIDN来设定提议区域的目标存在性和类别可能性得分，从而在图像类别标签的监督下实现弱监督检测。

In this work, we propose the CBNet, as shown in figure 1 (b), a weakly supervised polyp detection (WSPD) model designed to solve the challenges: (1) non-entire polyp detection. (2) unfilterable noise. (3) failed polyp mask without point annotation. (4) flat polyp. Our approach mainly includes two innovative module: Cooperation Based Region Proposal Network (CBRPN), Adaptive ROI Fusion Module (ARFM) and a common module: Multiple Instance Detection Network (MIDN). Specifically, inspired by the complementary strengths and weaknesses, we designed the CBRPN mainly for challenges (2) ~ (3) while initially addressing challenge (1), which includes a proposals generator and SAM. The former generates a series of rough proposals by graph-based segmentation and merging strategies. The latter generates pseudo masks based on point prompts (derived from the gradient matrix of the backbone). After that masks and proposals are computed intersection over union (IOU) one by one to leave proposals that contain polyps relatively entire and remove the noise without polyps. Furthermore, we employ ARFM to solve challenge (4) and further address challenge (1), which consists of two parts: classification layer and ROI fusion. During the training of the module, the parameters of the classification layer are frozen to ensure that the backbone network can focus on polyps from the whole image for further completeness, and judge the morphology of them from the global features to improve classification accuracy. In the fusion part, the differences between polyps and background will be refined through combining deep-shallow region features to improve the detection of flat polyps. This module adaptively adjusts features learned in the second stage, resulting in a more comprehensive global-local feature. Finally, we introduce MIDN for setting the target existence and category possibility scores of the proposal region to achieve weakly supervised detection under the supervision of image category labels.

总之，我们的主要贡献有四个方面：

* 我们提出了CBNet框架，一个采用协作机制的框架，在只有图像级注释的弱监督环境中完成对息肉的检测。
* 在提案生成阶段，我们通过分析了分类器和SAM特性，证实了的协作作用并据此设计了cooperation based region proposal network，实现了提案噪声过滤减轻了网络学习息肉特征的压力，同时移除了定位不完整的提案降低了过拟合的风险。
* 为了更好地捕捉平面息肉，进一步保证位置完整性，我们设计了自适应ROI融合模块，从全局-局部层面学习息肉特征，从深浅层面细化息肉与背景之间的细微差异。
* 我们的方法在三个数据集(即ClinicDB, Kvasir, private)上进行，不仅在弱监督方法中获得了最先进的性能，而且在某些方面超过了完全监督的性能。

In summary, our main contributions are listed four-fold:

* We propose the CBNet, a framework that employ collaborative mechanism to to detect polyps in the weakly supervised setting with only image-level annotation.
* In the proposals generation stage, we confirmed the collaborative nature of the classifier and SAM. Accordingly, we conducted cooperation based region proposal network (CBRPN), which implements proposal noise filtering to ensure the correctness of feature extraction, and removes the proposal with incomplete polyp to reduce the risk of over-fitting.
* To better capture the flat polyps and further ensure positional integrity, we design the adaptive ROI fusion module (ARFM) to learn polyp features from global-local level and refine slight disparities between polyps & background from deep-shallow level.
* Our method is conducted on three datasets (i.e. ClinicDB, Kvasir, private), which not only obtains state-of-the art performance in weakly supervised methods, but also exceeds fully supervised performance in some respects.

## **Related Work**

Region Proposal Generation.

区域提议生成方法可以分为传统方法和基于cnn的方法，传统方法又可以分为基于边缘的方法[26,31]和基于超像素的方法[27,28]。基于边缘的方法通过图像边缘评估窗口框以优化其位置，这在处理边界模糊的息肉时很困难。基于超像素的方法如选择性搜索(SS)可以适应边界模糊的息肉，但在复杂多变的息肉中存在较高的虚警率。由于息肉与背景的相似性，这些方法的局部性导致误报率高，检出率低。基于cnn的方法，如区域提议网络(Region Proposal Network, RPN)[32]已经被提出来优化和过滤提议以更快地检测，并在该领域取得了很大进展。然而，为了保证高性能，它需要边界框标注并设置大量的超参数进行训练，这偏离了只有图像级标注的弱监督要求。与之不同的是，我们探索了一种新的提案生成方法，提出了一种新的CBRPN，它可以同时利用SS和RPN的优点，并在同一网络中显示出更好的效果。

Region proposal generation methods could be categorized into traditional methods and CNN-based ones where the former category can be further divided into two approachs, i.e., edge-based[26, 31] and superpixel-based[27, 28]. Edge-based methods evaluate window boxes by image edges to refine their location, which struggle in dealing with boundary blurred polyps. Superpixel-based methods such as Selective Search (SS), can adapt to polyps with blurred boundaries, but has a high false alarm rate in complex polyps with various variations. And due to the similarity between polyp and background, the locality of these methods lead to high false alarms and low detection rates. CNN-based methods such as Region Proposal Network (RPN) [32]have been proposed to optimize and filter proposals for faster detection and made a great progress in this area. Nevertheless, to ensure the high performance, it requires bounding box annotations and sets a large number of hyperparameters for training, which deviates from the weak supervision requirement that only image-level annotations are available (see supplementary material for more information). Different from them, we explore a new ways to generate proposals and propose a novel CBRPN, it can takes the advantages of both SS and RPN and shows a better result in the same network.

Weakly Supervised Cooperation Detection

由于缺少实例级注释，弱监督检测方法很容易在对象部分上过拟合。为了解决这一问题，许多基于合作的工作被提出，如C-MIDN[33]和P-MIDN[34]。尽管取得了令人满意的结果，但过度拟合仍然是一项具有挑战性的任务，因为协作的分割网络明确地针对特定领域进行了跟踪，当应用于不同类型的成像数据时，它们的性能会显著下降。近年来，源自自然图像处理领域的SAM备受关注，并因其出色的泛化能力被引入医学领域[9,24,25]。但是，在仅使用类别注释的WSPD中应用SAM的问题还没有研究过。为了实现同样的目标，我们提出了新的CBRPN，减少了过度拟合的可能性，并成功地将SAM用于类别监督的WSPS。同时，ARFM在更好地检测扁平息肉的同时，也进一步避免了过拟合。两者的复杂性和特殊性都较低，但性能较好。

Due to the absence of instance-level annotations, weakly supervised detection methods are easy to over-fit on object parts. To address this issue, many cooperation-based works such as C-MIDN[33] and P-MIDN[34], have been proposed. Despite achieving promising results, over-fitting is still a challenging task because segmentation networks in collaboration are explicitly tailed for the specific domain and their performance can degrade significantly when applied to different types of imaging data. Recently, SAM which originated from the field of natural image processing attracted a lot of attention, and was introduced into the medical area [9,24,25], owing to the excellent ability of generality. However, it is still unexplored to employ SAM in weakly supervised polyp segmentation only with category annotations. Towards the same goal, we propose the novel CBRPN reduces the possibility of over-fitting and successfully employs SAM for category-supervised WSPS. Meanwhile ARFM is also designed to further avoid over-fitting while better detecting flat polyps. Both of them have fewer complexity and particularity while performing better.

MedSAM在包含超过一百万对医学图像-掩模对的前所未有的数据集上微调，以提升SAM在医学图像分割领域的适用性。由于SAM本质上是一种可提示的分割方法，需要使用点或边界框来指定分割目标，所以MedSAM为每张训练数据都提供了bounding box作为输入提示。然而结果表明，当边界框提示不明确时，分割将会出现困难（例如眼底血管分割）。

**Method**

Overview

我们网络的整体架构如图2所示。提出的CBNet在训练和测试中包含两个阶段，其中第一阶段是用于提案生成的区域提案网络(CBRPN)，第二阶段是用于息肉检测的WS网络(ARFM & MIDN)。对于输入图像I，将其同时馈入SSW、骨干和SAM，分别用于生成粗糙的窗口框BSSW，获得息肉中最显著点(x, y)，以及产生用于过滤的框BSAM(根据伪掩码)。之后，将BSSW和BSAM发送给ITF去噪，保留的提案形成BCBRPN。然后来自第一阶段的F1-F5特征和细化的提案被ARFM处理，它包含三个依次堆叠跨层增强块和紧随其后的融合操作以得到Xfusion。此外，我们还加入了参数冻固定的卷积层和全局平均池化层来提高分类的准确性通过Ximg。最后，Xfusion和Ximg进一步通过MIDN和常见的softmax给提案设置分数，并通过非最大抑制获得检测结果。

The overall architecture of our network is shown in Figure 2. The proposed CBNet consists of two stages during training and testing, where the first stage is the region proposal network (CBRPN, see Section 3.3) for proposal generation and the second stage is a WS network (ARFM, see Section 3.4 & MIDN, see Section 3.5) of polyp detection. For an input image I ∈RH×W×C（height, width, channels）, it is simultaneously fed into SSW, pre-trained backbone (see Section 3.1) and SAM for generating a coarse set of window boxes BSSW, obtaining the most significant point (point\_col , point\_row) in the polyp, and producing the other box set BSAM (according to the pseudo-mask) for filtering, respectively. After that, BSSW andBSAM will be sent to iou threshold filter to remove some noise, and the retained proposals form BCBRPN. Then, features F1-F5 and refined proposals BCBRPN from the first stage are processed by ARFM, which contains three cross-layer augmentation blocks stacked sequentially and the following fusion operation to get Xfusion. In addition, a convolution layer with fixed parameters Conv\_P and the global average pooling are added in parallel for Ximg to enable the accuracy of classification. Finally, the features Xfusion andXimg further go through the MIDN and a common softmax to set proposal scores and obtain the detection results through non-maximum suppression.

For more clarity, the training and the testing of our CBNet are summarized in Algorithm 1.

Pre-trained Backbone

We build our method on a pre-trained CNN that has been pre-trained on the ImageNet and fine-tunes it on polyp data with only image-level supervision (i.e. no bounding box annotations) to get features F={F1,F2, ... ,FN} (where N is the number of feature blocks) and conv\_P. We give details of usage in other sections.

Cooperation-Based RPN

由于弱监督息肉检测所需解决的问题与多实例学习研究目标相吻合，通常它被视为多实例学习问题来处理that需要为训练数据生成区域候选框。基于超像素的方法由于能适应边缘模糊的目标且不需要任何注释通常被用来生成初始的候选框。然而他们是粗糙的因为其中有很多不或仅包含息肉极少部分的噪声。幸运的是，SAM能为息肉生成准确的伪标签来优化提案，但需要要额外的点注释提示。在分类任务中，像素会根据与目标类的相关性做出响应，所以目标物体中必有一个响应最高的像素，这一点的坐标恰好可以作为SAM的提示。因此，我们尝试将二者的优势结合起来解决提案噪声和缺失点注释时SAM失效的问题，同时减少不完整息肉提案的数量在训练前降低过拟合风险。

Since the goals consistency between weakly supervised polyp detection and multi-instance learning (MIL), it is usually treated as MIL that requires giving proposals to the training data. Superpixel-based SSW is commonly used to generate the initial candidate boxes due to their ability to adapt to targets with blurred edges and do not require any annotation. However, they are rough because many of them have only the background or contain only a small percentage of polyps. Fortunately, SAM can generate accurate pseudo-labels for polyps to optimise the proposal, but requires additional point prompts. In the classification task, pixels will respond according to their relevance with the target class, so there must be a pixel with the highest response in the object, and the coordinates of this point can exactly be used as a prompt for SAM. Therefore, we try to take advantages of both SSW and SAM to filter background noise as well as reduce incomplete polyp proposals before training for reducing the over-fitting risk.

In detail, the CBRPN takes the image I and feature F1-F5 as input, generates an initial set proposals BSSW∈RB1×4 (B1: the number of proposals from SSW, 4: coordinates of each proposal for top, left, right and bottom)throughSSW, and produces the coordinate (point\_col, point\_row) for point according to the predicted category, formulated as below:

point\_row , point\_col= (1)



where point\_col & point\_row∈R, BSSW, Fk & wck denote the kth feature map and their weights corresponding to class c, fSSW is fast setting selective search windows, argmax (·) is the max operation, funrval\_index are functions to find the max value and the coordinate of it, respectively. wck is defined as：



(2)

Where Z represents the number of pixels in the feature map, is the derivation operation. yc and Fkij are the gradient of the cth score and the pixel value of the Kth feature map at coordinates (i , j), respectively.

Since (point\_row , point\_col) can be regarded as the input prompt of the SAM, thereby another set of candidate boxes for filtering can be defined as:

(3)



where BSAM∈RB2×4 (B2:number of proposals from SAM, 4:coordinates of each proposal for top, left, right and bottom). fSAM is the segment anything model with pre-training weights loaded. Thus, the significant noise can be filtered by ITF. The processes and rules can be formulated as:



(4)

where BCBRPN∈RB×4 (B:number of proposals after filtering, 4:coordinates of each proposal for top, left, right and bottom), mi & wj are the proposals from BSAM and BSSW, respectively. ∩is the intersection of two sets, ||, || and || represent the number of pixels in mi∩mj, mi and mj, respectively.τis a threshold for the judgement of trade-offs, and K denotes the stack of proposals. Since the classification network may learn data bias[35] during training, the prompt point may appear in the background, which cause BCBRPN to be empty. When this occurs, Eq. 5 can be re-written as BCBRPN=BSSW.

By calculating the IOU between BSSW and BSAM and then removing proposals with low overlap through the ITF, the BCBRPN can suppress noise without polyps while strengthening the target learning by reducing false positives proposals.

Adaptive ROI Fusion Module

The ROI feature in existing WS network typically directly generates by the last convolution layer, then send it to region softmax to calculate probability score. However, the features of this layer are limited and inaccurate, shown in Figure 2 (a) 2th row Conv\_5, which is inappropriate for medical images, e.g., the polyp images with a typical resolution of 224×224 in the morphological classification task. In addition, the flat polyp is usually small and constant down-sampling will drown it in the background, and it is often ignored due to the unclear boundary. To address the above problem, we develop an adaptive ROI fusion module (ARFM) inspired by the path aggregation mechanism (Refine-FPN) [36]. Specifically, ARFM has an agg and a fusion structure with detailed configuration inside as shown in Figure 2 (b)、(c). We use {F2, F3, F4, F5} to denote feature levels generated by backbone. The augmented mechanism starts from the lowest level F2 and gradually approaches F5. From F2 toF5, the spatial size is gradually being halved. We use {F^2, F^3, F^4, F^5} to denote newly produced feature maps corresponding to {F2, F3, F4, F5}. Note that F^2=F2, without any processing. The calculation process are as follows:

(5)



where F^i ∈Rb×C’×H’×W’ (i={2, 3, 4, 5} denote ith aggregation feature map, b is the batch\_size, C’={128, 256, 512, 512}, H’ ={H/2, H/4, H/8, H/16}, W’={W/2, W/4, W/8, W/16}), convks represents the convolution operation with specific kernel k and stride s.

Further, we pool features from all levels for proposals to the same size (7×7) as:



(6)

where R-i ∈Rb×128×7×7( b: batch\_size ), Ri is the ith region feature cropped according to the proposal from BCBRPN, roi\_align is an adaptive pooling operation to adjust Ri to a uniform size 7×7. Then, they are fused for following prediction. Considering the influence of position information on subsequent modules, we design different fusion strategies: cnn-based and vector-based, detailed formulas are described as:



(7)

where XfusionC, XfusionR ∈RB×C (B: the number of proposals from BCBRPN, C: the number of categories), GAP is the global average pooling (GAP), FC is the fully connected layer, convks follow the same definition as Eq. 6.

Since ROI features only contain local region of the image, which reduces classification accuracy and leads to over-fitting. To overcome this problem, we add the specific classification layer from the backbone to this module in parallel for global image feature Ximg, which is calculated according to:

(8)



where ximg ∈RC (C is the number of morphological categories), Conv\_Pdenotes the convolution that loads and freezes the pre-trained weights.

Multiple Instance Detection Network

Current WS object detection methods usually choose WSDDN as the criterion, which solves the problem that maps proposals scores at the instance level to category labels at the image level. Following the same double branch, we build multiple instance detection network (MIDN) to set scores for proposals. Specifically, MIDN consists of a location branch and a classification branch, the former selects which proposal region is more likely to contain the entire polyp fragment while the latter predicts which class to associate to the proposal region . Hence, the different scores of proposals can be represented:



(9)

where SMIC，SMIR ∈[0,1]B\*C , softmaxC and softmaxR are both softmax operation that is responsible for mapping the feature matrix to the category dimension (dim=1) and the proposal dimension (dim=0), respectively. Finally, these score vectors are element-wise producted and added to obtain the image-level classification scores. The rule can be formulated as:



(10)

where SMIC\*R∈[0,1]B\*C, SMI∈[0,1]C. Additionally, the feature Ximg of the whole image also be feed to a softmax to get the image-level Simg ∈[0,1]C:

(11)



Loss Function

**BCE Loss**: Binary Cross Entropy Loss is a measure used to evaluate the distance between the prediction and category label.We employ it as the loss function to train our network:



(12)

where N indicates the number of predicted object groups, yi is the value of ith category in the one-hot label, p (Simgi) and p (SMIi) are the probabilities belong to ith class predicted by model.

**Experiment**

Datasets

**ClinicDB [38]:** The dataset comprises 612 images sourced from 29 colonoscopy video sequences, each with a resolution of 288×384. It was developed in partnership with the Hospital Clinic of Barcelona, Spain, and consists of 322 flat, 282 pedicle, and 66 edge polyps.

**Kvasir [37]:**This dataset was assembled by Vestre Viken Health Trust in Norway which includes 1,000 polyp images along with their ground truth annotations from colonoscopy videos. The bounding boxes for ground truth were initially marked by medical doctors and later confirmed by experienced gastroenterologists. The images span a range of resolutions from 332 × 487 to 1920 × 1072 pixels. The number of three morphologies is 255, 851 and 53, respectively.

**Private:**Our internal dataset extract 290 static image from OlympiusEurope colonoscopy videos, which consisting of 177 patients, annotated and validated by experienced endoscopists. Each frame in the dataset is paired with a morphological category label (87, 158, 45) to guide the training, along with the ground truth bounding box for evaluation. The images vary in resolution, spanning from 564 × 480 to 600 × 530 pixels.

Implementation Details

Training Details

The slightly modified classifier layer of VGG16 serves as our backbone. Specifically, we replace the classifier of three FC operations with a GAP layer and a reshape after the original last feature layer. In the first stage, we only train the backbone for a total of 25 epochs with a learning rate of 1e-4 and a batch size of 32. In the second stage, we train the CBNet for a weight decay of 5e-4, a batch size of 1, and 15 epochs with learning rate 1e-5 following 5 epochs with 1e-6. As discussed in Eq 4, we empirically set τ as 0.5. We divided data into two splits on each dataset: training, and test. The training split comprises about 80% of data; test about 20% each.

Evaluation metrics

To evaluate detection performance, we employ three performance measures. The first one follows the standard PASCAL VOC protocol, calculate average precision (AP) and mean AP (mAP) at IOU thresholds of 10%, 30%, 50% between the detected boxes and the ground truth ones. Additionally,we report CorLoc for location, a commonly-used weakly supervised detection measure [39], which means percentage of images where at least one instance of the target object class is correctly localized with the most confident detected bounding box overlapping at least 50%. As for classification, we use the standard PASCAL VOC protocol and report accuracy for different iou thresholds.

Quantitative Results

**Main result:** As shown in Table 1, traditional weakly supervised methods have a limited ability in dealing with challenging polyps, thereby having much worse scores than fully supervised methods and ours. Although fully supervised methods have a better performance on Kvasir, they could not well detect small targets, resulting in inaccurate box predictions, e.g., lower mAP on other datasets. Besides, the high CorLoc but low ACC for them indicates that they are inefficient in learning discriminative category representation, leading to false detections or missed detections. Compared with them, our proposed CBNet offers higher mAP despite having slightly low CorLoc due to the lack of instance-level annotations, which show that CBNet can learn a better target representation to deal with the flat polyps and background noise issue. More detail for each category can been in Table.......

**ROC results:**We also plot the ROC curves of different methods on the ClinicDB dataset and private dataset for flat polyps (other categories in 附件) in Figure 5.

As can be seen, the performance of our CBNet is significantly better than all other methods, where the area under the ROC curve (AUC) of our CBNet is much larger than those of both the fully supervised methods and weakly supervised methods, e.g. xxxx AUC of xxx v.s. xxxx AUC of xxx on the xxx dataset.

Visual Results

In Figure 6, we present some visual detection results of different methods. As we can see, the second test image contains a flat polyp, all the methods including both fully supervised and weakly methods falsely detect the polyp as a target except our CBNet, showing its excellent representation ability in complex flat polyps. We attribute this success to the unique structure of our ARFM, i.e., it aggregating different level features such that it has the ability to provide more information for following modules and recognize polyp as target rather than background.

To gain more in-depth in-sight into what backbone has learned, we also visualize the gradient-weighted class activation mapping (Grad-CAM), and as in Figure 7. It can be observed that the baseline has the ability to successfully locate polyps, but only focus on a small part of the target and may occasionally fail e.g., the third row, the fifth row, and the last row.

Ablation Study

Considering that the detection of backbone is CAM-based and has a nature performance gap (demonstrated in impact of MIDN) with the MIL-based method, we selected WSDDN as the reference in order to ensure the fairness of comparison.

**Impact of CBRPN:** To investigate the impact of the proposed CBRPN, we conducted ablation studies by use SAM's proposal as a result, directly using proposals from SAM and filtering them based on area before using. The results are summarized in Table 2. As can be seen, compared to other proposal generation methods, our CBRPN achieves higher mAP. And the lower flat AP is due to insufficient learning ability in the subsequent modules, as demonstrated in the ablation experiments of the ARFM module. More detailed analysis are in 附件。

In addition, we also compared the average number of proposals used for training and the average overlap with ground truth boxes. The results in Table 3 show that during training, CBRPN minimises the number of proposals and maximises average overlap with ground truth.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| proposal source | mAP | Flat AP | Pedicle AP | Edge AP |
| SSW | 4.71 | 2.79 | 11.36 | - |
| SAM | - | - | - | - |
| SAM(filter) | - | - | - | - |
| SAM&SSW | 10.38 | 1 | 30.16 | - |
| SAM(filter)&SSW | 10.84 | 0.18 | 26.45 | 5.9 |

|  |  |  |
| --- | --- | --- |
| proposal source | mean number | mean overlap |
| SSW |  |  |
| SAM |  |  |
| SAM&SSW |  |  |
| SAM(filter)&SSW |  |  |

**Impact of the design of the ARFM:**

We investigate the effect of ARFM on the original reference network as well as different proposal filtering strategies. The results on ClinicDB and Kvasir are shown in Table 4. Comparing strategy 1 v.s. 2, we can find that the addition of AFFM contributes to a gain of about 28.85% mAP for ClinicDB and 2.47% mAP for Kvasir, as well as increasing the AP of each morphology in different degree.

Again, from the 3th strategy to 4th strategy ARFM outperforms 1th & 2th strategy in all metrics. Finally, comparing strategies 3 and 4 we find that SAM self-filtering based on area can further improving performance on Kvasir, but is negative on ClinicDB. The probable reason for this is that the latter has many small polyps and the inappropriate area threshold filters them out. In a word, the proposed ARFM can help network achieves better performance in all metrics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| strategy | dataset | mAP | flat AP | Pedicle AP | Edge AP |
| 1. SSW | ClinicDB | 6.69 | 5.65 | 11.45 | 2.99 |
| Kvasir | 4.71 | 2.79 | 11.36 | - |
| 1. SSW+ARFM | ClinicDB | 35.54 | 50.37 | 20.68 | 35.38 |
| Kvasir | 7.18 | 4.88 | 16.65 | - |
| 1. SAM&SSW+ARFM | ClinicDB | 51.94 | 77.29 | 42.97 | 35.58 |
| Kvasir | 11.02 | 6.58 | 26.48 | - |
| 1. SAM(filter)&SSW+ARFM | ClinicDB | 49.17 | 63.88 | 45.15 | 38.46 |
| Kvasir | 13.44 | 14.25 | 18.33 | 7.74 |

**Impact of MIDN:** To demonstrate the importance of multiple instance learning for weakly supervised polyp detection, we compare MIL-based WSDDN with the CAM-based baseline, and the results are shown in Table 5. The lower mAP indicates that CAM-based backbone has insufficient learning ability, which is the reason why we chose WSDDN as the reference in our ablation experiments.

|  |  |
| --- | --- |
| method | mAP |
| baseline |  |
| WSDDN |  |

**Conclusion**

In this paper, we propose a novel CBNet for image-level weakly supervised polyp detection. CBNet adopts CBRPN to automatically generate point prompts for SAM taking advantage of the local properties in the classification and add iou filter to improving the quality of proposals. We also propose the ARFM to enhance the region feature, which further helps to detect flat polyps and avoid over-fitting. Experiments on public datasets and internal demonstrate the superiority of our CBNet.

Although CAD has demonstrated impressive capabilities in polyp detection [6-12], they need to be combined with datasets with instance -level bounding box annotations, which is very laborious and challenging due to the complexity of colonic images and the diversity of polyps. ~~as shown in Fig. 2 (a).~~ In order to reduce the labeling burden, researchers hope to make detectors work in a weakly supervised (WS) fashion，e.g. learning a detector with only category labels rather than bounding boxes. Recently, several promising works combining MIL with deep learning [17-19] have greatly pushed the boundaries of natural images to successfully apply WS to the medical field.

通过**迁移学习**[9]来利用从大规模自然图像数据集中学习到的知识来解决息肉检测问题，或者例如在息肉数据集上微调通用的分割模型，或者学习仅使用类别标签而不使用边界框的检测器。

Although CAD has demonstrated impressive capabilities in polyp detection and segmentation[6-12],they need to be combined with datasets with instance -level bounding box annotations, which is very laborious and challenging due to the complexity of colonic images and the diversity of polyps. ~~as shown in Fig. 2 (a).~~

In order to reduce the labeling burden, researchers hope to solve the polyp detection problem by leveraging the knowledge learned from large-scale natural image datasets through transfer learning[9] or make detectors work in a weakly supervised fashion, e.g. fine-tuning the general segmentation model on polyp datasets or learning a detector with only category labels rather than bounding boxes.

最近，**SAM**已经成为视觉基础模型中最引人注目的工作，因为它展示了出色的性能和跨不同任务的通用性。一些有前途的将开箱即用模型应用于息肉检测任务[9,24,25]。然而，如文献[24]所述，SAM本质上是一种**提示性**方法，因此**不适用**于息肉检测。具体来说，虽然SAM可以从框提示生成高质量的息肉掩模，但如果只有点提示或根本没有提示，它将失去准确识别息肉的能力。在非提示设置下，可以看出SAM在息肉分割任务中的表现并不好。因此，如何有效地将SAM应用到息肉分割中还需要进一步探索。

Recently, SAM has been the most high-profile work in vision foundation models because it demonstrates outstanding performance and versatility across diverse tasks. Several promising applied the out-of-the-box model to polyp detection task[9,24,25]. Nevertheless, as noted in [24] that SAM is essentially a promptable method and thus not applicable in polyp detection. Specifically, while SAM can produce high-quality polyp masks from box prompts, it will lose the ability to accurately identify polyps if there are only point prompts, as shown in Figure 1 (a) right.

由于实例级监督的不可用，普遍的工作将其视为一个多示例学习(MIL)问题来处理，这需要预先生成候选框。在众多研究中，选择性搜索[14]和边缘盒[15]是创建候选提案的标准方法。然而，后来的研究认为他们大多数是负例，这不仅影响了检测效果，而且降低了预测速度[13]。考虑到这一点，他们设计了区域提议网络(RPN)[16]. 确实，与传统方法相比，它们在速度上有了显著的提高。不幸的是，这种策略需要大量的先验参数，如尺度和比例。更糟糕的是，这些参数的有效性在很大程度上控制了建议的质量，进而影响检测结果。我们统计了两个公开的以及一个私人息肉数据集的相关信息。如图2 (b)-(c)所示，不同数据集的息肉大小和box宽高比的范围变化很大。这些差异对参数的设置提出了严峻的挑战，如果只设置高频参数会导致召回率降低造成漏检；如果设置所有参数，则会导致大量冗余的候选框和更复杂的计算。此外，这些参数是动态的，因为它们与结肠镜和息肉之间的距离密切相关。进一步地，为了确保RPN的有效性，网络训练需要实例级注释，这不仅偏离了弱监督的要求，更无法应用于WSPD。**考虑到这些，我们认为传统方法可能比RPN更适合WSPD。加入长尾数据（背景框远多于息肉框）**

Since instance-level supervision is not available, prevailing work regards it as a multiple instance learning (MIL) problem to handle, whereby pre-generated proposals are required. Among various studies Selective search [14] and Edge Boxes [15] is 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[13]. With that in mind, they designed the region proposal network (RPN)[16]. Indeed, it showed a significant improvement in speed compared to traditional methods. Unfortuantely, 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. As Fig. 2 (b)-(c) shows the range of polyp scales and box ratios in different datasets is extremely variable. These variations in differentiation pose a severe challenge to parameters setting, as setting only the high-frequency parameter will cause a lower recall rate resulting in missed detections, while result in a large number of redundant candidate proposals and more complex calculations if all the parameters are set. In addition, these numbers are constantly dynamic because they are closely linked to the distance between the colonoscope and polyps. 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 WSPD.In consideration of these we argue that rather than RPN probably traditional methods are more adequate for WSPD.

尽管有几项很有前途的工作将MIL与深度学习[17-19]相结合，突破了自然图像的界限将弱监督成功地应用于医学领域。然而，与自然图像领域类似，它们也很容易过度拟合对象部分（如图1a left），因为它们本质上都高度依赖于区域分类器，而最具辨别力的分类证据可能来自整个对象区域，但也可能来自关键部分。幸运的是，在弱监督分割（WSS）中，检测区域的完整性更容易确保。考虑到他们内在的联系，一些结合分割指导检测的努力[20-21]已经被做出。然而，WSS需要检测热图作为伪标签，但其过于粗糙不足以作为指导标记。得益于迁移学习，最近的视觉大模型方面的进展如（SAM）使这一想法成为现实。糟糕的是，直接的息肉掩码预测是失败的（如图1a right），而高质量掩码需要输入提示（例如点、框）[22]。具体来说，边界框是最有效的点仅次于它。然而对于WSPD，点是唯一可用的选项。考虑到前面提到的事实，即分类器更倾向于关注对象中最具鉴别力的部分，我们认为它是自动获取输入提示的最强大的助手 (详见第3.1节)。**总之，WSPD和SAM不应该独立工作，而是天生的合作精神，应该共同努力克服自己的内在弱点**（如图1b）。

Despite several promising works combining MIL with deep learning [17-19] have greatly pushed the boundaries of natural images to successfully apply WS to the medical field. However, similar to the natural image domain, they also easily overfit the object parts as shown in Fig. 1 (a) left, because they both highly rely on regional classifiers in essence, while the most discriminating classification evidence may come from the entire object region, but may also from the crucial parts. Fortunately, the completeness of a detected region is easier to ensure in weakly supervised segmentation (WSS). For the inherent relations, some efforts [20-21] have been done to combine segmentation in order to guided the detection task. Nevertheless, WSS demands the detection heatmap as pseudo-labels, which is too coarse to be sufficient as guidemark. Benefiting from transfer learning, the recent advances in large vision models such as segment anything model (SAM) brings this assumption to reality. Badly, direct polyp mask prediction is failed as shown in Fig. 1 (a) right, and input prompts (e.g., points, boxes) are desired for high-quality masks of the specified location [22]. Specifically, the bounding box is the most effective, with the point next to it. However, for WSPD, point is the only available option. Considering the fact mentioned previously that the classifiers are more tend to focus on the most discriminative parts of the object, we believe that it is the most powerful assistant to automatically get the input prompts (detailed in section 3.1). In a word, instead of working independently, WSPD and SAM are naturally cooperative and should work together to overcome their intrinsic weaknesses as shown in Fig. 1 (b).

尽管在弱注释检测模型方面取得了一些显著进展，但准确可靠的息肉检测很容易被小而平坦的息肉所欺骗。正如我们在表1中的分析，在图像中，与背景相比，息肉区域相对较小，这使得前景（息肉）和背景（结肠直肠壁）之间的严重区域不平衡，导致息肉被大背景淹没，造成模型过度拟合不相关信息。而后，扁平息肉通常缺乏清晰可见的边界，并且表现出与周围结直肠组织相似的方面，这使得传统的网络难以准确区分前景息肉和外来背景。这背后的原因在于，与低分辨率深层特征相比，高分辨率浅层特征可以很好地学习细节特征和小结构，但检测会不断进行下采样以获得细粒度的表示，这会损害浅层特征，削弱对小而平坦目标的识别[23]。为了解决这一问题，我们引入了特征语义流融合模块（FSFF）来对齐多尺度特征之间的空间信息以丰富特征信息(详见第3.2节)。

Although some notable progress made in weak annotation detection models, accurate and reliable polyp detection can be easily fooled by small and flat polyps. As our analysis in Table 1, in the image, the polyp regions are relatively small when compare to the background leads to the severely regional imbalance between the foreground (polyp) and background (colorectal wall), which results in polyps being overwhelmed by the large background causing the model to overfit irrelevant information. Subsequently, typically the flat polyp lacks clearly visible borders and exhibits a similar aspect to the tissue of the surrounding colorectum, which makes conventional networks struggle to accurately distinguish foreground polyps from extraneous backgrounds. The reason behind this is that high-resolution shallow features work well to learn detailed features as well as small structures compared to low-resolution deep features, but the detection is constantly downsampled to obtain a fine-grained representation, which damages the shallow features and weakens recognition of small, flat targets[23]. To address this, we introduce feature semantic flow fusion (FSFF) module to align spatial information between multi-scale features to achieve a more efficient transfer of semantic information from shallow to deep layers to enrich feature information(detailed in section 3.2).

总之，我们的主要贡献有四个方面：

* 在提案生成阶段，我们通过对主流数据集的多维分析，证实了RPN在弱监督息肉检测任务中的不适用性。因此，我们使用传统方法为WSPD生成提案。
* 我们提出了CBIA-WSPD，一个基于互补协作机制的框架，在只有图像级注释的弱监督环境中增强对彼此的帮助。
* 为了更好地捕捉扁平和小息肉，我们设计了特征语义流融合模块，以从不同层获取更高质量的特征信息，通过聚合多尺度特征层的空间信息。
* 我们的方法在三个数据集（即ClinicDB、Kvasir、private）上进行，这些数据集在WSPD中获得了最先进的性能，并在某些方面与完全监督的检测具有竞争力。

In summary, our main contributions are listed four-fold:

* In the proposals generation stage, we confirmed the inapplicability of RPN in the task of weakly supervised polyp detection by performing multi-dimensional analyses on popular datasets. Therefore, we retain the traditional method to generate proposals for WSPD.
* We propose the CBIA-WSPD, a framework that employ complementary collaborative mechanism to enhance valuable assistance to each other in the weakly supervised setting with only image-level annotation.
* To better capture the flat and small polyps, we design the feature semantic flow fusion module to get higher quality feature information from different level by aggregating spatial information from multi-scales feature layers.
* Our method is conducted on three datasets(i.e. ClinicDB, Kvasir, private) , which obtains state-of-the art performance in WSPD and are competitive with fully supervised detection in certain aspects.

Our proposed weakly supervised polyp detection network achieves outstanding performance with only image category annotation through a novel cooperative mechanism. The contributions for the multimedia are shown as follows: 1) we successfully apply weak supervision to the domain of polyp detection, reducing the burden of data annotation and expanding the scope of weak supervision in multimedia interpretation. 2) We designed the collaborative mechanism can provide high-quality proposals to improve detection performance, which offers a new study direction for generating high-quality region proposals and solving local over-fitting.  3) Our work can detect small and flat objects, thus it has the potential to be applied to other multimedia interpretation tasks, e.g., small target recognition, defect detection.

**Reference**

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,”CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209–249, 2021.
2. Alrushaid N, Khan FA, Al-Suhaimi E, Elaissari A. Progress and Perspectives in Colon Cancer Pathology, Diagnosis, and Treatments. Diseases. 2023 Oct 24;11(4):148. doi: 10.3390/diseases11040148. PMID: 37987259; PMCID: PMC10660546.
3. P. Wang, P.X. Liu, J.R.G. Brown, T.M. Berzin, G.Y. Zhou, S. Lei, X.G. Liu, L.P. Li, X. Xiao, Lower Adenoma Miss Rate of Computer-Aided Detection-Assisted Colonoscopy vs Routine WhiteLight Colonoscopy in a Prospective Tandem Study, Gastroenterology, 159 (2020) 1252-+.https://doi.org/10.1053/j.gastro.2020.06.023
4. Jiaxin Mei and Tao Zhou and Kaiwen Huang and Yizhe Zhang and Yi Zhou and Ye Wu and Huazhu Fu:A Survey on Deep Learning for Polyp Segmentation: Techniques, Challenges and Future Trends. arXiv preprint arXiv:2311.18373(2023)
5. Hassan, C., Spadaccini, M., Iannone, A., Maselli, R., Jovani, M., Chandrasekar, V. T., Antonelli, G., Yu, H., Areia, M., Dinis-Ribeiro, M., et al., “Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis,” Gastrointestinal endoscopy 93(1), 77–85 (2021).
6. Yuncheng Jiang, Zixun Zhang, Yiwen Hu, Guanbin Li, Xiang Wan, Song Wu:ECC-PolypDet: Enhanced CenterNet with Contrastive Learning for Automatic Polyp Detection. arXiv preprint arXiv:2401.04961(2024)
7. Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins, “Esfpnet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video,” in Medical Imaging: Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 12468. SPIE, 2023, p. 1246803.
8. M. M. Rahman and R. Marculescu, “Medical image segmentation via cascaded attention decoding,” in IEEE/CVF WACVW, 2023, pp. 6222–6231.
9. Y. Li, M. Hu, and X. Yang, “Polyp-sam: Transfer sam for polyp segmentation,” arXiv preprint arXiv:2305.00293, 2023.
10. Hemin Ali Qadir, Younghak Shin, Jacob Bergsland, Ilangko Balasingham:Accurate Real-time Polyp Detection in Videos from Concatenation of Latent Features Extracted from Consecutive Frames. arXiv preprint arXiv:2303.05871(2023).
11. N. K. Tomar, D. Jha, U. Bagci, and S. Ali, “Tganet: Text-guided attention for improved polyp segmentation,” in MICCAI. Springer, 2022, pp. 151–160.
12. J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, “Stepwise feature fusion: Local guides global,” in MICCAI. Springer, 2022, pp.110–120.
13. 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（09）：2644-2660
14. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV 104(2), 154–171 (2013)
15. Zitnick, C.L., Doll´ar, P.: Edge boxes: Locating object proposals from edges. In:ECCV. pp. 391–405 (2014)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI 39(6), 1137–1149 (2017)
17. Philip Müller and Felix Meissen and Johannes Brandt and Georgios Kaissis and Daniel Rueckert.Anatomy-Driven Pathology Detection on Chest X-rays.MICCAI.2023
18. Su Z, Tavolara TE, Carreno-Galeano G, Lee SJ, Gurcan MN, Niazi MKK. Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images. Med Image Anal. 2022 Jul;79:102462. doi: 10.1016/j.media.2022.102462. Epub 2022 Apr 17. PMID: 35512532; PMCID: PMC10382794.
19. Qu J, Wei X, Qian X. Generalized pancreatic cancer diagnosis via multiple instance learning and anatomically-guided shape normalization. Med Image Anal. 2023 May;86:102774. doi: 10.1016/j.media.2023.102774. Epub 2023 Feb 21. PMID: 36842410.
20. Xu Y Q， Zhou C L， Yu X， Xiao B and Yang Y. 2021. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. IEEE Transactions on Image Processing，30：3029-3040［DOI：10.1109/TIP.2021.3056887］
21. Gao W， Wan F， Yue J， Xu S C and Ye Q X. 2022. Discrepant multiple instance learning for weakly supervised object detection. Pattern Recognition， 122： #108233 ［DOI： 10.1016/j. patcog. 2021.108233］
22. Kirillov, A., et al., Segment anything. arXiv preprint arXiv:2304.02643, 2023.
23. X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, and Y. Tong,“Semantic flow for fast and accurate scene parsing,” in ECCV. Springer,2020, pp. 775–793
24. Ma, J., He, Y., Li, F. et al. Segment anything in medical images. Nat Commun 15, 654 (2024).
25. [Tao Zhou](https://arxiv.org/search/cs?searchtype=author&query=Zhou,+T), [Yizhe Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang,+Y), [Yi Zhou](https://arxiv.org/search/cs?searchtype=author&query=Zhou,+Y), [Ye Wu](https://arxiv.org/search/cs?searchtype=author&query=Wu,+Y), [Chen Gong](https://arxiv.org/search/cs?searchtype=author&query=Gong,+C).Can SAM Segment Polyps? arXiv preprint arXiv:2311.18373(2023)
26. C. Zitnick and P. Dollar, “Edge boxes: Locating object proposals from edges,” in Proc. 13th Eur. Conf. Comput. Vis., 2014, pp. 391–405.
27. J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders,“Selective search for object recognition,” Int. J. Comput. Vis.,vol. 104, pp. 154–171, 2013.
28. A. Humayun, F. Li, and J. M. Rehg, “RIGOR: Recycling inference in graph cuts for generating object regions,” in Proc. IEEE Conf.Comput. Vis. Pattern Recog., 2014, pp. 336–343.
29. M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, “BING: Binarized normed gradients for objectness estimation at 300fps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 3286–3293.
30. Tang P， Wang X G， Wang A T， Yan Y L， Liu W Y， Huang J Z and
31. Yuille A. 2018. Weakly supervised region proposal network and object detection//Proceedings of the 15th European Conference on Computer Vision. Munich， Germany： Springer：370-386［DOI:10.1007/978-3-030-01252-6\_22］
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI 39(6), 1137–1149 (2017)
33. Gao W， Wan F， Yue J， Xu S C and Ye Q X. 2022. Discrepant multiple instance learning for weakly supervised object detection. Pattern Recognition， 122： #108233 ［DOI： 10.1016/j. patcog. 2021.108233］
34. Xu Y Q， Zhou C L， Yu X， Xiao B and Yang Y. 2021. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. IEEE Transactions on Image Processing，30：3029-3040［DOI：10.1109/TIP.2021.3056887］
35. Kim, E.; Lee, J.; and Choo, J. 2021. BiaSwap: Removing dataset bias with bias-tailored swapping augmentation. In ICCV
36. Li, Xiaolian et al. “Refine-FPN: Instance Segmentation Based on a Non-local Multi-feature Aggregation Mechanism.” Neural Processing Letters 55 (2022): 3411-3428.
37. Jha, D., et al. Kvasir-seg: A segmented polyp dataset. in MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. 2020. Springer.
38. Bernal, J., et al., WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics, 2015. 43: p. 99-111.
39. Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari.Weakly supervised localization and learning with generic knowledge. International Journal of Computer Vision (IJCV), 100(3):27 –293, 2012.