MonoBox: Tightness-Free Box-Supervised Polyp Segmentation Using Monotonicity Constraint

Qiang Hu[1], Zhenyu Yi[2], Ying Zhou[1], Fan Huang[3], Mei Liu[4], Qiang Li[1], Zhiwei Wang[1]*

[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[2] School of Engineering Sciences, Huazhong University of Science and Technology
[3] Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.
[4] Tongji Medical College, Huazhong University of Science and Technology
{huqiang77, zwwang}@hust.edu.cn

Abstract

We propose MonoBox, an innovative box-supervised segmentation method constrained by monotonicity to liberate its training from the user-unfriendly box-tightness assumption. In contrast to conventional box-supervised segmentation, where the box edges must precisely touch the target boundaries, MonoBox leverages imprecisely-annotated boxes to achieve robust pixel-wise segmentation. The 'linchpin' is that, within the noisy zones around box edges, MonoBox discards the traditional misguiding multiple-instance learning loss, and instead optimizes a carefully-designed objective, termed *monotonicity constraint*. Along directions transitioning from the foreground to background, this new constraint steers responses to adhere to a trend of monotonically decreasing values. Consequently, the originally unreliable learning within the noisy zones is transformed into a correct and effective monotonicity optimization. Moreover, an adaptive label correction is introduced, enabling MonoBox to enhance the tightness of box annotations using predicted masks from the previous epoch and dynamically shrink the noisy zones as training progresses. We verify MonoBox in the box-supervised segmentation task of polyps, where satisfying box-tightness is challenging due to the vague boundaries between the polyp and normal tissues. Experiments on both public synthetic and in-house real noisy datasets demonstrate that MonoBox exceeds other anti-noise state-of-the-arts by improving Dice by at least 5.5% and 3.3%, respectively.

Code — https://github.com/Huster-Hq/MonoBox

Introduction

Colorectal Cancer (CRC) threats to human health worldwide, and colonoscopy is a golden-standard of identifying and resecting early polyps (Haggar and Boushey 2009; Ji et al. 2024). Recently, numerous deep learning-based polyp segmentation methods (Fan et al. 2020; Zhao, Zhang, and Lu 2021; Dong et al. 2021; Zhang et al. 2022; Ji et al. 2023) have been proposed to assist the accurate resection and workload reduction. However, they are mostly fully-supervised and thus require pixel-level mask labels, which are time-consuming and expensive to acquire.

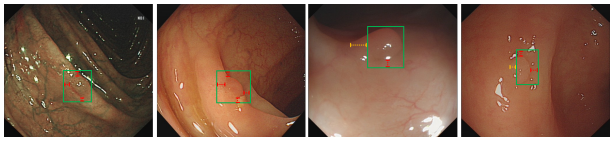

Figure 1: Examples of non-tight box annotations produced by endoscopists in the real annotation process. The red dashed lines indicate regions where the annotation is too wide, and the yellow dashed lines indicate regions where the annotation is too narrow.

To reduce the annotation cost, weakly-supervised segmentation (WSS) methods are studied, aiming to train segmentation models with more cost-effective labels, such as image-level categories (Ahn, Cho, and Kwak 2019), points (Cheng, Parkhi, and Kirillov 2022), and bounding boxes (Tian et al. 2021; Wang and Xia 2021). Among them, box-supervised segmentation (BSS) methods achieve the closest performance to fully-supervised methods, and thus attract dominant research attentions. The prevailing idea of BSS is called multiple-instance learning (MIL), which views each pixel as an instance, and defines a pixel-width image column or row as a positive bag if it crosses the annotated box, or negative bag otherwise. By pooling the instances' predictions as the corresponding bag-level prediction, the segmentation model can be trained to produce pixel-level results in the optimization of bag classification. However, the existing MIL-based BSS methods (Tian et al. 2021; Wang and Xia 2021; Cheng et al. 2023; Wei et al. 2023; Wang et al. 2023) mostly, if not all, are based on a box-tightness assumption, that is, the box's edges must precisely touch the target to make sure the bag-level labels are accurate. This severely inhibits the practical application of polyp datasets. In the polyp context, as shown in Figure 1, the characteristics of polyps, such as blurred boundaries, low contrast with normal tissues, and small sizes, bring ambiguities to annotators and lead to inaccurate (i.e., non-tight) box annotations. Therefore, improving the tolerance of BSS methods to inaccurate boxes can alleviate the annotation difficulty and thus is of urgency for the clinical usage of polyp segmentation.

The above objective can fall into the scope of noise learning, where the noise is induced by box non-tightness. A

*Corresponding author.

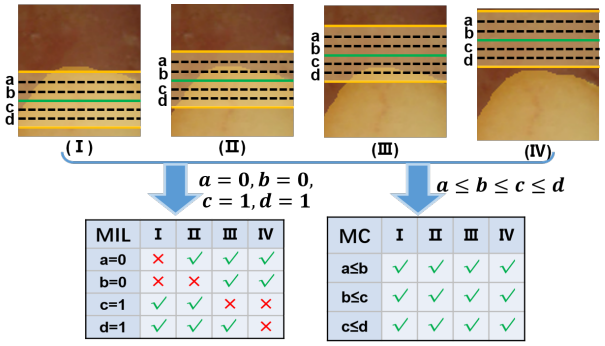

Figure 2: In cases (I, II, III, IV) of four typical noisy box annotations (green line), our proposed monotonicity constraint (MC) provides correct constrains for four sampled bags (black dashed line) from the unconfident region (the region between two yellow lines), but the traditional MIL lead to incorrect constraints. For brevity, we only visualize the local region of the upper boundary of the box annotation.

straightforward solution is to rectify the boxes to enhance the tightness (Song, Kim, and Lee 2019; Xu et al. 2021), but this usually requires a set of clean data as reference, which is beyond the scope of this work. Without clean reference, one promising solution is to apply noise-tolerant classification losses (Wang et al. 2019; Ma et al. 2020) for addressing the inaccurate bag-level labels in MIL. In addition to the loss term, sampling more accurate bags is another direction (Wang and Xia 2022; Zhu et al. 2023). However, no matter improving classification losses or sampling strategies in MIL, these solutions ignore the spatial correlation of noises (i.e., incorrect bags), showing weakness in the task of segmentation.

In this paper, we propose MonoBox that constrains monotonicity for tightness-free BSS, which is motivated by two characteristics of the spatial distribution of incorrect bags: (1) the confusing bags are sampled nearby the non-tight box edges, and (2) the probability of sampling a positive bag decrease from inside to outside across the box edges. Specifically, we first define the regions near the box edges as unconfident regions and the others as confident ones. For bags sampled from the confident regions, we adopt traditional MIL constraints. For bags sampled from the unconfident regions, we propose a novel monotonicity constraint (MC), which does not force hard predictions aligning the unreliable labels, but instead encourages a monotonicity trend that the inner response should be higher than the outer response. Principally, MC finds a soft but reliable surrogate objective when the precise ground-truth is absent, and the objective conforms to the expected spatial distribution pattern. As evidenced by the four typical noisy cases in unconfident regions (see Figure 2, for the bags (i.e., a,b,c,d) across the inaccurate box edge, the MIL's classification loss may result in misguiding supervisions, while MC can always derive meaningful gradients for optimization.

Moreover, as the model gradually gains the ability of distinguishing polyp pixels, the MC's imposing areas should dynamically change correspondingly. Thus, we further introduce a label correction strategy to gradually replace the noisy boxes with the predicted masks' bounding boxes, and shrink the unconfident regions, as the training progresses. This improves the box tightness and increases the learning efficacy of MonoBox.

In summary, our major contributions are as follows.

- We propose a tightness-free box-supervised polyp segmentation method, namely MonoBox, which can be effectively trained to precisely segment polyps using noisy box annotations.
- We propose a new monotonicity constraint (MC) to convert the confusing supervisions into reliable ones on the unconfident regions, and a label correction strategy to dynamically improve the box-tightness during training.
- We conduct extensive and comprehensive experiments on both public and in-house datasets. The public dataset contains our synthetic controllable noises, and the in-house shows real non-tightness patterns. The comparison results demonstrate that our method can generally improve the robustness of MIL-based BSS methods for non-tight box annotations, and its superiority over the anti-noise state-of-the-arts with an increase of Dice by at least 5.5% and 3.3% on the public and in-house dataset, respectively.

Related Work

Fully-Supervised Polyp Segmentation

The polyp segmentation methods mostly rely on the fully-supervised learning to train their models. For example, PraNet (Fan et al. 2020) generated a global map as the initial guidance region and then used the reverse attention module to refine the segmentation results by using the multi-level features. SANet (Wei et al. 2021) adopted a shallow attention module and a color exchange operation to remove background noise and reduce the interference of image color to segmentation respectively. Polyp-PVT (Dong et al. 2021) utilized a transformer encoder, a camouflage identification module, and a similarity aggregation module, to effectively suppress noises in the features and significantly improve their expressive capabilities. However, these methods all require pixel-level polyp annotations, which is difficult and time-consuming to acquire than box-level annotations, impeding the application on large-scale datasets.

Weakly-Supervised Segmentation Using Boxes

For reducing the cost of annotation, box-supervised segmentation (BSS) is extensively studied. These methods can be broadly classified into two categories. One is generating pseudo-labels and the other is based on multiple-instance learning (MIL).

Earlier, most of the methods were based on generating pseudo-labels. For example, BoxSup (Dai, He, and Sun 2015) and Box2Seg (Kulharia et al. 2020) used MCG (Williams et al. 2004) and GrabCut (Rother, Kolmogorov, and Blake 2004), respectively to generate refined pixel-level masks as pseudo labels based on ground truth

(GT) boxes. Mahani *et al.* (Mahani et al. 2022) proposed to reduce the loss weight of high entropy regions according to the predicted results to reduce error propagation during training. However, these methods rely heavily on the quality of pseudo labels and lack a stable and accurate guidance.

Recently, an idea based on MIL has been adopted by a wide range of methods due to its effectiveness (Hsu et al. 2019; Tian et al. 2021; Lan et al. 2023; Cheng et al. 2023; Wang et al. 2023; Sun et al. 2024). It takes the compactness of the box as a priori and considers the maximum value of any row or column of pixels within the box as a positive and the maximum value of any row or column of pixels not passing through the box as a negative. The implementation details vary from method to method, such as projection into a 1D vector (Tian et al. 2021; Lan et al. 2023; Cheng et al. 2023) or reconstruction into a 2D box-like mask (Wei et al. 2023; Wang et al. 2023). With MIL loss as the pipeline, some auxiliary constraints, such as color pairwise affinity (Tian et al. 2021), multi-scale consistency (Wei et al. 2023) are introduced to further improve the performance. Among them, IBoxCLA (Wang et al. 2023) proposed Contrastive Latent-Anchors (CLA), which enhances the feature contrast between the polyp and the surrounding normal tissue, and achieves state-of-the-art (SOTA) in box-supervised polyp segmentation. However, these methods are extremely dependent on the box tightness assumption, which reduces the cost-effective advantage of BSS methods.

Learning With Noisy Labels

The task that training accurate models using noisy labels has been an active research area. In the classification realm, various techniques were proposed, such as label correction (Ma et al. 2018; Song, Kim, and Lee 2019), noise-tolerant loss function (Ghosh, Kumar, and Sastry 2017; Ma et al. 2020), and data cleaning (Han et al. 2018; Jiang et al. 2018). Most of them can be translated to segmentation task which can be regarded as pixel-level classification. In the detection realm, He *et al.* (He et al. 2019) proposed KL loss for learning localization variance to alleviate the interference of ambiguity boxes on the detector. Xu *et al.* (Xu et al. 2021) introduced a meta-learning method to deal with noisy labels by utilizing a few clean samples. OA-MIL (Liu et al. 2022) and SSD-Det (Wu et al. 2023) proposed to leverage clean class labels as guidance signals for refining inaccurate bounding boxes. However, there are drawbacks when applying to segmentation. The solutions for classification do not consider the structure information of the object, while those for detection can not well guarantee the tightness in corrected boxes.

At present, there are only a few studies explicitly aiming at BSS with noisy box labels. PolarT (Wang and Xia 2022) adopt MIL on the polar transformed image, which reduces the number of incorrect bags but biases the model toward simple instances. FSRM (Zhu et al. 2023) first generated pseudo-masks based on the noisy boxes, and then relied on the pseudo-masks to guide the bag sampling in MIL. However, the noises also can be inherited in the pseudo-masks, negatively impacting the following MIL learning consequently. In this paper, we aim to seek a noise-tolerant constraint to replace the unreliable MIL-loss on the non-tight regions, and view the constraint as a plug-and-play module to boost the robustness of the current MIL-based SOTAs to tightness-free box annotations.

Method

Figure 3(b) gives the overview of MonoBox, which employs the most advanced and efficient image-level MIL fashion (Wei et al. 2023; Wang et al. 2023) by optimizing a proxy map rather than individual sampled bags for box-supervised segmentation (BSS), and introduces a new monotonicity constraint (MC) to improve the optimization reliability on the noisy regions. Meanwhile, MonoBox designs a label correction strategy to dynamically improve the tightness of box annotations. In the following, we first recap the optimization of proxy map, and then explain the key components of MonoBox, and at last briefly discuss the extensibility of MC to other box-supervised MIL frameworks of segmentation.

Optimizing Proxy Map for BSS

Given a colonoscopy image $I \in \mathbb{R}^{3\times H\times W}$ and its GT box annotation $B = (x_{lt}, y_{lt}, x_{rb}, y_{rb})$, where (x_{lt}, y_{lt}) and (x_{rb}, y_{rb}) represent the coordinates of the top-left and bottom-right points of box, respectively, the GT box-filled mask $b \in \{0,1\}^{H\times W}$ is created by assigning 1 within B and 0 outside B, and a segmented map $m \in (0,1)^{H\times W}$ is obtained by applying the segmentation model on I.

The GT mask b and the predicted map m have discrepant representation, which impedes a direct optimization between them. To bridge the gap, m should firstly be converted into a proxy map p using the following equation:

$$p[i,j] = \texttt{max}(m[i,:]) \cdot \texttt{max}(m[:,j]), \tag{1}$$

where i and j represent the pixel indexes. The proxy map p decouples the shape information from the segmentation map, and thus can be optimized directly using the GT box-filled mask, as illustrated in Figure 3(b). The typical optimization objective is the consistency constraint (e.g., Dice) between b and p, which is formulated as follows:

$$\mathcal{L}_{CC} = -\frac{2\times|b\cap p|}{|b|+|p|}. \tag{2}$$

The consistency constraint relies on the tightness as a prior, which is hard to be satisfied for polyps and thus renders the above consistency constraint unreliable.

Separating Confident and Unconfident Regions

Based on the fact that the noise in the m caused by non-tight box distributes around the edges, we divide the proxy map into confident and unconfident regions. Specifically, as shown in Figure 3(a), there are four unconfident regions $\{R_u^l, R_u^r, R_u^t, R_u^b\} \in \mathbb{R}^{1\times H\times W}$, corresponding to the left, right, top, and bottom edges of the GT box, respectively. The areas of four unconfident regions are determined by the width and height of the GT box annotation and an unconfident scale λ. For example, R_u^l can be calculated as:

$$R_u^l(x,y) = \begin{cases} 1, & x \in [x_{lt} - \lambda\cdot w, x_{lt} + \lambda\cdot w] \\ & \quad and\ y \in [y_{lt} - \lambda\cdot h, y_{rb} + \lambda\cdot h], \\ 0, & otherwise. \end{cases} \tag{3}$$

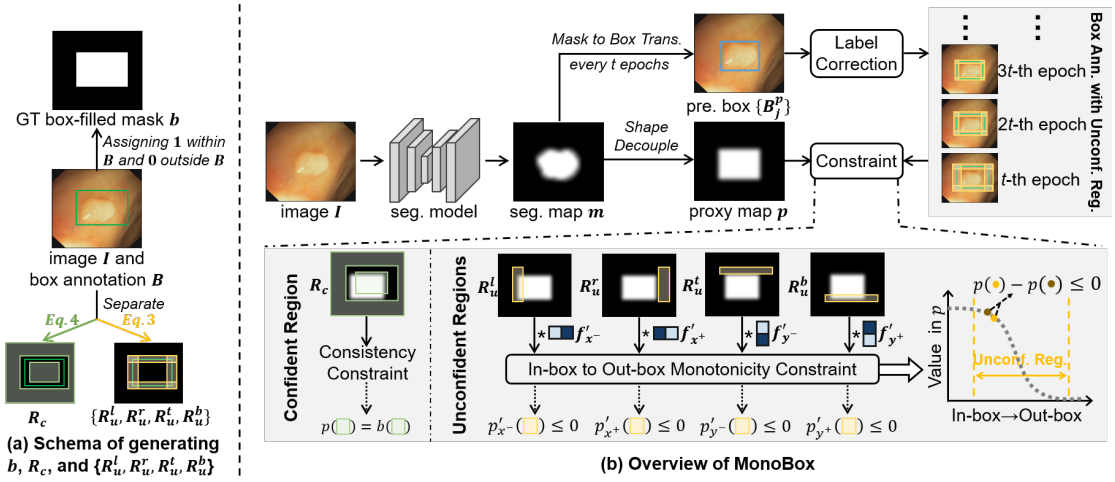

Figure 3: (a) Schema of generating GT box-filled mask (assigning 1 within box annotations and 0 outside box annotations), confident region (as Eq. 4) and unconfident regions (as Eq. 3). (b) Overview of our proposed MonoBox. For the proxy map decoupled from the segmented map, we first define the confident and unconfident regions, and then adopt the consistency constraint and monotonicity constraint on them, respectively. Moreover, we utilize a strategy of label correction strategy to dynamically improve the tightness of box annotations

where $w = x_{rb} - x_{lt}$ and $h = y_{rb} - y_{lt}$ are the width and height of the GT box, respectively. The confident region is mutually exclusive with four unconfident regions, which can be formulated as follows:

$$R_c = \mathbf{J} - (R_u^l \cup R_u^r \cup R_u^t \cup R_u^b), \tag{4}$$

where $\mathbf{J}$ is a map of size $H \times W$ with all ones. On the confident region, we adopt the consistency constraint, which is calculated in Eq. 2. On the unconfident region, we use our proposed monotonicity constraint, which is detailed in the next part.

Monotonicity Constraint on Unconfident Regions

Monotonicity constraint (MC) encourages a monotonicity trend that the box-inner response should be higher than the box-outer response. This is inspired by the fact that the closer to the box center, the higher probability of sampling positive bags, which is also satisfied for those non-tight boxes. This fact implies that, in the first-order derivative function of the direction pointing to the box center, the gradient should not exceed zero value. To this end, we first design four operation kernels, f'_{x^-}, f'_{x^+}, f'_{y^-}, and f'_{y^+}, which are formulated as:

$$f'_{x^-} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T f'_{x^+} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}^T f'_{y^-} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} f'_{y^+} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \tag{5}$$

and then we take f'_{x^-}, f'_{x^+}, f'_{y^-}, and f'_{y^+} to compute the first-order gradient maps on the proxy map p for R_u^l, R_u^r, R_u^t, and R_u^b, respectively. p'_{x^-}, p'_{x^+}, p'_{y^-}, and p'_{y^+} are the four corresponding gradient maps, and utilized to calculate losses by comparing them with zeros for monotonicity constraint (MC loss). Taking the MC loss in R_u^l as an example, the formula is as follows:

$$\mathcal{L}_{MC}^l = \sum_{i,j|R_u^l[i,j]=1} max\left(p'_{x^-}[i,j], 0\right). \tag{6}$$

Similarly, we calculate the MC loss for the other three unconfident regions, denoted as $\mathcal{L}_{MC}^r$, $\mathcal{L}_{MC}^t$, and $\mathcal{L}_{MC}^b$. Therefore, the complete MC loss is formulated as:

$$\mathcal{L}_{MC} = \mathcal{L}_{MC}^l + \mathcal{L}_{MC}^r + \mathcal{L}_{MC}^t + \mathcal{L}_{MC}^b. \tag{7}$$

Label Correction for Dynamic Training

We introduce label correction to maximize the learning efficacy. This is motivated by the observation that the segmentation model guided by the above MC loss can gradually gain the ability to classify polyp pixels, yielding reasonable segmentation results closer to the true clean box. Specifically, we transform the segmentation results into boxes $\{B_j^p, j = 0, 1, ..., M\}$ by finding connected regions and calculating the tightest box for each region. Next, we match the predicted boxes $\{B_j^p\}$ with the GT boxes $\{B_i\}$. The rule is that a GT box matches with only one predicted box that has the largest IoU with it, and the IoU must exceed a threshold τ. We merge the matched GT and predicted boxes to get the corrected boxes, and left the unmatched GT boxes as they are. The label correction increases the tightness of most GT boxes, and thus lowers noise degree. Therefore, the unconfident scale λ in Eq. 3 should be decreased accordingly. During the training, we evoke the label correction every t epochs, and dynamically adjust λ after each label correction by half reduction.

Generality for MIL-Based BSS

Although the implementation of the MC is based on proxy map optimization, we would like to remark that it is easy to expand to other MIL-based variants (Hsu et al. 2019; Tian et al. 2021) of BSS. Specifically, for the bags sampled from the unconfident regions, we only need to remember the row or column index of each bag, and the monotonicity constraint is a contrastive constraint of bag pairs with

adjacent indexes, that is, the prediction of bag with inside index should be higher than that with outside index. Such process is just simplified by use of the proxy map and the four derivative kernels in our implementation of MonoBox.

Experiments

Implementation Details

MonoBox is implemented using PyTorch (Paszke et al. 2019) and trained using a single NVIDIA GeForce RTX 3090 GPU with 24GB memory. We use AdamW (Loshchilov and Hutter 2017) as the optimizer, and set both the learning rate and weight decay to 0.0001. We resize the input image into 352 × 352 and set the batch size to 16. We train the models for 50 epochs in total and evoke the label correction every 10 epochs, i.e., $t = 10$. The unconfident scale λ in Eq. 3 and the IoU threshold τ in label correction are set to 0.2 and 0.7, respectively.

Datasets and Evaluation Metrics

Public Synthetic Noisy Dataset. We select five public polyp datasets used in previous work (Wang et al. 2023): ClinicDB (Bernal et al. 2015), Kvasir-SEG (Jha et al. 2020), ColonDB (Tajbakhsh, Gurudu, and Liang 2015), EndoScene (Vázquez et al. 2017), and ETIS (Silva et al. 2014). Following (Wang et al. 2023), we set 550 samples in ClinicDB and 900 samples in Kvasir as the train-set, and the remaining samples from these two datasets and all samples from the other three datasets as the test-set. Note that the original annotations for these datasets are ground truth (GT) masks, and we convert the GT masks to boxes by finding the tightest bounding boxes of the connected components.

We consider the boxes converted above to the tight and clean. For simulating tightness-free boxes, we perturb box coordinates of these clean boxes. Specifically, let (x_c, y_c, w, h) denote the center x coordinate, center y coordinate, width, and height of an clean box. We simulate an noisy bounding box $(\hat{x_c}, \hat{y_c}, \hat{w}, \hat{h})$ by randomly shifting and scaling the box as follows:

$$\begin{cases} \hat{x_c} = x_c + \triangle x \cdot w, & \hat{y_c} = y_c + \triangle y \cdot h, \\ \hat{w} = (1 + \triangle w) \cdot w, & \hat{h} = (1 + \triangle h) \cdot h, \end{cases} \tag{8}$$

where $\triangle x$, $\triangle y$, $\triangle w$, and $\triangle h$ follow the normal distribution $\mathcal{N}(0, \sigma^2)$, σ is the noise level.

In-House Real Noisy Dataset. CThe in-house dataset cosists 18,656 colonoscopy images of polyp. The dataset is from a local hospital, which is split to the train-set (17,350 images) and test-set (1,306 images). The train-set is provided with box annotations (Figure 1 shows four box-annotated examples), and the test-set is provided with pixel-level annotations by two experts. Written informed consent was not required for this study as documented clinical colonoscopic images were collected retrospectively and appropriately by anonymizing and deidentifying. Therefore, the study design was exempted from full review by the Institutional Review Board.

BSS	Methods	Real			Synthetic		
		Dice	IoU	HD(px)	Dice	IoU	HD(px)
WeakPolyp	UB	n/a	n/a	n/a	0.774	0.694	3.317
	LB	0.775	0.671	3.164	0.700	0.597	4.146
	SSD-Det	0.770	0.651	3.321	0.710	0.599	4.107
	NCE+RCE	0.783	0.684	3.224	0.721	0.615	3.989
	PolarT	0.766	0.654	3.328	0.701	0.591	4.288
	FSRM	0.785	0.689	3.174	0.722	0.613	3.853
	Ours	**0.804**	**0.714**	**2.916**	**0.763**	**0.660**	**3.411**
IBoxCLA	UB	n/a	n/a	n/a	0.827	0.750	3.048
	LB	0.803	0.716	2.412	0.735	0.628	4.310
	SSD-Det	0.802	0.713	2.324	0.743	0.634	3.898
	NCE+RCE	0.811	0.721	2.203	0.748	0.641	3.737
	PolarT	0.797	0.711	2.538	0.734	0.622	3.664
	FSRM	0.816	0.720	2.242	0.748	0.642	3.664
	Ours	**0.849**	**0.764**	**1.920**	**0.803**	**0.714**	**3.338**

Table 1: Comparison segmentation results between MonoBox and other anti-noise methods. The best performance is marked in bold. 'LB' means directly training the backbone with noisy datasets, 'UB' means training the backbone with clean datasets.

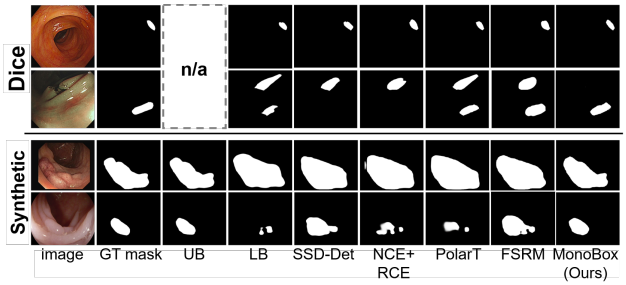

Figure 4: Visualization results of the different anti-noise methods using IBoxCLA as the model on the Real noisy dataset and the Synthetic noisy dataset.

Evaluation Metrics. We first use a threshold of 0.5 to binarize the segmented map of models to obtain the binary mask, and then we employ three widely-used evaluation metrics, including Dice, IoU, and Hausdorff distance (HD) to evaluate the similartity between the segmented and the ground truth (GT) masks. Among these metrics, Dice and IoU are similarity measures at the regional level, which mainly focus on the internal consistency of segmented objects. HD can better evaluate the segmentation results at the boundaries.

Comparison with Anti-Noise SOTAs

We view MonoBox as a plug-and-play anti-noise method, and compare it with four state-of-the-art (SOTA) anti-noise methods, i.e., SSD-Det (Wu et al. 2023), NCE+RCE (Ma et al. 2020), PolarT (Wang and Xia 2022), and FSRM (Zhu et al. 2023) on the Real noisy dataset and the Synthetic noisy dataset with noise level $\sigma = 0.2$. We train these methods using two box-supervised segmentation (BSS) backbones, i.e., WeakPolyp (Wei et al. 2023) and IBoxCLA (Wang

Components		Real			Synthetic		
MC	LC	Dice	IoU	HD	Dice	IoU	HD
✗	✗	0.803	0.716	2.412	0.735	0.628	4.310
✓	✗	0.838	0.758	2.053	0.787	0.700	3.655
✗	✓	0.809	0.720	2.447	0.744	0.628	4.002
✓	✓	**0.849**	**0.764**	**1.920**	**0.803**	**0.714**	**3.338**

Table 2: Ablation study on the effectiveness of the two proposed components, i.e., Monotonicity Constraint (MC) and Label Correction (LC).

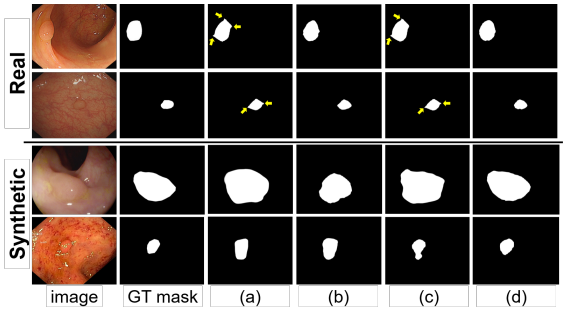

Figure 5: Visualization results of different variants. (a)-(d) correspond to the 1st-4th rows in Table 2. Yellow arrows show the predicted small thorn-like incorrect regions.

et al. 2023). Note that, we denote UB and LB as backbones trained using the clean and noisy datasets, respectively, and since there are no clean annotations in the Real noisy dataset, we do not report results of UB for the Real noisy dataset.

The comparison results are provided in Table 1. Comparing the results of LB and UB of the two backbones, we can see that on Synthetic noisy dataset, the noise leads a decrease of Dice by 7.4% and 9.2% for WeakPolyp and IBoxCLA, respectively. This exhibits the high sensitivity of the existing BSS methods to noisy boxes. When using MonoBox for optimization, the performance margin compared to the UB is significantly narrowed for both backbones.

Moreover, our method achieves the best performance among all anti-noise methods. IBoxCLA with our method exceeds that with the second-best method, i.e., FSRM, by 3.3% and 5.5% in Dice on the Real and Synthetic noisy datasets, respectively. This is because FSRM aims to correct the original noisy boxes using the model's predictions, but the predictions guided by traditional MIL are unreliable under the supervision of noisy labels. In contrast, the monotonicity constraint of MonoBox can provide reliable supervision even under noisy labels.

Figure 4 shows the visualization results of different methods using IBoxCLA as the backbone. It can be seen that the model trained with our method can accurately find the inconspicuous polyps and segment more reasonable boundaries.

Ablation Study

Effectiveness of Key Components. To verify the effectiveness two key components of MonoBox, i.e., monotonicity constraint (MC) and label correction (LC), we train three variants of MonoBox by disabling MC and/or LC, and IBoxCLA is used as the BSS backbone. The segmentation results are presented in Table 2 and Figure 5. Based on these results, two key conclusions can be made as follows:

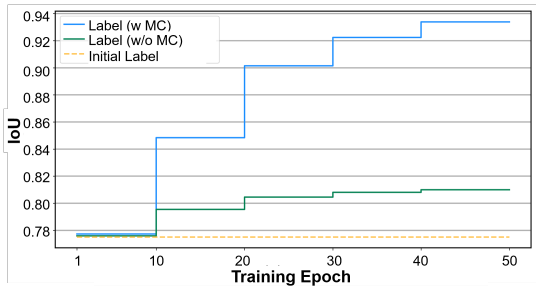

Figure 6: Label accuracy curve across the training process. The ordinate represents the IoU between the training box label and the clean box label. Following the implementation details, the label correction is performed every 10 epochs.

(1) By comparing the first two rows in Table 2 corresponding to (a) and (b) in Figure 5, we find that MC significantly improves the tolerance of the model to noisy boxes. From Figure 5(a), we find that on the Real noisy dataset, the baseline tends to produce small thorn-like regions (indicated by yellow arrows) at the boundary. This is because the doctors often annotated over-sized boxes, and thus the model has to output splinters to touch the incorrect boundaries. This phenomenon does not occur in the Synthetic noisy dataset, because we use gaussian distributed noise with zero-centered mean value, as shown in Eq. 8. Notably, as shown in Figure 5(b), MC can mitigate the misdirection of incorrect box annotations and eliminate erroneous thorn-like regions.

(2) By comparing the first and third rows in Table 2, we find that using LC alone brings only limited improvement and fails to address erroneous thorn-like regions. However, when LC is combined with MC, it can lead more improvement and eliminate the thorn-like regions, as shown in Figure 5(d). To clearly show the mechanism of LC, we visualize the label accuracy curve with/without MC across the training process on the Synthetic noisy dataset in Figure 6. Without MC, LC hardly improves the accuracy of labels, while with MC, LC significantly and continuously improves the label accuracy. This is because MC optimizes the constraint in unconfident regions, having the predicted masks more precisely, which can guide LC to improve the tightness of boxes, which in turn can guide more accurate predictions. Therefore, MC and LC complement each other and the combination of them achieves the best performance.

Different Strategies for Unconfident Region. MonoBox essentially defines the unconfident region and improves the reliability of constraint in this region through monotonicity constraint (MC). Likewise, there are also some strategies that address the constraint of the unconfident region, such as excluding this region in loss calculation, or computing the loss using soft labels. To further verify the effectiveness of MC, we conduct an experiment to compare the performance

Methods	Real			Synthetic		
	Dice	IoU	HD	Dice	IoU	HD
UB	n/a	n/a	n/a	0.827	0.750	3.048
LB	0.803	0.716	2.412	0.735	0.628	4.310
Exclusion	0.809	0.720	2.371	0.731	0.620	3.940
Soft Label	0.816	0.730	2.222	0.749	0.650	3.960
MC (Ours)	**0.838**	**0.758**	**2.053**	**0.787**	**0.700**	**3.655**

Table 3: Comparison results between the MC and two strategies for addressing the unconfident regions.

λ	Dice	HD	τ	Dice	HD	t	Dice	HD
0.1	0.836	2.108	0.5	0.842	1.994	2	0.840	2.057
0.2	**0.849**	**1.920**	0.6	0.846	1.953	5	0.844	1.987
0.3	0.844	1.964	0.7	**0.849**	**1.920**	10	**0.849**	**1.920**
0.4	0.838	2.011	0.8	0.845	1.988	20	0.842	2.043

Table 4: Ablation on λ, τ and t on the Real noisy dataset

of MC with two strategies, denoted as Exclusion and Soft Label, respectively. We choose 2D-Gaussian Label (Qadir et al. 2021) as the implementation of Soft Label. Note that, to avoid the influence of other components, we disable LC in this experiment. The quantitative comparison results are presented in Table 3. As can be seen, using Exclusion hardly improves the performance of the baseline. This is because simply excluding the unconfident regions could waste the possibly carried valuable information. Using Soft Label can improve the performance compared to LB, but there is still a large gap compared to UB. In contrast, our MC significantly improves the baseline's noise tolerance and achieves better performance than both strategies by remarkable margins.

Hyperparameters Choices. Table 3 compares our results on the Real noisy dataset with different choices of hyperparameters: confindent scale λ in Eq. 3, IoU threshold τ and interval epoch t in label correction. As can be seen, MonoBox using $\lambda = 0.2$, $\tau = 0.7$, and $t = 10$ shows the best performance, respectively. Note that, MonoBox does not rely on hyperparameter tuning when applied to different datasets. As shown in Figure 7, we keep hyperparameter choices constant ($\lambda = 0.2$, $\tau = 0.7$, $t = 10$) on datasets with different noise levels, MonoBox can maintain stable performance and outperform other SOTA methods.

Performance under Different Levels of Noise

We conduct an experiment to analyze the stability of MonoBox and the previous anit-noise SOTAs against different level of synthetic box noises. Specifically, taking IBoxCLA as the BSS backbone, we train the methods multiple times using four Synthetic noisy datasets with different noise levels σ. Figure 7 illustrates the performance trends in terms of Dice of these methods. As can be seen, when adding a slight noise ($\sigma = 0.1$), the segmentation backbone suffers a significant decrease in Dice from 0.827 to 0.753 (UB vs. LB). Compared to LB, MonoBox greatly alleviates noise interference, improving Dice by 5.7% (0.810 vs. 0.753) compared to LB and almost competitive to UB (0.810 vs. 0.827).

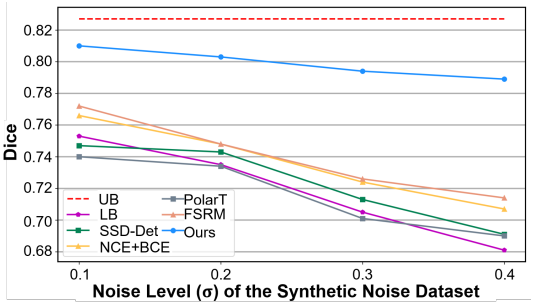

Figure 7: Dice of different methods under the different levels of noise on the Synthetic noisy dataset.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
UB	0.321	0.551	0.324	0.156	0.343	0.435
LB	0.248	0.480	0.225	0.117	0.282	0.329
FSRM	0.259	0.488	0.240	0.124	0.293	0.344
MonoBox(Ours)	**0.308**	**0.532**	**0.304**	**0.149**	**0.331**	**0.417**

Table 5: Comparison results on the COCO test-dev split.

As the noise level increasing, all methods show varying degrees of performance degradation. However, the curve of MonoBox is more stable and flatter compared to other methods, which shows that our method is more robust to noise.

Generality for Other Scenarios and Methods

MonoBox is not only applicable to any scenario, but also can be easily implemented with other MIL-based BSS methods. To verify this, we conduct an experiment on COCO (Lin et al. 2014), set BoxInst (Tian et al. 2021) as the backbone, and expand the MonoBox to adapt its MIL-based loss, i.e., projection loss. The tightness-free box annotations are generated by Eq. 8. As shown in Table 5, MonoBox significantly narrows the performance gap between LB and UB and outperforms FSRM by a large margin, which verifies that generality of MonoBox on natural scene and other MIL-based BSS methods.

Conclusion and Future Work

In this paper, we propose MonoBox, which addresses the limitations of the existing BSS method on tightness-free box annotations. Inspired by the specific spatial distribution of the noise, we propose MC to provide more reliable optimizations in model training. Moreover, we propose LC to improve the tightness of the box annotations and dynamically optimize the training process. Our approach is general and can easily cooperate with modern MIL-based BSS methods. The comprehensive experiments on the public synthetic noisy dataset and the in-house real noisy dataset demonstrate that MonoBox can effectively solve realistic problems in clinical practice and has great application value. Nevertheless, MonoBox currently adopts a uniform unconfident scale for all samples during training, we will explore methods to adaptively perceive the appropriate unconfident scale of each training sample in the future.

Acknowledgments

This work was supported in part by National Key R&D Program of China (Grant No. 2023YFC2414900), Key R&D Program of Hubei Province of China (No.2023BCB003), Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.

References

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2209–2218.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111.

Cheng, B.; Parkhi, O.; and Kirillov, A. 2022. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2617–2626.

Cheng, T.; Wang, X.; Chen, S.; Zhang, Q.; and Liu, W. 2023. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3145–3154.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1635–1643.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273. Springer.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Haggar, F. A.; and Boushey, R. P. 2009. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(04): 191–197.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, Y.; Zhu, C.; Wang, J.; Savvides, M.; and Zhang, X. 2019. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2888–2897.

Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. 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*, 451–462. Springer.

Ji, G.-P.; Fan, D.-P.; Chou, Y.-C.; Dai, D.; Liniger, A.; and Van Gool, L. 2023. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1): 92–108.

Ji, G.-P.; Liu, J.; Xu, P.; Barnes, N.; Khan, F. S.; Khan, S.; and Fan, D.-P. 2024. Frontiers in Intelligent Colonoscopy. *arXiv preprint arXiv:2410.17241*.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

Kulharia, V.; Chandra, S.; Agrawal, A.; Torr, P.; and Tyagi, A. 2020. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, 290–308. Springer.

Lan, S.; Yang, X.; Yu, Z.; Wu, Z.; Alvarez, J. M.; and Anandkumar, A. 2023. Vision transformers are good mask auto-labelers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23745–23755.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, C.; Wang, K.; Lu, H.; Cao, Z.; and Zhang, Z. 2022. Robust Object Detection with Inaccurate Bounding Boxes. In *European Conference on Computer Vision*, 53–69. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, 6543–6553. PMLR.

Ma, X.; Wang, Y.; Houle, M. E.; Zhou, S.; Erfani, S.; Xia, S.; Wijewickrema, S.; and Bailey, J. 2018. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, 3355–3364. PMLR.

Mahani, G. K.; Li, R.; Evangelou, N.; Sotiropolous, S.; Morgan, P. S.; French, A. P.; and Chen, X. 2022. Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qadir, H. A.; Shin, Y.; Solhusvik, J.; Bergsland, J.; Aabakken, L.; and Balasingham, I. 2021. Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction. *Medical Image Analysis*, 68: 101897.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314.

Silva, J.; Histace, A.; Romain, O.; Dray, X.; and Granado, B. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9: 283–293.

Song, H.; Kim, M.; and Lee, J.-G. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, 5907–5915. PMLR.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *The Twelfth International Conference on Learning Representations*.

Tajbakhsh, N.; Gurudu, S. R.; and Liang, J. 2015. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2): 630–644.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5443–5452.

Vázquez, D.; Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; López, A. M.; Romero, A.; Drozdzal, M.; Courville, A.; et al. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017.

Wang, J.; and Xia, B. 2021. Bounding box tightness prior for weakly supervised image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 526–536. Springer.

Wang, J.; and Xia, B. 2022. Polar transformation based multiple instance learning assisting weakly supervised image segmentation with loose bounding box annotations. *arXiv preprint arXiv:2203.06000*.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 322–330.

Wang, Z.; Hu, Q.; Shi, H.; He, L.; He, M.; Dai, W.; Li, T.; Zhang, Y.; Li, D.; Liu, M.; et al. 2023. IBoxCLA: Towards Robust Box-supervised Segmentation of Polyp via Improved Box-dice and Contrastive Latent-anchors. *arXiv preprint arXiv:2310.07248*.

Wei, J.; Hu, Y.; Cui, S.; Zhou, S. K.; and Li, Z. 2023. Weakpolyp: You only look bounding box for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 757–766. Springer.

Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S. K.; and Cui, S. 2021. Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 699–708. Springer.

Williams, B.; Poulter, N.; Brown, M.; Davis, M.; McInnes, G.; Potter, J.; Sever, P.; and McG Thom, S. 2004. Guidelines for management of hypertension: report of the fourth working party of the British Hypertension Society, 2004—BHS IV. *Journal of human hypertension*, 18(3): 139–185.

Wu, D.; Chen, P.; Yu, X.; Li, G.; Han, Z.; and Jiao, J. 2023. Spatial Self-Distillation for Object Detection with Inaccurate Bounding Boxes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6855–6865.

Xu, Y.; Zhu, L.; Yang, Y.; and Wu, F. 2021. Training robust object detectors from noisy category labels and imprecise bounding boxes. *IEEE Transactions on Image Processing*, 30: 5782–5792.

Zhang, W.; Fu, C.; Zheng, Y.; Zhang, F.; Zhao, Y.; and Sham, C.-W. 2022. HSNet: A hybrid semantic network for polyp segmentation. *Computers in biology and medicine*, 150: 106173.

Zhao, X.; Zhang, L.; and Lu, H. 2021. Automatic polyp segmentation via multi-scale subtraction network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 120–130. Springer.

Zhu, Z.; Shi, J.; Zhao, M.; Wang, Z.; Qiao, L.; and An, H. 2023. Contrast Learning Based Robust Framework for Weakly Supervised Medical Image Segmentation with Coarse Bounding Box Annotations. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, 110–119. Springer.