TextPolyp: Point-supervised Polyp Segmentation with Text Cues

Yiming Zhao[1], Yi Zhou[2], Yizhe Zhang[1], Ye Wu[1] and Tao Zhou[1] (✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China.
taozhou.ai@gmail.com
[2] School of Computer Science and Engineering, Southeast University, China.

Abstract. Polyp segmentation in colonoscopy images is essential for preventing Colorectal cancer (CRC). Existing polyp segmentation models often struggle with costly pixel-wise annotations. Conversely, datasets can be annotated quickly and affordably using weak labels such as points. However, utilizing sparse annotations for model training remains challenging due to the limited information. In this study, we propose a TextPolyp approach to tackle this issue by leveraging only point annotations and text cues for effective weakly-supervised polyp segmentation. Specifically, we utilize the Grounding DINO algorithm and Segment Anything Model (SAM) to generate initial pseudo-labels, which are then refined with point annotations. Furthermore, we employ a SAM-based mutual learning strategy to effectively enhance segmentation results from SAM. Additionally, we propose a Discrepancy-aware Weight Scheme (DWS) to adaptively reduce the impact of unreliable predictions from SAM. Our TextPolyp model is versatile and can seamlessly integrate with various backbones and segmentation methods. Importantly, the proposed strategies are used exclusively during training, incurring no additional computational cost during inference. Extensive experiments confirm the effectiveness of our TextPolyp approach. Our code is available at https://github.com/taozh2017/TextPolyp.

Keywords: Polyp segmentation· Weakly-supervised segmentation· SAM

1 Introduction

Colorectal cancer (CRC), one of the most common types of cancer worldwide, is highly associated with colon polyps. Early identification and eradication of polyps can prevent further detriment to adjacent tissues and greatly reduce the incidence of colorectal cancer. Given its importance, numerous polyp segmentation models [6,29,17,25,27] have demonstrated promising performance. However, as shown in Fig. 1, several existing models are fully-supervised and heavily rely on pixel-wise annotations, which is time-consuming and expensive.

To address the challenge, weakly-supervised methods based on sparse annotations in various forms have attracted wide attention and proven to be effective solutions for handling limited annotated data. For example, Chen *et al.* [4]

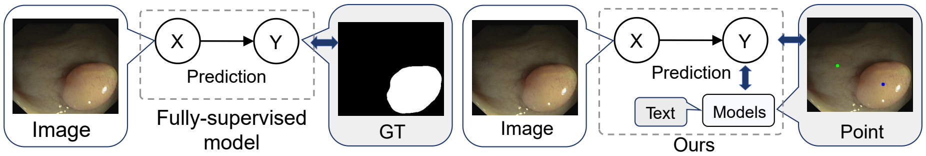

Fig. 1. Comparison of fully-supervised polyp segmentation method with our point-supervised model.

introduced a causal CAM method that utilizes image-level labels to overcome the challenges of unclear boundaries between the target foreground and background. Liu *et al.* [16] presented a weakly-supervised segmentation method that only requires scribble supervision for detecting COVID-19 infections in CT slices. This method is enhanced with uncertainty-aware self-ensembling and transformation-consistent techniques. In addition to image-level and scribble labels, Wei *et al.* [23] proposed a weakly-supervised polyp segmentation model based on bounding box annotations. However, variations in polyp size, shape, and indistinct boundaries make segmentation with sparse annotations significantly more challenging. Building upon the fusion of classical edge detection techniques, Bui *et al.* [3] proposed MEGANet, tailored specifically for polyp segmentation within colonoscopy images. Recently, the Segment Anything Model (SAM) [14] has demonstrated remarkable performance in segmentation. Some studies [28,12,10] have incorporated SAM into different segmentation. While SAM showcases zero-shot transfer capabilities, it may not excel in certain downstream tasks, such as medical image segmentation. Risab [2] utilized SAM with text prompting for robust and more precise polyp segmentation. Furthermore, rare works focus on utilizing SAM to enhance weakly-supervised polyp segmentation. Therefore, further research is necessary to investigate the effective integration of SAM into weakly-supervised polyp segmentation approaches.

In this paper, we propose TextPolyp, a novel method for weakly-supervised polyp segmentation that relies solely on point annotations with text cues. Initially, we utilize large models to generate pseudo-labels through Text-induced Pseudo-label Generation and refine these labels using a strategy for optimization. Subsequently, we employ a SAM-based Mutual Learning strategy to engage in mutual consistency supervision to boost segmentation capabilities. Additionally, we introduce a Discrepancy-aware Weight Scheme (DWS) to dynamically adjust weights and mitigate the impact of imprecise predictions from SAM. Notably, TextPolyp is a plug-and-play module that harnesses SAM to enhance the performance of existing segmentation models or attractive backbones. Through comprehensive experiments, we demonstrate the effectiveness of TextPolyp, achieving comparable performance across various baseline models.

2 Methodology

Fig. 2 presents the overview structure of our proposed model. We utilize commonly employed and effective backbones or segmentation methods (*e.g.*, UNet [18],

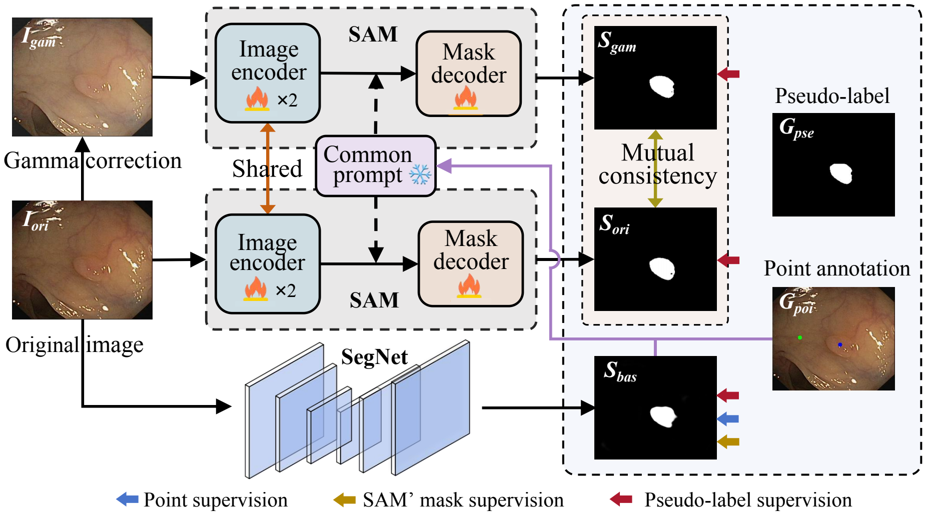

Fig. 2. Overview of the proposed framework. Our approach involves utilizing SegNet and SAM to produce masks (S_{bas} and S_{ori}) from a given image, while the gamma-corrected image also serves as input for SAM to generate S_{gam}. Additionally, we combine S_{bas} with the point annotation G_{poi} to form a box prompt for enhancing the quality of both S_{ori} and S_{gam}. Mutual consistency supervision is implemented between S_{ori} and S_{gam}. Finally, S_{ori} and S_{gam} along with the point annotation and pseudo-label, guide the training of S_{bas}, which represents the ultimate output of our model.

Res2Net [7], PraNet [6], etc.) as the baseline segmentation model (denoted "SegNet"). Specifically, we employ the text-induced pseudo-label generation module to produce pseudo-labels. Subsequently, for the input image $I_{ori} \in R^{H\times W\times 3}$, we apply gamma correction for I_{ori} to obtain $I_{gam} \in R^{H\times W\times 3}$. Afterward, I_{ori} and I_{gam} are fed into the SAM-based mutual learning network to acquire two segmentation masks. Further, we combine the segmentation map S_{bas} from the SegNet with the point annotations to form the box prompt, which is then supplied to SAM to produce S_{ori} and S_{gam}. Finally, the segmentation maps can be supervised by point annotations and pseudo-labels.

2.1 Text-induced Pseudo-label Generation

In alignment with standard weakly-supervised segmentation tasks, we initially acquire pseudo-labels and utilize them for network training. To enhance the supervisory signal obtained from point annotations, we incorporate the Grounding DINO [15] algorithm and SAM to generate the initial pseudo-labels. Specifically, as illustrated in Fig. 3 (a), we input a generic text description of polyps (*e.g.*, "A colorectal polyp is an abnormal growth on the lining of the colon or rectum. Some polyps are flat while others have a stalk. Polyps come in various shapes, which are sometimes flat and round, yet usually irregular. The color of the polyp is similar to the surrounding normal tissue, and the polyp lacks a

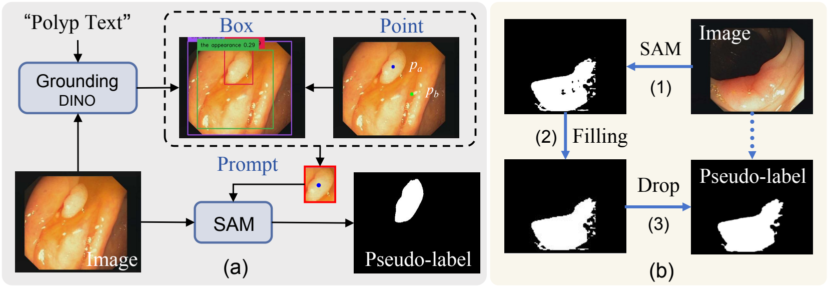

Fig. 3. Illustration of the processing steps for pseudo-label generation.

distinct boundary") into the Grounding DINO to produce a variety of detection boxes $\mathcal{B} = \{B_i\}_{i=1}^{N}$ for each image. However, some boxes are inaccurate or cover numerous background regions. Hence, it is crucial to filter out misidentified boxes to ensure the most precise inclusion of polyps. Additionally, in situations where $B_i \cap B_j = B_i$ or $B_i \cap B_j = B_j$ with $i \neq j$, we must decide whether to retain B_i or B_j to eliminate redundant boxes. To address this issue, we refine these detected boxes from Grounding DINO using point annotations. Specifically, we utilize the foreground point p_a (the blue point in Fig. 3 (a)) within the point annotation to identify the box containing polyps and exclude misidentified boxes. Simultaneously, we use p_b to eliminate extraneous boxes. Consequently, we can obtain refined boxes $\mathcal{B}^* = \{B_i \in \mathcal{B} \mid p_a \in B_i, p_b \notin B_i\}$, which are fed into the SAM along with the point annotations p_a and p_b to obtain the initial pseudo-labels.

Pseudo-label Refinement. While SAM demonstrates remarkable segmentation proficiency, it lacks reasonable judgment in output results without any prompts, leading to low-quality pseudo-labels. As depicted in Fig. 3 (b), SAM not only focuses on the entire image but also emphasizes details, including impurities within polyp images, which can significantly impact the accuracy of pseudo-labels. Consequently, we introduce a sound methodology to rectify pseudo-labels. When a small background region is present within the boundary of the polyp area, we regard this background as an integral part of the polyp and proceed with step (2) to fill the target foreground. Furthermore, we exclude small foreground areas (radius ≤ 5 pixels) that are isolated and proceed with step (3) to eliminate these areas.

2.2 SAM-based Mutual Learning

It is crucial to supply SAM with effective and accurate prompts for producing more promising segmentation results.

Prompt Generation. As depicted in Fig. 2, to improve the quality of segmentation masks produced by SAM, we utilize the point annotations and segmentation maps obtained by the SegNet to produce bounding box prompts. However, the bounding box derived from SegNet's segmentation map may not accurately encompass the polyp region. On the other hand, although the point annotation provides limited information about the target pixels, it accurately pinpoints the position of the polyp, enabling precise target localization. Consequently, we devise an integrated strategy to generate a suitably sized bounding box by

$$\begin{cases} Box_p = \{P(x,y) \pm \text{axis}, \quad \text{axis} = \delta/2\}, \; if \quad \max(Box_s|w|, Box_s|h|) < \delta, \\ Box_p = \{P(x,y) \pm \text{axis}, \quad \text{axis} = \delta\}, \quad\; if \quad \max(Box_s|w|, Box_s|h|) \geq \delta, \end{cases} \tag{1}$$

where Box_s denotes the bounding box obtained from the segmentation map of the SegNet, and $Box_s|w|$ and $Box_s|h|$ represent the width and height of the box, respectively. The tuning parameter δ is used to derive a relatively accurate box prompt. Moreover, $P(x,y)$ represents the coordinate of the point. Consequently, we can derive an integrated box Box_p. Then, to ensure that the current box encompasses the target regions (*i.e.*, polyp), we expand the original box Box_s as $Box_s^{'} = \{Box_s(|w|) \pm \delta/20, Box_s(|h|) \pm \delta/20\}$. Therefore, the resulting intersecting region serves as the final prompt box, denoted as $Box = Box_p \cap Box_s^{'}$.

Mutual Consistency Supervision. In light of the weak supervision labels, we develop a dual SAM structure. This architecture allows the two outputs from the two SAM decoders to engage in mutual learning and supervision. Specifically, given an image $I_{ori} \in R^{H \times W \times 3}$, we implement gamma correction on I_{ori} to obtain $I_{gam} \in R^{H \times W \times 3}$. Then, I_{ori} and I_{gam} are fed to the dual SAM networks for training and get two segmentation masks S_{ori} and S_{gam}. After that, we put I_{ori} and I_{gam} into the encoders individually and unfreeze the last two layers to fine-tune SAM, thereby allowing them to acquire distinctive features. Additionally, they share the same prompt to ensure overall consistency.

Moreover, to ensure the overall information is complementary between the original image branch and the gamma branch, we modify the mask generation process by applying the Sigmoid activation function and fine-tuning the mask decoder to produce the segmentation masks (S_{ori} and S_{gam}). It permits each branch to get the supervision of the other branch, thus retaining richer features. Inspired by [9], we adopt ℓ_1 loss and the structural similarity index measure (SSIM) for the supervision, which can be defined as $\mathcal{L}_{CON}(S_1, S_2) = \alpha\mathcal{L}_1(S_1, S_2) + (1-\alpha)\mathcal{L}_{ssim}(S_1, S_2)$, where S_1 and S_2 denote two prediction maps, and α is a trade-off parameter. Therefore, the mutual consistency loss is expressed as follows:

$$\mathcal{L}_{mut} = \mathcal{L}_{CON}(S_{ori}, S_{gam}). \tag{2}$$

2.3 Discrepancy-aware Weight Scheme

To minimize the impact of inaccuracies in segmentation masks generated by SAM, we propose a Discrepancy-aware Weight Scheme (DWS). This scheme can adaptively weight the loss between the segmentation masks produced by SAM and SegNet, aiming to reduce the influence of locations with significant discrepancies between the two maps from dual SAM. Specifically, given the distinct outputs (S_{ori} and S_{gam}), we assume that the overlapping areas of the two maps represent the more reliable polyp regions. Therefore, for regions with substantial differences between the maps, we reduce their weights to enhance overall precision.

To achieve this, we first calculate the average map, *i.e.*, $S_{avg} = \frac{1}{2}(S_{ori} + S_{gam})$, and then determine the discrepancies between each segmentation mask and the average. Using these differences, we can derive the weight map for S_{ori} and S_{gam} as follows:

$$\omega_{ori} = \exp^{-|S_{ori}-S_{avg}|}, \quad \omega_{gam} = \exp^{-|S_{gam}-S_{avg}|}. \tag{3}$$

We then integrate each weight map into the loss function to minimize the influence of unreliable prediction locations. Given a predicted map P and a supervised map S, we incorporate each weight map into the formulations of the BCE loss and IoU loss as follows:

$$\begin{cases} \mathcal{L}^{\omega}_{BCE} = -\dfrac{1}{H \times W} \displaystyle\sum_{i\in H, j\in W} [s_{ij} \cdot \log(p_{ij}) + (1 - s_{ij}) \log(1 - p_{ij})] \cdot \omega_{ij}, \\ \mathcal{L}^{\omega}_{IoU} = 1 - \dfrac{\sum_{i\in H, j\in W}(p_{ij} \cdot s_{ij} \cdot \omega_{ij})}{\sum_{i\in H, j\in W}(p_{ij} + s_{ij}) \cdot \omega_{ij} - \sum_{i\in H, j\in W}(p_{ij} \cdot s_{ij} \cdot \omega_{ij})}. \end{cases} \tag{4}$$

For convenience, we denote $\mathcal{L}_{SEG} = \mathcal{L}^{\omega}_{BCE} + \mathcal{L}^{\omega}_{IoU}$. Due to the presence of completely failed segmentation masks obtained by SAM, we first discard the erroneous segmentation maps and then utilize the remaining maps to constrain the SegNet model. As a result, the loss function can be formulated as follows:

$$\mathcal{L}_{bas} = \mathcal{L}_{SEG}(S_{bas}, S_{ori}) + \mathcal{L}_{SEG}(S_{bas}, S_{gam}), \tag{5}$$

where ω_{ori} and ω_{gam} are applied for $\mathcal{L}_{SEG}(S_{bas}, S_{ori})$ and $\mathcal{L}_{SEG}(S_{bas}, S_{gam})$, respectively.

Overall Loss Function. We employ partial cross entropy loss [24] (denoted as $\mathcal{L}_{PCE}$) to quantify the disparities between the segmentation masks and point annotations. As a result, we have the loss function for the point supervision as $\mathcal{L}_{poi} = \mathcal{L}_{PCE}(S_{bas}, G_{poi})$, where G_{poi} denotes point annotations. Besides, we compute the consistency loss between S_{bas} and G_{pse}, *i.e.*, $\mathcal{L}_{pse} = \mathcal{L}_{CON}(S_{bas}, G_{pse})$. Moreover, we compute the consistency loss between the pseudo-labels (G_{pse}) and the SAM's masks, which can be expressed as follows:

$$\mathcal{L}_{psa} = \mathcal{L}_{CON}(S_{ori}, G_{pse}) + \mathcal{L}_{CON}(S_{gam}, G_{pse}). \tag{6}$$

Finally, the overall loss function can be formulated by

$$\mathcal{L}_{total} = \mathcal{L}_{bas} + \mathcal{L}_{psa} + \mathcal{L}_{mut} + \mathcal{L}_{poi} + \mathcal{L}_{pse}. \tag{7}$$

3 Experiments

Datasets. Five public colonoscopy datasets are adopted in this study, namely CVC-300 [21], Kvasir [13], ETIS-LaribPolypDB [19], CVC-ClinicDB [1], and CVC-ColonDB [20]. Following the setting in [6], 1450 images (900 Kvasir images and 550 CVC-ClinicDB images) are selected for the training set, and the remaining images are used for testing.

Table 1. Quantitative results on two seen datasets. “Poi.” means the supervision with point annotation, “Bac.” means backbone, and “Seg.” is the existing model designed for polyp segmentation.

Methods		**CVC-ClinicDB**				**Kavsir**			
		Dice	IoU	S_α	F_β^w	Dice	IoU	S_α	F_β^w
Poi.	PSOD [8]	0.343	0.261	0.412	0.333	0.605	0.480	0.668	0.603
	SCOD [11]	0.606	0.505	0.753	0.579	0.613	0.546	0.742	0.610
	SCWS [24]	0.666	0.567	0.759	0.566	0.599	0.501	0.645	0.522
	WSSOD [26]	0.617	0.458	0.729	0.587	0.563	0.412	0.676	0.546
Bac.	Ours+UNet [18]	0.739	0.650	0.828	0.719	0.749	0.656	0.819	0.737
	Ours+Res2Net [7]	0.822	0.744	0.879	0.807	0.818	0.737	0.862	0.814
	Ours+PVT [22]	0.833	0.761	0.890	0.819	0.848	0.776	0.882	0.848
Seg.	Ours+PraNet [6]	0.829	0.758	0.888	0.814	0.831	0.767	0.870	0.827
	Ours+Polyp-PVT [5]	0.838	0.766	0.893	0.825	0.853	0.786	0.886	0.854

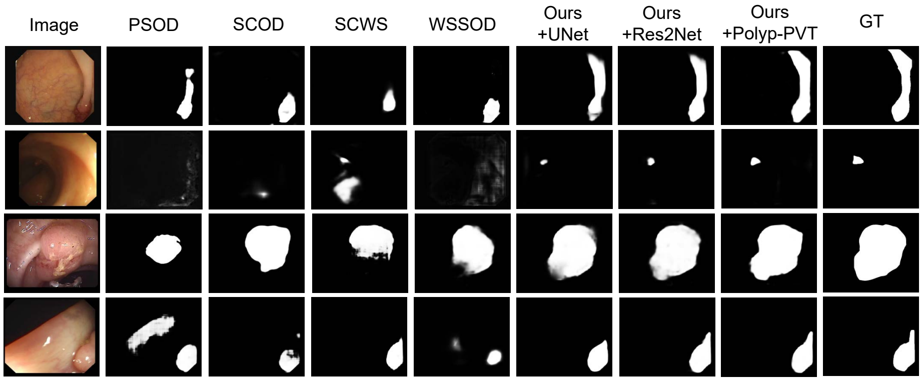

Fig. 4. Visualization results of different methods on the polyp segmentation.

Implementation Details. The overall framework is implemented in PyTorch and conducted on NVIDIA GeForce RTX3090 GPU. All input images are uniformly resized to 320×320. To optimize the training process, we adopt triangular warm-up and decay strategies optimized by SGD with the momentum of 0.9, and weight decay of 5e-4, and the learning rate is set to a maximum of 1e-2 and a minimum of 1e-5. Our model is trained for 100 epochs with a batch size of 8. Additionally, δ and α are set to 120 and 0.85, respectively. For evaluation, we adopt four commonly adopted metrics [6], namely Dice coefficient (Dice), Intersection over Union (IoU), S-measure (S_α), and weighted F-measure (F_β^w).

Results Comparison. We compare the proposed TextPolyp with recent methods that are trained with weak annotations. PSOD [8] is an existing point-supervised method, and we retrain it on polyp datasets following its original settings. The other methods, SCOD [11], SCWS [24], and WSSOD [26] are weakly-supervised approaches based on scribble annotations. We replace the scribble supervision with points and retrain them according to the original experiment setup. TextPolyp is trained using three different backbones, *i.e.*, UNet [18], Res2Net [7], and PVT [22]. Additionally, we apply TextPolyp to ex-

Table 2. Quantitative results on three unseen datasets.

Methods		CVC-300				CVC-ColonDB				ETIS-LaribPolyp			
		Dice	IoU	S_α	F_β^w	Dice	IoU	S_α	F_β^w	Dice	IoU	S_α	F_β^w
Poi.	PSOD [8]	.485	.365	.588	.475	.304	.214	.351	.284	.255	.196	.330	.242
	SCOD [11]	.623	.517	.768	.618	.518	.423	.686	.521	.351	.275	.580	.347
	SCWS [24]	.621	.506	.725	.589	.546	.449	.675	.444	.428	.352	.588	.324
	WSSOD [26]	.655	.520	.781	.612	.479	.359	.667	.442	.407	.290	.642	.343
Bac.	Ours+UNet [18]	.697	.621	.832	.681	.517	.437	.709	.510	.361	.297	.643	.352
	Ours+Res2Net [7]	.831	.752	.912	.797	.698	.611	.813	.662	.645	.557	.802	.579
	Ours+PVT [22]	.844	.765	.911	.816	.723	.641	.830	.698	.665	.591	.810	.627
Seg.	Ours+PraNet [6]	.832	.756	.914	.800	.718	.628	.819	.686	.642	.561	.808	.599
	Ours+Polyp-PVT [5]	.846	.770	.916	.817	.730	.644	.832	.701	.673	.591	.816	.629

Table 3. Ablation results with different settings.

Bac.	Settings	CVC-ClinicDB			CVC-300			CVC-ColonDB		
		Dice	IoU	S_α	Dice	IoU	S_α	Dice	IoU	S_α
UNet	w/o SAM	0.572	0.436	0.725	0.571	0.454	0.747	0.447	0.343	0.681
	w/o Gamma branch	0.679	0.596	0.796	0.662	0.564	0.808	0.502	0.428	0.703
	w/o DWS	0.734	0.644	0.822	0.676	0.584	0.815	0.513	0.434	0.708
	Ours	0.739	0.650	0.828	0.697	0.621	0.832	0.517	0.437	0.709
Res2Net	w/o SAM	0.726	0.641	0.828	0.694	0.600	0.825	0.574	0.482	0.734
	w/o Gamma branch	0.781	0.700	0.860	0.792	0.708	0.890	0.659	0.572	0.793
	w/o DWS	0.810	0.724	0.874	0.818	0.731	0.898	0.681	0.596	0.803
	Ours	0.822	0.744	0.879	0.831	0.752	0.912	0.698	0.611	0.813

isting polyp segmentation models, including PraNet and Polyp-PVT. Table 1 and Table 2 present quantitative results for different methods. It is evident that our TextPolyp model, when integrated with the base UNet backbone, outperforms other specifically designed weakly-supervised segmentation methods, highlighting the effectiveness of our approach. Furthermore, when TextPolyp is combined with different backbones, our model with PVT yields superior performance compared to that with UNet or Res2Net. Additionally, our model incorporating Polyp-PVT outperforms the one with PraNet. Fig. 4 illustrates some predictions based on different methods. Compared to other point-supervised and scribble-supervised methods, the segmentation models incorporating TextPolyp not only locate the position of polyps accurately but also achieve more complete segmentation for larger polyps, resulting in more accurate masks.

Ablation Study. To assess the effectiveness of our method, we conduct three ablation experiments using two different backbones: UNet and Res2Net. • **Effectiveness of SAM:** We exclude the SAM module and train the weakly-supervised model solely with point annotations, denoted as “w/o SAM”. As shown in Table 3, the degradation in performance without SAM indicates its effectiveness. SAM successfully combines the SegNet to achieve improved segmentation performance. • **Effectiveness of Mutual Consistency Supervision:** By removing the gamma branch (denoted as “w/o Gamma branch”), our method loses the capability for mutual supervision, and the DWS component is also eliminated. The results demonstrate a decline in performance without

the gamma branch, underscoring its critical role in achieving optimal results.
• **Effectiveness of DWS:** We remove the DWS component from our proposed model, denoted as "w/o DWS". In this scenario, we only utilize the common BCE loss and IoU loss to measure the discrepancy between predicted maps from the SegNet and SAM's mask maps. As illustrated in Table 3, the results confirm the effectiveness of the proposed DWS component. Through these ablation experiments, we validate the effectiveness of SAM, mutual consistency supervision, and the DWS component in our approach.

4 Conclusion

We propose TextPolyp, a versatile point-supervised polyp segmentation model derived from SAM. TextPolyp initially leverages Grounding DINO and SAM to generate pseudo-labels. Furthermore, we present the discrepancy-aware weight scheme with Dual-SAM, utilizing two maps from SAM to evaluate reliable pixels for model training. Experimental results show that TextPolyp surpasses other weakly-supervised methods. Notably, TextPolyp only necessitates point annotations and text cues, reducing the dependence on pixel-level annotations for segmentation models. Additionally, TextPolyp serves as a plug-and-play module that can seamlessly integrate into various backbones and segmentation methods.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 62172228, 62201263, 62106043, and 62201265).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
2. Biswas, R.: Polyp-sam++: Can a text guided sam perform better for polyp segmentation? arXiv preprint arXiv:2308.06623 (2023)
3. Bui, N.T., Hoang, D.H., Nguyen, Q.T., Tran, M.T., Le, N.: Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 7985–7994 (2024)
4. Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S.: C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11676–11685 (2022)
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-PVT: Polyp segmentation with pyramidvision transformers. CAAI Artificial Intelligence Research (2023)
6. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference

on Medical Image Computing and Computer-assisted Intervention. pp. 263–273. Springer (2020)

7. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(2), 652–662 (2021)
8. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 670–678 (2022)
9. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279 (2017)
10. He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. Advances in Neural Information Processing Systems (2023)
11. He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 781–789 (2023)
12. Huang, Z., Liu, H., Zhang, H., Xing, F., Laine, A., Angelini, E., Hendon, C., Gan, Y.: Push the boundary of sam: A pseudo-label correction framework for medical segmentation. arXiv preprint arXiv:2308.00883 (2023)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling. pp. 451–462. Springer (2020)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
16. Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D.: Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. Pattern Recognition **122**, 108341 (2022)
17. Pan, X., Ma, C., Mu, Y., Bi, M.: Glsnet: A global guided local feature stepwise aggregation network for polyp segmentation. Biomedical Signal Processing and Control **87**, 105528 (2024)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 234–241. Springer (2015)
19. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery **9**(2), 283–293 (2014)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)
21. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering **2017**(1), 4037190 (2017)

22. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 568–578 (2021)
23. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 757–766. Springer (2023)
24. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3234–3242 (2021)
25. Yue, G., Zhuo, G., Yan, W., Zhou, T., Tang, C., Yang, P., Wang, T.: Boundary uncertainty aware network for automated polyp segmentation. Neural Networks **170**, 390–404 (2024)
26. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12546–12555 (2020)
27. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 120–130. Springer (2021)
28. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? arXiv preprint arXiv:2304.07583 (2023)
29. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023)