QueryNet: A Unified Framework for Accurate Polyp Segmentation and Detection

Jiaxing Chai[1], Zhiming Luo[1,2(✉)], Jianzhe Gao[1], Licun Dai[1], Yingxin Lai[1], and Shaozi Li[1,2]

[1] Department of Artificial Intelligence, Xiamen University, Xiamen, Fujian, China
[2] Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University, China
✉Correspondences: zhiming.luo@xmu.edu.cn

Abstract. Recently, deep learning-based methods have demonstrated effectiveness in the diagnosing of polyps, which holds clinical significance in the prevention of colorectal cancer. These methods can be broadly categorized into two tasks: Polyp Segmentation (PS) and Polyp Detection (PD). The advantage of PS lies in precise localization, but it is constrained by the contrast of the polyp area. On the other hand, PD provides the advantages of global perspective but is susceptible to issues such as false positives or missed detections. Despite substantial progress in both tasks, there has been limited exploration of integrating these two tasks. To address this problem, we introduce QueryNet, a unified framework for accurate polyp segmentation and detection. Specially, our QueryNet is constructed on top of Mask2Former, a query-based segmentation model. It conceptualizes object queries as cluster centers and constructs a detection branch to handle both tasks. Extensive quantitative and qualitative experiments on five public benchmarks verify that this unified framework effectively mitigates the task-specific limitations, thereby enhancing the overall performance. Furthermore, QueryNet achieves comparable performance against state-of-the-art PS and PD methods. Code is available at Github.

Keywords: Unified Framework· Polyp Segmentation· Polyp Detection

1 Introduction

Early diagnosis of polyps is of great clinical significance in preventing colorectal cancer. However, manual diagnosis entails considerable costs and is susceptible to challenges such as missed detections and false positives, affecting the examinations and treatments. Deep learning-based methods offer promise in overcoming these limitations, thereby enhancing the precision and efficiency of polyp treatment. Currently, these methods can been divided into two distinct tasks: Polyp Segmentation (PS) and Polyp Detection (PD). Specifically, PS affords pixel-level localization and detailed anatomical information, facilitating preoperative decision-making. PD aims to identify and recognize the presence of polyps, serving as a convenient tool for promptly assessing potential abnormalities with its efficiency and low computational demands.

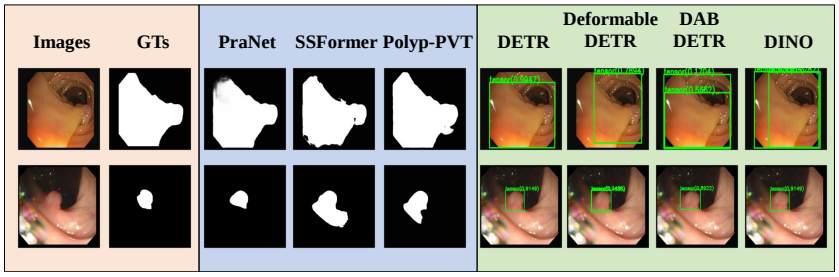

Fig. 1. The visualization results of different PS and PD methods on hard samples. Blue part shows the segmentation results. Green part shows the detection results.

Although significant progress has been achieved in both PS [8,12,18,2,14] and PD [19,10,20,3], these methods still exhibit frustrating task-specific limitations. Specifically, as shown in Fig. 1, in the scenarios with low contrast of polyp region (2^{nd} row), PS models tend to misclassify surrounding mucosa as polyps, whereas PD models can still accurately delineate the boundaries of the polyp. Besides, in cases where the polyp morphology is excessively large (1^{st} row), PD models may generate incomplete or duplicated detections, while PS could successfully identify the entire polyp. Therefore, we propose a hypothesis: integrating PS and PD into a unified framework could leverage the advantages of both approaches, ultimately eliminating task-specific limitations and enhancing overall performance.

In this paper, we propose QueryNet, a unified framework aimed at harnessing the full potential of coupling PS and PD tasks. Specifically, our QueryNet is built upon the Mask2Former [6], which is a query-based segmentation model. Object queries could be treated as instance samples with multiple spatial information. Therefore, it could locate and retrieve targets in different types of features. Exploiting this characteristic, we constructed the detection branch to allow segmentation to benefit from exploring more intricate contextual relationships. Additionally, to enhance the support of segmentation for detection, we introduce Mask-refinement Transformer Decoder by improving the feature representation from segmentation-related transformer decoder. Extensive experiments validate the mutual benefits between these two tasks, affirming the feasibility of a unified framework. To summarize, our contributions are three-fold:

1. To the best of our knowledge, our QueryNet is the first unified framework for accurate polyp segmentation and detection. Our work explores the feasibility of a unified model in the field of polyps.
2. We introduced Mask-refinement Transformer Decoder, extending the structure to enhance the utilization of segmentation-related features for detection, thereby enabling segmentation to support detection.
3. Extensive experiments indicate that the unified framework can inherit the advantages of PS and PD, thereby alleviating the limitations of each single task. Besides, QueryNet achieves comparable results against state of the art on five benchmark datasets.

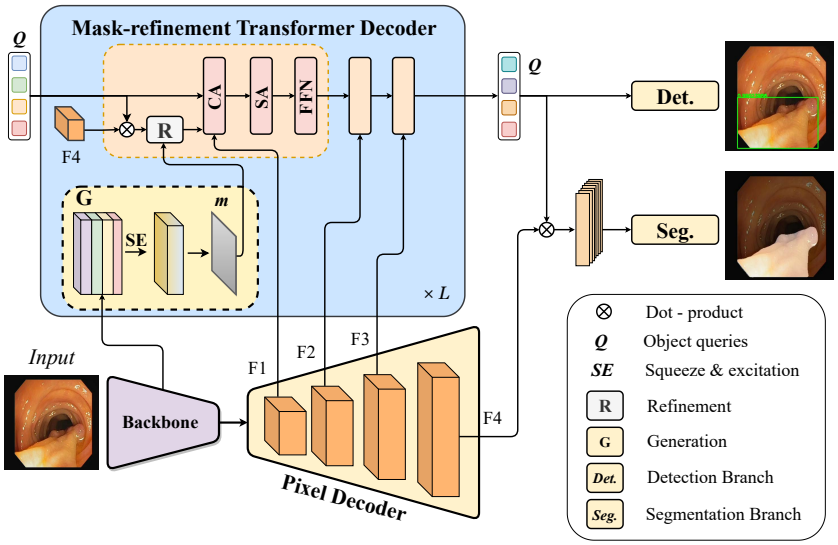

Fig. 2. Overview of the proposed QueryNet.

2 Method

As depicted in Fig. 2, our QueryNet comprises five components: backbone, pixel decoder, Mask-refinement Transformer Decoder, segmentation branch, and detection branch. We will begin by presenting the overall feature flow of the network. Then, each component will be elaborated as follows.

2.1 Overall Architecture

Firstly, the input image, denoted as "*Input*", is processed by the backbone to extract multi-scale features. These backbone features are subsequently passed to the pixel decoder, resulting in decoded features (F_1, F_2, F_3, F_4) with uniform dimensions. The pivotal step involves updating the object queries Q, accomplished by the Mask-refinement Transformer Decoder (MrTD). MrTD takes both the backbone and pixel decoder features as inputs, and detailed information about MrTD is provided in Sec. 2.2. The final object queries are then fed into the detection head to obtain detection results. By multiplying these object queries with the F_4 features, masks are generated. These masks are further forwarded to the segmentation head, resulting in the final segmentation results.

2.2 Mask-refinement Transformer Decoder

As depicted in Fig. 2, the original transformer decoder consists of three main operations: Cross Attention (CA), Self-Attention (SA), and Feed-Forward Network

(FFN). Object queries effectively assimilate valuable information from pixel decoder features (F_1, F_2, F_3) through CA. We perform these operations L times in order to fully exploit the information embedded within these pixel decoder features.

The original transformer decoder accelerates the convergence process by employing attention mask in CA. However, the original attention mask does not suit detection task. Because attention mask are obtained from scratch by dot-producting object queries and the pixel decoder feature F_4 in different decoder layers. This would cause discontinuous changes in the attention mask within the transformer decoder, consequently leading to the discontinuities of the receptive field in different transformer decoder.

To this end, we propose the Mask-refinement Transformer Decoder (MrTD) to improve the way attention masks are generated. The details of MrTD are shown in Blue Box of Fig. 2. Specifically, compared to the original transformer decoder, MrTD has two additional key components.: Generation G and Refinement R. In G, we concatenate the multi-scale encoder features, pass them through a Squeeze-and-Excitation (SE) layer [9], and finally obtain the refinement mask m through a convolutional layer. Subsequently, R utilizes m to refine the original attention mask by a logical OR operation.

MrTD could effectively addresses the issue of discontinuity in receptive field changes by employing the refined attention mask. The continuous and complete receptive field supports the detection to capture more comprehensive contextual information in transformer decoder, so as to better utilize segmentation-related features. Besides, this task-interaction module could facilitate intersection between segmentation and detection, thereby enabling features well perceive multi-task supervision in joint training.

2.3 Segmentation Branch

As suggested in [7], pixels are classified into different clusters. Each query could be conceptualized as a cluster center for an instance. The product of $Q \in R^{N \times D}$ and $F_4 \in R^{D \times H_4 \times W_4}$ denotes the similarity between each pixel and the cluster centers. Then, we apply softmax activation *softmax* to obtain the mask prediction M,

$$M = softmax(QF_4) \in R^{N \times H_4 \times W_4}, \tag{1}$$

here, N represents the number of object queries, D represents the dimensionality of each query, and H_4 and W_4 denote the height and width of F_4, respectively. Simultaneously, the object queries go through a Multi-Layer Perception MLP to predict the class probabilities for each query. Multiplying the class probabilities with the M yields the final prediction results P_m. This process can be formulated as follows:

$$P_m = MLP(Q)^T M \in R^{k \times H \times W}, \tag{2}$$

where H and W represent the height and width of the original image, k represents the number of classes. We omit the upsampling operation in our representation for simplicity.

2.4 Detection Branch

As discussed in Sec. 2.3, object queries could be conceptualized as cluster centers, which encode spatial coordinate information related to different instances. Therefore, we could extract this spatial information from object queries by using a *MLP*. This process can be formulated as follows:

$$P_b = \sigma(MLP(Q)) \in R^{N\times 4}, \tag{3}$$

here, P_b represents the prediction bounding boxes. We apply a sigmoid function σ to normalize all coordinates, in order to avoid the troubles caused by sampling operations.

In the detection branch, we are able to compute the coordinate loss associated with object queries. This detection-related supervisory information is then disseminated to each feature during the gradient update process. In this way, the segmentation could perceive the detection-related signal from these features during forward propagation. Consequently, segmentation is able to reap the advantages of detection, thereby effectively overcoming the task-specific limitations.

2.5 Loss Function

Our overall losses consist of two parts: segmentation loss and detection loss. For segmentation, we adopt dice loss and cross entropy loss. For detection, we adopt giou Loss and l1 Loss. Given P_m and P_b, the overall loss function can be formulated as:

$$L_o = \underbrace{L_d(P_m, G_m) + L_{ce}(P_m, G_m)}_{Segmentation} + \underbrace{L_g(P_b, G_b) + L_l(P_b, G_b)}_{Detection}, \tag{4}$$

where L_d, L_{ce}, L_g, L_l represent dice loss, cross entropy loss, GIoU loss and L1 loss, respectively. G_m represents the ground truth masks. G_b represents the ground truth bounding boxes.

3 Experiments

3.1 Datasets and Metrics

Datasets. We evaluate the performance of our QueryNet on five benchmark polyp datasets: CVC-ClinicDB [1], Kvasir [11], CVC-ColonDB [16], ETIS [15], and CVC-300 [17]. We follow the same setting as in [8]: 900 images from Kvasir and 550 images from the CVC-ClinicDB are used for training, the remaining images are used to test the learning ability and the other three datasets are used to test the generalization ability. **Metrics.** We use mainstream metrics to measure segmentation and detection tasks. Specifically, for segmentation, we use Dice and mean IoU; For detection, we use Precision (Pre) and Recall (Re).

Table 1. Quantitative results on the ***seen*** datasets compared to previous *state-of-the-art* methods. Seg. means segmentation methods. Det. means detection methods. The best and second best results are bolded in red and blue, respectively.

Type	Methods	Pub.	CVC-ClinicDB				Kvasir-SEG			
			Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.
Seg.	PraNet	MICCAI'20	89.40	83.49	-	-	89.14	82.91	-	-
	UACANet	ACM MM'21	**93.63**	88.87	-	-	91.38	86.13	-	-
	SSFormer-L	MICCAI'22	90.65	85.56	-	-	92.19	87.08	-	-
	Polyp-PVT	CAAI'23	93.38	88.37	-	-	**92.23**	86.91	-	-
	PVT-CASCADE	WACV'23	93.57	**88.89**	-	-	92.22	**87.24**	-	-
Det.	DETR	ECCV'20	-	-	**95.71**	**98.52**	-	-	**91.07**	**84.29**
	Deformable DETR	ICLR'21	-	-	95.52	94.11	-	-	90.19	76.03
	DAB-DETR	CVPR'22	-	-	94.02	92.64	-	-	90.65	80.17
	DINO	ICLR'23	-	-	95.45	92.65	-	-	90.20	76.03
	QueryNet(Ours)	-	**94.21**	**89.40**	**97.05**	**97.05**	**93.28**	**88.35**	**91.74**	**82.64**

Table 2. Quantitative results on the ***unseen*** datasets compared to previous *state-of-the-art* methods. Seg. means segmentation methods. Det. means detection methods. The best and second best results are bolded in red and blue, respectively.

Type	Methods	CVC-ColonDB				ETIS				CVC-300			
		Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.
Seg.	PraNet	74.70	66.08	-	-	66.55	58.14	-	-	87.50	79.74	-	-
	UACANet	75.93	68.67	-	-	77.01	69.04	-	-	**91.27**	**85.07**	-	-
	SSFormer-L	81.29	**73.52**	-	-	**80.11**	**72.80**	-	-	90.35	83.79	-	-
	Polyp-PVT	81.32	72.92	-	-	78.13	69.69	-	-	89.79	82.82	-	-
	PVT-CASCADE	**81.60**	73.47	-	-	78.59	70.83	-	-	89.15	82.25	-	-
Det.	DETR	-	-	76.37	80.79	-	-	**74.67**	**75.33**	-	-	90.16	91.67
	Deformable DETR	-	-	**79.90**	**82.63**	-	-	72.64	70.19	-	-	90.46	**91.77**
	DAB-DETR	-	-	77.55	78.16	-	-	73.63	71.15	-	-	88.52	90.00
	DINO	-	-	77.54	78.16	-	-	71.35	68.27	-	-	**91.67**	91.67
	QueryNet(Ours)	**82.78**	**75.93**	**83.51**	**85.26**	**81.89**	**73.99**	**74.88**	**77.40**	**92.05**	**85.97**	**91.80**	**93.33**

3.2 Implementations Details

Our model is implemented based on the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU. We resize the input images to the size 352×352. Random horizontal flipping and random rotation are used to avoid overfitting. The AdamW optimizer is used for optimization with an initial learning rate of 1e-4. The whole network takes approximately 5 hours to converge over 150 epochs with a batch size of 8.

3.3 Results

We compare our QueryNet with several advanced segmentation and detection methods. **Segmentation:** PraNet [8], UACANet [12], SSFormer [18], Polyp-PVT [2] and PVT-CASCADE [14]. **Detection**: DETR [4], Deformable DETR [22], DAB-DETR [13] and DINO [21]. We reproduce these segmentation models by using their released source codes, while these detection models are reproduced by using the MMDetection Framework [5]. All models are trained by five times on the same device, and the averaged results are reported for comparison.

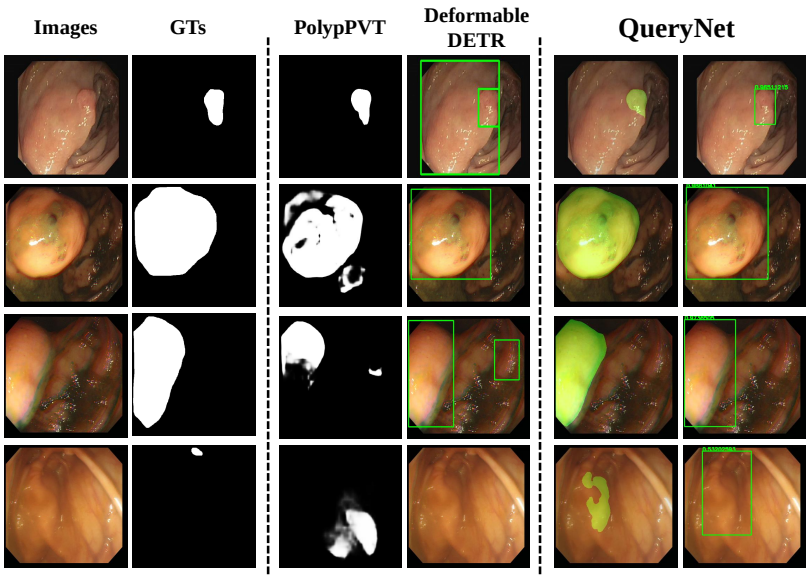

Fig. 3. Qualitative results of different methods.

Quantitative Analysis: Tab. 1 shows the intra-domain comparative results, while Tab. 2 reports the inter-domain results. It can be seen that our QueryNet exhibits robust learning and generalization abilities across all datasets of both segmentation and detection tasks. In particular, our QueryNet achieves a predominant performance on the detection task on the challenging CVC-ColonDB dataset. Compared to the second best method (Deformable DETR), our QueryNet achieves an significant improvement of 3.61% and 2.63% in terms of Precision and Recall. This indicates that segmentation could bring strong generalization ability for the detection.

Qualitative Analysis: Fig. 3 shows the visualization results of different models. It can be seen that our method has better performance for challenging polyps. The first row shows that the detection model misidentifies pseudo-polyps as polyps, while our model could achieve precise detection. This proves that the detection part could benefit from the segmentation. The second row displays that the segmentation model incorrectly segments the background due to the complex texture of the polyp. However, our model could benefit from the detection to make the segmentation results more precise and complete. Besides, the unified framework also possesses enhanced feature representation capabilities compared to single specialized model. For example, the third row presents a more challenging sample where both models fail, while our model could still accurately recognize the polyp. Undeniably, our model also exhibits flaws. As shown in the fourth row, when the target is nearly invisible, our model may oc-

Table 3. Ablation study of QueryNet on CVC-ClinicDB and CVC-ColonDB datasets. Results with underlines are computed from prediction masks. "Seg." indicates performing only segmentation task, while "Det." indicates performing only the detection task. MrTD stands for Mask-refinement Transformer Decoder.

No.	Settings			CVC-ClinicDB (***seen***)				CVC-ColonDB (***unseen***)			
	Seg.	Det.	MrTD	Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.
1	✓			91.91	87.06	<u>92.53</u>	<u>95.00</u>	78.04	71.59	<u>75.85</u>	<u>82.23</u>
2		✓		-	-	90.43	92.12	-	-	71.40	76.68
3	✓	✓		93.10	88.10	93.12	94.11	80.23	72.79	82.32	81.07
4	✓		✓	93.06	87.84	<u>95.65</u>	<u>97.06</u>	79.72	72.71	<u>81.73</u>	<u>82.33</u>
5		✓	✓	-	-	95.59	95.59	-	-	82.63	83.66
6	✓	✓	✓	94.21	89.40	97.05	97.05	82.78	75.93	83.51	85.26

casionally misidentify unrelated regions as polyps. This suggests that our model does inherit certain limitations from PS and PD in perceiving small polyps.

3.4 Ablation Study

In this section, we conduct ablation experiments of our QueryNet on CVC-CliniDB (seen) and CVC-ColonDB (unseen) datasets. We take the Mask2Former (only Seg.) as the baseline. And we haven't made any parameter changes for fairness. The ablation results are shown in Tab. 3. We can observe that: **(1)** Unified framework could benefit from joint segmentation and detection. On the seen dataset, Dice in No.3 increased by 1.19% compared to No.1, Precision increased by 2.69% compared to No.2; No.6 shows a 1.15% increase in Dice compared to No.4, and Precision increased by 1.46% compared to No.5; **(2)** Segmentation could provide strong generalization capabilities for detection. The Precision in No.3 exhibits a great improvement compared to No.2, achieving a remarkable increase of 10.92% on unseen dataset. **(3)** MrTD could effectively assist detection to utilize segmentation-related features. Compared to No.2, the Precision in No.5 significantly improves with a gain of 5.16% and 11.23% on *seen* and *unseen* datasets, respectively. **(4)** MrTD could enable the unified framework to well utilize multi-task supervision in joint training. All the indicators in No.6 have increased compared No.3, especially the detection part.

4 Conclusion

In this paper, we propose a unified framework for accurate polyp segmentation and detection, which could inherit the advantages of PS and PD to mitigate the task-specific limitations, thereby enhance the overall performance. Specifically, we exploit the characteristics of object queries and construct the detection branch, allowing the model to support both tasks. Additionally, we introduced MrTD to enable the segmentation to better support the detection. Extensive experiments validated the feasible of the unified framework, showcasing its potential in the field of polyps. Besides, our QueryNet also achieves comparable results on five benchmark datasets. We hope that our work could provide a fresh perspective for the community.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62276221, No. 62376232); the Open Project Program of Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University (No. FKLBDAITI202203).

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., de Miguel, C.R., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society **43**, 99–111 (2015), https://api.semanticscholar.org/CorpusID:1961788
2. Bo, D., Wenhai, W., Deng-Ping, F., Jinpeng, L., Huazhu, F., Ling, S.: Polyp-pvt: Polyp segmentation with pyramidvision transformers. CAAI AIR (2023)
3. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers (2020)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
8. 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)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Itoh, H., Misawa, M., Mori, Y., Kudo, S.E., Oda, M., Mori, K.: Positive-gradient-weighted object activation mapping: visual explanation of object detector towards precise colorectal-polyp localisation. International Journal of Computer Assisted Radiology and Surgery **17**(11), 2051–2063 (2022)
11. 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: International Conference on Multimedia Modeling. pp. 451–462. Springer (2020)

12. Kim, T., Lee, H., Kim, D.: Uacanet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2167–2175 (2021)
13. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
14. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6222–6231 (January 2023)
15. 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**, 283–293 (2014)
16. 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)
17. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017)
18. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: Local guides global. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 110–120. Springer (2022)
19. Yao, L., He, F., Peng, H., Wang, X., Zhou, L., Huang, X.: Improving colonoscopy polyp detection rate using semi-supervised learning. Journal of Shanghai Jiaotong University (Science) **28**(4), 441–449 (2023)
20. Yu, J., Wang, H., Chen, M.: Colonoscopy polyp detection with massive endoscopic images. arXiv preprint arXiv:2202.08730 (2022)
21. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection (2022)
22. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)