Parallel Res2Net-based Network with Reverse Attention for Polyp Segmentation

ChengHui Yu[a], JiangPeng Yan[a,b] and Xiu Li[a]

[a] Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China
[b] Department of Automation, Tsinghua University, Beijing 100084, China

Abstract

Colorectal cancer (CRC) is a commonly found and highly fatal carcinoma at the later stage. Colonoscopy is recommended for the early detection and prevention of CRC by finding and removing the leading CRC precursors, the polyps. Polyp segmentation, which helps extract polyps from colonoscopy images, is an essential step for diagnosing and developing an automatic real-time polyp classification system. In the EndoCV2021 challenge, aiming to enhance PraNet [1], we utilized a parallel res2Net-based network with reverse attention and proposed a new post-processing workflow to predict polyp segmentation masks.EndoCV2021 and three more datasets were used to test the performance and generalization ability of the proposed segmentation methodology with seven metrics. Quantitative and qualitative evaluation results show that we develop a generalizable model with excellent real-time efficiency (∼32fps) and precision / accuracy / sensitivity of 81.14%-88.94% / 94.11%-98.57% / 77.63%-91.48%, respectively. Compared to the original PraNet, our proposed method improves segmentation precision significantly by a level of 10%-29% and is more accurate with an increase of 2%-6%.

Keywords

Colonoscopy, Polyp segmentation, Deep learning

1. Introduction

Colorectal cancer (CRC) is the third most common cancer and the fourth most common cause of cancer-related death [2]. Polyps, considered the most prominent CRC precursors, could be easily removed before they advance into malignant carcinoma. The location, size, and appearance of polyps obtained during colonoscopy are crucial for CRC clinical diagnosis and follow-up treatment decisions. Therefore, numerous researches have explored the development of polyp detection and segmentation [3].

Deep learning is widely used in medical imaging analysis and computer-aided detection (CAD) systems due to its excellent feature extraction ability [4, 5]. Since MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge, more and more datasets and challenges have been launched, which further promote the application of deep learning-based endoscopic vision. Among the deep learning networks, UNet-based [6] and FCN-based [7] models are widely

3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2021) in conjunction with the 18th IEEE International Symposium on Biomedical Imaging ISBI2021, April 13th, 2021, Nice, France
ych20@mails.tsinghua.edu.cn (C. Yu); yanjp17@mails.tsinghua.edu.cn (J. Yan); li.xiu@sz.tsinghua.edu.cn (X. Li)
0000-0002-0767-1726 (J. Yan); 0000-0003-0403-1923 (X. Li)

applied in medical image segmentation for their excellent segmentation precision. However, these models concentrate more on detecting the polyp regions and less on the examination of the relationship between polyp areas and boundaries. Subsequent approaches on area-boundary investigation for polyp segmentation also have the problem of incomplete calculations [8] or time-consuming [9].

In 2020, Fan et al. [1] reported the state-of-the-art PraNet, a parallel reverse attention network to generate a recurrent cooperation mechanism between polyp areas and boundaries. PraNet first uses a parallel partial decoder to predict rough areas for polyps, then applies a reverse attention module to model the previous boundaries. It works similarly as an endoscopist, looking at the polyp's general location, then extracting its edge and mask from the local features.

To further improve PraNet, our proposed methodology starts with a parallel res2Net-based network with reverse attention to segment polyp from colonoscopy image. In addition, the segmentation results are post-processed to eliminate the uncertain pixels and distinct the segmentation edge. By doing so, we clarify the polyp boundaries and diminish large numbers of false-positive cases, thus enhancing the precision, Jaccard, and Dice value. Four datasets are used to assess the performance of the model. Evaluation results demonstrate that our method offers the advantages of high generalization capacity, real-time efficacy, precision, and accuracy. Comparing to PraNet, our method is much more precise and accurate. Therefore, the tremendous emotional stress of the patients and additional tests can be avoided in clinical practice.

2. Methods

2.1. Pre-processing

Different from data augmentation, we first split the data randomly into five equal sub-datasets to employ the five cross-validation training strategy. In this strategy, five different models were trained each time, using four datasets to find the appropriate model parameters and prevent over-fitting.

The images have a different size in the subfolder of the EndoCV2021 dataset [10]. We then resized images to the same size with a multi-scale training strategy to feed into the neural network. Using this strategy to scale the input images to different sizes, the model can better adapt to images of various sizes and increase the number of training sets.

2.2. Methods

The model we used is a Res2Net-based network, which extracts multi-level features from polyp images. Figure 1 shows the outline framework. The architecture could be divided into three processes. First, we used the feature extractors to generate five-level features that included two low-level and three high-level features. After that, the high-level features will be up-sampled and paralleled in connection to the partial decoder. We next put the high-level features and output of the partial decoder (PD) to reverse attention (RA) component, which could adjust segmentation to accurate results. Each process will be described as follows.

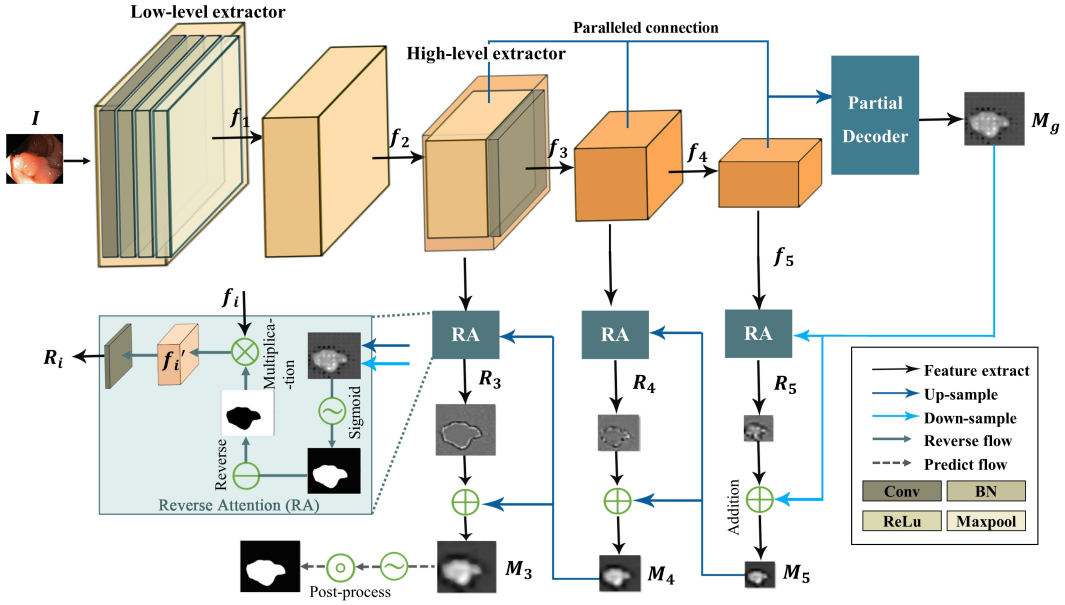


Figure 1: Framework of the model, which extracted features by a parallel network with three reverse attention components.

The Res2Net-based [11] backbone network was utilized to extract multi-level features. As Figure 1 shown, we designed five extractors, and each of them contains several network layers based on a Res2Net-based network. In details, we called the first two extractors low-level extractor and the others high-level extractor, which extract low-level features$\{f_1, f_2\}$ and high-level features$\{f_3, f_4, f_5\}$ respectively.

The high-level features would occupy fewer computing resources and contribute more to the network, thus are concentrated more than low-level features[12] in our model. Therefore, we apply paralleled connection on high-level features rather than all features. High-level features will be up-sampled and using the Concat function to be connected parallelly to be more specific. To aggregating these deep features, a partial decoder $p()$[12] component is computed by $p(f_3, f_4, f_5)$. The previous partial detector generates a global map M_g with the paralleled connection of high-level features. Using the deepest CNN network, high-level extractors will roughly extract the feature information.

Figure 2(a) shows a traditional medical image segmentation framework, which generating map M by adopting a full decoder to integrate all level features. However, Wu et al. [12] proved the high-level features would occupy fewer computing resources and contribute more to the network. Besides, they conducted visualization experiments on multi-level feature maps and found that the fifth layer still reserved edge information. Inspired by their work, our model concentrates more on high-level features and uses five-level extractors to elicit features. As Figure 2(b) shown, we applied paralleled connection on high-level features rather than all features. Therefore, high-level features are up-sampled and connected parallelly by the Concat function. To aggregating these deep features, a partial decoder p()[12] component is computed

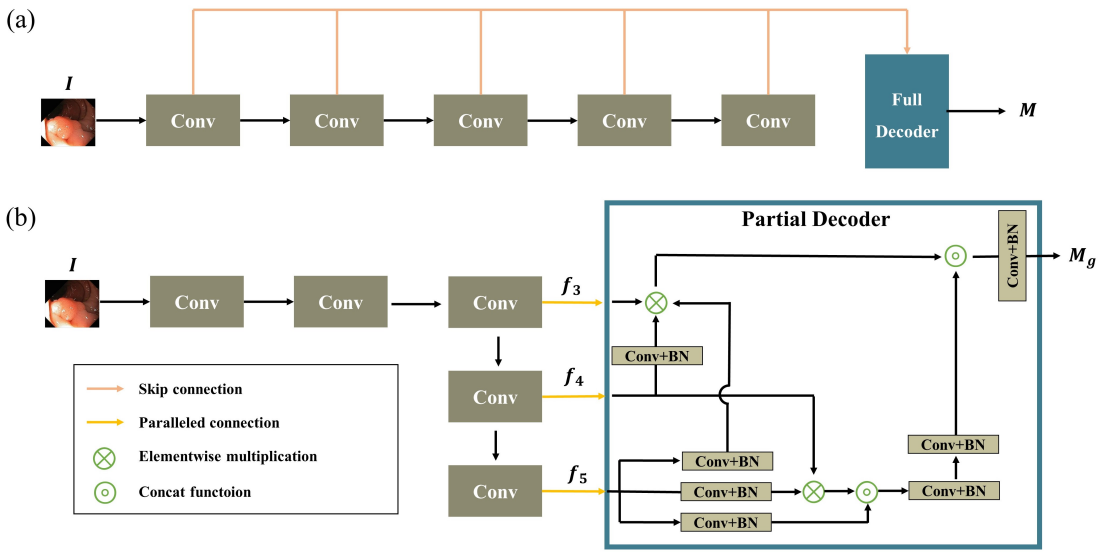


Figure 2: (a) Traditional encoder-decoder framework, (b) The paralleled partial decoder framework.

by $p(f_3, f_4, f_5)$. This partial detector generates a global map M_g. With the paralleled connection of high-level features, high-level extractors will roughly extract the feature information using the deepest CNN network.

To further detail the information, we adopt a reverse attention component for each high-level feature branch. A shallow-forward strategy is applied to ensure that the maps M_i generated by each branch are gradually refined. Specifically, the deep maps $\{M_4, M_5, M_g\}$ will be delivered to the RA component of the shallow part to compute reverse attention features $\{R_3, R_4, R_5\}$. Each deep map and reverse attention feature element-wise is multiplied and generates the shallow maps $\{M_3, M_4, M_5\}$. The map gradually sharpens by sequentially fusing with shallow maps.

Our loss function formula is represented as $L =_i l(G, M_i^{up})$, where the deep feature maps $\{M_i, i = 3, 4, 5, g\}$ will be up-sampled to the same size of the ground-truth G, and calculate the loss between each $M_i^{up}$and G by function l. Function l is defined as (1).

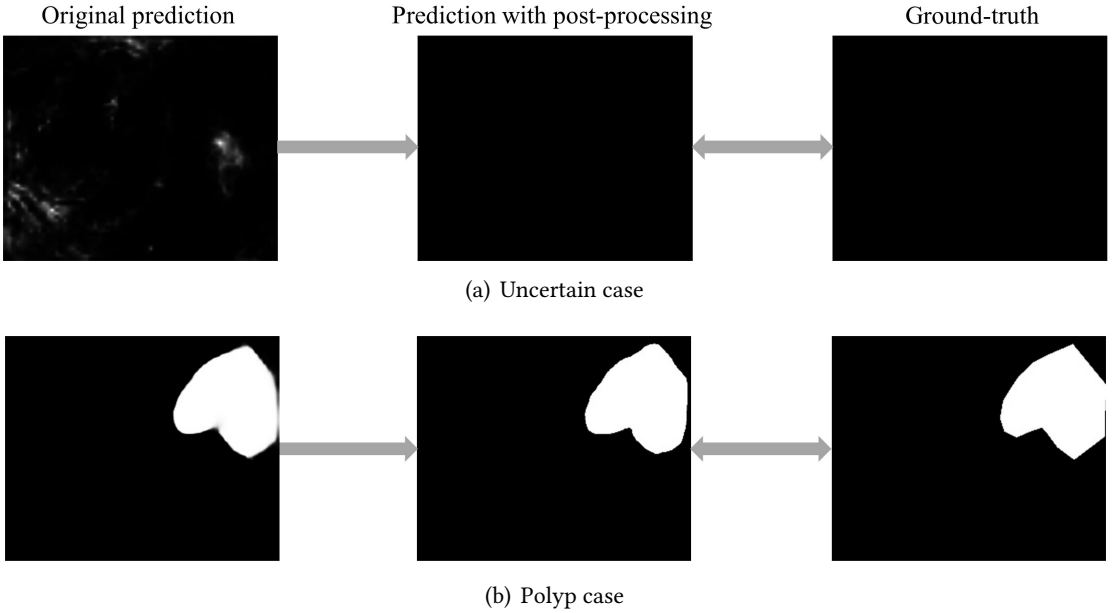
$$l = l_{IOU}^w + l_{BCE}^w \tag{1}$$

Among them, l_{IOU}^w and l_{BCE}^w express the weighted IOU loss and BCE loss, respectively. Comparing with standard IOU and BCE loss, the weighted loss would concentrate more on hard pixels instead of allocating the same weights for all pixels.

2.3. Post-processing

The original prediction, generated by the deep paralleled network with reverse attention modules, could roughly divide into polyp cases and pending cases. We can distinguish two cases easily by the presence or absence of a clear polyp site. The polyp case contains more white pixels and is more concentrated than the pending case, while most of the gray pixels distribute on the edge

of white pixels. By contrast, polyps have no presence in the pending case, so their white pixels are sparse, and gray pixels are scattered irregularly. Therefore, we introduced post-processing to the original predictions to clarify the polyp case boundary and eliminate the uncertainty.



(a) Uncertain case

(b) Polyp case

Figure 3: (a) shows a pending case of the original prediction, while (b) is a polyp case. Each pending case compares with the polyp case visual difference of the original prediction, post-processing result, and corresponding ground-truth.

Concretely, the smallest polyp size, which is the tiniest white pixel in the dataset's polyp case, is appointed as the reference value. We subsequently use the number of white pixels to judge the type of case from the original prediction. When the number of white pixels in the original prediction is more than the reference value, we consider it the polyp case because it is sufficient to form the smallest polyp. Thereafter, we set a threshold to polarize the gray pixels into black and white pixels for polyp cases—this conversion distinct the boundary and inner area of polyp clearly through polarized the pixels. For a pending case, where white pixels are rare, we consider it a non-polyp image and convert all non-black pixels to the background. The original prediction, post-processed results of each case, and the corresponding ground-truth are shown in Figure 3.

3. Experiments

3.1. Dataset and Implementation

Our model mainly used the EndoCV2021 challenge dataset [10] for endoscopic images for polyp segmentation in this work. The dataset contains 1,449 endoscopic images with ground-truth masks in C1-C5 folders. We split the dataset into 80% for training and 20% for validation.

To verify the generalization abilities of our method, we also utilize three publicly available endoscopy datasets, Kvasir-seg [13], CVC-Clinic [14], and EndoScene-CVC300 [15] for testing.

The deep models are implemented based on PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU using Adam optimizer with a learning rate of $1e-4$. The batch size is set to 16, while input images are resized to 512512 with multi-scale training parameters $\{0.75, 1, 1.25\}$.

3.2. Evaluation Metrics

To quantitatively evaluate our model's performance, seven metrics provided by the challenge for the segmentation task are employed: Jaccard (Jac), Dice, F2-score, Precision (Positive Predictive Value, PPV), Recall (Rec), Accuracy (Acc), and Hausdorff distance (Hdf). A toolbox provided by the challenge organizer at https://github.com/sharibox/EndoCV2021-polyp_det_seg_gen [16, 17] is utilized to calculate scores between each prediction and ground-truth. Also, we add the Frame Per Second (Fps) metric to evaluate the real-time performance of our model.

3.3. Experimental Results

Figure 4 presents our model's loss on the training set for varying epochs. When the epoch is less than 50, the graph describes a decreasing trend, while it remains stable after the epoch greater than 50. That is to say, the loss gradually converges downward, which proves that our model is feasible. Moreover, we experiment with different input images' sizes to attempt to obtain the appropriate size on this dataset. We can observe that when the input size is small, the loss value after convergence is relatively large, and vice versa. Among them, the size of 512512 reaches the lowest convergence value in the training phase.

Table 1 provides our model's segmentation results on EndoCV2021 C1 and CVC-ClinicDB datasets with different input image sizes. Consistent with the results in Figure 4, the larger input size performs better because it contains more semantic information, which may be lost after resizing to the smaller input size. Also, we employ the fps metric for real-time performance evaluation. Our model's real-time inference speed can reach ∼30fps even using different datasets or input sizes. That is to say, our model not only performs real-time efficiency well but also can be extended to colonoscopy videos.

Table 1
Quantitative results on the C1 and CVC-ClinicDB datasets with different size of the input image

	Size	Jac	Dice	F2	PPV	Rec	Acc	Hdf	Fps
EndoCV2021 C1	608	**0.5973**	**0.6979**	**0.7402**	**0.6756**	0.8124	0.9306	**0.3690**	31.4
	512	0.5804	0.6785	0.7196	0.6597	**0.8159**	**0.9342**	0.4528	31.6
	448	0.5389	0.6505	0.7169	0.6052	0.8340	0.9311	0.3831	31.6
	384	0.4634	0.5751	0.6492	0.5400	0.7980	0.9105	0.5173	**32.1**
CVC-ClinicDB [14]	608	**0.5519**	**0.6481**	**0.6819**	**0.6361**	**0.7427**	0.9302	0.4874	35.4
	512	0.5190	0.6133	0.6591	0.6072	0.7215	**0.9340**	**0.4433**	35.6
	448	0.4634	0.5699	0.6260	0.5565	0.7164	0.9108	0.5192	35.8
	384	0.4184	0.5268	0.5909	0.4804	0.6802	0.9093	0.5290	**36.2**

Quantitative results of model performance, including seven metrics, are presented Table 2. The results demonstrate that our model's varying component is a crucial contributor to the

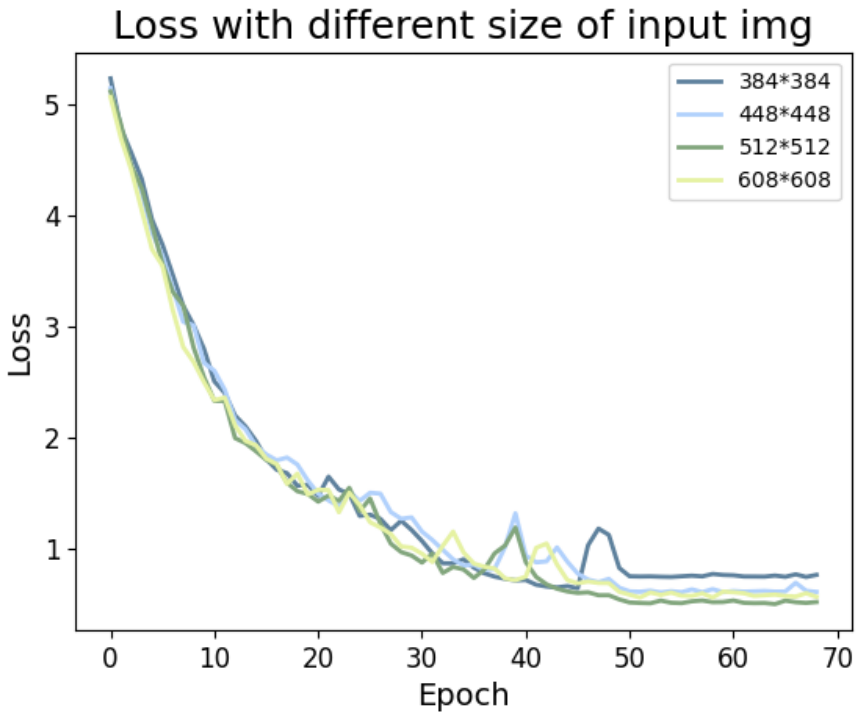


Figure 4: The model loss with different sizes of the input image on the training datasets.

improvement of the model segmentation ability. The precision, accuracy, and sensitivity are 81%-89%, 94%-98%, and 64%-80%, respectively, across different source datasets. The post-processing component enhances the Jaccard, Dice, precision, and accuracy metrics. In other words, our model can better distinguish the patients without polyps. Comparing with PraNet, the accuracy increases 2%-6%, and segmentation precision enhances a significant 10%-29%. Besides, Table 2 also shows that the method achieves excellent generability by delivering well segment precision on datasets never train the model, the Kvasir-seg, and EndoScene-CVC300 datasets.

4. Conclusion

In this manuscript, we introduce a deep model for the EndoCV2021 polyp segmentation task. The model adopts a parallel res2Net-based network with reverse attention for polyp segmentation and concentrates on extracting the high-level features. The results are predicted with a post-processing component. The experimental results indicate that our model (1). has excellent generalization ability on different datasets, (2). delivers real-time efficiency, and (3). offers better precision and accuracy that can better identify negative polyp individuals. The 512512 input size is a balance between our GPU memory and image resolution on the EndoCV2021 challenge datasets. Future works will focus on the data augmentation of polyps and lighter feature extractors to further improve the segmentation results.

Table 2
Ablation study result of all validation, Kvasir-seg, and EndoScene-CVC300 datasets

	Methods	Jac	Dice	F2	PPV	Rce	Acc	Hdf
EndoCV2021 C1-C5	Backbone+RA(BR)	0.4499	0.5400	0.5840	0.5376	0.7250	0.9082	0.4373
	Backbone+PD(BP)	0.5126	0.5993	0.6074	0.5931	0.7489	0.8821	0.3851
	BR+PD(BRP)	0.5210	0.6032	**0.6341**	0.5953	**0.7763**	0.9086	**0.3659**
	BRP+post	**0.5699**	**0.6322**	0.6199	**0.8894**	0.6373	**0.9576**	0.4134
Kvasir-seg [13]	Backbone+RA(BR)	0.6378	0.7408	0.7954	0.7023	0.8686	0.9057	0.4518
	Backbone+PD(BP)	0.6951	0.7911	0.8104	0.7547	0.8735	0.9181	0.4375
	BR+PD(BRP)	0.6972	0.7940	**0.8276**	0.7746	**0.8743**	0.9236	**0.4311**
	BRP+post	**0.7240**	**0.8082**	0.7986	**0.8687**	0.8089	**0.9411**	0.4439
EndoScene-CVC300 [15]	Backbone+RA(BR)	0.4703	0.5838	0.7081	0.4209	0.9023	0.9204	0.4818
	Backbone+PD(BP)	0.5597	0.6767	0.7730	0.5828	0.9087	0.9481	0.4215
	BR+PD(BRP)	0.5867	0.7020	**0.7933**	0.6158	**0.9148**	0.9654	**0.4139**
	BRP+post	**0.6377**	**0.7309**	0.7640	**0.8114**	0.8029	**0.9857**	0.4884

Acknowledgments

Special thanks are given to the EndoCV2021's organizing committee, clinical collaborators, program committee, and GPU cloud-based inference admin. This research was partly supported by the National Natural Science Foundation of China (Grant No. 41876098), the National Key R&D Program of China (Grant No. 2020AAA0108303), Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798), Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (Grant No. HW2018008). The authors would like to also thank Zhe Xu and Yu Yang in the Intelligent Computing Lab, the Department of Shenzhen International Graduate School, Tsinghua University for their valuable discussions.

References

[1] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2020) 263–273. doi:10.1007/978-3-030-59725-2_26.

[2] I. Mármol, C. S. de Diego, A. P. Dieste, E. C. 3OrcID, M. J. R. Yoldi, Colorectal carcinoma: A general overview and future perspectives in colorectal cancer, International journal of molecular sciences 18 (2017) 197. doi:10.3390/ijms18010197.

[3] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, J. B. Pagador, Deep learning to find colorectal polyps in colonoscopy: A systematic literature review, Artificial intelligence in medicine 108 (2020). doi:10.1016/j.artmed.2020.101923.

[4] J. Yan, S. Chen, Y. Zhang, X. Li, Neural architecture search for compressed sensing magnetic resonance image reconstruction, Computerized Medical Imaging and Graphics 85 (2020) 101784.

[5] Z. Xu, J. Luo, J. Yan, X. Li, J. Jayender, F3rnet: Full-resolution residual registration network for multimodal image registration, arXiv preprint arXiv:2009.07151 (2020).

[6] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image

segmentation, International Conference on Medical image computing and computer-assisted intervention (2015) 234–241. doi:10.1007/978-3-319-24574-4_28.

[7] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, D. Stoyanov, Fully convolutional neural networks for polyp segmentation in colonoscopy, Medical Imaging 2017: Computer-Aided Diagnosis 10134 (2017) 101340F. doi:10.1117/12.2254361.

[8] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, M. Sivaprakasam, Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019) 7223–7226. doi:10.1109/EMBC.2019.8857339.

[9] Y. Fang, C. Chen, Y. Yuan, K.-y. Tong, Selective feature aggregation network with area-boundary constraints for polyp segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019) 302–310. doi:10.1007/978-3-030-32239-7_34.

[10] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, M. A. Riegler, P. Halvorsen, C. Daul, J. Rittscher, O. E. Salem, D. Lamarque, T. de Lange, J. E. East, Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv (2021).

[11] S. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1. doi:10.1109/TPAMI.2019.2938758.

[12] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3907–3916. doi:10.1109/CVPR.2019.00403.

[13] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, International Conference on Multimedia Modeling (2020) 451–462. doi:10.1007/978-3-030-37734-2_37.

[14] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Comput Med Imaging Graph 43 (2015) 99–111.

[15] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, Journal of healthcare engineering 2017 (2017).

[16] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, et al., An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific reports 10 (2020) 1–15.

[17] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002.