Kvasir-SEG: A Segmented Polyp Dataset

Debesh Jha[1,2(✉)], Pia H. Smedsrud[1,3,4], Michael A. Riegler[1,7],
Pål Halvorsen[1,6], Thomas de Lange[4,5], Dag Johansen[2],
and Håvard D. Johansen[2]

[1] SimulaMet, Oslo, Norway
debesh@simula.no
[2] UIT The Arctic University of Norway, Tromsø, Norway
[3] Augere Medical AS, Oslo, Norway
[4] University of Oslo, Oslo, Norway
[5] Oslo University Hospital, Oslo, Norway
[6] Oslo Metropolitan University, Oslo, Norway
[7] Kristiania University College, Oslo, Norway

Abstract. Pixel-wise image segmentation is a highly demanding task in medical-image analysis. In practice, it is difficult to find annotated medical images with corresponding segmentation masks. In this paper, we present Kvasir-SEG: an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated by a medical doctor and then verified by an experienced gastroenterologist. Moreover, we also generated the bounding boxes of the polyp regions with the help of segmentation masks. We demonstrate the use of our dataset with a traditional segmentation approach and a modern deep-learning based Convolutional Neural Network (CNN) approach. The dataset will be of value for researchers to reproduce results and compare methods. By adding segmentation masks to the Kvasir dataset, which only provide frame-wise annotations, we enable multimedia and computer vision researchers to contribute in the field of polyp segmentation and automatic analysis of colonoscopy images.

Keywords: Medical images · Polyp segmentation · Semantic segmentation · Kvasir-SEG dataset · Fuzzy C-mean clustering · ResUNet

1 Introduction

Colorectal cancer is the second most common cancer type among women and third most common among men [25]. Polyps are precursors to colorectal cancer and therefore important to detect and remove at an early stage. Polyps are found in nearly half of the individuals at age 50 that undergo a colonoscopy screening, and their frequency increase with age [21]. Polyps are abnormal tissue growth from the mucous membrane, which is lining the inside of the GI tract, and can sometimes be cancerous. Colonoscopy is the gold standard for detection and
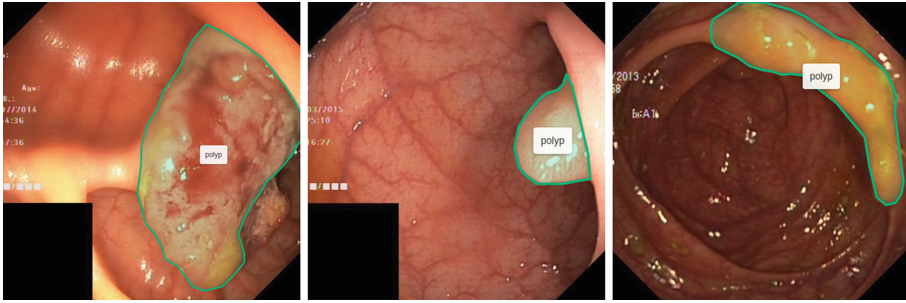
Fig. 1. Example frames from the Kvasir dataset where we additionally have marked the polyp tissue with green outlines. (Color figure online)

assessment of these polyps with subsequent biopsy and removal of the polyps. Early disease detection has a huge impact on survival from colorectal cancer [11]. In addition, several studies show that polyps are often overlooked during colonoscopies, with polyp miss rates of 14 to 30% depending on type and size of the polyps [26]. Increasing the detection of polyps has been shown to decrease risk of colorectal cancer [12]. Thus, automatic detection of more polyps at an early stage can play a crucial role in prevention and survival from colorectal cancer. This is the main motivation behind the development of a polyp segmentation dataset.

Image segmentation is the technique of dividing images into meaningful Regions of Interests (ROIs) that are simple to analyze and interpret. Further research in medical image segmentation can assist processes such as monitoring pathology, improving the diagnostic ability by increasing accuracy, precision, and reducing manual intervention [15]. In particular, for Computer Assisted Interventions (CAIs), pixel-wise semantic segmentation methods have a huge potential to become part of fast, accurate and cost-effective systems.

The goal of image segmentation is to assign a label to each pixel of the image so the pixels with the same label share specific characteristics, e.g., the pixels covered by the outline in the Fig. 1 show a polyp. Manual segmentation by physicians is still the gold standard for most of the medical imaging modalities, for example, Magnetic Resonance Imaging for evaluating hippocampal atrophy in Alzheimer's Disease [5] and tumor segmentation of glioma [27]. However, manual image segmentation is tedious, time-consuming, and subject to physician's bias and inter-observer variation. Therefore, there is a need for an automated and efficient image segmentation technique. Methods for automated and efficient image segmentation are difficult to develop as state-of-the art machine learning methods often require large number of annotated and labelled quality data, which is difficult to obtain in this field. Annotating medical data such as polyp images manually requires a lot of time and effort. It also requires medical experts, gastroenterologists in our case, which can be expensive and inaccessible. Also, there are problems related to the collection of medical images, with concern to privacy and security for patients and hospitals. Riegler et al. [19] have raised

several open questions about the medical world that need to be addressed, where they emphasized the need for test datasets, including annotations and ground truth that meet current medical standards. Although there are a few available datasets, open-access datasets for comparable evaluations are missing in this field. We therefore provide the Kvasir-SEG dataset and propose a baseline model for evaluation.

The main contribution of this paper are as follows:

1. We extend the Kvasir dataset [16] with polyp images along with their corresponding segmentation masks and bounding boxes. The ROIs are the pixels depicting polyp tissue. These are represented by a white foreground in the segmentation masks. The ROIs are generated from manual annotations verified by an experienced gastroenterologist. The bounding boxes are the set of coordinates that encloses the polyp regions. The Kvasir-SEG dataset is made publicly available and open access.
2. In this article, we include a first attempt to use the Kvasir-SEG dataset for pixel-wise semantic segmentation based analysis. For the experiment, we have used Fuzzy C-mean clustering (FCM) [6] and Deep Residual U-Net (ResUNet) [28] architecture. We achieved promising results with our proposed methods when evaluated on the same dataset. We evaluated the proposed method using Dice Coefficients and mean Intersection over Union (IoU). These metrics were selected for fair comparison, and we encourage the use of these and similar metrics in future work on the dataset. The promising results demonstrated in this paper serves as baseline and motivation for further research and evaluation done on the same dataset.
3. Multiple datasets are prerequisites for comparing computer vision based algorithms, and this dataset is useful both as a training dataset or as a validation dataset. This dataset can assist the development of state-of-the-art solutions on images captured by colonoscopes. Further research in this field has the potential to help reduce the polyp miss rate and thus improve examination quality.

This paper is organized as follows: Sect. 2 discusses related datasets. We discuss the Kvasir-SEG dataset in Sect. 3. In Sect. 4, we define the suggested metrics for the segmentation of polyps. Section 5 describes the baseline experiments, results and discussion. We conclude our work and give future directions in Sect. 6.

2   Related Work

There are only few available polyp datasets that consist of ground truth and corresponding segmentation mask. These are CVC-ColonDB [24], ASU-Mayo Clinic Colonoscopy Video © Database [3], ETIS-Larib Polyp DB [23], and CVC-Clinic DB [2].

CVC-ColonDB [24] is the second largest database available and consists of annotated video sequences from colonoscopy videos. From 15 short colonoscopy

sequences, 1200 image frames are extracted. Out of these images, only 300 frames are annotated. These annotated frames were specifically chosen to maximize the the visual differences between them. Use of the CVC-ColonDB requires registration.

The ASU-Mayo Clinic Colonoscopy Video © Database [3] is the first and largest available dataset captured using standard colonoscopes. The training dataset consists of 18,781 frames extracted from 20 short videos. Of these, there are 10 videos of polyps (positive) and 10 videos without polyps (negative). Ground truth and its corresponding segmentation masks are provided with the more than 3,500 frames showing polyps. For testing, 18 videos without ground truth are included. The images in the dataset are very similar to each other, which raise the problem of overfitting [16]. The ASU-Mayo Clinic Colonoscopy database is copyrighted, and is only available through direct contact with the administrators at Arizona State University.

The ETIS-Larib Polyp DB [23] consists of 196 frames of polyps extracted from colonoscopy videos and their corresponding masks. This database is available through registration.

The CVC-Clinic DB [2] consists of 612 image frames extracted from 29 different colonoscopy sequences and their corresponding ground truth as segmentation masks. The use of this database is public and open access.

The CVC-Clinic DB, ETIS-Larib and ASU-Mayo Clinic Colonoscopy Video DB were used at Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015 Automatic Polyp Detection in Colonoscopy Videos Sub-Challenge. More details about the dataset and competition can be found in the paper by Bernal et al. [4].

The literature review shows that there are few available datasets. However, an open-access dataset for comparable evaluation is missing in this field. Therefore, it was a logical next step to extend the Kvasir dataset with segmentation masks. The presented data and baseline work can be a important source for addressing the problem of standard datasets for evaluation, and help develop robust and efficient systems.

3   The Kvasir-SEG Dataset

The Kvasir-SEG dataset is based on the previous Kvasir [16] dataset, which is the first multi-class dataset for gastrointestinal (GI) tract disease detection and classification.

3.1   The Original Kvasir Dataset

The original Kvasir dataset [16] comprises 8,000 GI tract images from 8 classes where each class consists of 1000 images. We replaced the 13 images from the polyp class with new images to improve the quality of the dataset. These images were collected and verified by experienced gastroenterologists from Vestre Viken Health Trust in Norway. The classes include anatomical landmarks, pathological
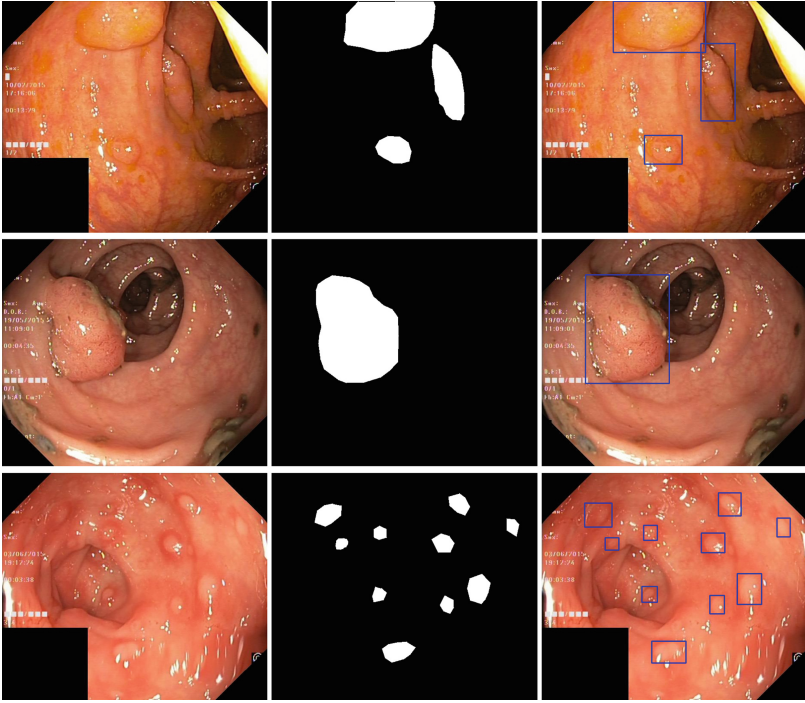
Fig. 2. Examples of polyp images and their corresponding masks from Kvasir-SEG. The third image is generated from the original image using the bounding box information from the JSON file.

findings and endoscopic procedures. A more detailed explanation about each image classes, the data collection procedure and the dataset details can be found in [16].

The Kvasir dataset was used for the Multimedia for Medicine Challenge (the Medico Task) in 2017 [20] and 2018 [17] at the MediaEval Benchmarking Initiative for Multimedia Evaluation[1] to develop and compare methods to reach clinical level performance on multiclass classification of endoscopic findings in the large bowel. However, the dataset was limited to frame classification only, due to only a frame-wise annotations. Thus, Pozdeev et al. [18] trained their model on the CVC-ClinicDB, and tried to predict the segmentation masks for the Kvasir dataset, but could not report the experimental scores because of missing ground truth.

3.2   The Kvasir-SEG Dataset Details

To address the high incidence of colorectal cancer, we selected the polyp class of the Kvasir dataset for the initial investigation. The Kvasir-SEG dataset contains

---

[1] http://www.multimediaeval.org.

annotated polyp images and their corresponding masks. As shown in Fig. 2, the pixels depicting polyp tissue, the ROI, are represented by the foreground (white mask), while the background (in black) does not contain positive pixels. Some of the original images contain the image of the endoscope position marking probe from the ScopeGuide (Olympus).

The Kvasir-SEG dataset is made up of two folders: one for images and one for masks. Each folder contains 1000 images. The bounding boxes for the corresponding images are stored in a JSON file. Therefore, the kvasir-SEG dataset has image folder, masks folder and JSON file. The image and its corresponding mask have the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. The open-access dataset can be easily downloaded for research purposes at: https://datasets.simula.no/kvasir-seg/.

3.3   Mask Extraction

We uploaded the entire Kvasir polyp class to Labelbox [22] and created all the segmentations using this application. The Labelbox is a tool used for labelling the ROI in image frames, i.e., the polyp regions for our case. A team consisting of an engineer and a medical doctor manually outlined the margins of all polyps in all 1000 images. The annotations were then reviewed by an experienced gastroenterologist.

Figure 1 shows example frames from the kvasir dataset where we have additionally marked the polyp tissue with green outline. After annotation, we exported the files to generate masks for each annotation. The exported JSON file contained all the information about the image and the coordinate points for generating the mask. To create a mask, we used ROI coordinates to draw contours on an empty black image and fill the contours with white color. The generated masks are a 1-bit color depth images with white foreground and black background. Figure 2 shows example images, their corresponding segmentation masks and bounding boxes from the Kvasir-SEG dataset.

4   Suggested Metrics

Different metrics for evaluating and comparing the performance of the architectures exist. For medical image segmentation tasks, the perhaps most commonly used metrics are Dice coefficient and IoU. These are used in particular for several medical related Kaggle competitions [10]. In this medical image segmentation approach, each pixel of the image either belongs to a polyp or non-polyp region. We calculate the Dice coefficient and mean IoU based on this principle.

Dice coefficient: Dice coefficient is a standard metric for comparing the pixelwise results between predicted segmentation and ground truth. It is defined as:

$$\text{Dice coefficient}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

where A signifies the predicted set of pixels and B is the ground truth of the object to be found in the image. Here, TP represents true positive, FP represents false positive, and FN represents the false negative.

Intersection over Union: The Intersection over Union (IoU) is another standard metric to evaluate a segmentation method. The IoU calculates the similarity between predicted (A) and its corresponding ground truth (B) as shown in the equation below:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} = \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{2}$$

In Eq. 2, t is the threshold. At each threshold value t, a precision value is calculated based on the above equation and parameters, which is done by calculating the predicted object to all the ground truth objects. There are other parameters such as recall, specificity, precision, and accuracy which are mostly used for frame-wise image classification tasks. The detailed explanation about these parameters can be found in the Kvasir dataset paper [16].

5   Evaluation

The Kvasir-SEG dataset is intended for research and development of new and improved methods for segmentation, localization, and classification of polyps. To show that the dataset is useful for these purposes, we conducted several experiments, which we will described next.

5.1   Baseline Models

As our baseline, we have conducted initial investigations using two different methods. The first method is based on the efficient FCM [6] unsupervised clustering algorithm. The second method is based on the deep-learning ResUNet [28] architecture, utilizing the advantage of the residual block.

When using basic CNN architectures to predict outcomes in computer-vision tasks, millions of labelled training data are often needed to counteract overfitting and ensuring the model's ability to generalize when tested on new data [9]. Because large datasets of medical images are hard to come by, using CNNs for medical-image segmentation systems remains challenging. Image augmentation techniques [7] and encoder-decoder architecture such as ResUNet [28] are popular methods to use CNNs with smaller training sets.

5.2   Implementation Details

Before applying the FCM algorithm, several pre-processing steps were applied to the dataset. First, we converted the image to grayscale and applied median blur to reduce noise. Then, we applied the Median-based Otsu method [14], which gave us the ROI. Next, we converted image pixels between 0 and 1 and subtracted

Table 1. Quantitative performance of ResUNet model on Kvasir-SEG dataset.

ResUNet	Loss	Dice coefficient	Mean IoU
Train	0.059389	0.940609	0.920957
Validation	0.196520	0.803479	0.792339
Test	0.212236	0.787763	0.777771

the image with its blurred version. We then used a threshold value and created an image with edges in it. Afterwards, we performed the dilation operation, which increases the foreground (white) region of the image. We subtracted edges from the image and clipped the image pixels value between 0 and 1. After that, we reshaped the image into 1D, which is the input to the FCM. Finally, the output of the FCM was reshaped into a 2D binary mask.

For our experiment with the ResUNet model, we used image augmentation techniques like flipping, random crop, scaling, rotation, brightness, cutout, and random erasing to increase the size of our training dataset. After all pre-processing was completed, we resized our images to 320×320 pixels. We used 80% of the dataset for training, and 10% for validation. The remaining of 10% was used for testing. We used five convolutional blocks both in the encoder and the decoder of the ResUNet model. The batch size was set to 8, and we trained the model for 150 epochs. The proposed model converged at 91 epochs. We used a Nadam optimizer with the learning rate of 0.0001, $\beta 1$ of 0.9 and $\beta 2$ of 0.999. We chose Dice coefficient as the loss function and Relu as non-linearity. We used a threshold value t of 0.5 to convert the predicted masks pixels to foreground or background.

For our deep-learning implementations, we used the Keras framework [8] and Tensorflow [1] as a backend. We performed our experiment on a single Volta 100 GPU on a powerful Nvidia DGX-2 [6] AI system capable of 2 PFLOPS tensor performance. The system is part of Simula Research Laboratories heterogeneous cluster and has dual Intel(R) Xeon(R) Platinum 8168 CPU@2.70 GHz, 1.5 TB of DDR4-2667 MHz DRAM, 32 TB of NVMe scratch space, and 16 of NVIDIAs latest Volta 100 GPGPUs [7] interconnected using Nvidia's NVlink fully non-blocking crossbars switch capable of 2.4 TB/s of bisectional bandwidth. The system was running Ubuntu 18.04.3 LTS OS and had the latest Cuda 10.1.243 installed.

5.3   Results and Discussions

The FCM clustering algorithm achieved a Dice coefficient of 0.239002 and a mean IoU of 0.314187. The ResUNet model achieved a Dice coefficient of 0.787763 and mean IoU of 0.777771 (see Table 1) using the test dataset. We have included the training, validation and testing scores for the ResUNet model in Table 1. Examples of qualitative result comparisons for the FCM algorithm and ResUNet model on the Kvasir-SEG dataset is shown in the Fig. 3.
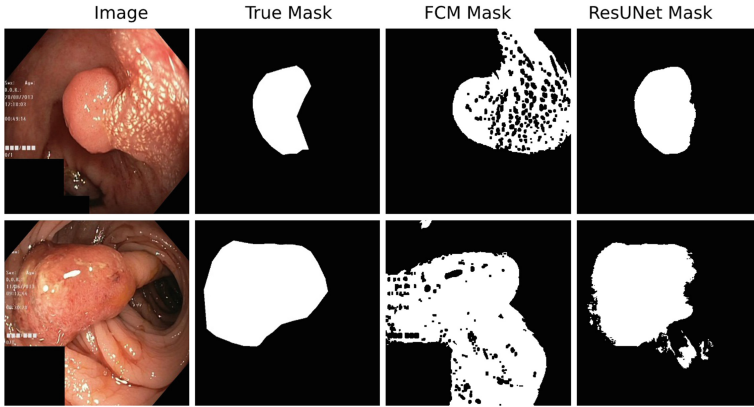
Fig. 3. Qualitative comparison provided by both methods: Coloumn one shows the original image, column two shows the ground truth of the corresponding image. Column three shows the result of FCM clustering and column four shows the results of ResUNet. (Color figure online)
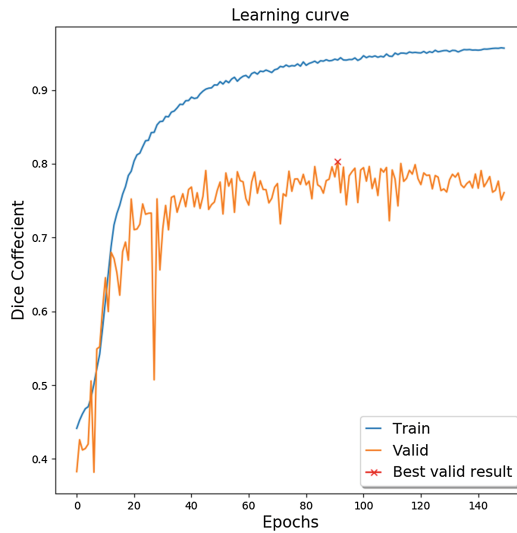
Fig. 4. The learning curve of the proposed ResUNet model on Kvasir-SEG dataset showing Dice coefficient versus number of epochs. (Color figure online)

Considering the quantitative and qualitative results (see Table 1 and Fig. 3), the study shows a superior performance of the ResUNet model over the FCM algorithm in segmenting the polyp pixels. It should be noted that the FCM algorithm uses no data augmentation because it does not have any learning mechanism or learning parameters, whereas the ResUNet utilizes the advantage of the data augmentation techniques. Another important reason why FCM

clustering approach did not perform well as it uses color as a significant feature for discriminating normal tissue and polyp. However, in practice, it is difficult to distinguish between polyps and other conditions inside the GI tract on the basis of color features because of their similar appearances. We achieved promising results with the ResUNet model.

There are no directly comparable papers with regards to our results. Nevertheless, compared to the work of Kang et al. [13] which obtained the IoU of 0.6607 and the work of Pozdeev et al. [18] that showed dice ranging from 0.6200 to 0.8600, we can say our results are either comparable or better. We think that the performance of our ResUNet model can be improved by providing it with more diverse polyp images. The plot of the learning curve of the ResUNet model is shown in Fig. 4. The red mark in the learning curve "x" denote the best model. This best model was used for testing the previously unseen test dataset. The presented results are good; however, we believe that more research is required to achieve performance applicable to the clinic.

6   Conclusion

In this paper, we present Kvasir-SEG: a new polyp segmentation dataset developed to aid multimedia researchers in carrying out extensive and reproducible research. We also present a FCM clustering algorithm and a ResUNet-based approach for automatic polyp segmentation. Our results show that the ResUNet model is outperforming the FCM clustering.

The Kvasir-SEG dataset is released as open-source to the multimedia and medical research communities, hoping it can help evaluate and compare existing and future computer vision methods. This could boost the performance of computer vision methods, an important step towards building clinically acceptable CAI methods for improved patient care.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: Proceeding of the ACM Symposium on Operating Systems Design and Implementation (SOSP), pp. 265–283 (2016)
2. 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. Comput. Med. Imag. Graph. **43**, 99–111 (2015)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012)

4. Bernal, J., et al.: Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imag. **36**(6), 1231–1249 (2017)
5. Boccardi, M., et al.: Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimer's Dis. **26**(s3), 61–75 (2011)
6. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recogn. **40**(3), 825–838 (2007)
7. Chollet, F.: Building powerful image classification models using very little data. Keras Blog (2016)
8. Chollet, F.: Keras: The Python Deep Learning Library. Astrophysics Source Code Library (2018)
9. Dravid, A.: Employing deep networks for image processing on small research datasets. Microsc. Today **27**(1), 18–23 (2019)
10. Goldbloom, A., Hamner, B., et al.: Kaggle: your home for data science. Competition, Kaggle Inc. (2019). https://www.kaggle.com. Accessed 12 July 2019
11. Haggar, F.A., Boushey, R.P.: Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clin. Colon Rectal Surg. **22**(04), 191–197 (2009)
12. Kaminski, M.F., et al.: Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. Gastroenterology **153**(1), 98–105 (2017)
13. Kang, J., Gwak, J.: Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. IEEE Access **7**, 26440–26447 (2019)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
15. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. Ann. Rev. Biomed. Eng. **2**(1), 315–337 (2000)
16. Pogorelov, K., et al.: Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of Multimedia Systems Conference (MMSYS), pp. 164–169. ACM (2017)
17. Pogorelov, K., et al.: Medico multimedia task at MediaEval 2018. In: CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval) (2018)
18. Pozdeev, A.A., Obukhova, N.A., Motyko, A.A.: Automatic analysis of endoscopic images for polyps detection and segmentation. In: IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 1216–1220. IEEE (2019)
19. Riegler, M., et al.: Multimedia and medicine: teammates for better disease detection and survival. In: Proceedings of ACM Multimedia (ACM MM), pp. 968–977. ACM (2016)
20. Riegler, M., et al.: Multimedia for medicine: the medico task at Mediaeval 2017. In: CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval) (2017)
21. Rundle, A.G., Lebwohl, B., Vogel, R., Levine, S., Neugut, A.I.: Colonoscopic screening in average-risk individuals ages 40 to 49 vs 50 to 59 years. Gastroenterology **134**(5), 1311–1315 (2008)
22. Sharma, M., Rasmuson, D., Rieger, B., Kjelkerud, D., et al.: Labelbox: the best way to create and manage training data. Software, LabelBox Inc. (2019). https://www.labelbox.com/. Accessed 21 May 2019
23. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)

24. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imag. **35**(2), 630–644 (2015)
25. Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: Cancer J. Clin. **65**(2), 87–108 (2015)
26. Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. **101**(2), 343 (2006)
27. Visser, M., et al.: Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage: Clin. **22**, 101727 (2019)
28. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-net. IEEE Geosci. Rem. Sens. Lett. **15**(5), 749–753 (2018)