Artificial Intelligence for Colonoscopy: Past, Present, and Future

Wallapak Tavanapong, *Senior Member, IEEE*, JungHwan Oh, Michael A. Riegler, Mohammed Khaleel, Bhuvan Mittal, and Piet C. de Groen

***Abstract*—During the past decades, many automated image analysis methods have been developed for colonoscopy. Real-time implementation of the most promising methods during colonoscopy has been tested in clinical trials, including several recent multi-center studies. All trials have shown results that may contribute to prevention of colorectal cancer. We summarize the past and present development of colonoscopy video analysis methods, focusing on two categories of artificial intelligence (AI) technologies used in clinical trials. These are (1) analysis and feedback for improving colonoscopy quality and (2) detection of abnormalities. Our survey includes methods that use traditional machine learning algorithms on carefully designed hand-crafted features as well as recent deep-learning methods. Lastly, we present the gap between current state-of-the-art technology and desirable clinical features and conclude with future directions of endoscopic AI technology development that will bridge the current gap.**

***Index Terms*—Artificial intelligence, medical image analysis, real-time systems, machine learning, colonos copy.**

I. Introduction

OVER the past two decades, automated analysis of endoscopic images recorded during colonoscopy has become a research area of great interest. Colonoscopy is the gold-standard for prevention of colorectal cancer (CRC), because during colonoscopy endoscopists can examine the entire colon and remove all premalignant lesions. Therefore, timely enrollment in a colonoscopy-based screening program in principle should prevent most CRC. Yet, despite stool-based and colonoscopy-based screening programs in many countries, CRC still causes significant morbidity and mortality [1]. In 2020, there were 935,173 deaths worldwide [1] and around 53,200 deaths in the U.S. [2].

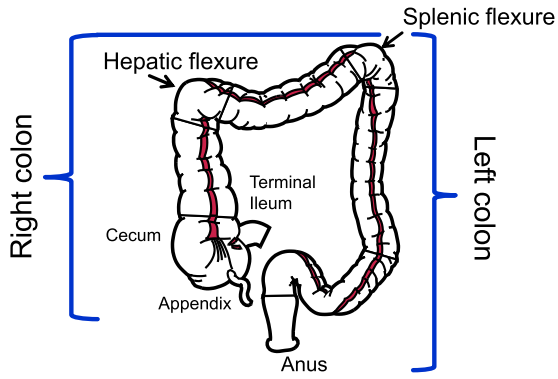

Fig. 1. Diagram showing the colon anatomy.

The colon is about five feet (150 cm) long and nested inside the human abdomen. Fig. 1 shows the anatomy of the colon. In the first phase of colonoscopy the endoscopist advances a flexible endoscope with a single wide-angle camera lens at the tip from the anus upstream with the intent to reach the cecum. The second phase starts at the point of maximum intubation; from this point the endoscope is gradually withdrawn. Careful examination behind colon folds and angulations is performed during the withdrawal phase by flexing the tip and torquing the shaft of the instrument to maximize mucosal coverage and avoid missing any abnormality located outside the longitudinal or axial view with the tip of the instrument in the neutral, straight position. At the same time premalignant lesions are removed. Both inspection and removal of lesions can vary from easy to difficult; successful completion of both, especially within a limited time, requires an advanced skill set, which explains why colonoscopy is an operator-dependent procedure.

In the early years of colonoscopy image analysis, image processing was typically used to extract carefully designed features as input to traditional machine learning methods for decision making. The last decade has seen a significant growth in supervised deep-learning (DL) methods for colonoscopy with automated feature learning from raw training images for prediction. Two surveys focusing on development of analysis methods [3], [4] were written by computing researchers. Readers interested in analysis methods for colonoscopy including pre-

Manuscript received 19 August 2021; revised 28 February 2022; accepted 12 March 2022. Date of publication 22 March 2022; date of current version 9 August 2022. This work was supported in part by NIH under Grant 1R01DK106130-01A1. *(Corresponding author: Wallapak Tavanapong.)*

Wallapak Tavanapong and Mohammed Khaleel are with the Iowa State University, Ames, IA 50011-2140 USA (e-mail: tavanapo@iastate.edu; mkhaleel@iastate.edu).

JungHwan Oh and Bhuvan Mittal are with the University of North Texas, Denton, TX 76203 USA (e-mail: junghwan.oh@unt.edu; bhuvanmittal@my.unt.edu).

Michael A. Riegler is with SimulaMet, Oslo, Norway and UiT The Arctic University of Norway, 9019 Tromsø, Norway (e-mail: michael@simula.no).

Piet C. de Groen is with the University of Minnesota, Minneapolis, MN 55455 USA (e-mail: degroen@umn.edu).

Digital Object Identifier 10.1109/JBHI.2022.3160098

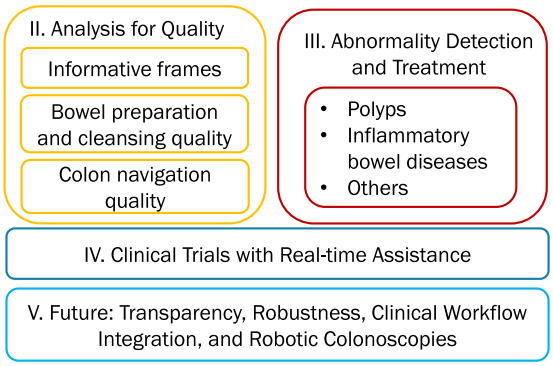

Fig. 2. Overview of the topics (in Sections II-V) summarized in this survey.

and post-procedure analysis (e.g., content-based video retrieval, efficient storage, efficient video interaction and browsing) as well as analysis for other types of minimum invasive endoscopy surgeries are referred to [3]. Readers with interest in deep-learning methods for polyp image detection, polyp region localization and segmentation prior to 2020 are referred to [4]. The latter survey also includes information about publicly available polyp datasets and performance metrics.

Unlike [3], this survey focuses on methods aimed for real-time assistance during live colonoscopy procedures. It does not cover analysis for wireless capsule endoscopy [5] or other types of endoscopy procedures [3]. Unlike [4], we summarize methods beyond polyp detection, localization, and segmentation. Polyps must first appear in the field of view of the camera before any image analysis methods can find them. This requires a good bowel preparation by the patient and most importantly good quality inspection skills by the endoscopist after reaching the cecum. That is to (1) clean remaining fecal debris, (2) see adequate amount of frames in focus (non-blurry frames), (3) look everywhere behind folds and difficult to reach areas, and (4) perform high quality, complete polypectomy [6]. Fig. 2 outlines the topics discussed in this survey.

The future of AI-assisted colonoscopy was forecast by leading domain experts in their surveys [7]–[11]. They agree that AI systems for endoscopy are forthcoming, and anticipate that AI-assisted polyp detection systems will become widely available clinically in the next five years [8]. GI Genius owned by Cosmo Pharmaceuticals and commercialized by Medtronic received FDA approval for use in the U.S. in 2021 and this system is already in place in some hospitals. ODIN Vision's CADDIE is in an on-going clinical trial in the U.K. [12]. However, the domain experts also expressed concerns about deployment in clinical practice. We categorize these concerns into **robustness**, **transparency**, and **cost-effective integration** of AI systems into clinical workflow.

We use the term "robustness" broadly to cover a number of issues. The training datasets reported in the literature are much smaller than the amount of data generated during routine colonoscopy screening and may not represent the real world. Training images tend to represent optimal conditions, e.g., a picture with a clean colon in perfect focus. *How well does the model pre-trained on small datasets under optimal conditions generalize to real-world data under sub-optimal conditions e.g., polyps partially occluded with feces?*

We use the term "transparency" to include adequate disclosure about ground truth training data such as the number of training images, the diversity of the training data, inherent biases in the training data, and explainability of deep models in making predictions. Our contributions are as follows.

- We summarize existing research aimed for real-time assistance during colonoscopy in three subcategories: analysis of the quality of the colon inspection, analysis for abnormalities and treatment, and the clinical trials using real-time AI-assisted technology. The summary of the quality of the colon inspection methods and the feedback used in clinical trials were not included in the existing surveys.
- As deep learning models are prevalent in present and future AI systems for colonoscopy, it is important to focus on improving robustness and transparency of deep-learning models for colonoscopy in clinical use. This topic has received the least research attention and was not included in detail in the existing surveys. We summarize existing methods that were applied to colonoscopy.
- We discuss future research directions including robustness and transparency, integration with clinical workflow, and robotic colonoscopies.

Many methods were proposed and evaluated over the years. Because there are few publicly available annotated datasets, many researchers used their private datasets for performance evaluations. The available public datasets [13]–[18] are mostly for polyp detection and segmentation. They are relatively small and have images taken under an optimal condition. They do not yet represent a large variety of colonoscopy images in clinical use. Due to these limitations, we do not compare existing methods directly, but present them in a chronological order. Except the topic of colon navigation techniques via 3D reconstruction, we also omit performance reports based on evaluations using small private test datasets (i.e., fewer than 3,000 images or fewer than 10 full length colonoscopy videos).

II. Analysis for Objective Quality Measurements

Objective measurements of quality of colonoscopy are important to reduce subjective biases and differences among endoscopists [19]. We focus on three key measures of quality of colonoscopy [20]: the amount of blurry (non-informative) images during the withdrawal phase, the quality of bowel preparation by patients prior to colonoscopy and the effort to remove remaining debris by the endoscopist, and the quality of the endoscope navigation inside the colon. The latter remains very challenging to solve, but has recently gained more interest due to its significance to the clinical outcome.

A. Informative Frame Analysis

An informative frame in a colonoscopy video can be broadly defined as a frame in focus and useful for analysis of the colon mucosa [21]. If most frames during the withdrawal phase of

the procedure are non-informative or blurry, then a significant part of the mucosa may not have undergone adequate inspection. Furthermore, distinguishing non-informative frames from informative ones early can improve accuracy of analysis of colonoscopy video frames for other purposes such as detection of abnormalities. Several features can distinguish non-informative frames from informative ones: corner and edge features matched with the previous frame, the percentage of edge pixels, and the mean and standard deviation of intensity in HSV (hue-saturation-value) color space were investigated in [22]. A Random Forest classifier was used for classification. An enhanced edge detection-based method was proposed in [23], [24]. Non-Informative frames usually do not contain many edges. However, very bright regions due to specular reflections can produce false edges. Therefore, the proposed method includes bright region segmentation to identify and remove false edges.

A Convolutional Neural Network (CNN) model was used for the first time for this problem in [25]. Inadequate or improper bowel preparation is characterized by remaining debris and cleansing agent which are causes of non-informative frames. SimpleNet (CNN implemented from scratch by the authors), AlexNet [26], GoogLeNet [27] and ResNet [28] were compared in terms of accuracy and speed using a dataset of about 12,000 frames. The experimental results showed that the CNN methods were fast at detecting non-informative frames with accuracies of 70 to 95%.

Hand-crafted and deep learning features from a pre-trained Inception-v3 model were combined in [29] to classify non-informative images. Although the required computation time was high, experiments based on around 17,000 frames showed an average Area-Under-the-Curve of 93.9% and an average F1 score of 77.5%. Resnet18 with Long-Short-Term-Memory (LSTM) or Gated-Recurrent-Unit (GRU) was proposed to learn from the temporal sequence of frames to predict the informativeness [30]. Gradient-weighted Class Activation Map (Grad-CAM) [31] interpretation was used to localize the informativeness within a frame. The Resnet18 extracted features were input to three separate classifiers, namely, the fully connected network, LSTM, and GRU.

B. Bowel Preparation and Cleansing

Bowel preparation (cleansing) is a key precondition for a successful colonoscopy. The degree of bowel cleansing affects successful disease detection. Therefore, an accurate assessment of bowel preparation quality is important. The Boston Bowel Preparation Scale (BBPS) [32] is a widely used bowel preparation quality assessment score. BBPS measures the individual cleanliness of three colon segments (ascending colon, transverse colon and descending colon) with a score ranging from 0 (dirtiest) to 3 (cleanest); the addition of the segmental scores provides the overall BBPS score.

Informative frames were classified by Support Vector Machine (SVM) into frames with and without remaining debris in [33]. A CNN with two DenseNet layers which have a feature reuse mechanism embedded before the softmax classifier was proposed to estimate BBPS scores [34]. This method achieved an accuracy of 90% based on the public Nerthus dataset [16]. EndoAngel based on a CNN architecture outputs bowel preparation scores every 30 seconds during the withdrawal phase of colonoscopy [35]; an accuracy of 89% was achieved over 20 colonoscopy videos.

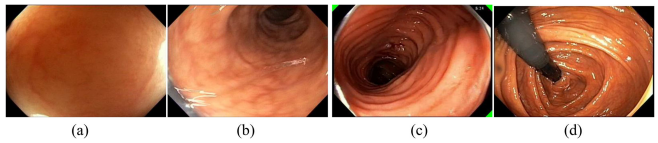

Fig. 3. Examples (a) wall view; (b) lumen view; (c) spiral score and feedback; (d) retroflexion for viewing a difficult-to-reach area.

C. Analysis of Navigation Quality

Circumferential or 360 degrees inspection of the colon mucosa throughout the withdrawal phase of colonoscopy leads to high quality of colonoscopy and greatly reduced mortality from CRC [6]. A few objective measurements have been proposed and analysis methods to derive these metrics were introduced.

1) Inspection Coverage of Colon Mucosa: Liu *et al.* [36] proposed the first objective metric called "Quadrant Coverage Histogram (QCH)" based on the domain knowledge that both distant and close up inspection should be performed during the withdrawal phase of colonoscopy. To compute QCH, an SVM classifier separates informative frames into two classes: "wall view" and "lumen view". Wall views are informative frames without the lumen, which represents close up inspection of the colon wall (Fig. 3(a)). Lumen views are informative frames with the colon lumen seen in the distance. Given a lumen view, the quadrant of the colon the endoscopist is focusing on is estimated to be the opposite quadrant where the lumen locates. For instance, in Fig. 3(b), the lumen is in the top right quadrant and the inspected quadrant is the lower left quadrant. QCH score is the average number of quadrants seen in a given duration (time window). A QCH of one indicates that only one side of the colon is inspected by the endoscopist.

Later on, "spiral score" was proposed [37] where a "spiral" is defined as a completion of inspection of four different quadrants of the colon considering only the lumen views. The spiral score is a count of the number of "spirals" performed thus far. The more "spirals," the more likely a high-quality inspection of the colon. Fig. 3(c) shows the spiral score as the white text on the top right corner and three little green triangles indicating the quadrants that had been inspected. Hong *et al.* improved the method for calculating spiral scores [38] based on detection of colon fold edges and the center of the innermost haustral fold. Feedback showing the spiral scores was used in a single center clinical trial on ten GI-trainees [39] over 159 colonoscopy procedures. The study found the spiral score feedback resulted in statistically significant improvement in quality of the colon examination. The spiral score was one of the key factors of the first single automated score of quality of colonoscopy. The automated score was developed using a mixed stepwise logistic regression model and validated on 200 full colonoscopy procedures [40].

The spiral score is a coarse estimation of how well the endoscopist looks everywhere during colonoscopy. Several attempts were made to obtain a more detailed estimate via 3D reconstruction of a virtual colon from a sequence of colonoscopy images. None of techniques have reported real-time performance in live colonoscopy procedures. The research challenge is the lack of detailed ground truth of camera depths and motion parameters during colonoscopy.

Zhou *et al.* [41] proposed a method to generate a small 3D colon segment by using optical flow analysis to align neighboring 3D circles (approximating colon folds) and determine the distance among them. The limitations of this method are as follows. Some colon segments (e.g., transverse colon) are not circular. Partial occlusion of colon folds is typical in practice. Lastly, colon fold thickness is not modeled, which is important as polyps may be hidden behind folds. Two methods for 3D reconstruction given a single image [42] and sequential images [43] were proposed. These techniques do not have the aforementioned limitations. For reconstruction from a single image, the method [42] first estimates colon fold contours and places the detected folds in 3D space via reverse projection and depth estimation from non-specular pixels. Next surface of the colon folds and surface between folds are generated to complete the reconstruction of a 3D virtual colon segment from a single image. The percentage of the colon mucosa area not seen in the field-of-view of the camera and a 3D map of the unseen areas are estimated from a reconstructed 3D colon via a simulation of a simple fly through inside the virtual colon from the first colon fold to the last fold without lateral tip deflection [44]. These metrics are of interest to objectively estimate how well the endoscopist inspects the colon and which areas have not been inspected. The method in [43] tracks detected colon fold edges across a sequence of images and reconstructs a corresponding 3D colon segment and camera motion parameters.

Mahood and Dur proposed a deep-learning method that reconstructs 3D surface of a colon from a single image [45]. The proposed network takes an over-segmented input image and outputs the predicted depth map. The network architecture consists of three major components. One component consists of five convolutional layers followed by four fully connected layers. Another component is a fully connected layer that takes neighborhood pairwise superpixel similarities from the over-segmented input image. The output of both components are input to the third component—the conditional random field layer. In [46], the authors improved upon their previous method [45] using a fully convolutional network to generate convolutional feature maps and nearest neighborhood upsampling to generate superpixel feature vectors.

Colonoscopy Coverage Deficiency via Depth (C2D2) was proposed to predict the colonoscopy coverage [47]. Coverage score per colon segment was defined as a fraction of the colon mucosa in the field of view of the camera to all visible mucosa area in a colon segment. The colon segment model is simplified by excluding fold thickness from the model. Hence, the mucosa area behind folds is not taken into account in the calculation of the coverage score. C2D2 uses ResNet-18 to estimate a depth image directly from an RGB input image and estimates camera intrinsics as well as the camera pose (translation vector and rotation matrix) between this frame and its preceding frame. To predict the coverage score for a segment, two additional neural networks are used. The first network is a 2D CNN (ResNet-50) with the global spatial average pooling (GAP) layer before the fully connected (FC) layer. The second network takes a sequence of frame feature vectors output by GAP of the first network and predicts the coverage score for a sequence of 300 frames. This network has a 1D CNN followed by GAP (in the temporal domain) before the FC layer. The total time for all stages was less than 17.07 ms per frame. The Google-Synthetic dataset used is available upon request; it has 187,369 (RGB, depth) image pairs with a train-test split of 134,025 and 53,344, respectively [47]. For qualitative evaluation, two domain experts were asked whether they agreed with the reconstruction results on 301 real colonoscopy sequences. Each sequence was rated by one domain expert. The experts agreed with the results in 93% of the sequences.

Armin *et al.* [48] proposed a CNN that predicts the colon center line (a set of points in the middle of the colon lumen) and camera direction from a sequence of colonoscopy frames. The network is based on VGG16, but takes a pair of consecutive frames as input. By modeling a colon segment as a cylinder, a colonoscopy frame is projected onto the cylinder and unrolled into a radial strip called a "band image". Band images of consecutive frames are then stitched together based on average motion flows to form a "visibility map". Ma *et al.* [49] proposed RNNSLAM integrating a localization and mapping method and depth and pose estimation neural networks to reconstruct 3D colon segments. Blau *et al.* [50] proposed an unsupervised learning technique for estimating examination coverage on colon segments modeled as bent cylinders. The work by Zhang *et al.* [51] and Mathew *et al.* [52] utilizes pre-procedure CT scans for reconstruction of 3D colon segments. Abrahams *et al.* [53] proposed to predict blindspots at acute bends in the colon assuming a known colon centerline, the camera's pose relative to the model, and a torus colon model with fixed-diameter circular cross-sections and straight or bent centerline. Lastly, Ma *et al.* [54] made available a Colon 10K dataset for evaluation of methods for finding the region in the colon in the current colonoscopy given an image taken from the same patient in a previous colonoscopy.

Several challenging research problems remain, for instance, 1) modeling deep haustral folds where polyps may be hidden, 2) handling low-texture and intensity variations, and presence of instruments, debris, and water, 3) modeling non-circular colon segments and geometric distortion of the colon, and 4) quantitative evaluation of the reconstruction of the colon from a full-length colonoscopy procedure since there is no quantitative ground truth of the true structure of the colon during colonoscopy.

2) Retroflexion Detection: Retroflexion is an endoscope maneuver where the tip of a flexible endoscope equipped with a wide angle lens is deflected more than 90 degrees from the axial direction of the shaft of the endoscope. Retroflexion allows examination of the colon in difficult to reach areas such as the hepatic flexture and peri-anal mucosa in the rectum. Fig. 3(d)

shows retroflexion in the right colon. Rectal retroflexion was suggested as an essential part of the colonoscopy examination [55]. Studies reported that retroflexion improved the yield of polyps [55]–[57]. The meta analysis study [55] of six studies compared colonoscopy with right-sided retroflexion and without. The study concluded that retroflexion in the right colon improved the detection of adenomas in the right colon and recommended that it be strongly considered in the guidelines for standard of care for colonoscopy [55]. The challenge for detecting retroflexion automatically is the short duration of some retroflexions (about 1-2 seconds) and the dark appearance of the endoscope within a dark lumen. During retroflexion, the endoscope may be bent, appear in gray color or blurry due to rotation of the scope, or be partially occluded from view. The scope may also appear in a small portion of the screen blending with the black background at the edge of the endoscopic field of view. Wang *et al.* proposed pre-processing steps and hand-crafted features as input to SVM and Decision Tree classifiers to predict whether an image shows retroflexion or not [58]. Although promising, the required compute time did not allow real-time detection. Thus better methods are needed for real-time detection and quality estimation of retroflexion. For AI systems with a focus on quality of colonoscopy, an accurate estimation of amount or percent of all mucosa seen posts technical challenges, mainly due to variation in individual colon shape, the unpredictable nature of colonic contractions, and a lack of ground truth for training and verification of new AI methods. Estimates of effort of inspection, such as the spiral score, a coarse heat map of inspected mucosa or detection of retroflexion in the right colon and rectum, are starting to address the critical issue of mucosal coverage. However, more detailed methods combining new imaging with advanced mapping and AI-based interpretation systems that include AI-assisted polyp detection are needed to provide more detailed objective evidence of amount of colonic mucosa seen and number of polyps present.

III. Abnormality and Treatment Detection

A. Polyp Detection and Segmentation

Colon polyps are generally classified based on their appearance as pedunculated, sessile, or flat. Pedunculated polyps have short or long stalks. Sessile polyps grow on the surface of the colon without a stalk. Flat polyps grow along the surface of the colon. In general, sessile polyps, the head of pedunculated polyps, and flat polyps have an elliptical shape when small. Some polyps may transition into CRC. Complete removal of polyps during colonoscopy prevents the transition to CRC. Polyps vary in their appearance, shape, size, amount of protrusion, and location in the colon; to complicate matters, the same polyp may appear differently in different images due to amount of colon insufflation, degree of colon muscular contraction, angle of view, and distance from the camera. Objects between the lens such as remaining debris or instruments, may prevent polyp visualization.

Detection of colon polyps using computer assisted methods has been an active topic for research over the last two decades. During that time the focus has shifted from proof of concept work toward real-time deployment; e.g., how to achieve high detection rates while maintaining high precision in real-time [4], [59]–[61]. Early on research was focused on polyp features such as shape, color, and texture. Most methods consisted of feature engineering and used the handcrafted features for learning [62], [63]. That changed around 2016 when methods based on deep neural networks, in particular CNNs, were applied to polyp detection [64]–[67]. Performance comparison studies were reported in [4], [66], [68]. One of the most influential and first works was done by Wang *et al.* [59]. They presented algorithms and software modules for near real-time polyp detection. In addition to the algorithm a software system called Polyp-Alert was presented, which was the first complete system for automatic polyp detection. Since this report, many other studies have been completed. YOLO and similar methods [69] use deep-learning architecture for detection and localization of colon polyps. Different implementations of YOLO are mostly known and applied for their real-time capabilities. For example, Lee *et al.* [70] used YOLOv2 in their polyp detection and localization algorithm. Wan *et al.* used the latest YOLOv5 to [71] to perform polyp detection. Both articles show that YOLO-based methods have good sensitivity and near real-time performance.

Once accurate automated detection and localization of polyps was achieved, research efforts focused on pixel-wise classification or segmentation methods. Segmentation methods are intended to provide exact polyp boundaries and use every single pixel of a polyp for training. Therefore, smaller datasets can be used for training. Jha *et al.* [72] proposed a new architecture, ResUNet++. They also proposed a DoubleUNet architecture for solving the segmentation task. For a polyp segmentation task, performance metrics include Dice Coefficient, Jaccard Coefficient, precision, recall, and overall accuracy [60]. DoubleUNet is a combination of two stacked U-Nets [73] and variations of this architecture are commonly used for polyp segmentation [74]–[78]. Others used fully convolutional dilation networks to perform the analysis [79], [80]. Ali *et al.* [67] evaluated segmentation approaches against their robustness for artifacts that are part of clinical endoscopy videos and images [81].

A boundary-aware network (BA-Net) for segmentation was proposed by Wang *et al.* [82]. The architecture is based on an encoder-decoder network which captures high-level context and at the same time preserves spatial information. In [83], [84] also boundaries are taken into account to improve U-Net-based architectures. The main goal of boundary-based approaches is to take into account the information of the boundary of polyps compared to the polyp itself. Polyp segmentation using SegNet, a deep learning based segmentation model, can process around 25 frames per second [85], [86], which is seen as the border for real-time feedback during colonoscopy. Bernel *et al.* [66] compared the performance of eight different methods for polyp localization and segmentation and provided an analysis of various detection methods. Their best overall performance was a precision of 85.6,% a recall of 76.8%, and an F1 score of 81%. Their work was based on a dataset of 38 videos (20 training, 18 testing) with many near-duplicate frames. Puyal *et al.* [87] proposed a hybrid 2D/3D CNN to take advantage of both spatial and temporal information.

There are many more recent approaches and most also rely on the well-known U-Net architecture as a basis with different modifications or in different variations [88]–[91]. Interest in image segmentation of polyps remains very high and new work is appearing almost daily [92]–[100]. Generalizability of the models for polyp segmentation has become an important factor to consider. The Polyp Segmentation challenge 2021 (EndoCV 21) provided a new dataset that consisted of polyps from different centers to specifically address generalizability. Two new architectures by Thambawita *et al.* [101] performed best in the challenge. One was a triple U-Net (TriUNet) consisting of three U-Nets combined. The second one is called DivergentNets and is a combination of five different segmentation networks where each of the networks learns a different view on the data. The DivergentNet method achieved an Intersection-Over-Union or Jaccard Index of 97.6%, an F1 score of 98.6%, a recall of 98.6%, and a precision of 98.6%.

Most of the polyp segmentation datasets are rather small in terms of the number of different polyps or the number of total frames or videos. In the EndoCV 21 challenge, the PolypGen dataset was used [102]. It contains data from six different clinics and more than 300 patients. In total 3,446 annotated polyp labels with precise segmentation masks of the polyps are included. All have been verified by six senior gastroenterologists. The best reported Dice coefficient on this dataset is around 82%, implying that there is still room for improvement [102]. Kvasir-SEG is another diverse, large dataset with segmentation for 1,000 different polyps [18]. Works that performed segmentation on the Kvasir-SEG dataset report a mean Dice coefficient between 0.787 [18] and 0.918 [103].

Considering the vast amount of research on polyp segmentation, it is challenging to keep track of the open problems and what the real improvements are. Based on insights from the articles referenced in this survey, we identified the following open challenges. (1) Generalizability of segmentation methods needs to be improved. (2) Current metrics are not representing performance requirements for clinical practice. (3) Segmentation datasets are still small and they do not often represent different centers/cameras. The datasets are often imbalanced or not diverse enough, in addition to the lack of the clinical outcome for many cases. Even if performance metrics indicate great performance of most of the proposed methods, it is not clear how this performance translates into clinical practice and how it relates to imbalanced data, which is an important gap that needs to be addressed by the community.

B. Detection of Inflammatory Bowel Diseases

Ulcerative colitis (UC) is a chronic inflammatory disease of the large intestine which may extend upstream from rectum to cecum. It is characterized by periods of relapses and remissions affecting more than 750,000 in North America [104]. The therapeutic goals of UC are to first induce and then maintain disease remission. Endoscopic disease severity may better predict future outcomes of UC than symptoms. The challenges to evaluate the severity of UC objectively are non-uniform nature of symptoms associated with UC, and large variations in their patterns [105]. To assist UC diagnosis, Nosato *et al.* proposed a method [105] that uses geometrical features such as the textures of the colonic mucosa and their appearance in the colonoscopy images. The features are expressed by Higher-order Local Auto-Correlation (HLAC) [106] and Multivariate Data Analysis for classification of UC severity levels. In addition, a color conversion technique is used to enhance the ability to efficiently observe the colon conditions. Nosato *et al.* also proposed a method [107] to retrieve multi-scale objects related to UC from colonoscopy images based on HLAC. This method generates integral HLAC feature tables that are calculated using the HLAC extraction method [108].

To extract distinct textures for UC severity classification, a hybrid approach [109] uses a feature based on the accumulation of pixel value differences in combination with an existing feature such as Local Binary Pattern. A K-nearest-neighbor classifier was used to classify images into five categories: Severe, Moderate, Mild, Scar, and Normal.

Alammari *et al.* proposed a CNN-based method to objectively classify UC severity levels [110]. The first step classifies a frame into one of the 'severe', 'moderate', 'mild,' and 'normal' classes, and calculates the severity score automatically for a given video based on these classification results. Around 50,000 frames and 15,000 frames were used to train and test their CNNs, respectively. The frame-level test accuracy of 45% was reported to classify four classes.

Tejaswini *et al.* [111] proposed an improvement of [110] in two ways for better accuracy. First, essential preprocessing was added to discard out-of-focus frames, and frames containing large amounts of water or bubbles, excessive specular reflection areas, or very high uneven illumination. Second, each class of UC severity was subdivided, and more classes were generated to accommodate large variations in patterns. Each of three classes of UC such as 'Mild', 'Moderate', and 'Severe' are subdivided to 'blood' and 'non-blood' classes based on the amount of blood appearing in a frame. 'Normal' class is not divided to 'blood' and 'non-blood' classes since it is not 'Normal' if it includes any amount of blood. Thus there are a total of seven classes: 'Normal', 'Mild-blood', 'Mild-non-blood', 'Moderate-blood', 'Moderate-non-blood', 'Severe-blood', and 'Severe-non-blood'. In the next step, each of these seven classes are subdivided to 'flat' and 'non-flat' classes based on the visual contents from different viewing directions. The proposed CNN has these 14 classes to classify into the four UC classes. Around 30,000 frames and 15,000 frames were used for training and testing, respectively. The frame-level test accuracy of 61% was reported for the four class classification, which is a 15% improvement over [110]. A method to classify UC severity by detecting the vascular (vein) patterns which are defined as the amount of blood vessels in a frame was proposed [112]. To detect these vascular patterns, image pre-processing methods and three CNNs were used for classification for four UC severity levels. Around 53,000 frames and 15,000 frames were used for training and testing, respectively. The frame level test accuracy of 80% was reported, which is a 19% improvement over their previous work [111].

A GoogLeNet based model was trained using 26,304 colonoscopy images from a cumulative total of 841 patients with UC [113]. The area under the receiver operating characteristic (AUROC) was used to evaluate CNN performance in classifying

the normal mucosa and mucosal healing states (mild) using an independent test set of 3,981 images from 114 patients with UC. The study showed a high performance with AUROCs of 0.86 and 0.98 to identify normal and mild, respectively. However, this work did not consider the clinically important differences among mild, moderate and severe UC classes. To classify four different degrees of severity of the colonoscopy images with ulcerative colitis, a method using Efficient Attention Mechanism Network (EAM-Net) [114] and UC-DenseNet [115] was proposed [116]. Using 14,306 colonoscopy images, the accuracies were improved from 1% to 7% compared to the existing methods.

C. Detection of Other Types of Abnormality

In [117], a SVM based method to classify normal and abnormal colonoscopy images was proposed. It uses the image-to-class (I2C) distance measure [118] for calculation of distances among the classes. Also, it uses an extension of LBP (Local Binary Pattern) called 'discriminative feature learning' to extract the input features for SVM, which is a combination of distance metric learning [119] and discriminative subspace learning [120].

To detect colon diseases, a combination of Cross-Wavelet Transform (XWT) [121] and MSVM (Multiclass Support Vector Machine) was proposed in [122]. XWT is an extension of a conventional wavelet transform, which outputs high dimensional features. Principal Component Analysis was used to reduce the feature dimensions for MSVM.

To distinguish abnormal images with lesions that need resection (adenoma and serrated adenoma), a method using features extracted from color, texture and morphology (3D shape) of the lesions was proposed [123]. The color-GLCM (Gray Level Co-occurrence Matrix), Invariant Local Binary Patterns [124], Invariant Gabor Texture Descriptors [125], and 3D configuration Shape-from-Motion [126] features were investigated.

Narrow Band Imaging (NBI) is a video endoscopic system that uses RGB rotary filters placed in front of a white light source to narrow the bandwidth of the spectral transmittance. It provides a limited penetration of light to the mucosal surface, and enhances the micro-vessels and their fine structure on the colorectal surface. The NBI International Colorectal Endoscopic (NICE) classification system divides NBI images into Types 1–3 based on three characteristics: (i) lesion color; (ii) microvascular architecture; and (iii) surface pattern. Type 1 includes hyperplastic lesions, Type 2 includes adenoma or mucosal/submucosal scanty invasive carcinoma, and Type 3 includes deep submucosal invasive carcinoma. Kuo *et al.* proposed a two-layered SVM classifier that separates NBI images into these three types [127]. It uses the features derived from the Bank of Binarized Statistical Image Features [128].

Shang *et al.* trained multiple 121-layer DenseNet models [115] with different combinations of five training datasets (NBI Colonoscopy, white-light Colonoscopy, Esophagogastroduodenoscopy, Skin Lesion, and ImageNet) [129]. The test dataset defines non-adenomatous polyp images as benign and adenomatous polyps and cancer images as malignant. A model using MobileNetV2 [130] and DenseNet-121 [115] was proposed [131] to detect abnormalities. A summary report about the main findings from videos of gastrointestinal (GI) tract examinations can be generated using Class Activation Maps [132].

For medical image classification, a combination of data augmentation, multi-epoch fusion, and adaptive threshold selection was proposed in [133]. Data augmentation methods were randomly selected from RandomContrast, RandomBrightness, RandomGamma, Blur, MotionBlur, InvertImg, Rotate, or RandomScale. For multi-epoch fusion, the weights of each layer in the last four epochs were averaged to generate the final model. In adaptive threshold selection, various combinations of threshold values were tested to find the best one. From the datasets (Kvasir [17] and Nerthus [16]) of more than 10,000 images (16 classes), the F1 score of 0.907 and MCC (Matthew correlation coefficient) score of 0.952 were reported.

In [134], five methods in which each method is a different combination of existing classifier(s) were proposed. In Method 1, the supervised learning classifier from Weka software [135] to build a linear logistic regression model was combined with LogitBoost [136]. In Method 2, the Logistic Model Tree classifier from Weka software was used. Method 3 used only ResNet-152 [28]. In Method 4, ResNet-152 was combined with DenseNet-161 [115] using simple averaging of the final class probabilities. In Method 5, multi-layer perceptron (MLP) was used to combine the outputs from ResNet-152 and DenseNet-161 instead of the simple averaging because simple averaging does not produce an accurate classification when the two models provide different outcomes. ResNet-152 and DenseNet-161 were trained separately, and the MLP was trained using their outputs. These five methods were evaluated on the 2018 Medico dataset [137], CVC-356-plus (a modified version of CVC-356 [138]), CVC-612-plus (a modified version of CVC-612 [138]), and CVC-12k [139]. MCC scores of 0.63 to 0.94 were reported as results.

A two-stream model for endoscopic image analysis, which fuses two streams of deep feature inputs by mapping their inherent relations through a relational network model, was proposed [140]. Extracted features from earlier layers and from later layers of the pre-trained CNN model were combined to facilitate the final prediction. Their accuracy, precision, recall, F1-score, and MCC were between 0.88 and 0.99 on two public datasets (Kvasir [17] and Nerthus [16]).

A two-stage knowledge distilled framework was proposed to detect polyp, Meckel's diverticulum, ulcer, and bleeding in colonoscopy frames [141]. The accuracies between 83 and 94% on 3,799 colonoscopy images were reported. The accuracy for detection of Meckel's diverticulum is better (around 13%) than the existing work, but the accuracy for detecting polyp, ulcer, and bleeding is very similar with the others. MobileNet from the Jetson-inference software package [142] was used [143] to classify sessile polyps, pedunculated polyps, lipoma, diverticulum, bleeding, vascularized mucosa, water jet, multi-tool head, forceps, and snare) in colonoscopy frames. Accuracy was not reported.

A semi-supervised learning approach using an unsupervised jigsaw learning task [144] in combination with supervised training (ResNet-18 [28]) was proposed in [145] to classify two classes: 'neoplastic/precancerous' and 'non-neoplastic' polyps.

Using the histologic labels, adenomas and serrated adenomas were assigned to the neoplastic/precancerous class, while hyperplastic polyps were assigned to the non-neoplastic class. Several percentage improvement was reported in correctly classifying lesions when compared to a fully-supervised baseline.

D. Detection of Biopsy and Therapeutic Treatment

Colonoscopy not only allows for detailed examination of the entire colon, but also removal of all premalignant lesions during the procedure. Most often diagnostic or therapeutic operations are performed during the withdrawal phase when instruments are inserted via a working channel within the shaft of the endoscope. A variety of instruments (e.g., forceps, snares, and cytology brushes, needles for sclerotherapy or mucosal injection, and aspiration catheters) can be used. Within a single procedure, the head and the cable of the instrument typically appear in the field of view (FoV) of the camera. Detection of operations is useful for obtaining more fine-grained quality metrics such as withdrawal time without time spent for treatment and quality of treatment.

Cao *et al.* investigated methods for detecting instrument images using hand-crafted features [146], [147]. The detected consecutive instrument frames are grouped to form an *operation shot* as a segment of visual data that corresponds to a diagnostic or therapeutic operation. The proposed methods were not fast enough to run in real-time [147]. Zhang *et al.* proposed a faster method for prediction of instrument frames and detects an *instrument scene* or *operation scene* defined as a video segment corresponding to a single purpose diagnostic or therapeutic action [148]. One scene may consist of one or more operation shots such as several biopsy shots taken in close proximity in the colon. This technique, although fast, also cannot be used in real-time. The aforementioned methods thus far use hand-engineered features of cable body and cannot detect when only instrument heads appear in the FoV since instrument heads have totally different appearances.

Zhang *et al.* introduced EndoCNN with four pairs of convolutional and pooling layers, followed by a fully connected layer and a softmax layer to classify four instrument classes and one non-instrument class [149]. The authors also proposed a similarity-based data augmentation method that recommends selected unlabeled images for manual labeling to add to the seed training dataset. On the test dataset of 36,210 images, the average F1 score is 0.95 when using the similarity-based data augmentation to expand a small seed training dataset to 52,000 images. The model can run in real-time and detect instruments when only the head portions of instruments are visible as well.

IV. Clinical Trials With Real-Time AI-Assisted Colonoscopy

Although AI for colonoscopy has received much research attention over the years, there have been relatively few systems tested in clinical trials. There are nine reports of clinical trials of real-time AI-assisted colonoscopy, seven single-center [39], [150]–[155] and two three-center clinical trials [156], [157]. Four trials [39], [151], [154], [156] provided feedback on quality of colonoscopy. The systems in the remaining four trials provided feedback solely for polyps. All these trials show that AI-assisted systems improve colonoscopy outcome either by increasing quality or detecting more polyps. The first clinical trial reported in 2012 used EMIS software that detected the start and end of each procedure automatically in real-time. It measured multiple intra-procedure quality metrics: clear withdrawal time without blurry frames, amount of stool seen on images during insertion and withdrawal, BBPS scores, and the spiral score. The provided feedback consisted only of the aforementioned "spiral score." In [150], a sound alert was made when computer automated detection (CADe) [85] detected a polyp. The detected polyp bounding box was shown on a second monitor. In a later trial using the same CADe system, the same type of feedback was shown directly on the diagnostic monitor [153]. CADe design is based on SegNet [86], a deep learning method for image segmentation. In the trial reported by [157], GI-Genius provided a green prompt surrounding the detected polyp region on the diagnostic screen. The trial by Su *et al.* [151] used audio prompts when continuous blurry frames were detected. The detected polyp location was displayed on a second monitor. Su *et al.* utilized five neural-network models, four of which used features extracted from existing pre-trained models as input to shallow fully-connected neural networks [151]. They predicted cecum images to identify the beginning of the withdrawal phase of a colonoscopy, removal of the endoscope from the patient, BBPS scores, and withdrawal stability through prediction of blurry frames and similarities between frames, respectively. For polyp detection, the DL model based on YOLOv2 [69] was used.

Gong *et al.* [154] used EndoAngel to monitor withdrawal speed and colonoscopy withdrawal time using three CNN models. A warning was presented when endoscope slipping was detected (continuous blurry frames). Ten frames prior to the beginning of the slipping were displayed at the bottom of the screen until pictures similar to the ten frames were detected. The authors did not elaborate whether the frames were shown on the same diagnostic screen or another monitor. A nurse pushed a button to indicate the start of the withdrawal time if ten consecutive frames showing cecum images were not detected. The trial by Maeda *et al.* [155] required the use of endo-cytoscope and Narrow Band Imaging to study the effectiveness of their AI-system on predicting ulcerative colitis activity.

Current feedback commonly used in clinical trials shows bounding boxes surrounding detected polyps, but not the detailed polyp contours. Under ideal circumstances, polyps are removed with a margin of normal tissue surrounding the polyp; therefore polyp detection is more important than polyp segmentation. Yet, segmentation may be critical to assess completeness of resection.

V. Future of AI for Colonoscopy

Leading domain experts are optimistic about the prospective of using AI systems in daily practice for real-time assistance during colonoscopy [7]–[11]. However, they have reservations regarding three issues: robustness, transparency, and integration

with clinical workflow. We will examine the two former issues in more detail but limit the discussion of the latter given the breadth of the topic. Lastly, we will briefly discuss the potential of AI as a driver of autonomic or robotic instruments.

As DL systems are prevalent technologies for AI for colonoscopy, availability of large ground truth datasets under optimal and sub-optimal conditions is critical to advance the performance of AI assisted systems. And even with availability of optimal ground truth datasets we must define the boundaries of valid use and realize when AI-based results may have limited value. Models based on private datasets will need to disclose the patterns the models recognize and the prevalence of these patterns in the datasets. This will help to understand the limitations of the models trained on such private datasets.

A. Robustness

The effectiveness of specific CNNs is highly dependent on the training dataset [158]. Obtaining a sufficiently large and representative training dataset of population data during routine colonoscopy screening is difficult due to variations in colon anatomy, the quality of colon preparation, the navigation and inspection techniques of endoscopists, presence of unknown type and degree of disease, and endoscopists' intervention techniques and skill sets. Moreover, manual labeling of training data by domain experts is very expensive. Significant class imbalance is often found (i.e., images of the class of interest occur infrequently). For instance, the class imbalance ratio of images of non-instrument versus instrument class is about 44:1 for the instrument image classification problem [159]. What makes classification even more challenging is the fact that the common class may contain a great variety of image patterns. Significant class imbalance if not properly handled results in incorrect prediction of rare class images. Hence, creating a representative training dataset is critical and, unfortunately, often time-consuming.

Common approaches that have been applied to improve robustness of DL for colonoscopy are as follows.

- Synthetic Data Augmentation (SDA): SDA is the most commonly used method to improve image classification. SDA synthetically generates multiple variations of a training image. The most common SDA approach for colonoscopy applies user-specified image transformation methods such as rotation, zooming in/out, and cropping and translation [160]. SinGAN-Seg [161] is a generative adversarial network-based method recently proposed to generate synthetic images for polyp segmentation.
- Active Learning (AL): Given a small initial training dataset, AL methods minimize manual labeling efforts by using a query strategy to select necessary sample images (typically from an unlabeled dataset) for the domain experts to classify. A new classifier is constructed from the enlarged training dataset. This process is repeated until a stopping criterion is satisfied. Several query strategies were explored (e.g., selecting samples at the border separating different classes and selecting outlier samples). Zhou *et al.* [162] and Zhang *et al.* [159] proposed AL methods that were tested on colonoscopy datasets.
- A few-shot learning method trained on a large number of normal images and fewer abnormal images (less than 100 frames) was applied on polyp image detection [163].

More advanced SDA methods (e.g., linear and non-linear mixing of randomly cropped, labeled images and feature space augmentation) have been shown to improve classification performance for generic images [164]. Nevertheless, SDA methods are inherently limited by patterns in the training dataset before augmentation. Given a large unlabeled dataset from routine colonoscopy screening, effective AL methods are more likely to add diverse patterns seen in practice. Recent AL methods utilizing variational auto-encoder were proposed [165], [166]. Other approaches are 1) semi-supervised learning [167], [168], 2) zero and few-shot learning [169]–[171], and 3) domain-specific or out-of-domain transfer learning via supervised learning on annotated medical data [129], [172], [173].

B. Transparency of DL Models

DL methods automatically extract important image features from the training data and build the function for prediction using the extracted features. It is important to be able to determine for a given image whether the correct pixels/regions or features are used to predict the class assigned to the image. **Local interpretation** explains the prediction decision for a given image. **Global interpretation** explains an entire DL model, giving all patterns the model can recognize or patterns detected by individual or groups of neurons in various layers of the model. Global interpretation reveals the overall capability and limitations of a model. Ideally, local and global explanations are readily available, convincing and easy to understand. Limitations of a DL model may include causes such as the choice of model, training data that do not represent the target population, biases in the training data, and labeling errors. Interpretation tools may show some of the limitations of a DL model, and therefore may be useful for clinical decision makers as they provide some insights in the DL "black box" while reviewing different AI systems for possible implementation in clinical practice. For endoscopists these tools may help understand the DL classification process by showing evidence in support of or against DL-based recommendations.

1) Local Interpretation: We divide local interpretation methods into two sub-categories: pixel-based interpretation and concept-based interpretation.

Pixel-based interpretation methods assign relevance scores to individual pixels to reflect how well they support the predicted class and output the heatmap of the relevance scores. The heatmap does not explicitly convey relationships between highly relevant pixels and corresponding semantic concepts in the images of the predicted class because we do not know what features in the training data cause the final classification. Pixel-based interpretation methods mostly work on image classification problems except for a single work [174] that applied this method to polyp segmentation. Since image segmentation decides which pixel belongs to which region, the interpretation should inform the reasons for selection or rejection of the pixel as

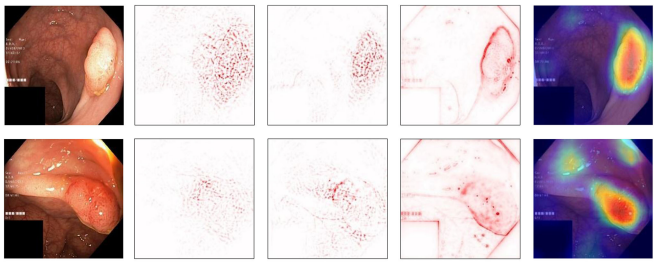

Fig. 4. Examples of pixel-based interpretation on polyp images of the Kvasir V2 public dataset [17]: (a) input image, (b) gradient, (c) LRP, (d) Deep Taylor, and (e) Grad-CAM.

part of region of interest, such as a polyp. Three main approaches for computing the relevance scores for image classification are as follows.

- **Relevance score backpropagation** methods obtain the output score of the predicted class and redistribute the score via backpropagation to the input layer. Examples include Layer-wise Relevance Propagation (LRP) [175] and Class Activation Map (CAM) [132].
- **Gradient based** methods calculate relevance scores of individual pixels as the absolute values of the gradients of the predicted class score with respect to a given input image [176]. Gradient-by-input methods calculate the relevance score of each pixel by multiplying the gradients by the output of a particular convolution layer. If needed, up-sampling the generated heatmap to the size of the input image is applied. Grad-CAM [31], Grad-CAM++ [177], DeepLIFT [178], Deep Taylor [179], and Integrated Gradient [180] are examples of this approach. These methods have received much research attention in recent years as they are applicable to any CNN architecture and allow fast interpretation calculation via a single backpropagation. In general, as the gradient based methods identify the most discriminative pixels in an input image, the interpretation output may cover only part of the discriminative object in the image, making it difficult to understand the basis for the classification.
- **Attention-based** methods learn the weights of an attention map to get the classifier to focus more on the relevant parts of the input for classification. The learned attention map is then used to create a heatmap of the relevant pixels in the input image [181]. Training the attention map adds additional computational cost, but the interpretation of a test image is fast. However, the effectiveness of the attention-based interpretation has not yet been proven [182], [183].

Fig. 4(b)–(e) show pixel-based interpretation examples by various methods for polyp image classification of Fig. 4(a). The redder the pixels are, the likelier these pixels are used by the classifier to predict the image as a polyp image. However, the interpretations do not provide insight about which edge, color, or shape patterns determine image classification. We also do not know how well such patterns are represented in the training data; are they representing rare or commonly seen polyps? That type of information is useful to improve confidence in classification results.

Concept-based interpretation methods highlight regions that represent similar concept(s) learned from the training data for the predicted class. This approach provides some knowledge about the interpretation and the relevant training data. For instance, Li *et al.* proposed to learn image-level prototypes (representation of concepts in the training data) for a DL classifier by minimizing classification loss, image reconstruction loss, and loss reflecting the distance between the learned image prototypes and training images [184].

- **Self-interpretable classification models:** These models learn to automatically extract prototypes or generalized representations of a class and use them for both classification and interpretation [185]–[187]. The prototype based self-interpretable deep classification model has a tendency to offer slightly lower classification accuracy compared to a non-prototype approach as found by the authors of [187] and [188]. Improving of accuracy can be achieved via other means such as transfer learning.
- **Contrastive explanation:** The methods in this category present images most similar to the input image but of a different class. Given image regions and corresponding text explanation for each training image, neural networks were trained to select the most suitable contrastive explanation [189], [190].
- **Hierarchical interpretation:** Wang *et al.* [191] used a manually labeled dataset [192] of color, texture, objects, scenes for generic objects to build a hierarchy of concepts for image-level and class-level interpretation. Their method requires that each concept has a set of binary segmentation mask images and the concept label as ground truth. The method cannot detect other concepts beyond the ones manually labeled. Khaleel *et al.* developed a method that automatically learns concepts at different semantic levels (e.g., color, texture, and object) from the training dataset [188] and produces a hierarchy of the concepts found in a test image as interpretation. Fig. 5 shows concept-based interpretation examples.

2) Global Interpretation: The methods in this category attempt to reveal what image properties the neural network neurons or layers detect or what patterns the model recognizes. Recent global interpretation methods are described in survey [193]. Zeiler and Fergus proposed a visualization method that shows patterns detected at intermediate layers by applying deconvolution and un-pooling operations [194]. Their method does not reveal relationships among the patterns across layers beyond spatial locations. Ghorbani *et al.* proposed to construct high-level concepts that are meaningful to humans, and coherent and important for classification [195]. Bau *et al.* proposed to dissect a CNN network by identifying which neurons in the CNN detects which concept using the intersection-over-union score between the predicted and ground truth mask [192].

To the best of our knowledge, there are no large studies that objectively evaluate any of these interpretation methods with leading domain experts in gastroenterology. We believe that global interpretation that provides patterns recognized by

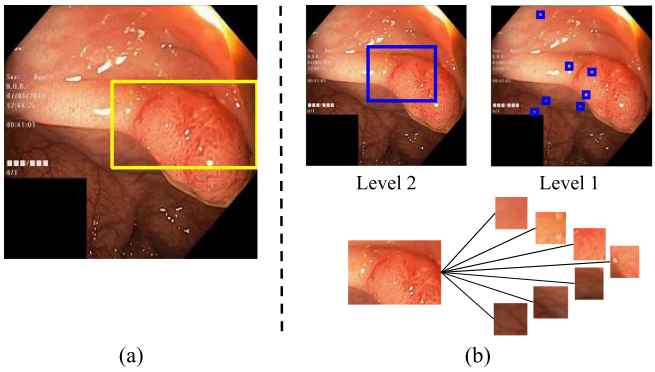

Fig. 5. Examples of (a) ProtoPNet [187] and (b) two-level hierarchical concept-based interpretation [188] on polyp images of the Kvasir V2 public dataset [17]. The blue and yellow boxes indicate the image object that is used for prediction. Level 2 shows the object level concepts (e.g., the polyp object). Level 1 shows the low-level concepts (different shades of red colors and texture) that make up the polyp object. Thicker connecting lines indicate stronger influence of the lower-level to higher-level concepts.

a model will be useful for adoption of an AI system. Local interpretation is useful for a retrospective review of performance of an AI system.

C. Integration Within Clinical Workflow

There are several potential benefits of integrating AI into the clinical workflow. First, presenting AI-generated information obtained in real-time during colonoscopy to the endoscopist is critical to improve outcome of the patient who is undergoing the procedure. Any feedback at that time potentially can change endoscopist behavior. The current focus in colonoscopy is on detection and segmentation of possible polyps, and the classification of detected polyps in likely benign or pre-malignant class. However, polyp detection, segmentation and classification can only occur for lesions within the field of view. AI can also provide information about areas of the colon not well or not at all seen [37], [47]. This information, when presented in a timely fashion, can stimulate the endoscopist to improve image clarity, remove remaining fecal matter or reposition the endoscope tip to allow visualization behind haustrae or sharp angulations. At present we do not know what information should be presented to the endoscopist, in what format, where on the monitor, and for how long.

Second, the information obtained can be used to pre-fill an endoscopy report; current endowriters require extensive clicking of entry fields to provide detailed information about preparation, findings, interventions and complications. Future methods will be able to determine all of this, select and mark appropriate image or video documentation, and document all within structured and at the same time in a easily readable format for humans. This will result in more complete procedure documentation and allow more time for actual patient-physician contact at potentially lesser cost. AI can also be used to objectively score inflammatory bowel diseases to allow comparison of patients seen by different endoscopists among centers anywhere in the world; this allows a single universal classification which may facilitate treatment optimization for patients with inflammatory bowel diseases and accelerate clinical trials of new drugs in these diseases [109], [196].

Third, AI-based information provides objective information about the quality of individual endoscopist or an endoscopy group, the average colon preparation of the patient population, the amount of time patients are within colonoscopy, disease trends of patients seen, number and type of specific instruments used, etc. Thus ample information will become available to manage endoscopic skill sets among the endoscopy team members, optimize schedules, manage the practice, maintain adequate supplies and predict practice trends.

D. Autonomic and Robotic Instruments

Knowledge of location of the endoscope tip and the location and nature of any lesions allows steering of endoscope and instruments. We foresee a gradual introduction of DL-based automation, initially under direct human supervision. Eventually standalone instruments completely driven by autonomous software may result in colonoscopy robots [197]. For instance, current manipulation of the endoscope tip is manually via dials in order to steer the tip of the endoscope in the direction of the upstream lumen; there is no reason to believe that DL cannot do this as well if not better than human operators. Patient movement, breathing and pulsating heart or vessels may move the endoscope tip away from a polyp that needs to be removed; DL-based software may automatically correct for these movements facilitating complete polyp removal. Current video capsule endoscopy does not allow steering of the capsule, obtaining samples or remove lesions; all of this in theory can be addressed, and DL is expected to play a major role in this [3]. With miniaturization and better battery technology any hardware can be located inside the body whereas the software driving a robotic capsule able to change position or remove lesions is residing outside the patient. Indeed, it is likely that predominantly hybrid robots will be applied in the colon where the tools are inside and the operating system outside the patient, either connected via a wire, also allowing power transmission, such as via the anus, or a wireless solution, requiring a battery-operated robot [198].

VI. Conclusion

We present a summary of research over the past two decades and the progress made towards real-time AI-assisted colonoscopy. Recent clinical trials have shown that feedback during live procedures improves quality of patient care by detecting more polyps. More work is to be done as described in the future research directions. Data privacy complicates matters as sharing of detailed medical image data is not allowed. Finally, having all the tools and implementing them in clinical practice does not mean that the problem at hand is solved. Perfect AI-scores for cleaning and circumferential inspection of the colon are not the same as having carefully inspected all mucosa. All it means is that the endoscopist has met the expectations of the AI-based classifiers. What eventually is needed are trials that show that AI-based techniques implemented during colonoscopy lower the incidence, morbidity and mortality of

CRC [20]. Those have been and will continue to be the ultimate indicators of successful CRC prevention; therefore AI assisted systems need to show that their implementation lowers these CRC benchmarks.

Acknowledgment

Wallapak Tavanapong and JungHwan Oh have equity interest and management roles in EndoMetric Corp. Piet C. de Groen serves on the Scientific Advisory Board of EndoMetric Corp. Findings, opinions, and conclusions expressed in this article do not necessarily reflect the view of the funding agency.

References

[1] World Health Organization International Agency for Research on Cancer, "Colorectal cancer source: Globocan 2020," [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf

[2] American Cancer Society, "Colorectum cancer statistics," 2020. [Online]. Available: https://cancerstatisticscenter.cancer.org/?_ga=2.95264916.902125337.1581945528-1873365005.1581945528#!/cancer-site/Colorectum

[3] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based processing and analysis of endoscopic images and videos: A survey," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1323–1362, 2018.

[4] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *Artif. Intell. Med.*, vol. 108, pp. 1019–1023, 2020.

[5] V. Prasath, "Polyp detection and segmentation from video capsule endoscopy: A review," *J. Imag.*, vol. 3, no. 1, p. 1, 2016.

[6] S. Xirasagar, Y. Wu, M.-H. Tsai, J. Zhang, S. Chiodini, and P. C. de Groen, "Colorectal cancer prevention by a CLEAR principles-based colonoscopy protocol: An observational study," *Gastrointestinal Endoscopy*, vol. 91, no. 4, pp. 905–916, 2020.

[7] C. Le Berre *et al.*, "Application of artificial intelligence to gastroenterology and hepatology," *Gastroenterology*, vol. 158, no. 1, pp. 76–94, 2020.

[8] D. Chahal and M. F. Byrne, "A primer on artificial intelligence and its application to endoscopy," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 813–820, 2020.

[9] P. Sharma, A. Pante, and S. A. Gross, "Artificial intelligence in endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 4, pp. 925–931, 2020.

[10] S. A. Hoogenboom, U. Bagci, and M. B. Wallace, "Artificial intelligence in gastroenterology. The current state of play and the potential. How will it affect our practice and when?," *Techn. Innov. Gastrointestinal Endoscopy*, vol. 22, no. 2, pp. 42–47, 2020.

[11] A. P. Abadir, M. F. Ali, W. Karnes, and J. B. Samarasena, "Artificial intelligence in gastrointestinal endoscopy," *Clin. Endoscopy*, vol. 53, no. 2, pp. 132–141, 2020.

[12] "Multi-centre, open-label, randomised, prospective trial to assess efficacy and safety of the caddie artificial intelligence system for improving endoscopic detection of colonic polyps in real-time," [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT04325815

[13] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[14] D. Vázquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, 2017.

[15] J. J. Bernal *et al.*, "Polyp detection benchmark in colonoscopy videos using GTCreator: A novel fully configurable tool for easy and fast annotation of image databases," in *Proc. 32nd CARS Conf.*, Berlin, Germany, Jun. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01846141

[16] K. Pogorelov *et al.*, "Nerthus: A bowel preparation quality video dataset," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 170–174.

[17] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 164–169.

[18] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Model*. Cham, Switzerland: Springer, 2020, pp. 451–462.

[19] J. Oh *et al.*, "Measuring objective quality of colonoscopy," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 9, pp. 2190–2196, Sep. 2009.

[20] P. C. de Groen, "Using artificial intelligence to improve adequacy of inspection in gastrointestinal endoscopy," *Techn. Innov. Gastrointestinal Endoscopy*, vol. 22, no. 2, pp. 71–79, 2020.

[21] J. Oh, S. Hwang, W. Tavanapong, P. C. de Groen, and J. Wong, "Blurry frame detection and shot segmentation for colonoscopy videos," *Proc. SPIE*, vol. 5307, pp. 531–542, 2004.

[22] M. A. Armin *et al.*, "Uninformative frame detection in colonoscopy through motion, edge and color features," in *Proc. Revised Sel. Papers Int. Workshop Comput.- Assist. Robot. Endoscopy*, 2015, vol. 9515, pp. 153–162.

[23] C. Ballesteros, M. Trujillo, and C. Mazo, "Automatic classification of non-informative frames in colonoscopy videos," in *Proc. Latin- Amer. Conf. Networked Electron. Media*, 2015, pp. 1–5.

[24] C. Ballesteros, M. Trujillo, C. Mazo, D. Chaves, and J. Hoyos, "Automatic classification of non-informative frames in colonoscopy videos using texture analysis," in *Proc. Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl.*, 2017, pp. 401–408.

[25] A. B. M. R. Islam, A. Alammari, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Non-informative frame classification in colonoscopy videos using CNNs," in *Proc. Int. Conf. Biomed. Imag., Signal Process.*, 2018, pp. 53–60.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[27] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[29] H. Yao, R. W. Stidham, R. Soroushmehr, J. Gryak, and K. Najarian, "Automated detection of non-informative frames for colonoscopy through a combination of deep learning and feature extraction," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2019, pp. 2402–2406.

[30] T. G. W. Boers *et al.*, "Detection of frame informativeness in endoscopic videos using image quality and recurrent neural networks," in *Proc. SPIE*, vol. 11313, 2020, Art. no. 1131315.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[32] E. Lai, A. Calderwood, G. Doros, O. Fix, and B. Jacobson, "The boston bowel preparation scale: A valid and reliable instrument for colonoscopy-oriented research," *Gastrointest Endoscopy*, vol. 69, no. 3, pp. 620–625, Mar. 2009.

[33] M. Cho, H. J. Kong, J. H. Kim, B. Jeon, K. Hong, and S. Kim, "Residue detection in the large intestine from colonoscopy video using the support vector machine method,," in *Proc. Int. Conf. Control, Automat. Syst.*, 2018, pp. 398–401.

[34] Y. Zhu, Y. Xu, W. Chen, T. Zhao, and S. Zheng, "A CNN-based cleanliness evaluation for bowel preparation in colonoscopy," in *Proc. Int. Congr. Image Signal Process., Biomed. Eng. Informat.*, 2019, pp. 1–5.

[35] J. Zhou *et al.*, "A novel artificial intelligence system for the assessment of bowel preparation (with video)," *Gastrointestinal Endoscopy*, vol. 91, no. 2, pp. 428–435, 2020.

[36] D. Liu, Y. Cao, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Quadrant coverage histogram: A new method for measuring quality of colonoscopic procedures.," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2007, pp. 3470–3473.

[37] X. Liu, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Automated measurement of quality of mucosa inspection for colonoscopy," *Procedia Comput. Sci.*, vol. 1, no. 1, pp. 951–960, 2010.

[38] D. Hong, "3D colon segment and endoscope motion reconstruction from colonoscopy video," Ph.D. dissertation, Iowa State Univ., 2012.

[39] N. Srinivasan *et al.*, "Real-time feedback improves the quality of colonoscopy by trainees: A controlled clinical trial: ACG/AstraZeneca award: 1492," *Amer. J. Gastroenterol.*, vol. 107, 2012, Art. no. S596.

[40] F. Enders, W. Tavanapong, M. J. Szewczynski, J. Oh, J. Wong, and P. de Groen, "Tu1018 objective evaluation of colonoscopy: Development and validation of an automated score," *Gastroenterology*, vol. 146, no. 5, pp. S–728, May 2014.

[41] J. Zhou, A. Das, F. Li, and B. Li, "Circular generalized cylinder fitting for 3D reconstruction in endoscopic imaging based on MRF," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–8.

[42] D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "3D reconstruction of virtual colon structures from colonoscopy images," *Computerized Med. Imag. Graph. J.*, vol. 38, no. 1, pp. 22–33, 2013.

[43] W. Tavanapong, D. Hong, J. Wong, P. de Groen, and J. Oh, "Reconstruction of a 3D virtual colon structure and camera motion for screening colonoscopy," *Med. Res. Arch.*, vol. 5, no. 6, 2017.

[44] D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "A new technology to visualize what the endoscopist does not see during colonoscopy," *Gastrointest Endoscopy*, vol. 69, no. 5, 2009, Art no. Ab366.

[45] F. Mahmood and N. J. Durr, "Topographical reconstructions from monocular optical colonoscopy images via deep learning," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 216–219.

[46] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Med. Image Anal.*, vol. 48, pp. 230–243, 2018.

[47] D. Freedman *et al.*, "Detecting deficient coverage in colonoscopies," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3451–3462, Nov. 2020.

[48] M. A. Armin, N. Barnes, F. Grimpen, and O. Salvado, "Learning colon centreline from optical colonoscopy, a new way to generate a map of the internal colon surface," *Healthcare Technol. Lett.*, vol. 6, no. 6, pp. 187–190, 2019.

[49] R. Ma *et al.*, "RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102100. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521001468

[50] Y. Blau, D. Freedman, V. Dashinsky, R. Goldenberg, and E. Rivlin, "Unsupervised 3D shape coverage estimation with applications to colonoscopy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3364–3374.

[51] S. Zhang, L. Zhao, S. Huang, R. Ma, B. Hu, and Q. Hao, "3D reconstruction of deformable colon structures based on preoperative model and deep neural network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1875–1881.

[52] S. Mathew, S. Nadeem, and A. Kaufman, "Visualizing missing surfaces in colonoscopy videos using shared latent space representations," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, 2021, pp. 329–333.

[53] G. Abrahams, A. Hervé, J. E. Bernth, M. Yvon, B. Hayee, and H. Liu, "Detecting blindspots in colonoscopy by modelling curvature," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 12508–12514.

[54] R. Ma *et al.*, "Colon 10k: A benchmark for place recognition in colonoscopy," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, 2021, pp. 1279–1283.

[55] J. M. Hanson *et al.*, "Rectal retroflexion: An essential part of lower gastrointestinal endoscopic examination," *Dis. Colon Rectum*, vol. 44, no. 11, pp. 1706–1708, 2001.

[56] H. S. Lee, S. W. Jeon, H. Y. Park, and S. J. Yeo, "Improved detection of right colon adenomas with additional retroflexion following two forward-view examinations: A prospective study," *Endoscopy*, vol. 49, no. 4, pp. 334–341, 2017.

[57] J. Cohen, D. Grunwald, L. B. Grossberg, and M. S. Sawhney, "The effect of right colon retroflexion on adenoma detection: A systematic review and meta-analysis," *J. Clin. Gastroenterol.*, vol. 51, no. 9, pp. 818–824, 2017.

[58] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Near real-time retroflexion detection in colonoscopy," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 143–152, Jan. 2013.

[59] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *Comput. Methods Programs Biomed.*, vol. 120, no. 3, pp. 164–179, 2015.

[60] D. Jha *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning.," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[61] M. Yamada *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 14465.

[62] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.

[63] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Proc. Bildverarbeitung für die Medizin* 2009, 2009, pp. 346–350.

[64] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[65] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[66] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[67] S. Ali *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020.

[68] J. Bernal and A. Histace, *Computer-Aided Analysis of Gastrointestinal Videos*. New York, NY, USA: Springer, 2021. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-64340-9

[69] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[70] J. Y. Lee *et al.*, "Real-time detection of colon polyps during colonoscopy using deep learning: Systematic validation with four independent datasets," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020.

[71] J. Wan, B. Chen, and Y. Yu, "Polyp detection from colorectum images by using attentive YOLOv5," *Diagnostics*, vol. 11, no. 12, 2021, Art. no. 2264.

[72] D. Jha *et al.*, "ResUNet : An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia*, 2019, pp. 225–2255.

[73] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Conf. Comput. Based Med. Syst.*, 2020, pp. 558–564.

[74] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2020, pp. 263–273.

[75] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Double encoder-decoder networks for gastrointestinal polyp segmentation," in *Proc. Int. Conf. Pattern Recognit.*, Cham: Springer, 2021, pp. 293–307.

[76] D. Jha *et al.*, "Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 37–43.

[77] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, "Hardnet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3552–3561.

[78] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "HarDNET-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS," 2021, *arXiv:2101.07172*.

[79] Y. Guo, J. Bernal, and B. J. Matuszewski, "Polyp segmentation with fully convolutional deep neural networks-extended evaluation study," *J. Imag.*, vol. 6, no. 7, pp. 69–90, 2020.

[80] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proc. Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory Appl.*, 2019, pp. 632–641.

[81] A. Sharib *et al.*, "Endoscopy artifact detection (EAD 2019) challenge dataset," 2021, *arXiv:1905.03209*.

[82] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," 2020, *arXiv:2005.00966*.

[83] D. Wang, M. Hao, R. Xia, J. Zhu, S. Li, and X. He, "MSB-Net: Multi-scale boundary net for polyp segmentation," in *Proc. IEEE 10th Data Driven Control Learn. Syst. Conf.*, 2021, pp. 88–93.

[84] Y. Fang, D. Zhu, J. Yao, Y. Yuan, and K.-Y. Tong, "ABC-Net: Area-boundary constraint network with dynamical feature selection for colorectal polyp segmentation," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11 799–11 809, May 2021.

[85] P. Wang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, 2018.

[86] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[87] J. G.-B. Puyal *et al.*, "Endoscopic Polyp Segmentation Using a Hybrid 2D/3D CNN," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham, Switzerland: Springer, 2020, pp. 295–305.

[88] B. Sushma, C. K. Raghavendra, and J. Prashanth, "CNN based U-net with modified skip connections for colon polyp segmentation," in *Proc. 5th Int. Conf. Comput. Methodologies Commun.*, 2021, pp. 1762–1766.

[89] L. T. T. Hong, N. C. Thanh, and T. Q. Long, "CRF-EfficientUNet: An improved UNet framework for polyp segmentation in colonoscopy images with combined asymmetric loss function and CRF-RNN layer," *IEEE Access*, vol. 9, pp. 156987–157001, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9622208/

[90] S. Geetha, C. Gopakumar, A. Sreenivas, A.M. John, and A.S. Arathy, "Lite-deep: Improved auto encoder-decoder for polyp segmentation," in *Proc. 8th Int. Conf. Smart Comput. Commun.*, 2021, pp. 6–11.

[91] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty augmented context attention for polyp segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Jul. 2021, pp. 2167–2175.

[92] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3059968.

[93] M. Baldeon-Calisto and S. K. Lai-Yuen, "AdaResU-Net: Multiobjective adaptive convolutional neural network for medical image segmentation," *Neurocomputing*, vol. 392, pp. 325–340, 2020.

[94] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "Covid TV-Unet: Segmenting COVID-19 chest CT images using connectivity imposed U-Net," *Comput. Methods Programs Biomed.*, vol. 1, 2021, Art. no. 100007.

[95] Y. Meng *et al.*, "CNN-GCN aggregation enabled boundary regression for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 352–362.

[96] H. Wu, G. Chen, Z. Wen, and J. Qin, "Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3489–3498.

[97] D. Jha *et al.*, "A comprehensive study on colorectal polyp segmentation with ResUNet, conditional random field and test-time augmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2029–2040, Jun. 2021.

[98] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Focus u-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy," *Comput. Biol. Med.*, vol. 137, 2021, Art. no. 104815.

[99] A. Liu, X. Huang, T. Li, and P. Ma, "Co-Net: A collaborative region-contour-driven network for fine-to-finer medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1046–1055.

[100] K. Yang *et al.*, "Automatic polyp detection and segmentation using shuffle efficient channel attention network," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 917–926, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016821003148

[101] V. Thambawita, S. A. Hicks, P. Halvorsen, and M. A. Riegler, "DivergentNets: Medical image segmentation by network ensemble," 2021, *arXiv:2107.00283*.

[102] S. Ali *et al.*, "PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment," 2021, *arXiv:2106.04463*.

[103] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 14–24.

[104] U.S. National Library of Medicine, "Ulcerative colitis," [Online]. Available: https://ghr.nlm.nih.gov/condition/ulcerative-colitis

[105] H. Nosato, H. Sakanashi, E. Takahashi, and M. Murakawa, "An objective evaluation method of ulcerative colitis with optical colonoscopy images based on higher order local auto-correlation features," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2014, pp. 89–92.

[106] N. Otsu and T. Kurita, "A new scheme for practical flexible and intelligent vision systems," in *Proc. IAPR Workshop Comput. Vis.*, 1988, pp. 431–435.

[107] H. Nosato, H. Sakanashi, E. Takahashi, and M. Murakawa, "Method of retrieving multi-scale objects from optical colonoscopy images based on image-recognition techniques," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2015, pp. 1–4.

[108] F. C. Crow, "Summed-area tables for texture mapping," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 1984, pp. 207–212.

[109] A. Dahal, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Detection of ulcerative colitis severity in colonoscopy video frames," in *Proc. 13th Int. Workshop Content-Based Multimedia Indexing*, 2015, pp. 1–6.

[110] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Classification of ulcerative colitis severity in colonoscopy videos using CNN," in *Proc. Int. Conf. Inf. Manage. Eng.*, 2017, pp. 232–237.

[111] S. V. L. L. Tejaswini, B. Mittal, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Enhanced approach for classification of ulcerative colitis severity in colonoscopy videos using CNN," in *Proc. Int. Symp. Vis. Comput.*, 2019, pp. 25–37.

[112] M. F. Mokter, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Classification of ulcerative colitis severity in colonoscopy videos using vascular pattern detection," in *Proc. MICCAI Workshop Mach. Learn. Med. Imag., Conjunction With Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 552–562.

[113] T. Ozawa *et al.*, "Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 416–421, 2019.

[114] Q. Wang *et al.*, "ECA-Net: Efficient channel attention for deep convolutional neural networks," 2020, *arXiv:1910.03151*.

[115] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[116] X. Luo, J. Zhang, Z. Li, and R. Yang, "Diagnosis of ulcerative colitis from endoscopic images based on deep learning," *Biomed. Signal Process. Control*, vol. 73, pp. 1–12, Mar. 2022.

[117] S. Manivannan and E. Trucco, "Learning discriminative local features from image-level labelled data for colonoscopy image classification," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2015, pp. 420–423.

[118] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[119] Z. Wang, Y. Hu, and L.-T. Chia, "Image-to-class distance metric learning for image classification," in *Proc. Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science Series), Berlin, Heidelberg, vol. 6331, K. Daniilidis, P. Maragos, and N. Paragios Eds., 2010, pp. 706–719.

[120] X. Zhen, L. Shao, and F. Zheng, "Discriminative embedding via image-to-class distances," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[121] D. Dey, B. Chatterjee, S. Chakravorti, and S. Munshi, "Cross-wavelet transform as a new paradigm for feature extraction from noisy partial discharge pulses," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 17, no. 1, pp. 157–166, Feb. 2010.

[122] M. Biswas, A. Bhattacharya, and D. Dey, "Classification of various colon diseases in colonoscopy video using cross-wavelet features," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw.*, 2016, pp. 2141–2145.

[123] P. Mesejo *et al.*, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.

[124] R. Nava, G. Cristóbal, and B. Escalante-Ramírez, "Invariant texture analysis through local binary patterns," 2011, *arXiv:1111.7271*.

[125] F. Riaz, F. B. Silva, M. D. Ribeiro, and M. T. Coimbra, "Invariant gabor texture descriptors for classification of gastroenterology images," *IEEE Trans. Biomed. Eng.*, vol. 10, no. 59, pp. 2893–2904, Oct. 2012.

[126] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[127] W. Kuo, L. Wang, and P. Chen, "Preliminary results of computer aided system with the 2nd-generation narrow-band imaging for endoscopic screening of colorectal neoplasms," in *Proc. Int. Conf. Appl. Syst. Innov.*, 2017, pp. 854–857.

[128] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 1363–1366.

[129] H. Shang, Z. Sun, X. Fu, Z. Zhang, and W. Yang, "What and how other datasets can be leveraged for medical imaging classification," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2019, pp. 814–818.

[130] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[131] P. Harzig, M. Einfalt, and R. Lienhart, "Automatic disease detection and report generation for gastrointestinal tract examination," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2573–2577.

[132] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[133] Y. Chang, Z. Huang, W. Chen, and Q. Shen, "Gastrointestinal tract diseases detection with deep attention neural network," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2568–2572.

[134] V. Thambawita *et al.*, "An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification," *ACM Trans. Comput. Healthcare*, vol. 1, no. 3, pp. 1–29, 2020.

[135] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newslett.*, vol. 11, pp. 10–18, 2008.

[136] J. Friedman, R. Tibshirani, and T. Hastie, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, pp. 337–407, 2000.

[137] K. Pogorelov and M. Riegler, "Medico multimedia task at mediaeval 2018," in *CEUR Workshop Proc.*, vol. 2283, pp. 1–4, 2018.

[138] D. Vázquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, 2017, Art. no. 4037190.

[139] J. J. Bernal *et al.*, "Polyp detection benchmark in colonoscopy videos using GTCreator: A novel fully configurable tool for easy and fast annotation of image databases," in *Proc. 32nd CARS Conf.*, Berlin, Germany, 2018.

[140] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two-stream deep feature modelling for automated video endoscopy data analysis," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 742–751.

[141] J. Chen *et al.*, "On-site colonoscopy auto-diagnosis using smart Internet of Medical Things," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2021.3116699.

[142] Accessed: Sep., 2021. [Online]. Available: https://github.com/dusty-nv/jetson-inference

[143] A. Ciobanu, M. Luca, T. Barbu, V. Drug, A. Olteanu, and R. Vulpoi, "Experimental deep learning object detection in real-time colonoscopies," in *Proc. Int. Conf. E-Health Bioeng.*, 2021, pp. 1–4.

[144] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving Jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 69–84.

[145] M. Golhar, T. L. Bobrow, M. P. Khoshknab, S. Jit, S. Ngamruengphong, and N. J. Durr, "Improving colonoscopy lesion classification using semi-supervised deep learning," *IEEE Access*, vol. 9, pp. 631–640, 2021.

[146] Y. Cao, D. Li, W. Tavanapong, J. Oh, J. Wong, and P. C. de Groen, "Parsing and browsing tools for colonoscopy videos," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 844–851.

[147] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 7, pp. 1268–1279, Jul. 2007.

[148] C. Zhang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Cable footprint history: Spatio-temporal technique for instrument detection in gastrointestinal endoscopic procedures," in *Proc. WorldComp Int. Conf. Image Process., Comput. Vis. Pattern Recognit.*, 2015, pp. 308–314.

[149] C. Zhang, W. Tavanapong, J. Wong, P. C. de Groen, and J. Oh, "Real-time instrument scene detection in screening GI endoscopic procedures," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, 2017, pp. 720–725.

[150] P. Wang *et al.*, "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study," *Gut*, vol. 68, no. 10, pp. 1813–1819, 2019.

[151] J.-R. Su *et al.*, "Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: A prospective randomized controlled study (with videos)," *Gastrointestinal Endoscopy*, vol. 91, no. 2, pp. 415–424, 2020.

[152] B. P. Mohan *et al.*, "Real-time computer aided colonoscopy versus standard colonoscopy for improving adenoma detection rate: A meta-analysis of randomized-controlled trials," *EClinicalMedicine*, vol. 29/30, Dec. 2020, Art. no. 100622.

[153] P. Wang *et al.*, "Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomised study," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 4, pp. 343–351, Apr. 2020.

[154] D. Gong *et al.*, "Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): A randomised controlled study," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 4, pp. 352–361, Apr. 2020.

[155] Y. Maeda *et al.*, "Evaluation in real-time use of artificial intelligence during colonoscopy to predict relapse of ulcerative colitis: A prospective study," *Gastrointestinal Endoscopy*, vol. 95, no. 4, pp. 747–756, 2022.

[156] W. Tavanapong, J. Oh, G. Kijkul, J. Pratt, J. Wong, and P. de Groen, "Real-time feedback for colonoscopy in a multicenter clinical trial," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, 2020, pp. 13–18.

[157] A. Repici *et al.*, "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial," *Gastroenterology*, vol. 159, no. 2, pp. 512–520, Aug. 2020.

[158] I. Bello *et al.*, "Revisiting ResNets: Improved training and scaling strategies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, [Online]. Available: https://papers.nips.cc/paper/2021/file/bef4d169d8bddd17d68303877a3ea945-Paper.pdf

[159] C. Zhang, W. Tavanapong, G. Kijkul, J. Wong, P. C. de Groen, and J. Oh, "Similarity-based active learning for image classification under class imbalance," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 1422–1427.

[160] H. A. Qadir, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "A framework with a fully convolutional neural network for semi-automatic colon polyp annotation," *IEEE Access*, vol. 7, pp. 169537–169547, 2019.

[161] V. Thambawita, "DeepSynthBody: The beginning of the end for data deficiency in medicine," in *Proc. Inter. Conf. Appl. Artif. Intell.*, 2021, pp. 1–8.

[162] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4761–4772.

[163] Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, and G. Carneiro, "Few-shot anomaly detection for polyp frames from colonoscopy," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer, 2020, pp. 274–284.

[164] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[165] J. Choi *et al.*, "VaB-AL: Incorporating class imbalance and difficulty with variational bayes for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6749–6758.

[166] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8166–8175.

[167] X. Luo, "SSL 4MIS," 2020. [Online]. Available: https://github.com/HiLab-git/SSL4MIS

[168] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8801–8809.

[169] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1705–1714.

[170] Y. Liu *et al.*, "Photoshopping colonoscopy video frames," in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 1–5.

[171] A. Ravichandran, R. Bhotika, and S. Soatto, "Few-shot learning with embedded class models and shot-free meta training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 331–339.

[172] T. Agrawal, R. Gupta, S. Sahu, and C. E. Wilson, "SCL-UMD at the medico Task-MediaEval 2017: Transfer learning based classification of medical images," in *Proc. SCL-UMD Medico Task-MediaEval: Transfer Learn. Based Classification Med. Images*, 2017.

[173] R. Zoetmulder, E. Gavves, M. Caan, and H. Marquering, "Domain- and task-specific transfer learning for medical segmentation tasks," *Comput. Methods Programs Biomed.*, vol. 214, 2022, Art. no. 106539. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721006131

[174] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Med. Image Anal.*, vol. 60, 2020, Art. no. 101619.

[175] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Proc. Int. Conf. Artificial Neural Netw.*, Cham, Switzerland: Springer, 2016, pp. 63–71.

[176] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Workshop Int. Conf. Learn. Representations*, 2014.

[177] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam : Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[178] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 3145–3153.

[179] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, 2017.

[180] M. Sundararajan, A. Taly, and Q. Yan, "Gradients of counterfactuals," 2016, *arXiv:1611.02639*.

[181] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.

[182] S. Serrano and N. A. Smith, "Is attention interpretable?," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2931–2951.

[183] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, 2019, pp. 11–20.

[184] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[185] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 7786–7795.

[186] B. Hosseini and B. Hammer, "Interpretable multiple-kernel prototype learning for discriminative representation and feature selection," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1863–1872.

[187] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," *Proc. Int. Conf. NeurIPS*, 2018, pp. 8–14.

[188] M. Khaleel, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Hierarchical visual concept interpretation for medical image classification," in *Proc. IEEE 34th Int. Symp. Comput. Based Med. Syst.*, 2021, pp. 25–30.

[189] A. Dhurandhar *et al.*, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. the 32nd Int. Conf. NeurIPS*, 2018, pp. 590–601.

[190] A. Kanehira and T. Harada, "Learning to explain with complemental examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8603–8611.

[191] D. Wang, X. Cui, and Z. J. Wang, "Chain: Concept-harmonized hierarchical inference interpretation of deep convolutional neural networks," 2020, *arXiv:2002.01660*.

[192] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6541–6549.

[193] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, 2018, pp. 80–89.

[194] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 818–833.

[195] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[196] R. W. Stidham *et al.*, "Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis," *JAMA Netw. Open*, vol. 2, no. 5, May 2019, Art. no. e193963.

[197] G. Ciuti *et al.*, "Frontiers of robotic colonoscopy: A comprehensive review of robotic colonoscopes and technologies.," *J. Clin. Med.*, vol. 9, no. 6, May 2020, Art. no. 1648.

[198] M. F. Hale *et al.*, "Magnetically steerable gastric capsule endoscopy is equivalent to flexible endoscopy in the detection of markers in an excised porcine stomach model: Results of a randomized trial," *Endoscopy*, vol. 47, no. 7, pp. 650–653, Jul. 2015.