Artificial Intelligence In Medicine 108 (2020) 101923

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Deep learning to find colorectal polyps in colonoscopy: A systematic literature review

Luisa F. Sánchez-Peralta[a,*], Luis Bote-Curiel[a], Artzai Picón[b], Francisco M. Sánchez-Margallo[a], J. Blas Pagador[a]

[a] Jesús Usón Minimally Invasive Surgery Centre, Ctra. N-521, km 41.8, 10071 Cáceres, Spain
[b] Tecnalia, Parque Científico y Tecnológico de Bizkaia, C/ Astondo bidea, Edificio 700, 48160 Derio, Spain

ARTICLE INFO

Keywords:
Colorectal cancer
Deep learning
Detection
Localization
Segmentation

ABSTRACT

Colorectal cancer has a great incidence rate worldwide, but its early detection significantly increases the survival rate. Colonoscopy is the gold standard procedure for diagnosis and removal of colorectal lesions with potential to evolve into cancer and computer-aided detection systems can help gastroenterologists to increase the adenoma detection rate, one of the main indicators for colonoscopy quality and predictor for colorectal cancer prevention. The recent success of deep learning approaches in computer vision has also reached this field and has boosted the number of proposed methods for polyp detection, localization and segmentation. Through a systematic search, 35 works have been retrieved. The current systematic review provides an analysis of these methods, stating advantages and disadvantages for the different categories used; comments seven publicly available datasets of colonoscopy images; analyses the metrics used for reporting and identifies future challenges and recommendations. Convolutional neural networks are the most used architecture together with an important presence of data augmentation strategies, mainly based on image transformations and the use of patches. End-to-end methods are preferred over hybrid methods, with a rising tendency. As for detection and localization tasks, the most used metric for reporting is the recall, while Intersection over Union is highly used in segmentation. One of the major concerns is the difficulty for a fair comparison and reproducibility of methods. Even despite the organization of challenges, there is still a need for a common validation framework based on a large, annotated and publicly available database, which also includes the most convenient metrics to report results. Finally, it is also important to highlight that efforts should be focused in the future on proving the clinical value of the deep learning based methods, by increasing the adenoma detection rate.

1. Introduction

Colorectal cancer (CRC) is defined as a carcinoma, usually an adenocarcinoma, in the colon or rectum. Colorectal cancer is considered primarily as a "lifestyle" disease; its incidence is higher in countries with a diet high in calories and animal fat and with a largely sedentary population [1]. CRC accounts for a 10% of overall new cancer cases worldwide, with a higher incidence rate in developed countries [2]. Only in the United States, it has increased from over 132,000 estimated new cases and nearly 50,000 estimated deaths in 2015 [3] to over 145,000 estimated new cases and 51,000 estimated deaths in 2019 [4]. In Europe, CRC represents the second most common cancer and also the second cause of death from cancer [5].

Nevertheless, an early detection of the CRC increases the 5-year survival rate from 18% when CRC is detected in the highest grade to 88.5% when it is detected in an initial grade due to symptoms. Furthermore, screening programs achieve a significant increase of the survival rate as they allow to start treatment even before the appearance of those symptoms, so up to 222 deaths out of 1000 patients detected with symptomatic CRC could be avoided [6].

Abbreviations: ADR, adenoma detection rate; CAD, computer aided detection; CDDN, cascaded deep decision network; CI, confidence interval; CNN, convolutional neural network; CRC, colorectal cancer; DWD, distance-weighted discrimination; FCN, fully convolutional network; GAN, generative adversarial network; IoU, Intersection over Union; LSTM, long short term memory; mAP, mean average precision; PIVI, preservation and incorporation of valuable endoscopic innovations; RNN, recurrent neural network; ReLU, rectified linear unit; RPN, region proposal network; SVM, support vector machine

* Corresponding author.
E-mail addresses: lfsanchez@ccmijesususon.com (L.F. Sánchez-Peralta), lbote@ccmijesususon.com (L. Bote-Curiel), artzai.picon@tecnalia.com (A. Picón), msanchez@ccmijesususon.com (F.M. Sánchez-Margallo), jbpagador@ccmijesususon.com (J.B. Pagador).

https://doi.org/10.1016/j.artmed.2020.101923
Received 27 August 2019; Received in revised form 3 March 2020; Accepted 1 July 2020

Colonoscopy is a standard technique for visual exploration of the colon and rectum by inserting a flexible endoscope through the patient anus [7] and is considered the gold standard for detection and removal of colorectal lesions, associated with important reduction of CRC mortality [8]. The European Society of Gastrointestinal Endoscopy recommends the use of high definition white light endoscopes for detection of colorectal neoplasms in middle risk population [9], as it is estimated that 70–80% of CRC has a sporadic origin [10]. Therefore, it is clear the need for increasing the adenoma detection rate (ADR), defined as the proportion of patients with at least one colorectal adenoma detected among all patients examined by the gastroenterologist, which is both a colonoscopy quality measure and a validated predictor for CRC prevention, having an inverse relationship [11].

Recently, the success of deep learning [12] has also boosted the applications on medical imaging analysis [13], achieving expert performance in several cases [14–16]. Deep learning approaches rely on the ability of networks with several layers to automatically learn hierarchical features characterizing the input data through the application of non-linear operations together with backpropagation for training. Deep learning architectures stack blocks of different types of layers (fully connected, convolutional, pooling or activation layers) to simultaneously be sensitive to minute details and insensitive to large irrelevant details.

Computer-aided detection (CAD) systems have the potential to revolutionize the endoscopic practice by (1) improving the adequacy of inspection technique; (2) providing automatic detection of precursor lesions of CRC; and (3) facilitating real-time diagnosis with optical biopsy [17]. In this review, we focus on the second topic. Traditionally, CAD systems for polyp detection have been based on the manual extraction of polyp features, or so called hand-crafted methods: shape-based [18], texture-based [19,20], depth of valleys-based [21,22] or combined-based [23] methods. Nevertheless, results of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge proved that convolutional neural networks (CNNs) are the state-of-the-art regarding polyp detection methods [24].

Within the scope of this systematic review, three main tasks are considered:

1. Detection: identifying whether a polyp is shown or not in the frame, but information on the polyp location is not given.
2. Localization: identifying the position of the polyp within a given frame, but exact shape of the polyp is not relevant.
3. Segmentation: marking the exact polyp area in a given frame.

Polyp classification (benign vs malign, Paris classification, NICE classification, etc.) is out of the scope of the current systematic review. Besides, classification in this case is done once the presence of the polyp is confirmed and in this review we place the focus on the prior stage (identifying the presence of the polyp in the frame). Therefore, we analyze the published methods for detection, localization and segmentation of colorectal polyps based on deep learning approaches. In this sense, methods are classified according to their main aim, the used database, their approach and the reported metrics. The state-of-art approaches are compared, showing advantages and disadvantages of the different categories, to identify the most auspicious trends.

There are several surveys on deep learning for medical imaging analysis [13,25–31], where at most methods for colorectal polyps identification are roughly analyzed. A more specific review was done by Prasath [32] but the focus was placed on video capsule endoscopy rather than on colonoscopy images. More recently, Ahmad et al. [33] summarized the evidence for clinical applications of computer-aided diagnosis and artificial intelligence in colonoscopy of key studies in a narrative manner. Therefore, to the authors' knowledge, there is no previous systematic review on the proposed topic with a comparative analysis of retrieved works.

Nevertheless, the lack of a common framework makes difficult to provide a reliable comparison of the state-of-art methodologies. The main limitation comes from the different datasets and/or testing set used for reporting results as well as selected metrics. Ideally, all methods should be applied on a common database [34]. Regretfully, this ideal situation is not real in many of the published works, where authors use different datasets, in a different manner, and reporting on different test sets with different metrics. Therefore, in the current review, methods are compared accordingly to their reported results, identifying characteristics that might influence on them. To overcome this heterogeneous situation, efforts have been lately performed by the organization of challenges under the hosting of international congresses in order to unify criteria. It is of special relevance the Endoscopic Vision Challenge [35], which has organized a subchallenge focused on detection, localization and segmentation of polyps in 2015 [24], 2017 and 2018. In these cases, methods are easily and reliably compared. Besides, Vázquez et al. [36] also propose a benchmark for endoluminal scene segmentation of colonoscopy images, with the aim of boostering comparative research.

The main contributions of this systematic review are:

1. We analyze publicly available datasets of colonoscopy images.
2. We provide a comprehensive analysis of polyp detection, localization and segmentation methods based on deep learning, discussing advantages and disadvantages of the different categories.
3. We analyze and discuss the reporting metrics used.
4. We identify future challenges and recommendations based on the findings of the review to be addressed by the scientific community to advance the field.

This review is organized as follows. In Section 2 we present the material and methods to carry out the systematic literature review, including search strategy, study selection and data extraction and management. In Section 3 we show the results of the search and summarize the works found. Then, datasets of colonoscopy images and how they are used are described in Section 4. In Section 5 and its corresponding subsections, we present and discuss the methods for detection (Section 5.1), localization (Section 5.2) and segmentation (Section 5.3) of colorectal polyps using deep learning approaches. We conclude this section with the advantages and disadvantages of each category (Section 5.4). After this, metrics are analyzed in Section 6, to conclude the paper with an analysis from a clinical perspective (Section 7), future challenges and recommendations (Section 8) and final conclusions (Section 9).

2. Material and methods

2.1. Search strategy

The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [37] has been followed to perform this systematic review. The basic search string was ("colon" OR "colorectal") AND ("cancer" OR "polyp") AND ("deep learning") AND ("detection" OR "localization" OR "segmentation") and searches were performed on February 2nd 2019 using ACM Digital Library, IEEE Digital Library, Web of Science, PubMed, Science@Direct, Scopus and Springer Link databases. The search string syntax was adapted when necessary, depending on the database requirements. Search was performed on title, abstract and keywords. Previously identified articles were also included in the process.

2.2. Study selection

Studies included in the analysis were full text articles or full-length proceedings published in English. The exclusion criteria were papers published before 2015; in other languages; about a different topic; applying deep learning in a field other than CRC; pure clinical studies;

use of endoscopic capsules as imaging source; not using white light imaging; not using deep learning techniques; short proceedings; and meta-analyses or reviews. Retrieved abstracts were read by two authors (LBC and LFSP), searching for the full text when information in the abstract was not enough to determine its inclusion or exclusion. Full texts of selected abstracts were retrieved (LFSP) and independently revised (AP, LBC, JBP and LFSP) for final agreement on inclusion criteria. The reference list of selected works was also scrutinized for potential interesting articles (LFSP).

2.3. Data extraction

Information from papers selected for analysis was extracted for comparison (JBP, AP and LFSP). Papers were initially categorized according to their main objective (detection, localization or segmentation). For each work, we also identified findings related to the used dataset, the data augmentation approach, the proposed method, the reporting metrics and reproducibility aspects (Table 1).

2.4. Review and data management

Parsif.al[1] was used for abstract review and data management. Search results from each database were imported into the platform. Authors, title, year and journal were automatically extracted, facilitating the selection procedure and the creation of the PRISMA flow diagram. A custom data extraction form was developed in the platform and used for the data extraction process.

3. Results of the systematic search

In all, 1,332 abstracts were found (Table 2 and Fig. 1). 35 papers were previously identified and used to select the search string based on their keywords, while 1,297 were retrieved from searches in the different databases. After removing duplicates, 1,123 abstracts were screened. 1,071 were excluded based on the exclusion criteria and 52 were revised in full text. From those, 33 were included in this review analysis. From manual inspection of references lists, 2 more works were added.

During the revision process, detection was also considered as classification between healthy tissue and polyp classes or polyp and non-polyp classes, but we excluded papers aiming at multi-class classification (polyp among other classes) (such as Pogorelov et al. [38] or Park and Sargent [39]) or polyp classification (neoplastic/non-neoplastic) (such as Ribeiro et al. [40]).

Full text documents were carefully read, and data included in Table 1 were extracted for each work when available in the document. Authors were contacted when deemed necessary for clarification. Table 3 summarizes the 35 articles included in the analysis, together with some relevant aspects. Some works address more than one task, so they appear once per task in the table, resulting in 39 cases. There are 5 book sections (usually proceedings of major conferences), 18 conference proceedings, 11 journal articles and one preprint. To unify criteria for results reporting in Table 1, metrics reported as percentages by authors have been expressed in the normalized range [0, 1].

As it might be expected, there is an overall increasing trend since 2016 in the application of deep learning techniques for polyp detection, localization and segmentation (Fig. 2).

Fig. 3 summarizes the networks used by the different authors. Architectures have been grouped into 4 clusters: (1) CNNs, such as AlexNet [75], VGG16 and VGG19 [76] or GoogLeNet [77], including also residual networks such as ResNet50 [78]; (2) fully convolutional networks (FCNs), based on any CNN architecture, including also the encoder-decoder architectures, such as SegNet [79] or U-Net [80]; (3) generative adversarial networks (GANs), and (4) recurrent neural networks (RNNs), including long short term memory (LSTM). It can be clearly seen that the use of CNNs surpasses the rest of networks in the three analyzed tasks.

A baseline comparison is needed to assess whether the proposed method actually improves the results of the state-of-art. Out of the 35 analyzed works, only 10 of them present a baseline against which the proposed method is compared. The baseline is usually the network on which the proposed method is based, or hand-crafted methods. In other 12 cases, authors provide a comparison of their method against other similar works. The most repeated comparison is against the methods participating in the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge [24]. The remaining 13 works only present their work, without any type of comparison.

As for reproducibility, we considered two aspects: the use of public datasets and the availability of the code. Most works (29) use only public datasets, while 3 use only private datasets and other 3 use both public and private ones. The use of proprietary datasets hampers the reproducibility and fair comparison of methods. The code of only three works have been found [36,70,72]. Therefore, Vázquez et al. [36] and Wickstrøm et al. [70] outstand in terms of reproducibility, as the code is available, and they use CVC-EndoSceneStill, which, as it will be explained in the following section, provides a division into training, validation and test sets.

4. Datasets of colonoscopy images

4.1. Currently available public datasets

The creation of large, annotated datasets has also contributed to the tremendous growth of deep learning for the last years. Although they are easily accessible for natural images with different ground truths (i.e. ImageNet [81,82], MSCoco [83,84] or Pascal VOC [85,86]), the limited size of medical imaging datasets is a well-known problem, especially for supervised learning [26]. This situation also applies to colonoscopy images datasets. The type of ground truth highly influences the size of the dataset, since manual annotation of frames is a cumbersome, time-consuming task [87]. Table 4 shows a summary of the currently publicly available datasets of colonoscopy images for polyp detection, localization and segmentation, although prior registration might be required by the dataset owner to grant access to the content. All datasets are mentioned in at least one of the analyzed papers. Although these datasets are widely used in the retrieved papers in this systematic review, some authors also use proprietary datasetss, which compromises a fair comparison of methods and raise concerns about reproducibility, as mentioned in the previous section. Datasets of natural images have been also included in Table 4 for scale comparison between computer vision and biomedical image datasets. Works retrieved in this review only employ these larger datasets to initialize the weights of the network before training.

The organization of challenges under the umbrella of major conferences has meant a great step towards the establishment of a common framework. As a result, most of the currently used datasets were provided within a challenge. The Gastrointestinal Image ANAlysis (GIANA) sub-challenge [91], as part of the EndoVis challenge [35], was last hosted at MICCAI2017 and MICCAI2018. CVC-VideoClinicDB [87,89] was provided as training dataset for the polyp detection task, while CVC-ColonDB, CVC-ClinicDB and CVC-ClinicHDSegment datasets [21,22,36] were provided for the polyp segmentation task. During MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge [24], two more datasets were released. On one side, ETIS-LARIB was provided as test set for detection on still frames. On the other hand, the ASU-Mayo Clinic [88] was intended for exploring detection on videos. In all cases, the provided ground truth is a binary mask indicating the polyp area. This segmentation is a precise manual delineation except for the CVC-VideoClinicDB dataset, where the polyp area has been

[1] https://parsif.al/

Table 1
Data extraction.

Category	Item	Description
General information	Type of publication	Journal article, conference proceeding, book section (usually proceedings of a major conference), preprint
	Published in	Title of the journal, conference or preprint repository
	Country	Country of the first author affiliation
Objective	Task	Aim of the work: detection, localization and/or segmentation
Data information	Dataset	Name of the public dataset and reference (if available) or proprietary
	Training set	Description of the samples used for training, indicating the number (with and without data augmentation, if applied)
	Validation set	Description of the samples used for validation, indicating the number
	Test set	Description of the samples used for testing, indicating the number
Data augmentation	Approach	Data augmentation approach to generate new training samples: creation of images, on the flow, patch-based or none
	Transformations	If data augmentation is applied, description of the transformations and ranges used
Model	Approach	Type of approach used: feature extractor, classification, patch-based, bounding-box or semantic segmentation
	Architecture	Type of the network (AlexNet, VGG, fully convolutional network, GAN, etc.)
	Loss function	Loss function used for optimization
	Training	Use of pre-trained models or training from scratch
Reporting	Reporting test	Samples used for reporting
	Metrics	Metrics used for reporting (accuracy, Dice, Intersection over Union, etc.)
	Baseline	Whether the authors establish an initial baseline model or compare their results to other works
Reproducibility	Dataset	Whether the dataset is proprietary or publicly available
	Code	Whether the code is publicly available or not

Table 2
Number of abstracts retrieved.

Database	# abstracts
ACM Digital Library	2
IEEE Digital Library	17
Web of Science	43
PubMed	14
Science@Direct	208
Scopus	772
Springer	241
Previously identified	35

approximated to the most convenient elliptical shape. CVC-EndoSceneStill [36] has not been used in any challenge, but it is publicly available. It compiles CVC-ColonDB [22] and CVC-ClinicDB [21] adding ground truth masks for other classes, establishing the distribution of images into training, validation and test sets and indicating the metrics for reporting. Lastly, and not used as much as the previously reported datasets, Kvasir [90] is a multi-class image dataset containing the polyp class among other labels for anatomical landmarks and pathological findings.

There are as well other public datasets used by authors in the current review. The Nerthus database [92], used by Pogorelov et al. [54],

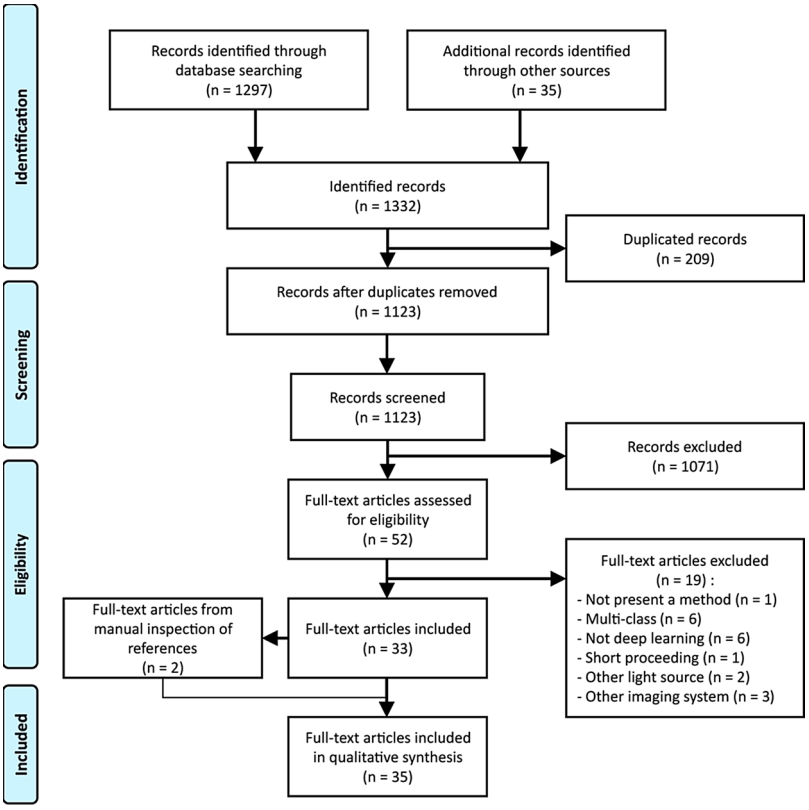

Fig. 1. Literature flow diagram.

Table 3
Summary of reviewed works.

Objective	Type	Approach	Authors	Datasets	Data augmentation	Network	Main results	Baseline	Code
Detection	Hybrid	Feature extractor, patch-based	Taha et al. [41]	CVC-ColonDB	Patches	AlexNet	Prec = 0.927; Rec = 0.960	Comparison	NF
		Feature extractor, bounding-box	Shin et al. [42]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	Patches	CNN	Acc = 0.9126; Prec = 0.9271; Rec = 0.9082; Spec = 0.9176	Yes	NF
		Patch-based	Tajbakhsh et al. [43]	ASU-Mayo	Patches	AlexNet	Sens = 0.7 @ FPPF = 0.02	Yes	NF
			Yuan et al. [44]	ASU-Mayo	Patches	AlexNet	Acc = 0.9147; Rec = 0.9176	No	NF
			Tajbakhsh et al. [45]	ASU-Mayo	Patches	AlexNet	Sens = 0.7 @ FPPF = 0.02	Yes	NF
			Tajbakhsh et al. [46]	ASU-Mayo	Patches	CNN	Sens = 0.5 @ FPPF = 0.002	Yes	NF
			Tajbakhsh et al. [47]	ASU-Mayo	Patches	CNN	Sens = 0.5 @ FPPF = 0.002	No	NF
		Classification, patch-based	Axyonov et al. [48]	CVC-ColonDB, ASU-Mayo	N/A	AlexNet	AUC = 0.92; Sens = 0.75 @ FPPF = 0.1	Comparison	NF
	End-to-end	Classification	Akbari et al. [49]	ASU-Mayo	N/A	CNN	Acc = 0.9028; Rec = 0.6832; Prec = 0.7434; Spec = 0.9497; FPPF = 0.06	Comparison	NF
			Aksenov et al. [50]	CVC-ColonDB, ASU-Mayo	Creation	CNN	TPR ≈ 0.98 @FPR = 0.025	No	NF
			Itoh et al. [51]	Proprietary	N/A	C3dNet	AUC = 0.83; Acc = 0.747; Sens = 0.881; Spec = 0.617	No	NF
			Misawa et al. [52]	Proprietary	N/A	C3dNet	AUC = 0.87; Spec = 0.63 @ Sens = 0.90	No	NF
			Murthy et al. [53]	ISBI2014 challenge	Creation	CDDN	Acc = 0.8743	Yes	NF
			Pogorelov et al. [54]	CVC-ColonDB, CVC-ClinicDB, CVC-VideoClinicDB, Kvasir, Nerthus	Creation	Xception, VGG19, ResNet50, GAN	Acc = 0.909; Spec = 0.94	No	NF
		Bounding-box	Mo et al. [55]	CVC-ClinicDB, CVC-VideoClinicDB, CVC-ColonDB, CVC-EndoSceneStill	N/A	Faster R-CNN (VGG16)	Acc = 0.985; Prec = 1; Rec = 0.985; F1-score = 0.971; F2-score = 0.992	Comparison	NF
			Urban et al. [56]	Proprietary	On the flow	VGG16, VGG19, ResNet50	Acc = 0.964; ROC-AUC = 0.991	No	NF
		Semantic segmentation	Mohammed et al. [57]	ASU-Mayo	Creation, On the flow	Y-Net	Prec = 0.874; Rec = 0.844; F1-score = 0.859; F2-score = 0.850	Yes	NF
			Brandao et al. [58]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	On the flow	FCN-AlexNet, FCN-GoogLeNet, FCN-VGG	Prec = 0.7361; Rec = 0.8631	Comparison	NF
Localization	Hybrid	Patch-based	Park et al. [59]	CVC-ClinicDB, ASU-Mayo	Patches	CNN	Prec = 0.6575; Rec = 0.8276	No	NF
		Bounding-box	Shin et al. [60]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo, CVC-VideoClinicDB	Creation	Faster R-CNN (Inception ResNet-v2)	Prec = 0.914; Rec = 0.803; F1-score = 0.833; F2-score = 0.815	Comparison	NF
			Yu et al. [61]	ASU-Mayo	Creation	3D-FCN	Prec = 0.881; Rec = 0.710	No	NF
			Zhang et al. [62]	ASU-Mayo, CVC-ClinicDB, ETIS-LARIB	On the flow	RYCO	Prec = 0.886; Rec = 0.716; Spec = 0.970	Yes	NF
		Bounding-box, semantic segmentation	Wang et al. [63]	CVC-ClinicDB, proprietary	N/A	SegNet	Sens = 0.9438; Spec = 0.9592	No	NF
		Feature extractor, patch-based	Billah et al. [64]	Mesejo, ASU-Mayo, CVC-ClinicDB, ETIS-LARIB	N/A	CNN	Acc = 0.9865; Sens = 0.9879; Spec = 0.9852	Comparison	NF
	End-to-end	Bounding-box	Mo et al. [55]	CVC-ClinicDB, CVC-VideoClinicDB, CVC-ColonDB, CVC-EndoSceneStill	N/A	Faster R-CNN (VGG16)	Prec = 0.862; Rec = 0.981; F1-score = 0.917; F2-score = 0.956	Comparison	NF
			Zheng et al. [65]	CVC-ColonDB, CVC-ClinicDB, ETIS-LARIB, proprietary	Creation	YOLO	Prec = 0.774; Sens = 0.740; F1-Score = 0.757; F2-score:0.747	No	NF
			Pogorelov et al. [66]	ASU-Mayo, proprietary	N/A	TensorBox, Darknet-YOLO	Acc TensorBox = 0.316; Acc Darknet-YOLO = 0.422	No	NF
		Classification	Urban et al. [56]	Proprietary	On the flow	VGG16, VGG19, ResNet50	Dice = 0.827 ± 0.003	No	NF
		Patch-based, classification, semantic segmentation	Pogorelov et al. [54]	CVC-ColonDB, CVC-ClinicDB, CVC-VideoClinicDB, Kvasir, Nerthus	Creation	Xception, VGG19, ResNet50, GAN	Acc = 0.946; Spec = 0.984	No	NF

(*continued on next page*)

Table 3 (*continued*)

Objective	Type	Approach	Authors	Datasets	Data augmentation	Network	Main results	Baseline	Code
Segmentation	Hybrid	Patch-based	Zhang et al. [67]	CVC-ColonDB	N/A	FCN-8s	Acc = 0.9754; Rec = 0.7566; Spec = 0.9881; Dice = 0.7014	Comparison	NF
	End-to-end	Semantic segmentation	Nguyen and Lee [68]	CVC-ClinicDB, ETIS-LARIB	Pixel level	Encoder-decoder	Dice = 0.889; IoU = 0.8935; Acc = 0.984	Comparison	NF
			Wichakam et al. [69]	CVC-EndoSceneStill	On the flow	Compressed FCN-8s	Prec = 0.8848; Rec = 0.7814; IoU = 0.6936; Dice = 0.9594	Yes	NF
			Wickstrøm et al. [70]	CVC-EndoSceneStill	On the flow	Enhanced FCN-8, enhanced SegNet	IoU = 0.767; Acc = 0.949	Comparison	Yes[a]
			Xiao et al. [71]	CVC-ClinicDB	N/A	DeepLab-v3 + LSTM	IoU = 0.9321	Comparison	NF
			Zhou et al. [72]	ASU-Mayo	N/A	U-Net + +	IoU = 0.3345	Yes	Yes[b]
			Bardhi et al. [73]	CVC-ColonDB, CVC-ClinicDB, ETIS-LARIB	On the flow	SegNet	Acc = 0.967	No	NF
			Brandao et al. [58]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	On the flow	FCN-AlexNet, FCN-GoogLeNet, FCN-VGG	Prec = 0.7023; Rec = 0.5420	Comparison	NF
			Li et al. [74]	CVC-ClinicDB	Creation	FCN	Acc = 0.9698; Rec = 0.7732; Spec = 0.9905	No	NF
			Vázquez et al. [36]	CVC-EndoSceneStill	On the flow	FCN-8	IoU Polyp = 0.5160; MGA = 0.9677	Yes	Yes[c]

N/A: Not applicable; CNN: convolutional neural network; FCN: fully convolutional network; Prec: precision; Rec: recall; Sens: sensitivity; IoU: Intersection over Union; MGA: mean global accuracy; MCA: mean classes accuracy; Spec: specificity; FPPF: false positive per frame; TPR: true positive rate; FPR: false positive rate; NF: not found.

[a] https://github.com/Wickstrom/Thesis.
[b] https://github.com/MrGiovanni/UNetPlusPlus.
[c] https://github.com/jbernoz/deeppolyp.

provides a classified set of videos depending on the Boston bowel preparation scale, therefore not providing any polyp information and being excluded from Table 4. Mesejo et al. [93] also provide a labelled dataset of 76 videos of different lesions (serrated adenomas, hyperplastic lesions and adenomas), which is used by Billah et al. [64]. Since optical biopsy for polyp classification is out of the scope of the current review, the dataset has not been included in Table 4.

Regarding clinical variability, few datasets indicate the type of polyps included. The Paris classification [94,95] is a general framework for the endoscopic classification of superficial lesions of the oesophagus, stomach, and colon. Fig. 4 shows the different types of polyps, both in schematic view and actual endoscopic images. Pedunculated and sessile polyps are easier to detect than flat polyps and CAD systems to assist their detection would be more useful for gastroenterologists, but regretfully they are underrepresented in the public datasets [96].

4.2. Use of the datasets and data augmentation

Authors do not follow a standard methodology to distribute the dataset into training, validation and test sets, except for those using CVC-EndoSceneStill, because the distribution is provided by the dataset owner; or those following the rules of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge. Works for detection use a greater number of images than those for segmentation, which might be because labelling frames is easier than manually segmenting polyps for ground truth creation (Fig. 5). Test sets are usually one or two degrees of magnitude smaller than the training sets. Due to this heterogeneity in training, validation and test sets, it is no easy to make a fair comparison of the methods and their reported metrics.

Data augmentation is the process whereby the training dataset is artificially increased in size, which in medical imaging is typically done with transformations that are applied to only the image in the case of detection (as each image only have a label that remains unaffected) or to the image and mask in the same way in the case of localization and segmentation. Augmentation methods commonly employ transformations such as rotations, reflections, and elastic deformations [97]. Data augmentation strategies are used by 24 out of the 35 analyzed papers (Table 3). The most common approach (8 works) is to enlarge the training set by creating new images through the application of transformations to the original images. Other authors (7 works) make data augmentation on the flow, i.e. transformations to the original image are randomly applied at training time, increasing the variability of the training set but without specifically creating new transformed images. Lastly, some other authors (8 works) train the models using patches extracted from the original images rather than using the full image. More recently, Nguyen and Lee [68] proposed a data augmentation approach at a pixel level for polyp segmentation.

To create new images, there is a wide variability on the transformations applied and their ranges (Table 5), as data augmentation is typically performed by trial and error and transformations are selected based on the imagination, time and experience of the researcher [98]. None of the authors justify the selection of neither the transformations nor the ranges. Few authors analyze the influence of data augmentation in the results. Vázquez et al. [36] compare results with and without different transformations, while Shin et al. [60] compare the influence of including more or fewer transformations. While the former identified that the combination of transformations leads to better results, the latter found that more augmentation does not guarantee better performance.

The different methods for data augmentation lead to a wide variability in terms of the actual training samples used (Fig. 5).

5. Comparison and discussion of methods

CNN [12,99,100] architectures are a type of neural networks which are specialized for data with grid-structured topology. CNNs are

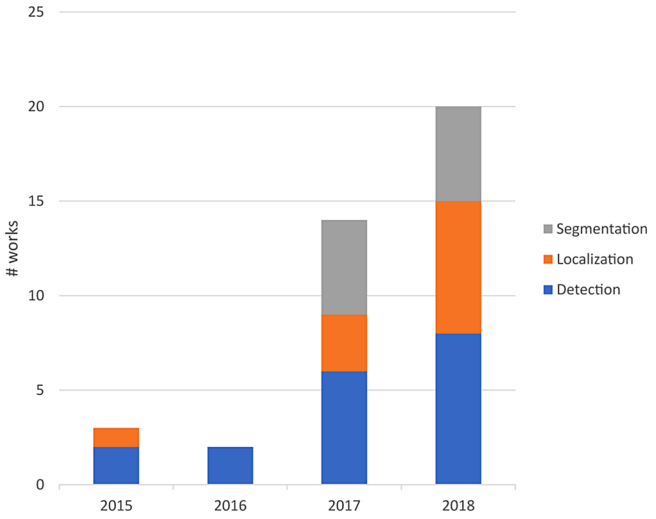

Fig. 2. Literature trends.

composed of different hierarchical stages that take advantage of local connections, shared weights, pooling layers and the use of many layers. The first set of stages may include a certain number of convolutional layers (namely two or three) followed by subsequent pooling layers. Convolutional layers exploit the local connections and shared weights by using the convolution operation instead of matrix multiplication and pooling layers subsample the data and merge similar features. Blocks of convolutional and pooling layers are then stacked to create a feature vector that represents the input data. Fully connected layers are latterly connected to this vector for the final object classification. One of the simpler and classical implementations of this stackable approach is VGG [76], where different stackable layers of 3×3 convolutions and maximum pooling layers are concatenated. However, this network presents a large number of parameters and slow convergence. He et al. [78] proposed the so called residual neural networks, which include skip connections so learnt filters are applied not to the final transformed image but to the residual over the input image instead, allowing deep neural networks to go deeper by mitigating the vanishing gradient problem and providing the first layers with larger scale gradients during backpropagation. Any CNN architecture can be extended into a FCN [101] by using a classification network as an encoder that is convolved over larger images producing spatially dense prediction tasks. However, these FCNs lacked the capability of generating fine shape delineation as high resolution reconstruction was calculated by interpolation. These capabilities were introduced by the addition of deconvolution layers, skip connections [80] and pooling indices [79]. This segmentation architecture was improved by using the fully convolutional DenseNet for image segmentation [102].

Based on the aforementioned works, different efforts have been followed during the last years for polyp detection, localization and segmentation, where deep learning-based approaches have proven to excel hand-crafted methods. An initial division of methods has been established by Bernal et al. [24]. Two different types of methods are of interest in this review: (1) hybrid methods that combine deep learning approaches with other hand-crafted methods and (2) end-to-end methods that use one single deep learning approach to obtain the result. A third type also mentioned by Bernal et al. [24], hand-crafted methods, lie out of the scope of this review. As secondary classification, we have grouped the approaches into five types, depending on their use of the deep learning network (Fig. 6):

1. Feature extractor. Deep learning architectures are used for automatically creating a feature vector, instead of the manual extraction of features. The computed vector is afterwards the input to a classical classifier, such as support vector machine (SVM) or distance-weighted discrimination (DWD) classifier [103], usually more robust than SVM. These methods are therefore always hybrid, as deep learning is combined with classical classifiers.
2. Classification. A classification network is used to label an image as containing a polyp or not, without position information of the polyp.
3. Patch-based. The method uses image patches or tiles and the presence of polyp is obtained for each patch. Location of the polyp might be obtained based on the patch location.
4. Bounding-box. The method provides the location of the polyp through a bounding-box (coordinates of the upper right corner, height and width), generally using a regression layer.
5. Semantic segmentation. Each pixel of the image is labelled as polyp or background. Networks based on encoder-decoder blocks are usually selected. The first half of the layers encode the image description highlighting the discriminative features for the entrusted task, while the second half is responsible for mapping the low-resolution encoding into full input resolution feature maps. These FCNs can be initialized from the weights of a classification network (encoder) that acts as a feature extractor.

All papers have been categorized using these primary and secondary classifications. While the primary classification is mutually exclusive, some works falls in more than one group of the secondary classification. Fig. 7 shows the distribution of papers. Since tasks and primary

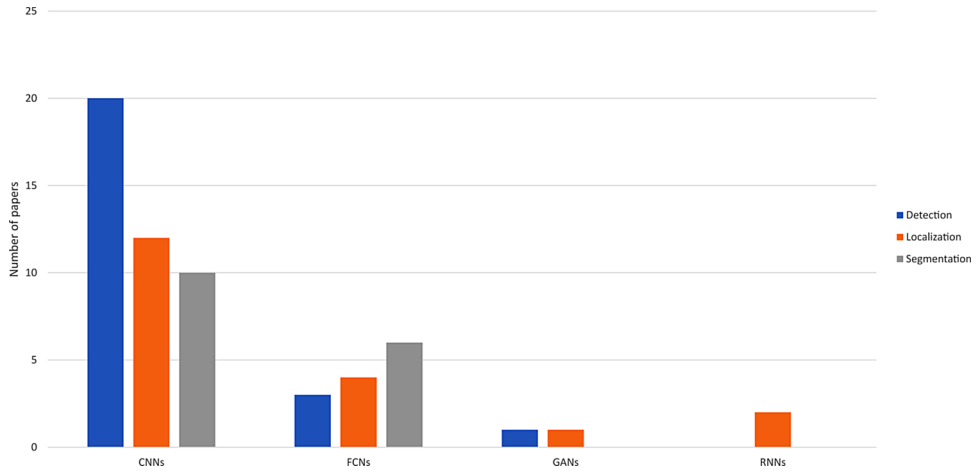

Fig. 3. Networks for each of the tasks. CNNs: Convolutional neural networks; FCNs: fully convolutional networks; GANs: generative adversarial networks; RNNs: recurrent neural networks.

Table 4
White-light datasets for polyp detection, localization and segmentation. Main natural image datasets are also included for comparison.

Dataset	Content	Ground truth	Delineation	Resolution	Type of polyps[a]	# patients	# items	Defined datasets	Training set	Validation set	Test set	Works[b]
CVC-EndoSceneStill [36]	WP images	BM (border, polyp, lumen and specular lights classes)	Manual by experts	500 × 574; 384 × 288	N/A	36	912	Yes	547	183	182	5
CVC-ColonDB [22]	WP images	BM (polyp)	Manual by experts	500 × 574	N/A	15	300	No	–	–	–	12
CVC-ClinicDB [21]	WP images	BM (polyp)	Manual by experts	384 × 288	0-Ip, 0-Is, 0-IIa, 0-IIb	23	612	No	–	–	–	20
ETIS-LARIB [24]	WP images	BM (polyp)	Manual by experts	1255 × 966	N/A	44 seqs	196	No	–	–	–	10
ASU-Mayo Clinic [88]	Video (WP and WO frames)	BM (polyp) (only training set)	Manual by experts	712 × 480; 856 × 480; 1920 × 1080	N/A	N/A	38 videos (8,591 WP; 27,979 WO)	Yes	20 videos (4,278 WP; 14,718 WO)	–	18 videos (4,313 WP; 13,261 WO)	22
CVC-VideoClinicDB [89,87]	Video (WP and WO frames)	BM (polyp) (only training set)	Elliptical approximation by experts	384 × 288	0-Is, 0-Ip, 0-IIa	18 seqs	36 videos	Yes	18 videos (11,954 frames)	–	18 videos (18,733 frames)	6
Kvasir [90]	Multi-Class Image Dataset, containing the polyp class	Label	N/A	from 720 × 576 up to 1920 × 1072	N/A	N/A	500 images in polyp class	No	–	–	–	3
ImageNet [81,82]	Natural images	Label and BB	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	14,197,122	Yes	Up to 1.2 million	≈50,000	Up to 150,000	12
MSCoco [83,84]	Natural images	BM (171 classes), BB, person keypoints, captions	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	330,000	Yes	Up to 118,287	Up to 5,000	Up to 40,670	1
Pascal VOC [85,86]	Natural images	BM (20 classes), BB	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	19,041 images 40,657 instances[c]	Yes	9,477 images 20,250 instances	9,564 images 20,407 instances	N/A	2

WP: With polyp; WO: without polyp; BM: binary mask; BB: bounding-box; seqs: sequences; N/A: not available.
[a] According to Paris classification.
[b] Total add more than retrieved works because several authors use more than one database.
[c] Only training and validation sets.

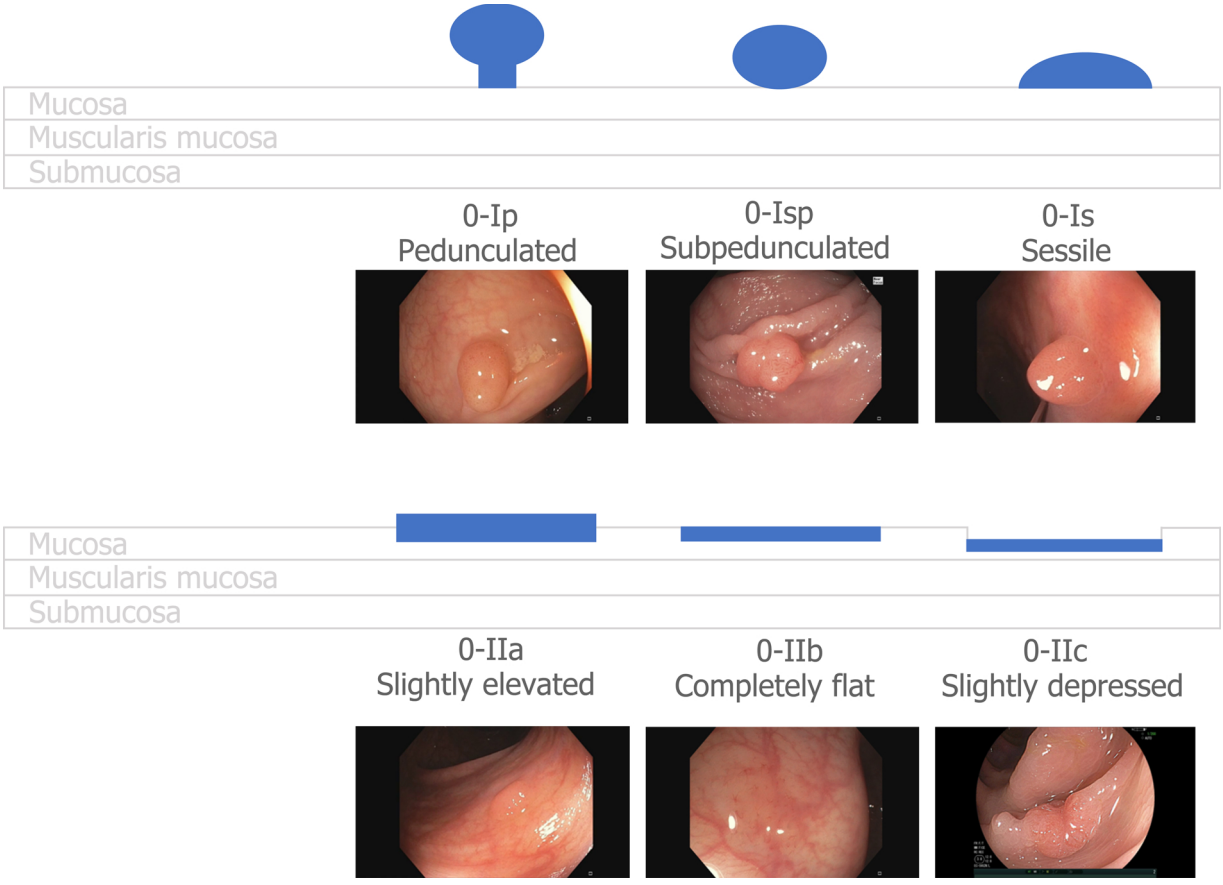

Fig. 4. Paris classification. Adapted from [24,95].

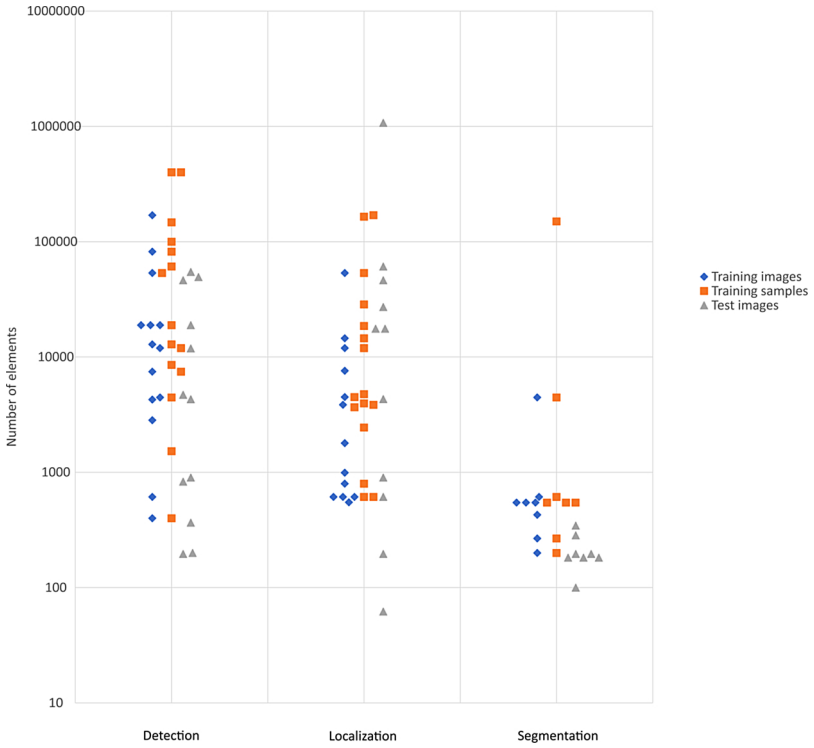

Fig. 5. Number of elements used for training and test, categorized by the main task of the work. 'Elements' refers to images for training and test sets as well as training samples, which might differ from training images if patches are extracted or data augmentation is applied.

Table 5
Transformations for data augmentation on full images.

Task	Work	Creation/flow	Rotation	Translation	Sampling	Zoom	Shearing	Warping	Flip	Gaussian noise	Contrast	Brightness	Crop	Increment
Detection	Aksenov et al. [50]	Creation	–	–	–	–	–	–	–	Yes	–	Yes	–	×1[a]
	Mohammed et al. [57]	Creation, on the flow	(10°,350°)	(−10, 10)	–	(1, 1.3)	–	–	H/V	–	–	–	Yes	×2
	Murthy et al. [53]	Creation	90°, 180°, 270°	–	–	–	–	–	H	–	–	–	–	×9
Localization	Shin et al. [60]	Creation	90°, 180°, 270°	–	–	–	–	–	H/V	–	–	–	–	×6
			90°, 180°, 270°	–	–	10%	10%, 30%,50%	–	H/V	–	–	–	–	×30
			90°, 180°, 270°	–	different resolutions	10%	10%, 30%,50%	–	H/V	–	–	Yes	–	×48
	Zhang et al. [62]	On the flow	Yes	–	–	–	–	–	H/V	Yes	Yes	Yes	–	×1
	Zheng et al. [65]	Creation	90°, 180°, 270°	–	–	–	–	–	–	–	–	–	–	×3
	Yu et al. [61]	Creation	90°, 180°, 270°	(−3, 3)	(−10, 10)	–	–	–	–	–	–	–	–	×22
Segmentation	Wichakam et al. [69]	On the flow	Up to 180°	(0, 20%) H/V	–	(-0.8, 1.2)	(0, 0.2)	–	H/V	–	–	–	–	×1
	Wickstrøm et al. [70]	On the flow	(-90°, 90°)	–	–	(0.8, 1.2)	(0, 0.4)	–	–	–	–	–	Yes	×1
	Bardhi et al. [73]	On the flow	Yes	Yes	0.5 dropout	–	Yes	–	H/V	Yes	–	–	Yes	×1
	Li et al. [74]	Creation	Yes	H/V	–	–	–	–	–	Yes	Yes	–	–	×333
	Vázquez et al. [36]	On the flow	(0°, 180°)	–	–	(0.9, 1.1)	(0, 0.4)	(0, 10)	–	–	–	–	–	×1
Detection and localization	Pogorelov et al. [54]	Creation	20° steps	–	–	–	–	–	H	–	–	–	–	×35
	Urban et al. [56]	On the flow	(0°, 90°)	–	–	–	Yes	–	H/V	–	–	–	–	×1
Detection and segmentation	Brandao et al. [58]	–	–	–	–	–	–	–	Random	–	–	–	–	×1

[a] They modify 40% of the images to increase variability, but do not create more new training examples. H: horizontal; V: vertical.

classification are mutually exclusive, six Venn's diagrams are necessary to show the overlap between the categories of the secondary classification. Venn's diagrams have been created with InteractiVenn [104].

End-to-end and hybrid methods have similar proportion in detection and localization (close to 50% each), but in segmentation, the end-to-end methods vastly surpass hybrid ones (9 vs 1). Regarding the secondary categorization, classification and patch-based ranked equally for the detection task; while bounding-box and semantic segmentation are the preferred approaches for localization and segmentation, respectively. Hybrid, patch-based methods and end-to-end, semantic segmentation methods are the most usual combinations.

Fig. 8 shows how the different types of methods have had presence along the years analyzed in this systematic review. In summary, the tendency goes towards the use of end-to-end methods over hybrid ones, as deep learning is gaining more and more capabilities to address complex problems as a whole, rather than being used as a component to codify the image into features that are afterwards further analyzed. Similarly, semantic segmentation has raised interest over the last two years as it comprises a straightforward and seamless method for polyp identification.

The use of data augmentation and pre-trained networks in the retrieved works have also been analyzed. Fine-tuning is a well-known alternative to training a network from scratch when the labelled training data is limited. In this case, pre-trained networks on a large labelled dataset from a different application is used as starting point [45]. Fig. 9 shows the number of training samples after applying data augmentation, if any, of each method according to the type of approach. There is no clear trend in the use of data augmentation strategies combined with pre-trained models in the considered approaches. It might be expected that using pre-trained models would be linked to a lower number of training samples, but findings do not show so.

Lastly, it is worth mentioning that all retrieved works apply supervised learning, relying its training on a labelled set of images. Other learning approaches already applied in different medical fields, such as unsupervised learning [105] or few-shot learning [106,107] have not been applied for polyp detection, localization or segmentation yet. Unsupervised learning might be more difficult to apply because the polyp area is usually a small portion of the image, while the rest presents a high level of similarity, as there is also healthy mucosa in an image labelled as with polyp.

In the following sections, methods are briefly described, grouped accordingly to the primary and secondary classifications. When methods fall into two groups of the secondary classification, they are each indicated in a different paragraph.

5.1. Methods for polyp detection

5.1.1. Hybrid methods

5.1.1.1. Feature extractor and patch-based. Several authors have compared the use of hand-crafted advanced features with simple fine-tuned classification CNNs, and in all cases the CNN-based approaches overcame the manual feature extraction. Shin et al. [42] demonstrate that features obtained with a classification CNN perform much better than hand-crafted features even when using state-of-art histogram of oriented gradients descriptors. They employ a basic architecture of three convolution:max pooling layers followed by a fully connected layer with 256 neurons image patches to feed an SVM.

Similarly, Taha et al. [41] perform a comparative testing of a shallow classifier fed either with hand-crafted methods or with features extracted using a pre-trained classification CNN. These authors employ AlexNet to obtain a 1,000-dimensional feature vector that is the input to an SVM.

In these works, images patches are used, so these methods can also be classified as patch-based detection but without localization information, as location of patches is not considered.

5.1.1.2. Patch-based. In this case, Yuan et al. [44] propose a 2-stage method. In first place, candidates are detected by the analysis of edges. Patches are then cropped around the polyp candidates and the resulting candidate patches are analyzed by a classification network based on AlexNet.

Tajbakhsh et al. in various works [43,45–47] extend this method by using a pre-detection stage based on edge maps and voting schemas to extract all suitable candidates that are classified by AlexNet, either trained from scratch or fine-tuned, depending on the work. Oriented patches are extracted based on shape features, conforming the set of candidates.

5.1.1.3. Classification and patch-based. In this case, Axyonov et al. [48] combine image contrasting and the K-means-with-connectivity-constraint segmentation method to identify regions with similar pixels, which are then classified into polyp or non-polyp region using AlexNet as classifier.

5.1.2. End-to-end methods

5.1.2.1. Classification. Akbari et al. [49] use a CNN made of four convolutional and pooling layers plus two fully connected layers. This CNN is combined with binarized weights and kernels to reduce the CNN size, so it is suitable for implementation in portable medical hardware with limited memory.

On the other hand, Aksenov et al. [50] combine various classification networks through classifier assembling to get higher accuracies, basing the final result on the average result of the three ensembled models. Each model presents a different number of layers as well as different configurations for the filters.

Itoh et al. [51] aim at detecting polyps as a prior stage to polyp size estimation. In their approach, they use a 3D CNN (C3dNet or C3D) [108] that exploits both the spatial structure and the temporal features present in colonoscopy videos by using 3D convolutional filters and 3D pooling layers. The input is a sequence of 16 consecutive frames for which the CAD system provides an output probability of being a sequence with or without polyp. The same network is used by Misawa et al. [52], who focused their effort on a clinical trial to measure the sensibility of the C3dNet.

A different approach is followed by Mo et al. [55], who employ a Faster R-CNN with VGG16 as backbone. In this case, they use an approximately joint optimization, which takes a mini-batch as input and optimizes both the classification and regression losses at the same time. The classification tail allows for detecting polyps in a frame by giving a probability level.

Classification networks have also been proven successful even on colon cancer screening programmes [56]. In this case, they test pre-trained VGG16, VGG19 and ResNet50 with a final binary classification layer, as well as custom, trained-from-scratch CNNs for polyp detection without further processing. Images were classified on polyp or non-polyp classes.

Lastly, Murthy et al. [53] introduce a cascaded deep decision network (CDDN) for classification. In the first learning stage, samples are classified by a pre-trained network. In the second training stage, samples classified with high confidence are discarded to place efforts on the most challenging samples. This process is repeated on successive stages, using previous stage's features, and obtaining a better separation over samples. In this particular case, a 2-stage network is proposed.

5.1.2.2. Classification and semantic segmentation. Pogorelov et al. [54] compare three detection approaches: (1) hand-crafted global features and a logistic model tree classifier; (2) fine-tuned well-known architectures, such as Xception [109], VGG19 and ResNet50; and (3) a pixel-wise segmentation GAN with a threshold on the number of positively labelled pixels. They use different training and testing sets for the three suggested methods and find that the GAN approach performs better.

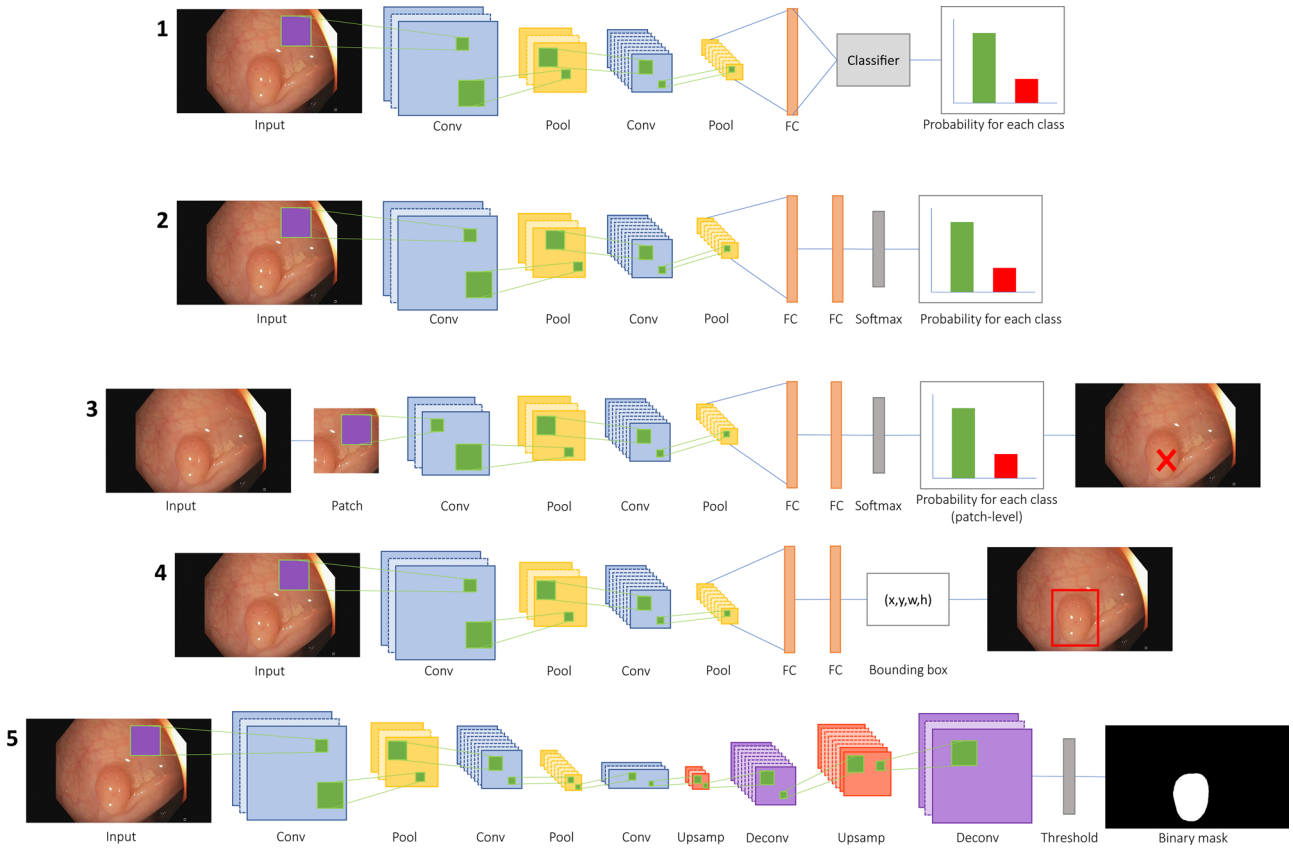

Fig. 6. Schematic representation of the five considered approaches in this review. From top to bottom, (1) feature extractor, (2) classification, (3) patch-based, (4) bounding-box, and (5) semantic segmentation. Each type of layer is represented by a different colour: convolutional layer (conv); pooling layer (pool), fully connected layer (FC), upsampling layer (upsamp) and deconvolutional layer (deconv). The receptive field is marked with a green square. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

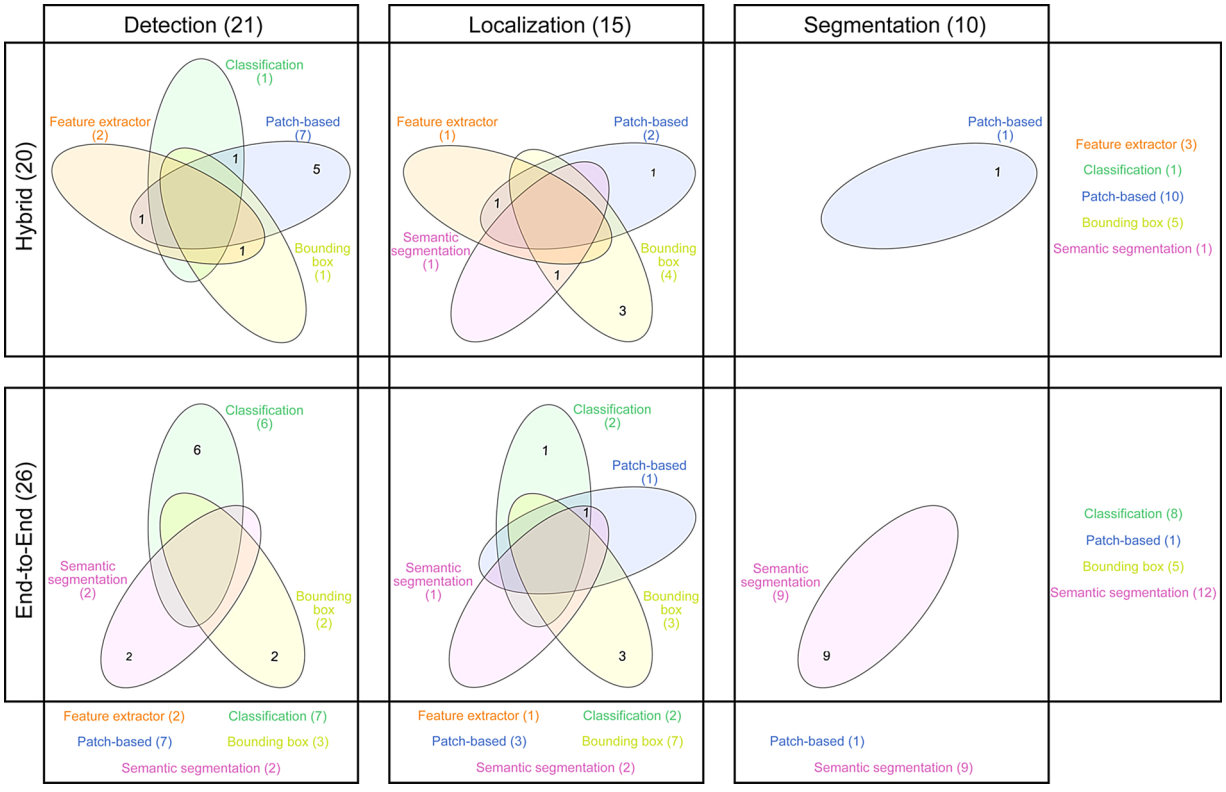

Fig. 7. Categorization of works per tasks.

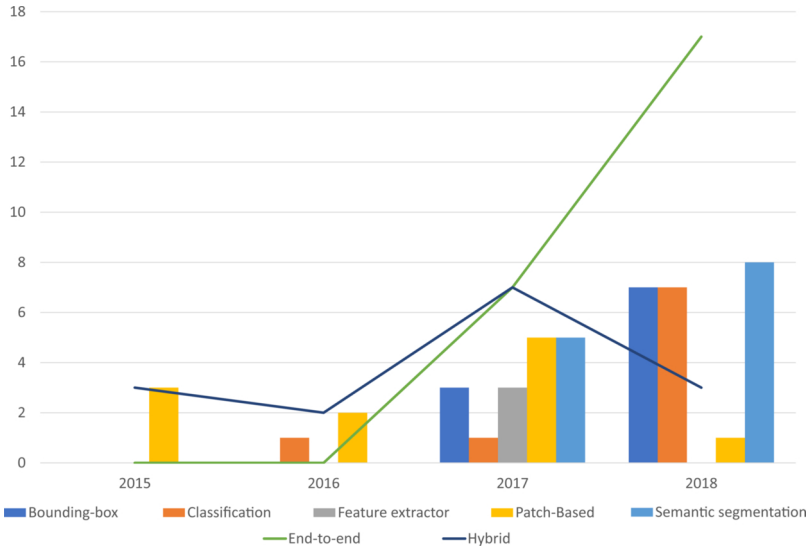

Fig. 8. Trends of the different approaches along the years.

5.1.2.3. Semantic segmentation. Mohammed et al. [57] design the Y-Net, which combines two encoders (both use VGG19 as backbone; pre-trained for encoder 1, while trained-from-scratch for encoder 2) and one decoder to produce a pixel-wise segmentation. Detection is considered when the Intersection over Union of the predicted result and the ground truth is greater than 0.90. In contrast, although Brandao et al. [58] also base detection on the semantic segmentation performed with different FCNs, they consider any degree of overlap between the predicted result and the ground truth. Therefore, not setting an overlap minimum might lead to better metrics in comparison to Mohammed et al. [57], who set their overlap threshold at 0.9.

5.1.3. Comparison of detection methods

Although it is difficult to compare results because of the different datasets, we focus firstly on accuracy as an important indicator to compare between models and approaches (Table 6). With this in mind, Mo et al. [55] deliver the best results in detection. These results are obtained using three different test sets (CVC-ClinicDB, CVC-ColonDB and CVC-EndoSceneStill) and averaging the results obtained for each one. Urban et al. [56] also achieve a high accuracy. In this case, the accuracy is identical in both cross-validation and independent test set of a proprietary dataset. Remarkable results in a test set from ASU-Mayo Clinic database are also those presented by Yuan et al. [44], who achieve an accuracy of 0.9147, with a recall of 0.9176. Akbari et al. [49] also use the same ASU-Mayo Clinic database but with a distinct test set. In this work, the accuracy obtained is 0.908, and the recall, precision and specificity are 0.6838, 0.7434 and 0.9497, respectively. Confidence intervals (CI) are indicated when provided by authors. This applies also to Sections 5.2.3 and 5.3.3.

5.2. Methods for polyp localization

5.2.1. Hybrid methods

5.2.1.1. Patch-based. Park et al. [59] use an architecture composed of three convolutional layers and three max-pooling layers to obtain a 60-dimension feature vector for each patch obtained at three different image scales. The resulting 180-dimension feature vector is used to classify the centre pixel of the patch either as polyp or non-polyp, through a fully connected network with 256 hidden nodes. A probability map is created based on these classified pixels. This map is smoothed with a 5×5 Gaussian filter and 9×9 median filter. Afterwards, it is thresholded, setting to 0 those pixels with probability lower than 0.65. Lastly, connected components of non-zero regions are identified, and the polyp centre is obtained by calculating the centre of mass of each connected component.

5.2.1.2. Feature extractor and patch-based. Billah et al. [64] use a 10-layer CNN to extract features from the last fully connected layer, which are then combined with wavelet features and all of them fed into an SVM for classification into polyp or non-polyp. Patches corresponding to a sliding window are the input, so location of the polyp is found by averaging regions with higher probabilities of being polyp.

5.2.1.3. Bounding-box. The method of Shin et al. [60] combines a region proposal network (RPN), a detector and post-learning approach in the so-called Faster R-CNN, using Inception ResNet-v2 [110] as backbone. Out of an image, the RPN proposes rectangular candidate regions that are the input to the detector, which classifies them into containing or not containing a polyp. The detection is further improved by the post-learning approach, based on false positive learning and off-line learning.

Yu et al. [61] integrate temporal information into the model. Their 3D-FCN is capable of learning more representative spatio-temporal features from colonoscopy videos and hence has more powerful discrimination capability. This 3D network consists of a 3D extension of a 2D fully convolutional segmentation network that uses a 16 frames video entrance to extract the temporal features. An offline 3D-FCN is firstly trained, which is combined with an online 3D-FCN incrementally updated for each input video to remove false positives. Outputs of these two 3D-FCNs are combined to obtain the final detection results.

Moreover, Zhang et al. [62] propose an evolution of a YOLO detection network [111], ResYOLO. The previous detection output of the network is integrated into the following prediction to assure prediction regularization. Besides, temporal information is incorporated as an online object tracker. They also analyzed that the inclusion of temporal information on the network improves the detection rates.

5.2.1.4. Bounding-box and semantic segmentation. SegNet is well-known for semantic segmentation of natural images, so it is the network selected by Wang et al. [63]. Although SegNet provides a pixel-wise labelling in the form of a probability map, there is a post-processing stage that transforms it into the corresponding bounding-box for the polyp class.

5.2.2. End-to-end methods

5.2.2.1. Bounding-box. As mentioned before, Mo et al. [55] use a Faster

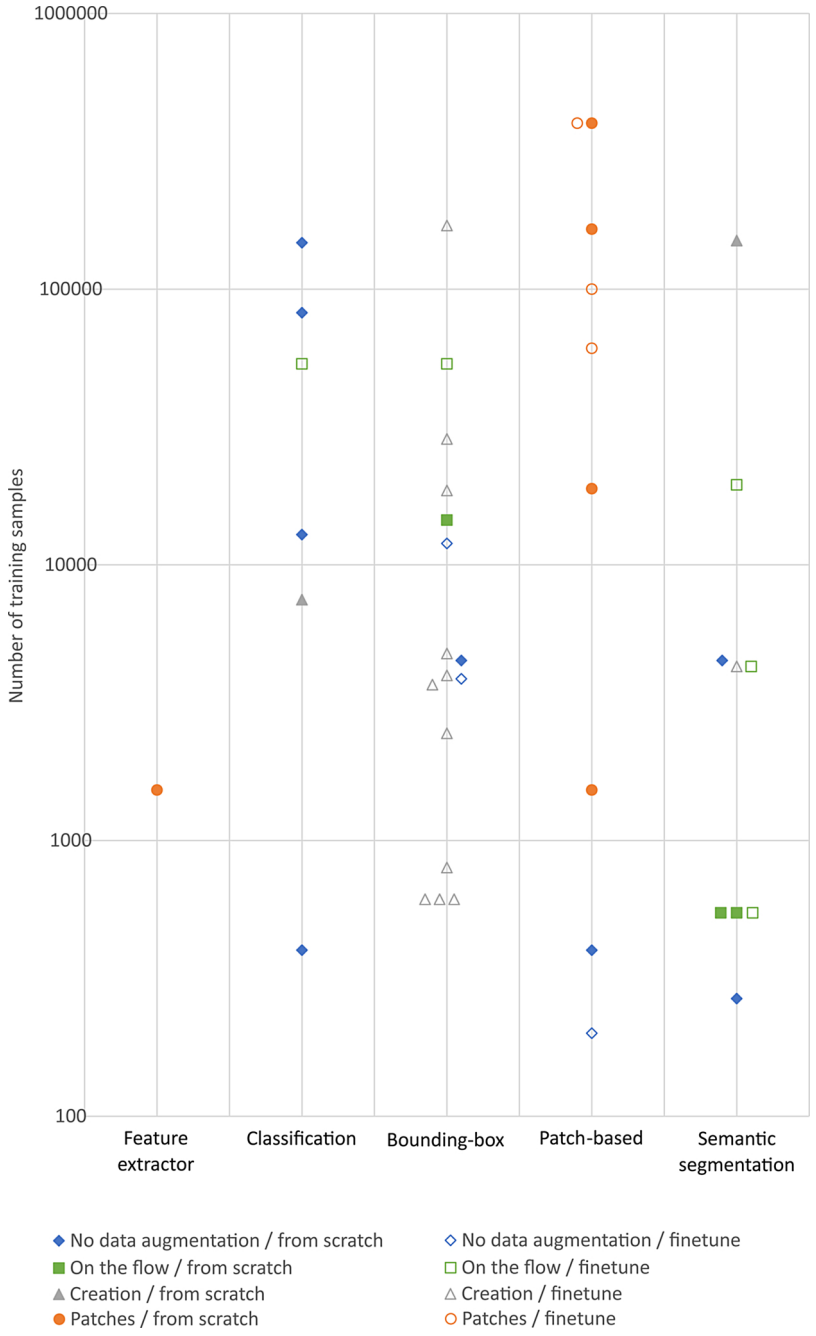

Fig. 9. Training samples per approach and type of data augmentation. Solid items correspond to methods where networks are trained from scratch, while hollow items correspond to fine-tuned networks. In the horizontal axis, the different approaches are indicated. The series corresponds to the different strategies for data augmentation.

Table 6
Summary of detection methods results.

Work	Accuracy	Precision	Recall	AUC	Specificity
Mo et al. [55]	0.985	1.000	0.985		
Urban et al. [56]	0.964			0.974	
Yuan et al. [44]	0.915		0.918		
Akbari et al. [49]	0.908	0.743	0.684		0.950

AUC:area under the curve

R-CNN with VGG16 as backbone. In this case, the regression tail provides the coordinates of the bounding-box indicating the location of the polyp.

Urban et al. [56] compare models trained from scratch to pre-trained models (VGG16 and VGG19 with a final regression layer), obtaining a higher Dice for the latter when ground truth and predicted bounding-boxes are compared.

On the other hand, YOLO is selected by Zheng et al. [65] as the detector network, without further modifications on the network and proceeding only with a fine-tuning of a pre-trained model.

Pogorelov et al. [66] compare two different approaches, TensorBox and Darknet-YOLO, both aiming at detecting objects in images. On one hand, TensorBox avoids multiple detections of the same object by using an RNN with an LSTM. On the other hand, Darknet-YOLO is based on a CNN, therefore encoding contextual information about classes as well as their appearance. This results in a better generalization of objects' representation. In both cases, the methods return sets of rectangles

marking possible polyp locations together with corresponding location confidence values.

5.2.2.2. Classification and patch-based. Pogorelov et al. [54] use sliding windows to feed the classification models (hand-crafted global features, fine-tuned networks and GAN approach grounded on V-GAN [112] modified by adding an activation layer to generate a per-pixel image segmentation, so detection is based on a minimum number of activated pixels) and then reconstruct a coarse localization map by grouping-back the processed patches.

5.2.3. Comparison of localization methods

Out of all localization methods (Table 7), Billah et al. [64] report the highest metrics. In this case, colour wavelet features and CNN features are combined, so it is not possible to determine which type of feature has a greater influence on the result. Besides, they do not report the testing data set, neither in terms of number of images nor their origin, which makes it more difficult to compare results. On the contrary, Mo et al. [55] clearly indicate the testing datasets (CVC-ClinicDB, CVC-ColonDB and CVC-EndoSceneStill), reporting a mean recall which is comparable to the previous work, but providing a more solid evidence and reproducibility of results. Nevertheless, no further works use the same validation set. In this regard, Yu et al. [61] and Zhang et al. [62] do follow the rules of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge in terms of testing dataset. This way, they provide a fair and straightforward comparison to other methods participating in the challenge [24]. Although none of them outperform ASU, the winning method, in precision, they do provide better recall, F1-score and F2-score, showing a more balanced performance.

Table 7
Summary of localization methods results.

Work	Accuracy	Recall	F1-score	F2-score
Billah et al. [64]	0.987	0.988		
Mo et al. [55]		0.981		
Yu et al. [61]	0.881	0.710	0.786	0.739
Zhang et al. [62][a]	0.886	0.716 (0.703, 0.730)	0.792 (0.780, 0.804)	0.744 (0.732, 0.757)
ASU [24]	0.935	0.611	0.739	0.657

[a] 95% confidence intervals are provided between brackets.

5.3. Methods for polyp segmentation

5.3.1. Hybrid methods

5.3.1.1. Patch-based. The focus of Zhang et al. [67] is placed on the use of texton-based spatial features for detailed classification that is used to remove false positives based on local textural analysis. In this case, results of the FCN-8s [101] are used to extract the image region proposals. These regions are refined using texton-based patch representation, which is followed by a random forest classifier to provide the final segmentation.

5.3.2. End-to-end methods

5.3.2.1. Semantic segmentation. Nguyen and Lee [68] take an encoder-decoder model as basis and then produce the polyp segmentation using a model combination by training the encoder-decoder model with three different resolutions databases.

Besides, Wichakam et al. [69] present a compressed FCN that reduce the number of parameters of the feature vector extracted by the network to minimize the computational time showing faster convergence and increased performance for polyp segmentation. The model is compressed by substituting two $7 \times 7 \times 4096$ convolutionalized layers by one $7 \times 7 \times 512$ convolutionalized layer, reducing the number of trainable weights in a significant manner.

The proposal of Wickstrøm et al. [70] is to enhance two traditional encoder-decoder networks (FCN-8s and SegNet, using both VGG16 as encoder) by including batch normalization after each layer and dropout after the three central encoders and decoders. They also analyze uncertainty and interpretability of the models.

On the other hand, Xiao et al. [71] combine LSTM with DeepLab-v3 in parallel. While the latter learns and extracts polyp features thanks its wide field-of-view and higher resolution, the former aims at preserving the information of the polyp location using the information stored in the memory cells These are regulated through the input, forget and output gates.

The U-Net has been modified into the U-Net++ by Zhou et al. [72]. The main difference is the inclusion of nested dense convolutional blocks that bridge the semantic gap between the feature maps of the encoder and decoder prior to fusion.

Bardhi et al. [73] directly use SegNet and train it from scratch on different datasets, while Brandao et al. [58] transform several traditional classification backbones for the segmentation networks into FCNs and prove that VGG16 backbone works better than GoogLeNet and AlexNet.

Similarly, Li et al. [74] propose an FCN and U-Net based segmentation network. The encoder is composed of 8 convolution layers, 8 rectified linear unit (ReLU) layers and 5 pooling layers, while the decoder includes 5 deconvolution layers, 5 concat layers, 6 convolution layers and 6 ReLU layers. Both stages are linked using skipping connections.

Finally, Vázquez et al. [36] provide an exhaustive benchmark showing that FCNs outperform previous results. They implement the FCN-8s architecture and test the influence of data augmentation and number of classes to be segmented on the network performance.

5.3.3. Comparison of segmentation methods

Table 8 summarizes the results of the segmentation methods. The highest Intersection over Union (IoU) values are obtained by Xiao et al. [71], in 345 images from CVC-ClinicDB, and Nguyen and Lee [68], reporting on the ETIS-LARIB dataset. In both cases, mean IoU is reported, so both polyp and background classes are considered, which explains to a great extent those values close to 1. This fact can be clearly seen in the work by Wickstrøm et al. [70]. IoU for the polyp class is 0.587 but mean IoU is equal to 0.767 thanks to the value of IoU for the background, as high as 0.946.

The comparison of methods that use CVC-EndoSceneStill is fair and straight-forward, as the dataset owners provide its division into training, validation and testing sets. Therefore, Vázquez et al. [36], Wichakam et al. [69] and Wickstrøm et al. [70] report results on the same 182 images. Regretfully, and although the CVC-EndoSceneStill benchmark provides a set of metrics (IoU and accuracy), not all authors calculate them. Wichakam et al. [69] follow instead the metrics given by the Pascal VOC challenge [85]. Vázquez et al. [36] report slightly higher values than Wickstrøm et al. [70] in terms of accuracy but it is on the contrary when IoU is considered. In this regard, Wichakam et al. [69] report an intermediate value of IoU.

Bardhi et al. [73] do not clearly state the division into training and testing datasets; therefore results, although showing high values, should be interpreted carefully. The only hybrid method for segmentation [67] obtains comparable results to the rest of end-to-end methods.

In the segmentation task, it is important to remark the influence of including the background class when calculating metrics. Since background usually means the largest area within a frame in comparison to the polyp class, background affects the results by increasing the metric value even when the segmented result is poor. This issue is further discussed in Section 6.

5.4. Advantages and disadvantages

Table 9 gathers the main advantages and disadvantages of the categories on which the works have been classified.

5.5. Loss functions

The final goal against which the network is optimized is completely defined by the loss function. While traditional loss functions for classification (such as negative log-likelihood – NLL – or cross entropy), regression (i.e. L1-loss or mean squared error – MSE) and distribution matching tasks (such as the Kullback-Leibler – KL – divergence) might be generalizable enough for most of the problems, the selection of the loss function is more relevant for complex problems such as the segmentation of unbalanced classes [113]. In these cases, an inappropriate formulation of the loss function might cause that the network converges into a minima where the task is not achieved because the network does not behave as expected. In this review, only 22.86% of the analyzed works present an analysis on the selected loss function and less than 15% of the authors use a custom loss function. These works are briefly commented below.

The loss in the Faster R-CNN used by Mo et al. [55] consist of a classification loss and a bounding-box regression loss, using parameterized coordinates to minimize the influence of scales during training. Besides, Zhang et al. [62] employ a loss function comprised of loss for grids labelled as object (polyp) and non-object. Furthermore, other three works include the Dice coefficient. On one hand, Wichakam et al. [69] optimize the network with a custom loss function simply computed as $1 - Dice$. On the other hand, while Zhou et al. [72] combine this coefficient with binary cross-entropy, Mohammed et al. [57] use the weight binary entropy instead.

When deciding the loss function for detection, it is important to know beforehand the proportion of polyp and non-polyp images in the training set. If it is balanced, traditional loss functions might be enough. Otherwise, it would be advisable to use a loss function intended to overcome the unbalance towards the positive class, such as the weighted binary cross entropy. We also agree on including overlap measures such as the Dice coefficient in the loss function for segmentation tasks when classes are highly unbalanced, as this type of functions have been proved to be more robust [113]. Since segmentation can be done as a pixel-wise binary classification of unbalanced classes, comments for detection can also be applied here.

6. Metrics

There is a wide variety of metrics found in this systematic review. In all cases, metrics are intended to compare the prediction of the method, either for detection, localization or segmentation, against the ground truth, which might be a label, a bounding-box or a binary mask. In this regard, many metrics are calculated based on the confusion matrix and its four basic elements (Fig. 10):

- True positives (TP): number of polyp items correctly predicted as polyp.
- True negatives (TN): number of non-polyp items correctly predicted as non-polyp.
- False positives (FP): number of non-polyp items incorrectly predicted as polyp.
- False negatives (FN): number of polyp items incorrectly predicted as non-polyp.

Table 8
Summary of segmentation methods results.

Work	Accuracy	IoU	F1-score	F2-score
Xiao et al. [71]		0.932		
Nguyen and Lee [68]		0.889		
Wickstrøm et al. [70]	0.949	0.767	0.786	0.739
Vázquez et al. [36]	0.968	0.516	0.792	0.744
Bardhi et al. [73]	0.967			
Wichakam et al. [69]		0.694	0.739	0.657

It is also important to point out that some authors compute the confusion matrix at a frame or image level, while others compute it at a pixel level. This mainly depends on the main objective of the work, being more usual to calculate the confusion matrix at frame level for detection and localization, and at pixel level for segmentation.

Table 10 compiles all metrics found during the analysis, providing the mathematical calculation and alternative names. The calculation of the Dice coefficient and IoU metrics based on the confusion matrix can be used when calculating the confusion matrix at pixel level by comparing two binary masks. These metrics are also more appropriate as they reduce the unbalance due to the large value of true negatives. Two definitions for accuracy have been found in this review. On one hand, only TP are considered in the numerator. This definition can be interpreted as class accuracy, as the sum of accuracies for all classes would reach 1 in an ideal situation. On the other hand, both TP and TN are considered in the numerator. In both cases, the denominator remains unchanged, being the total sum of elements (TP + TN + FP + FN). In the case where the formula is not explicitly provided, accuracy has been considered as correct classified items (TN + TP) divided by the total number of elements.

Table 11 summarizes the metrics employed for the different considered tasks. There is no standard to follow in terms of reporting metrics; therefore for one single task authors report different metrics. Overall, recall is the most used metric followed by precision, accuracy and F1-score in similar proportion. As for each of the analyzed tasks, recall at frame level is the predominant metric for detection and localization, while IoU is the most used metric for segmentation, closely followed by accuracy at pixel-level.

For detection and localization, recall stands out. This metrics penalizes a high number of FN but does not consider FP. In a clinical setting, both parameters are equally important, as if gastroenterologists are warned unnecessarily, the exploration time might be increased without benefits and they would eventually pay no attention to the CAD system ringing false alarms. Therefore, we consider a more suitable metrics the use of the F1-score, where importance of recall and precision is balanced.

Regarding segmentation metrics, it is important to point out that their selection must consider the properties of the segmentation under evaluation, so when the segment size is smaller than the background (less than 5% of the background in any of the axis, as it is usually the case in polyp segmentation), metrics based on the confusion matrix elements are not the more suitable and it is recommended to substitute them by distance metrics [114]. Therefore, when background is not considered in the calculation of the metric, a smaller increment in its value might be more significant than the same increment in a metric considering the background. Bearing this in mind, reporting accuracy, where true negatives corresponding to the background are included, does not seem to be the most adequate one, despite being one of the most common.

It is noteworthy to remark that traditional object detection metrics have not been found in any of the analyzed works. These metrics, such as average precision (AP) or mean average precision (mAP), measure the average precision of a model for a specific IoU and are commonly used on computer visions challenges [83,115]. This is relevant as these metrics provide a single value estimation on the performance of object detection models. However, although mAP can be used for computational performance estimation, it does not reflect the clinical performance due to its sensitivity for small objects and integration of the different. Thus, in clinical gastrointestinal works where mAP is used to measure the overall performance of the method, this has to be complemented with further analysis such as precision/recall curve analysis [116].

In general terms, metrics are reported for the overall testing set. Only few authors report metrics in a detailed manner, considering polyps characteristics. Misawa et al. [52] analyze the percentage of the video for each of the 50 polyps, for which their Paris classification, size,

Table 9
Advantages and disadvantages of each category of methods.

Category	Advantages	Disadvantages
End-to-end	1. Automatic learning of relevant features 2. Superior performance over hand-crafted methods 3. High degree of automation	1. Requires a large dataset for training 2. Selection of network, dataset and hyperparameters might highly influence the performance
Hybrid	1. Combines useful information of well-known hand-crafted features 2. More convenient with small datasets	1. Selection of features based on the researcher experience and knowledge
Feature extractor	1. Combine useful information of well-known hand-crafted features 2. Using pre-trained networks without fine-tuning does not require any labelled data	1. Not tune the network to the target dataset, potentially leading to suboptimal or even negative transfer results when domain shift is large
Classification	1. Cheapest label is required (polyp/non-polyp per frame)	1. Requires considerably more data to converge
Patch-based	1. Increments the size of the training set by obtaining several patches from one single image	1. Slow method in general
Bounding-box	1. Provides enough information for the clinician, focusing their attention on a suspicious area 2. Better computational speed and inter-patch classification coherence 3. Able to converge without pixel-level labels and still predicting location information 4. Bounding-box annotations are cheaper to label than pixel-level ones 5. Preferred type for localization task	1. Does not provide a pixel-level labelling 2. Dense objects might lead to overlapping bouding-boxes
Semantic segmentation	1. Preferred type for segmentation task	1. Accurate borders might be difficult to obtain

location and pathologic diagnosis are also provided. While most polyps present a detected ratio over 80%, only 3 flat polyps (0-IIa in the Paris classification) do not achieve that level, proving the difficulty for detection of this type of polyps.

Characteristics of the polyps are also considered for metrics reporting in the work of Urban et al. [56]. They compare polyp detection by gastroenterologists with and without using the CAD system. In their first study, they found that 9 sessile polyps were missed (i.e. detected with the CAD system and not found without it). Therefore, it might be expected that such system could improve the ADR. Nevertheless, in both studies flat polyps represent a minority: only 2 out of 45 in the first study and 3 out of 73 in the second one. On the other hand, Wang et al. [63] select to separately report results for small (< 0.5 cm), flat, isochromatic polyps, as they are associated with a higher missing rate, finding that the per-image-sensitivity decreased from 0.9438 (95% CI: 0.9380, 0.9496) in the overall dataset to 0.9165 (95% CI: 0.9021, 0.9309). Lastly, Mo et al. [55], although not providing detailed metrics, found that their method behaved differently for sequences showing small polyps, encountering difficulties for their detection.

7. Clinical perspective

The American Society for Gastrointestinal Endoscopy has a set of publications within the Preservation and Incorporation of Valuable endoscopic Innovations (PIVI) initiative to establish thresholds for incorporating innovative technologies into the clinical practice. While there is a PIVI paper related to the in-vivo real-time assessment of diminutive polyps, there is so far no statement regarding the application of CAD systems which could connect the technical development and metrics to clinical performance metrics, such as the ADR.

In this regard, the ADR is considered one of the main indicators for colonoscopy quality, which is influenced both by the endoscopist and the technical factors [117]. Despite presenting a wide variability, ranging from 12.5% to 68.1% in conventional colonoscopy, it has been proven that new technologies such as Endocuff, G-Eye or full-spectrum colonoscopy might help to increase this indicator, reaching values higher than 80% in some of the works systematically reviewed and meta-analyzed by Castaneda et al. [118]. Only two papers in this review [56,63] mention this clinical concept in their technical studies. So, additional efforts should be taken in future works to connect both technical and clinical outcomes to increase the acceptance of CAD systems based on deep learning in the daily clinical practice.

It has been already proven that a second observer, such as an experienced nurse, improves the ADR even in the event of an experienced gastroenterologist performing the colonoscopy [119]. Therefore, and as already Wang et al. [63] suggest, CAD systems might be used as an "extra pair of eyes" to avoid missing subtle lesions. In this regard, they have recently analyzed the influence of their method on the ADR [120]. They have carried out a non-blinded trial, where patients were prospectively randomized into diagnostic colonoscopies with or without

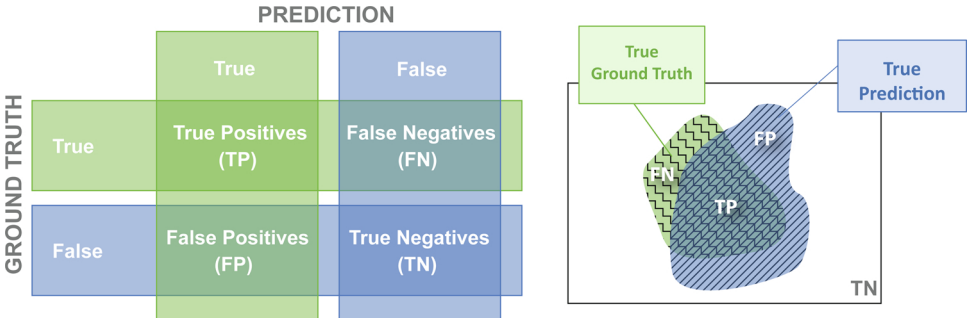

Fig. 10. Confusion matrix. Left side, confusion matrix calculated at frame level. On the right side, elements of the confusion matrix calculated at pixel level by overlapping of two binary masks.

Table 10
Definition of metrics used in the retrieved works.

Metric – alternative names	Calculation
Accuracy	$\text{Acc}_1 = \frac{\text{TP}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$
	$\text{Acc}_2 = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$
Precision	$\text{Prec} = \frac{\text{TP}}{\text{TP}+\text{FP}}$
Recall – true positive rate, sensitivity, pre-class accuracy	$\text{Rec} = \frac{\text{TP}}{\text{TP}+\text{FN}}$
Specificity – pre-class accuracy	$\text{Spec} = \frac{\text{TN}}{\text{TN}+\text{FP}}$
F1-score	$F1 = \frac{2\cdot\text{Prec}\cdot\text{Rec}}{\text{Prec}+\text{Rec}}$
F2-score	$F2 = \frac{5\cdot\text{Prec}\cdot\text{Rec}}{4\cdot\text{Prec}+\text{Rec}}$
Matthew correlation coefficient	$\text{MCC} = \frac{\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$
False positive rate	$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}$
False positive per frame	$\text{FP/Frames} = \frac{\text{FP}}{\#\text{frames}}$
Intersection over Union – Jaccard index	$\text{IoU(PR, GR)} = \frac{\lvert\text{PR}\cap\text{GT}\rvert}{\lvert\text{PR}\cup\text{GT}\rvert} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} = \frac{\text{Dice}}{2-\text{Dice}}$
Dice coefficient	$\text{Dice(PR, GR)} = \frac{2\cdot\lvert\text{PR}\cap\text{GT}\rvert}{\lvert\text{PR}\rvert+\lvert\text{GT}\rvert} = \frac{2\cdot\text{TP}}{2\cdot\text{TP}+\text{FP}+\text{FN}} = \frac{2\cdot\text{IoU}}{1+\text{IoU}}$
Receiver operating characteristic (ROC) Curve	Plot of TPR against FPR
Area under the curve (AUC) − area under the ROC curve (AUROC), ROS curve	$\text{AUC} = 1 - \frac{\text{FPR}-\text{FNR}}{2}$
Free response receiver operating characteristic curve	Plot of TPR against FPR/frame
Temporal coherence	$\text{TC} = \frac{\#\text{correctly detected consecutive frame pairs}}{\#\text{consecutive frame pairs}}$
Polyp detection rate	$\text{PDR} = \frac{\#\text{polyps detected at least in one frame}}{\#\text{polyps in the dataset}}$
Mean processing time per frame	Actual detection processing time taken by a method to process a frame and display the detection result
Reaction time	RT = frame of first detection − frame of first appearance
Mean distance	Mean Euclidean distance between polyp centres

TP: True negatives; TN: true negatives; FP: false positives; FN: false negatives; PR: predicted binary mask; GT: ground truth binary mask.

Table 11
Metrics for reporting detection, localization and segmentation tasks.

Task	Metric	Level	Total of works	References
Detection	Accuracy-2	Frame	7	[42,49,51,53–56]
	Precision	Frame	7	[41,42,49,54,55,57,58]
	Recall	Frame	9	[41,42,44,49,51,54,55,57,58]
	Specificity	Frame	5	[42,44,49,51,54]
	F1-score	Frame	3	[54,55,57]
	F2-score	Frame	2	[55,57]
	AUC	Frame	4	[48,51,52,56]
	FPR	Frame	2	[44,49]
	Dice	–	1	[49]
	MCC	Frame	1	[54]
	FROC	Frame	4	[43,45–47]
	ROC	Frame	1	[50]
Localization	Accuracy-2	Frame	2	[64,66]
	Accuracy-2	Patch	1	[54]
	Precision	Frame	7	[55,59–62,65,66]
	Precision	Patch	1	[54]
	Recall	Frame	9	[55,59–66]
	Recall	Patch	1	[54]
	Specificity	Patch	1	[54]
	Specificity	Frame	3	[62–64]
	F1-score	Frame	7	[55,59–62,65,66]
	F1-score	Patch	1	[54]
	F2-score	Frame	5	[55,60–62,65]
	Dice	Pixel	1	[56]
	MCC	Patch	1	[54]
Segmentation	Accuracy-1	Pixel	1	[36]
	Accuracy-2	Pixel	5	[67,68,70,73,74]
	Precision	Pixel	3	[58,69,74]
	Recall	Pixel	4	[58,67,69,74]
	Specificity	Pixel	2	[67,74]
	F1-score	Pixel	2	[69,74]
	IoU	Pixel	6	[36,68–72]
	Dice	Pixel	3	[67–69]

CAD assistance. In this trial, the main outcome was the ADR. Despite not obtaining the highest metrics for the localization task, they prove a significant increment from 20.3% to 29.1% in the ADR, showing the clinical potential of their method. Therefore, the final aim of any technical development should be to prove the benefit in the clinical practice, rather than only raking on the top considering technical metrics.

Another relevant aspect for CAD systems based on deep learning is matching real time constraints to facilitate clinical application in a live procedure. Efforts should therefore be oriented to exploit the use of videos, rather than isolated images, minimizing the processing time to keep it under the restriction of processing 25 or 30 frames per second. Current deep learning algorithms are able to run in nearly real-time speed [111,121–123] and have not been fully tested in real clinical conditions. As far as the authors know, there is only one commercial system for detection assistance based on artificial intelligence [124], but no technical information has been found on its algorithms. We highly advise for the design of real-time clinical essays to analyze the usability and real performance of the algorithms. On the contrary, images in the datasets usually show polyps in a clean, well-centred state. This is though not the situation in which the CAD system is useful. It would be highly interesting to provide "fly-by" explorations as well as difficult polyps to locate, such as partially hidden or located in folders, to mimic situations with higher clinical value. The difference between images available in the dataset and those in the clinical situations might lead to the fact that is not possible to guarantee that success in the dataset will be reproduced in the clinical environment.

Lastly, it is also important to recognize the limitation of CAD systems in the identification of polyps, as they might be missing due to two main situations: (1) they never appear on the visual field, due to an inappropriate bowel preparation, an inappropriate exposure technique or more importantly, because it is in the 20% of the colon surface that is never surveyed [125]; and (2) missed by the gastroenterologist due to a lack of training or short withdrawal time. While CAD systems might

help the gastroenterologist to not miss polyps, they will not counteract the difficulties of poor bowel preparation or exposure technique.

8. Recommendations and future challenges

Polyp detection, localization and segmentation have been boosted in the last years by the application of deep learning strategies. In this review, we have analyzed the datasets, methods and metrics used up to now. Nevertheless, and despite the success of deep learning over handcrafted methods, there are still challenges to be faced by the scientific community in the upcoming years. In this section we aim at pointing out trends and/or give recommendations on future research lines.

Lack of reproducibility has been raising concerns lately as a critical flaw, especially in the field of health as explained by McDermott et al. [126]. In their paper, they pose recommendations for data providers, researchers and journals and conferences, bearing in mind technical, statistical and conceptual replicability as the three main aspects of reproducibility. Many of our recommendations are aligned with their work, particularized to our field of interest.

8.1. Datasets

The collection of images to create a large dataset is one of the challenges that should be addressed by the clinical community, which would eventually help the technical researchers when developing CAD systems. In this regard, it would be useful to collaboratively work under common guidelines, such as the methodology proposed by Sánchez-Peralta et al. [127], that allows for the systematic acquisition and annotation of colonoscopy videos without modifying the clinical routine. This relates to three issues to consider: (1) collecting abnormal cases, (2) the quality of annotations and (3) the variability of acquisition systems. In the first case, since we are dealing with medical images, images from some particular classes might be more difficult to collect, since they are less frequent to be found in the clinical practice despite having even more relevance for the application of the CAD system than other more frequent classes. This is for instance the situation of flat polyps (types 0-IIb and 0-IIc in the Paris classification as showed in Fig. 4). While gastroenterologists find that CAD systems would be more useful to help in their detection, those types are underrepresented in the publicly available databases [96]. This fact might hinder the efficacy of CAD systems. On the other hand, medical datasets must be annotated by clinicians, taking considerable time. Tools such as GTCreator [87] have been designed to ease the process of ground truth creation, allowing for image and text-based annotations, so it would be advisable to use it to share and revise annotations among different clinicians. Strictly speaking, ground truth is not available for detection methods in colonoscopy as annotated datasets are made by expert calls that represent the best knowledge that can be extracted by an expert from the colonoscopy. For that, agreement with expert's praxis can be analyzed by the Cohen's kappa coefficient. In order to minimize this issue, datasets should be independently annotated by different clinicians and the inconsistencies should be handled by an additional clinician. This will allow to create a dataset closer to an actual ground truth. Furthermore, inter-observer variability is a well-known problem in manual segmentation of medical images [128]. It would be advisable that future datasets provide uncertainty maps to reflect the variability of experts opinion on the same image. Lastly, a collaborative dataset would increase the variability of the acquisition systems, therefore strengthening the CAD systems to accurately work regardless of the endoscope manufacturer.

When producing a dataset, it is also important to establish a set of criteria and guidelines, similarly to the challenge rules, so authors can follow them, facilitating the posterior comparison of methods. In this regard, the distribution of images into training, validation and test sets is a minimum where cross validation performed over the different subjects of the dataset or over fixed sets can be employed for deriving the statistical metrics and confidence intervals. It is essential to assure patient independence of the sets, so all images originated from a patient must fall into one of the sets. For testing, it is recommended that the dataset owners establish a bootstrapping methodology [129], what has been successfully used in other domains [130–132], identifying the number of testing images and the corresponding sample size and repetitions, clearly indicating which images must be considered in each iteration. Bootstrapping consists on analyzing different testing subsets and measuring the posterior distribution of the results. Authors are encouraged to report information on the posterior probability of the metrics or, at least, information on their mean and standard deviation, giving a more realistic vision of the method performance.

All currently available datasets exposed in Section 4.1 only contain medical images and videos from actual patients. None of the articles mentions the use of synthetic image datasets, which could be an alternative to increase the number of samples. There are already efforts in this direction, as in the work of Shin et al. [133]. They employ a conditional GAN to synthetize polyp frames from normal colonoscopy images, by using a filtering-based binary image as input, modified to include the position and size of the polyp. Even though the generated images are qualitatively realistic, they result on deterministic polyps without many variations on colour and texture. Hence, the potential of synthetic image can be further explored and exploited.

8.2. Metrics

Another element to define for fair comparison is the set of metrics to be used for reporting. As seen in Section 6, there is a lack of criteria among the authors to select the most convenient one. In this choice, there are two aspects that play a role. On one side, the type of ground truth available and on the other side, the task to be accomplished. While the former limits the available information (label per image/frame or binary mask), the latter relates to the information worthy to measure. For methods aiming at detecting and locating polyps, it would be advisable to calculate the F1-score at frame level, as it gives a balanced measure between missing polyps (or false negatives – FN) and false alarms (or false positives – FP) [24]. In order to trace easier parallels with the literature of computer vision when analyzing results, we would also recommend including mAP to perform a global technical evaluation of the algorithm, as commented in Section 6. However, mAP analysis should be complemented with a more detailed analysis to validate the real clinical performance of the model. As for segmentation methods, despite being useful, metrics based on elements of the confusion matrix at pixel-level do not detect whether the two masks are similar in shape [114], so it would be useful to complement them with distance metrics that are valid for small segments (such as Hausdorff distance or Mahalanobis distance). In this regard, it is also important to mention the recommendation to calculate agreement measurements with the experts. Specially in the case of detection methods, it would be highly desirable to compute inter-rater measurements, selecting the most suitable method depending on the type of variable (categorical or continuous) and the number of observers [134].

Clinically speaking and agreeing with the concern raised by Robinson et al. [135], we are of the opinion that reporting metrics per patient or per polyp might be more convenient than averaging results all over the test set, as long as the database provides information to identify which polyp and/or patient originate each frame. Polyps and/or patients might have an unequal presence in the database, for example polyp A and B having 8 and 2 frames, respectively, in the test set. If a detection method has 80% accuracy, it might be that all detected frames corresponds to polyp A. If metrics are averaged all over the test set, this situation cannot be identified, but if metrics are provided per polyp (and later averaged), accuracy would be 50%. This way, it would be possible to identify the cases in which the method presents flaws to further work on.

8.3. Data augmentation

In terms of data augmentation, it has been shown that transformations are selected based on subjective criteria upon the researcher experience and that there is no general strategy as the differences in Table 5 show. Efforts are therefore now focused on the identification of the most convenient transformations and ranges in a more objective way. Initiatives to find the most convenient data augmentation policies, such as Smart Augmentation [98], AutoAugment [136] or the use of a Bayesian data augmentation approach [137] have been mostly developed for classification of natural images. Thus, the application of these methodologies to colonoscopy images might boost performance of methods and is therefore worth research.

Besides, it would be also interesting that data augmentation transformations would address particularities of the colonoscopy images, such as illumination effects (specular lights and lack of uniformity); sensor acquisition effects (colour phantoms); image interlacing; sharpening (to improve the quality of the visualized image but increasing the image noise at the same time); information overlay or the presence of the black mask [138]. As these effects might negatively affect the CAD system performance, their inclusion in the training dataset could lead to the model invariance when they are present.

8.4. Network design

The utility of CNNs for polyp detection, localization and segmentation have been already widely explored using supervised learning. In the future, semi-supervised or unsupervised training should be further exploited, relying on smaller datasets which would be eventually easier to compile. On the other hand, LTSMs or RNNs, with small presence in the current review, will be more widely employed in the field of CRC detection. The capability of recurrent networks to model temporal relationships can help creating models tackling temporal information into account for colonoscopy videos. On the other hand, GANs can be employed in the future to generate synthetic data from polyp models to increase the variability of the dataset or to be able to over-express difficult to detect lesions into existing datasets.

Besides, networks have to be designed for taking advantage for the new very high-resolution colonoscopy devices. This can help detecting micro-patterns that can be missed by veteran gastroenterologists, specially from small or flat polyps. This implies challenges on the definition of a network capable of real-time inspection capabilities for very high resolution images.

Reproducibility of deep learning methods comes with associated difficulties. It should be clearly stated in the papers the following information:

- Dataset and distribution of training, validation and test subsets.
- Data pre-processing, if any.
- Model training, including learning parameters such as learning rate, early stopping or weight initialization. The use of seeds when applicable.
- Loss function, as it highly impacts the model performance.
- Hardware and software details, as software packages are continuously updated, and some models might require exceptional hardware conditions.

Randomization of parameters hampers reproducibility, e.g. when images are randomly transformed on the flow. Although ideally all particular seeds and values should be reported, or the code made available, so other researchers could reproduce the experiment, releasing at least the trained models could be an intermediate solution.

Some efforts on reproducibility have also been taken from major conferences organizations such as NeurIPS, with the request of a reproducibility checklist [139]. We strongly advise its use in future publications related to CAD systems based on deep learning.

9. Conclusions

CRC is one of the major causes of death by cancer worldwide. Early detection of precursor lesions has been proved to minimize its incidence, so screening programs are essential. Colonoscopy is a gold standard technique for the detection and treatment of polyps and adenomas. CAD systems might help endoscopists to identify lesions and minimize the ADR.

In the current work, we provided a systematic and comprehensive review of 35 works for detection, localization and segmentation of polyps using deep learning approaches since 2015. We further analyzed seven currently available public databases of colonoscopy images as well as the most common metrics used for reporting. Retrieved methods have been classified according to the approach they follow in a primary (end-to-end vs hybrid methods) and secondary (feature extractor, classification, patch-based, bounding-box and semantic segmentation) classifications. Although there is no common dataset or framework for easy and direct comparison of methods, some trends, advantages and disadvantages have been identified and discussed. Lastly, recommendations and future challenges have been identified.

Despite the great success of deep learning approaches, clinical validation and application is still a must. The creation of larger, more assorted, public datasets; new algorithms requesting less training samples and the creation of a common validation framework will maintain the upwards tendency and will end in the clinical application of CAD systems to assist gastroenterologists to increase the ADR and early detect CRC.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 Research and Innovation Programme under grant agreement No. 732111. The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein. The authors would also like to thank Dr. Federico Soria for his support on this manuscript and Dr. José Carlos Marín, from Hospital 12 de Octubre, and Dr. Ángel Calderón and Dr. Francisco Polo, from Hospital de Basurto, for the images in Fig. 4.

References

[1] World Health Organization. World cancer report 2014. 2014.
[2] International Agency for Research on Cancer. Colorectal cancer factsheet. Tech. rep. 2018http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf.
[3] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65(1):29. https://doi.org/10.3322/caac.21254.
[4] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34. https://doi.org/10.3322/caac.21387.
[5] Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 2018;103:356–87. https://doi.org/10.1016/j.ejca.2018.07.005.
[6] Wiegering A, Ackermann S, Riegel J, Dietz UA, Götze O, Germer CT, et al. Improved survival of patients with colon cancer detected by screening colonoscopy. Int J Colorectal Dis 2016;31(5):1039–45. https://doi.org/10.1007/s00384-015-2501-6.
[7] Williams CB. Insertion technique. In: Waye JD, Rex DK, Williams CB, editors. Colonoscopy. Principles and practice Blackwell Publishing Ltd; 2009. https://doi.org/10.1002/9781444316902.ch40.
[8] Berros Fombella JP, Aguilar Huergo S, García Teijido P. Enfermedades premalignas. In: Sociedad Espa nola de Oncología Médicaeditor. Manual SEOM de prevención y diagnóstico precoz del cáncer. 2017https://seom.org/seomcms/images/stories/recursos/Manual_SEOM_Prevencion_2017.pdf.
[9] Kamiński MF, Hassan C, Bisschops R, Pohl J, Pellisé M, Dekker E, et al. Advanced imaging for detection and differentiation of colorectal neoplasia: European Society

of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2014;46:435–49. https://doi.org/10.1055/s-0034-1365348.
[10] Müller MF, Ibrahim AEK, Arends MJ. Molecular pathological classification of colorectal cancer. Virchows Arch 2016;469(2):125–34. https://doi.org/10.1007/s00428-016-1956-3.
[11] Wieszczy P, Regula J, Kaminski M. Adenoma detection rate and risk of colorectal cancer. Best Pract Res Clin Gastroenterol 2017;31:441–6. https://doi.org/10.1016/j.bpg.2017.07.002.
[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
[13] Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89. https://doi.org/10.1109/ACCESS.2017.2788044.
[14] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–9. https://doi.org/10.1038/s41591-018-0268-3.
[15] Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9(1):3358. https://doi.org/10.1038/s41598-019-40041-7.
[16] Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. Comput Med Imaging Graph 2019;74:61–71. https://doi.org/10.1016/j.compmedimag.2019.02.005.
[17] Byrne MF, Shahidi N, Rex DK. Will Computer-aided detection and diagnosis revolutionize colonoscopy? Gastroenterology 2017;153(6). https://doi.org/10.1053/j.gastro.2017.10.026. 1460–1464.e1.
[18] Iwahori Y, Hagi H, Usami H, Woodham RJ, Wang A, Bhuyan MK, et al. Automatic polyp detection from endoscope image using likelihood map based on edge information. Proceedings of the 6th international conference on pattern recognition applications and methods – volume 1: ICPRAM 2017:402–9. https://doi.org/10.5220/0006189704020409.
[19] Iakovidis DK, Maroulis DE, Karkanis SA. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. Comput Biol Med 2006;36(10):1084–103. https://doi.org/10.1016/j.compbiomed.2005.09.008.
[20] Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F. Texture-based polyp detection in colonoscopy. Bildverarbeitung für die Medizin 2009 2009:346–50. https://doi.org/10.1007/978-3-540-93860-6_70.
[21] Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez de Miguel C, Vilari no F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007.
[22] Bernal J, Sánchez FJ, Vilariño F. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit 2012;45:3166–82. https://doi.org/10.1016/j.patcog.2012.03.002.
[23] Fu JJ, Yu Y-W, Lin H-M, Chai J-W, Chen CC-C. Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging. Comput Med Imaging Graph 2014;38(4):267–75. https://doi.org/10.1016/j.compmedimag.2013.12.009.
[24] Bernal J, Tajbakhsh N, Sánchez FJ, Matuszewski BJ, Chen H, Yu L, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans Med Imaging 2017;36(6):1231–49. https://doi.org/10.1109/TMI.2017.2664042.
[25] Biswas M, Kuppili V, Saba L, Edla DR, Suri HS, Cuadrado-godia E, et al. State-of-the-art review on deep learning in medical imaging. Front Biosci 2019;24:380–406. https://doi.org/10.2741/4725.
[26] Kim J, Hong J, Park H. Prospects of deep learning for medical imaging. Precis Future Med 2018;2(2):37–52. https://doi.org/10.23838/pfm.2018.00030.
[27] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42(11):1–13. https://doi.org/10.1007/s10916-018-1088-1.
[28] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19(1):221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.
[29] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
[30] Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol 2017;10(3):257–73. https://doi.org/10.1007/s12194-017-0406-5.
[31] Greenspan H, van Ginneken B, Summers RM. Deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35(5):1153–9. https://doi.org/10.1109/TMI.2016.2553401.
[32] Prasath VBS. Polyp detection and segmentation from video capsule endoscopy: a review. J Imaging 2017;3(1):1–15. https://doi.org/10.3390/jimaging3010001.
[33] Ahmad OF, Soares AS, Mazomenos E, Brandao P, Vega R, Seward E, et al. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. Lancet Gastroenterol Hepatol 2019;4(1):71–80. https://doi.org/10.1016/S2468-1253(18)30282-6.
[34] Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Comput Med Imaging Graph 2018;70:83–100. https://doi.org/10.1016/j.compmedimag.2018.10.002.
[35] EndoVis Grand Challenge. EndoVis grand challenge. 2019https://endovis.grand-challenge.org/endoscopic_vision_challenge/.
[36] Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering 2017;2017:4037190. https://doi.org/10.1155/2017/4037190.
[37] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62(10):e1–34. https://doi.org/10.1016/j.jclinepi.2009.06.006.
[38] Pogorelov K, Lange TD, Randel KR, Dang-Nguyen D-T, Johansen D, Riegler M. A holistic multimedia system for gastrointestinal tract disease detection. Proceedings of the 8th ACM on multimedia systems conference (MMSys'17) 2017:112–23. https://doi.org/10.1145/3083187.3083189.
[39] Park SY, Sargent D. Colonoscopic polyp detection using convolutional neural networks. Medical imaging 2016: computer-aided diagnosis, vol. 9785 2016. https://doi.org/10.1117/12.2217148.
[40] Ribeiro E, Uhl A, Wimmer G, Häfner M. Exploring deep learning and transfer learning for colonic polyp classification. Comput Math Methods Med 2016;2016:6584725. https://doi.org/10.1155/2016/6584725.
[41] Taha B, Dias J, Wergui N. Convolutional neural network as a feature extractor for automatic polyp detection. 24th IEEE international conference on image processing, ICIP 2017 2017:2060–4. https://doi.org/10.1109/ICIP. 2017.8296644.
[42] Shin Y, Balasingham I, Member S. Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification. 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2017:3277–80. https://doi.org/10.1109/EMBC.2017.8037556.
[43] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. On the necessity of fine-tuned convolutional neural networks for medical imaging. Deep learning and convolutional neural networks for medical image computing: precision medicine, high performance and large-scale datasets 2017:181–93. https://doi.org/10.1007/978-3-319-42999-1_11.
[44] Yuan Z, Izadyyazdanabadi M, Mokkapati D, Panvalkar R, Shin JY, Tajbakhsh N, et al. Automatic polyp detection in colonoscopy videos. Prog Biomed Opt Imaging – Proc SPIE 2017;10133:1–10. https://doi.org/10.1117/12.2254671.
[45] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 2016;35(5):1299–312. https://doi.org/10.1109/TMI.2016.2535302.
[46] Tajbakhsh N, Gurudu SR, Liang J. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. 2015 IEEE 12th international symposium on biomedical imaging 2015:79–83. https://doi.org/10.1109/ISBI.2015.7163821.
[47] Tajbakhsh N, Gurudu SR, Liang J. A comprehensive computer-aided polyp detection system for colonoscopy videos. 24th International Conference on Information Processing in Medical Imaging, IPMI 2015 2015:327–38. https://doi.org/10.1007/978-3-319-19992-4_25.
[48] Axyonov S, Zamyatin A, Liang J, Kostin K. Advanced pattern recognition and deep learning for colon polyp detection. Distributed computer and communication networks: control, computation, communications (DCCN-2016) 2016:27–34http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000616307.
[49] Akbari M, Mohrekesh M, Rafiei S, Reza Soroushmehr SM, Karimi N, Samavi S, et al. Classification of informative frames in colonoscopy videos using convolutional neural networks with binarized weights. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), vol. 2018 2018:65–8. https://doi.org/10.1109/EMBC.2018.8512226. arXiv:1802.01387.
[50] Aksenov S, Kostin K, Ivanova A, Liang J, Zamyatin A. An ensemble of convolutional neural networks for the use in video endoscopy. Sovrem Tehnol Med 2018;10(2):7–17. https://doi.org/10.17691/stm2018.10.2.01.
[51] Itoh H, Roth HR, Lu L, Oda M, Misawa M, Mori Y, et al. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. medical image computing and computer-assisted intervention. MICCAI 2018. Lecture notes in computer science, vol 11071 2018:611–9. https://doi.org/10.1007/978-3-030-00934-2_68.
[52] Misawa M, Kudo SE, Mori Y, Cho T, Kataoka S, Yamauchi A, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018;154(8):2027–9. https://doi.org/10.1053/j.gastro.2018.04.003.
[53] Murthy VN, Singh V, Sun S, Bhattacharya S, Chen T, Comaniciu D. Cascaded deep decision networks for classification of endoscopic images. Proceedings volume 10133, medical imaging 2017: image processing 2017. https://doi.org/10.1117/12.2254333. pp. 10133-1–10133-15.
[54] Pogorelov K, Ostroukhova O, Jeppsson M, Espeland HN, Griwodz C, De Lange T, et al. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. 2018. p. 381–6. https://doi.org/10.1109/BHI.2018.8333444.
[55] Mo X, Tao K, Wang Q, Wang G. An efficient approach for polyps detection in endoscopic videos based on faster R-CNN. 2018 24th international conference on pattern recognition (ICPR) 2018:3929–34https://arxiv.org/abs/1809.01263.
[56] Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Accuracy in screening colonoscopy. Gastroenterology 2018;155(4). https://doi.org/10.1053/j.gastro.2018.06.037. 1069–1078.e8.
[57] Mohammed A, Yildirim S, Farup I, Pedersen M, Hovde Ø. Y-Net: a deep convolutional neural network for polyp detection. 29th British machine vision conference (BMCV2018) 2018http://bmvc2018.org/contents/papers/0487.pdf.
[58] Brandao P, Mazomenos E, Ciuti G, Caliò R, Bianchi F, Menciassi A, et al. Fully convolutional neural networks for polyp segmentation in colonoscopy. Medical imaging 2017: computer-aided diagnosis, vol. 10134 2017:101340F. https://doi.org/10.1117/12.2254361.
[59] Park S, Lee M, Kwak N. Polyp detection in colonoscopy videos using deeply-

learned hierarchical features. 2015http://mipal.snu.ac.kr/images/0/0b/Polyp_short_report.pdf.
[60] Shin Y, Qadir HA, Aabakken L, Bergsland J, Balasingham I. Automatic colon polyp detection using region based deep CNN and post learning approaches. IEEE Access 2018;6:40950–62. https://doi.org/10.1109/ACCESS.2018.2856402.
[61] Yu L, Chen H, Dou Q, Qin J, Heng PA. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J Biomed Health Inform 2017;21(1):65–75. https://doi.org/10.1109/JBHI.2016.2637004.
[62] Zhang R, Zheng Y, Poon CC, Shen D, Lau JY. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recognit 2018;83:209–19. https://doi.org/10.1016/j.patcog.2018.05.026.
[63] Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2018;2:741–8. https://doi.org/10.1038/s41551-018-0301-3.
[64] Billah M, Waheed S, Rahman MM. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. Int J Biomed Imaging 2017;2017:1–10. https://doi.org/10.1155/2017/9545920.
[65] Zheng Y, Zhang R, Yu R, Jiang Y, Mak TWC, Wong SH, et al. Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2018:4142–5. https://doi.org/10.1109/EMBC.2018.8513337.
[66] Pogorelov K, Riegler M, Eskeland SL, de Lange T, Johansen D, Griwodz C, et al. Efficient disease detection in gastrointestinal videos – global features versus neural networks. Multimed Tools Appl 2017;76(21):22493–525. https://doi.org/10.1007/s11042-017-4989-y.
[67] Zhang L, Dolwani S, Ye X. Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons. In: Valdes Hernandez M, González-Castro V, editors. Medical image understanding and analysis. MIUA 2017. Communications in computer and information science, vol. 723 Springer; 2017. p. 707–17. https://doi.org/10.1007/978-3-319-60964-5_62.
[68] Nguyen Q, Lee S-W. Colorectal segmentation using multiple encoder-decoder network in colonoscopy images. 2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE) 2018:208–11. https://doi.org/10.1109/AIKE.2018.00048.
[69] Wichakam I, Panboonyuen T, Udomcharoenchaikit C. Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. MultiMedia modeling. MMM 2018. Lecture notes in computer science, vol 10704 2018:393–404. https://doi.org/10.1007/978-3-319-73603-7_32.
[70] Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. 2018 IEEE international workshop on machine learning for signal processing 2018. https://doi.org/10.1109/MLSP.2018.8516998.
[71] Xiao W-T, Chang L-J, Liu W-M. Semantic segmentation of colorectal polyps with DeepLab and LSTM networks. 2018 IEEE international conference on consumer electronics-Taiwan (ICCE-TW) semantic 2018:115–20. https://doi.org/10.1109/ICCE-China.2018.8448568.
[72] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet + +: a nested U-Net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support. DLMIA 2018, ML-CDS 2018. Lecture notes in computer science, vol. 11045 2018. https://doi.org/10.1007/978-3-030-00889-5_1.
[73] Bardhi O, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A. Automatic colon polyp detection using convolutional encoder-decoder model. 2017 IEEE international symposium on signal processing and information technology (ISSPIT) 2017:445–8. https://doi.org/10.1109/ISSPIT.2017.8388684.
[74] Li Q, Yang G, Chen Z, Huang B, Chen L, Xu D, et al. Colorectal polyp segmentation using a fully convolutional neural network. 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI 2017) 2017:1–5. https://doi.org/10.1109/CISP-BMEI.2017.8301980.
[75] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (NIPS2012) 2012. https://doi.org/10.1145/3065386.
[76] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015http://arxiv.org/abs/1409.1556.
[77] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 IEEE conference on computer vision and pattern recognition (CVPR) 2015. https://doi.org/10.1109/CVPR.2015.7298594.
[78] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR) 2016. https://doi.org/10.1109/CVPR.2016.90.
[79] Badrinarayanan V, Handa A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95. https://doi.org/10.1103/PhysRevX.5.041024.
[80] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention. MICCAI 2015. Lecture notes in computer science, vol. 9351 Springer; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
[81] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition 2009. https://doi.org/10.1109/CVPR.2009.5206848.
[82] Stanford Vision Lab. Stanford University. Princeton University, ImageNet dataset. http://www.image-net.org/.
[83] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 8693 LNCS (Part 5) 2014:740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
[84] COCO Consortium, MSCoco dataset. http://cocodataset.org/#home.
[85] Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes challenge: a retrospective. Int J Comput Vis 2014;111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5.
[86] PASCAL VOC project. PascalVOC dataset. http://host.robots.ox.ac.uk/pascal/VOC/index.html.
[87] Bernal J, Histace A, Masana M, Angermann Q, Sánchez-Montes C, de Miguel CR, et al. GTCreator: a flexible annotation tool for image-based datasets. Int J Comput Assist Radiol Surg 2019;14(2):191–201. https://doi.org/10.1007/s11548-018-1864-x.
[88] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 2016;35(2):630–44. https://doi.org/10.1109/TMI.2015.2487997.
[89] Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, et al. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. Computer assisted and robotic endoscopy and clinical image-based procedures 2017:29–41. https://doi.org/10.1007/978-3-319-67543-5.
[90] Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, et al. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. Proceedings of the 8th ACM multimedia systems conference (MMSys'17) 2017:164–9. https://doi.org/10.1145/3083187.3083212.
[91] Consortium for Open Medical Image Computing, Sub-Challenge Gastrointestinal Image ANAlysis (GIANA); 2017. https://giana.grand-challenge.org/Home/.
[92] Pogorelov K, Randel KR, de Lange T, Eskeland SL, Griwodz C, Johansen D, et al. Nerthus: a bowel preparation quality video dataset. Proceedings of the 8th ACM on multimedia systems conference (MMSys'17) 2017:170–4. https://doi.org/10.1145/3083187.3083216.
[93] Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Trans Med Imaging 2016;35(9):2051–63. https://doi.org/10.1109/TMI.2016.2547947.
[94] Participants in the Paris Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointest Endosc 2003;58(6):S3–43. https://doi.org/10.1016/S0016-5107(03)02159-X.
[95] Endoscopic Classification Review Group. Update on the Paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy 2005;37(6):570–8. https://doi.org/10.1055/s-2005-861352.
[96] Sánchez-Peralta LF, Sánchez-Margallo FM, Bote Chacón J, Soria Gálvez F, López-Saratxaga C, Picón Ruiz A, et al. Is it necessary to improve the colorectal polyps databases for detection CAD systems based on deep learning? Br J Surg 2018;105(S2):5–14.
[97] Eaton-Rosen Z, Bragman F, Ourselin S, Cardoso MJ. Improving data augmentation for medical image segmentation. 1st conference on medical imaging with deep learning (MIDL 2018) 2018. . https://openreview.net/pdf?id=rkBBChjiG.
[98] Lemley J, Bazrafkan S, Corcoran P. Smart augmentation learning an optimal data augmentation strategy. IEEE Access 2017;5:5858–69. https://doi.org/10.1109/ACCESS.2017.2696121.
[99] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016http://www.deeplearningbook.org/.
[100] Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, et al. A survey on deep learning: algorithms, techniques and applications. ACM Comput Surv 2018;51(5). https://doi.org/10.1145/3234150.
[101] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE conference on computer vision and pattern recognition (CVPR) 2015:3431–40. https://doi.org/10.1109/CVPR.2015.7298965.
[102] Jegou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW) 2017:1175–83. https://doi.org/10.1109/CVPRW.2017.156.
[103] Marron JS, Todd MJ, Ahn J. Distance-weighted discrimination. J Am Stat Assoc 2007;102(480):1267–71. https://doi.org/10.2307/27639976.
[104] Heberle H, Meirelles VG, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinform 2015;16(1):1–7. https://doi.org/10.1186/s12859-015-0611-3.
[105] Moriya T, Roth HR, Nakamura S, Oda H, Nagara K, Oda M, et al. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. Proceedings volume 10578, medical imaging 2018: biomedical applications in molecular, structural, and functional imaging 2018. https://doi.org/10.1117/12.2293414.
[106] Medela A, Picon A, Saratxaga CL, Belar O, Cabezón V, Cicchi R, et al. Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. 2019 IEEE 16th international symposium on biomedical imaging (ISBI) 2019.
[107] Medela A, Picon A. Constellation loss: improving the efficiency of deep metric learning loss functions for optimal embedding. 2019https://arxiv.org/abs/1905.10675.
[108] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE international conference on computer vision 2015 Inter 2015:4489–97. https://doi.org/10.

1109/ICCV.2015.510.
[109] Chollet F. Xception: deep learning with depthwise separable convolutions. 2017 IEEE conference on computer vision and pattern recognition (CVPR) 2017:1800–7. https://doi.org/10.1109/CVPR.2017.195.
[110] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence (AAAI-17) 2017:4278–84https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14806/14311.
[111] Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018http://arxiv.org/abs/1804.02767.
[112] Son J, Park SJ, Jung K-H. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. 2017http://arxiv.org/abs/1706.09318.
[113] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMR, Moradi M, editors. Deep learning in medical image analysis and multimodal learning for clinical decision support. DLMIA 2017, ML-CDS 2017. Lecture notes in computer science, vol. 10553 Springer; 2017. p. 240–8. https://doi.org/10.1007/978-3-319-67558-9_28.
[114] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.
[115] Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis 2010;88(2):303–38. https://doi.org/10.1007/s11263-009-0275-4.
[116] Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, et al. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 2019;14(3):1–16. https://doi.org/10.1371/journal.pone.0214133.t005.
[117] Rex DK. Polyp detection at colonoscopy: endoscopist and technical factors. Best Pract Res: Clin Gastroenterol 2017;31(4):425–33. https://doi.org/10.1016/j.bpg.2017.05.010.
[118] Castaneda D, Popov VB, Verheyen E, Wander P, Gross SA. New technologies improve adenoma detection rate, adenoma miss rate, and polyp detection rate: a systematic review and meta-analysis. Gastrointest Endosc 2018;88(2):209–22. https://doi.org/10.1016/j.gie.2018.03.022.
[119] Wang W, Xu L, Bao Z, Sun L, Hu C, Zhou F, et al. Differences with experienced nurse assistance during colonoscopy in detecting polyp and adenoma: a randomized clinical trial. Int J Colorectal Dis 2018;33(5):561–6. https://doi.org/10.1007/s00384-018-3003-0.
[120] Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019:1–7. https://doi.org/10.1136/gutjnl-2018-317500.
[121] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision – ECCV 2016. Springer International Publishing; 2016. p. 21–37.
[122] Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
[123] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. 2017 IEEE international conference on computer vision (ICCV) 2017:2980–8. https://doi.org/10.1109/ICCV.2017.322.
[124] Medtronic, GI Genius. https://www.medtronic.com/covidien/en-us/products/gastrointestinal-artificial-intelligence/gi-genius-intelligent-endoscopy.html.
[125] Edakkanambeth Varayil J, Enders F, Tavanapong W, Oh J, Wong J, de Groen PC. Colonoscopy: what endoscopists inspect under optimal conditions. Gastroenterology 2011;140(5):S–718. https://doi.org/10.1016/s0016-5085(11)62982-x.
[126] Mcdermott MBA, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L. Reproducibility in machine learning for health. Seventh international conference on learning representations ICLR2019 2019http://arxiv.org/abs/1907.01463v1.
[127] Sánchez-Peralta LF, Calderón ÁJ, Cabezón V, Ortega-Morán JF, Sánchez-Margallo FM, Polo F, et al. Systematic acquisition and annotation of clinical cases for the generation of a medical image database. Br J Surg 2019;106(S2):16.
[128] Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol 2019;29(3):1391–9. https://doi.org/10.1007/s00330-018-5695-5.
[129] Freedman DA. Bootstrapping regression models. Ann Stat 1981;9(6):1218–28. https://doi.org/10.1214/aos/1176345638.
[130] Picon A, Irusta U, Álvarez-Gila A, Aramendi E, Alonso-Atienza F, Figuera C, et al. Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia. PLoS One 2019;14(5):e0216756. https://doi.org/10.1371/journal.pone.0216756.
[131] He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artif Intell Med 2019;93:43–9. https://doi.org/10.1016/j.artmed.2018.05.001.
[132] Kooi T, Litjens G, Van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12. https://doi.org/10.1016/j.media.2016.07.007.
[133] Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 2018;6:56007–17. https://doi.org/10.1109/ACCESS.2018.2872717.
[134] Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res 2017;8(4):187–91. https://doi.org/10.4103/picr.PICR.
[135] Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H, et al. Automated quality control in image segmentation: application to the UK Biobank Cardiac MR Imaging study. J Cardiovasc Magn Reson 2019:1–17. https://doi.org/10.1186/s12968-019-0523-x.
[136] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. AutoAugment: learning augmentation policies from data. 2018http://arxiv.org/abs/1805.09501.
[137] Tran T, Pham T, Carneiro G, Palmer L, Reid I. A Bayesian data augmentation approach for learning deep models. 31st conference on neural information processing systems (NIPS 2017) 2017http://arxiv.org/abs/1710.10564.
[138] Bernal J, Sánchez FJ, Rodríguez de Miguel C, Fernández-Esparrach G. Building up the future of colonoscopy? A synergy between clinicians and computer scientists. Screening for colorectal cancer with colonoscopy. InTech; 2015. https://doi.org/10.5772/61012.
[139] Pineau J. The machine learning reproducibility checklist (version 1.0). 2019https://www.cs.mcgill.ca/jpineau/ReproducibilityChecklist.pdf.