Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge

Jorge J Bernal, Nima Tajkbaksh, F J Sánchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn R Rustad, Ilangko Balasingham, et al.

Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge

Jorge Bernal*, Nima Tajkbaksh*, F. Javier Sánchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang+, Aymeric Histace+

***Abstract*—Colonoscopy is the gold standard for colon cancer screening though still some polyps are missed, thus preventing early disease detection and treatment. Several computational systems have been proposed to assist polyp detection during colonoscopy but so far without consistent evaluation. The lack of publicly available annotated databases has made it difficult to compare methods and to assess if they achieve performance levels acceptable for clinical use. The Automatic Polyp Detection sub-challenge, conducted as part of the Endoscopic Vision Challenge (http://endovis.grand-challenge.org) at the international conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2015, was an effort to address this need. In this paper, we report the results of this comparative evaluation of polyp detection methods, as well as describe additional experiments to further explore differences between methods. We define performance metrics and provide evaluation databases that allow comparison of multiple methodologies. Results show that convolutional neural networks (CNNs) are the state of the art. Nevertheless it is also demonstrated that combining different methodologies can lead to an improved overall performance.**

***Index Terms*—Endoscopic vision, Polyp Detection, Hand-crafted features, Machine Learning, Validation Framework**

*Authors contributed equally to this work, +Last position is shared between the authors.

Jorge Bernal and F. Javier Sánchez are with Computer Science Department at Universitat Autònoma de Barcelona and Computer Vision Center, Spain

Nima Tajkbaksh and Jianming Liang are with Arizona State Univ., USA

Aymeric Histace, Quentin Angermann, Olivier Romain and Xavier Dray are with ETIS, ENSEA, Univ. of Cergy-Pontoise, CNRS, Cergy, France. Xavier Dray is also with Lariboisière Hospital-APHP, France.

Hao Chen and Lequan Yu are with Dpt of Computer Science and Engineering, Chinese University of Hong Kong, China

Bjørn Rustad and Ilangko Balasingham are with Oslo University Hospital. Bjørn Rustad is also with OmniVision, University of Oslo, Norway.

Konstantin Pogorelov is with Media Performance Group, Simula Research Laboratory and University of Oslo, Norway

Sungbin Choi is with Seoul National University, Seoul, South Korea

Bogdan J. Matuszewski is with School of Engineering, University of Central Lancashire, Preston, United Kingdom

Quentin Debard is with University of Nice-Sophia Antipolis, Nice, France

Lena Maier-Hein is with the junior group Computer-assisted Interventions, German Cancer Research Center (DKFZ), Germany

Stefanie Speidel is with the Institute for Anthropomatics, Karlsruhe Institute of Technology, Germany

Danail Stoyanov and Patrick Brandao are with the Centre for Medical Image Computing and Dept. of Computer Science, Univ. College London, UK

Henry Córdova, Cristina Sánchez-Montes and Gloria Fernández-Esparrach are with Endoscopy Unit, Gastroenterology Department, Hospital Clínic, IDIBAPS, CIBEREHD, University of Barcelona, Barcelona, Spain

Suryakanth R. Gurudu is with Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, Arizona, USA

I. Introduction

A. Clinical context

Colorectal cancer (CRC) is the third largest cause of cancer deaths in the United States among men and women, and it is expected to have resulted in about $49,196$ deaths in 2016 in the USA [1]. CRC arises from adenomatous polyps (or adenomas), that are growths of glandular tissue originating from the colonic mucosa. Though adenomas are initially benign, they might become malignant over time and spread to adjacent and distant organs such as lymph nodes, liver or lungs, being ultimately responsible for complications and death [2].

CRC prevention is first based on the detection of at-risk patients: those with symptoms (such as hematochezia and anemia), those with positive screening tests (such as a fecal occult blood test or a fecal immunochemical test), and those with a past history of adenoma or with a family history of advanced adenoma or CRC. In these groups of patients, a colonoscopy is proposed to detect polyps before any malignant transformation or at an early cancer stage. This stage refers to the most superficial colon layers, with no deep invasion, and it is associated with a 5-year survival rate over 90% [3], [1]. If any polyp found is characterized as a likely adenoma, its removal should be considered to confirm the diagnosis, to set its histological stage and to confirm its complete removal, giving clinicians clues to determine the need and timing of the next colonoscopy [4].

Though colonoscopy is the gold standard for colon screening, other alternatives, such as CT colonography [5] or wireless capsule endoscopy (WCE) [6], are also used to search for polyps. They are less invasive to patients and do not present perforation risk. Though, as colonoscopy, they require bowel preparation. Nevertheless in these cases, if a polyp is found, a colonoscopy must be considered to remove the suspicious lesion. These alternatives have specific limitations that may affect the outcome of the screening. For instance, CT colonography has a low small lesions (5 mm or less) detection rate due to resolution constraints [7] and it implies using ionising radiation. WCE allows to detect all kind of lesions but their observation depends on whether they are recorded during the progress of the camera through the gastrointestinal tract or not. Moreover, its diagnostic yield is highly dependent on the

cleanliness of the colon (whereas colonoscopy has some in-situ lavage capabilities). Last but not least, the analysis of the information provided by WCE can be highly time-consuming, as the recorded videos can last up to 8 hours [8].

Colonoscopy presents some drawbacks, polyp miss-rate being the most important among these. Colonoscopy rarely misses polyps bigger than 10 mm, but the miss-rate increases significantly with smaller sized and/or flat polyps [9], [10]. It has also to be noted that colonoscopies are seldom recorded, so a new procedure must be performed to revisit explored areas.

The outcome of the colonoscopy exploration depends on: 1) bowel preparation [11]; 2) specific choice of endoscope and video processor, affecting image quality and preventing the use of certain image enhancing tools; 3) clinicians' skills, as both endoscopist's experience and his/her actual concentration during the intervention may influence the degree of procedure completion (reaching the cecum or not) and the percentage of the colon that has been explored [12], [13] and 4) patient-specific issues, as due to colon movements and the appearance of folds and angulations during the exploration, some parts of the colon which may potentially present polyps may not be reached [9]. Moreover, patients' personal and family history can increase the risk of having a polyp and, in this case, the exploration should be even more thorough.

B. Technical strategies to improve polyp detection rate

Apart from the continuous improvement of clinicians' skills through training programs and practice [14], technical efforts are being undertaken to improve colonoscopy's outcome. We clustered them into two groups: improvement of devices and the development of computational support systems.

Amongst the device improvements, the following should be highlighted: 1) increase in image resolution and, consequently, textural information; 2) the use of wide-angle cameras showing more colon wall surface; 3) the development of zooming and magnification techniques [15] and 4) the development of new imaging methodologies such as autofluorescence imaging [16] or virtual chromoendoscopy (Olympus' Narrow Band Imaging [17], Fujinon's FICE [18] or Pentax's i-Scan [19]). This last group of techniques modify how the scene is observed by improving the contrast of endoluminal scene elements, which may help in lesion detection and also with in-vivo lesion diagnosis due to the enhanced visualization of lesion tissues [20]. These advances have fostered the cooperation between clinicians and computer scientists in the development and validation of computer-aided support systems for colonoscopy, aimed to help clinicians in all stages of CRC diagnosis. A significant part of this effort has been focused on computer assisted polyp detection. As it is indicated in [21], cooperation between technologists and clinicians is essential to develop clinically useful solutions, with both these groups understanding challenges and limitations in their respective domains.

Automatic polyp detection in colonoscopy videos has been an active research topic during the last 20 years and several approaches have been proposed. We present a review of the most relevant methods in Section II but, to the best of our knowledge, none of them has been adopted for a routine patient treatment. There might be several reasons behind this. First of all, in order for a given method to be clinically useful, it has to meet real time constraints; e.g. for videos acquired at 25 frames per second (fps) the maximum time available to process each image frame should be under 40ms. Secondly, some of them are built from a theoretical model of a polyp appearance [14], [22] and therefore limited to only certain polyp morphologies, which may not translate to the actual scene where polyp appearance varies greatly. Thirdly, the majority of methods are mainly focused on the polyps and they do not consider the presence of other elements such as folds, blood vessels or the lumen that can affect methods' performance [14]. Last but not least, some of these methods have been only trained and tested on selected good quality still image frames. The lack of temporal coherence and the great variability in polyp appearance due to camera progression and visibility conditions might impact their performance in the full sequences analysis, as they might cause instability in their response against similar stimuli.

Computational methods also have to deal with additional colonoscopy-specific challenges. For instance, they should consider the impact of image artifacts generated due to scene illumination (specular highlights, overexposed regions) or to specific configuration of the videoprocessor attached to the colonoscope, which might overlay information over the scene view. These artifacts, apart from altering the view of the scene, might not be stable within consecutive frames and therefore methods should both compensate their impact on the individual frame polyp detection and tracking in the full sequence analysis. Additionally, though an effort is made to ensure an adequate bowel preparation, some particles may still appear which, in some cases, could lead to false detections when isolated or to occlusion leading to miss detection or localization errors. As mentioned before, these methods have to cope with a great degree of variability in polyp appearance which depends on illumination conditions, camera position and on clinician skills when progressing through the colon. Finally, available methods have been typically validated on small and restricted databases, under specific endoscope device conditions (brand and resolution), in some cases even covering only one specific polyp type, shape or morphology hindering their actual performance in a more generic setting.

C. Motivation of the comparison study

Unfortunately, the lack of a common validation framework, which is a frequent problem in medical and endoscopy image analysis [21], has limited the effectiveness of the comparison between existing approaches, making it difficult to determine which of them could have actual advantage in clinical use. To cope with this, efforts have been made on publishing fully annotated databases [14], [22] and on organizing challenges as part of international conferences (ISBI, MICCAI), which offer a basis to discuss validation strategies.

Considering this and taking inspiration from recent works on quantitative comparative methods' analysis in areas such as laparoscopic 3D Surface Reconstruction [23] or liver segmentation [24], we present in this paper a complete validation

study of polyp detection methods performed as part of the 2015 MICCAI sub-challenge on Automatic Polyp Detection. This sub-challenge was organized jointly by three research teams: 1) Computer Vision Center/Universitat Autònoma de Barcelona and Hospital Clinic from Barcelona, Spain (CVC-CLINIC); 2) ETIS Lab (ENSEA/CNRS/University of Cergy-Pontoise) and Lariboisière Hospital-APHP at Paris, France (ETIS-LARIB), and 3) Arizona State University and Mayo Clinic, USA (ASU-Mayo).

The objective of this paper is to present a comparative study of polyp detection methods under a newly proposed validation framework. This validation framework was firstly introduced as part of MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy and we present in this paper the results of the mentioned sub-challenge. Beyond this, we also propose additional experiments to assess even more in-depth the performance of an automatic polyp detection method. These new experiments are focused on exploring the actual clinical applicability of a given method by assessing up to what extent they are affected by some of the technical and clinical challenges reported in the literature or whether they incorporate temporal coherence features or not. Finally we also go beyond the individual analysis of methods and propose combination strategies in order to study whether a combination method may lead to improved individual performance.

The remainder of the paper is structured as follows: In Section II we present the methods proposed by each of the participating teams in the challenge, including them in the context of existing published methods. In Section III we describe the complete validation framework. Results from the comparative study are presented in Section IV. Section V provides an in-depth analysis of the results and discusses some topics related to challenge organization. Finally, the concluding remarks are drawn in Section VI.

II. Automatic Polyp Detection Methods

A. Historical review of computational polyp detection methods

After analyzing approaches reported in the literature, we propose to cluster methods into three groups: 1) **hand-crafted**; 2) **end-to-end learning** and 3) **hybrid** approaches. This taxonomy represents the different historical trends of polyp detection methods, as in early 2000s, the majority of the methods used a given texture descriptor to guide a classification method but, subsequently, some researchers decided to go for hand-crafted features, aiming at a real time implementation. As technology evolved and the computational capabilities increased, techniques such as neural networks that were developed in the past and abandoned due to excessive computational cost have now resurfaced.

Regarding hand-crafted methods, the majority are based on exploiting low-level image processing methods to obtain candidate polyp boundaries (using Hessian filters in the work of Iwahori et al. [25], intensity valleys in the work of Bernal et al. [14] or Hough transform in the work of Silva et al, [26]) and then use resulting information to define cues unique to polyps. For instance, the work of Zhu et al. [27] analyzes curvatures of detected boundaries whereas the method of Kang et al. [28] is focused on searching ellipsoidal shapes typically associated with polyps. Finally, the method of Hwang et al. [29] combines curvature analysis and shape fitting in their strategy.

Concerning end-to-end learning, texture and color information were formerly used as descriptors such as in the work of Karkanis et al. [30] which proposed the use of color wavelets, the work of Ameling et al. [31] that exploits the use of co-ocurrence matrices or the work of Gross et al. [32], which proposed the use of local binary patterns. Active learning methodologies have also been introduced as in the work of Angermann et al. [33] to reinforce the tradeoff between performance and computation time. Some of the most recent methods use deep learning tools to aid in polyp detection tasks, as in the work of Park et al. [34] or in the work of Ribeiro et al. [35]. In these very recent developments, differences among methods are based on the selection of a specific network architecture and databases used for training.

Finally, there are several hybrid methods which combine both methodologies for polyp detection, such as in the works of Tajbaksh et al. [22], which combines edge detection and feature extraction to boost detection accuracy, the work of Bae et al. [36], that propose a system based on imbalanced learning and discriminative feature learning; the work of Silva et al. [26], which uses hand-crafted features to filter non-informative image regions and the work of Ševo et al. [37], which combines edge density and convolutional networks.

As mentioned in Section I, the great majority of the methods are tested on private databases though we can observe that more recent publications such as the work of Park et al. [34] or the work of Ribeiro et al. [35] have started to use publicly available databases such as the ones used in the MICCAI 2015 Sub-challenge on Automatic Polyp Detection. Related to this, apart from new proposals, some of the referenced methods have been adopted by participants, such as the works of Bernal et al. [14], Silva et al. [26] or the work of Tajbaksh et al. [22]. We provide in the next subsection a brief description of participating methods highlighting their most relevant contributions to the field. We grouped the methods following the taxonomy defined earlier in this subsection.

B. MICCAI 2015 Polyp Detection Sub-challenge methods

1) Hand-crafted features:

- **CVC-CLINIC:** This method [14] is based on a model of appearance considering polyps as protruding surfaces, being their boundaries defined from intensity valleys detection. Their proposal includes a pre-preprocessing stage to mitigate the impact of other valley-rich structures (blood vessels, specular highlights). To build final energy maps highlighting polyp presence, four different constraints (continuity, completeness, concavity, and robustness against spurious structures) are imposed to candidate boundaries to differentiate polyps from other structures.

2) End-to-end learning:

- **CUMED:** The architecture of the proposed network contains two sections including a downsampling path and an upsampling path [38]. The former contains convolutional

TABLE I
SUMMARY OF INFORMATION FROM THE TEAMS THAT TOOK PART IN MICCAI 2015 CHALLENGE ON AUTOMATIC POLYP DETECTION.

Team acronym	Full team details	Methodology	Published	Still-frame analysis	Video analysis	Training (seconds)	Testing (seconds)	System tested
ASU	Arizona State University (USA)	Hybrid	Yes [22]	No	Yes	N/A	2.7	2.4 GHz Intel quad core processor and an NVIDIA GeForce GTX 760 video card
CUMED	Department of Computer Science and Engineering, Chinese University of Hong Kong (China)	End-to-end learning (CNNs)	No	Yes	Yes	10800	0.2	A standard PC with a 2.50 GHz Intel(R) Xeon(R) E5-1620 CPU and a NVIDIA GeForce GTX Titan X GPU
CVC-CLINIC	Computer Vision Center and Universitat Autònoma de Barcelona (Spain)	Hand-crafted	Yes [14]	Yes	Yes	N/A	10	Intel core i7-4790 at 3.6GHz
ETIS-LARIB	ETIS, ENSEA, University of Cergy-Pontoise, CNRS, Cergy (France)	Hybrid	Yes [26]	Yes	No	196	2.14	Intel i5 4200U 2.30 GHz
OUS	Oslo University Hospital, OUS Norway, University of Oslo (Norway)	End-to-end learning (CNNs)	No	Yes	Yes	86400	5	Intel i5, 4 cores at 2.8 GHz, 4 GB RAM. Graphic card with 4 GB memory used for training
PLS	Polyp Localize and Spot Team, Media Performance Group, Simula Research Laboratory and University of Oslo (Norway)	Hybrid	No	Yes	Yes	0.33 per image	0.145	2 Intel(R) Xeon(R) CPU E5-2650 at 2.00GHz CPU, 64 GB of RAM, NVIDIA Corporation GK110, GeForce GTX TITAN
SNU	Seoul National University, Seoul (South Korea)	End-to-end learning (CNNs)	No	Yes	Yes	360	0.8-1	NVIDIA TITAN X GPU
UNS-UCLAN	School of Engineering, University of Central Lancashire, Preston (UK) and University of Nice-Sophia Antipolis, Nice (France)	End-to-end learning (CNNs)	No	Yes	No	18000	5	i7-5930K @ 3.5GHz (6 cores), 64 GB RAM, NVIDIA GeForce GTX TITAN X

and max-pooling layers while the latter contains convolutional and upsampling layers, increasing the resolutions of feature maps and output prediction masks. To alleviate the problem of vanishing gradients and encourage the back-propagation of gradient flow in deep neural networks, the auxiliary classifiers are injected to train the network. Furthermore, they can serve as regularization to reduce over-fitting and improve the discriminative capability of features in intermediate layers [39], [40].
The classification layer, after fusing multi-level contextual information, produces the detection results. Network training is formulated as a pixel-wise classification problem with respect to ground-truth masks. The highlight of this approach is that it explores multi-level feature representations with fully CNNs in an end-to-end way, taking an image as input and directly providing the score map. In addition, feature-rich hierarchies from a large scale auxiliary dataset are transferred into the model to reduce over-fitting and further boost detection performance [41].

- **UNS-UCLAN:** This method, inspired by reported works [42], [43], [44], uses three CNNs trained at different image scales, namely 1, 0.5, and 0.25, of the original training images. For all the scales the CNNs use the same architecture, but they are trained independently on the RGB images at their corresponding scale. After this initial training phase, the last fully connected part of each CNN is removed and the outputs from the 'convolutional part' of all the three networks are fed as input to a single Multi-Layer Perceptron (MLP) network. This additional network is trained independently from the three CNNs. In this approach CNNs are used as feature extraction engines operating at different spatial scales, and the MLP performs the classification based on these features.
The method's output is the polyp incidence probability map, which is then processed to locate dominant probability peaks, as peaks locations and probability values are returned as the final output of the system. The training was performed exclusively on the CVC-CLINIC database.
- **OUS:** This method is based on the popular AlexNet model [44] for CNNs and its slight modification CaffeNet, which is pre-trained on the ILSVRC 2012 [45] dataset. Computations are achieved using the Caffe library [46]. The original model is modified to take input patches of size 96×96, and the kernel size of the two first pooling layers is decreased from 3 to 2, while the last pooling layer is removed. The output layer is modified to give two outputs, polyp or non-polyp. In order to increase the training examples, data augmentation is performed in the form of random mirroring, rotation, up- and down-scaling, cropping, and brightness adjustment. Final polyp presence or absence was determined by using a sliding-window strategy, with three scalings for still frame analysis and two for full video sequence analysis.
- **SNU:** This methodology proposes a two-step approach: detection and localization. For both steps, CNNs were used. Starting from GoogleNet (pre-trained on the ImageNet dataset), a CNN fine-tuning was performed. Input image is resized to 224x224 pixels prior training and data augmentation (rotation and scaling) is also performed. Training set images are augmented by using several degrees of random rotation and scaling. Detection is considered as a simple binary classification task whereas, for localization, CNN are applied on polyp-positive images

which are then segmented into a uniform-sized 8x8 grid (64 grids per image). Then, for each image, one grid is overlaid in black and then CNNs are applied thereafter to perform the binary classification task. The 64 overlaid grid images are then sorted by classification score to calculate final polyps' position.

3) Hybrid approaches:

- **PLS:** The proposed full localization scheme consists of two parts, detection and localization. Regarding detection, two sets of images, one containing polyps, and the other without polyps, are used for training. Global image features [47] are used as they are easy and fast to calculate. Based on similarity scores between input frame and training ones and results ranks, the detection subsystem decides in real-time to which class (polyp or no polyp) the input frame belongs to.
  The localization scheme is implemented as a sequence of preprocessing filters (RGB to YCbCr color space conversion, removal of borders and sub-images, flare masking and low-pass filtering) and uses the polyp's physical shape to find its exact position, approximating polyps by elliptical shape regions presenting local features that differentiate them from surrounding tissues. The final decision regarding polyp location is taken by means of the maximum values in the energy map computed using the elliptical shape of the polyp's usual appearance. Finally, the method outputs four possible locations per frame.
- **ETIS-LARIB :** This method [26] is inspired by the psycho-visual methodology used by clinicians when performing an endoscopic examination. First, a detection of the Regions of Interests (ROI) that may contain a polyp is performed using shape and size image features. This first pre-selection allows a first and fast scanning of the image. Due to being circular/elliptical shapes associated to polyps, a Hough transform was used for this first filtering stage. Once ROIs are detected, a second analysis, based on texture is achieved in order to remove those ROIs with no actual polyp content. To achieve this, an ad-hoc classifier based on a boosting-based learning process using texture features computed from co-ocurrence matrices (standard Haralick features) is proposed.
- **ASU:** This method [22] consists of two stages. In the first stage, a set of polyp candidates is generated using geometric features. Specifically, given a colonoscopy frame, a crude set of edge pixels is first obtained. This edge map is then refined using a classification and feature extraction scheme [48]. The goal of the edge classification scheme is to remove as many non-polyp boundary pixels as possible from the initial edge map. The geometry of the retained edges is then used in a voting scheme that localizes polyps candidates as objects with curved boundaries in the refined edge maps. The voting scheme further estimates a bounding box for each generated candidate based on the generated voting map. In the second stage, an ensemble of CNNs -each of them specialized in one type of features- is applied to each candidate bounding box [49]. Finally, the outputs of the CNNs are averaged to generate a confidence score for a given polyp candidate.

Table I shows a summary of the different methods participating at MICCAI 2015 Challenge on Automatic Polyp Detection. As each method was tested under different conditions, computation times are given to complete the information on the training and testing processes.

III. Validation study

We introduce in this section the complete validation study proposed to assess and compare the performance of different polyp detection methods.

A. Definitions and general performance metrics

We define **Polyp detection** as the capability of a given method to determine polyp presence in a colonoscopy frame (**Polyp presence detection**) and, once this is determined, it is able to provide the location of the polyp within the image (**Polyp localization**). Consequently, a good polyp detection method should select images (video frames) containing polyps and ignore all others and it should indicate the position of all polyps present in an image. There are some terms defined next which are key to set performance metrics. As we deal with images from real patients examinations, we will find two different cases: images with polyps and images without polyps.

In the first case, if detection output is within the polyp, the method is said to be providing a **True Positive (TP)** or correct alarm. It has to be noted that only one TP will be considered per polyp, no matter how many detections fall within the polyp. Any detection that falls outside the polyp is considered a **False Positive (FP)** or false alarm. The absence of alarm in images with a polyp is considered a **False Negative (FN)**, counting one per each polyp in the image that has not been detected. Regarding images without polyps, we define as a **True Negative (TN)** whenever the method does not provide any output for this particular image. Any detection provided for frames without a polyp counts as a **False Positive (FP)**. Considering these definitions, we propose the use of the frame-based performance metrics presented in Table II.

B. Databases

Three different databases are used in the context of the validation study presented in this paper. Two publicly available databases were proposed for still frame analysis, **CVC-CLINIC** and **ETIS-LARIB**. CVC-CLINIC [14] contains 612

TABLE II
PERFORMANCE METRICS FOR POLYP DETECTION.

Metric	Abbreviation	Calculation
Precision	Prec	$Prec = \frac{TP}{TP+FP}$
Recall	Rec	$Rec = \frac{TP}{TP+FN}$
Specificity	Spec	$Spec = \frac{TN}{FP+FN}$
F1-measure	F1	$F1 = \frac{2 \times Prec \times Rec}{Prec+Rec}$
F2-measure	F2	$F2 = \frac{5 \times Prec \times Rec}{4 \times Prec+Rec}$

TABLE III
CONTENT OF ASU-MAYO CLINIC COLONOSCOPY VIDEO © DATABASE.

Training database						Testing database					
Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]
1	22	0/682 [712 × 480]	11	10	245/324 [1920 × 1080]	1	19	0/599 [712 × 480]	11	15	338/452 [1920 × 1080]
2	27	0/838 [712 × 480]	12	30	910/910 [1920 × 1080]	2	20	0/625 [712 × 480]	12	4	134/0134 [1920 × 1080]
3	25	0/769 [712 × 480]	13	17	374/519 [1920 × 1080]	3	20	0/628 [712 × 480]	13	10	312/312 [1920 × 1080]
4	23	0/712 [712 × 480]	14	16	391/501 [856 × 480]	4	20	0/607 [712 × 480]	14	60	0/1815 [712 × 480]
5	61	0/1843 [712 × 480]	15	36	1106/1200 [856 × 480]	5	30	693/918 [856 × 480]	15	59	0/1795 [712 × 480]
6	64	0/1925 [712 × 480]	16	11	209/339 [1920 × 1080]	6	40	1218/1218 [856 × 480]	16	54	0/1627 [712 × 480]
7	51	0/1550 [712 × 480]	17	13	234/418 [856 × 480]	7	18	445/555 [712 × 480]	17	60	0/1807 [712 × 480]
8	58	0/1740 [712 × 480]	18	18	189/259 [1920 × 1080]	8	14	335/446 [856 × 480]	18	61	0/1835 [712 × 480]
9	60	0/1802 [712 × 480]	19	20	235/616 [1920 × 1080]	9	13	290/396 [1920 × 1080]			
10	54	0/1639 [712 × 480]	20	13	385/410 [856 × 480]	10	60	548/1805 [1920 × 1080]			

Standard Definition (SD) frames and comprises 31 different polyps from 31 sequences. ETIS-LARIB database contains 196 High Definition (HD) frames and comprises 44 different polyps from 34 sequences. More details on these databases are presented in Table IV. It has to be noted that all images contain at least a polyp; both databases were built to cover as many different polyp appearances as possible. Ground truth consisting of a polyp mask was generated using the same procedure for both databases: Images were annotated by expert videoendoscopists from the corresponding associated clinical institution. These experts (one per hospital) were asked to outline the boundaries of any polyps present in the image. These boundaries are used to generate a binary mask representing the actual polyp area within the image, also to be used for validation purposes. Examples from these two databases are shown in the first two columns of Fig. 1.

The **ASU-Mayo Clinic Colonoscopy Video © Database** [22] comprises a set of short and long colonoscopy videos, collected at the Department of Gastroenterology at Mayo Clinic, Arizona. This database consists of 38 different, fully annotated videos. The videos were selected to display maximum variation in colonoscopy procedures including different resolutions and examination strategies (careful vs. fast inspection) and also include frames containing biopsy instruments or device information. Ground truth consisting of binary masks (polyp frames) and black frames (non-polyp frames) were created by volunteer students at Arizona State University and have been reviewed and corrected by a trained expert. Table III outlines information about the videos in that database, including for each video duration in seconds (Length), number of frames with polyps and the total number of frames (Polyp/Total) and the image resolution (Res). An example from this database is shown in the third column of Fig. 1.

TABLE IV
SUMMARY OF CONTENT OF STILL FRAME VALIDATION DATABASES. SD STANDS FOR STANDARD DEFINITION, HD STANDS FOR HIGH DEFINITION.

Database	Purpose	Institution	Content	Device
CVC-CLINIC	Training	Hospital Clinic, Barcelona, Spain	612 SD frames (388 × 284) from 31 sequences	Olympus Q160AL and Q165L, Exera II videoprocessor
ETIS-LARIB	Testing	Lariboisière Hospital, Paris, France	196 HD frames (1225 × 966) from 34 sequences	Pentax 90i series, EPKi 7000 videoprocessor

C. Statistical analysis

In order to account for statistically significant differences in performance between methods, we propose first to perform a Saphiro-Wilk test to find out whether the available data follows a normal distribution or not. In the first case (normal distribution) statistically significant differences across methods will be assessed using an analysis of variance (ANOVA) to detect differences regarding proposed metrics. In the second case (no normal distribution), the Kruskal-Wallis test will be used. All tests are done at a confidence level $1 - \alpha = 0.95$.

Considering the scope of the analysis presented in the paper, the metric that will be used to compare different methods will be F1-score, as it presents a balance between missed polyps and false alarms. We perform a statistical study of this metric only in videos with polyp (and potentially non-polyp) frames, as the number of samples in the still-frame analysis is not big enough to provide with statistically relevant conclusions and the analysis in videos with no polyps would cause the F1 score to be zero for all methods. We also perform statistical analysis of detection latency but, for the sake of a proper statistical comparison, this analysis is only done for those teams which detect the polyp in all sequences.

D. MICCAI 2015 Sub-challenge validation study

Two different scenarios were presented to the participants of the challenge: (i) still frame analysis and (ii) full video analysis. In the following, we present specific information of

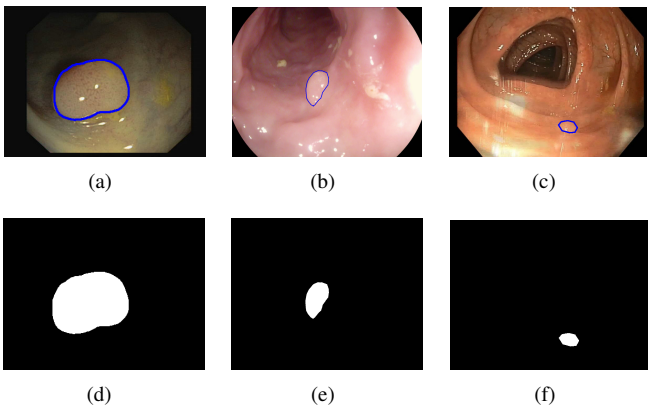


Fig. 1. Illustration of the content of the CVC-CLINIC (first column), ETIS-LARIB (second column) and ASU-Mayo Clinic (third column) databases. The first column shows the original images with the corresponding reference polyp contour shown as a blue line and the second contains binary masks representing the ground truth

the two presented scenarios, including validation databases and performance metrics used in each of them.

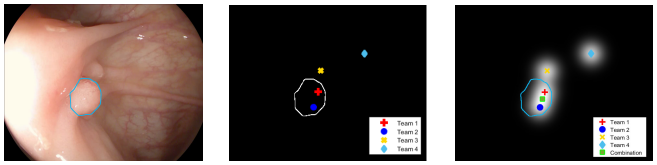
1) Still frame analysis: The objective of this analysis was to explore localization capabilities of a polyp detection method. We aim to test how different methods perform in challenging high-definition (HD), high-quality images showing great variability in polyp appearances. In this case, each image contains at least one polyp and images have been selected in order to have shots in which polyp appearance can be mistaken with other elements of the scene (folds, vessels).

Two different databases were used in this study: **CVC-CLINIC** is used for the **training** stage whereas **ETIS-LARIB** is used during the **testing** stage. Participating methods are compared using performance metrics exposed in Table II. Additionally, in case a given method provides confidence values a Precision-Recall curve is also provided otherwise the operating point will represent its performance.

2) Video analysis: In this second scenario we aim to explore full polyp detection capabilities (localization and presence detection) of a given method in full sequences from actual colonoscopy procedures. In this case, polyp detection methods have to deal, apart from appearance variability, with potential polyp presence or absence in each image and, moreover, with variability in image quality (blurring, bowel preparation). Additionally, the videos in the second scenario may contain images with extra-endoluminal elements such as device information or surgical instruments. We also have to consider that, as in real procedures, nor all the sequences or all the frames contain a polyp.

The **ASU-Mayo Clinic Colonoscopy Video © Database** was used in this experiment. Apart from using common performance metrics exposed in Table II, we proposed an additional performance metric to assess whether how fast a given detection method reacts to polyp presence. In this context, **Detection Latency (DL):** $DL = first\_detection - first\_appearance$ represents the delay in frames between the first appearance of the polyp in the video sequence ($first\_appearance$) and the first actual detection of the polyp by a method ($first\_detection$). Considering this, a clinically useful support system should have a DL close to zero. Finally we also provide with Receiver Operating Curves (ROC), though again, each method's representation depends on whether they provide detections' confidence values or not.

From a general organization perspective, all teams taking part in the challenge were to use the same data for both their training and testing stages. Participants were provided with labelled training data on June 15th whereas unlabelled testing data (still frames and full sequences) was released on July 24^{th}. In order to take part in the challenge, each participating team was asked to provide a unique CSV file for the analysis of the ETIS-LARIB database and/or one CSV file for each of the 18 testing videos in the ASU-Mayo database, depending on the sub-category the team would take part in. Each row in the CSV file represents a detection candidate region. Additionally, teams could also provide a confidence value (value between 0 and 1) for the performance curves drawing purposes, though this was not mandatory. Finally, though 8 different teams took part in the challenge, not all of them participated in all categories. ASU did not take part in the still-frame analysis sub-category whereas ETIS-LARIB and UNS-UCLAN did not take part in the video analysis one.



(a) (b) (c)

Fig. 2. Synthetical examples of different ways to perform a combination of methods: (a) original image; (b) result of combination by union, and (c) result of combination by saliency map creation. Outputs from different teams are represented by different colors and shapes. In all images, the contour of the polyp is represented as a blue curve.

E. Additional validation experiments

1) Combination of methods: In this study we propose to go beyond the analysis of individual methods by providing quantitative elements on how potential combinations of some of the presented approaches may lead to an improved performance. Inspired by [50], we have studied two options of fusing the methods, namely: 1) **combination by union** and 2) **saliency map** creation.

The first one consists of adding, for a particular frame, the outputs from all submitted methods. Saliency maps creation proposes a combination of the output of the methods in order to generate heat maps which aim to represent those areas in the image where most of the methods coincide in their decision regarding polyp location, following the methodology proposed in [14]. We show in Fig. 2 a graphical comparison between both strategies.

In this case, we treat the output of each method as a 'fixation' or vote, and we create saliency maps from this set of discrete fixations/votes. These fixation points are interpolated by a Gaussian function to build up the final saliency map for a given image as follows:

$$s\left(x,y\right) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2\pi\sigma_s^2}.exp\left(-\frac{\left(x-x_n^f\right)^2+\left(y-y_n^f\right)^2}{2\sigma_s^2}\right), \quad (1)$$

where: x and y denote, respectively, the horizontal and vertical positions of a given image pixel; x_n^f and y_n^f represent the horizontal and vertical coordinates of a detection point (fixation); N indicates the total number of detected points and finally σ_s denotes the standard deviation of the Gaussian function, determined as proposed in [14]. We determine the location of the global maximum of the saliency map as the final output of the combination of methods for a particular frame.

Two versions of saliency map creation have been implemented: (**saliency by union**) calculates the saliency maps for each frame considering all the methods that have provided any output whereas (**saliency voting**) only calculates the saliency maps if the majority of the teams in the studied combination provide an output for the specific frame.

We provide results for each combination strategy in the two challenge scenarios (still frame analysis and video analysis) using the same frame-based performance metrics.

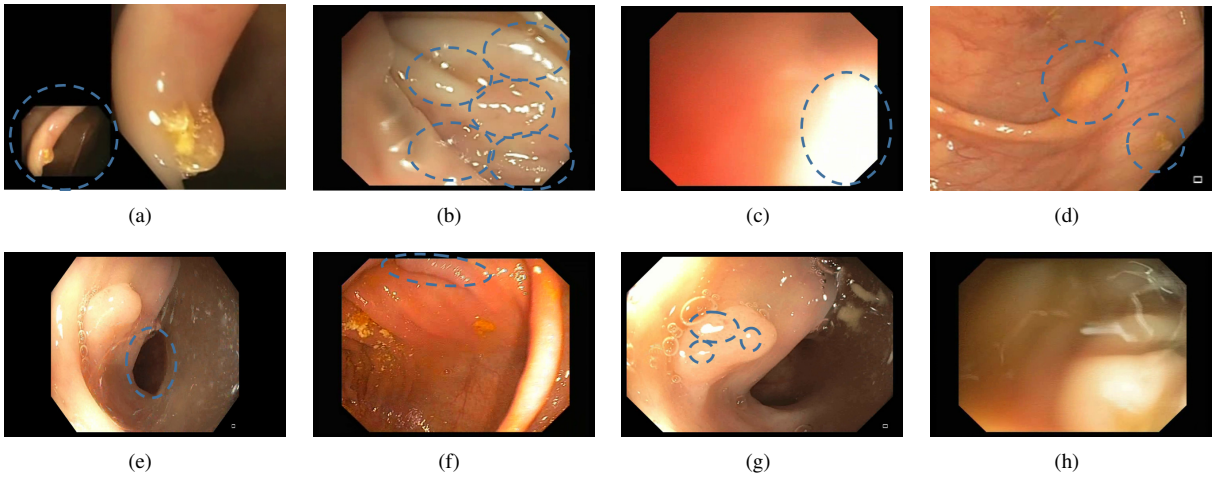


Fig. 3. Examples of the 8 technical and clinical challenges selected for the study: (a) Presence of overlay information; (b) High presence of specular highlights; (c) Overexposed regions; (d) Intestinal content; (e) Luminal region; (f) Polyp cannot be seen completely in the image; (g) Specular highlights within the polyp and (h) Impact of low visibility quality

2) Impact of image challenges on method's performance: This experiment aims to study the impact of some of the technical and clinical challenges reported in Section I over the performance of a polyp detection method. In order to study this, we proposed clinicians and computer scientists from the contributing teams to define the main image challenges present in colonoscopy frames that were to be studied. The following ones were selected: 1) Presence of overlay information in images (including patient information and camera shots); 2) Presence of specular highlights; 3) Appearance of overexposed regions; 4) Occurrence of intestinal content (fecal particles, bubbles); 5) Presence of the luminal region; 6) Lack of visibility of the whole polyp in the image; 7) Presence of specular highlights within the polyp region and 8) Images with low visibility (due to blurring or excessive intestinal content). Fig. 3 shows examples of each of the challenges.

A graphical user interface was built for experts to label individually each frame from the testing videos of **ASU-Mayo Clinic Colonoscopy Video © Database** according to the mentioned image challenges. For the sake of statistical representativeness of the results, we did not perform the same experiment for **ETIS-LARIB** database due to its smaller size.As some of them may lead to subjective interpretations we collected three different annotations per frame and the final decision of a frame for each challenge was taken by majority voting from the three experts. We present statistics about the presence of the different image challenges in Table V.

We can observe how roughly half of the frames contain a high number of specular highlights, some degree of intestinal content and overexposed regions. Regarding polyp frames, which equate to a 25% (4313) of the frames, we can observe that about a 30% of them (1360) do not show completely the polyp and that nearly all of them (3959) present specularities within the polyp region. Finally, it is interesting to mention that more than a 70% of the images were considered of low visibility quality, which indicates how the methods are tested in clear challenging conditions.

Once we have final annotations, we broke down the methods performance analysis into two groups: frames with polyps and frames without polyps. For the first case, we analyze differences between performance for frames with and without a specific image challenge regarding Precision, Recall and F1-score whereas for the second the same kind of analysis was done regarding Specificity score.

3) Impact of polyp morphology on methods' performance: This experiment assesses whether methods' performance depends on the polyp morphology. This analysis examines if the methods perform differently for polyps with different associated morphological type. Such analysis could be useful to check whether existing methods are able to cope with different morphologies as well as determining which method to choose if a given morphology is predicted before the examination. In order to study these potential differences in performance, we propose to categorize each of the polyps that appear in the testing databases using the Paris classification criteria [51]. We show graphical examples of each type in Fig. 4.

To account for differences in performance related to polyp morphology we will use Precision, Recall and F1 scores as defined in Table II. We also study differences in latency score for the case of video sequences analysis.

4) Temporal coherence on method's response: One capability that a computational method should have when dealing with video analysis is temporal stability in its response. That is, if a given method detects a polyp in a given frame and considering normal camera movement, its output for the following frame

TABLE V
BREAKDOWN OF CLINICAL AND TECHNICAL CHALLENGES WITHIN ASU-MAYO TEST DATABASE.

Challenge	Number of frames	Challenge	Number of frames
Overlay information	517[02.94%]	Massive specular highlights presence	8638[49.15%]
Overexposed regions	8270[47.05%]	Intestinal content	10013[56.97%]
Visible luminal region	4963[28.24%]	Images showing an incomplete polyp	3286[18.69%]
Specular highlights within polyp	93[19.88%]	Low visibility	12788[72.67%]

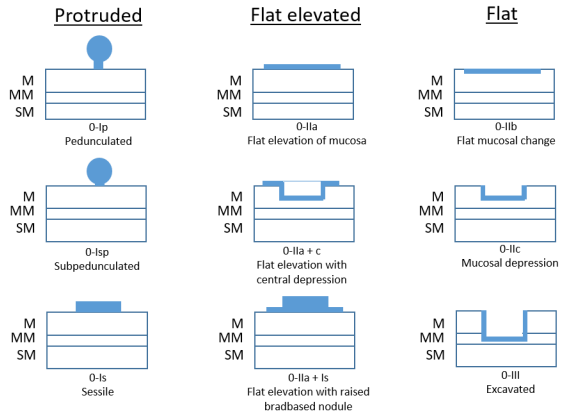


Fig. 4. Graphical representation of Paris classification of endoscopic polyps. M stands for mucosa, MM for muscularis mucosa and SM for submocusa.

should provide a relevant detection. As we can observe in the example shown in Fig. 5, none of the methods presented in the challenge incorporated per se temporal stability capabilities in their methodologies but we consider that it is important to assess up to what extent they provide this kind of stability. Moreover, and as a consequence of this stable temporal output, the method should provide with correct detection in the majority of the frames in which the polyp appears.

In order to study this we perform two evaluations. Regarding detection stability in consecutive frames, for each testing video from **ASU-Mayo Clinic Colonoscopy Video © Database** that contained a polyp, we extracted the pairs of consecutive frames containing a polyp. We analysed methods' output for each pair of consecutive polyp frames and calculated as metric the percentage of these pairs in which the method provided correct output - detection inside the polyp mask - for both frames. With respect of overall detection stability in a sequence, we study Recall scores over the different sequences, analyzing mean and standard deviation values to account for intra and inter-sequence stability on detection performance.

5) Analysis of the direct output of the methods: As mentioned in Section III-D, participating teams were only asked to provide CSV files indicating detection output for the testing frames (x and y position). This file is created from the output of the different methods and we propose here to analyze this actual output. As a first study, we asked the teams participating in the still frame analysis challenge sub-category to provide their actual output for each frame of **ETIS-LARIB** database.

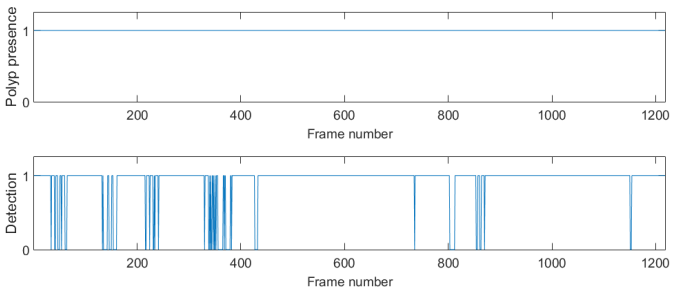


Fig. 5. Example of non-temporal coherence of polyp detection methods. The example represents the performance of the CVC-CLINIC method for the testing video 6 of ASU-Mayo Clinic Colonoscopy Video © Database. Image at the top shows ground polyp presence per frame (1 is polyp, 0 is no polyp) whereas bottom image shows detection score (1 correct, 0 no correct).

In this context, we foresee the output of a method to be interpreted as a likelihood or heat map, in which brighter (hotter) areas of the image represent parts of the image more likely to contain a polyp. By analyzing these maps, we could observe up to what extent method's attention is only focused on the polyp. To measure this, we propose Concentration Ratio (CR) to compare these maps as proposed in [14]; CR measures, for each frame, the rate of total energy in the image (calculated as the sum of the each pixel's value from the energy map image) falling within the polyp. High CR values are interpreted as a method actually focusing on the polyp, being lower values related to sparser energy maps.

IV. RESULTS

In this section, we present the performance achieved by each method in the several experimental studies proposed in the paper, including those part of the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy.

A. MICCAI 2015 Challenge Results

We present main still frame analysis results in both Fig. 6 (a) and in Table V. CUMED offers the best performance in all metrics evaluated, being the team which detected the most polyps (144) frames along with providing the lowest number of false alarms (55). The comparison between the performance of CNN-based approaches shows the importance of specific network configuration, as relevant differences in both number of detected polyp frames and false alarms can be observed - for instance, the number of detected polyp frames falls into a range between 131 (OUS) and 20 (SNU) -. Finally we can observe a performance gap between end-to-end learning and hybrid/hand-crafted methods, which provide less correct detections and significantly more false alarms. It has to be noted that as PLS provides four locations per image, the number of false alarms for this method is inherently higher than for other methods.

Considering full video analysis, the study shows superior performance by CNN-based methods -see Table VII and in Fig. 6 (b)- , with CUMED being the method providing a higher number of polyp frames detected (3081). In this case, it has to be noted that CUMED does not outperform all other methods

TABLE VI
SUMMARY OF STILL FRAME ANALYSIS RESULTS.

	TP	FP	FN	Prec	Rec	F1	F2
Individual team analysis							
CUMED	144	55	64	72.3	69.2	70.7	69.8
CVC-CLINIC	102	920	106	10.0	49.0	16.5	27.5
ETIS-LARIB	103	1373	105	6.9	49.5	12.2	22.3
OUS	131	57	77	69.7	63.0	66.1	64.2
PLS	119	630	89	15.8	57.2	24.9	37.6
SNU	20	176	188	10.2	9.6	9.9	9.7
UNS-UCLAN	110	226	98	32.73	52.8	40.4	47.1
Combination of teams analysis							
Best concatenation result (all teams)	188	3410	20	5.2	90.4	9.9	21.2
Best saliency result (CUMED, OUS)	159	38	49	80.7	76.4	78.5	77.2
Best saliency voting (CUMED, OUS)	159	38	49	80.7	76.4	78.5	77.2

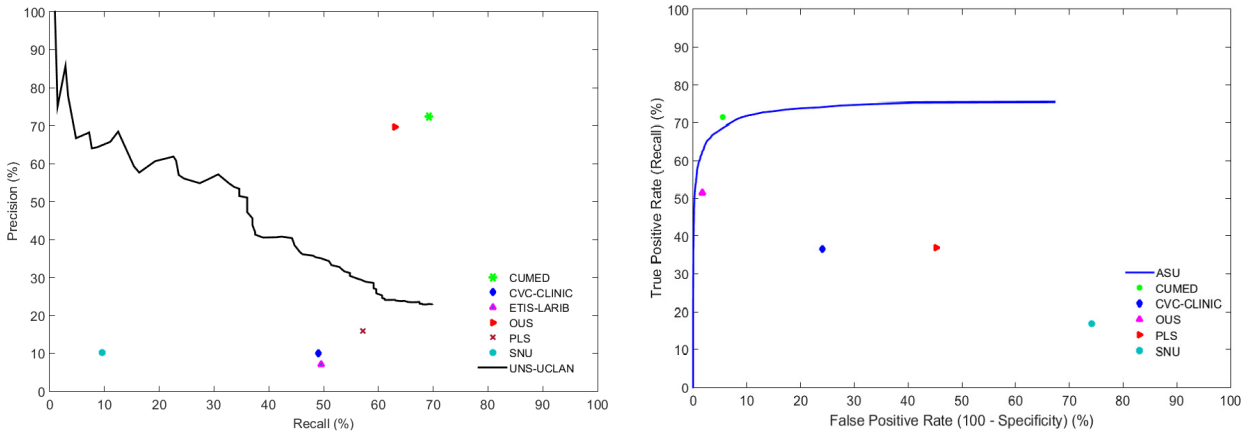
TABLE VII
SUMMARY OF THE MOST RELEVANT RESULTS REGARDING VIDEO SEQUENCE ANALYSIS.

All videos									
	TP	FP	TN	FN	Prec [%]	Rec [%]	Spec [%]	F1 [%]	F2 [%]
Individual team analysis									
ASU	2636	184	13149	1677	93.5	61.1	98.6	73.9	65.7
CUMED	3081	769	13010	1232	80.0	71.4	94.4	75.5	73.0
CVC-CLINIC	1578	3456	10927	2735	31.3	36.6	75.9	33.8	35.4
OUS	2222	229	13245	2091	90.6	51.5	98.3	65.7	56.4
PLS	1594	10103	12258	2719	13.6	36.9	54.8	19.9	27.5
SNU(Only videos with polyps)	721	3285	1140	3592	17.9	16.7	25.7	17.3	16.9
Combination of teams analysis									
Best concatenation result (all teams)	3576	14741	10064	737	19.5	82.9	40.6	31.6	50.3
Best saliency result (all teams)	3294	4070	10064	1019	44.7	76.4	71.2	56.4	66.9
Best saliency voting result (ASU,CUMED)	3316	557	12915	997	85.6	76.8	95.9	81.0	78.4
Videos with frames with and without polyp									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
Individual team analysis									
ASU	1218	92	1864	1431	92.9	45.9	95.3	61.5	51.1
CUMED	1439	600	1692	1210	70.6	54.3	73.8	61.4	57.0
CVC-CLINIC	195	1343	1430	2454	12.7	7.4	51.6	9.3	8.0
OUS	651	55	1914	1998	92.2	24.6	97.2	38.8	28.8
PLS	328	6953	920	2321	4.5	12.4	11.7	6.6	9.2
SNU	282	2085	1140	2367	11.9	10.6	35.3	11.2	10.9
Combination of teams analysis									
Best combination by union (all teams)	1949	11128	493	700	14.9	73.6	4.2	24.8	41.2
Best saliency by union (all teams)	1588	2557	493	1061	38.3	59.9	16.2	46.7	53.9
Best saliency voting result (CUMED,ASU)	1698	439	1649	951	79.4	64.0	79.0	70.9	66.7
Videos with only polyp frames									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
Individual team analysis									
ASU	1418	40	no	246	97.2	85.2	N/A	90.8	87.4
CUMED	1642	149	no	22	91.7	98.7	N/A	95.0	97.2
CVC-CLINIC	1383	272	no	281	83.5	83.1	N/A	83.3	83.2
OUS	1571	167	no	93	90.4	94.4	N/A	92.3	93.6
PLS	1266	3150	no	398	28.7	76.1	N/A	41.6	57.2
SNU	439	1200	no	1225	26.8	26.4	N/A	26.6	26.5
Combination of teams analysis									
Best combination by union (all teams)	1664	4978	N/A	0	25.0	100.0	N/A	40.0	62.6
Best saliency by union (all teams)	1662	2	N/A	2	99.8	99.8	N/A	99.8	99.8
Best saliency voting (all teams)	1662	2	N/A	2	99.9	99.9	N/A	99.9	99.9
Videos without polyp frames									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
Individual team analysis									
ASU	N/A	52	11286	N/A	N/A	N/A	99.5	N/A	N/A
CUMED	N/A	20	11318	N/A	N/A	N/A	99.8	N/A	N/A
CVC-CLINIC	N/A	1841	9497	N/A	N/A	N/A	83.8	N/A	N/A
OUS	N/A	7	11331	N/A	N/A	N/A	99.9	N/A	N/A
PLS	N/A	0	11338	N/A	N/A	N/A	100.0	N/A	N/A
Combination of teams analysis									
Best combination by union (all teams)	N/A	1920	9454	N/A	N/A	N/A	83.2	N/A	N/A
Best saliency by union (PLS,OUS)	N/A	20	11318	N/A	N/A	N/A	99.8	N/A	N/A
Best saliency voting (CUMED,PLS,OUS)	N/A	0	11338	N/A	N/A	N/A	100.0	N/A	N/A

in all considered metrics, as ASU provides a better balance between true and false alarms (higher F1-score) at the cost of detecting less polyp frames (2636 vs. 3081).

We present in Table VII a complete breakdown of video analysis results, dividing them into 3 groups according to the degree of polyp presence in the sequences: 1) videos containing frames with and without polyps; 2) videos containing only frames with polyps, and 3) videos containing non-polyp frames. In all cases, results again show a superior performance of CUMED in terms of total number of polyp frames detected. Deepening the analysis, we observe a decrease in the difference in performance observed in global analysis between hand-crafted methods (CVC-CLINIC) and CNN-based methods when videos with only polyp frames are analyzed. This can be related to those methods being designed to highlight polyp-like structures in the image (localization) but not for determining specific polyp presence. The analysis of sequences without polyp frames shows that PLS offers the best performance, which is possibly due to the presence of a specific polyp presence module in this approach.

As mentioned in Section III-C, a statistical analysis is performed to account for differences in performance between



(a) (b)

Fig. 6. Performance curves: (a) Precision-Recall curve for the analysis of the ETIS-LARIB database and (b) Receiver Operating Characteristic (ROC curves) for the analysis of the ASU-Mayo database. For the ROC curve, SNU operating point is calculated from the videos the team provided results for. Methods are represented with a line in cases where the confidence value has been provided for each detection, otherwise the operating point is used.

methods. Results of the Saphiro-Wilk test over the F1 results for each video and method indicates a normal data distribution ($p-value > 0.05$). Considering this, we perform a subsequent ANOVA analysis and multicomparison test to compare the different methods. The ANOVA test detects significant differences across F1 values ($p - value = 5.4e^{-10}$), which are explored in the multicomparison test shown in Fig. 8. Results of this test show the superior performance of ASU, providing CUMED with a comparable performance different from the rest in a statistically significant way. CUMED and OUS also show performances comparable to each other. Finally CVC, PLS, and SNU also present comparable performances.

We present in Fig. 7 detection latency results. We can observe how there are only two teams (ASU and CUMED) which present latency scores for all the videos. We perform a statistical analysis to account for the differences between them. The result of the Saphiro-Wilk test indicates a non-normal data distribution ($p - value < 0.05$) and, consequently, the Kruskal-Wallis test is performed to account for statistically significant differences. In this case the test confirms the null hypothesis that both data samples come from the same distribution $p-value = 0.76$, which can be observed in Fig.8. Concerning the rest of the methods, we can observe that they do not detect the polyps in all the videos which is also a cause of the difference in performances shown in Table VII.

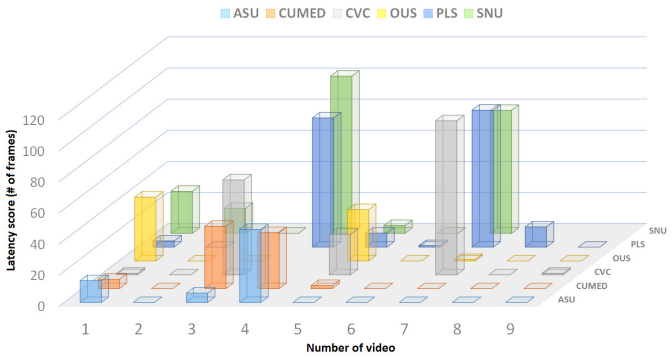


Fig. 7. Detection latency for polyp-containing videos.

B. Additional validation experiments

1) Combination of methods: We have included in Table IV-A and in Table VII the best performance achieved after applying each of the proposed method combination strategies. The most important though logical conclusion extracted is that a combination of methods leads to better detection results. As expected, any combination of methods leads to an increase of the total number of detected polyps. This shows that different methods detect different polyps and that even those with lower performance can contribute positively to the overall detection.

We can observe in Fig. 9 (d-f) that if we do not include all teams in the combination, the number of correct polyp frame detections could be affected. We can also observe in Fig. 9 (a-c) that the combination of the two best methods in each category surpasses the individual methods' performance, which indicates the potential of saliency map methods to build up more reliable systems. It is clear that the combination by union strategy increases the total number of detected polyp

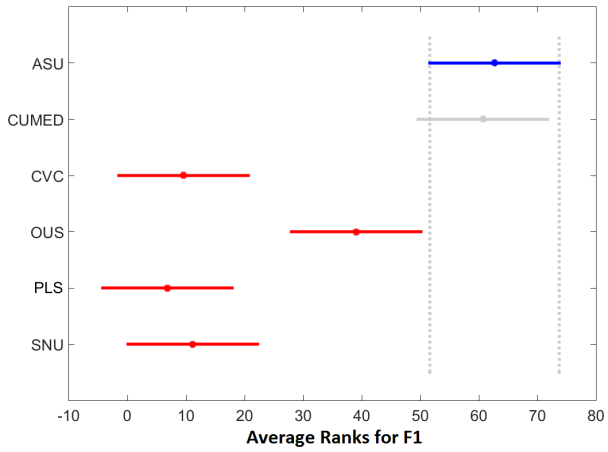
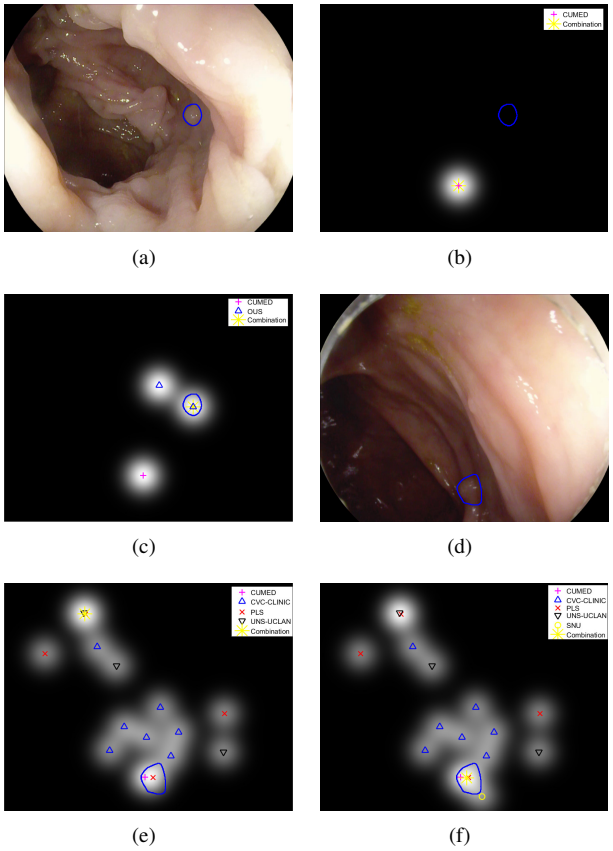


Fig. 8. Multicomparison test for the analysis of the F1 score in videos showing frames with and without polyps. Each method is represented as a horizontal line whose center is located in corresponding method's mean F1 score and whose width corresponds to the variance calculated according anova1 fit model. The best ranked group is represented by a blue horizontal line, comparable methods are shown in grey, and methods that are different in a statistically significant way are shown in red.



(a) (b) (c) (d) (e) (f)

Fig. 9. Examples of the benefits of using a saliency-map-based approach. The first row shows the impact of combining the two best methods that surpass their individual performances: (a) original image; (b) saliency map with the position of detection points superimposed (best method, CUMED); (c) saliency map with the position of detection points superimposed (two best method, CUMED and OUS). The second row shows the positive impact of the worst performing method: (d) original image; (e) saliency map with the position of detection points superimposed (all method); (f) saliency map with the position of detection points superimposed (all method but SNU). The polyp contour is represented in blue. Each method is represented by a different color and shape.

frames at the cost of vastly increasing the number of false alarms and, consequently, other strategies should be explored to achieve a clinically useful system.

Considering this, we observe that the use of saliency maps leads to a better balance between correct and false alarms. Regarding the two saliency-map sub-strategies, the voting strategy leads to a slightly better performance for the case of still frame analysis, specifically observed in the reduction of FP. This can be explained as being due to those poorly performing methods providing outputs for almost all frames. Once their contribution is not considered as majority is not achieved for a particular frame, these false alarms vanish.

Taking into account these results, if we consider a potential combination of methods as the solution for polyp detection, we would propose saliency maps with a voting sub-strategy as the strategy that leads to a better compromise between correct detections and false alarms, though other potential combinations can be explored. For instance, we can think of a system which also includes the specific detection modules that some approaches have presented here (PLS, SNU), with polyp localization within a given image being then obtained using CNN-based approaches (CUMED, OUS).

2) Impact of image challenges on individual methods' performance: We present in Table VIII a summary of the results of the experiment assessing the impact of several image challenges on individual methods' performance. With respect to polyp frames, the first conclusion to be extracted is that low visibility images and the presence of specular highlights within the polyp affect all methods in the same way. We interpret the impact of image quality as being both mucosa wall and its elements, such as polyps, better visually defined in good visibility images hence helping in polyp detection. We associate the positive impact of specular highlights inside the polyp to polyps appearing commonly as protruding elements in the scene and, as a consequence of this, specularities appear in their surface as their reflect light back to the camera [52].

There are some image challenges that generally seem to make polyp frames detection difficult such as the presence of overlay information and overexposed regions, with the latter being more prevalent in the explored images. The clear view of the luminal region also negatively affects detection capabilities, which we interpret as result of lumen presenting strong boundaries and contrast in comparison with the mucosa, which is a feature that polyps also exhibit. Surprisingly, the presence of intestinal content affects positively Recall and F1-score; this could be explained by the fact that this image challenge appears clearly different from polyps (weak contours, different color). Finally, the degree of completeness of the polyp seems to present a low impact on the performance of the methods, specially regarding F1-score.

Regarding non-polyp frames, we can observe that the presence of overlay information and overexposed regions helps methods to discard frames without a polyp. Intestinal content leads to more false alarms, as does the presence of the luminal region and the presence of specular highlights; the three of them may falsely indicate the presence of a polyp, as the may also present contrast to mucosa or an indication of protrudness. We can also observe how methods tend to provide a higher number of false alarms for good quality images, which we interpret as a result of structures likely to be confused with polyps being better visually defined.

With respect to individual methods, we observe that those including boundary information (ASU, CVC-CLINIC and SNU) in their methodologies are specially damaged by the presence of structures with strong boundaries such as lumen or overlay information. End-to-end learning approaches are less affected in non-polyp frames analysis and they benefit from the presence of specular highlights in polyp frames.

3) Impact of polyp morphology on methods' performance: We present in Table X and in Table IX results of the study on the impact on polyp morphology on method's performance. It has to be noted that we only provide results for the morphologies that appear in each particular database, as described in Section III. We can observe that for both still frames and video sequences analysis, methods' performance do depend on polyp morphology. With respect to still frame analysis, we can observe that Recall scores are higher for sessile polyps (including sub-types 0-ISp and 0-IIa+Is) than for those less elevated (including flat) ones. We associate this to appearing sessile polyps more different to the mucosa and hence attract-

TABLE VIII
IMPACT OF CLINICAL AND TECHNICAL CHALLENGES ON INDIVIDUAL METHODS' PERFORMANCE.

Clinical and Technical Challenges								
Team	Overlay information	High specular highlights presence	Overexposed regions	Intestinal content	Luminal region	Incomplete polyp	Specular highlights inside polyp	Low visibility
Polyp frames: Recall [differences in %]								
ASU	-14.3	48.2	-10.2	7,7	-11.10	-0.2	43.5	-11.9
CUMED	-3.9	42.1	-24.7	8.3	-17.1	-1.3	23.0	-30.1
CVC-CLINIC	-28.8	24.7	-20.7	28.9	-35.5	3.4	26.9	-11.6
OUS	-27.3	49.6	-31.3	18.8	-21.4	15.9	41.0	-14.1
PLS	-12.0	21.3	-3.7	28.4	-16.1	1.0	35.3	-13.4
SNU	-18.5	13.9	-2.1	2.5	-10.7	-2.5	14.8	5.6
Polyp frames: F1-score [differences in %]								
ASU	-11.1	50.8	-8.4	5.5	-8.2	0.3	43.5	-9.3
CUMED	-18.6	32.1	-28.0	8.3	-16.9	1.1	16.3	-18.0
CVC-CLINIC	-31.6	23.3	-23.4	30.3	-40.8	1.0	27.9	-13.1
OUS	-25.6	58.9	-30.4	15.0	-16.1	13.0	44.2	-11.9
PLS	-7.4	20.2	-5.1	18.7	-15.1	2.5	25.9	-4.0
SNU	-20.6	15.3	-1.9	3.2	-12.9	-4.1	15.6	6.6
Non-polyp frames: Specificity [differences in %]								
ASU	0.8	-0.4	0.5	-0.7	-0.5	N/A	N/A	0.3
CUMED	-7.9	-1.2	2.6	0.1	-3.6	N/A	N/A	4.8
CVC-CLINIC	-63.8	-26.0	12.5	-24.5	2.1	N/A	N/A	32.7
OUS	0.1	-0.2	-0.1	-0.1	-0.1	N/A	N/A	0.1
PLS	13.7	-3.9	28.2	-17.0	-32.2	N/A	N/A	42.5
SNU	40.8	5.6	-15.7	-14.6	17.9	N/A	N/A	20.9

ing the attention of the different methods. We can also observe how CVC-CLINIC and ETIS-LARIB , despite offering worse overall performance, are able to detect all kind of polyps though they obtain worse Precision scores.

Concerning video sequences, differences regarding degree of polyp elevation follow the same trend; in this case we can observe big differences in Recall for all methods but, in this case, Precision is not greatly affected but for the case of CVC-CLINIC, which is logical due to its big dependence on boundary presence to guide polyp detection; boundaries between mucosa and the polyp are less distinguishable for the case of slightly elevated polyps. Finally, this positive increase in Recall score associated to sessile polyps also has an impact in latency score; all teams achieve smaller latency scores for those videos containing polyps of this morphological type (videos 2, 6, 8 and 9) in Fig. 7.

4) Temporal coherence on method's response: We present results of our temporal coherence study on Table XI(a). For both consecutive frame and within sequence detection stability, we can observe that results follow the same trend than the analysis of individual frames, being CUMED and ASU the teams which present a higher degree of temporal coherence, despite none of them including temporal information as part of their methodology. We can also observe how CUMED and ASU are able to correctly detect polyp frames in more than half of the polyp frames that each sequence contain, which can be associated to them being more capable to cope with polyp appearance variability within a same sequence.

TABLE IX
IMPACT OF POLYP MORPHOLOGY IN METHODS' PERFORMANCE: VIDEO SEQUENCE ANALYSIS. ONLY FRAMES CONTAINING A POLYP ARE CONSIDERED FOR METRICS CALCULATION.

	0-Is (4 polyps, 2212 images)				**0-IIa (5 polyps, 2101 images)**			
	Prec	**Rec**	**F1**	**Lat**	**Prec**	**Rec**	**F1**	**Lat**
ASU	97.4	73.7	83.9	0.5	97.2	47.9	64.1	13.4
CUMED	92.1	86.8	89.4	0.0	77.3	55.3	64.4	16.8
CVC-CLINIC	81.9	62.5	70.9	0.3	19.3	9.3	12.5	46.7
OUS	90.9	77.6	83.7	0.0	92.1	24.1	38.2	18.7
PLS	24.6	62.6	35.3	3.5	10.2	9.9	10.1	46.0
SNU	26.6	23.3	24.9	79.0	15.9	9.7	12.1	42.4

5) Analysis of the direct output of the methods: We present mean and standard deviation values of CR in Table XI(b). As we can observe, CUMED achieves the higher mean CR value across all frames from **ETIS-LARIB** database, concentrating around half of the total energy of the image inside the polyp. It has to be noted that, in this case, differences between methods can be associated to several reasons. First of all, it is clear that methods detecting correctly more polyps will be prone to concentrate more energy inside them hence the superior performance of CUMED, which was the best performing team in still frame analysis. Second, we also have to consider how the actual output of the method looks like, as it can have an impact in the specific metric considered.

We observe in Fig. 10 how some methods do not provide probabilistic energy maps but binary masks approximating the polyp region. Due to these regions having pre-determined shapes, two problems may appear. First, it is highly unlikely that actual polyps fit those shapes, so that any pixel-wise metric value can be damaged by the shape choice. Second, if methods' evaluation is based on the calculation of detection scores from single-position values and this position is calculated as the centroid of the pre-determined shape in case of large regions partially covering the polyp, it may happen that the detection position falls outside the polyp when in fact part

TABLE X
IMPACT OF POLYP MORPHOLOGY IN METHODS' PERFORMANCE: STILL FRAME ANALYSIS.

	0-Is (27p,127im)			0-IIa (9p,45im)			0-IIb (4p,6im)			0-IIa+c (2p,6 im)			0-Isp (1p,6im)			0-IIa+Is (1p,6im)		
Team	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
CUMED	79.5	79.5	79.5	57.4	60.0	58.7	33.3	11.1	16.6	42.8	50.0	46.1	100.0	100.0	100.0	83.3	83.3	83.3
CVC-CLINIC	7.7	44.8	13.1	10.4	48.9	17.1	35.7	27.8	31.2	42.8	100.0	60.0	66.6	100.0	80.0	18.2	100.0	30.7
ETIS-LARIB	5.4	50.4	9.7	13.9	46.6	21.4	6.7	22.2	10.4	75.0	50.0	60.0	11.6	83.3	20.4	18.2	100.0	30.7
OUS	67.3	76.4	71.6	66.6	40.0	50.0	0.0	0.0	0.0	100.0	66.6	80.0	100.0	100.0	100.0	85.7	100.0	92.3
PLS	15.3	58.3	24.2	15.0	60.0	24.1	0.0	0.0	0.0	33.3	100.0	50.0	28.6	100.0	44.4	25.0	100.0	40.0
SNU	11.8	11.8	11.8	4.4	4.4	4.4	0.0	0.0	0.0	16.6	16.6	16.6	0.0	0.0	0.0	33.3	33.3	33.3
UNS-UCLAN	24.2	74.0	36.5	15.3	55.5	24.0	0.0	0.0	0.0	75.0	100.0	85.7	75.0	100.0	85.7	33.3	100.0	50.0

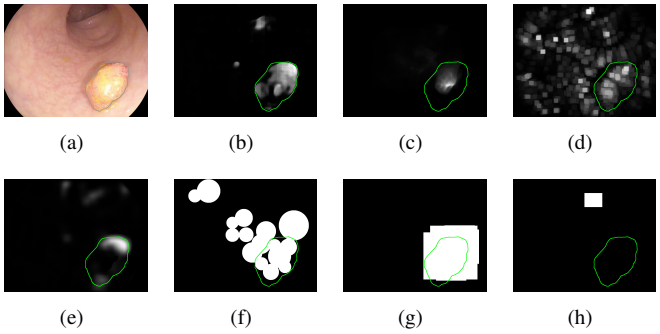


Fig. 10. Comparison of energy maps provided by each method: (a) original image (b) CUMED (c) CVC-CLINIC (d) PLS (e) UNS-UCLAN (f) ETIS-LARIB (g) OUS,s and (h) SNU. In each of the images, a green line represents the reference polyp mask.

of the polyp region was covered by method's output.

Consequently, we think it is not fair to include those methods (ETIS-LARIB, OUS and SNU) in a CR-based comparison. We did statistically compare the different energy map-based methods. Preliminary results shown in Fig. 11 indicate again a superior performance of CUMED over the rest. Regarding the statistical significance of the differences, the Saphiro-Wilk test over CR values indicates a normal distribution of data. Therefore ANOVA and multicomparison tests are performed to study potential differences across methods. The ANOVA test detects significant differences across CR values ($p - value = 2.51e^{-61}$), which are explored in the multicomparison test shown in Fig. 11. CUMED's performance is statistically different from the rest of the approaches, which present comparable performances between them.

V. DISCUSSION

A. Impact of the methodology used on method's performance

The main result of this comparative study is that methods including some degree of machine learning outperform classic hand-crafted methods, specially regarding specificity scores in non-polyp videos. This correlates with the trend actually observed on computer vision research; methods traditionally were hybrid, using hand-crafted features and machine learning to classify a given input image according to the specific problem to solve. There is an extensive amount of hand-crafted features defined within the computer vision community, covering from general ones such as HOG or SIFT features to others more domain-specific, such as the ones presented by CVC-Clinic team. Designing hand-crafted features to solve specific problems can be complicated and highly time consuming, as well as limiting the widespread use of a new developed technology.

CNNs allow to learn jointly problem-specific features and the classifier to differentiate among classes. Their great power comes from the ability of learning problem-specific features in an increasing depth of complexity and abstraction. As it has been shown in this paper, we can observe a superior performance of CNN-based approaches over hand-crafted methods. We can also observe differences in performance between CNN-based methods which shows how obtaining good performance of these networks depends strongly on defining the proper architecture and having quality data to feed the network. Regarding the design of the network, there are several details to take into account which go from pure architecture decisions (number of layers, number and size of the filters of each layer or activation functions) to how the training is done (choice of optimization method, setting a learning rate, data preprocessing). A proper selection of these

TABLE XI
TEMPORAL COHERENCE AND CONCENTRATION RATIO RESULTS. FOR EACH METHOD, MEAN AND STANDARD DEVIATION VALUES OF CORRESPONDING METRIC ARE PROVIDED.

(a) Temporal Coherence

Method	Consecutive frames	Within sequence
ASU	54.8 ± 21.4	61.2 ± 21.9
CUMED	63.7 ± 23.9	67.7 ± 23.2
CVC-CLINIC	34.1 ± 37.6	30.3 ± 37.0
OUS	48.7 ± 30.8	47.2 ± 34.1
PLS	27.4 ± 26.1	31.7 ± 30.4
SNU	10.2 ± 06.3	11.7 ± 11.7

(b) Concentration Ratio

Method	Value
CUMED	48.5 ± 26.7
CVC-CLINIC	17.9 ± 24.1
PLS	11.9 ± 14.4
UNS-UCLAN	18.7 ± 18.8

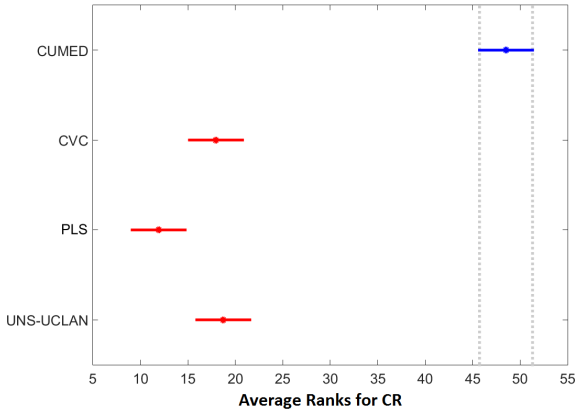


Fig. 11. The multicomparison test for the analysis of the CR score in the ETIS-LARIB database. Each method is represented as a horizontal line centered in corresponding method's mean CR score and whose width corresponds to the variance calculated according anova1 fit model. The best ranked group is represented by a blue line, comparable methods are shown in grey, and statistically significant different methods are shown in red.

parameters may lead to a boost in performance achieved.

Apart from differences related to parameter tuning, we can also observe that one important difference between CNN-based proposals rely on the type of data used to define the network. CUMED uses only colonoscopy data whereas OUS and SNU networks are pre-trained over general image databases. CNNs are structured in layers and each of them captures a different kind of features from the data; first layers capture basic image features such as boundaries whereas deeper ones capture more meaningful and abstract features built over the previous ones. Considering this, features learnt on the first layers might work well in several domains but those learnt in the last layers are more application dependent. One big requirement to use CNNs is to have a large amount of labelled data that may not be available for the case of medical imaging analysis. One widespread solution (used in SNU methodology) is to train the network on a very big database such as ImageNet (with over 10^6 images and 10^3 classes) and then fine tune the network to adapt for a more specific domain. The problem relies on ImageNet containing images potentially very different that the ones that the polyp detection system will have to deal with and, as results show, this may limit the use of those pre-trained networks. In this sense, we can observe how methods using colonoscopy data from end to end (CUMED) obtain better performance than those trained in general datasets such as OUS or SNU, which indicate that efforts should be made to build up domain-specific networks in order to obtained desired performance levels.

Concerning a general comparison among methods regardless their methodology, we can also observe from Table VII that recall scores improve if we only consider frames with polyps. As non-polyp frames are included in the study, performance of hand-crafted and some of the hybrid methods decrease with respect to all metrics. We explain this as being due to some of the methods being specifically tuned for polyp-like structures detection, but not on specific polyp presence or absence; this can be observed in the high number of false positives that these approaches provide as they seem to find these polyp-like structures in frames without a polyp. We link this to non-polyp frames containing structures which guide polyp detection methods, such as, boundary information which also appears associated in other endoluminal structures such as folds or vessels. We also observe strong differences in the performance of hand-crafted methods when dealing with polyp frames in the two proposed scenarios (still frames and videos). This could be related to the fact that the ETIS-LARIB database presents a high number of lateral polyp views, deviating from the model of appearance which the hand-crafted method is based on [14].

B. Impact of clinical and technical image challenges on method's performance

We have presented in this paper a preliminary study on whether image challenges defined and reported by clinicians and technicians do impact the performance of an automatic polyp detection method. Results exposed in Table VIII show that all of them, in a certain degree, should be tackled in order the automatic system to efficiently assist clinicians during the procedure. The most straightforward conclusion from this experiment is that image quality matters, as methods' performance decrease when only bad quality images are considered. The presence of extra-endoluminal structures such as overlay information or overexposed regions do also affect negatively the performance of automatic methods. This indicates that efforts should be made during the exploration in order a computational support system to efficiently assist clinicians. We can also observe that results do improve if luminal region is not present in the image; this correlates with actual exploration guidelines in which a thorough inspection of the mucosa is prescribed in order to efficiently detect polyps.

It is also interesting to observe how there are some cases in that image challenges considered for both technical and clinical domains do not suppose an actual technical challenge. For instance, we would expect that the presence of intestinal content or the observation of specular highlights over the polyp would impact negatively the performance of an automatic method; results show that studied methods are indeed positively affected by their presence. We associate this to these image challenges appearing clearly different from polyps.

Moreover, and also related to this, there are some image challenges which may provide unexpected results and which would need to be better defined to avoid potential subjectivity. Though we have gathered three observations per frame to mitigate this, some of the image challenges should be defined appropriately to avoid discrepancies between observers. For instance, the presence of intestinal content, image quality or, specially, the high specular highlights presence should be redefined as, for the former, we should also consider its size and type (solid, bubbles) and, for the latter, we may consider not only the number but their size and position in the image.

Apart from the image challenge experiment, we have also performed another one to assess the impact of polyp morphology on methods' performance. Even considering that this experiment is limited to the actual morphologies that the databases contain, differences can already be observed in a way such methods obtain better performance as the polyp protrudes more from the mucosa. With respect to polyps with flat morphology cited by clinicians as the most difficult ones to the detect [9], [10] we observe, for the case of still frames analysis in which they are present, that there are methods that are already able to detect them, despite its low presence in training and testing databases.

C. Towards clinical applicability

One of the objectives of the challenge and, consequently, of this paper, was to assess if any of the participating methods is close to clinical applicability. In order to assess this, we have proposed several studies to observe certain features that a given clinically applicable computational polyp detection method should have. The main feature that a clinically applicable system should have is that it should detect all polyps regardless their appearance (high detection rate (DR), measured as the percentage of polyps detected in at least one frame out of the total of polyps present in the testing videos). This detection

TABLE XII
SUMMARY OF INDIVIDUAL METHOD'S PERFORMANCE. FOR LATENCY AND TEMPORAL COHERENCE MEAN AND STANDARD DEVIATION VALUES FROM THE ANALYSIS OF THE 9 VIDEOS WITH POLYPS ARE PROVIDED.

Method	DP [%]	Latency [frames]	Rec [%]	F1 [%]	TempC [%]
ASU	100.0	7.4 ± 15.6	61.1	73.9	54.8 ± 21.4
CUMED	100.0	9.3 ± 16.4	71.4	75.9	63.7 ± 23.9
CVC-CLINIC	77.8	26.8 ± 39.0	36.6	33.7	34.1 ± 37.6
OUS	88.9	9.4 ± 17.2	51.5	65.7	48.7 ± 30.8
PLS	88.9	24.7 ± 37.8	36.9	19.9	27.4 ± 26.1
SNU	77.8	32.5 ± 40.9	10.6	11.2	10.2 ± 06.3

should also be fast enough to be of an actual help; speed here is not only considered in terms of computation time but also in response to a stimuli as a computational method should react to polyp presence as soon as it appears in the image (associated to a low latency score). The actual response of a given method should be stable over time (high temporal coherence) in order to provide an smooth assistance to clinicians in polyp search. Finally, and considering the scope of application of the methods, the number of false alarms should be kept low (high F1 score associated to an also high Recall value) as the contrary would suppose indicating the clinicians to further explore uninteresting regions of the image.

Considering these criteria, we present in Table XII a summary of the main results presented in this paper for the video analysis challenge. Columns are ordered according to the, under our point of view, relevance of the specific criteria. As it can be seen, there are only two methods (CUMED, ASU) that may be actually considered for a potential clinical use as they do detect all polyps. Concerning the rest of criteria, both do perform similarly: ASU presents a lower latency which could be compensated by CUMED's higher temporal coherence degree. Concerning frame-based metrics we can observe that ASU leads to a better balance between true and false alarms though CUMED detects polyps in more frames.

It has to be noted that we have not included in Table XII information regarding computation time for comparison purposes as they have not been tested under the same configuration and, consequently, provided times may vary in an actual final deployed system. Nevertheless, a clinically useful method should operate under real-time constraints. Considering that videos are recorded on 25 or 30 frames per second, processing of a new frame should not take more than 40 ms (33 ms for NTSC systems) in order not to suppose a delay in overall procedure time. All methods studied in this paper have computation times higher than these threshold values and, consequently, do not comply with real-time constraints though the processing of all frames might not be needed considering due to the small variation between consecutive frames due to usual smooth camera movement.

D. Possible improvements in validation framework

During the analysis of the performance of each of the methods, we have discovered several aspects to be considered for future iterations of this study.

The first one deals with the variability of the image quality provided in the training and testing stages. In this study, the databases used for validation come from three different sources presenting differences in image size or acquisition system, as we have source data from both OLYMPUS and PENTAX devices. This was done on purpose, as it is impossible to predict under which specific scenario a given system can be potentially used, as there is no standard regarding resolution or manufacturer and a given method should perform similarly regardless of the specific conditions. But it is true that these variabilities may have affected the performance of the different methods, as training was done using images with resolutions different from the ones used for testing. These differences in resolution can imply to have a greater level of texture detail which can impact the performance of systems trained with SD images (i.e. edge detection could be greatly affected by the presence of small texture details inside the polyp).

Also related to database content, and after observing that polyp morphology can impact methods' performance, an effort could also be made on enlarging the databases to cover those types that are not currently present. It is important to mention that performance of learning-based approaches for certain morphologies could be affected by the lack of frames of this particular type in the training set. In our experiment, this only happens for still frames analysis as CVC-CLINIC database does not contain polyps of types 0-IIa+c and 0-IIa+Is which are indeed present in ETIS-LARIB database. Nevertheless it has to be noted that these types are only present in 12 frames out of the 196 frames of the database and, consequently, global performance should not be greatly affected by this issue. Finally, not all types are represented in the database (for instance, proposed databases have no examples of types 0-Ip, 0-IIc or 0-III); it would surely be helpful to study how computational methods deal with those polyp types reported as the ones with higher associated miss-rate [9], [10].

The second are of improvement deals with how actual results are calculated. The majority of results presented in this paper have been calculated from the CSV files provided by participants in the challenge. Though they are useful to represent the actual performance of the method, we think it is also necessary to analyze how these files are generated (the actual output of the method they come from) in order to have a deep understanding on how a given method performs. In this sense, we proposed a preliminary study comparing the amount of actual image energy that is kept inside the polyp. As it was shown in Fig. 10, there are big differences in how the actual output of the methods is calculated, inherent in each teams' methodology. Therefore, if we wanted to present a fair comparison between methods over their direct output, specific guidelines should be given to participants in order to gather comparable data.

Finally, we think that Precision-Recall and ROC curves should be used for methods' comparison as well. In order to provide these curves for all teams, confidence values should have been provided; in this case, only one team per subcategory (UNS-UCLAN in still-frame analysis and ASU-Mayo for full video analysis) provided this information whereas the rest only provided what we assume are results obtained using the best configuration of each particular method. Nevertheless, we have presented both curves in Fig. 6 along with quantitative

data in both Table IV-A and Table VII.

VI. Conclusions

We present in this paper a complete validation study comparing the performance of different polyp detection methods. Eight different teams took part in this challenge, ranging from methodologies based on hand-crafted methods to trending techniques such as CNNs. We propose the use of uniform performance metrics and common, publicly available, fully annotated databases to objectively assess their performance.

The analysis of the results obtained by each method shows a superior performance by methods using machine learning as part of their methodologies, obtaining promising performance in both still frames and full-sequence sets. The global analysis of methods' performance shows that some of them are close to be clinically applicable as they are able to detect polyps in all sequences with a small reaction time. We have also shown how there is a clear link between clinical and technical challenges and that mitigating them is key to improve methods' performance. As it was expected, our preliminary study proves that image quality and careful mucosa inspection do have a positive impact in methods' performance.

A deep analysis of the results shows that, as different methods detect different polyps, there is room for improvement by combining some of the methods into a new solution. Going along this line, we have performed a first study on how to combine some of the methods in order to improve detection performance. Preliminary results show how the combination of the best methods can be used to exceed best individual scores, indicating the potential of creating clinically useful systems integrating capabilities from several individual methods.

Beyond presenting challenge results, this study shows areas in which methods might focus to increase their performance, such as the ability to work equally under different conditions, the necessity of include spatial and temporal coherence in full sequences analysis or by considering the presence of other elements of the scene to help in polyp detection task. More importantly, this study also shows how efforts should be made between clinicians and computer scientists to build up image acquisition protocols that can help to better observe (clinicians) and analyze (computational methods) the endoluminal scene. Finally and concerning availability of data to test methods, the study shows that granting access to large available labelled data is needed for a comprehensive validation of a polyp detection method and that this might lead to a boost in performance of end-to-end learning methods. We believe efforts should also be made to create and use data from new imaging technologies such as magnification endoscopy or virtual chromoendoscopy, due to increased visualization performance already observed by clinicians.

After analyzing the complete validation study, we have detected several areas in which the study can be extended to provide with an even deeper comparative analysis of the performance of polyp detection methods. More precisely, future studies should tackle some of the issues detected such as the variability in source data resolution and size and should aim to cover all different polyp morphological types. Moreover, a consensus should be reached on how the information provided by each method is to be interpreted to allow a comparison beyond simple detection positions. This may result in, apart from a more complete analysis, a deeper understanding on how each method works and in which scenarios each of them show the most benefit, thinking of potential optimized combinations of them to finally build up a clinically useful method.

ACKNOWLEDGEMENTS

The authors would like to thank EndoVis challenge organizers for their continuous help and guidance through both challenge and paper preparation. The idea of organizing a competition for polyp detection in colonoscopy was first conceived by Dr. Jianming Liang, and the foundational framework was established by Drs. Tajbakhsh and Liang before the first challenge at ISBI-2015. The associated ground truth images in the ASU-Mayo Clinic Colonoscopy Video © Database were created by Saiswathi Javangula, Ireen Khan, Kamran Bodushev, Sarah Fallah-Adl, and Tracy Phan. The ASU-Mayo Clinic Colonoscopy Video © Database is copyrighted and its use is granted to the work for the challenge on polyp detection in colonoscopy as reported this IEEE TMI paper. For any other uses, a prior agreement must be obtained from Arizona State University. This work was supported by several grants through ASU-Mayo Clinic partnerships, by the Spanish Government through the funded project iVENDIS (DPI2015-65286-R), by the FSEED and by the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya, 2014-SGR-1470 and 2014-SGR-135. The authors would further like to acknowledge support from the European Union through the ERC starting grant COMBIOSCOPY under the New Horizon Framework Programme grant agreement ERC-2015-StG-37960.

References

[1] A. C. Society, "Colorectal cancer," online, January 2016.
[2] M. Gschwantler and S. e. a. Kriwanek, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics," *European journal of gastroenterology & hepatology*, vol. 14, no. 2, pp. 183–188, 2002.
[3] A. Jemal, R. Siegel *et al.*, "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.
[4] R. Jover, M. Kalager11 *et al.*, "Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy (esge) guideline," *Endoscopy*, vol. 45, pp. 842–851, 2013.
[5] D. Burling, I. C. for CT Colonography Standards *et al.*, "Ct colonography standards," *Clinical Radiology*, vol. 65, no. 6, pp. 474–480, 2010.
[6] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, p. 417, 2000.
[7] C. D. Johnson, M.-H. Chen *et al.*, "Accuracy of ct colonography for detection of large adenomas and cancers," *New England Journal of Medicine*, vol. 359, no. 12, pp. 1207–1217, 2008.
[8] M. J. Farnbacher, H. H. Krause *et al.*, "Quickview video preview software of colon capsule endoscopy: reliability in presenting colorectal polyps as compared to normal mode reading," *Scandinavian journal of gastroenterology*, vol. 49, no. 3, pp. 339–346, 2014.
[9] A. Leufkens, M. Van Oijen *et al.*, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study." *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.
[10] J. C. Van Rijn, J. B. Reitsma *et al.*, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *The American journal of gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006.
[11] B. Lebwohl, F. Kastrinos *et al.*, "The impact of suboptimal bowel preparation on adenoma miss rates and the factors associated with early repeat colonoscopy," *Gastrointestinal endoscopy*, vol. 73, no. 6, pp. 1207–1214, 2011.

[12] S.-H. Lee, I.-K. Chung *et al.*, "An adequate level of training for technical competence in screening and diagnostic colonoscopy: a prospective multicenter evaluation of the learning curve," *Gastrointestinal endoscopy*, vol. 67, no. 4, pp. 683–689, 2008.

[13] M. A. Armin, H. De Visser *et al.*, "Visibility map: A new method in evaluation quality of optical colonoscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 396–404.

[14] J. Bernal, F. J. Sánchez *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[15] M. Bruno, "Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis," *Gut*, vol. 52, no. suppl 4, pp. iv7–iv11, 2003.

[16] H. Inomata, N. Tamai *et al.*, "Efficacy of a novel auto-fluorescence imaging system with computer-assisted color analysis for assessment of colorectal lesions," *World journal of gastroenterology: WJG*, vol. 19, no. 41, p. 7146, 2013.

[17] H. Machida, Y. Sano *et al.*, "Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study," *Endoscopy*, vol. 36, no. 12, pp. 1094–1098, 2004.

[18] R. Coriat, A. Chryssostalis *et al.*, "Computed virtual chromoendoscopy system (fice): a new tool for upper endoscopy?" *Gastroentérologie clinique et biologique*, vol. 32, no. 4, pp. 363–369, 2008.

[19] A. Hoffman, F. Sar *et al.*, "High definition colonoscopy combined with i-scan is superior in the detection of colorectal neoplasias compared with standard video colonoscopy: a prospective randomized controlled trial," *Endoscopy*, vol. 42, no. 10, pp. 827–833, 2010.

[20] N. Gupta, A. Bansal *et al.*, "Accuracy of in vivo optical diagnosis of colon polyp histology by narrow-band imaging in predicting colonoscopy surveillance intervals," *Gastrointestinal endoscopy*, vol. 75, no. 3, pp. 494–502, 2012.

[21] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 3, pp. 172–186, 2015.

[22] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *Medical Imaging, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[23] L. Maier-Hein, A. Groch *et al.*, "Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction," *IEEE Trans Med Imaging*, vol. 33, no. 10, pp. 1913–1930, Oct 2014.

[24] T. Heimann, B. Van Ginneken *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 8, pp. 1251–1265, 2009.

[25] Y. Iwahori, T. Shinohara *et al.*, "Automatic polyp detection in endoscope images using a hessian filter," *Proceedings of MVA*, pp. 21–24, 2013.

[26] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[27] H. Zhu, Y. Fan, and Z. Liang, "Improved curvature estimation for shape analysis in computer-aided detection of colonic polyps," in *International MICCAI Workshop on Computational Challenges and Clinical Opportunities in Virtual Colonoscopy and Abdominal Imaging*. Springer, 2010, pp. 9–14.

[28] J. Kang and R. Doraiswami, "Real-time image processing system for endoscopic applications," in *Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on*, vol. 3. IEEE, 2003, pp. 1469–1472.

[29] S. Hwang, J. Oh *et al.*, "Polyp detection in colonoscopy video using elliptical shape feature," in *2007 IEEE International Conference on Image Processing*, vol. 2. IEEE, 2007, pp. II–465.

[30] S. Karkanis, D. K. Iakovidis *et al.*, "Computer-aided tumor detection in endoscopic video using color wavelet features," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 3, pp. 141–152, 2003.

[31] S. Ameling, S. Wirth *et al.*, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009*. Springer, 2009, pp. 346–350.

[32] S. Gross, T. Stehle *et al.*, "A comparison of blood vessel features and local binary patterns for colorectal polyp classification," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2009, pp. 72 602Q–72 602Q.

[33] Q. Angermann, A. Histace, and O. Romain, "Active learning for real time detection of polyps in videocolonoscopy," *Procedia Computer Science*, vol. 90, pp. 182–187, 2016.

[34] S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2016, pp. 978 528–978 528.

[35] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2016, pp. 253–258.

[36] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE transactions on medical imaging*, vol. 34, no. 11, pp. 2379–2393, 2015.

[37] I. Ševo, A. Avramović *et al.*, "Edge density based automatic detection of inflammation in colonoscopy videos," *Computers in biology and medicine*, vol. 72, pp. 138–150, 2016.

[38] H. Chen, X. J. Qi, J. Z. Cheng, and P. A. Heng, "Deep contextual networks for neuronal structure segmentation," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *arXiv preprint arXiv:1409.5185*, 2014.

[40] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: Deep contour-aware networks for accurate gland segmentation," *arXiv preprint arXiv:1604.02677*, 2016.

[41] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE journal of biomedical and health informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.

[42] S. Park, M. Lee, and N. Kwak, "Polyp detection in colonoscopy videos using deeply-learned hierarchical features," *Seoul National University*, 2015.

[43] S. Demyanov, "A convolutional neural network toolbox," https://github.com/sdemyanov/ConvNet.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[45] O. Russakovsky, J. Deng *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[46] Y. Jia, E. Shelhamer *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654889

[47] M. Riegler, M. Larson, M. Lux, and C. Kofler, "How'how'reflects what's what: Content-based exploitation of how users frame social images," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 397–406.

[48] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection using global geometric constraints and local intensity variation patterns," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 179–187.

[49] ——, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 79–83.

[50] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[51] H. Inoue, H. Kashida, S. Kudo, M. Sasako, T. Shimoda, H. Watanabe, S. Yoshida, M. Guelrud, C. Lightdale, K. Wang *et al.*, "The paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to december 1, 2002," *Gastrointest Endosc*, vol. 58, no. 6 Suppl, pp. S3–43, 2003.

[52] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.