Neural Computing and Applications (2022) 34:10375–10396
https://doi.org/10.1007/s00521-021-06496-4

S. I. : EFFECTIVE AND EFFICIENT DEEP LEARNING

Real-time polyp detection model using convolutional neural networks

Alba Nogueira-Rodríguez[1,2] · Rubén Domínguez-Carbajales[3] · Fernando Campos-Tato[1] · Jesús Herrero[4] · Manuel Puga[4] · David Remedios[4] · Laura Rivas[4] · Eloy Sánchez[4] · Águeda Iglesias[4] · Joaquín Cubiella[4] · Florentino Fdez-Riverola[1,2] · Hugo López-Fernández[1,2,5,6] · Miguel Reboiro-Jato[1,2] · Daniel Glez-Peña[1,2]

Received: 31 December 2020 / Accepted: 30 August 2021 / Published online: 21 September 2021

Abstract

Colorectal cancer is a major health problem, where advances towards computer-aided diagnosis (CAD) systems to assist the endoscopist can be a promising path to improvement. Here, a deep learning model for real-time polyp detection based on a pre-trained YOLOv3 (You Only Look Once) architecture and complemented with a post-processing step based on an object-tracking algorithm to reduce false positives is reported. The base YOLOv3 network was fine-tuned using a dataset composed of 28,576 images labelled with locations of 941 polyps that will be made public soon. In a frame-based evaluation using isolated images containing polyps, a general F_1 score of 0.88 was achieved (recall = 0.87, precision = 0.89), with lower predictive performance in flat polyps, but higher for sessile, and pedunculated morphologies, as well as with the usage of narrow band imaging, whereas polyp size < 5 mm does not seem to have significant impact. In a polyp-based evaluation using polyp and normal mucosa videos, with a positive criterion defined as the presence of at least one 50-frames-length (window size) segment with a ratio of 75% of frames with predicted bounding boxes (frames positivity), 72.61% of sensitivity (95% CI 68.99–75.95) and 83.04% of specificity (95% CI 76.70–87.92) were achieved (Youden = 0.55, diagnostic odds ratio (DOR) = 12.98). When the positive criterion is less stringent (window size = 25, frames positivity = 50%), sensitivity reaches around 90% (sensitivity = 89.91%, 95% CI 87.20–91.94; specificity = 54.97%, 95% CI 47.49–62.24; Youden = 0.45; DOR = 10.76). The object-tracking algorithm has demonstrated a significant improvement in specificity whereas maintaining sensitivity, as well as a marginal impact on computational performance. These results suggest that the model could be effectively integrated into a CAD system.

Keywords Colorectal cancer · Polyp detection · Deep learning · Real time

1 Introduction

Colorectal cancer (CRC) is one of the biggest health issues, being the third most common type of cancer worldwide [1] and having a high mortality. The gold standard method for CRC screening is colonoscopy [2] and the establishment of population-based CRC screening programs is an efficient strategy to reduce CRC-associated mortality and incidence [3]. These screening programs are aimed at detecting and removing adenomatous polyps that can potentially result in CRC [4]. Also, early CRC diagnosis is known to increase the 5-year survival rate from 18 to 88.5% [5]. The adenoma detection rate (ADR) is a recommended quality measure of colonoscopy that represents the proportion of examinations performed by an endoscopist that detect one or more adenomas. It has been found that each 1% increase in ADR is associated with a 3% decrease in the risk of interval CRC [6]. The adenoma detection rates ranged from 7.4 to 52.5% and polyps are missed during colonoscopic examination at a rate that varies from 6 to 27% [6, 7]. Polyps can be missed due to several factors, including their characteristics (size and morphology), quality of bowel preparation, physicians' experience, or physicians' fatigue [8]. In addition, some studies revealed that having a second observer during endoscopy increases polyp detection rate (PDR) [9, 10], although it is not clear if the same applies to increasing ADR [11]. Given these circumstances, there has been a great interest in the development of computer-aided diagnosis (CAD) systems based on artificial intelligence to assist endoscopists in detecting polyps in colonoscopy images.

Extended author information available on the last page of the article

In recent years, deep learning (DL), as part of a broader family of machine learning (ML) methods based on artificial neural networks (ANN), has gained a lot of attention in the field of medical image analysis due to its superior performance in image classification when compared to previous techniques, being nowadays considered the state-of-the-art. In this regard, a recent review identified more than three-hundred studies using DL-based approaches in medical image analysis [12], and a meta-analysis published in 2019 demonstrated that the diagnostic performance of DL-based models is equivalent to that of health-care professionals [13]. Naturally, there has been a growing interest in the application of DL to the analysis of colonoscopy images for polyp detection and/or classification, and several reviews on this topic have been also published [14–18]. In fact, the first randomized clinical trials of CAD systems for polyp detection based on DL have been performed recently [11, 19–23]. These clinical trials demonstrated that such automatic polyp detection systems can detect polyps initially missed by endoscopists and can increase both PDR and ADR. Moreover, a recent meta-analysis studied these clinical trials, demonstrating that there is an increase in both PDR and ADR when AI-assisted colonoscopy is performed [24]. Although the cost–benefit ratio and the effect on the incidence of interval colorectal should be further investigated, these results demonstrate that DL is a promising technology for the development of automatic polyp detection systems that could effectively assist endoscopists during colonoscopy by providing them a second, unbiased opinion.

In this context, the PolyDeep project (http://www.polydeep.org) aims to create a CAD system capable of detecting colorectal polyps in real time and classifying them. This project has been developed by the authors of this work, who have created a new dataset of videos and images of colorectal polyps with samples donated selflessly by patients of the Digestive Service of the Complexo Hospitalario Universitario de Ourense (CHUO) (Ourense, Spain).

In this work, the first results of this project on training and evaluating a real-time automatic polyp detection system based on DL are reported. In the proposed model, a pre-trained You Only Look Once (YOLO) v3 model was used and fine-tuned with an in-house database of 28,576 polyp images from 941 different polyps. Polyp images were acquired under both white light (WL) and narrow band imaging (NBI). To reduce the false positive bounding boxes, an efficient object-tracking algorithm was implemented and evaluated.

2 State of the art

As stated in the introduction, there are several recent reviews [14–18] analysing novel approaches reporting DL-based systems for polyp detection [14]. Some of the published works only perform polyp detection without localization, meaning that they report systems aimed to predict whether there are one (or more) polyps in a given video frame, but without indicating the exact location of the polyp. Other systems also try to locate polyps in the frames, showing their positions with either a bounding box (a square or a circle) or a binary mask. The first task is usually called polyp localization, while the later polyp segmentation. Polyp localization or polyp segmentation systems are usually able to predict the locations of one or more polyps in the same frame.

For any system based on one of these approaches to be useful in clinical practice, it must be able to operate in real time and, therefore, it must process at least 25 frames per second. According to the updated information of our review, available at this GitHub repository (https://github.com/sing-group/deep-learning-colonoscopy) as of August 2021, only 13 out of 33 studies report their ability to operate in real time.

As noted in the available reviews, comparing the performance of the different studies is difficult due to the high degree of heterogeneity in several aspects. First, most of the studies use private databases of varying sizes, while others use one or more publicly available databases such as CVC-ClinicDB [25], CVC-ColonDB [26, 27], CVC-PolypHD [26, 27], ETIS-Larib [28], or the ASU-Mayo Clinic Colonoscopy Video dataset [29]. Even when the same performance metrics are used to evaluate those systems (*i.e.* sensitivity, predictive positive value or PPV, specificity, negative predictive value or NPV), some studies report frame-based metrics while others calculate polyp-based metrics. Another drawback is the fact that some studies report performances using datasets of manually selected images, usually having a higher quality than those frames found in routine colonoscopy videos. This selection bias may produce overestimated performance results that cannot be achieved in actual practice. The validation schemes also vary between studies, with some of them reporting only training performance or even lack of test datasets. Finally, some studies evaluate the performance of a single model, while others try different convolutional neural network (CNN) architectures and hyperparameter configurations to identify the most promising combination.

From the 33 studies listed in our repository, Lee et al. [30] reported the best overall performance on a public dataset (CVC-ClinicDB), with an F_1 of 0.94, a sensitivity of 90.2%, and a PPV of 98.2% (frame-based metrics). Four studies reporting polyp-based metrics using private datasets achieved sensitivities superior to 93% [31, 32]. From these, Misawa et al. [31] reported a PPV of 48% and a specificity of 40%, and Urban et al. [32] reported PPV values of 35% and 60%, associated with two different datasets where they achieved sensitivities of 100% and 93%, respectively.

From a technical perspective, multitudes of frameworks have been proposed for object detection following different strategies. These strategies fall in two main types: (i) region proposal based (RPB) frameworks, and (ii) regression/classification based (RCB) frameworks. The RPB frameworks divide the object detection task into an initial region proposal generation phase followed by a classification of the proposed regions into different object categories. On the other hand, the RCB frameworks generate categorized regions in a single step, treating the problem as a regular regression or classification problem [33]. The multi-stage composition of the RBP frameworks, which includes region proposal generation, feature extraction with CNN, classification, and bounding box regression, negatively affects the time required to process a frame, conditioning its application in real time [33]. Therefore, the use of frameworks following this strategy was discarded for this work, since real-time processing capability is a mandatory feature for a model that detects polyps during endoscopy. Regarding RCB frameworks, the two most relevant implementations are YOLO [34] and SSD [35]. Between these two alternatives, YOLO was finally used because of the performance and prediction time of the YOLOv3 version (the most recent at the time of starting our PolyDeep project), which seems to outperform SSD [36].

Most of the existing studies perform polyp localization using architectures like R-CNN (and its variants Fast/Faster R-CNN) [37–40], Single Shot MultiBox Detector (SSD) [41–43], and YOLO [30, 44–47]. Although these architectures allow performing simultaneous object detection (*i.e.* localization) and classification (*i.e.* determining the type of object detected), most of the studies train them on one class to perform only polyp localization. There are, however, exceptions to this. For instance, Liu et al. [39] trained a Faster R-CNN network to locate polyps and adenomas, while Tian et al. [48] and Ozawa et al. [49] trained their networks to differentiate between several polyp classes. On the other hand, works indicating the location of polyps as binary marks use encoder–decoder architectures for object segmentation like SegNet or Unet [50, 51]. Recently, Qadir et al. [52] reported a real-time polyp detection model based on an encoder–decoder architecture named MDeNetplus that converts binary masks into bounding boxes for polyp localization. Finally, Urban et al. [32] presented polyp localization as a regression problem, training different network architectures (VGG16, VGG19, and ResNet50) to predict the size and location of the bounding boxes. A recent work compared the performance of different architectures performing polyp detection [53], showing the inability of Faster R-CNN and RetinaNet to operate in real time (8 and 16 FPS respectively). In this benchmarking, YOLO networks achieved state-of-the-art performance, being able to work in real time (> 40 FPS).

Remarkably, many authors pick a state-of-the-art performing architecture and apply different post-processing techniques and algorithms to increase the polyp detection performance, typically considering temporal information with the aim of reducing false positives [54]. For instance, Qadir et al. [55] introduced a false positive reduction unit that exploits the temporal dependencies among frames in colonoscopy videos, integrating past and future frames when making a decision on a specific frame. Similarly, Xu et al. [54] proposed an inter-frame similarity correlation unit to both reduce false positive identifications and correct false negatives (i.e. correct missed identifications). In this line, we propose a post-processing algorithm to filter YOLOv3 predictions with the goal of reducing false positives.

3 Methods

3.1 YOLOv3 object detection model

YOLO models divide the image into an $S \times S$ grid, whose cells are assigned with the responsibility of detecting an object if its centre is within the boundaries of the cell. Each cell predicts k bounding boxes with a *confidence score* that combines the probability of the box containing an object and the accuracy of the bounding box, calculated as the intersection over union (IoU) between the predicted box and the ground truth. The location of the centre of each bounding box (<*x, y*> coordinates) is predicted as an offset relative to the bounds of the corresponding grid cell, while the size of the bounding box (*width* and *height*) is predicted relative to the image size. Regarding the class prediction, each cell predicts C class probabilities conditioned to the presence of an object in the cell. In summary, each bounding box consists of five values (x, y, *width*, *height*, and *confidence*), each cell consists of B bounding boxes and C class predictions, and, therefore, the size of the YOLO output layer is $S \times S \times (B * 5 + C)$.

During the last years, several versions of the YOLO framework have been published, each one introducing some modifications to the base ideas of the original YOLO network with the aim of improving both the performance and prediction time of the previous versions. Thus, for example, YOLOv2 (also known as YOLO9000) [56] introduces the concept of the anchor boxes, which are used as the reference for the bounding boxes location instead of the cell. Each cell has K associated anchor boxes, whose dimensions are determined by a k-means procedure. YOLOv2 also moves the class prediction from cell level to boundary box level, allowing the location of objects with different classes associated to the same cell. Therefore, the size of the YOLOv2 output layer becomes $S \times S \times B * (5 + C)$. The most relevant innovation of the next version, YOLOv3 [36], is the prediction at multiple scales by using anchor boxes from the last layer but also from two previous layers, achieving better results with objects of different sizes. Both YOLOv2 and YOLOv3 also replace the backbone network, evolving the original Darknet network used by YOLO.

Recently, and within a short period of time, the YOLOv4 [57] and YOLOv5 [58] networks have been published. However, it should be noted that these networks have been published by independent authors different from the author of the three original YOLO networks, who has decided to abandon their development. One of the consequences of this is that, contrary to what one might think, YOLOv5 is not an evolution of YOLOv4, but both are derived from YOLOv3. Of these two new versions, only YOLOv4 has received official support from the original author of YOLO.

Regarding the technical aspect, YOLOv4 makes use of techniques such as bag of freebies and bag of specials, among others, to improve performance and speed compared to YOLOv3. In addition, it also uses CSPDarknet53 as a backbone, which is a version of Darknet that uses Cross-Stage-Partial-connections (CSP) [59]. On the other hand, YOLOv5 natively reimplements YOLO for PyTorch [60], achieving better support and performance. In addition, it introduces several improvements, such as the use of a CSP backbone, the use of mosaic data augmentation and auto learning bounding box anchors.

As explained before, the YOLOv3 was selected to accomplish this work, part of the PolyDeep project that our research groups are developing. Specifically, the GluonCV 0.7.0 implementation for the Apache MXNet 1.4.1 framework was used. This framework was preferred over other popular alternatives, such as TensorFlow [61] or CNTK (Microsoft Cognitive Toolkit) [62], because it showed superior performance in several benchmarks consulted. GluonCV is an extension of Apache MXNet that provides implementations of state-of-the-art DL algorithms in computer vision. It also includes the Model Zoo library, which contains several pre-trained models. From these models, YOLOv3 was selected with a 416 × 416 input layer. Although GluonCV also provides YOLOv3 implementations with input layers of 320 × 320 and 608 × 608, the 416 × 416 implementation was selected because it attains the best trade-off between performance and prediction time [36].

Since the volume of data available for model development and evaluation was not very large, the use of transfer learning was considered necessary. For this reason, the YOLOv3 model pre-trained with the PASCAL VOC dataset [63] available in the Model Zoo of GluonCV was used. This dataset includes images from a general domain used in the PASCAL VOC 2007 and 2012 challenges. The original object classes of this model were discarded and replaced by a single one (*polyp*), reusing the weights of the *aeroplane* object class, which achieved best results as not reusing class weights in the preliminary model tests done.

3.2 Data augmentation

As is commonly known, DL models require a large amount of data to be trained, which can be difficult to obtain in medical domains due to the high costs associated to data annotation. To overcome this problem, several strategies have been proposed by the scientific community, being data augmentation one of the most commonly used.

Data augmentation encompasses various techniques that allow the generation of new images from those already available by applying, in general, relatively simple transformations. Apart from the obvious benefit of having more images available, data augmentation is interesting because it reduces overfitting, since using images with small alterations prevents the model from memorising them, thus achieving a more generalised model.

Data augmentation techniques can be classified into two categories, depending on the strategy followed: data warping and oversampling. Data warping techniques alter an image without changing its label, applying alterations such as colour or geometric transformations, image rotation, image flipping, among others. On the other hand, oversampling techniques generate new synthetic images from those available. This includes mixing images, feature space augmentations, and generative adversarial networks [64].

In a previous review done by Nogueira-Rodríguez et al. [14], it was observed that the transformations most commonly used in polyp detection studies using DL were image rotation and mirroring, although translation, cropping, scaling and zooming modifications were also used. In this work, the default data augmentation pipeline implemented in the GluonCV library for YOLOv3 was used (class `gluoncv.data.transforms.presets.yolo.YOLO3DefaultTrainTransform`), which includes several of the aforementioned transformations. This pipeline was only modified to eliminate the initial random colour distortion step, since it produced worse results in preliminary tests. Therefore, the data augmentation pipeline finally used during the proposed model training consist of the following sequential steps:

1. *Random image expansion*: the size of the image is expanded up to four times the original size with a probability of 0.5. The new space added is filled with the colour (0.485, 0.456, 0.406) in the RGB colour model.
2. *Random cropping with bounding box constraints*: the image is cropped using the target bounding box as reference. The size ratio between the cropped region and the original image ranges from 0.3 to 1, with a maximum aspect ratio of 2. This operation also uses a constraint to discard image crops that have an IoU with the bounding box under 0.9, 0.7, 0.5, 0.3, 0.1, or 0 (*i.e.* ignoring this constraint), selecting this value randomly. This data augmentation technique is taken from the training of the SSD network [35].
3. *Resizing with random interpolation method*: the image is resized to the size of the neural network input (*i.e.* 416 × 416, see Sect. 3.1). The interpolation method is selected at random from among nearest neighbours interpolation, bilinear interpolation, area-based (resampling using pixel area relation), bicubic interpolation over 4 × 4 pixel neighbourhood, and Lanczos interpolation over 8 × 8 pixel neighbourhood.
4. *Random horizontal flip*: the image is horizontally mirrored with a probability of 0.5.

Figure 1 shows the results of applying this data augmentation process to a sample polyp image. Although for illustration purposes in this example all the transformations were applied, it must be taken into account that during model training they are randomly activated and, therefore, some transformations may not be applied.

3.3 Object-tracking algorithm

In order to filter out isolated bounding boxes in video predictions (*i.e.* those that appear only during a few frames), an efficient object-tracking algorithm was developed as a candidate bounding box (CBB) post-filter. As can be seen in Fig. 2, this algorithm is integrated into the network through a module downstream of the object detection network (YOLOv3 in this work), which is responsible for applying the algorithm to filter the network output, maintaining the information needed for the algorithm from the previous frames and manipulating the image to generate the final output.

The algorithm is run on every video frame after CBBs were produced by the neural network, and selects those CBBs whose average of the max IoU found with CBBs on previous frames (up to a specified window size) is above an specified threshold. The basic idea is that only a bounding box, which is consistently produced in similar positions in a set of frames, should be considered. In this sense, confirmed bounded boxes are those with a high average IoU with a bounding box found on each previous frame.

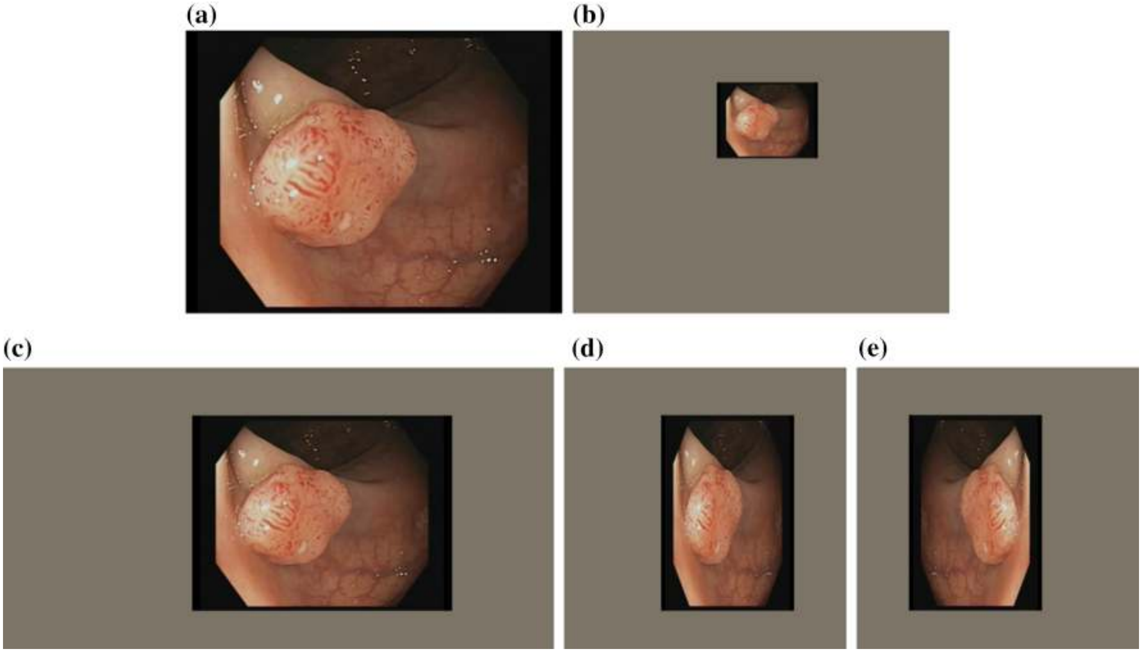

Fig. 1 Example of the data augmentation steps applied to a sample polyp image. The steps, which are listed in order of application, are: **a** original image, **b** random image expansion, **c** random cropping with bounding box constraints, **d** resizing with random interpolation method, and **e** random horizontal flip

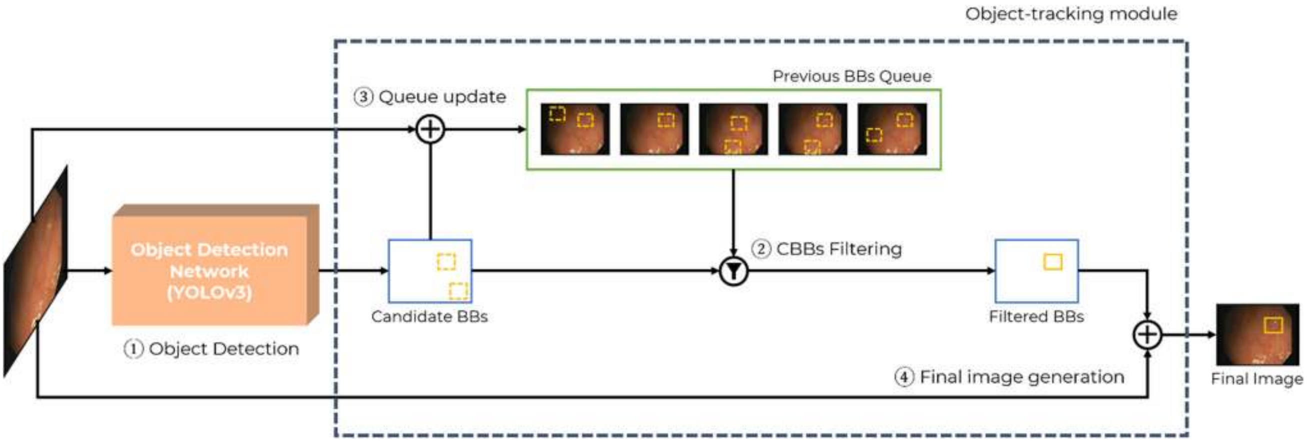

Fig. 2 Representation of the network architecture in which an object-tracking module has been added for filtering the CBBs proposed by the object detection network

Figure 3 shows an example of the proposed algorithm working with a window size of three and a threshold for the average of max IoU of 0.5, where the CBB 1 in the current frame will pass the filter criterion and the CBB 2 does not.

Figure 4 shows the pseudocode of the object-tracking algorithm. CBBs (candidateBBs) in the current frame are filtered out producing a subset of them (filteredBBs), by taking into account all CBBs in the previous frames (previousBBsQueue, of size windowSize) for computing their max IoUs average, which must be above the specified threshold (averageIoUThreshold).

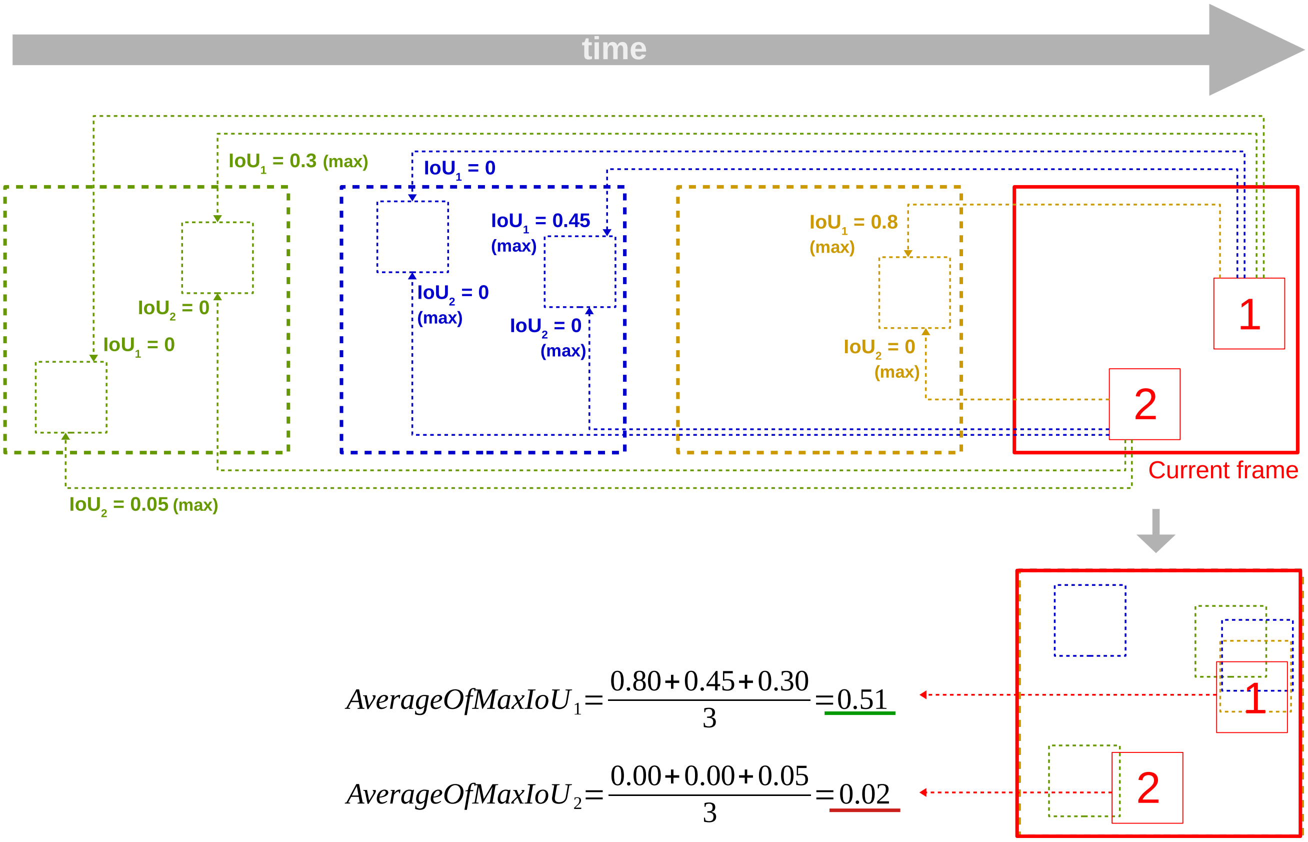

Fig. 3 Example of the object-tracking algorithm filtering candidate bounding boxes (CBBs). The current frame has two CBBs (1 and 2). A window of the previous three frames is used in order to compare current CBBs with previous ones via IoU to finally compute, for each CBB, the average of the max. IoU found on each previous frame. Given a threshold of 0.5, CBB box 1 passes the filter, whereas CBB 2 does not

```
ALGORITHM objectTrackingFilter
in: candidateBBs,
in: averageIoUThreshold,
in: windowSize,

in-out: previousBBsQueue,

out: filteredBBs

BEGIN
   filteredBBs = {}

   for BB in candidateBBs do
      sumIoU <- 0

      for f from 0 to windowSize - 1 do
         maxIoU <- 0

         for previousBB in previousBBsQueue[f] do
            IoU = IoU(previousBB, BB)
            if IoU > maxIoU then
               maxIoU <- IoU
            endif
         done

         sumIoU <- sumIoU + maxIoU
      done

      averageIoU <- sumIoU / windowSize

      if averageIoU > averageIoUThreshold
         filteredBBs <- filteredBBs U {BB}
      endif
   done

   pull(previousBBsQueue)
   push(previousBBsQueue, candidateBBs)
END
```

Fig. 4 Pseudocode of the object-tracking algorithm that filters candidate bounding boxes (candidateBBs) based on the average of the max. IoU (averageIoU) with the bounding boxes of each previous frame (previousBBsQueue) up to a window size (windowSize)

As can be seen, to do this, the algorithm iterates over the list of CBBs (candidateBBs) looking for the BB of each frame in the window with the one that has the highest IoU value. Bounding boxes on each frame are stored as a list in previousBBsQueue. The maximum IoU (maxIoU) of each frame is accumulated to calculate a final average value (averageIoU) that is compared to a threshold (averageIoUThreshold). In the case of exceeding it, the candidate BB will be added to the final list of bounding boxes (filteredBBs). Finally, the oldest frame is pulled from previousBBsQueue, and the list of CBBs (candidateBBs) for the current frame are pushed.

3.4 Pipeline development

In this work, pipelines used for model development and evaluation were implemented with Compi [65]. Compi is a framework for the development of pipeline-based, command-line applications, which provides developers with several useful features, such as the automatic generation of the user interface, the packaging of applications together with their dependencies in Docker images, or an easy way to distribute applications through the public Compi Hub repository [66].

In Compi, a computational pipeline is defined in an XML file. Within this XML file the pipeline is divided into several tasks, which, apart from containing the code, can define dependencies on other tasks, the interpreter to be used to execute the code (by default, Compi uses Bash) or the input parameters required by the task, among others. This division into tasks allows Compi to optimize the use of resources when the pipeline is executed, reducing the time required to complete the execution of the pipeline. In addition, Compi also allows partial executions, which include only a subset of the tasks or resuming an execution from a previous task after an execution failure.

The ability to optimise pipeline execution, together with the ease of managing pipeline dependencies, which facilitates its reuse and reproducibility, were the two main reasons for which this framework was chosen to develop the analysis pipeline. However, other additional features have also been used, such as the automatic generation of graphical representations of the pipeline.

4 Data

Colonoscopy videos were collected from study participants who underwent CRC screening colonoscopy from January 2018 to November 2019 in CHUO. People with positive results in faecal occult blood tests were invited to participate in the study and signed the informed consent. After that period, a dataset of 330 explorations from different patients was obtained. All colonoscopy images and videos were recorded using state-of-the-art endoscopy suites equipped with Olympus EVIS EXERA III CV-190 processors and Olympus 185 and 190 series colonoscopes (Olympus, Tokyo, Japan). Exploration videos were initially saved in MP4 format in the colonoscopy equipment, and later uploaded into a tool for further annotation and image extraction. These videos were annotated identifying the region of the video where 941 unique polyps are visible and linking these annotations to the corresponding polyp histology. The videos were also annotated to indicate the presence of other conditions, such as the existence of surgical instruments, on-screen information, water, and the imaging modality (WL or NBI). Normal mucosa video regions were also annotated. Currently, the necessary procedures to make this dataset publicly available through the biobank of the Instituto de Investigación Sanitaria Galicia Sur (IISGS) (https://www.iisgaliciasur.es/home/biobank-iisgs) are being carried out.

4.1 Image dataset

The annotated colonoscopy videos were processed in two ways to create the image dataset. First, a set of 16,691 images was systematically extracted from the polyp video regions at a rate of 1 frame per second (taking the middle frame of each second). Regions that contain several polyps, on-screen information, blurry images, or any other non-NBI tag were not considered. In addition, a second set of 11,885 manually selected images from the polyp video regions (with good quality and where the polyp is clearly visible) was pooled together with the previous set.

After this procedure, 28,576 polyp images (21,046 WL and 7530 NBI) were obtained. These images were annotated by a team of experienced endoscopists (JC, JH, MP, DR, LR, and ES) with the locations of the polyps (bounding boxes). All the endoscopists participated in the Galician CRC screening program [67]. The median number of colonoscopies performed annually by the endoscopists enrolled in the screening program (71 from 7 hospitals) was 278 (IQR 56–507) and the median adenoma detection rate was 65.3% (IQR 60.0–70.1%). This initial dataset of 941 polyps was split into the following partitions:

> *Development* dataset (Dataset 1). Labelled images from the 70% of polyps that are further split into:
>
> > Training dataset (Dataset 1.1). Labelled images from the 70% of the polyps to train the model (50% of total). Total: 460 polyps, 13,873 images (see Table 1).
> > Validation dataset (Dataset 1.2). Labelled images from the remaining 30% of the polyps to assess performance during model learning (20% of total). Total: 198 polyps, 6045 images (see Table 1).
>
> *Image testing* dataset (Dataset 2). Labelled images from the remaining 30% of the polyps not belonging to Dataset 1 containing a diverse set of polyps. Total: 283 polyps, 8658 images (see Table 1).

These splits were made at polyp level to ensure that all images of the same polyp go to the same split. Moreover, in order to keep partition sizes, in terms of number of images, near to that at polyp level, a stratified sampling (stratification by the number of images of each polyp) was carried out. Table 1 summarizes the distribution of both datasets, indicating the number of polyps and images in each partition, as well as the disaggregated counts by polyp histology, polyp morphology, polyp size and imaging technology.

4.2 Video dataset

Two video datasets were also created. First, a *video validation* dataset (Dataset 3) was created including (i) 426 polyp video segments, those used to extract the images of Dataset 1.2 (validation dataset), and (ii) 90 normal mucosa video segments, containing regions without polyps from a random sample of explorations. This dataset is used to perform the object-tracking algorithm tuning (Sect. 5.3). Second, a *video testing* dataset (Dataset 4) was created including (i) 628 polyp video segments, those used to extract the images of Dataset 2 (image testing dataset), and (ii) 171 normal mucosa video segments, containing regions without polyps from a random sample of explorations. This dataset is used for the polyp-based evaluations (Sect. 5.1).

5 Experimental methodology

5.1 Evaluation metrics

The evaluation of the polyp location network is carried out both at frame and at polyp level, by using testing datasets composed of images with located polyps, as well as video segments of polyp and normal mucosa, respectively.

For the frame-based evaluation, recall, precision, and the associated F_1 score are calculated. Briefly, recall is the proportion of true bounding boxes that are correctly detected, precision is the proportion of predicted bounding boxes that are also true bounding boxes, and F_1 is the harmonic mean of recall and precision.

Concretely, given a set of *true* bounding boxes, which are those annotated by the endoscopist in all frames, and a set of *predicted* bounding boxes, which are those bounding boxes produced by the network with a confidence score above a threshold (as it is explained later in Sect. 5.2), true positive, false positive, and false negative bounding boxes are defined as follows: a true positive bounding box ($\mathrm{TP}_{\mathrm{bb}}$) is a bounding box of a polyp that is predicted by the network with an IoU above 0.5 with a true bounding box. IoU is the ratio of the overlapping area divided by the area of the union of the real and the predicted bounding boxes. A false positive bounding box ($\mathrm{FP}_{\mathrm{bb}}$) is a predicted bounding box that does not overlap with any true bounding box or overlaps with true bounding boxes all with IoU below 0.5. A false negative bounding box ($\mathrm{FN}_{\mathrm{bb}}$) is a true bounding box that does not overlap with any predicted bounding boxes, or overlaps with predicted bounding boxes all with

Table 1 Summary of the image datasets. Totals in rows and columns are highlighted in bold

	Polyps				Images			
	Total	Train	Val	Test	Total	Train	Val	Test
Histology								
Adenoma (A + TSA + SSA)	**715**	355	153	207	**22,144**	10,813	4440	6891
Hyperplastic	**110**	52	21	37	**3850**	1879	977	994
Others	**116**	53	24	39	**2582**	1181	628	773
Morphology								
Flat	**221**	104	44	73	**7125**	3023	1432	2670
Sessile	**420**	212	87	121	**12,690**	6469	2727	3494
Pedunculated	**190**	92	44	54	**6371**	3184	1464	1723
Others (Depressed, Ulcerated, N/A)	**110**	52	23	35	**2390**	1197	422	771
Size								
$\geq$ 5 mm	**423**	218	86	119	**14,044**	6842	3078	4124
$<$ 5 mm	**484**	227	104	153	**14,110**	6884	2890	4336
N/A	**34**	15	8	11	**422**	147	77	198
Imaging								
WL	**921***	450*	193*	278*	**21,046**	10,194	4562	6290
NBI	**614***	295*	119*	200*	**7530**	3679	1483	2368
$\sum$	**941**	**460**	**198**	**283**	**28,576**	**13,873**	**6045**	**8658**

A Adenoma, *TSA* traditional serrated adenoma, *SSA* sessile serrated adenoma

*Number of polyps that contain at least one image with the corresponding imaging technology (the same polyp could have both images with different imaging technologies)

IoU below 0.5. Given these definitions, the following metrics are defined as:

$$\text{recall}_{\text{bb}} = \frac{\text{TP}_{\text{bb}}}{(\text{TP}_{\text{bb}} + \text{FN}_{\text{bb}})}$$

$$\text{precision}_{\text{bb}} = \frac{\text{TP}_{\text{bb}}}{(\text{TP}_{\text{bb}} + \text{FP}_{\text{bb}})}$$

$$F_{1\text{bb}} = \frac{2 * \text{recall}_{\text{bb}} * \text{precision}_{\text{bb}}}{(\text{recall}_{\text{bb}} + \text{precision}_{\text{bb}})}$$

In addition to these measures, Average Precision (AP) was used during model development to select the best model. AP is defined as the average precision for different confidence score thresholds and it is equivalent to the Area Under the Precision–Recall curve (AUPRC).

For the polyp-based evaluation, sensitivity, specificity, the associated Youden's index, as well the associated likelihood ratios and diagnostic odds ratio (DOR) are calculated. Briefly, sensitivity is the proportion of polyp videos marked as positive, whilst specificity is the proportion of normal mucosa videos marked as negative. Youden index combines sensitivity and specificity with the same weight (giving a score between 0 and 1), being 0 if the same frame positivity is obtained for both polyp and normal mucosa videos and 1 if the classification is perfect (100% sensitivity and specificity). Likelihood ratios also combine sensitivity and specificity. The positive likelihood ratio gives the ratio of the probability of a polyp video being marked as positive when compared to normal mucosa videos, where the highest possible value is desired. Conversely, the negative likelihood ratio gives the ratio of the probability of a polyp video being marked as negative when compared to normal mucosa videos, where a value near to zero is desired. Finally, DOR combines the previous likelihoods ratios, for which a high value is preferred.

For this polyp-based evaluation, two different criteria are defined to mark videos as positive:

> *Full video criterion*: a video is marked as positive if the proportion of the video frames containing predicted bounding boxes is above a given threshold, and marked as negative otherwise. For example, consider a video of five frames where predicted bounding boxes are present in the following layout: [*no, no, yes, yes, no*]. The proportion of positives is 2/5 and therefore, with a threshold of 50%, the video will be marked as negative.
> *Sliding window criterion*: a video is marked as positive if there is at least one segment of the video of size *w* (window) where at least a ratio *p* (frames positivity) of its frames contain one or more predicted bounding boxes above the confidence score threshold. If there is not such a segment, the video is marked as negative. For instance, considering the previous example (with predicted bounding boxes disposed as [*no, no, yes, yes, no*]), with *w = 3* and *p = 0.60*, the video will be marked as positive,

since at frame 4, the window has a frames positivity above 0.6 (window at frame 4 is [*no, yes, yes*]).

Given the previous criteria, a true positive video (TP_v) is a polyp video that is marked as positive, a true negative video (TN_v) is a normal mucosa video marked as negative, a false positive video (FP_v) is a normal mucosa video marked as positive, and a false negative video (FN_v) is a polyp video marked as negative. Given these definitions, the following metrics are defined as:

$$\mathrm{sensitivity_v} = \frac{\mathrm{TP_v}}{(\mathrm{TP_v} + \mathrm{FN_v})}$$

$$\mathrm{specificity_v} = \frac{\mathrm{TN_v}}{(\mathrm{TN_v} + \mathrm{FP_v})}$$

$$\mathrm{Youden_v} = \mathrm{sensitivity_v} + \mathrm{specificity_v} - 1$$

$$\mathrm{LR+} = \frac{\mathrm{sensitivity_v}}{(1 - \mathrm{specificity_v})}$$

$$\mathrm{LR-} = \frac{(1 - \mathrm{sensitivity_v})}{\mathrm{specificity_v}}$$

$$\mathrm{DOR} = \frac{\mathrm{LR+}}{\mathrm{LR-}}$$

Finally, for comparing proportions (sensitivity and/or specificity) for significant differences, 2-sample tests for equality of proportions with continuity correction (two-proportion z-tests) are used.

5.2 CNN model evaluation and selection

Model development is carried out by training the neural network with the training dataset (see Dataset 1.1 in Sect. 4.1). Training is performed iteratively for 50 cycles or epochs, where all training images plus augmented images generated with data augmentation techniques (see Sect. 3.2) are presented to the network in batches of eight images. After predicting each batch, network parameters are adjusted via error backpropagation with a learning rate of 0.001.

After each training epoch, the predictive performance of the neural network is assessed via the validation dataset (see Dataset 1.2 in Sect. 3.1). The model with the maximum AP (see Sect. 5.1) in the validation dataset is selected as the *best* model, which is then configured with the *confidence score threshold* established to the value that gives the maximum F_1.

5.3 Object-tracking algorithm tuning

The object-tracking algorithm described in Sect. 3.3 has two parameters that need to be adjusted: window size and average of max. IoU threshold. To do this, the performance of the object-tracking algorithm was evaluated on a grid of these two parameters (window size = {2, 5, 10, 15, 20}, average of max. IoU threshold = {0.05, 0.12, 0.25, 0. 50, 0.75}) using the video segments associated to the polyps in the validation dataset and additional video segments of normal-mucosa segments. The performance of each object-tracking configuration was evaluated as follows:

1. Annotate each video without applying the object-tracking filtering.
2. Annotate each video applying the object-tracking filtering.
3. Compute the percentage of bounding boxes removed by the object-tracking filtering in normal-mucosa and polyp videos, separately. This will give us: P_n (percentage of bounding boxes removed in normal-mucosa videos) and P_p (percentage of bounding boxes removed in polyp videos).
4. Compute the risk ratio as follows:

$$\mathrm{riskratio} = \frac{P_n}{P_p}$$

The risk ratio indicates the relative risk of a bounding box to be removed in a normal mucosa video with respect to that risk in a polyp video. Higher P_n values combined with lower P_p values correspond to higher risk ratios, which is what the object-tracking filtering is expected to do (remove more bounding boxes in normal-mucosa videos than in polyp videos). The combination of parameters with the highest risk ratio will be used to perform the model evaluation described in the next subsection.

5.4 Model evaluation

As explained before, two evaluations are carried out. First, a *frame-based* evaluation, where the objective is to correctly localize a polyp in a set of images containing a polyp (see Dataset 2 in Sect. 4.1). A frame-based evaluation is used to both select the best model and evaluate it in an independent set of images. Second, a *polyp-based* evaluation, where the objective is to determine the presence of polyps in a set of short videos divided into (i) *polyp videos* containing polyps in all frames, and (ii) *normal mucosa videos* without any polyps (see Dataset 4 in Sect. 4.2). A comparison of the performance with and without the activation of the object-tracking algorithm (see Sects. 3.3 and 5.3) is also presented.

5.5 Software pipeline

In this work, four different Compi pipelines were developed (Supplementary Material 1): (i) *dataset split* for development (training and validation datasets) and testing,

(ii) *model development*, (iii) *model evaluation* (with images), and (iv) *video annotation*. The source code of the Compi pipelines is available at this public repository: https://github.com/sing-group/polydeep-object-detection.

The *dataset split* pipeline (`ttv.xml`) is a preamble pipeline containing the steps necessary to prepare the dataset for the subsequent model development and evaluation. This pipeline starts with the *check-dataset* step, which is a control step that checks whether the dataset has been downloaded locally and, if not, alerts the user and ends the execution. Then, the *create-ttv* step analyses the dataset to generate training, validation, and test subsets based on the number of images of each polyp (see Sect. 4.1). As a result, this step generates three files with the polyp identifiers that correspond to each subset of the data.

The *model development* pipeline (`train.xml`) is responsible for the model development. This pipeline also starts with the *check-dataset* step, which is followed by another control step (*check-gpu*) that checks the sanity of the GPU memory, as a failure in it could cause erroneous results. After that, the *model-development* step performs the model training and validation, generating several files with the performance results. This step is followed by the *generate-plot-data* that post-processes previous result files in order to plot them in the three following steps: *plot-loss*, *plot-map*, and *plot-metrics*, which generate several plots that summarize the model evaluation results. Finally, a *cleanup* step removes some auxiliary files generated during the pipeline execution.

Then, the *model evaluation* pipeline (`test.xml`) is responsible for evaluating the model selected in the development phase with a new/unseen dataset of images. This pipeline has the same control steps as the *model development* pipeline (*i.e. check-dataset* and *check-gpu*) and a new control step (*check-neural-network*) to check that the trained model exists. Then, the *test* step evaluates the model with the test dataset created with the *dataset split* pipeline, generating a file with the performance results.

Finally, the video annotation pipeline (`video-annotation.xml`) allows applying the model to video segments of polyp and normal-mucosa regions to obtain annotated videos. This pipeline starts with the same control steps that the *model evaluation* pipeline. Then, this pipeline comprises three main stages: (i) download videos (*download-polyp-videos* and *download-normal-mucosa-videos*), (ii) extract video segments (*extract-polyp-segments* and *extract-normal-mucosa-segments*), and (iii) use the model to detect polyps in each frame of the video segments (*predict-polyp-segments* and *predict-normal-mucosa-segments*). A final a *cleanup* step removes some auxiliary files generated during the pipeline execution.

6 Results

6.1 CNN model evaluation and selection

As explained in Sect. 5.2, the neural network was trained for 50 epochs. AP was recorded for both training and validation datasets after each epoch. Figure 5 shows the learning curve of the training process. The best model is achieved at the 37th epoch with the maximum AP of 0.9201. For this model, the confidence threshold that maximizes $F_{1\ \mathrm{bb}}$ is 0.1906 ($F_{1\ \mathrm{bb}} = 0.9085$, recall$_{\mathrm{bb}}$ = 0.9047, precision$_{\mathrm{bb}}$ = 0.9124). This corresponds to the model selected for the evaluation with the testing datasets.

6.2 Object-tracking algorithm tuning

The performance of the object-tracking algorithm was evaluated on a grid of parameters (window size = {2, 5, 10, 15, 20}, average of max. IoU threshold = {0.05, 0.12, 0.25, 0. 50, 0.75}) using the video segments associated to the polyps in the validation dataset and additional video segments of normal-mucosa. Table 2 presents the risk ratio of each parameter combination (see Sect. 5.3). The highest risk ratio is achieved with a window size of 5 and an average of max. IoU threshold of 0.12. In this case, the mean bounding box removal with object tracking in normal-mucosa videos is 45%, while in polyp videos is 8.52%.

6.3 Model evaluation

6.3.1 Frame-based evaluation

For the frame-based evaluation on polyp detection, an overall recall, precision, F_1 and AP of 0.8720, 0.8901, 0.8810, and 0.8880 were obtained, respectively. Performance by subsets of polyps, according to their morphology, histology, size, and imaging technology was also measured (Table 3).

Regarding histology, the obtained $F_{1\ \mathrm{bb}}$ score is similar for both adenoma and hyperplastic subtypes (0.8830 and 0.8949, respectively). Regarding morphology, there is more variation, with the F_1 score obtained when testing with images of polyps of flat morphology being much lower (0.7779) than the average. However, images of pedunculated polyps were more accurately predicted, achieving the highest F_1 score of 0.9453. Small polyps seem not to be more difficult to detect, since the network performance was even better in smaller ones than in bigger ones (F_1 score of 0.8946 and 0.8653, respectively). Finally, regarding imaging technologies, images with NBI activated were slightly better classified (F_1 score of 0.9031) than those with WL (F_1 score of 0.8725).

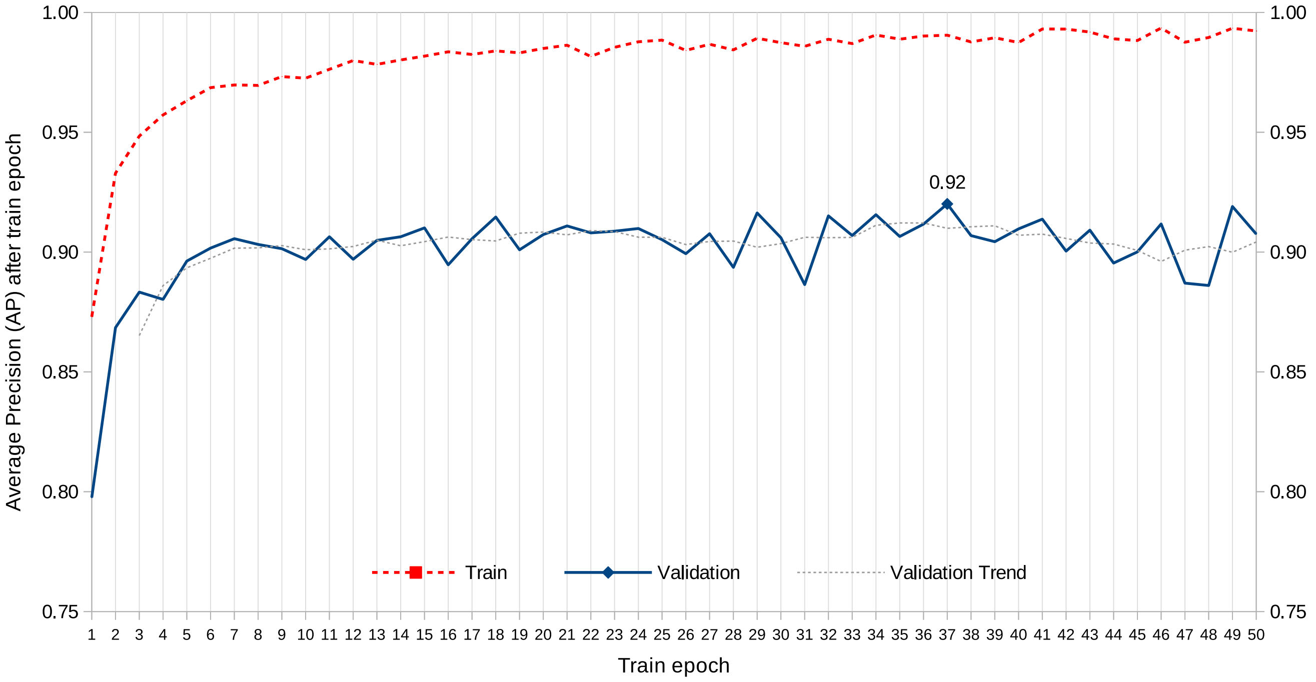

Fig. 5 Learning curves of the training process. *X*-axis shows the epoch and *Y*-axis the AP metric for both training and validation datasets. Epoch number 37 is the one with the best AP in the validation dataset (AP = 0.9201). Validation trend is the average of the five values around the point

Table 2 Risk ratios calculated after applying the object-tracking algorithm with different parameter combinations. The highest risk ratio (5.28) is highlighted in bold

		window size				
		2	5	10	15	20
average of max. IoU threshold	0.05	3.15	2.55	1.96	1.67	1.49
	0.12	4.84	**5.28**	4.43	3.92	3.52
	0.25	4.49	3.88	3.12	2.69	2.4
	0.50	3.15	2.55	1.96	1.67	1.49
	0.75	2.01	1.56	1.29	1.18	1.12

6.3.2 Polyp-based evaluation

For the polyp-based evaluation, where network predictions are obtained for each frame of polyp and normal mucosa videos, performance measures are obtained applying the two predefined criteria to identify positive videos (see Sect. 5.1). Moreover, metrics under activation of the object-tracking algorithm (Sects. 3.3 and 5.3) are also reported.

Results applying the full video criterion, where a proportion of the video frames with predicted bounding boxes is needed to be marked as positive, are shown in Table 4. Different parameters for this definition, *i.e.* different proportion thresholds, were tested.

As it can be seen in Table 4, sensitivity ranges from 80.41 to 99.36% without object tracking, and from 71.02 to 98.73% when it is used. Specificity ranges from 20.47 to 99.42% without object tracking, and from 52.63 to 99.42% when it is used. Under this definition of positive video, sensitivity is affected by the activation of object tracking as the criteria becomes more stringent, whereas specificity increases in general, which is an expected behaviour. Best performances attending to Youden index are found for the definition if a threshold of 50% is used, for both without

Table 3 Performance metrics for the overall image test dataset, for different subsets attending to the histology, morphology, size of the polyp, as well as for the imaging technology used: *WL* white light, *NBI* narrow band imaging

	recall	precision	F_1
Overall	0.8720	0.8901	0.8810
Histology			
Adenoma	0.8764	0.8898	0.8830
Hyperplastic	0.8742	0.9167	0.8949
Morphology			
Flat	0.7779	0.8429	0.8091
Sessile	0.9018	0.8952	0.8985
Pedunculated	0.9489	0.9418	0.9453
Size			
≥ 5 mm	0.8637	0.8669	0.8653
< 5 mm	0.8784	0.9115	0.8946
Imaging			
WL	0.8593	0.8861	0.8725
NBI	0.9058	0.9005	0.9031

and with object tracking configurations, reaching values of 0.90 (Sens = 93.79% 95; Spec = 96.49%) and 0.89 (Sens = 89.97%, Spec = 98.83%), respectively, being differences in sensitivity statistically significant (p-value < 0.05). The maximum sensitivity is reached when the positive criterion is less stringent (*i.e.* minimum ratio of frames with predicted bounding boxes of 15%), obtaining a similar sensitivity (p-value = 0.3842) around 99% in both object tracking configurations, but with a significantly (p-value < 0.001) better specificity when using object tracking (52.63% vs. 20.47%).

As an illustration of the polyp-based test and the full video criterion, Fig. 6 shows the clear affinity of the network to produce locations in polyp videos in comparison to normal mucosa videos. Although the object-tracking filter, which removes candidate bounding boxes, decreases the frequency of bounding boxes in videos, this decrease is different with respect to the type of video: 51% of candidate bounding boxes are removed in normal mucosa videos, whereas only 9% of bounding boxes are removed in polyp videos, giving a risk ratio of 5.65.

Results applying the sliding window criterion, where a segment containing a minimum proportion of bounding boxes is needed to mark a video as positive, are shown in Table 5. Different parameters for this definition, *i.e.* different frames positivity thresholds and window sizes, were tested.

As it can be seen in Table 5, sensitivity ranges from 74.52 to 90.13% without object tracking, and from 72.61 to 89.81% when it is used. Specificity ranges from 33.33 to 78.36% without object tracking, and from 54.97 to 83.04% when it is used. Whereas sensitivity is not affected by the activation of object tracking, specificity increases in general, which is an expected behaviour. Best performances attending to Youden index are found for the definition of positive if a window size of 50 frames (2 s) and a frames positivity threshold of 75% is used, for both without and with object tracking configurations, reaching values of 0.53 (Sens = 74.52%, Spec = 78.36%) and 0.55 (Sens = 72.61%, Spec = 83.04%), respectively. No statistically significant differences were found for sensitivity, whereas specificity was significantly higher with object tracking in two of four configurations. The maximum sensitivity is reached when the positive criterion is less stringent (*i.e.* window size of 25 and frames positivity of 50%), reaching

Table 4 Performance metrics for video classification attending to the usage of object tracking and under different parameters of the full video criterion

Threshold (%)	Without object tracking			With object tracking (window: 5, mean IoU $>$ 0.12)		
	15	50	75	15	50	75
Sens (%)	99.36	93.79*	80.41***	98.73	89.97	71.02
95% CI	(98.37–99.75)	(91.62–95.42)	(77.13–83.33)	(97.51–99.35)	(87.37–92.08)	(67.35 to 74.43)
Spec (%)	20.47	96.49	99.42	52.63***	98.83	99.42
95% CI	(15.10–27.13)	(92.56–98.38)	(96.76–99.90)	(45.17–59.98)	(95.84–99.68)	(96.76–99.90)
Youden	0.20	0.90	0.80	0.51	0.89	0.70
LR+	1.25	26.73	137.51	2.08	76.92	121.44
LR−	0.03	0.06	0.20	0.02	0.10	0.29
DOR	40.15	415.32	697.97	86.11	757.82	416.59

Threshold: minimum ratio of frames that should contain bounding boxes in the video to be marked as positive. p-values to test for significance for greater values when comparing object tracking activation are marked with '*' (p-value < 0.05), '**' (p-value < 0.01), or '***' (p-value < 0.001)

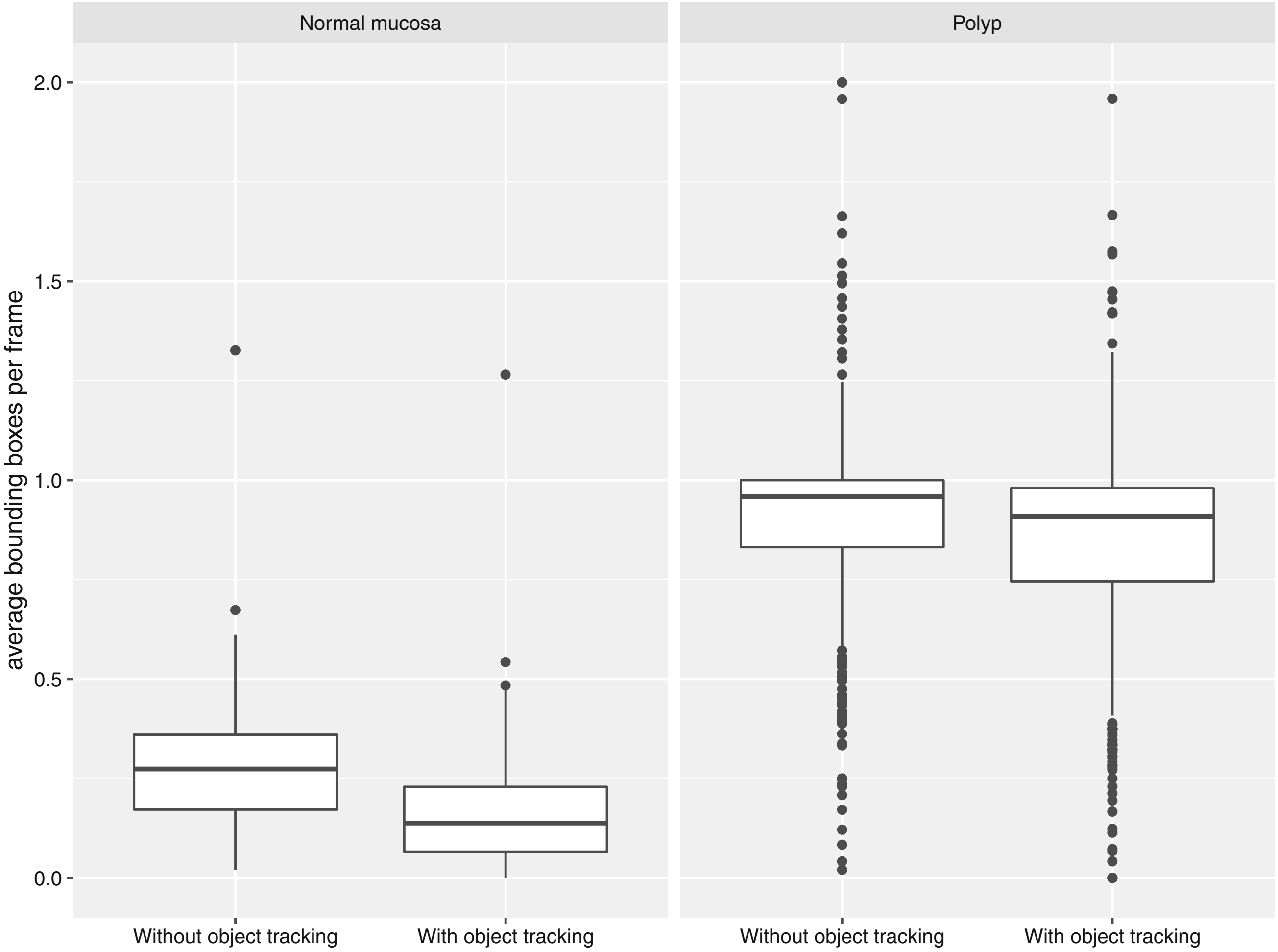

Fig. 6 Average bounding boxes per frame under different conditions (without and with object-tracking filtering, and with normal mucosa and polyp videos)

a similar sensitivity around 90% (p-value = 0.9251) in both object tracking configurations, but with significantly greater specificity (p-value < 0.01) when using object tracking (54.97% vs. 33.33%).

As an illustration of the polyp-based test and the sliding window criterion, Fig. 7 shows where predicted bounding boxes are generated (without object tracking in grey, object tracking activated in red) across two example videos, one with polyp (top) and another of normal mucosa (bottom) (see videos in Supplementary Material 2 and 3). As it can be seen, the polyp video contains more frames with predicted bounding boxes than the normal mucosa video. Lines show the count of frames with predicted bounding boxes across a sliding window of the last 25 frames without (grey) and with (red) object-tracking filtering. The polyp video contains several continuous regions where all 25 last frames contain predicted bounding boxes, whereas the normal mucosa video does not contain any segment of 25 frames where all of them contain predicted bounding boxes. Moreover, the red line tends to be more decoupled from the grey line in normal mucosa videos than in polyp videos, showing the affinity of the object-tracking to remove boxes in normal mucosa videos, as it was also pointed out by the risk-ratio (Sect. 6.2).

Similarly to the previous analysis, but by taking all the testing videos, Fig. 8 shows the distribution of the 25-frames-length video segments, according to the number of their 25 frames that contain predicted bounding boxes, comparing polyp videos against normal mucosa videos. As it is shown in 8, video segments whose 25 frames contain predicted bounding boxes are the most frequent segments in polyp videos, whereas segments with *none* of their 25 frames containing predicted bounding boxes are the most frequent in normal mucosa videos.

Finally, Fig. 9 shows representative examples of polyp identifications. The upper row corresponds to valid identifications (true positives) of sessile hyperplastic, flat adenoma, and pedunculated adenoma polyps. The bottom row shows false positive identifications in different normal-mucosa areas.

Table 5 Performance metrics for video classification attending to the usage of object tracking and under different parameters of the sliding window criterion

	Without object tracking				With object tracking (window: 5, mean IoU > 0.12)			
PT	50%		75%		50%		75%	
WS	25	50	25	50	25	50	25	50
Sens (%)	90.13	77.55	89.17	74.52	89.81	76.91	88.38	72.61
95% CI	(87.79–92.46)	(74.12–80.64)	(86.74–91.60)	(70.97–77.77)	(87.20–91.94)	(73.46–80.04)	(85.63–90.65)	(68.99–75.95)
Spec (%)	33.33	57.89	64.33	78.36	54.97***	69.00*	66.08	83.04
95% CI	(26.70–41.00)	(50.40–65.04)	(56.90–71.12)	(71.60–83.87)	(47.49–62.24)	(61.72–75.46)	(58.70–72.75)	(76.70–87.92)
Youden	0.23	0.35	0.53	0.53	0.45	0.46	0.54	0.55
LR+	1.35	1.84	2.50	3.44	1.99	2.48	2.60	4.28
LR−	0.30	0.39	0.17	0.33	0.18	0.33	0.17	0.33
DOR	4.56	4.75	14.85	10.59	10.76	7.41	14.81	12.98

PT frames positivity threshold: minimum ratio of frames in a video segment that should contain bounding boxes to be marked as positive, *WS* Window Size: minimum length in frames of the video segment with a bounding box ratio above PT to be marked as positive. *LR+* positive likelihood ratio, *LR−* negative likelihood ratios, *DOR* diagnostic odds ratio = LR+/LR−

p-values to test for significance for greater values when comparing object tracking activation are marked with '*' (p-value < 0.05), '**' (p-value < 0.01), or '***' (p-value < 0.001)

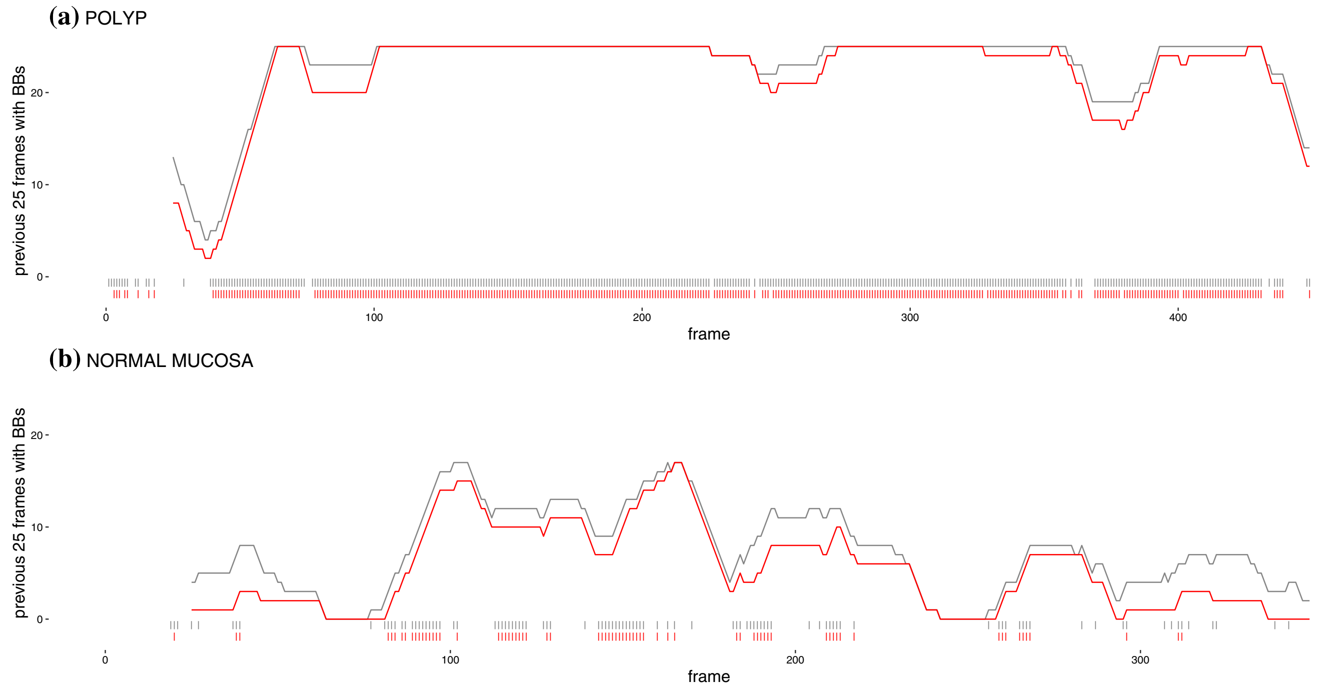

Fig. 7 Prediction over a polyp video (top) and a normal mucosa video (bottom). Bars show frames where there are predicted bounding boxes and lines show the number of the last 25 frames that contain bounding for both without (grey) and with (red) object-tracking filtering

6.4 Performance

We have tested a video annotation in a desktop computer with a 3.4 GHz CPU (AMD Ryzen 5 2600), 16 Gb of RAM memory, a NVIDIA GeForce RTX 2080 Ti 11 Gb GPU, for polyp detection, and a NVIDIA GeForce GTX 1050 Ti GPU, as main GPU for the operative system. Table 6 shows the average times to carry out several steps:

Time 1: network prediction, which is the time consumed by the neural network to predict a frame.

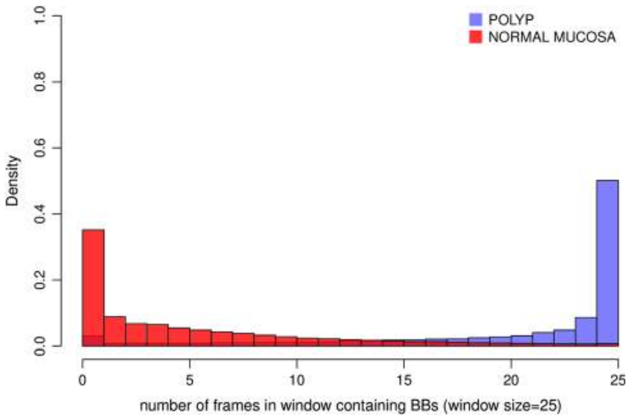

Fig. 8 Distribution of video segments by the number of frames containing predicted bounding boxes across all the testing video dataset

Time 2: frame processing, which includes the object tracking execution, if active, as well as painting the frame and the predicted bounding boxes on screen.

Time 3: input/output, which includes frame read and write.

As it can be seen in Table, the overall per-frame time including all tasks (last row in Table) is around 0.041 s, which yields a frame rate of about 24 frames per second, sufficient to be considered real time. Object tracking does not seem to affect performance significantly (+ 2% of the time consumed).

7 Discussion

The ability to operate in real time is mandatory for any polyp detection system to be useful in clinical practice. Therefore, Fig. 10 shows the F_1 scores for polyp detection only for those recent studies based on DL that can operate in real time, along with the results obtained in the present study. As it can be seen, the results here reported for the frame-based evaluation (F_1 = 0.88) are very close to the third one by Wang et al. [50] (F_1 = 0.91). Remarkably, the authors of such study have been the first ones in performing prospective clinical trials with polyp detection systems [11].

In the frame-based analysis, recall was lower for flat polyps (0.78) compared to the recall for sessile and pedunculated polyps (0.90, and 0.95, respectively). This finding is in line with the results of Lee et al., who also obtained a lower recall for flat polyps using a private dataset [30]. In the present study, there are far less train images for flat polyps (3023) than for sessile and pedunculated (6469 and 3184, respectively), a fact that may be an explanation for the lower recall. Differences are not found regarding histology and polyp size. Recall is 0.87 for both adenomatous and hyperplastic polyps, and 0.86 and 0.88 for polyps with a size $\geq$ 5 mm and $<$ 5 mm, respectively.

After initial evaluation on image datasets, polyp detection systems must demonstrate their performance when facing colonoscopy videos, a closer scenario to real clinical

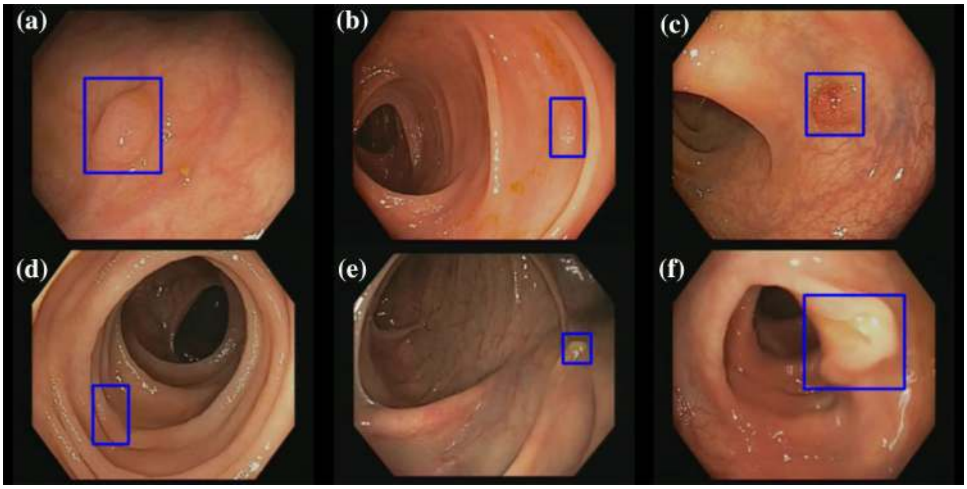

Fig. 9 Representative examples of polyp identifications: **A–B–C** True positives corresponding to sessile hyperplastic, flat adenoma, and pedunculated adenoma polyps; **D–E–F** false positive identifications in different normal-mucosa areas

Table 6 Average per-frame time in milliseconds to perform several computations

	Without object tracking	With object tracking
Time 1	22.21 ± 2.52	21.33 ± 2.50
Time 1 + Time 2	35.68 ± 1.53	36.01 ± 1.33
Time 1 + Time 2 + Time 3 (all)	41.10 ± 2.33	41.94 ± 2.40

NP network prediction, *OT* object tracking

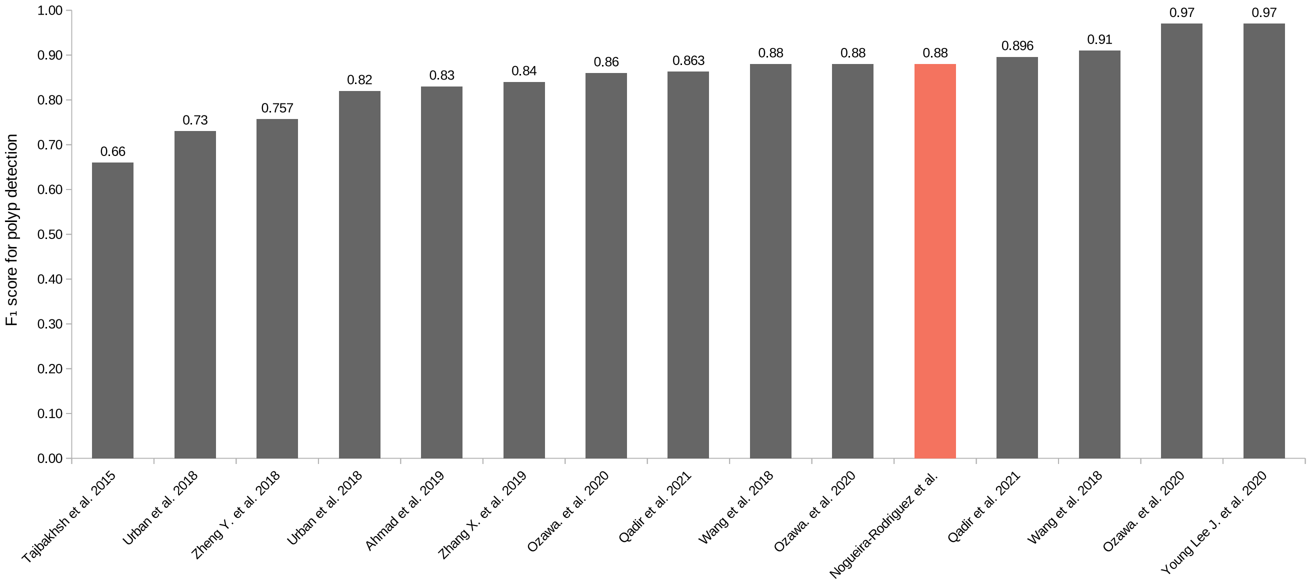

Fig. 10 F_1 scores for polyp detection for recent studies based on DL that can operate in real-time

settings. Wang et al. [50] evaluated their system with two video datasets: one containing short videos of polyps (similar to the video dataset here presented), and one containing 54 full-range unaltered colonoscopy videos. Similarly, Urban et al. [32] and Lee et al. [30] performed evaluations on unaltered colonoscopy videos. Misawa et al. [31] also used a set of 155 polyp-positive short videos and 391 polyp-negative videos. Similarly, a video testing dataset (Dataset 4) was created here by including polyp video segments at different moments of those 283 polyps belonging to Dataset 2 (image testing dataset), yielding a total of 628 polyp videos, along with 171 normal mucosa videos. To perform a polyp-based evaluation using colonoscopy videos as input, a criterion for polyp detection must be defined. Misawa et al. [31] defined polyp detection “as the system output over the cut-off value for $>75\%$ of the duration of each short video”. This criterion, also applied by Lee et al. [30] with a threshold of $>50\%$ and presented in Sect. 5.1 as full video criterion, is only applicable to short videos of a defined length. However, for its application in real time a temporal window size must be defined, since the length of the short polyp videos is not known in advance. This fact led to the sliding window criterion, where only a video segment of a defined length (*e.g.* 2 s) containing a minimum proportion of bounding boxes is needed to be considered as a positive polyp detection. This criterion can be applied in real time, considering the last frames in the temporal window size.

Using the first definition (full video criterion; Table 4), Misawa et al. [31] reported a sensitivity of 95% and a specificity of 40% (Youden = 0.34). The polyp detection system presented here obtained a sensitivity of 80.41% (95% CI 77.13- 83.33) and a specificity of 99.42% (95% CI 96.76–99.90) (Youden = 0.8). When the threshold is lowered to $>50\%$, a sensitivity of 93.79% (95% CI 91.62–95.42) is achieved, keeping specificity in 96.49% (95% CI 92.56–98.38) (Youden = 0. 9).

When the second definition (sliding window criterion; Table 5) is used, which is not comparable with other studies as it has not been previously described, sensitivity ranges from 74.52 to 90.13% and specificity ranges from 33.33 to 78.36%. As explained in the results section, the best performance according to the Youden index is found for the definition of positive if a window size of 50 frames (2 s) and a frames positivity threshold of 75% is used, and the maximum sensitivity is reached when the positive criterion is less stringent (*i.e.* window size of 25 and frames positivity of 50%). This criterion seems to be realistic and applicable to real-time polyp detection systems.

Nevertheless, to reduce false positive rates, an efficient object-tracking algorithm was implemented and evaluated. When object tracking is combined with the second definition of positive videos, sensitivity ranges from 72.61 to 89.81% and specificity ranges from 54.97 to 83.04% when it is used. As it can be seen, sensitivity is not affected by the activation of object tracking while specificity increases in general. In addition, the risk ratio analysis shows the clear affinity of the object-tracking algorithm to remove boxes on normal mucosa videos: 51% of candidate bounding boxes are removed in normal-mucosa videos, whereas only 9% of bounding boxes are removed in polyp videos, giving a risk ratio of 5.65. Therefore, the application of a post-processing object tracking filter seems to be beneficial for reducing the false positive rate. This approach has been applied in previous studies. For instance, Qadir et al. [55] added a false positive reduction unit to their network, able to exploit temporal dependencies between frames and correct the outputs. More recently, Lee et al. [30] applied a median filter to reduce false positive output frames in the YOLOv2 network predictions and Xu et al. [54] proposed an inter-frame similarity correlation unit to both reduce false positive identifications and correct false negatives.

The polyp detection system here presented was developed and evaluated using a private dataset, which is comparable in size to other private datasets used in similar studies [66]. Remarkably, this dataset includes both WL and NBI polyp-images, normal-mucosa images, and, unlike other public datasets for polyp detection, it includes additional information regarding polyp histology, morphology and size. Public datasets are essential for the development of new systems and, especially, to enable fair comparisons between polyp detection systems using the same testing datasets. Due to these reasons, the necessary procedures to make this dataset publicly available through the biobank of the Instituto de Investigación Sanitaria Galicia Sur (IISGS) (https://www.iisgaliciasur.es/home/biobank-iisgs) are currently being performed. This way, the dataset described and used here will be publicly accessible for other researchers, allowing not only the development and evaluation of polyp detection systems, but also systems for automatic polyp classification. The publication of this dataset will increase the public datasets available, which has been recently also increased with the addition of the PICCOLO dataset [68].

Finally, the work presented here has some limitations that should be acknowledged and that will guide the future work. First, as explained before, flat polyps are underrepresented, similarly to what happens in other datasets [68]. Since these types of polyps are less frequently found, special efforts (*e.g.* multicentre acquisitions) will be needed to increase their availability. Second, the lack of validation with external datasets. Although an independent test set of images (unseen during model development) was used for a frame-based evaluation, and the fact that a polyp-based evaluation on test videos was performed, it would be positive to assess the generalization capability of the polyp detection system here presented using a private dataset by using public testing datasets. Last of all, as it happens with every other polyp detection system reported, the usefulness must be assessed by developing clinical trials that can ultimately determine whether these systems can increase the adenoma or polyp detection rates of the endoscopists.

8 Conclusions

This work has described the development of a DL model for real-time polyp detection, which could be integrated, in the future, into a CAD system. YOLOv3 was selected as the base architecture for the development of this model due to its balance between performance and prediction time and complemented with an object-tracking filtering step able to reduce false positives.

Due to the low availability of public datasets [14], as part of this work, a database of images and videos of polyps taken during colonoscopies and manually annotated by expert endoscopists was created. This database, which currently contains 28,576 polyp images taken from the videos of 941 polyps, will soon be published through the biobank of the Instituto de Investigación Sanitaria Galicia Sur (IISGS).

The evaluation of this model was done using both a frame-based analysis and a polyp-based analysis. In polyp-based analysis, two different criteria were used to determine when polyp was effectively detected in a video. On the one hand, a criterion taken from other similar works [30, 31] was used, which is based on the percentage of positive frames in the whole video (full video criterion). On the other hand, a new criterion based on the percentage of positive frames in a sliding window (sliding window criterion) was also used, which is considered more rigorous and realistic.

Under the frame-based evaluation, the model obtained an F_1 score of 0.88, which is comparable to the results obtained by the best models developed in the field. In a polyp-based evaluation using polyp and normal mucosa videos, with a positive criterion defined as the presence of at least one 50-frames-length (window size) segment with a ratio of 75% of frames with predicted bounding boxes (frames positivity), our system with object-tracking activated, achieved 72.61% of sensitivity (95% CI 68.99–75.95) and 83.04% of specificity (95% CI 76.70–87.92) (Youden = 0.55, diagnostic odds ratio (DOR) = 12.98). When the positive criterion is less stringent (window size = 25, frames positivity = 50%), sensitivity reaches around 90% (sensitivity = 89.91%, 95% CI 87.20–91.94; specificity = 54.97%, 95% CI 47.49–62.24; Youden = 0.45; DOR = 10.76). Experiments also showed a general improvement when using object-tracking filtering.

Based on these results, and taking into account that the model is able to process a frame in 0.041 s, we consider that the developed model is valid to be tested in a real-time environment and integrated into a CAD system.

Regarding future work, several approaches can be explored. Firstly, more research in object-tracking filtering will be conducted, since it has demonstrated to be a promising way to improve the specificity while maintaining sensitivity of the detection model. On the other hand, the sensitivity could be improved by increasing the number of samples of the less frequent polyp histologies and morphologies. Finally, in order to improve model validation, the model can be tested with public datasets, such as ETIS-Larib [28], and it can also be tested under a clinical trial.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Funding for open access charge: Universidade de Vigo/CISUG. This work was partially supported by Ministerio de Economía, Industria y Competitividad, Gobierno de España under the scope of the PolyDeep project (DPI2017-87494-R), by Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding ED431C2018/55-GRC Competitive Reference Group, and through the individual scientific employment program-contract with Hugo López-Fernández (2020.00515.CEECIND). A. Nogueira-Rodríguez is supported by a pre-doctoral fellowship from Xunta de Galicia (ED481A-2019/299). SING group thanks the CITI (Centro de Investigación, Transferencia e Innovación) from the University of Vigo for hosting its IT infrastructure.

Declarations

Conflicts of interest The authors declare no conflict of interest.

Ethics approval The project under all patient samples of this paper collected was approved by the Pontevedra-Vigo-Ourense (Spain) Research Ethics Committee.

Availability of data and material Publication of the annotated images and videos database through the biobank of the Instituto de Investigación Sanitaria Galicia Sur (IISGS) (https://www.iisgaliciasur.es/home/biobank-iisgs) is planned.

Code availability The source code of the Compi pipelines is available at this public repository: https://github.com/sing-group/polydeep-object-detection.

Informed Consent Colonoscopy videos were collected from study participants who underwent CRC screening colonoscopy from January 2018 to November 2019 in Complexo Hospitalario Universitario de Ourense (CHUO). People with positive results in faecal occult blood tests were invited to participate in the study and sign the informed consent to use their colonoscopy videos in scientific publications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-021-06496-4.

References

1. Cancer today, https://gco.iarc.fr/today/online-analysis-table?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&group_cancer=1&include_nmsc=1&include_nmsc_other=1. Accessed 28 Dec 2020
2. US Preventive Services Task Force, Bibbins-Domingo K, Grossman DC, Curry SJ, Davidson KW, Epling JW, García FAR, Gillman MW, Harper DM, Kemper AR, Krist AH, Kurth AE, Landefeld CS, Mangione CM, Owens DK, Phillips WR, Phipps MG, Pignone MP, Siu AL (2016) Screening for colorectal cancer: US preventive services task force recommendation statement. JAMA 315:2564. https://doi.org/10.1001/jama.2016.5989.
3. Cubiella J, González A, Almazán R, Rodríguez-Camacho E, Zubizarreta R, Peña-Rey Lorenzo I (2020) Overtreatment in nonmalignant lesions detected in a colorectal cancer screening program: a cross-sectional analysis. Res Sq. https://doi.org/10.21203/rs.3.rs-113901/v1
4. Zauber AG, Winawer SJ, O'Brien MJ, Lansdorp-Vogelaar I, van Ballegooijen M, Hankey BF, Shi W, Bond JH, Schapiro M, Panish JF, Stewart ET, Waye JD (2012) Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 366:687–696. https://doi.org/10.1056/NEJMoa1100370
5. Wiegering A, Ackermann S, Riegel J, Dietz UA, Götze O, Germer C-T, Klein I (2016) Improved survival of patients with colon cancer detected by screening colonoscopy. Int J Colorectal Dis 31:1039–1045. https://doi.org/10.1007/s00384-015-2501-6
6. Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, Zauber AG, de Boer J, Fireman BH, Schottinger JE, Quinn VP, Ghai NR, Levin TR, Quesenberry CP (2014) Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 370:1298–1306. https://doi.org/10.1056/NEJMoa1309086
7. Ahn SB, Han DS, Bae JH, Byun TJ, Kim JP, Eun CS (2012) The miss rate for colorectal adenoma determined by quality-adjusted. Back-to-Back Colonoscopies Gut Liver 6:64–70. https://doi.org/10.5009/gnl.2012.6.1.64
8. Pannala R, Krishnan K, Melson J, Parsi MA, Schulman AR, Sullivan S, Trikudanathan G, Trindade AJ, Watson RR, Maple JT, Lichtenstein DR (2020) Artif Intell Gastrointest Endosc VideoGIE 5:598–613. https://doi.org/10.1016/j.vgie.2020.08.013
9. Aslanian HR, Shieh FK, Chan FW, Ciarleglio MM, Deng Y, Rogart JN, Jamidar PA, Siddiqui UD (2013) Nurse observation during colonoscopy increases polyp detection: a randomized prospective study. Off J Am Coll Gastroenterol ACG 108:166–172. https://doi.org/10.1038/ajg.2012.237
10. Lee CK, Park DI, Lee S-H, Hwangbo Y, Eun CS, Han DS, Cha JM, Lee B-I, Shin JE (2011) Participation by experienced endoscopy nurses increases the detection rate of colon polyps during a screening colonoscopy: a multicenter, prospective, randomized study. Gastrointest Endosc 74:1094–1102. https://doi.org/10.1016/j.gie.2011.06.033
11. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Zhang D, Li Y, Xu G, Tu M, Liu X (2019) Real-time automatic detection system increases colonoscopic

polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68:1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

12. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
13. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 1:e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
14. Nogueira-Rodríguez A, Domínguez-Carbajales R, López-Fernández H, Iglesias Á, Cubiella J, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D (2021) Deep neural networks approaches for detecting and classifying colorectal polyps. Neurocomputing 423:721–734. https://doi.org/10.1016/j.neucom.2020.02.123
15. Sánchez-Peralta LF, Bote-Curiel L, Picón A, Sánchez-Margallo FM, Pagador JB (2020) Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. Artif Intell Med 108:101923. https://doi.org/10.1016/j.artmed.2020.101923
16. Sánchez-Montes C, Bernal J, García-Rodríguez A, Córdova H, Fernández-Esparrach G (2020) Review of computational methods for the detection and classification of polyps in colonoscopy imaging. Gastroenterología y Hepatología (English Edition) 43:222–232. https://doi.org/10.1016/j.gastre.2019.11.003
17. Chao W-L, Manickavasagan H, Krishna SG (2019) Application of artificial intelligence in the detection and differentiation of colon polyps: a technical review for physicians. Diagnostics 9:99. https://doi.org/10.3390/diagnostics9030099
18. Azer SA (2019) Challenges facing the detection of colonic polyps: what can deep learning do? Medicina 55:473. https://doi.org/10.3390/medicina55080473
19. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, Lei L, Li L, Guo Z, Lei S, Xiong F, Wang H, Song Y, Pan Y, Zhou G (2020) Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 5:343–351. https://doi.org/10.1016/S2468-1253(19)30411-X
20. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Craviotto V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rosch T, Hassan C (2020) Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 159:512-520.e7. https://doi.org/10.1053/j.gastro.2020.04.062
21. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, Wang Z, Zhou W, An P, Huang X, Jiang X, Li Y, Wan X, Hu S, Chen Y, Hu X, Xu Y, Zhu X, Li S, Yao L, He X, Chen D, Huang L, Wei X, Wang X, Yu H (2020) Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol 5:352–361. https://doi.org/10.1016/S2468-1253(19)30413-3
22. Liu W-N, Zhang Y-Y, Bian X-Q, Wang L-J, Yang Q, Zhang X-D, Huang J (2020) Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. Saudi J Gastroenterol 26:13. https://doi.org/10.4103/sjg.SJG_377_19
23. Su J-R, Li Z, Shao X-J, Ji C-R, Ji R, Zhou R-C, Li G-C, Liu G-Q, He Y-S, Zuo X-L, Li Y-Q (2020) Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc 91:415-424.e4. https://doi.org/10.1016/j.gie.2019.08.026
24. Ashat M, Klair JS, Singh D, Murali AR, Krishnamoorthi R (2021) Impact of real-time use of artificial intelligence in improving adenoma detection during colonoscopy: a systematic review and meta-analysis. Endosc Int Open 09:E513–E521. https://doi.org/10.1055/a-1341-0457
25. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 43:99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007
26. Bernal J, Sánchez J, Vilariño F (2012) Towards automatic polyp detection with a polyp appearance model. Pattern Recogn 45:3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002
27. Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, Drozdzal M, Courville A (2017) A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng 2017:1–9. https://doi.org/10.1155/2017/4037190
28. Silva J, Histace A, Romain O, Dray X, Granado B (2014) Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int J CARS 9:283–293. https://doi.org/10.1007/s11548-013-0926-3
29. Tajbakhsh N, Gurudu SR, Liang J (2016) Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 35:630–644. https://doi.org/10.1109/TMI.2015.2487997
30. Lee JY, Jeong J, Song EM, Ha C, Lee HJ, Koo JE, Yang D-H, Kim N, Byeon J-S (2020) Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Sci Rep 10:8379. https://doi.org/10.1038/s41598-020-65387-1
31. Misawa M, Kudo S-E, Mori Y, Cho T, Kataoka S, Yamauchi A, Ogawa Y, Maeda Y, Takeda K, Ichimasa K, Nakamura H, Yagawa Y, Toyoshima N, Ogata N, Kudo T, Hisayuki T, Hayashi T, Wakamura K, Baba T, Ishida F, Itoh H, Roth H, Oda M, Mori K (2018) Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 154:2027-2029.e3. https://doi.org/10.1053/j.gastro.2018.04.003
32. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, Baldi P (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155:1069-1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037
33. Zhao Z, Zheng P, Xu S, Wu X (2019) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30:3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
34. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). pp 779–788. https://doi.org/10.1109/CVPR.2016.91
35. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) SSD: single shot MultiBox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016. Springer, Cham, pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
36. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv:1804.02767 [cs]
37. Sornapudi S, Meng F, Yi S (2019) Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. Appl Sci 9:2404. https://doi.org/10.3390/app9122404
38. Luo Y, Zhang Y, Liu M, Lai Y, Liu P, Wang Z, Xing T, Huang Y, Li Y, Li A, Wang Y, Luo X, Liu S, Han Z (2020) Artificial intelligence-assisted colonoscopy for detection of colon polyps: a

prospective. Randomized Cohort Study J Gastrointest Surg. https://doi.org/10.1007/s11605-020-04802-4
39. Liu X, Li Y, Yao J, Chen B, Song J, Yang X (2019) Classification of polyps and adenomas using deep learning model in screening colonoscopy. In: 2019 8th international symposium on next generation electronics (ISNE), pp 1–3. https://doi.org/10.1109/ISNE.2019.8896649
40. Wittenberg T, Zobel P, Rathke M, Mühldorfer S (2019) Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks. Curr Dir Biomed Eng 5:231–234. https://doi.org/10.1515/cdbme-2019-0059
41. Ma Y, Li Y, Yao J, Chen B, Deng J, Yang X (2019) Polyp location in colonoscopy based on deep learning. In: 2019 8th international symposium on next generation electronics (ISNE), pp 1–3 (2019). https://doi.org/10.1109/ISNE.2019.8896576
42. Misawa M, Kudo S, Mori Y, Cho T, Kataoka S, Maeda Y, Ogawa Y, Takeda K, Nakamura H, Ichimasa K, Toyoshima N, Ogata N, Kudo T, Hisayuki T, Hayashi T, Wakamura K, Baba T, Ishida F, Itoh H, Oda M, Mori K (2019) Tu1990 artificial intelligence-assisted polyp detection system for colonoscopy, based on the largest available collection of clinical video data for machine learning. Gastrointest Endosc 89:646–647. https://doi.org/10.1016/j.gie.2019.03.1134
43. Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, Hu W, Wang L, Duan H, Si J (2019) Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 14:e0214133. https://doi.org/10.1371/journal.pone.0214133
44. Ma Y, Chen X, Sun B (2020) Polyp detection in colonoscopy videos by bootstrapping via temporal consistency. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI), pp 1360–1363. https://doi.org/10.1109/ISBI45749.2020.9098663
45. Zhang R, Zheng Y, Poon CCY, Shen D, Lau JYW (2018) Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recogn 83:209–219. https://doi.org/10.1016/j.patcog.2018.05.026
46. Zheng Y, Zhang R, Yu R, Jiang Y, Mak TWC, Wong SH, Lau JYW, Poon CCY (2018) Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. In: 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, Honolulu, pp 4142–4145. https://doi.org/10.1109/EMBC.2018.8513337
47. Pacal I, Karaboga D (2021) A robust real-time deep learning based automatic polyp detection system. Comput Biol Med 134:104519. https://doi.org/10.1016/j.compbiomed.2021.104519
48. Tian Y, Pu LZCT, Singh R, Burt AD, Carneiro G (2019) One-stage five-class polyp detection and classification. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp 70–73. https://doi.org/10.1109/ISBI.2019.8759521
49. Ozawa T, Ishihara S, Fujishiro M, Kumagai Y, Shichijo S, Tada T (2020) Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. Therap Adv Gastroenterol. https://doi.org/10.1177/1756284820910659
50. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, Hu X, Liu P, Song Y, Zhang D, Yang X, Li L, He J, Yi X, Liu J, Liu X (2018) Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2:741–748. https://doi.org/10.1038/s41551-018-0301-3
51. Wang L, Chen R, Hu Y (2018) IDDF2018-ABS-0261 Polyp detection using an unet based model. Gut 67:A85–A85. https://doi.org/10.1136/gutjnl-2018-IDDFabstracts.182
52. Qadir HA, Shin Y, Solhusvik J, Bergsland J, Aabakken L, Balasingham I (2021) Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction. Med Image Anal 68:101897. https://doi.org/10.1016/j.media.2020.101897
53. Jha D, Ali S, Tomar NK, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P (2021) Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access 9:40496–40510. https://doi.org/10.1109/ACCESS.2021.3063716
54. Xu J, Zhao R, Yu Y, Zhang Q, Bian X, Wang J, Ge Z, Qian D (2021) Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit. Biomed Signal Process Control 66:102503. https://doi.org/10.1016/j.bspc.2021.102503
55. Qadir HA, Balasingham I, Solhusvik J, Bergsland J, Aabakken L, Shin Y (2020) Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video. IEEE J Biomed Health Inform 24:180–193. https://doi.org/10.1109/JBHI.2019.2907434
56. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6517–6525. https://doi.org/10.1109/CVPR.2017.690
57. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934 [cs, eess]
58. Nelson J, JUN 10, J.S., Read, 2020 4 Min: YOLOv5 is here. https://blog.roboflow.com/yolov5-is-here/. Accessed 9 Aug 2021
59. Wang C-Y, Mark Liao H-Y, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H (2020) CSPNet: a new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). IEEE, Seattle, pp 1571–1580. https://doi.org/10.1109/CVPRW50498.2020.00203
60. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems. Curran Associates, Inc.
61. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M. Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on operating systems design and implementation. USENIX Association, USA, pp 265–283
62. Seide F, Agarwal A (2016) CNTK: microsoft's open-source deep-learning toolkit. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, p 2135. https://doi.org/10.1145/2939672.2945397.
63. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vis 111:98–136. https://doi.org/10.1007/s11263-014-0733-5
64. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6:60. https://doi.org/10.1186/s40537-019-0197-0
65. López-Fernández H, Graña-Castro O, Nogueira-Rodríguez A, Reboiro-Jato M, Glez-Peña D (2021) Compi: a framework for portable and reproducible pipelines. PeerJ Comput Sci 7:e593. https://doi.org/10.7717/peerj-cs.593
66. Nogueira-Rodríguez A, López-Fernández H, Graña-Castro O, Reboiro-Jato M, Glez-Peña D (2021) Compi Hub: a public repository for sharing and discovering Compi pipelines. In: Panuccio G, Rocha M, Fdez-Riverola F, Mohamad MS, Casado-Vara R (eds) Practical applications of computational biology & bioinformatics, 14th international conference (PACBB 2020).

Springer, Cham, pp 51–59. https://doi.org/10.1007/978-3-030-54568-0_6.

67. Cubiella J, González A, Almazán R, Rodríguez-Camacho E, Fontenla Rodiles J, Domínguez Ferreiro C, Tejido Sandoval C, Sánchez Gómez C, de Vicente Bielza N, Lorenzo IP-R, Zubizarreta R (2020) pT1 colorectal cancer detected in a colorectal cancer mass screening program: treatment and factors associated with residual and extraluminal disease. Cancers 12:2530. https://doi.org/10.3390/cancers12092530.

68. Sánchez-Peralta LF, Pagador JB, Picón A, Calderón ÁJ, Polo F, Andraka N, Bilbao R, Glover B, Saratxaga CL, Sánchez-Margallo FM (2020) PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. Appl Sci 10:8501. https://doi.org/10.3390/app10238501

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alba Nogueira-Rodríguez[1,2] · Rubén Domínguez-Carbajales[3] · Fernando Campos-Tato[1] · Jesús Herrero[4] · Manuel Puga[4] · David Remedios[4] · Laura Rivas[4] · Eloy Sánchez[4] · Águeda Iglesias[4] · Joaquín Cubiella[4] · Florentino Fdez-Riverola[1,2] · Hugo López-Fernández[1,2,5,6] · Miguel Reboiro-Jato[1,2] · Daniel Glez-Peña[1,2]

✉ Daniel Glez-Peña
dgpena@uvigo.es

[1] CINBIO, Universidade de Vigo, Department of Computer Science, ESEI - Escuela Superior de Ingeniería Informática, 32004 Ourense, Spain

[2] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain

[3] Servicio de Sistemas y Tecnologías de la Información, Complexo Hospitalario Universitario de Ourense, Ourense, Spain

[4] Department of Gastroenterology, Complexo Hospitalario Universitario de Ourense, Instituto de Investigación Sanitaria Galicia Sur, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Ourense, Spain

[5] Instituto de Investigação E Inovação Em Saúde (I3S), Universidade Do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

[6] Instituto de Biologia Molecular E Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal