scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

A multi-centre polyp detection and segmentation dataset for generalisability assessment

Sharib Ali[1,2,3 ✉], Debesh Jha[4,5,6], Noha Ghatwary[7], Stefano Realdon[8,19], Renato Cannizzaro[8,9,19], Osama E. Salem[10], Dominique Lamarque[11], Christian Daul[12], Michael A. Riegler[4,5], Kim V. Anonsen[13], Andreas Petlund[14], Pål Halvorsen[4,15], Jens Rittscher[2,3], Thomas de Lange[14,16,17,19] & James E. East[3,18,19]

Polyps in the colon are widely known cancer precursors identified by colonoscopy. Whilst most polyps are benign, the polyp's number, size and surface structure are linked to the risk of colon cancer. Several methods have been developed to automate polyp detection and segmentation. However, the main issue is that they are not tested rigorously on a large multicentre purpose-built dataset, one reason being the lack of a comprehensive public dataset. As a result, the developed methods may not generalise to different population datasets. To this extent, we have curated a dataset from six unique centres incorporating more than 300 patients. The dataset includes both single frame and sequence data with 3762 annotated polyp labels with precise delineation of polyp boundaries verified by six senior gastroenterologists. To our knowledge, this is the most comprehensive detection and pixel-level segmentation dataset (referred to as *PolypGen*) curated by a team of computational scientists and expert gastroenterologists. The paper provides insight into data construction and annotation strategies, quality assurance, and technical validation.

Background & Summary

About 1.3 million new cases of colorectal cancer (CRC) are detected yearly in the world, with about 51% mortality rate, and CRC is the third most common cause of cancer mortality[1]. Approximately, 90% of CRCs result from slow transformation of the main benign precursors, adenomas or serrated polyps to CRC, but only a minority of them progress to CRC[2,3]. It is particularly challenging to assess the malignant potential for lesions smaller than 10 mm. As a consequence most detected lesions are removed with subsequent CRC mortality reduction[4]. The removal of the lesions depends also of an exact delineation of the boundaries to assure complete resection. If the lesions are detected and completely removed at a precancerous stage, the mortality is nearly null[5].

[1]School of Computing, University of Leeds, LS2 9JT, Leeds, United Kingdom. [2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, Oxford, United Kingdom. [3]Oxford National Institute for Health Research Biomedical Research centre, OX4 2PG, Oxford, United Kingdom. [4]SimulaMet, Pilestredet 52, 0167, Oslo, Norway. [5]Department of Computer Science, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019, Tromsø, Norway. [6]Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, USA. [7]Computer Engineering Department, Arab Academy for Science and Technology, Smart Village, Giza, Egypt. [8]Oncological Gastroenterology - Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, 2, 33081, Aviano, PN, Italy. [9]Department of Medical, Surgical and Health Sciences, University of Trieste, 34127, Trieste, Italy. [10]Faculty of Medicine, University of Alexandria, 21131, Alexandria, Egypt. [11]Université de Versailles St-Quentin en Yvelines, Hôpital Ambroise Paré, 9 Av. Charles de Gaulle, 92100, Boulogne-Billancourt, France. [12]CRAN UMR 7039, Université de Lorraine and CNRS, F-54010, Vandœuvre-Lès-Nancy, France. [13]Oslo University Hospital Ullevål, Kirkeveien 166, 0450, Oslo, Norway. [14]Augere Medical, Nedre Vaskegang 6, 0186, Oslo, Norway. [15]Oslo Metropolitan University, Pilestredet 46, 0167, Oslo, Norway. [16]Medical Department, Sahlgrenska University Hospital-Mölndal, Blå stråket 5, 413 45, Göteborg, Sweden. [17]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, 41345, Göteborg, Sweden. [18]Translational Gastroenterology Unit, Experimental Medicine Div., John Radcliffe Hospital, University of Oxford, OX3 9DU, Oxford, United Kingdom. [19]These authors contributed equally: Stefano Realdon, Renato Cannizzaro, Thomas de Lange, James E. East. ✉e-mail: s.s.ali@leeds.ac.uk

Dataset	Findings	Size	Availability
Kvasir-SEG[34]	Polyps	1000 images†	open academic
HyperKvasir[35]	GI findings including polyps	110,079 images and 374 videos	open academic
Kvasir-Capsule[36]	GI findings including polyps◇	4,741,504 images	open academic
CVC-ColonDB[37]	Polyps	380 images† †	by request•
ETIS-Larib Polyp DB[38]	Polyps	196 images†	open academic
EDD2020[15,39]	GI lesions including polyps	386 images	open academic
CVC-ClinicDB[40]	Polyps	612 images†	open academic
CVC-VideoClinicDB[41]	Polyps	11,954 images†	by request•
ASU-Mayo polyp database[42]	Polyps	18,781 images†	by request•
KID[43]	Angiectasia, bleeding, inflammations◇	2371 images, 47 videos	open academic•
Atlas of GI Endoscope[13]	GI lesions	1295 images	unknown•
El salvador atlas[12]	GI lesions	5071 video clips	open academic•♣
PolypGen (Ours)[18,44]	Multi-centre colon polyps	1537 images† & 2225 video sequence	open academic

Table 1. An overview of existing gastrointestinal (GI) lesion datasets including polyps: number of images or videos along with the availability type is provided. †Including ground truth segmentation masks ‡Contour ◇Video capsule endoscopy •Not available anymore. ♣Medical atlas for education with several low-quality samples of various GI findings.

Unfortunately, there is a considerable limitation related to various human skills[6,7] confirmed in a recent systematic review and meta-analysis demonstrating miss rates of 26% for adenomas, 9% for advanced adenomas and 27% for serrated polyps[8]. A thorough and detailed assessment of the neoplasia is essential to assess the malignant potential and the appropriate treatment. This assessment is based on size, morphology and surface structure. Currently, the Paris classification, prone to substantial inter-observer variation even among experts, is used to assess the morphology[9]. The surface structure classified by the Kudo pit pattern classification system or the Narrow-Band Imaging International Colorectal Endoscopic (NICE) classification system also help to predict the risk and degree of malignant transformation[10]. This classification system may to some extent also predict the histopathological classification into either adenomas, sessile serrated lesions (SSLs), hyperplastic polyps or traditional serrated adenoma (TSA)[10]. Unfortunately, these macroscopic classification systems are prone to substantial inter-observer variations, thus a high performing automatic computer-assisted system would be of great important both to increase detection rates and also reduce inter-observer variability. To develop such a system large segmented image databases are required. While current deep learning approaches has been instrumental in the development of computer-aided diagnosis (CAD) systems for polyp identification and segmentation, most of these trained networks suffer from huge performance gap when out-of-sample data have large domain shifts. On one hand, training models on large multi-centre datasets all together can lead to improved generalisation, but at an increased risk of false detection alarms[11]. On the other hand, training and validation on centre-based splits can improve model generalisation. Most reported works are not focused on multi-centre data at all. This is mostly because of the lack of comprehensive multi-centre and multi-population datasets. In this paper, we present the *PolypGen* dataset that incorporates colonoscopy data from 6 different centres for multiple patient and varied populations. Attentive splits are provided to test the generalisation capability of methods for improved clinical applicability. The dataset also is suitable for exploring federated learning and training of other time-series models. PolypGen can be pivotal in algorithm development and in providing more clinically applicable CAD detection and segmentation systems.

Although there are some publicly available datasets for colonoscopic single frames and videos (Table 1), lack of pixel-level annotations and preconditions applied for access of them pose challenges in its wide usability for method development. Many of these datasets are by request which requires approval from the data provider that usually takes prolonged time for approval and the approval is not guaranteed. Similarly, some datasets do not include pixel-level ground truth for the abnormality location which will cause difficulty in development or validation of CAD systems (*e.g.*, El salvador atlas[12] and Atlas of GI Endoscope[13]). Moreover, most of the publicly available datasets include limited number of images frames from one or a few centres only (*e.g.*, datasets provided in[14,15]). To this end, the presented *PolypGen* dataset is composed of a total of 8037 frames including both single and sequence frames. The provided comprehensive dataset consists of 3762 positive sample frames collected from six centres and 4275 negative sample frames collected from four different hospitals. The *PolypGen* dataset comprises of varied population data, endoscopic system and surveillance expert, and treatment procedures for polyp resections. A t-SNE plot for positive samples provided in the Fig. 1 demonstrates the diversity of the compiled dataset.
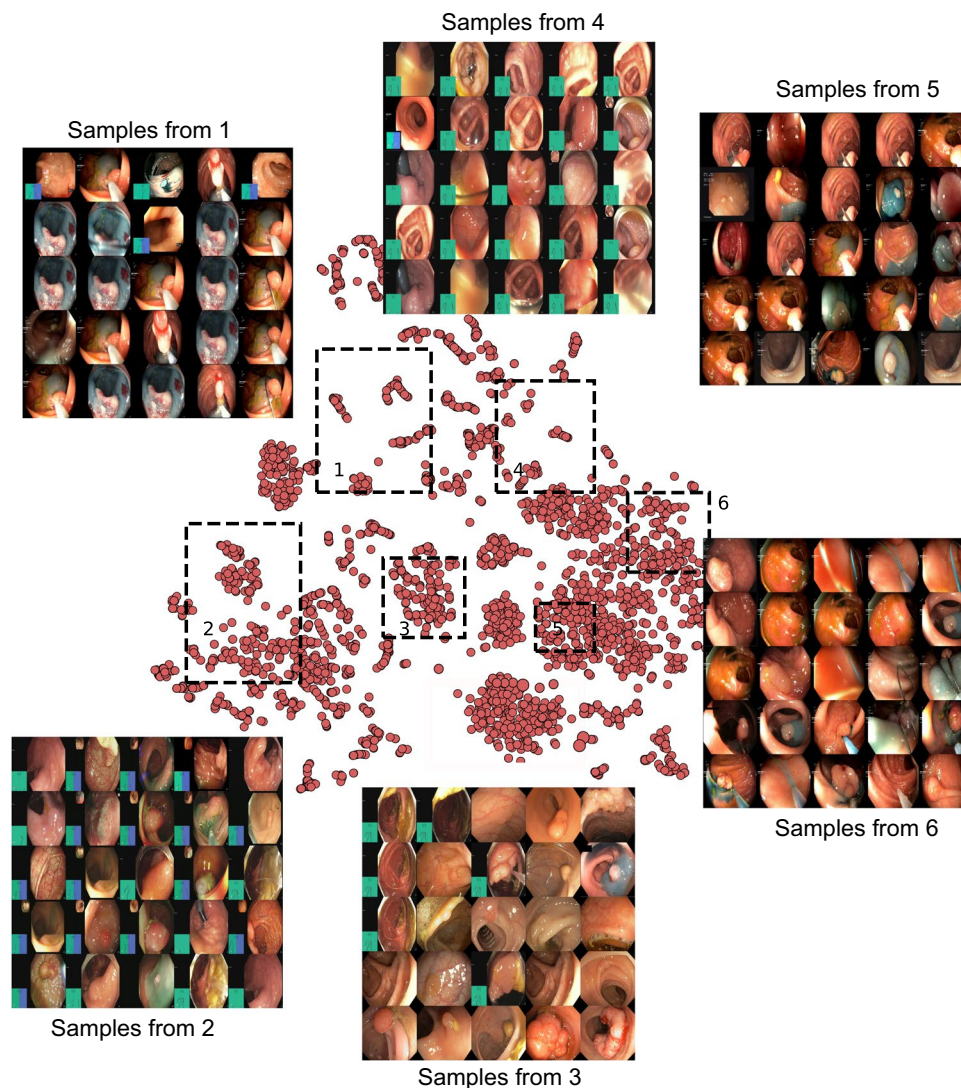
Fig. 1 t-SNE plot for positive samples: 2D t-SNE embedding of the "PolypGen" dataset based on deep autoencoder extracted features. Each point is an image in the positive samples of the dataset. For each of the six boxed regions (dashed black lines) 25 images were randomly sampled for display in a 5 × 5 image grid. Here, the 1st boxed region represents mostly the sequence data. Interestingly, the 3rd, the 4th, and the 6th boxed regions mostly represent both polyp and non-polyp data and heterogeneously distributed. Samples from 2nd and the 5th boxed regions shows mostly protruded polyps but with differently positioned endoscopy locations. Some samples in these also include the colonoscopy frames with dyes.

Methods

Ethical and privacy aspects of the data. Our multi-centre polyp detection and segmentation dataset consists of colonoscopy video frames that represent varied patient population imaged at six different centres including Egypt, France, Italy, Norway and the United Kingdom (UK). Each centre was responsible for handling the ethical, legal and privacy of the relevant data. The data collection from each centre included either two or all essential steps described below:

1. Patient consenting procedure at each individual institution (required)
2. Review of the data collection plan by a local medical ethics committee or an institutional review board
3. Anonymization of the video or image frames (including demographic information) prior to sending to the organizers (required)

Table 2 illustrates the ethical and legal processes fulfilled by each centre along with the endoscopy equipment and recorders used for the data collected.

Study design. PolypGen data was collected from 6 different centres. More than 300 unique patient videos/frames were used for this study. The general purpose of this diverse dataset is to allow robust design of deep learning

centres	System info.	Ethical approval	Patient consenting type
Ambroise Paré Hospital, Paris, France	Olympus Exera 195	N° IDRCB: 2019-A01602-55	Endospectral study
Istituto Oncologico Veneto, Padova, Italy	Olympus endoscope H190	NA	Generic patients consent
Centro Riferimento Oncologico, IRCCS, Italy	Olympus VG-165, CV180, H185	NA	Generic patients consent
Oslo University Hospital, Oslo, Norway	Olympus Evis Exera III, CF 190	Exempted†	Written informed consent
John Radcliffe Hospital, Oxford, UK	GIF-H260Z, EVIS Lucera CV260, Olympus Medical Systems	REC Ref: 16/YH/0247	Universal consent
University of Alexandria, Alexandria, Egypt	Olympus Exera 160AL, 180AL	NA	Written informed consent

Table 2. Data collection information for each centre: Data acquisition system and patient consenting information. †Approved by the data inspectorate. No further ethical approval was required as it did not interfere with patient treatment.

models and their validation to assess their generalizability capability. In this context, we have proposed different dataset configurations for training and out-of-sample validation and proposed unique generalization assessment metrics to reveal the strength of deep learning methods. Below we provide a comprehensive description of dataset collection, annotation strategies and its quality, ethical guidelines and metric evaluation strategies.

Video acquisition, collection and dataset construction. A consortium of six different medical data centres (hospitals) were built where each data centre provided videos and image frames from at least 50 unique patients. The videos and image samples were collected and sent by the senior gastroenterologists involved in this project. The collected dataset consisted of both polyp and normal mucosa colonoscopy acquisitions. To incorporate the nature of polyp occurrences and maintain heterogeneity in the data distribution, the following protocol was adhered for establishing the dataset:

1. Single frame sampling from each patient video incorporated different view points
2. Sequence frame sampling consisted of both visible and invisible polyp frames (at most cases) with a minimal gap
3. While single frame data consisted of all polyp instances in that patient, sequence frame data consisted of only a localised targeted polyp
4. Positive sequence included both positive and negative polyp instances but from video with confirmed polyp location while for negative sequence only patient videos with normal mucosa were used

An overview of the number of samples comprising positive samples and negative samples is presented in Fig. 2a. The total positive samples of 3762 frames are released that comprises of 484, 1166, 457, 677, 458 and 520 frames from centres C1, C2, C3, C4, C5 and C6, respectively. These frames consist of 1537 single frames (1449 frames from C1-C5 also provided in EndoCV2021 challenge and 88 frames from C6), and 2225 sequence frames with majority of sequence data sampled from centres C2 (865), C4 (450), and C6 (432). The number of polyp counts for pixel-level annotation of small ($\leq 100 \times 100$), medium (between $>100 \times 100$ pixels and $\leq 200 \times 200$ pixels), large ($\geq 200 \times 200$ pixels) sized polyps from each centre including no polyp frames but frames in close proximity of polyp are represented as histogram plot (Fig. 2b). The total annotations for polyp that we release is 3447. All these polyp samples are verified by expert gastroenterologists.
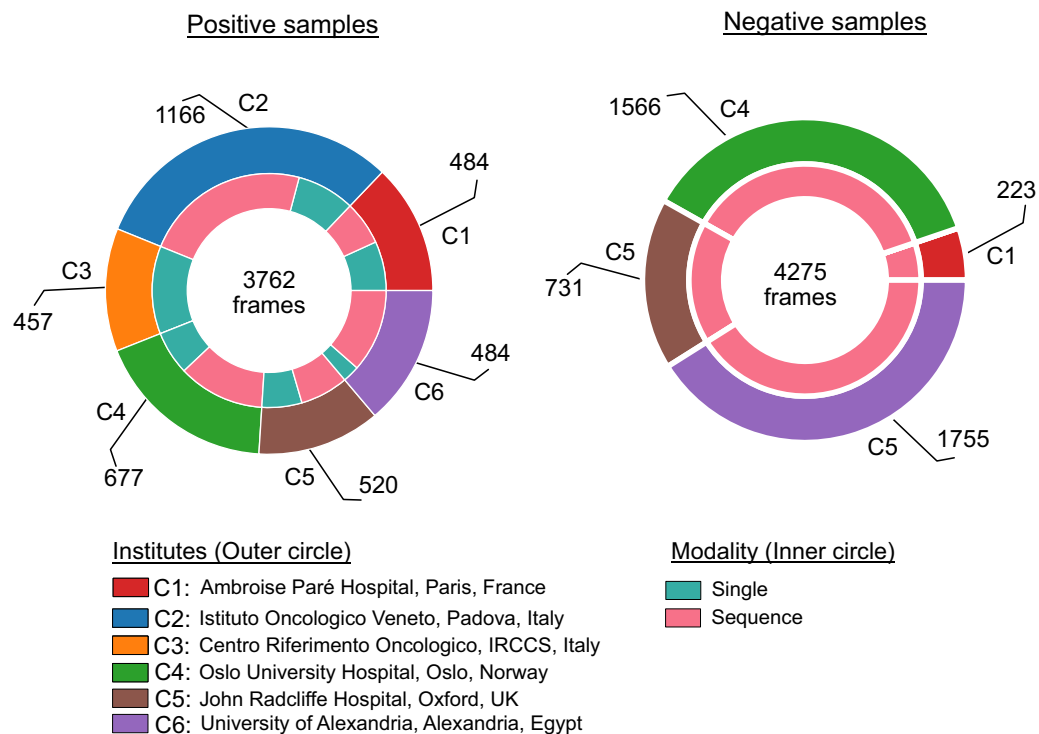
We have provided both still image frames and continuous short video sequence data with their corresponding annotations. The positive and negative samples in the dataset of the polyp generalisation (PolypGen) are further detailed below.

Positive samples. Positive samples consist of video frames from the patient with a diagnosed polyp case. The selected frames may or may not have the polyp in it but may be located near to the chosen frame. Nevertheless, a majority of these frames consists of at least one polyp in the frame. For the sequence positive samples, the continuity of the appearance and disappearance of the polyp similar to real scenario has been taken into account and thus these frames can have a mixture of polyp instances and frames with normal mucosa. Table 3 is provided to detail the characteristics of 23 sequence data included in our dataset. It can be observed from Fig. 4 that varied sized polyps are included in the dataset with variable view points, occlusions and instruments. Exemplary pixel-level annotations of positive polyp samples for each centre and their corresponding bounding boxes are presented in Fig. 3.

Negative samples. Negative samples mostly refer to the negative sequences released in this dataset, i.e. no polyp frames. These sequences are taken from patient videos which consisted of confirmed absence of polyps (*i.e.*, normal mucosa) in the acquired videos or at areas away from the polyp occurrences. It includes cases with anatomies such as colon linings, light reflections and mucosa covered with stool that may be confused with polyps (see Fig. 5 and corresponding negative sequence attributes in Table 4).

Annotation strategies and quality assurance. A team of 6 senior gastroenterologists (all over 20 years of experience in endoscopy), two experienced post-doctoral researchers, and one PhD student were involved in the data collection, data sorting, annotation and the review process of the quality of annotations. For details on data collection and data sorting please refer to Section **Video acquisition, collection and dataset construction**.

a. Sample distribution in PolypGen dataset

Positive samples

Negative samples

Institutes (Outer circle)
- **C1:** Ambroise Paré Hospital, Paris, France
- **C2:** Istituto Oncologico Veneto, Padova, Italy
- **C3:** Centro Riferimento Oncologico, IRCCS, Italy
- **C4:** Oslo University Hospital, Oslo, Norway
- **C5:** John Radcliffe Hospital, Oxford, UK
- **C6:** University of Alexandria, Alexandria, Egypt

Modality (Inner circle)
- Single
- Sequence

b. Annotated size-based polyp counts in positive samples
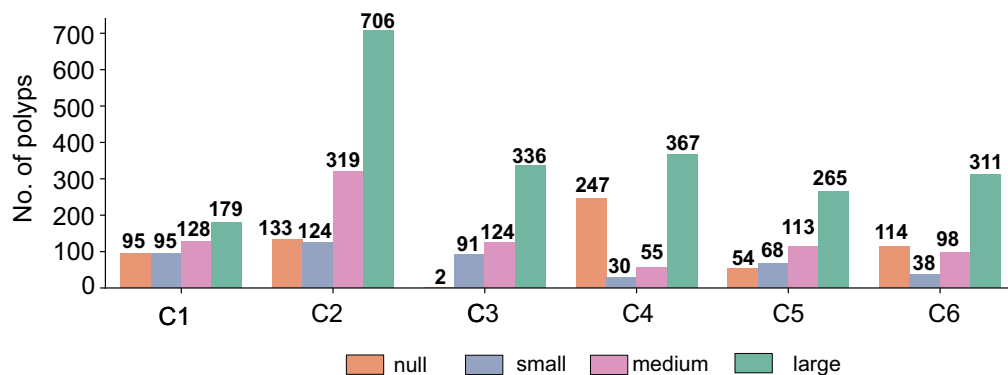
Fig. 2 PolypGen dataset: (**a**) Positive (both single and sequence frames) and negative samples (sequence only) from each centre, and (**b**) polyp size-based histogram plot for positive samples showing variable sized annotated polyps in the dataset (small is $\leq 100 \times 100$ pixels; medium is $> 100 \times 100 \leq 200 \times 200$, and large is $> 200 \times 200$ pixels). Null represents no polyp present in the sample.

All annotations were performed by a team of three experienced researchers using an online annotation tool called Labelbox[16]. The dataset was divided equally between the three reviewers for the annotation process where each research annotated a specific group of frames. However, all the annotated frames were revised by all the senior gastroenterologists team. Each annotation was later cross validated for accurate segmentation margins by the team and by the centre expert. Further, an independent binary review process was then assigned to a senior gastroenterologists, in most cases experts from different centres were assigned. A protocol for manual annotation of polyp was designed to minimise the heterogeneity in the manual delineation process. The protocol was in detail discussed together with the clinical experts and the annotators during several weekly meeting. Here, we only present a brief on the important aspects of the annotation that should be taken care during annotations. Example samples were provided by expert endoscopists to the annotators especially this was the case in the video annotations. The set protocols are listed below (refer Fig. 3 for final ground truth annotations):

Sequence	Description	Artifact
seq. 1	Normal mucosa	Light reflections; green patch
seq. 2	5 mm polyp at 6 o'clock	Partially covered with stool; reflections; green patch
seq. 3	Polyp at distance, 4 o'clock	Light reflection from liquid; green patch
seq. 4	2–3 mm polyp	Liquid covering half of the image; green patch
seq. 5	5 mm polyp catched by a snare	Partial occlusion by biopsy instrument
seq. 6	Polyp covering half of the circumference	Cap; green patch
seq. 7	Normal mucosa	Light reflection; some remnant stool; green patch
seq. 8	Typical flat cancer	Light reflection; green patch
seq. 9	2 mm polyp at 2 o'clock	Light reflection; green patch
seq. 10	Subtle small protrusions	some remnant stool
seq. 11	Polyp at 2–3 o'clock	Light reflections in the periphery
seq. 12	Dye lifted 4–5 mm polyp	Low contrast
seq. 13	6–7 mm polyp catched with a snare	Low contrast; small reflections
seq. 14	Paris 1 p polyp, large long stak, JNET 2a	Lifted by Indigo Carmine, snare place around the stalk
seq. 15	Paris 1 s JNET2a polyp and 1 Paris 1 sp to the left	Lifted by Indigo Carmine
seq. 16	Paris 1 p polyp, large long stalk, JNET 2a	Lifted by Indigo Carmine
seq. 17	Paris 1 sp polyp	Light reflections make surface assessment impossible
seq. 18	Difficult interpretation	Blurry image and reduced view
seq. 19	Paris 1 p polyp, large long stalk, JNET 2a	Less contrast and slightly occluded
seq. 20	half of the polyp visible	Blurry image, with some blood on the mucosa
seq. 21	Two adenomas polyp	Blurry image
seq. 22	adenomas polyp	Blurry image makes exact diagnosis impossible
seq. 23	serrated polyp	Perfect clean mucosa, minor light reflections

Table 3. Positive sample sequence summarised attribute: Total of 23 sequences are provided as positive sample sequence for patients with polyp instances during colonoscopy examination. Here JNET refers to Japan NBI Expert Team classification score. These sequences depict different sized polyps and location with different artifacts and varying visibility. Sequences referring to one selected image is shown in Fig. 4.

- Clear raised polyps: Boundary pixels should include only protruded regions. Precaution has to be taken when delineating along the normal colon folds
- Inked polyp regions: Only part of the non-inked appearing object delineation
- Polyps with instrument parts: Annotation should not include instrument and is required to be carefully delineated and may form more than one object
- Pedunculated polyps: Annotation should include all raised regions unless appearing on the fold
- Flat polyps: Zooming the regions identified with flat polyps before manual delineation. Also, consulting centre expert if needed.
- Video sequence annotation: One sample from expert gastroenterologist were provided for sequences that showed difficulty in distinguishing between mucosa and polyp. Polyps that are distant and not clearly visible were also not annotated as polyps.
- Tackling with occlusion: Polyps that were occluded with stool or instrument were required to exclude the parts of mucosa that were obstructed.
- Cancerous mucosa: Mucosa that were already cancerous but not appear as polyps were excluded from the annotation. However, raised mucosal surface that charactised adenomatous polyps were included.

Each of these annotated masks were reviewed by expert gastroenterologists. During this review process, a binary score was provided by the experts depending on whether the annotations were clinically acceptable or not. Some of the experts also provided feedback on the annotation and these images were placed into ambiguous category for further rectification based on expert feedback. These ambiguous category was then jointly annotated by two researchers and further sent for review to one expert. The outcome of these quality checks are provided in Fig. 6. It can be observed that large fraction (30.5%) of annotations were rejected (excluding ambiguous batch, total annotations were 2213, among which only 1537 were accepted and 676 frames were rejected). Similarly, the ambiguous batch that included correction of annotations after the first review also recorded 34.17% rejected frames on the second review.

Data Records

A sub-set of this dataset (from C1 - C5 except C6) forms the dataset of our EndoCV2021 challenge[17] (*Addressing generalisability in polyp detection and segmentation*) training data (https://endocv2021.grand-challenge.org), *i.e.*, an event held in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France. The current released data consists of additional positive and negative frames for both single and sequence data and a 6th centre data (C6). The presented version does not consists of training and test splits and users are free to apply their own strategies as applicable to the nature of their work. To access the complete dataset, users are requested to create a Synapse account (https://www.synapse.org/) and then the compiled dataset

C1: Ambroise Paré Hospital, Paris, France

C2: Istituto Oncologico Veneto, Padova, Italy

C3: Centro Riferimento Oncologico, IRCCS, Italy

C4: Oslo University Hospital, Oslo, Norway

C5: John Radcliffe Hospital, Oxford, UK

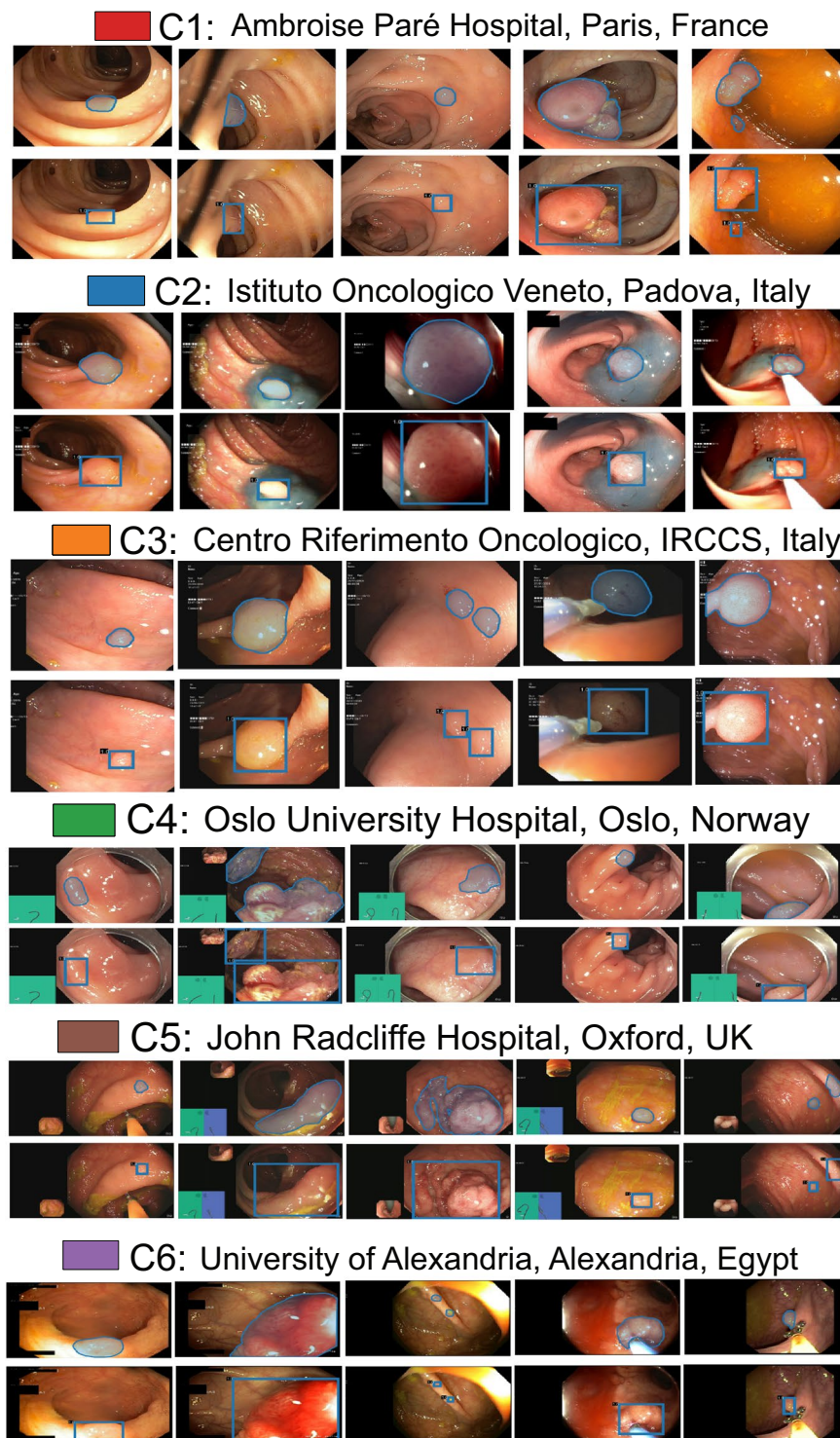C6: University of Alexandria, Alexandria, Egypt

Fig. 3 Sample polyp annotations from each centre: Segmentation area with boundaries and corresponding bounding box/boxes overlaid images from all six centres. Samples include both small sized polyp ($<$10000 pixels) including some flat polyp samples to large sized ($\geq$40000 pixels) polyps and polyps during resection procedure such as polyps with blue dyes.

can be downloaded at (https://www.synapse.org/#!Synapse:syn45200214)[18] which has been published under Creative Commons 4.0 International (CC BY) licence. Dataset can only be used for educational and research purposed and must cite this paper. All collected data has been obtained through a written patient consent or through an ethical approval as tabulated in Table 2.

The folder structure of the compiled multi-centre polyp detection and segmentation dataset is as presented in Fig. 7. The main folder "PolypGen" is divided into folders a) Positive and b) Negative. The positive folder is
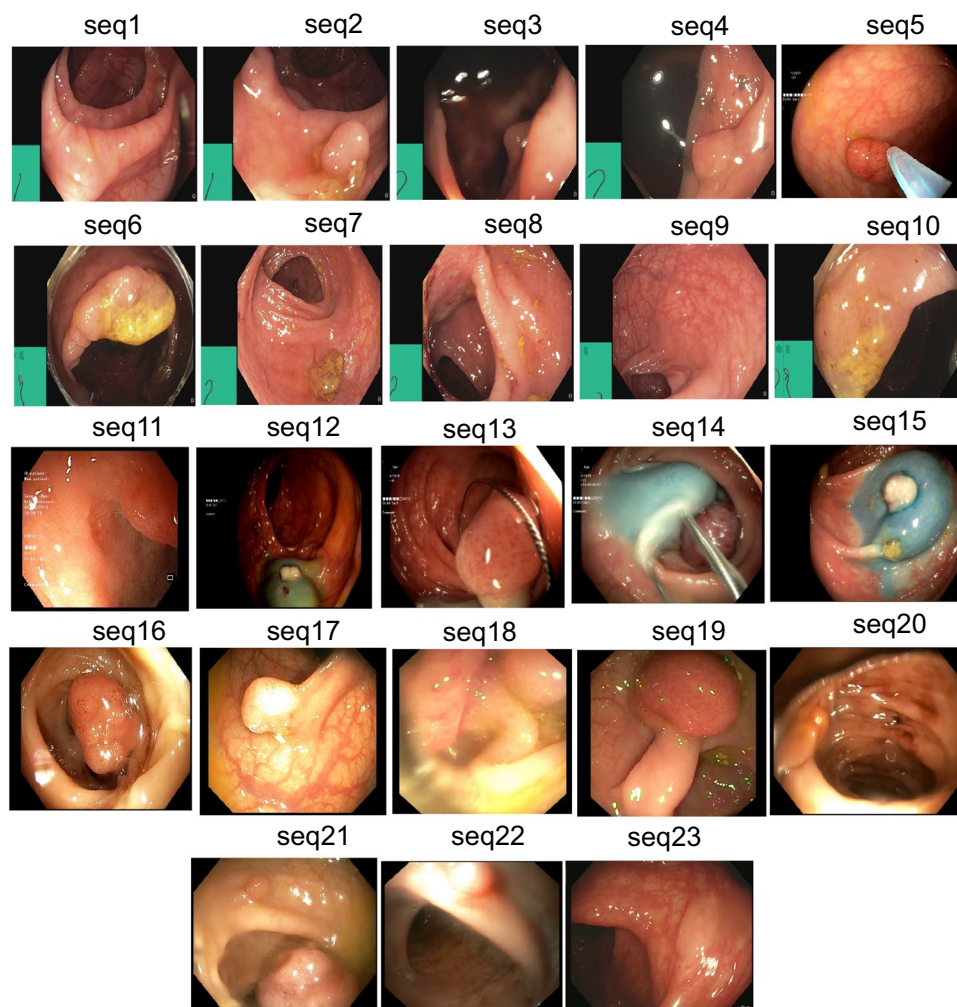
Fig. 4 Positive sequence data: Representative samples chosen from 23 sequences of the provided positive samples data. A summary description is provided in Table 3. Parts of images have been cropped for visualization.

then subdivided (sub-folder level 0) into "Single_Frames_centre_Split" and "Sequence_Frames". Each of these folders are then further subdivided (sub-folder-1). Single frame data is further categorized as centre-wise split "data-C1" to "data-C6", where C1 representing centre 1 and C6 representing the 6th centre, while sequence frames are categorized into "seq_1" to "seq_23" (legend color in Fig. 7 represents their corresponding centre). Finally, a second sub-folder level 2 includes four folders consisting of original images (.jpg format), annotated masks (.jpg format), bounding boxes (.txt) in the standard PASCAL VOC format[19], and images with box over-lay (.jpg). No mix of centrewise sequence is done in any of the sequence data, and the data can consist of both positive negative polyp image samples. No polyp in positive samples mask empty bounding box files and mask with null values. Both masks and bounding box overlaid images are of same size as that of the original images. The negative frames folder on the other hand does consist of only sequence frames, i.e., sub-folder level 0 only as these sequence samples consists of patients with no polyps during the surveillance. To further assist the users, we have also provided the folder structure inside the main folder "PolypGen". The size of images provided in the dataset can range from 384×288 pixels to 1920×1080 pixels. The size of masks correspond to the size of the original images, however, the polyp sizes in the provided dataset is variable (as indicated in Fig. 2). Since we followed a full anonymisation protocol, no gender or age information is provided.

Technical Validation

For the technical validation, we have included single frames data (1449 frames) from five centres (C1 to C5) in our training set and tested on out-of-sample C6 data on both single (88 frames) and sequence frames (432 frames). Such out-of-sample testing on a completely different population and endoscopy device data allows to comprehensively provide evidence of generalisability of current deep learning methods. The training set was split into 80% training only and 20% validation data. Here, we take most commonly used methods for segmentation in the biomedical imaging community[20–24], including that for polyps. For reproducibility, we have included the train-validation split as *.txt* files as well in the *PolypGen* dataset folder. However, users can choose any set of different combined training or split training schemes for generalisation tests as they prefer, e.g., training on
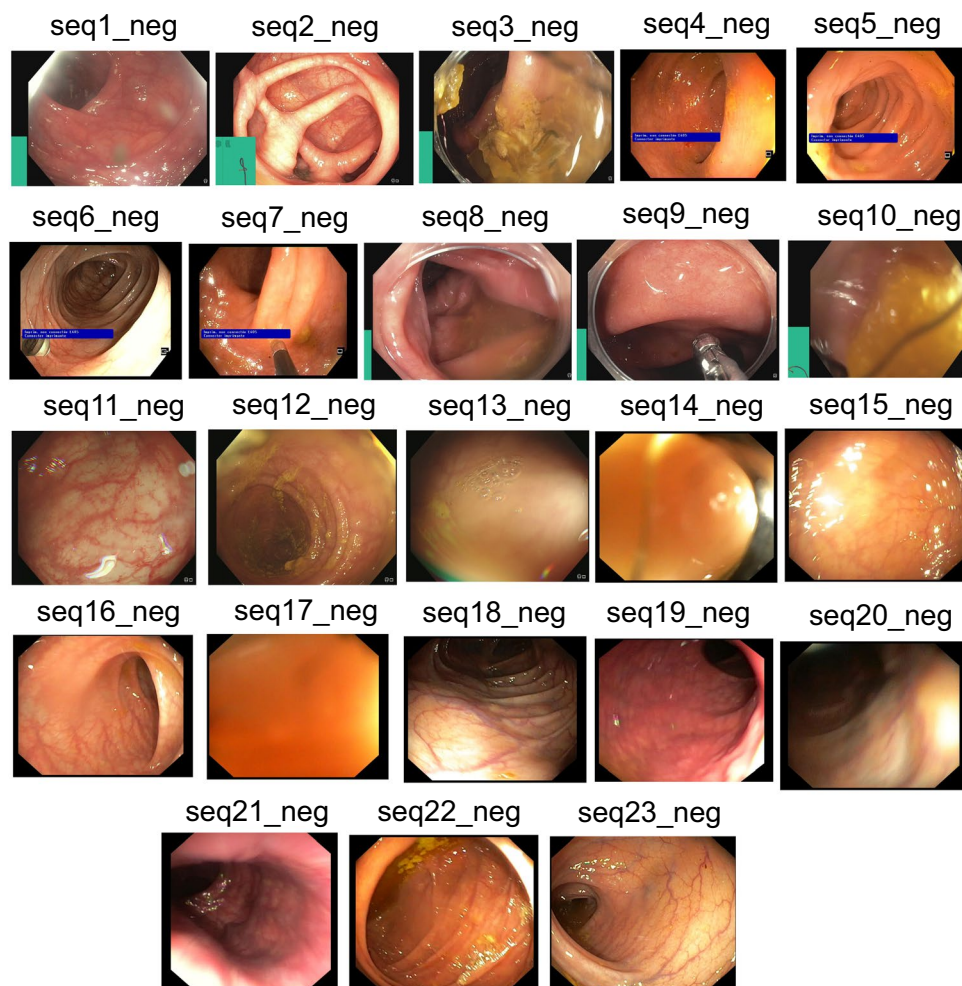
Fig. 5 Negative sequence data: Representative samples chosen from each sequence of the provided negative samples data. A summary description is provided in Table 4. Parts of images have been cropped for visualization.

random three centres and testing on remaining three centres. Also, the dataset is suitable for federated learning (FL) approaches[25] that uses decentralised training and allow to aggregate the weights from the central server for improved and generalisable model without compromising data privacy.

Benchmarking of state-of-the-art methods.   To provide generalisation capability of some state-of-the-art (SOTA) methods, we have used a set of popular and well-established semantic segmentation CNN models on our PolypGen dataset. Each model was run for nearly 500 epochs with batch size of 16 for image size of 512×512. All models were optimised using Adam with weight decay of 0.00001 learning rate of 0.01 and allowing the best model to be saved after 100 epochs. Classical augmentation strategies were used that included scaling (0.5, 2.0), random cropping, random horizontal flip and image normalisation. All models were run on Quadro RTX6000.

Evaluation metrics for segmentation.   We compute standard metrics used for assessing segmentation performances that includes Jaccard Index $\left(\text{JI} = \frac{TP}{TP + FP + FN} \right)$, F1-score (aka Dice similarity coefficient, DSC), F2-score, precision (aka positive predictive value, PPV or $p = \frac{TP}{TP + FP}$), recall $\left(r = \frac{TP}{TP + FN} \right)$, and overall accuracy $\left(\text{Acc.} = \frac{TP + TN}{TP + TN + FP + FN} \right)$ that are based on true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) pixel counts. Precision-recall tradeoff is also given by the Dice similarity coefficient (DSC) or F1-score and F2-scores:

$$\text{F}_{\beta} = (1 + \beta^2) \cdot \frac{p \cdot r}{(\beta^2 \cdot p) + r},$$

(1)

Sequence	Description	Artifact
seq1_neg	Normal vascular pattern	Light reflections in the periphery; not clean lens
seq2_neg	Normal vascular pattern	Contracted bowel; light reflections
seq3_neg	Mucosa not satisfactory visualized	Stool covers the field of view
seq4_neg	Reduced vascular pattern	Light reflections and small amount of stool
seq5_neg	Reduced vascular pattern	Light reflections
seq6_neg	Normal vascular pattern	Light reflections; biopsy forceps
seq7_neg	Normal vascular pattern	Very close to the luminal wall
seq8_neg	Normal vascular pattern	Blurry; semi-opaque liquid; cap
seq9_neg	Normal vascular pattern	Blurry; semi-opaque liquid; cap
seq10_neg	Not possible to assess the mucosa	Blurry; occluded
seq11_neg	Normal vascular pattern	Light reflections in the periphery; bubble on the lens
seq12_neg	Normal vascular pattern	Not clean lens, mucosa covered by stool
seq13_neg	Probably normal vascular pattern; Not possible to assess the mucosa	Air bubbles; remnant stool; too close to the mucosa, blur, reflections
seq14_neg	Clean bowel, normal vascular pattern	Very close to the mucosa in all
seq15_neg	Clean bowel, normal vascular pattern	Some bubbles and light reflections
seq16_neg	Clean bowel, normal vascular pattern	Some bubbles and light reflection
seq17_neg	Clean bowel, normal vascular pattern	Very close to the mucosa in all
seq18_neg	Clean bowel, normal vascular pattern, well distended	Some stool residues
seq19_neg	Clean bowel, normal vascular pattern, well distended	Some liquid residues
seq20_neg	Clean bowel, normal vascular pattern, well distended	Some stool residues and reflections
seq21_neg	Clean bowel, normal vascular pattern	Very close, minor stool residues in last images
seq22_neg	Clean bowel, normal vascular pattern, well distended	Some liquid and stool residues, reflections
seq23_neg	Perfect clean bowel, normal vascular pattern, well distended	Some light reflections

Table 4. Negative sample sequence summarised attribute: Total of 23 sequences are provided as negative sample sequence for patients with no polyp during colonoscopy examination. These sequences depict different artifacts and varying visibility of vascular pattern and occlusion of mucosa.

where F_β-score is computed as weighted harmonic means of precision and recall.

Another commonly used segmentation metric that is based on the distance between two point sets, here ground truth (G) and estimated or predicted (E) pixels, to estimate ranking errors is the average Hausdorff distance (d_{AHD}) and defined as:

$$d_{AHD}(G, E) = \left(\frac{1}{|G|} \sum_{g \in G} \min_{e \in E} d(g, e) + \frac{1}{|E|} \sum_{e \in E} \min_{g \in G} d(g, e) \right)/2. \qquad (2)$$

Since boundary-distance-based metrics are insensitive to the object size and sensitive to the object shape, we include two additional metrics which are average surface distance (ASD) and Normalised surface dice (NSD). ASD is the average of all distances (Euclidean) between pixels on the predicted object segmentation border and its nearest neighbour on the reference segmentation border. All obtained distances are averaged, yielding an average distance value ASD for symmetric case is:

$$ASD(G, E) = \frac{1}{|G| + |E|} \left(\sum_{g \in G} \min_{e \in E} d(g, e) + \sum_{e \in E} \min_{g \in G} d(g, e) \right). \qquad (3)$$

The normalized surface dice (NSD)[26] computes the fractional correctly predicted segmentation boundary using an additional threshold accounting for amount of class-specific distance deviation. In our experiments we have set it to 10. If d be the distance and τ be the acceptable deviation (threshold) with $d(B_E, B_G)$ be the computed distance for predicted mask boundary w.r.t the nearest-neighbour distances to the reference segmentation boundary then for symmetric case NSD is given by:

$$d'(B_E, B_G) = \{d \in d(B_E, B_G) | d \leq \tau\} \quad \text{and} \quad NSD = \frac{|d'(B_E, B_G)| + |d'(B_G, B_E)|}{|d(B_E, B_G)| + |d(B_G, B_E)|}. \qquad (4)$$

NSD is bounded between 0 and 1 where 1 refers to the segmentation boundary below deviation threshold τ.

Polyp segmentation benchmarking. Long *et al.*[20] presented a Fully Convolutional Network (**FCN**) that uses downsampling and upsampling for image segmentation. The model is dived into two sectors, the first sector is responsible for extracting detailed feature maps through downsampling the spatial resolution of the image.
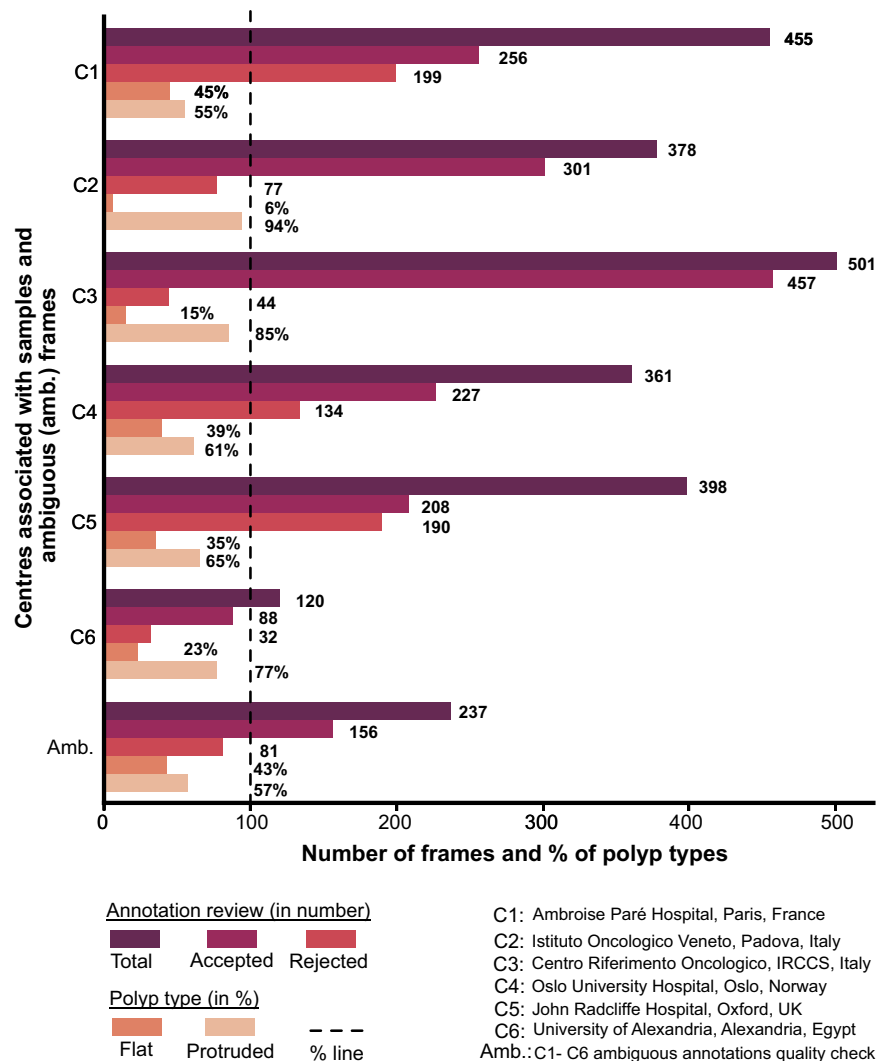
Fig. 6 Annotation quality review: Total curated frames along with accepted and rejected frame numbers during annotation quality review by experts for single frame data. Annotated frames with % of flat and protruded polyps categorised during annotation are also provided.

The second sector is responsible for retrieving the location information through upsampling. The **U-Net** architecture[21] has shown tremendous success in medical image segmentation[27] including endoscopy[15,28].

U-Net is generally an encoder network followed by a decoder network, where, convolution blocks followed by max-pooling downsampling are applied to the image to encode feature presentations at different multiple levels. Afterwards, the decoder projects semantically the distinguishing characteristics learnt by the encoder. The decoder is composed of upsampling and concatenation followed by a standard convolution function. The skip connections between downsampling and upsampling paths in the U-Net (i.e. which makes it symmetric) is the main difference between the U-Net and the FCN[29]. Pyramid Scene Parsing Network (**PSPNet**)[22] is designed to incorporate global context information for the task of scene parsing using context aggregation depending on different regions to take advantage of global context information. Both local and global findings provide a more reliable final prediction. The pyramid pooling module and CNN backbone with dilated convolutions are both present in the PSPNet encoder. Similarly, dilated convolutions enabled the construction of semantic segmentation networks to effectively control the size of the receptive field that was incorporated in an a family of very effective semantic segmentation architectures, collectively named DeepLab[23]. The DeepLabV3 capture multi-scale information by employing the atrous convolution at multiple rates in a cascade or parallel multi-scale context through spatial pyramid pooling. Moreover, ResUNet[30] incorporates the benefits of both the ResNet and U-Net which allowed the design of a network with fewer parameters and improved segmentation performance. Figure 8 provides an illustrative figure for the architecture of SOTA methods as explained.

All of these networks has been explored for polyp segmentation in literature[31–33]. Here, we benchmark our dataset on these popular deep learning model architectures. Out-of-sample generalisation results for both single frame (Table 5) and sequence data (Table 6) has been included in our technical validation of the presented data.
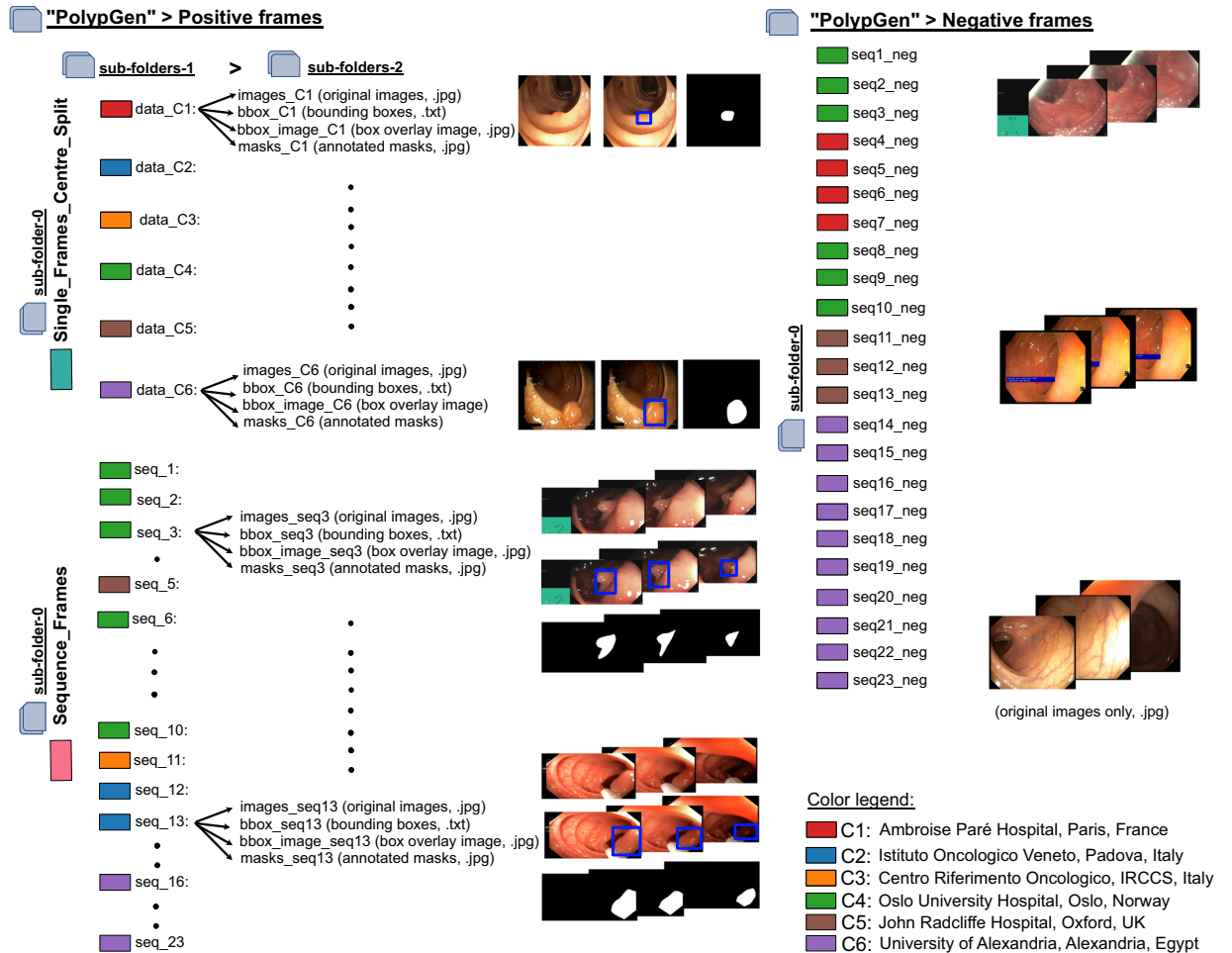
Fig. 7 Folder structure: PolypGen dataset is divided into two folders – positive frames and negative frames. Later, it is divided into three different levels for positive frames and only one level for negative samples. (On left) Sub-folder structure with different folder names and the format of data present in each sub-folder 2 is provided. Sample images are also shown. (On right) Sub-folders present in negative folder is shown with two sample sequences (from centre 4 and centre 1). Each data source centre is shown as color legends.

Validation summary. Our technical validation suggest that DeepLabV3+ with ResNet101 has the best performance on most metrics except for FPS suggesting larger latency in inference (Tables 5, 6). The highest score was 0.82 for DSC and the least 9.67 for d_{AHD} with DeepLabV3+ with ResNet101 on single frame data. However, the second best inference speed (FPS of 47) and score (DSC = 0.81, d_{AHD} = 9.95) was obtained again using DeepLabV3+ but with ResNet50 backbone. Similarly, for sequence C6 out-of-sample test generalisation the highest score of 0.65 for DSC with highest recall of 0.73 and the least 9.08 for d_{AHD} was obtained with DeepLabV3+ with ResNet101 backbone. With the same ResNet101 backbone ResUNet resulted in very close performance of 0.65 DSC but with highest precision on 0.92 and d_{AHD} of 9.20. However, even with the same ResNet101 backbone ResUNet (40 FPS) has better speed compared to DeepLabV3+ (33 FPS). In addition, we also evaluated using normalised surface dice (NSD) and mean average surface distance (MASD) both of which demonstrated similar performance trend for most methods. NSD was the lowest for the DeepLabV3+ and ResNetUNet with ResNet101 backbone, 0.64 and 0.63, respectively for single frames, and 0.57 each for sequence frames. The lowest MASD was reported for DeepLabV3+ with ResNet101 backbone with 23.29 and 18.59, respectively, for single and sequence frames. We also ran size-stratified DSC estimates for each algorithms. For medium and large polyps, the DSC score for majority of methods were not affected with DSC of 0.87 for large polyps and 0.84 for medium in the case of DeepLabV3+ with ResNet101 backbone. However, a steep decrease in DSC was observed for the small polyps with DSC value of only 0.46. Also for classical PSPNet and FCN8 networks, DSC difference was estimated to be over 0.10 between large and medium polyp sizes while for both ResNet-UNet and DeepLabV3+ had much smaller difference.

While for the single frames data DSC is above 0.80 for ResNetUNet and DeepLabV3+ using ResNet101 backbone, however, the same architectures only provided DSC around 0.70 on the sequence dataset. This can be primarily because of the larger number of frames in the sequence dataset (nearly 5 times) which if 432 versus only 88 single frames, and the heterogeneous image quality in the sequence data. Additionally, it is to be noted
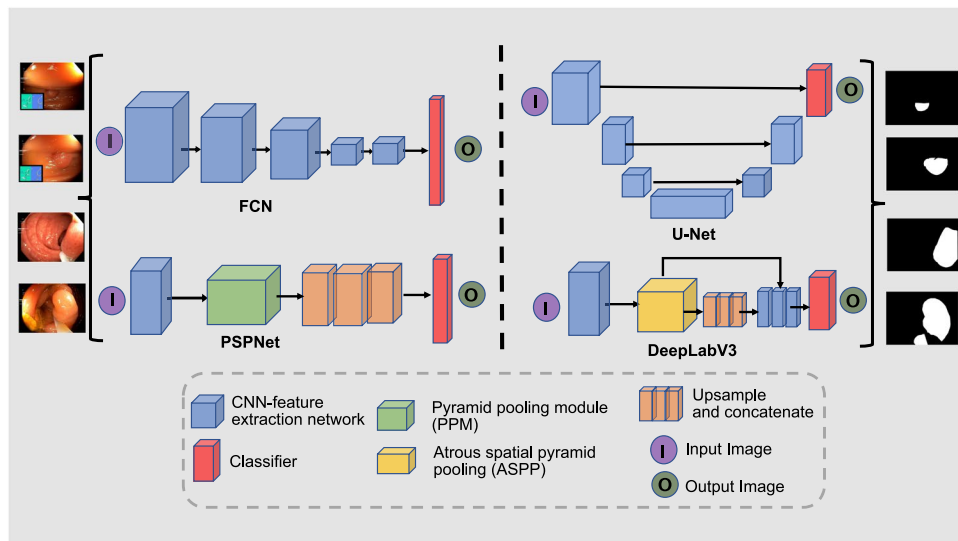
Fig. 8 Architecture layout for state-of-the-art methods: FCN (fully convolutional Network), U-Net, PSPNet (Pyramid Scene Parsing Network) and DeepLabV3.

Method	JI ↑	DSC ↑	F2 ↑	PPV ↑	Recall ↑	Acc. ↑	d_{AHD} ↓	NSD ↑	MASD ↓	FPS ↑
FCN8[20]	0.68 ± 0.30	0.76 ± 0.30	0.75 ± 0.31	0.90 ± 0.15	0.74 ± 0.31	0.97	10.69	0.49	37.61	44
U-Net[21]	0.55 ± 0.34	0.63 ± 0.36	0.64 ± 0.36	0.76 ± 0.31	0.66 ± 0.37	0.96	13.89	0.45	93.02	21
PSPNet[22]	0.72 ± 0.27	0.80 ± 0.26	0.79 ± 0.27	0.88 ± 0.20	0.79 ± 0.28	**0.98**	10.39	0.56	34.12	31
DeepLabV3+[23] (ResNet50)	**0.75 ± 0.28**	**0.81 ± 0.27**	**0.80 ± 0.28**	**0.92 ± 0.17**	0.79 ± 0.29	**0.98**	**9.95**	0.62	41.04	47
ResNetUNet[21,24] (ResNet34)	0.73 ± 0.29	0.79 ± 0.29	0.77 ± 0.29	**0.92 ± 0.20**	0.78 ± 0.29	**0.98**	10.04	0.59	35.83	**87**
DeepLabV3+[23] (ResNet101)	**0.75 ± 0.28**	**0.82 ± 0.27**	**0.80 ± 0.27**	**0.92 ± 0.18**	**0.81 ± 0.27**	**0.98**	**9.67**	**0.64**	**23.29**	33
ResNetUNet[21,24] (ResNet101)	0.74 ± 0.29	0.80 ± 0.28	**0.80 ± 0.28**	0.93 ± 0.14	**0.80 ± 0.29**	**0.98**	10.10	**0.63**	**27.38**	40

Table 5. Performance evaluation of SOTA segmentation methods on 88 single frames from centre 6 in an out-of-sample generalisation task. Top two values are presented in bold. **JI**: Jaccard index **DSC**: Dice coefficient **F2**: Fbeta-measure, with $\beta = 2$ **PPV**: positive predictive value **Acc.**: overall accuracy **d_{AHD}**: avg. Hausdorff distance **FPS**: frames per second ↑: best increasing ↓: best decreasing NSD: Normalised Surface Dice MASD: Mean Average Surface Distance.

Method	JI ↑	DSC ↑	F2 ↑	PPV ↑	Recall ↑	Acc. ↑	d_{AHD} ↓	NSD ↑	MASD ↓
FCN8[20]	0.56 ± 0.37	0.62 ± 0.38	0.59 ± 0.37	0.88 ± 0.28	0.63 ± 0.36	0.95	9.84	0.43	32.97
UNet[21]	0.43 ± 0.37	0.50 ± 0.39	0.47 ± 0.39	0.68 ± 0.41	0.62 ± 0.37	0.95	11.22	0.39	48.66
PSPNet[22]	0.58 ± 0.38	0.64 ± 0.38	0.61 ± 0.38	0.84 ± 0.33	0.68 ± 0.35	0.96	9.84	0.49	33.46
DeepLabV3+[23] (ResNet50)	0.60 ± 0.37	0.67 ± 0.37	0.64 ± 0.37	0.85 ± 0.31	**0.71 ± 0.33**	0.96	9.63	0.51	28.19
ResNetUNet[21,24] (ResNet34)	0.59 ± 0.37	0.66 ± 0.38	0.63 ± 0.38	0.87 ± 0.30	0.70 ± 0.35	0.96	9.78	0.50	27.10
DeepLabV3+[23] ResNet101	**0.65 ± 0.37**	**0.71 ± 0.37**	**0.68 ± 0.37**	**0.90 ± 0.26**	**0.73 ± 0.34**	**0.97**	**9.08**	**0.57**	**18.59**
ResNetUNet[21,24] (ResNet101)	**0.65 ± 0.37**	**0.70 ± 0.37**	**0.68 ± 0.37**	**0.92 ± 0.22**	**0.71 ± 0.35**	0.96	**9.20**	**0.57**	**22.70**

Table 6. Performance evaluation of SOTA segmentation methods on 432 sequence frames from centre 6 in an out-of-sample generalisation task. Top two methods are presented in bold. **JI**: Jaccard index **DSC**: Dice coefficient **F2**: Fbeta-measure, with $\beta = 2$ **PPV**: positive predictive value **Acc.**: overall accuracy **d_{AHD}**: Average Hausdorff distance ↑: best increasing ↓: best decreasing. NSD: Normalised Surface Dice MASD: Mean Average Surface Distance.

that while single frames have very clean polyp images, sequence can have different view points, sizes and quality and may or may not consist of polyps as in the real-world colonoscopy data. The boundary distances comparing with the estimated and ground truth mask boundaries using NSD, MASD and HD showed similar performance changes between different methods.

Most methods have close to 40 FPS performance, however, ResNetUNet with smaller backbone architecture ResNet34 provided real-time performance with 87 FPS but with DSC close to DeepLabV3+ with ResNet50

Method	Val.	JI ↑	DSC ↑	F2 ↑	PPV ↑	Recall ↑	Acc. ↑	d_{AHD} ↓
DeepLabV3+[23] (ResNet50)	C1	0.70±0.32	0.76±0.33	0.75±0.33	0.85±0.28	0.79±0.30	0.98	4.77
DeepLabV3+[23] (ResNet50)	C2	0.72±0.28	0.79±0.27	0.78±0.27	0.88±0.23	0.80±0.25	0.97	4.70
DeepLabV3+[23] (ResNet50)	C3	0.74±0.28	0.80±0.27	0.80±0.26	0.86±0.24	0.83±0.25	0.97	4.77
DeepLabV3+[23] (ResNet50)	C4	0.76±0.28	0.82±0.27	0.81±0.28	0.91±0.17	0.81±0.28	0.98	4.77
DeepLabV3+[23] (ResNet50)	C5	0.73±0.31	0.78±0.31	0.77±0.32	0.93±0.16	0.78±0.31	0.98	4.88

Table 7. Centre-wise performance evaluation of best approach while training using four centres, validating on individual centre, and testing on out-of-sample generalisation task (centre C6). Val.: Validation centre **JI**: Jaccard index **DSC**: Dice coefficient **F2**: Fbeta-measure, with $\beta = 2$ **PPV**: Positive predictive value**Acc.**: overall accuracy $\mathbf{d}_{AHD}$: Average Hausdorff distance ↑: best increasing ↓: best decreasing.

a. Best performing

Single frames | Sequence frames

b. Worse performing
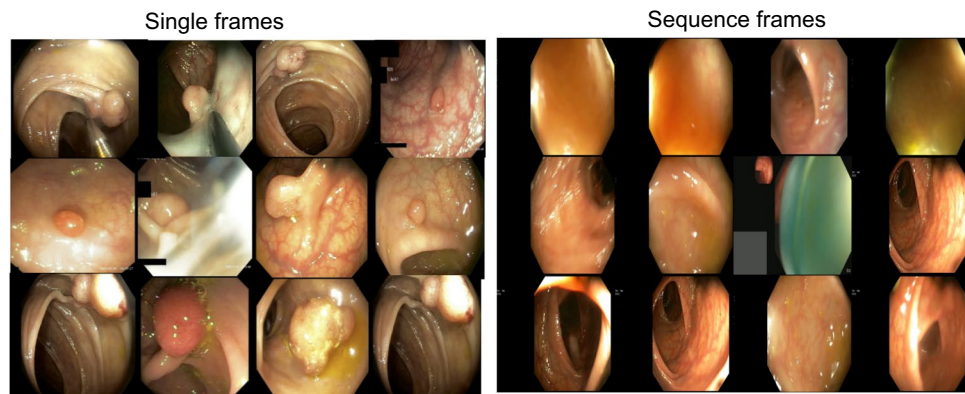
Single frames | Sequence frames
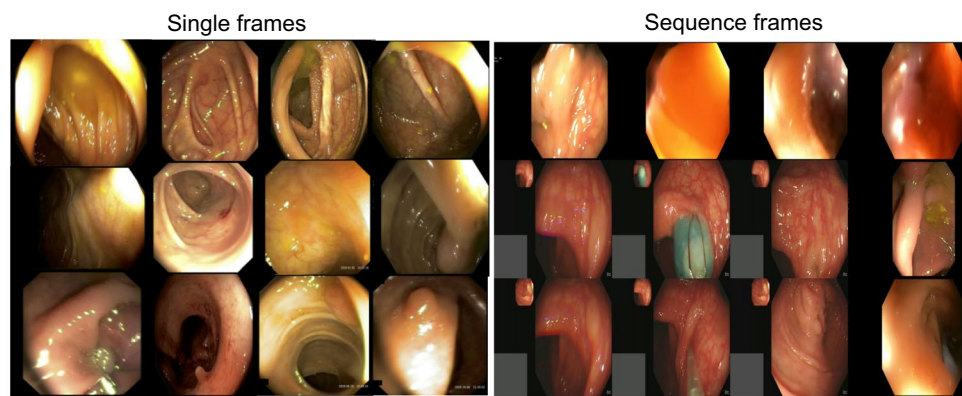
Fig. 9 Best and worse samples for single and sequence frames. 12 images for top performing frames (**a**) with highest DSC scores and bottom performing frames (**b**) with lowest DSC.

backbone. However, when evaluated on boundary distance-based metric MASD, it can be observed (Tables 5, 6) that the ResNetUNet have desired lower values in both single (35.83 vs 41.04) and sequence (27.10 vs 28.19) data compared to the DeepLabV3+. It can be inferred that MASD better captures the performance on small and medium sized polyps compared to other metrics.

The quantitative performances on this diverse multicentre dataset illustrates that utilising only baseline architectures such as vanilla UNet, PSPNet and FCN8 only provide sub-optimal results. This can be due to their model architectures that does not allow for capturing diverse polyp size and polyp variability and other changes such as color and contrast in images. However, using residual networks, atrous spatial pyramid pooling layers, and use of deeper backbones can improve generalisability of methods resulting in better performance on unseen centres. While small-sized polyp are generally ($\leq 100 \times 100$ pixels) performed poorly, these methods were able to capture the variabilities in both medium and large sized polyps. Additionally, as illustrated from our cross-validation results in Table 7 that even using networks that provide optimal results in this dataset, the choice of validation data affects the network performance largely.

From the qualitative results presented in Fig. 9, it can be concluded that the best performing frames in the single frames are those which have clear polyp views with appearances that is either very different to background

or at least appear as a lifted mucosa distinctly from the background mucosa. However, for the worse performing frames in the same, these appear to be most frames with either local strong illumination that confuses the network or embedded in the mucosa background color, i.e., almost flat or less protruded. Similarly, for the sequence frames, most frames that had good score were mostly with no polyp in it, while the networks failed to detect sessile or flat polyps.

Limitations of the dataset. The positive sample catalogue consists of both polyp and non-polyp images for completeness (see Fig. 2b). However, it has been made sure that these mimic the real-world dataset and taken such that non-polyp images are in close proximity to at least one polyp region. The entire dataset has been carefully reviewed by senior gastroenterologists. The accuracy, reliability and completeness of the annotations are subjective to the annotators. One additional limitation of this dataset is that the ambiguous annotations were mostly removed. For future versions, we will aim to quantify the level of disagreement among experts for each frame instead.

Usage Notes

All released dataset has been published under Creative Commons CC-BY licence. Dataset has been released only for educational, research and commercial purpose. Anyone using this dataset for their research or commercial application need to adhere to CC-by (Credit must be given to the creator) by citing this paper and acknowledging them.

The released dataset has been divided into positive and negative samples. Additionally, positive samples are divided into single frames and sequence frames. Users are free to use the samples according to their method demand. For example, for fully convolutional neural networks we adhere to the use of positive samples as done in our technical validation, while for the recurrent techniques that exploit temporal information users may use both positive and negative sequence data.

Code availability

To help users with the evaluate the generalizability of detection and segmentation method a code is available at: https://github.com/sharib-vision/EndoCV2021-polyp_det_seg_gen. The code also consists of inference codes that to assist in centre-based split analysis. Benchmark codes of the polypGen dataset with provided training and validation split in this paper for segmentation is also available at: https://github.com/sharib-vision/PolypGen-Benchmark. All the method codes are also available at different GitHub repositories provided in the Table 1.

References

1. Bray, F. *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* **68**, 394–424 (2018).
2. Leslie, A., Carey, F., Pratt, N. & Steele, R. The colorectal adenoma–carcinoma sequence. *British Journal of Surgery* **89**, 845–860 (2002).
3. Loeve, F. *et al*. National polyp study data: evidence for regression of adenomas. *International journal of cancer* **111**, 633–639 (2004).
4. Kaminski, M. F. *et al*. Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. *Gastroenterology* **153**, 98–105 (2017).
5. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet* **383**, 1490–502, https://doi.org/10.1016/S0140-6736(13)61649-9 (2014).
6. Hetzel, J. T. *et al*. Variation in the detection of serrated polyps in an average risk colorectal cancer screening cohort. *The American journal of gastroenterology* **105**, 2656 (2010).
7. Kahi, C. J., Hewett, D. G., Norton, D. L., Eckert, G. J. & Rex, D. K. Prevalence and variable detection of proximal colon serrated polyps during screening colonoscopy. *Clinical Gastroenterology and Hepatology* **9**, 42–46 (2011).
8. Zhao, S. *et al*. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: A systematic review and meta-analysis. *Gastroenterology* **156**, 1661–1674 e11, https://doi.org/10.1053/j.gastro.2019.01.260 (2019).
9. Van Doorn, S. C. *et al*. Polyp morphology: an interobserver evaluation for the paris classification among international experts. *American Journal of Gastroenterology* **110**, 180–187 (2015).
10. Saito, Y. *et al*. Multicenter trial to unify magnified nbi classification using web test system. *Intestine* **17**, 223–31 (2013).
11. Liu, W. *et al*. Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. *Saudi J Gastroenterol.* **26**, 13–19 (2020).
12. Murra-Saca, J. E Salvador atlas of Gastrointestinal Video Endoscopy online academic site as a learning resource. In *16th International Conference on Gastroenterology and Digestive Disorders* (2021).
13. Stiegmann, G. V. Atlas of Gastrointestinal Endoscopy. *Archives of Surgery* **123**, 1026–1026, https://doi.org/10.1001/archsurg.1988.01400320112031 (1988).
14. Mesejo, P. *et al*. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging* **35**, 2051–2063 (2016).
15. Ali, S. *et al*. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis* **70**, 102002, https://doi.org/10.1016/j.media.2021.102002 (2021).
16. Labelbox, https://labelbox.com/.
17. Ali, S. *et al*. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *ArXiv* **abs/2202.12031** (2022).
18. Ali, S. *et al*. PolypGen. *Synapse* https://doi.org/10.7303/syn26376615 (2021).
19. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
20. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
21. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
22. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).

23. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818 (2018).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
25. Reddi, S. J. *et al.* Adaptive federated optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (OpenReview.net, 2021).
26. Nikolov, S. *et al.* Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J Med Internet Res* **23**, e26151, https://doi.org/10.2196/26151 (2021).
27. Sevastopolsky, A. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. *Pattern Recognition and Image Analysis* **27**, 618–624 (2017).
28. Ali, S. *et al.* An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports* **10**, 2748, https://doi.org/10.1038/s41598-020-59413-5 (2020).
29. Ozturk, O., Saritürk, B. & Seker, D. Z. Comparison of fully convolutional networks (FCN) and U-Net for road segmentation from high resolution imageries. *International journal of environment and geoinformatics* **7**, 272–279 (2020).
30. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* **15**, 749–753 (2018).
31. Guo, Y., Bernal, J. & Matuszewski, J. B. Polyp segmentation with fully convolutional deep neural networks–extended evaluation study. *Journal of Imaging* **6**, 69 (2020).
32. Jha, D. *et al.* Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access* **9**, 40496–40510 (2021).
33. Nguyen, N.-Q., Vo, D. M. & Lee, S.-W. Contour-aware polyp segmentation in colonoscopy images using detailed upsamling encoder-decoder networks. *IEEE Access* (2020).
34. Jha, D. *et al.* KVASIR-SEG: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462 (Springer, 2020).
35. Borgli, H. *et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* **7**, 1–14 (2020).
36. Smedsrud, P. H. *et al.* Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data* **8**, 1–10 (2021).
37. Bernal, J., Sánchez, J. & Vilarino, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* **45**, 3166–3182 (2012).
38. Silva, J., Histace, A., Romain, O., Dray, X. & Granado, B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery* **9**, 283–293 (2014).
39. Ali, S. *et al* (eds.). *Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020*, vol. 2595 of *CEUR Workshop Proceedings* (CEUR-WS.org, 2020).
40. Bernal, J. *et al.* Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* **43**, 99–111 (2015).
41. Bernal, J. & Aymeric, H. MICCAI endoscopic vision challenge polyp detection and segmentation (2017).
42. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging* **35**, 630–644 (2015).
43. Koulaouzidis, A. *et al.* Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open* **5**, E477 (2017).
44. Ali, S., Ghatwary, N. M., Jha, D. & Halvorsen, P. (eds.). *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) co-located with with the 18th IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France, April 13, 2021*, vol. 2886 of *CEUR Workshop Proceedings* (CEUR-WS.org, 2021).

Acknowledgements

Author contributions

S. Ali conceptualized, initiated, and coordinated the work. He led the data collection, curation, and annotation processes and conducted most of the analyses and writing of the paper. T. de Lange assisted in writing of the introduction, clinical correctness of the paper and provided feedback regarding description of sequences presented in the manuscript. D. Jha and N. Ghatwary assisted in data annotation and parts of technical validation. S. Realdon, R. Cannizzaro, O. Salem, D. Lamarque, C. Daul, T. de Lange, M. Riegler, P. Halvorsen, K. Anonsen, J. Rittscher, and J. East were involved directly or indirectly in facilitating the video and image data from their respective centres. Senior gastroenterologists and collaborators S. Realdon, R. Cannizzaro, O. Salem, D. Lamarque, T. de Lange, and J. East provided timely review of the annotations and required feedback during dataset preparation. All authors read the manuscript, provided substantial feedback, and agreed for submission.

Competing interests

J. E. East has served on clinical advisory board for Lumendi, Boston Scientific and Paion; Clinical advisory board and ownership, Satisfai Health; Speaker fees, Falk. A. Petlund is the CEO and T. de Lange serves as chief medical scientist at Augere Medical, Oslo, Norway. All other authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Correspondence and requests for materials should be addressed to S.A.

Reprints and permissions information is available at www.nature.com/reprints.

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