A translational pathway of deep learning methods in GastroIntestinal Endoscopy

Sharib Ali[1,3,*], Mariia Dmitrieva[1], Noha Ghatwary[7], Sophia Bano[8], Gorkem Polat[10], Alptekin Temizel[10], Adrian Krenzer[11], Amar Hekalo[11], Yun Bo Guo[12], Bogdan Matuszewski[12], Mourad Gridach[24], Irina Voiculescu[13], Vishnusai Yoganand[14], Arnav Chavan[15], Aryan Raj[15], Nhan T. Nguyen[16], Dat Q. Tran[16], Le Duy Huynh[17], Nicolas Boutry[17], Shahadate Rezvy[18], Haijian Chen[19], Yoon Ho Choi[20], Anand Subramanian[21], Velmurugan Balasubramanian[22], Xiaohong W. Gao[18], Hongyu Hu[23], Yusheng Liao[23], Danail Stoyanov[8], Christian Daul[9], Stefano Realdon[5], Renato Cannizzaro[6], Dominique Lamarque[4], Terry Tran-Nguyen[2], Adam Bailey[2,3], Barbara Braden[2,3], James East[2,3], and Jens Rittscher[1]

[1]Institute of Biomedical Engineering and Big Data Institute, Old Road Campus, University of Oxford, Oxford, UK
[2]Translational Gastroenterology Unit, Experimental Medicine Div., John Radcliffe Hospital, University of Oxford, Oxford, UK
[3]Oxford NIHR Biomedical Research Centre, Oxford, UK
[4]Université de Versailles St-Quentin en Yvelines, Hôpital Ambroise Paré, France
[5]Instituto Onclologico Veneto, IOV-IRCCS, Padova, Italy
[6]CRO Centro Riferimento Oncologico IRCCS Aviano Italy
[7]Computer Engineering Department, Arab Academy for Science and Technology, Alexandria, Egypt
[8]Wellcome/EPSRC Centre for Interventional and Surgical Sciences(WEISS) and Department of Computer Science, University College London, London, UK
[9]CRAN UMR 7039, University of Lorraine, CNRS, Nancy, France
[10]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
[11]Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany
[12]School of Engineering, University of Central Lancashire, UK
[13]Department of Computer Science, University of Oxford, UK
[14]Mimyk Medical Simulations Pvt Ltd, Indian Institute of Science, Bengaluru, India
[15]Indian Institute of Technology (ISM), Dhanbad, India
[16]Medical Imaging Department, Vingroup Big Data Institute (VinBDI), Hanoi, Vietnam
[17]EPITA Research and Development Laboratory (LRDE), F-94270 Le Kremlin-Bicêtre, France
[18]School of Science and Technology, Middlesex University London, UK
[19]Department of Computer Science, School of Informatics, Xiamen University, China
[20]Dept. of Health Sciences & Tech., Samsung Advanced Institute for Health Sciences & Tech. (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea
[21]Claritrics India Pvt Ltd, Chennai, India
[22]School of Medical Science and Technology, Indian Institute of Technology, Kharagpur, West Bengal, India
[23]Shanghai Jiaotong University, Shanghai, China
[24]Ibn Zohr University, Computer Science HIT, Agadir, Morocco
[*]sharib.ali@eng.ox.ac.uk

ABSTRACT

The Endoscopy Computer Vision Challenge (EndoCV) is a crowd-sourcing initiative to address eminent problems in developing reliable computer aided detection and diagnosis endoscopy systems and suggest a pathway for clinical translation of technologies. Whilst endoscopy is a widely used diagnostic and treatment tool for hollow-organs, there are several core challenges often faced by endoscopists, mainly: 1) presence of multi-class artefacts that hinder their visual interpretation, and 2) difficulty in identifying subtle precancerous precursors and cancer abnormalities. Artefacts often affect the robustness of deep learning methods applied to the gastrointestinal tract organs as they can be confused with tissue of interest. EndoCV2020 challenges are designed to address research questions in these remits. In this paper, we present a summary of methods developed by the top 17 teams and provide an objective comparison of state-of-the-art methods and methods designed by the participants for two sub-challenges: i) artefact detection and segmentation (EAD2020), and ii) disease detection and segmentation (EDD2020). Multi-center, multi-organ, multi-class, and multi-modal clinical endoscopy datasets were compiled for both EAD2020 and EDD2020 sub-challenges. An out-of-sample generalisation ability of detection algorithms was also evaluated. Whilst most teams focused on accuracy improvements, only a few methods hold credibility for clinical usability. The best performing teams provided solutions to tackle class imbalance, and variabilities in size, origin, modality and occurrences by exploring data augmentation, data fusion, and optimal class thresholding techniques.

1 Introduction

Endoscopy is a widely used imaging technique for both diagnosis and treatment of patients with complications in hollow organs such as oesophagus, stomach, colon, bladder, kidney and nasopharynx. During the endoscopic procedure, an endoscope, a long thin tube with a light source and a camera at its tip, is inserted into the organ cavity. The imaging procedure is usually displayed on a monitor on-the-fly and is often recorded for post analysis. Each organ imposes very specific constraints to the use of endoscopes, but the most common obstructions in all endoscopic surveillance consists of artefacts caused by motion, specularities, low contrast, bubbles, debris, bodily fluid and blood. These artefacts hinder the visual interpretation of clinical endoscopists[1]. Missed detection rates of precancerous and cancerous lesions are another limitation. Gastrointestinal (GI) cancer (especially colo-rectal cancer) has high mortality rates and 5-year relative survival rates for stage IIB is around 65%[2]. In general, the missed detection rates in endoscopic surveillance is considerably high, at over 15%[3]. Therefore, the requirement for technology that can be effectively used in clinical settings during endoscopy imaging is necessary.

While a dedicated endoscopic procedure is followed for each specific organ, often these procedures are very similar, in particular for the GI tract organs like the oesophagus, stomach, small intestine, colon and rectum. Notably, some precancerous abnormalities such as inflammation or dysplasia and even cancer lesions in these GI organs naturally look very similar. Often automated methods are only trained for a specific abnormality, organ and imaging modality[4], whereas multiple different types of abnormalities can be present in different organs and several imaging protocols are used during endoscopy. Also, methods that are built for colonoscopy cannot be used during a gastroscopy (in the oesophagus, stomach and small intestine), despite the nature and occurrence of many abnormalities being similar in these organs. Artefacts are prevalent in all endoscopy surveillance and are usually confused with lesions, which can lead to unreliable outcomes.

A pathway to develop and reliably deploy methods in clinical settings is by benchmarking methods on a curated multi-center, multi-modal, multi-organ and multi-disease dataset and through a thorough evaluation of built methods using standard imaging metrics and metrics that can test their clinical applicability, for example ranking based on accuracy, robustness and computational efficiency[1]. Most publicly available datasets are specific to a particular organ, modality or a single abnormality class, e.g., polyp detection and segmentation challenges[5,6]. While dedicated organ specific challenges help to identify one particular disease type, they do not resemble the clinical workflow where the endoscopists are interested in biopsy and treatment of such abnormalities when of potential threat. For polyp class, it is required to identify different stages of polyp such as benign, dysplastic or cancer. Recently, it was shown that polyps and artefacts can be confused mostly due to specularity[7]. Artefacts are the fundamental and inevitable issue in endoscopy that often add confusion in detecting tissue abnormalities in these organs. It is therefore vital to accelerate research in identifying these classes and restore frames where possible[8] or reduce the false detections by adding uncertainties for such confusions[7].

The Endoscopy Computer Vision Challenge (EndoCV2020)[1] is a crowd-sourcing initiative to address fundamental problems in clinical endoscopy and consists of: 1) Endoscopy artefact detection and segmentation (EAD2020), and 2) Endoscopy disease detection and segmentation (EDD2020). EndoCV2020 releases diverse datasets that include multi-center, multi-modal, multi-organ, multi-disease/abnormality, and multi-class artefacts. Among the two sub-challenges, EAD2020 is an extended sub-challenge of EAD2019[9], however, unlike EAD2019 it includes both frame and sequence data with an addition of nearly 1000 frames and 41,832 annotations.

In this paper, we summarise and analyse the results of the top 17 (out of 43) teams participating in the EndoCV2020 challenge. Additionally, we benchmark these methods with the current state-of-the-art detection and segmentation methods.

[1] https://endocv.grand-challenge.org

Each method is also evaluated for its efficacy to detect and segment multi-class instances. In addition to the standard computer vision metrics used to evaluate methods during the challenge, we perform a holistic analysis of individual methods to measure their clinical applicability.

2 Related work

With the advancements in deep learning for computer vision, object detection and segmentation algorithms have shown rapid development in recent years. This is due to the hidden feature representations provided by Convolutional Neutral Networks (CNNs) that show significant improvement over hand-crafted features. CNN-based methods quickly gained the attention of the Medical Imaging community and are now widely used for automating the diagnosis and treatment for a range to imaging modalities, e.g. radiographs, CT, MRI, and endoscopy imaging. Below we present an overview of the recent deep learning-based object detection and segmentation techniques and discuss the related work in the context to medical image analysis with a particular focus on endoscopy imaging applications.

2.1 Detection and localisation

Object detection and localisation refers to determining the instances of an object (from a list of predefined object categories) that exist in an image. Object detection approaches can be broadly divided into three categories: single-stage, multi-stage and anchor-free detectors. A brief survey of these is presented below.

Single-stage detectors Single-stage networks perform a single pass on the data and incorporate anchor boxes to tackle multiple object detection on the same image grid such as in YOLO-v2[10]. Similarly, Liu et al.[11] proposed the Single Shot MultiBox Detector (SSD) with additional layers to allow detection of multiple scales and aspect ratios. RetinaNet was introduced by Lin et al.[12] where the authors introduced focal loss that puts the focus on the sparse hard examples enabling a boost in performance and speed.

The domain of Gastroenterology has started to benefit from the success of single-stage object detectors. Wang et al.[13] proposed a model that is based on SegNet[14] architecture to detect polyps during colonoscopy. Urban et al.[15] used YOLO to detect polyps from colonoscopy images in real-time. Horie et al.[16] used SSD to detect superficial and advanced oesophagal cancer. RetinaNet was the most popular detector in the first EAD challenge held in 2019. RetinaNet detector with focal loss was used by some top performing teams[17,18]

Multi-stage detectors Multi-stage detectors use a region proposal network to find regions of interest for objects and then a classifier to refine the search to get the final predictions. A two-stage architecture R-CNN using the classical region proposal method was proposed by Girshick et al.[19] whose speed was improved later by integrating an end-to-end trainable region proposal network (RPN), widely known as Faster R-CNN[20]. Due to the high precision of the Faster R-CNN, its architecture has become the base for many successful models in the object detection and segmentation domains, such as Cascade R-CNN[21] and Mask R-CNN[22]. Although these two-stage networks have shown successful results on public datasets such as Pascal VOC[23] and COCO[24], they are slow compared to the single-stage object detectors due to their region proposal mechanism.

In the field of Gastroenterology, Yamada et al.[25] used Faster R-CNN with VGG16 as the backbone to detect challenging lesions which are generally missed by colonoscopy procedures. Their reported prediction speed was not suitable for real-time examination. Shin et.al.[26] detected Polyps using the Fast R-CNN architecture with a region proposal network and an inception ResNet backbone. The two-stage detectors tend to yield better results than their single-stage contemporaries and have performed better at medical image analysis challenges. In the EAD2019 challenge, the top performing team[27] used a Cascade R-CNN with a FPN module and a ResNet backbone. Similarly, Pengyi Zhang et.al.[28] who used Mask aided R-CNN with an ensemble of different ResNet backbones finished second.

Anchor-free detectors A newly emerging detector type are the anchor-free detectors. Single and multi-stage detectors rely on the presence of anchors. Anchor free architectures claim to detect objects while skipping the process of anchor definition. They rely on a different geometrical characteristics like the center or corner points of objects[29,30]. Duan et al.[30] utilized the upper left and lower right corner to mark an object. The authors used classical backbones to generate a heatmap from the feature map showing potential spots of the object corners. A corner pooling technique was then used to create the classic bounding box of object detection. Zhou et al.[31] used a similar approach but instead they used a single point as the center of the bounding box.

Because of real-time dependencies in medical applications like the detection of polyps which have to be removed directly[32], anchor-free detectors are receiving more attention. Wang et al.[32] designed an anchor-free automatic polyp detector which achieved the state-of-the-art results while maintaining real-time applicability. Liu et al.[33] showed an anchor-free detector with state of the art performance while maintaining real-time performance.

2.2 Semantic segmentation

Semantic segmentation involves pixel-level partitioning of an image into multiple segments where each segment represents a pre-defined object or scene category. Based on the success of deep learning approaches on medical imaging data for segmentation, we can divide these approaches broadly into the following groups:

Models based on fully convolutional networks Fully Convolutional Network (FCN) architectures include only convolutional layers that enable them to take any arbitrary size input image to output a segmentation mask of the same size. These models are mostly based on the architecture developed by Long et al.[34] for semantic image segmentation.

Sun et al.[35] proposed a multi-channel FCN (MC-FCN) to segment liver tumors from multi-phase contrast-enhanced CT images. Kaul et al.[36] proposed FocusNet for skin cancer and lung lesion segmentation. A benchmark study for polyp segmentation using FCNs was conducted by[37]. Similarly, Patrick et al.[38] used FCN architecture with VGG backbone for a polyp segmentation task. The same group explored integration of depth information to improve segmentation accuracy in their FCN-based model[39].

Models based on encoder-decoder architecture U-Net[40], an encoder-decoder architecture, has become widely popular in medical image analysis community. U-Net based models have shown tremendous success, from cell segmentation[41] to liver tumor segmentation[42] and beyond[43,44].

In endoscopy imaging, U-Net-based models were used for instrument segmentation on GI endoscopy data[45]. Khan and Choo[46] developed a model based on U-Net architecture for endoscopy artefact segmentation. Bano et al.[47] directly used U-Net architecture for segmenting placental vessels from Fetoscopy imaging. Motion induced segmentation exploiting U-Net in the framework was used to segment kidney stones in the Uteroscopy data[48].

Models based on pyramid-based architecture In both detection and segmentation tasks, a crucial part is being able to identify objects and features of varying scales and sizes. One approach to this problem is to incorporate convolutional feature maps of varying resolutions during classification, which yields information about different scales of the image, making it easier to detect both small and big objects. Such architectures are referred to as *pyramid networks*. PSPNet[49] is one of such design that incorporates global context information for the task of scene parsing using a pyramid pooling module. A similar pyramid-based approach can be found in the task of object detection with Feature Pyramid Network (FPN)[50]. FPN extracts feature maps on a per-resolution-basis from the two bottom-up and top-down pathways of a pretrained architecture. The output maps can then be upsampled and concatenated to output a segmentation map[51].

Guo et al.[52] used PSPNet as part of an ensemble model including a U-Net and SegNet architecture for the task of automated polyp segmentation in colonoscopy images. Jia et al.[53] trained a two-stage polyp detector named PLPNet which utilizes FPN for multiscale feature representation using both CVC-ColonDB and CVC-ClinicDB. Their experimental results show that PLPNet outperforms other architectures in most regions on CVC-612 dataset and performs similarly on the ETIS dataset. Zhang and Xie[54] utilized an FPN combined with a Cascade R-CNN for artefact detection in endoscopic images.

Models based on dilated convolution architecture One of the challenges in the construction of semantic segmentation networks is to effectively control the size of the receptive field, providing adequate contextual information for pixel-level decisions while, at the same time, maintaining high spatial resolution and computational efficiency. The *dilated* or *atrous* convolution was proposed to address these challenges[55]. Chen at al.[56] proposed a family of very effective semantic segmentation architectures, collectively named DeepLab (also an *encoder-decoder* network), all using the dilated convolution. DeepLabv3+ uses atrous kernels within the spatial pyramid pooling (ASPP) module and depth-wise separable convolution to improve the computational efficiency.

Guo et al.[57] proposed the fully convolutional network based on atrous kernels to segment polyps in endoscopy images, with their network winning the GIANA 2017 challenge[6]. Nguyen et al.[58] augmented DeepLabv3+ architecture, showing its favourable performance when compared with other state-of-the-art methods on the CVC-ClinicDB and ETIS-Larib datasets. Ali et at.[59] used DeepLabv3+ with ResNet50 backbone to segment Barrett's area in esophageal endoscopy data. Yang and Cheng[60] developed a model based on DeepLabv3+ for multi-class artefact segmentation used with different backbone architectures.

3 The EndoCV challenge: Dataset, evaluation and submission

In this section, we present the dataset compiled for the EndoCV2020 challenge, the protocol used to obtain the ground truth for this data, evaluation metrics that were defined to assess participants methods and a brief summary on the challenge setup and ranking procedure.

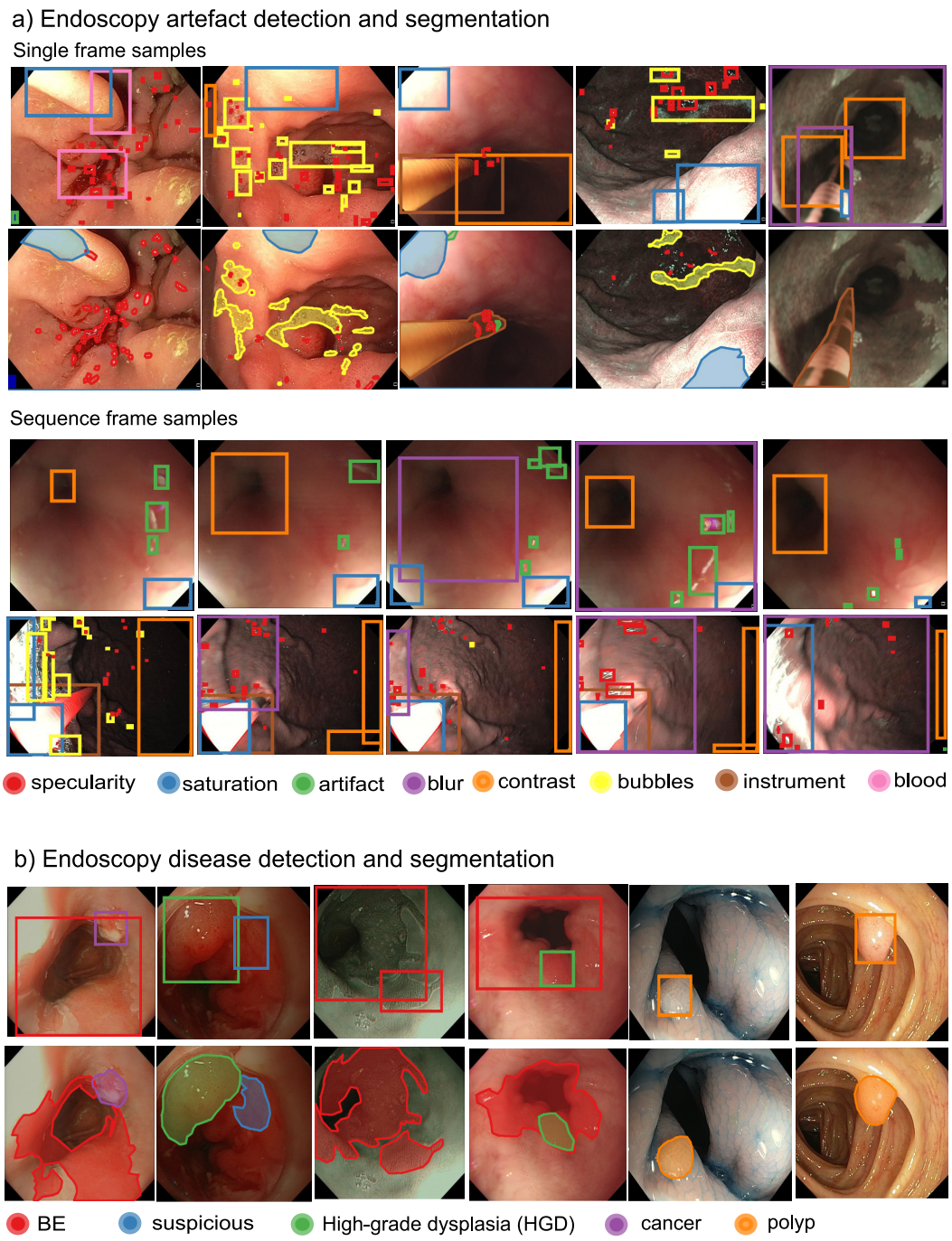

Figure 1. EndoCV2020 train data samples. (a) Endoscopy artefact detection and segmentation sub-challenge (EAD2020) samples. Both single frame samples (top) and sequence frames (bottom) were released. While detection annotations involve 8 classes, segmentattion classes were limited to 5 distinct class instances, mostly large indefinable shapes that include specularity, ,saturation, imaging artefact, bubbles and instrument. It can be observed that for sequence data most artefact instances follow upto few sequential frames so it is desirable to achieve such training datasets. (b) Endoscopy disease detection and segmentation training samples for sub-challenge EDD2020. First four samples belong to oesophageal endoscopy while last two frames were acquired during colonoscopy. It can be observed that disease classes in oesophagus confuse often, mostly the patient choice here is Barrett's where clearly suspected and high-grade dysplasia appear jointly. Similarly, for colonoscopy data protuded polyps are can easily be confused with the surrounding ridge-like openings and specular areas.

Table 1. Breakdown of number of samples and annotations released for EndoCV2020 challenge.

EndoCV	Tasks	# of classes	# of frames		# of annotations	
			Train	Test	Train	Test
EAD2020	Detection task	8	single: 2299 seq.: 233	single: 237 seq.: 80	31069	7750
	Segmentation task	5	573	162	3248	398
	Generalisation task	8	na	99	na	3013
EDD2020	Detection task	5	386	43	749	68
	Segmentation task	5	386	43	749	68

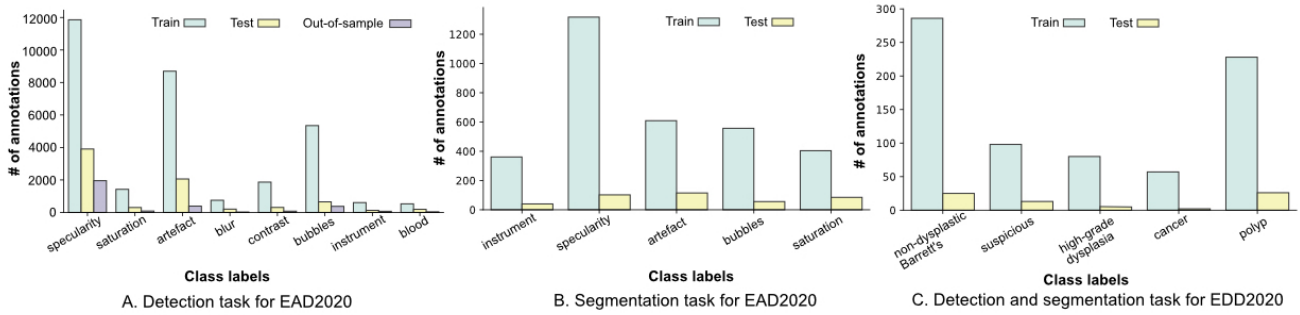

Figure 2. EndoCV2020 train and test per-class sample proportion: Train and test annotations for sub-challenge on artefact (A,B) and disease (C) detection and segmentation for each class label.

3.1 Dataset and challenge tasks

The EndoCV2020 challenge consists of two sub-challenges critical in clinical endoscopy. The EAD2020[2] sub-challenge comprises of diverse endoscopy video frames collected from seven institutions worldwide, including three different modalities and five different human organs (see **Supplementary figure**). Endoscopy video frames were annotated for detection and localisation of eight different artefact class occurrences identified by clinical experts in the challenge team. These include specularity, saturation, misc. artefacts, blur, contrast, bubbles, instrument and blood. Over 45,478 annotations were performed for this challenge on both single frame and sequence video data. Example annotations are shown in Figure 1. Training data for the detection task consisted of 2,532 frames with 31,069 bounding boxes while 573 frames with 16,240 binary masks were released for the segmentation task (except for blur, blood and contrast). Due to the nature of occurrence of various artefact classes (see Figure 1), the proportion of annotations for each class is different (Figure 2). However, the proportion of training and test samples per-class were matched in the test data.

Separately, EDD2020[3] is a new disease detection and segmentation sub-challenge that consists of five disease categories[61]. The provided training set consisted of total 385 video frames with 817 individual annotations. The annotations included non-dysplastic Barrett's oesophagus (NDBE), suspicious, high-grade dysplasia (HGD), cancer, and polyp categories (also see Figure 1). These disease classes were from three different endoscopic modalities (white light, narrow-band imaging, and chromoendoscopy) acquired from four different clinical centers, investigating four different GI organs. To our knowledge, this is the first comprehensive dataset for the multi-class detection and segmentation tasks. More details on the dataset are provided in the **Supplementary figure**. The detailed breakdown of training set and test set for each specific task is provided in Table 1.

EndoCV2020 posed three specific challenge tasks (see Figure 3) that included: 1) detection and localisation task, 2) semantic segmentation task and 3) out-of-sample generalisation task. For detection and generalisation tasks participants were provided with both frame label annotations for single and sequence images for the EAD2020 challenge while only single frames were released for EDD2020. The generalisation task was only evaluated for the EAD2020 and only consisted of test data from an unseen institution that was not present in any training set. It is to be noted that test samples for all other tasks were taken from different patients as well even though they were collected from the same centers as that in the training set. EAD2020 attracted nearly 700 participants with 29 teams on the leaderboard and EDD2020 recorded nearly 550 participants with 14 teams on the leaderboard. Participation was permitted in either one or both sub-challenges.

[2] https://ead2020.grand-challenge.org
[3] https://edd2020.grand-challenge.org

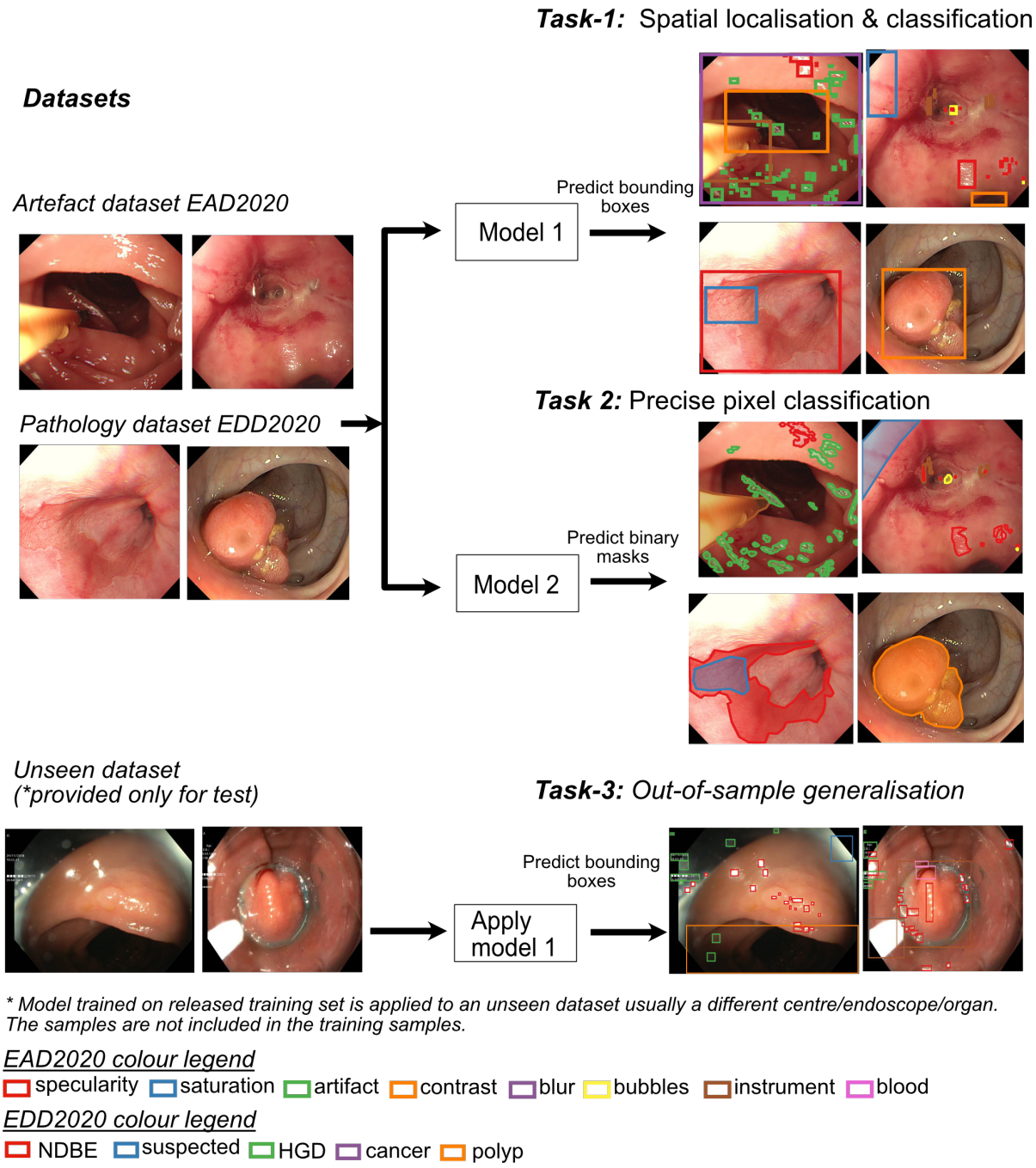

Figure 3. EndoCV2020 challenge task descriptions for each sub-challenge. The three tasks of the EndoCV2020 challenge includes: (a) The "detection" task aimed at the coarse localization and classification. Given an input image (left) a detection model (middle) outputs the artefact/disease class and coordinates of the containing bounding box. (b) The "segmentation" task is aimed at precise delineation of artefact/disease object boundaries. The model predicts binary output images denoting the presence ('1') or absence ('0') of each class. (c) The "out-of-sample generalization" task is aimed at assessing the ability of a model trained on different dataset to generalise on an unseen dataset usually coming from a different center.

3.1.1 Annotation protocol

A team of two clinical experts determined the class labels for the artefact detection challenge while for the disease detection challenge we consulted with four clinical experts regarding the class labels in the GI tract endoscopy. For each sub-challenge at least two clinical experts sampled the video frames from a small sub-set of video data collected from various institutions and multi-patient data cohort (see **Supplementary figure**). These frames were then taken as reference to produce bounding box

annotations for the remaining train-test dataset by four experienced postdoctoral fellows. Finally, further validation by three clinical endoscopists independently was carried out to assure the reference standard. The ground-truth labels were randomly sampled (1 per 20 frames) during this process. However, after the completion of this phase the entire annotation was discussed and reviewed together with the clinical experts. The priorities was given to indecisive frame annotations to have a collective opinion from experts. Following annotation strategies were used by clinical experts and researchers:

- For the same region, multiple boxes (for detection/generalisation) or pixel-wise delineation (for semantic segmentation) were performed if the region belonged to more than 1 class
- The minimal box sizes were used to describe the class region and similarly possible small annotation areas for semantic segmentation were merged instead of having multiple small boxes/regions
- Each class type was determined to be distinctive and general across all datasets

3.2 Evaluation metrics

The challenge problems fall into three distinct categories. For each there already exist well-defined evaluation metrics used by the wider imaging community which we use for evaluation here. Codes related to all evaluation metrics used in this challenge are also available online[4].

3.2.1 Spatial localisation and classification task

Metrics used for multi-class disease detection:

- IoU - intersection over union: This metric measures the overlap between two bounding boxes A and B, where A is segmented region and B is annotated GT. It is evaluated as the ratio between the overlapped area $A \cap B$ over the total area $A \cup B$ occupied by the two boxes:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \tag{1}$$

  where $\cap$, $\cup$ denote the intersection and union respectively. In terms of numbers of true positives (TP), false positives (FP) and false negatives (FN), IoU (aka Jaccard J) can be defined as:

$$IoU/J = \frac{TP}{TP+FP+FN} \tag{2}$$

- mAP - mean average precision: mAP of detected class instances is evaluated based on precision (p) defined as $p = \frac{TP}{TP+FP}$ and recall (r) as $r = \frac{TP}{TP+FN}$. This metric measures the ability of an object detector to accurately retrieve all instances of the ground truth bounding boxes. Average precision (AP) is computed as the Area Under Curve (AUC) of the precision-recall curve of detection sampled at all unique recall values $(r_1, r_2, ...)$ whenever the maximum precision value drops:

$$\text{AP} = \sum_n \left\{ (r_{n+1} - r_n)\, p_{\text{interp}}(r_{n+1}) \right\}, \tag{3}$$

  with $p_{\text{interp}}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$. Here, $p(r_n)$ denotes the precision value at a given recall value. This definition ensures monotonically decreasing precision. The mAP is the mean of AP over all N classes given as

$$\text{mAP} = \frac{1}{N} \sum_{i=0}^{N} \text{AP}_i \tag{4}$$

  This definition was popularised in the PASCAL VOC challenge[23]. The final mAP (mAP_d) was computed as an average mAPs for IoU from 0.25 to 0.75 with a step-size of 0.05 which means an average over 11 IoU levels is used for 5 categories in the competition (mAP $@[.25 : .05 : .75]$).

  Participants were finally ranked on a final mean score (score_d), a weighted score of mAP and IoU represented as:

$$\text{score}_\text{d} = 0.6 \times \text{mAP}_\text{d} + 0.4 \times \text{IoU}_\text{d} \tag{5}$$

Standard deviation between the computed mAPs ($\pm\sigma_{score_d}$) are taken into account when the participants have the same score_d. Scores on both single frame data and sequence data were first separately computed and then averaged to get the final $score_d$ of the detection task.

[4] https://github.com/sharibox/EndoCV2020

3.2.2 Segmentation task

Metrics widely used for multi-class semantic segmentation of disease classes have been used for scoring semantic segmentation. The final semantic score $score_s$ comprises of an average score of F_1-score (Dice Coefficient, DSC), F_2-score, precision and recall.

Precision, recall, F_β-scores: These measures evaluate the fraction of correctly predicted instances. Given a number of true instances #GT (ground-truth bounding boxes or pixels in image segmentation) and number of predicted instances #Pred by a method, precision is the fraction of predicted instances that were correctly found, $p = \frac{\#\text{TP}}{\#\text{Pred.}}$ where #TP denotes number of true positives and recall is the fraction of ground-truth instances that were correctly predicted, $r = \frac{\#\text{TP}}{\#\text{GT}}$. Ideally, the best methods should have jointly high precision and recall. F_β-scores gives a single score to capture this desirability through a weighted (β) harmonic means of precision and recall, $\text{F}_\beta = (1+\beta^2) \cdot \frac{p \cdot r}{(\beta^2 \cdot p)+r}$.

Participants are ranked based on the value of their semantic performance score given by:

$$\text{score}_\text{s} = 0.25 \times (p + r + \text{F}_1 + \text{F}_2) \tag{6}$$

Standard deviation between each of the subscores are computed and averaged to obtain the final $\pm\sigma_{score_s}$ which is used during evaluation for participants with same final semantics score.

3.2.3 Out-of-sample generalisation task

Out-of-sample generalization of disease detection is defined as the ability of an algorithm to achieve similar performance when applied to a completely different institution data. To assess this, participants were challenged to apply their trained models on video frames that were neither included in the training nor in the test data of the other tasks. Assuming that participants applied the same trained weights, the out-of-sample generalization ability was estimated as the mean deviation between the mAP score of the detection and out-of-sample generalization test datasets of each class i for deviation greater than a tolerance of $\{0.1 \times \text{mAP}_\text{d}^i\}$.

$$\text{dev}_\text{g} = \frac{1}{N} \sum_i {\text{dev}_\text{g}}^i \tag{7}$$

$${\text{dev}_\text{g}}^i = \begin{cases} 0, & \text{for } |{\text{mAP}_\text{d}}^i - {\text{mAP}_\text{g}}^i| / {\text{mAP}_\text{d}}^i \leq 0.1 \\ |{\text{mAP}_\text{d}}^i - {\text{mAP}_\text{g}}^i|, & \text{for } |{\text{mAP}_\text{d}}^i - {\text{mAP}_\text{g}}^i| / {\text{mAP}_\text{d}}^i > 0.1 \end{cases} \tag{8}$$

The best algorithm should have high mAP_g and low $\text{dev}_\text{g} (\to 0)$. Participants were finally ranked using a weighted ranking *score for out-of-sample generalisation* as $\text{R}_\text{gen} = 1/3 \cdot \text{Rank}(\text{dev}_\text{g}) + 2/3 \cdot \text{Rank}(\text{mAP}_\text{g})$ where $\text{Rank}(\text{mAP}_\text{g})$ is the rank of a participant when sorted by mAP_g in ascending order.

3.3 Challenge setup, and ranking procedure

The challenge proposal was submitted to the IEEE ISBI challenge organisers and was peer-reviewed by two reviewers. Upon the acceptance, the challenge website[5] was launched on 1st November 2019. Training datasets for each sub-challenge (EAD and EDD) were first provided (via AWS amazon S3 for EAD data and IEEE data portal for EDD data[6]). The test data was released nearly 20 days before the leaderboard closing through a docker container set-up. A docker based online leaderboard was established separately for EAD2020[7] and EDD2020[8] where each participating team was allowed to submit a maximum of 2 submissions per day on the final test data. A wiki-page[9] was set-up for the submission guidelines and a code repository with evaluation metrics used in the challenge was also provided[10].

For the ranking of different task categories, we used the metrics described in the above section 3.2. The participants were able to see only the final score in the leaderboard and all other sub-scores were hidden for the final test data. This was done to avoid any class specific refinement on the released test set. Notably, the detection task was bounded by two IoU thresholds (mAP @ IoU thresholds $[.25 : .05 : .75]$) and the overall IoU scores itself. For the detection task, participants were ranked on a final weighted score of mAP and IoU (see Eq. (5)), while for the segmentation task, participants were ranked based on a final weighted average of DSC or F1-score, F2-score, precision and recall (see Eq. (6)). For the generalisation task, both the mAP score gap dev_g and mAP on generalisation data mAP_g were taken into account.

4 Method summary of the participants

[5] https://endocv.grand-challenge.org
[6] https://ieee-dataport.org/competitions/endoscopy-disease-detection-and-segmentation-edd2020
[7] https://ead2020.grand-challenge.org/evaluation/leaderboard/
[8] https://edd2020.grand-challenge.org/evaluation/leaderboard/
[9] https://github.com/sharibox/EndoCV2020/wiki
[10] https://github.com/sharibox/EndoCV2020

Table 2. Endoscopy artefact detection and segmentation (EAD2020) method summary for top 13 teams (out-of 33 valid submissions).

Team EAD2020	Algorithm	Preprocessing	Nature	Basis-of-choice	Backbone	Data aug.	Pretrained	Computation		code
Detection								**GPU**	**Test time**	
polatgorkem (METU_DLCV)	Faster RCNN + CascadeRCNN + Retinanet	Resize Normalise	Ensemble	Accuracy++	ResNet50, ResNet101	Yes (R, F)†	COCO	RTX 2080	0.76	GorkemP/EAD
qzheng5 (CVML)	Faster RCNN	Resize Normalise	Context	Accuracy+	ResNet101	Yes (R, T, LD)†	COCO	GTX1060	0.20	CVML/EAD2020
xiaohong1	YOLACT + NMS-within-class	None	Context	Accuracy+ , speed+	ResNet101	None	ImageNet	Tesla K80	0.14	yolact
mathew666	Faster RCNN + NMS	None	Context	Accuracy+	ResNet101	Yes	NA	RTX 2080	NA	NA
VinBDI	EfficientDet D0	Resize (512x512)	Multiscale scalable	Speed++	EfficientNet B0	Yes (S, Sc, R, N, MU)†	COCO	RTX 2080TI	NA	endocv2020-seg
higersky	Cascade R-CNN	None	Cascading	Accuracy++	ResNeXt101	Yes	NA	GTX1080 Ti	NA	NA
StarStarG	Cascade R-CNN	Resize Normalise	Cascading	Accuracy++	ResNeXt101	Yes (F, S)†	NA	RTX 2080	NA	NA
anand_subu	RetinaNet	Resize Normalise	Context	Accuracy+ , speed+	ResNet101 (det); ResNet50 (gen)	Yes (R, Sh, F, C, B, St, H)†	ImageNet	GTX1050Ti	0.36 0.31	anand-subu/EAD2020
arnavchavan04	RetinaNet + FasterRCNN (FPN + DC5)	Resize (512x512)	Ensemble	Accuracy++	ResNet50; ResNeXt101	Yes (F, C, R)†	ImageNet	Tesla T4	NA	ubamba98/EAD2020
MXY	Cascase RCNN + FPN	Resize Normalise	Cascading	Accuracy+	ResNet101	Yes (F)†	ImageNet	RTX 2080 Ti	0.80	Carboxy/EAD2020
mimykgcp	Faster RCNN + + RetinaNet	Resize Normalise	Ensemble	Accuracy+ , speed+	ResNeXt101	Yes (RA)†	COCO	GTX 1080Ti	0.58	NA
DuyHUYNH (LRDE)	YOLOv3	Normalise	Multiscale	Accuracy+ , speed++	Darknet53	Yes (RA)†	COCO	GTX1080 Ti	0.07	dhuynh/endocv2020
Segmentation										
qzheng5 (CVML)	DeepLabv3+	Resize (513x513) Normalise	Encoder-decoder, mutiscale	Accuracy++	SE-ResNeXt50	(R, T, LD + TTA)†	ImageNet	GTX1080Ti	0.50; 5 (+TTA)	CVML/EAD2020
mouradai_ox	Pyramid dilated module	Resize (512x512) Normalise	Multiscale	Accuracy+ , speed+	ResNet50	Yes (T, R, LD)†	ImageNet	Colab	0.37	NA
arnavchavan04	FPN + EfficientNet	Resize (512x512)	Ensemble	Accuracy+	EfficientNet	Yes (F, C, R)†	ImageNet	Tesla T4	NA	ubamba98/EAD2020
VinBDI	U-Net + BiFPN	Resize (512x512)	Ensemble, Endcoder-decoder	Accuracy++ , speed+	EfficientNet B4; ResNet50	Yes (S, Sc, R, F)†	COCO ImageNet	RTX 2080TI	NA	endocv2020-seg
higersky	DeepLabv3+	None	Encoder-decoder, mutiscale	Accuracy+	ResNet101	Yes (F;S;Sc;Bl)†	ImageNet	GTX1080 Ti	NA	NA
anand_subu	U-Net	None	Encoder-decoder	Accuracy+	ResNet50	Yes (S, F, R, N, Cr, Bl, H, St, C, Sp)†	ImageNet	GTX1050Ti	0.17	anand-subu/EAD2020
DuyHUYNH (LRDE)	U-Net++	Normalise	Encoder-decoder	Accuracy+, speed+	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)†	ImageNet	GTX1080 Ti	0.97	dhuynh/endocv2020
mimykgcp	U-Net	Resize Normalise	Encoder-decoder	Accuracy+, speed+	ResNeXt50	Yes (RA)†	ImageNet	RTX 2070	0.25	NA

† B: brightness, C: contrast, F: Flip, H: hue, LD: Local deformation, N: noise, R: Rotation, RA: RandAugment, S: Shift, Sc: scaling Sh: shear, St: saturation, Mu: mixup, T: Translation, TTA: test-time augmentation

Table 3. Endoscopy disease detection and segmentation (EDD2020) method summary for top 7 teams (out-of 14 submission).

Team EDD2020	Algorithm	Preprocessing	Nature	Basis-of-choice	Backbone	Data aug.	Pretrained	Computation		code
Detection								**GPU**	**Test time**	
Adrian	YOLOv3+ Faster R-CNN	Resize	Ensemble	Accuracy+ , speed+	Darnet53 ResNet101	Yes (F, D)†	COCO public polyp dataset	Tesla P100	0.41	Adrian398/EDD
shahadate	Mask R-CNN	Resize Normalise	Multiscale	Accuracy , speed+	ResNet101	Yes (Sc, R, F, Cr, S, N)†	COCO	RTX2060	NA	EDD-Mask-rcnn
VinBDI	EfficientDet D0	Resize (512x512)	Ensemble	Speed++	EfficientNet B0	Yes (S, Sc, R, N, MU)†	COCO	RTX 2080TI	NA	endocv2020-seg
YH_Choi	CenterNet	NA	Context	Accuracy++	ResNet50	Yes(Du, R, F, C, B)†	PASCAL VOC2012	RTX 2080	2	NA
DuyHUYNH (LRDE)	U-Net++	Normalise	Encoder-decoder	Speed	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)†	ImageNet	GTX1080 Ti	1.53	dhuynh/endocv2020
mimykgcp (vishnusai)	Faster RCNN + RetinaNet	Resize (256x256) normalise	Ensemble	Accuracy+ , speed+	ResNeXt101	Yes (RA)†	COCO	GTX1080Ti	0.58	NA
Segmentation										
Adrian	YOLOv3 + Faster R-CNN + Cascade RCNN	Resize	Ensemble	Accuracy++	Darnet53 ResNet101	Yes (F, D)†	COCO public polyp dataset	Tesla P100		Adrian398/EDD2020
shahadate	MaskRCNN	Resize Normalise	Multiscale	Accuracy , speed+	ResNet101	Yes (Sc, R, F, Cr, S, N)†	COCO	RTX2060		EDD-Mask-rcnn
VinBDI	U-Net + BiFPN	Resized (512x512)	Ensemble Endcoder-decoder	Accuracy++ , speed+	EfficientNet B4 ResNet50	Yes (S, Sc, R, F)†	COCO ImageNet	RTX 2080 Ti	NA	endocv2020-seg
YH_Choi	U-Net	NA	Encoder-decoder	Accuracy+	ResNet50	Yes(Du, R, F, C, B)†	PASCAL VOC2012	RTX 2080	7	NA
DuyHUYNH (LRDE)	U-Net++	Normalise	Encoder-decoder	Accuracy+ , speed+	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)†	ImageNet	GTX1080 Ti	1.53	endocv2020
drvelmuruganb	SUMNet	NA	Encoder-decoder	Accuracy+ , speed++	VGG11	Yes(R, A, Sc, P, and Cr)†	ImageNet	GTX1080 Ti	0.16	drvelmuruganb/EDD2020
mimykgcp	U-Net	Resize Normalise	Encoder-decoder	Accuracy+	ResNeXt50	Yes (RA)†	ImageNet	RTX2070	1.25	NA

† A: affine, B: brightness, C: contrast, Cr: cropping, D: distortion, Du: duplication, F: flip, H: hue, LD: local deformation, Mu: mixup, N: noise, P: perspective transformation, R: rotation, RA: RandAugment library, S: shift, Sc: scaling, Sh: shear, St: saturation, T: translation, TTA: test-time augmentation

In this Section, we present summary of top participating teams for both EAD2020 and EDD2020 sub-challenges. Each of these teams has participated in either detection task or segmentation task or both.

4.1 EAD2020 Participating teams

- **Team *polatgorkem***[62] The team used an ensemble of three object detectors: Faster R-CNN (ResNet50 with FPN), Cascade R-CNN (ResNet50 with FPN), RetinaNet (ResNet101 with FPN). Class-agnostic NMS operation, where the model predictions were passed through the NMS procedure together for all classes, was applied to the output of each individual model. During ensemble, only the bounding boxes for which majority of the models agree were kept. False-positive elimination was applied as a post-processing step to eliminate same-type predicted boxes located close to each other. For each class, an IoU threshold was determined.

- **Team *CVML***[63] CVML team's model was inspired by DeepLabV3+. The team experimented with several changes including the backbone, the global pooling, the dilated kernels and the convolution kernels with dilation rates. Moreover, the squeeze-and-excitation module is added behind the balanced ASPP module to introduce attention gating at the output of the original encoder to better utilize the information available in the computed feature maps. In addition, the original multi-class classifier is replaced with 5 binary classifiers to enable segmentation of the overlapping objects. At test time, they used some post-processing techniques such as rotation, holes filling and removal of objects from the image boundary.

- **Team *mouradai_ox***[64] The team proposed a novel neural network called OxEndoNet to tackle the segmentation challenge. The network uses the pyramid dilated module (PDM) consisting of multiple dilated convolutions stacked in parallel. For each input image, pre-trained ResNet50 (on ImageNet) was used as the backbone to extract the feature map followed by multiple PDM layers to form an end-to-end trainable network. In the final architecture, they used four PDM layers; each layer used four parallel dilated convolutions with a filter size of 3×3 and dilation rates of 1, 2, 3, and 4. They fed the final PDM layer to a convolution layer followed by a bilinear interpolation to up-scale the feature map to the original image size.

- **Team *mimykgcp***[65] The team re-trained the ResNeXt101 backbone with the cardinality parameter set to 64. To enable detection of artefacts at different scales, an FPN was integrated into the object detectors. Data-Augmentation techniques based on RandAugment[66] were incorporated to improve the generalization capability. For the segmentation task, a U-Net with an ImageNet pre-trained ResNext50 backbone was used.

- **Team *DuyHUYNH***[67] For segmentation, the team exploited a model based on U-Net++ using pre-trained EfficientNet on ImageNet as the backbone. The model was trained to minimize F2-loss using the Adam optimizer. At the test-time the team used five transformations: horizontal, vertical flipping, and three rotations. For detection, the team used the bounding boxes deduced from the results of their segmentation model on the EDD dataset, while for EAD, they used YOLOv3 pre-trained on COCO.

- **Team *mathew666***[68]The team used Cascade RCNN architecture with the ResNeXt backbone in a FPN based feature extraction paradigm. Data augmentation with probability of 0.5 for horizontal flip was applied. The team also utilised multi-scale detection to tackle with variable sized object detection.

- **Team *arnavchavan04***[69] For the object detection task, the team used an ensemble of three models: Faster R-CNN (ResNext101 + FPN), RetinaNet (ResNet101 + FPN) and Faster R-CNN (ResNext101 + DC5). For the segmentation task, an ensemble of multiple depth EfficientNet models with FPN trained on multiple optimization plateaus (DICE, BCE, JACCARD) was designed. Data augmentation techniques like horizontal and vertical flip, cutout (random holes), random contrast, gamma, brightness, rotation along with CutMix[70] strategy for the segmentation task were incorporated to improve generalisation capability.

- **Team *anand_subu***[71] The team used RetinaNet with ResNet50 and ResNet101 backbones. For the segmentation task, the team used an ensemble network with U-Net with a ResNet50 backbone and DeepLabV3. All the backbones were pre-trained on the ImageNet. Real-time augmentation techniques like rotation, shear, random-image-flip, image contrast, brightness, saturation, and hue variations were incorporated while training to improve the generalization capability of the network.

- **Team *higersky***[72] The team implemented Hyper Task Cascade and Cascade R-CNN with ResNeXt101 backbone as a feature extractor and FPN module for multi-scale feature representation for the object detection task. They applied Soft-NMS[73] to avoid mistakenly discarded bounding-boxes. For the semantic segmentation task, the team incorporated

DeepLabV3+ with ResNet101 backbone and trained with BCE and DICE losses. The backbones for both tasks were pre-trained on ImageNet.

- **Team *MXY***[74] The team used a Cascade R-CNN with an ImageNet pre-trained ResNet101 backbone and a FPN module. Post-detection, soft-NMS was added to remove false predictions. The dataset was augmented by random resizing technique to improve the final output scores. The team used more weight for the losses of specularity, artefact, and bubbles classes to overcome classification difficulties between those classes.

- **Team *StarStarG*** The team used Cascade-RCNN as network architecture and adpoted COCO2017 pre-trained ResNeXt as backbone with FPN and multi-stage RCNN framework. The authors also integrated Deformable Convolutional Networks in backbone to improve model performance.

- **Tesam *xiaohong1***[75] The team built their detection and segmentation method upon Yolact-based instance segmentation system. Yolact[76] adds a segmentation component to the RetinaNet to ensure the tasks of detection, classification and delineation which are performed simultaneously. The network uses ResNet101 as an imageNet pretrained backbone.

4.2 EDD2020 Participating teams

- **Team *Adrian***[77] The team compared two different models: YOLOv3 with darknet-53 backbone and Faster R-CNN with ResNet-101 backbone. For post-processing, both algorithms in the final architecture were combined. For the second task, the team leveraged the state-of-the-art Cascade Mask R-CNN with ResNeXt-151 as a backbone. The team trained YOLOv3 using categorical cross-entropy for classification and default localization loss, while for Cascade Mask-RCNN, they used binary cross entropy for classification and mask, and L1 smooth for boundary box regression.

- **Team *Shahadate***[78] The team implemented a modified benchmark Mask R-CNN infrastructure model on the EDD2020 dataset. They used COCO trained weights and biases with the ResNet101 backbone as an initial feature extractor. The network head of the backbone model was replaced with new untrained layers that consisted of a fully-connected classifier with five classes and an additional background class. Non-maximum suppression was used to reduce overlapped detection. Finally, the team merged multiple bounding boxes for the same class label as one bounding box to match with the mask annotation.

- **Team *VinBDI***[79] For the object detection task, the team designed an ensemble of six EfficientDet models (with BiFPN modules) trained on six different EfficientNet backbones. A total of eleven augmentation techniques were incorporated to increase the output prediction scores of the model. For the segmentation task, an ensemble of U-Net and EfficientNet-B4 and BiFPN with the ResNet50 backbone was devised. The same team also participated in the EAD2020 sub-challenge.

- **Team *YH_Choi***[80] The team implemented a CenterNet-based model with the PASCAL VOC pretrained ResNet50 backbone for the object detection task. A similar backbone with U-Net was devised for the segmentation task. The dataset was randomly duplicated to tackle class-imbalance. To improve generalization performance, each image was augmented 86 times by randomly choosing augmentation techniques from the pool of rotation, flipping, contrast enhancement and brightness adjustment.

- **Team *drvelmuruganb***[81] For the segmentation of disease classes the team used an encoder-decoder based SUMNet architecture with the ImageNet pretrained VGG11 backbone. The authors also applied several augmentation strategies including variable brightness and HSV values, multiple crops and geometric transformations such as rotation, affine, scaling and projective were also applied to improve the accuracy.

5 Results

For the EAD2020 sub-challenge, we present the results of 12 participating teams for multi-class artefact detection task and 8 teams for segmentation task. Similarly, for EDD2020 sub-challenge, we have included top 6 teams for detection and 7 teams for segmentation of multi-class diseases. In this section we present the quantitative and qualitative results for each team based on the evaluation metrics discussed in Section 3.2. For the EAD2020 sub-challenge, 3 different test dataset were released: 1) single-frame data for detection and segmentation, 2) sequence dataset for detection only and 3) out-of-sample data for generalisation task only. For the detection task, the average of the aggregated sum of the detection scores for the single frame data and the sequence data were considered for final scoring. While, for the EDD2020 challenge only single frame detection and segmentation data were released. Below we present the result for each sub-challenges separately.

Table 4. EAD2020 results for the detection task on the single frame dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are computed at 11 IoU thresholds and averaged. Weighted detection score $score_d$ is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	$\mathbf{mAP}_{25}$	$\mathbf{mAP}_{50}$	$\mathbf{mAP}_{75}$	overall $\mathbf{mAP}_d$	overall $\mathbf{mIoU}_d$	$\mathbf{mAP}_\delta$	$\mathbf{score}_d \pm \delta$
polatgorkem	26.886	17.883	5.608	17.486	**36.579**	7.124	**25.123 ± 7.124**
qzheng5	33.134	20.084	5.570	19.720	27.185	8.82	22.706 ± 8.820
xiahong1	30.627	19.384	4.935	18.512	26.388	8.428	21.663 ± 8.428
mathew666	20.36	19.44	7.783	18.091	**32.692**	5.617	**23.931 ± 5.617**
VinBDI	38.429	25.426	7.053	24.069	12.644	**10.291**	19.499 ± 10.291
higersky	36.92	25.770	**9.452**	24.771	17.298	8.707	21.781 ± 8.707
StarStarG	**41.800**	**29.984**	**10.733**	**28.380**	16.250	**10.042**	23.528 ± 10.042
anand_subu	29.755	19.893	5.271	18.886	24.029	7.619	20.943 ± 7.619
arnavchavan04	**38.752**	**27.247**	**9.858**	**26.021**	21.165	9.342	**24.079 ± 9.342**
MXY	25.373	18.967	7.171	17.82	**28.056**	5.754	21.914 ± 5.754
mimykgcp	**39.897**	**26.296**	6.839	**25.082**	10.209	**10.765**	19.133 ± 10.765
DuyHUYNH	20.512	12.234	2.978	11.894	27.063	5.671	17.962 ± 5.671
baselines							
YOLOv3	22.798.	13.736	2.804	13.249	24.883	6.525	17.903 ± 6.525
RetinaNet (ResNet101)	15.27	8.927	2.061	8.754	23.202	4.275	14.533 ± 4.275

5.1 Quantitative results

5.1.1 EAD2020 sub-challenge

In this section, the results of the participant teams in the EAD2020 challenge to detect and segment artifacts are presented.

5.1.2 Detection task for EAD2020

Table 4 and Table 5 present the mAP values computed at different IoU thresholds (i.e., 25%, 50%, and 75%), overall mAP, overall IoU, and the final score for the detection of the artefacts from single frame and sequence data, respectively. Additionally, we also provide results of baseline methods that include YOLOv3 and RetinaNet with darknet53 and ResNet101 backbones, respectively. In Table 4 (i.e., single frame detection), it can be observed that the method presented by team *polatgorkem* surpassed the other teams by achieving the highest final score of 25.123 ± 7.124. Additionally, they gained an overall mIoU of 36.579 providing a high overlap ratio between the generated bounding box with ground truth per frame. The method proposed by the team *arnavchavan04* comes in the second place for the detection score on single frame data with a value of 24.079 ± 9.342. In the third place, the method by team *mathew666* obtained a final score of 23.931 ± 5.617 with the second place for the overall mIoU of 32.692. Similarly, for sequence data in Table 5, the method by team *polatgorkem* maintained the detection performance with a final score of 25.529 ± 10.326 and an overall mIoU of 29.117. While in the second place, team *VinBDI* suggested a method that obtained the second place in both the final score and overall mIoU with values 24.542 ± 13.972 and 23.426, respectively. The second scorer team *arnavchavan04* on single frame data was only 11th on the sequence data. In both tables, the top final scores outperformed the baseline methods in all evaluation metrics.

Furthermore, Table 6 shows the overall ranking for the teams in terms of Score (R_{score_d}), mAP (R_{mAP}), and generalisability performance (R_g) in addition to, mAP_d, mAP_{seq}, $score_d$, mAP_g and dev_g. As shown, the best mAP_d and mAP_g were achieved with the method proposed by team *StarStarG* with values 28.380 and 25.340. For the mAP_{seq}, the highest value reached 28.242 with the method proposed by the team *higersky*. Moreover, for the generalisation, the method by the baseline RetinaNet, team *DuyHUYNH* and YOLOv3 comes in the first, second and third places with values of 1.985, 4.807 and 4.397, respectively. On considering the mAP_g and dev_g together for the final ranking of the generalisation task, it can be observed that teams VinBDI and StarStarG were on the first place while team higersky secured second place.

We also present the detection score, AP, and IoU for each artefact class by participant teams for a single frame and sequence

Table 5. EAD2020 results for the sequence dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are averaged with 11 IoU thresholds. Weighted detection score $score_d$ is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	mAP_{25}	mAP_{50}	mAP_{75}	overall mAP_{seq}	overall $mIoU_{seq}$	mAP_{δ}	$score_d \pm \delta$
polatgorkem	38.464	24.803	4.138	23.137	**29.117**	10.326	**25.529 ± 10.326**
qzheng5	**48.21**	25.717	3.997	25.665	20.949	**14.222**	**23.779 ± 14.222**
xiahong1	46.087	25.813	2.684	25.136	18.398	**15.128**	22.441 ± 15.128
mathew666	31.599	21.878	3.053	19.623	20.858	9.718	20.117 ± 9.718
VinBDI	45.295	**26.723**	4.396	25.285	**23.426**	**13.972**	**24.542 ± 13.972**
higersky	**47.716**	**29.841**	**4.473**	**28.334**	12.865	14.579	22.147 ± 14.579
StarStarG	**46.965**	**30.202**	**5.432**	**28.107**	8.371	13.367	20.213 ± 13.367
anand_subu	38.352	25.535	3.843	23.014	20.703	10.859	22.089 ± 10.859
arnavchavan04	34.511	21.524	**4.886**	20.700	11.827	9.839	17.151 ± 9.839
MXY	31.391	19.838	3.620	18.601	**21.504**	8.688	19.762 ± 8.688
mimykgcp	44.972	26.78	4.400	**25.937**	6.892	13.697	18.319 ± 13.697
DuyHUYNH	28.632	15.524	0.815	15.468	16.968	9.381	16.068 ± 9.381
baselines							
YOLOv3	32.199	18.473	1.137	17.176	16.351.	10.596	16.846 ± 10.596
RetinaNet (ResNet101)	17.646	6.447	0.767	8.079	10.000	5.151	9.252 ± 5.151

of frames in Figure 4. As shown in Figure 4 (a) (i.e., single frame), there was a high detection score and AP for locating the saturation and contrast categories by the participants. Additionally, most of the teams had a high IoU with the ground truth when detecting the instrument class. On the other hand, the detection of the bubble class showed the worst performance for all measures by all the participating teams and even by the baseline methods.

5.1.3 Segmentation task for EAD2020

For the segmentation task, we compare the results of the participants w.r.t baseline segmentation methods: *FCN8, UNet-ResNet34, PSPNet, DeepLabv3(ResNet50), DeepLabv3+(ResNet50)*, and *DeepLabv3+(ResNet101)*. Table 7 presents the JC, DSC, F2, PPV, recall, and accuracy obtained by each team and baseline methods when segmenting the different artefacts. As shown, the method proposed by team *arnavchavan04* and team *VinBDI* had the best performance in terms of JC, DSC, F2 and PPV proving the ability to segment less false positive regions. While the team *arnavchavan04* ranked the first place with a score of 0.731 and accuracy of 0.977, the team *VinBDI* was placed on 2nd place with an overall accuracy of 0.978 and a semantic score of 0.730. Both teams had an equal performance for the DSC with a value of 0.673. However, the method suggested by team *qzheng5* and team *DuyHUYNH* segmented more true positive regions compared to other teams obtaining top recall values of 0.8352 and 0.828. The baseline methods showed a low performance in terms of final score compared to the methods proposed by the participants coming to the end of the ranking list. In conclusion, the method proposed by team *arnavchavan04* outperformed (i.e., R_{score_s}=1) in the overall performance for the different measures.

Furthermore, Figure 5 (a) presents the scores for DSC, PPV and Recall for each class by the teams and baseline methods. Segmenting the saturation maintained the high score among the teams and the baseline methods for the three measures. For the different teams, the highest recall was obtained by the specularity class while the lowest was shown when segmenting the artifact class. Moreover, similar to detection results, segmenting the bubble category showed a low performance for these three measures.

EDD2020 sub-challenge

In this section, we report the performance of the participating teams in the EDD2020 challenge for the detection and segmentation of the five different abnormalities from the endoscopic frames.

Table 6. EAD2020 team ranking based on different metric criteria for detection and generalisation task. Overall mAPs (mAP_d and mAP_{seq}) computed on single frame and sequence data are averaged. Final $score_d$ is then computed as the weighted value between the final IoU_d and the averaged mAP. Rankings for each metric are also provided based on ascending order of the scores except for deviation score for out-of-sample data. Three best scores for each metric criteria are in bold.

Team Names	**mAP_d**	**mAP_{seq}**	**final IoU**	**final $score_d$**	**mAP_g**	**dev_g**	**R_{score_d}**	**R_{mAP}**	**R_{gen}**
polatgorkem	17.486	23.137	**32.848**	**25.326**	21.008	9.359	**1**	9	6
qzheng5	19.720	24.174	23.751	**22.668**	23.749	8.522	**2**	6	5
xiahong1	18.512	25.136	22.393	**22.051**	24.579	8.169	**3**	7	**3**
mathew666	18.091	19.651	**26.783**	22.035	16.714	5.674	4	10	4
VinBDI	24.069	25.282	18.033	22.018	**24.140**	5.607	5	4	**1**
higersky	24.771	**28.252**	15.061	21.931	**24.850**	7.686	6	**2**	**2**
StarStarG	**28.380**	**28.107**	12.311	21.870	**25.340**	7.537	7	**1**	**1**
anand_subu	18.886	23.004	22.359	21.510	20.203	7.896	8	8	5
arnavchavan04	**26.021**	20.700	16.496	20.614	21.138	6.968	10	5	**3**
MXY	17.820	18.597	**24.779**	20.836	17.294	6.077	9	11	4
mimykgcp	**25.082**	**25.843**	8.536	18.691	23.929	7.999	11	**3**	4
DuyHUYNH	11.894	15.468	22.016	17.015	11.304	**4.807**	13	13	4
baselines									
YOLOv3	13.249	17.176	20.617	17.374	15.456	**4.397**	12	12	**3**
RetinaNet (ResNet101)	8.754	8.079	16.601	11.690	7.763	**1.985**	14	14	**3**

Table 7. Evaluation of the artefact segmentation task. Top three best scores for each metric criteria are in bold.

Team Names	**JC**	**DSC**	**F2**	**PPV**	**Rec**	**Acc**	**$Score_s$**	**R_{score_s}**
qzheng5	0.477	0.532	0.561	0.556	**0.835**	0.973	0.621	8
VinBDI	**0.628**	**0.673**	**0.670**	**0.837**	0.738	**0.978**	**0.730**	**2**
higersky	0.529	0.579	0.587	0.675	0.758	0.975	0.650	5
anandsubu	0.304	0.354	0.361	0.430	0.747	0.975	0.473	14
arnavchavan04	**0.622**	**0.673**	**0.683**	**0.800**	0.767	**0.977**	**0.731**	**1**
DuyHUYNH	0.502	0.557	0.583	0.593	**0.829**	0.974	0.640	6
mimykgcp	0.531	0.576	0.579	**0.723**	0.726	**0.977**	0.651	4
mouradai_ox	**0.581**	**0.632**	**0.647**	0.711	**0.800**	0.974	**0.697**	**3**
baselines								
FCN8	0.500	0.548	0.550	0.670	0.708	0.976	0.619	9
UNet-ResNet34	0.310	0.364	0.373	0.419	0.766	0.974	0.481	13
PSPNet	0.497	0.541	0.534	0.698	0.680	0.975	0.613	10
DeepLabv3 (ResNet50)	0.448	0.495	0.492	0.599	0.704	0.974	0.572	12
DeepLabv3+(ResNet50)	0.485	0.533	0.535	0.646	0.726	0.976	0.610	11
DeepLabv3+(ResNet101)	0.501	0.547	0.546	0.683	0.718	0.973	0.624	7

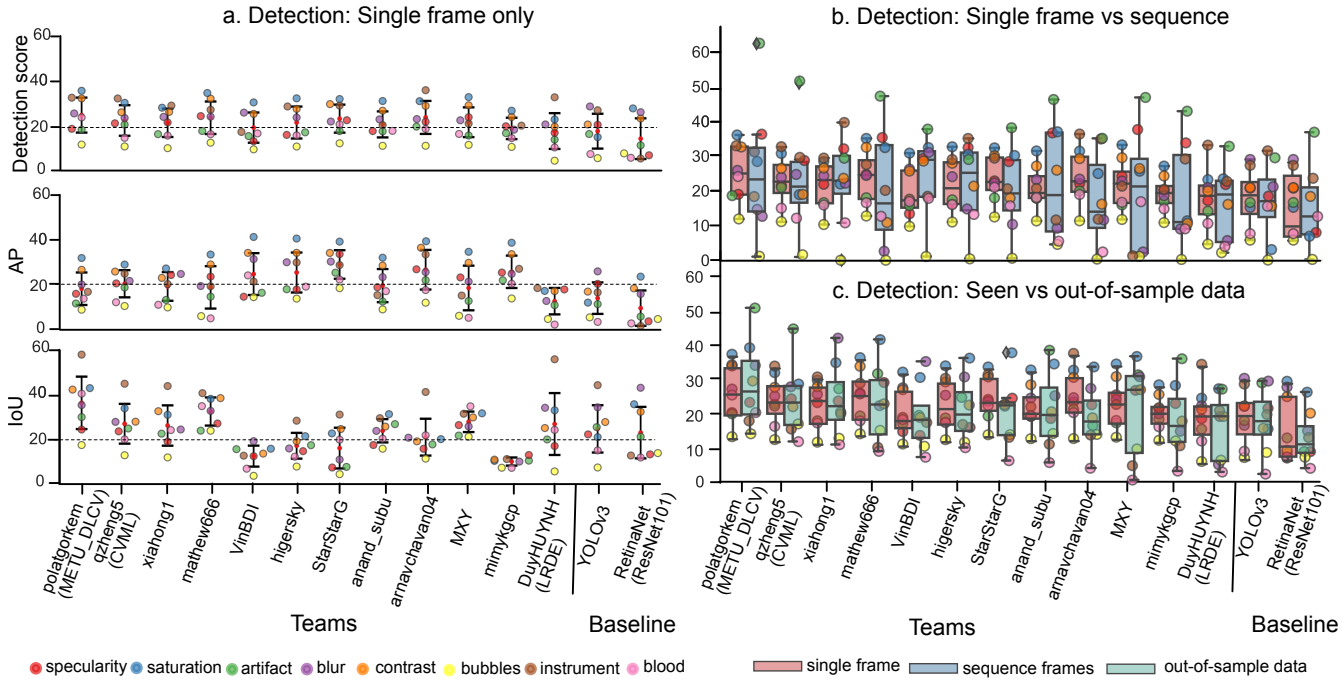

Figure 4. Detection and out-of-sample generalisation tasks for EAD2020 sub-challenge. a) Error bars and swarm plots for the intersection over union (IoU, top), average precision (AP, middle) and challenge detection score (mAP_d, bottom) for each team is presented on 237 single frame test data. b-c) Comparison of mAP_d w.r.t. mAP_{seq} (mAP on sequence test data with 80 frames) and mAP_g (mAP on out-of-sample data 99 frames) are provided. a-c) On the right, results from baseline detection methods: YOLOv3 and RetinaNet (with ResNet101 backbone) are also presented. Teams are arranged by decreasing overall detection ranking $\mathbf{R}_{score_d}$ (see Table 6).

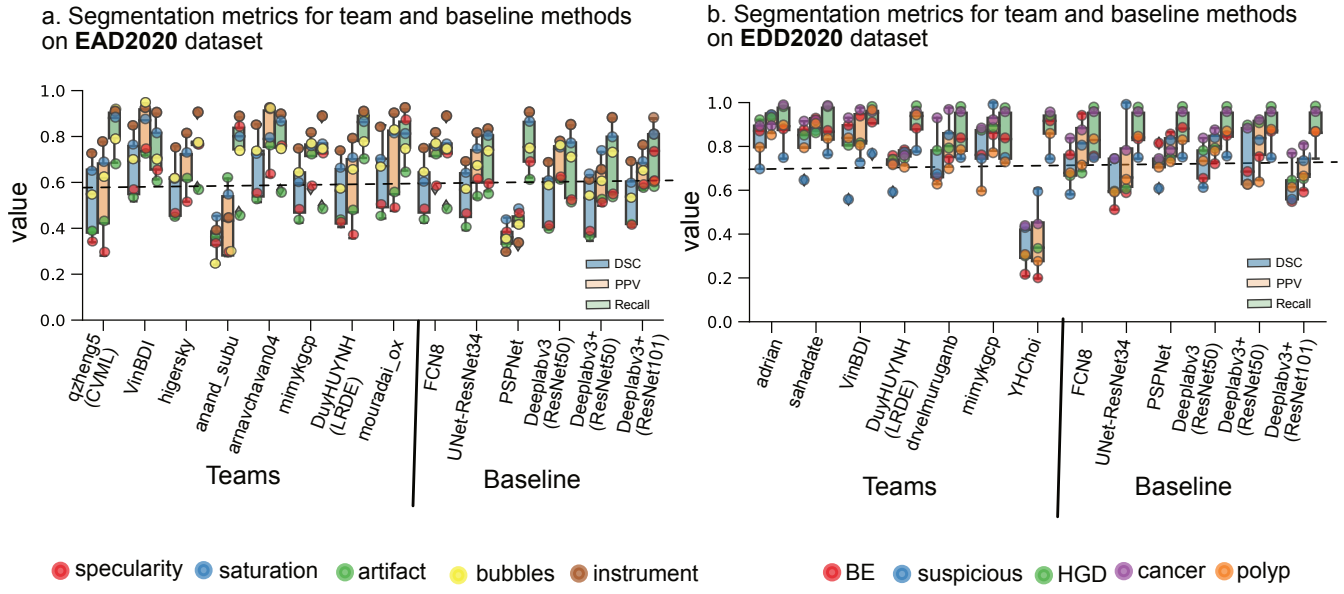

Figure 5. Semantic segmentation for EAD and EDD sub-challenges: Error bars with overlayed swarm plots for dice similarity coefficient (DSC), positive predictive value (PPV) or precision and recall are presented for each team and baseline methods for the EAD2020 (a) and EDD2020 (b) challenges. 6 different baseline methods are also provided for comparison.

Detection task for EDD2020

Table 8 shows the mAP values at different thresholds 25%, 50%, and 75%, overall mAP, overall IoU, mAP deviation, and the final score for the single frame detection task. We also compared results from participants with the state-of-the-art baseline

Table 8. EDD2020 results for the detection task on the single frame dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are computed at 11 IoU thresholds and averaged. Weighted detection score $score_d$ is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	mAP_{25}	mAP_{50}	mAP_{75}	overall mAP_d	overall $mIoU_d$	mAP_δ	$score_d \pm \delta$
adrian	**48.402**	**33.562**	**27.098**	**37.594**	**27.614**	8.523	**33.602 ± 8.523**
sahadate	**37.612**	**23.284**	**15.837**	**26.834**	**32.42**	8.325	**29.068 ± 8.325**
VinBDI	**43.202**	**26.981**	**17.001**	**30.219**	17.773	**9.478**	25.241 ± 9.478
YHChoi	23.183	11.082	8.8	15.783	24.623	6.216	19.319 ± 6.216
DuyHUYNH	23.959	9.587	5.659	12.479	13.829	6.284	13.019 ± 6.284
mimykgcp	34.884	20.982	4.463	20.742	2.27	**9.359**	13.353 ± 9.359
drvelmuruganb	31.018	18.421	11.768	21.79	7.322	7.424	16.002 ± 7.424
baselines							
YOLOv3	34.305	21.227	14.65	22.98	24.351	6.456	23.528 ± 6.456
RetinaNet (ResNet50)	26.833	14.441	9.907	17.552	25.58	6.464	20.763 ± 6.464
RetinaNet (ResNet101)	42.579	27.0	11.194	27.974	**26.434**	**11.949**	**27.358 ± 11.949**

methods (YOLOv3 and RetinaNet). Team *adrian* achieved the highest score among other participants and the baseline methods with a final $score_d$ of 33.602 ± 8.523. The team obtained also the best overall mAP of 37.594 and the second highest overall mIoU of 27.614. Team *sahadate* was able to maintain a high mAP performance at different IoU thresholds compared to the other teams and baseline methods achieving the top performance in terms of the overall mIoU of **32.42** and secured a second place for the final score (29.068 ± 8.325). Furthermore, the baseline method *RetinaNet* with the ResNet101 backbone performed better than most of the participating teams securing the third rank for the final score (27.358 ± 11.949) and overall mIoU (26.434).

Table 9 presents per class detection results for each team's method and the baseline methods. For the detection of NDBE and suspicious regions, the method presented by the team *mimykgcp* obtained the highest value of 50.089 and 4.592, respectively. On the other hand, team *adrian* was able to perform better in detecting HGD and cancer regions from the endoscopic frames with scores of 32.727 and 64.286, respectively on these classes. Finally, the best performance for detecting polyps was shown by the team *VinBDi* with a value of 63.26. The baseline methods failed in detecting suspicious regions and similarly by the methods proposed by the teams *YHChoi* and *drvelmuruganb*. In conclusion, the overall performance by the team *adrian* outperformed against the other teams with the ability to detect most abnormalities with the high detection scores. Similarly, RetinaNet with ResNet101 backbone (baseline method) performed better than most teams in detection of cancer and polyp classes.

5.1.4 Segmentation task for EDD2020

The results of the participant teams for disease segmentation from endoscopic frames are reported in this section. In table 10, we represent the JC, DSC, F2, PPV, recall and accuracy obtained by each team for the segmentation task. As illustrated, three teams *(Adrian, sahadate* and *nhanthanhnguyen94*) achieved a high overlapping ratio between the generated masks and the ground-truth annotations with a JC value of 0.8203, 0.798, and 0.7882, respectively. Moreover, they maintained the high performance for the DSC, F2, and PPV measures with comparable results as well.

Also, the method presented by the teams *VinBDI, DuyHUYNH* and *YHChoi* were able to segment more true positive regions reaching the top recall values of 0.912, 0.905 and 0.896, respectively. The overall accuracy result for all the participating teams and the baseline methods were very competitive.

Additionally, the scores for DSC, PPV, and recall for each class by the teams and baseline methods are shown in Fig. 5 b. Most participating teams and the baseline methods showed highest values for cancerous regions. Also, most teams showed higher DSC, PPV and recall for BE class instance as well (> 0.8 for top three teams). As presented, the majority of metrics had the least values for the suspicious class. Also, the segmentation of polyps did not show good performance in comparison to the detection results.

Table 9. Per class evaluation results for the detection task of the EDD2020 sub-challenge.

Teams EDD2020	NDBE	suspicious	HGD	cancer	polyp	δ
adrian	28.911	**1.776**	**32.727**	**64.286**	**60.269**	**22.841**
sahadate	**46.193**	1.099	22.727	10.0	54.152	**20.414**
VinBDI	**48.489**	**3.497**	**25.852**	10.0	**63.26**	**22.66**
YHChoi	26.9	0.0	22.727	0.0	29.289	13.057
DuyHUYNH	20.281	1.499	11.364	0.0	29.254	11.134
mimykgcp	**50.089**	**4.592**	**23.064**	5.852	20.112	16.429
drvelmuruganb	34.775	0.0	22.727	0.0	51.446	19.993
baselines						
YOLOv3 (darknet53)	38.839	0.0	6.97	**16.667**	52.426	19.712
RetinaNet (ResNet50)	23.636	0.0	18.182	0.0	45.943	17.086
RetinaNet (ResNet101)	29.483	0.0	22.727	**31.818**	**55.84**	17.909

Table 10. Evaluation of the disease segmentation methods proposed by the participating teams and the baseline methods. Top three evaluation criteria are highlighted in bold.

Team Names	JC	DSC	F2	PPV	Rec	Acc	Score$_s$	R$_{score_s}$
adrian	**0.820**	**0.836**	**0.842**	**0.921**	0.894	0.955	**0.873**	**1**
sahadate	**0.797**	**0.816**	**0.819**	**0.906**	0.883	0.955	**0.856**	**2**
VinBDI	**0.788**	**0.805**	**0.812**	**0.859**	**0.912**	0.952	**0.847**	**3**
DuyHUYNH	0.6843	0.7058	0.718	0.762	**0.905**	0.931	0.773	9
drvelmuruganb	0.7166	0.7349	0.734	0.819	0.857	0.959	0.786	6
mimykgcp	0.7561	0.7721	0.770	0.893	0.845	0.957	0.820	4
YHChoi	0.314	0.340	0.356	0.385	**0.896**	0.892	0.494	13
baselines								
FCN8	0.687	0.705	0.709	0.811	0.850	0.953	0.769	10
UNet-ResNet34	0.617	0.637	0.638	0.732	0.868	0.958	0.719	11
pspnet	0.698	0.721	0.723	0.797	0.876	0.959	0.779	8
DeepLabv3 (RetinaNet50)	0.704	0.724	0.724	0.810	0.878	**0.962**	0.784	7
DeepLabv3+ (RetinaNet50)	0.725	0.744	0.749	0.818	0.882	**0.960**	0.798	5
DeepLabv3+ (RetinaNet1010	0.608	0.627	0.629	0.698	0.880	**0.962**	0.709	12

Qualitative results

Detection task

Figure 6 shows the best (panel a) and the worse (panel b) performing frames from single frame dataset for EAD2020. It can be observed that specularity and artefacts are detected and well localised by top teams (see Figure 6 a). Similarly, in the bottom

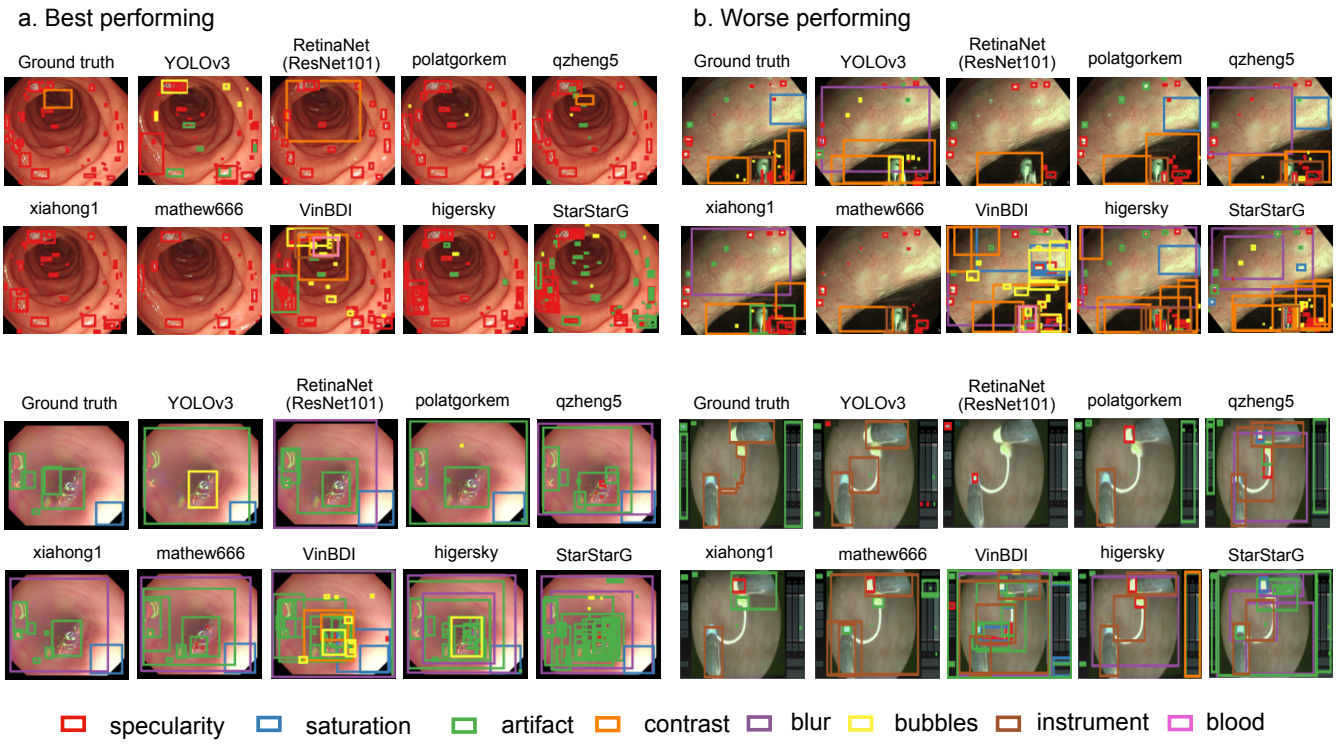

Figure 6. EAD2020 best and worse performing samples for the detection task. a) Best performing samples for 6 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

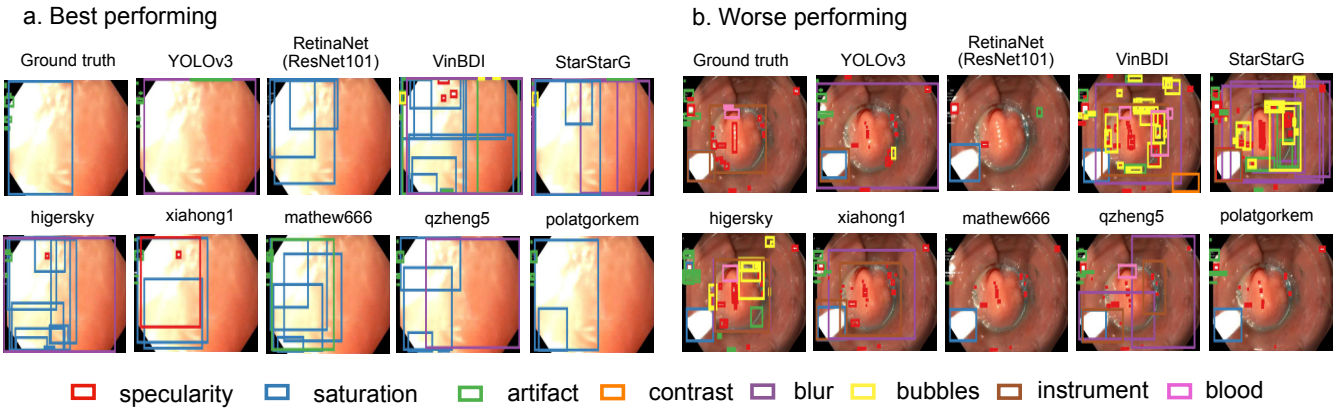

Figure 7. EAD2020 best and worse performing samples for the generalisation task. a) Best performing samples for 7 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

example, saturation is also detected by all the participants. Even though, blur is not present for this sample, most methods also detected it. While for the worse performing frame (see Figure 6 b), instrument class is confused with contrast or artefact on the top sample, while in the bottom sample instrument is detected by some teams but often either detected only partially or overlapped by different classes such as saturation or artefact.

For out-of-sample generalisation task, it can be seen in Figure 7 (a) that besides YOLOv3 baseline method, all the baselines and teams detected saturation part. While some teams (*mathew666*, *VinBDI*, *higersky*) detected multiple bounding boxes for the same class, they also detected blur class for this frame. While for worse performing frame (see Figure 7 (b)), instrument class (at the center of the image) is well localised only by the team *xiahong1* while most teams either partially detected the instrument (e.g., team *qzheng5*) or do not detect the instrument class at all (e.g., team *polatgorkem*). In both cases, the three

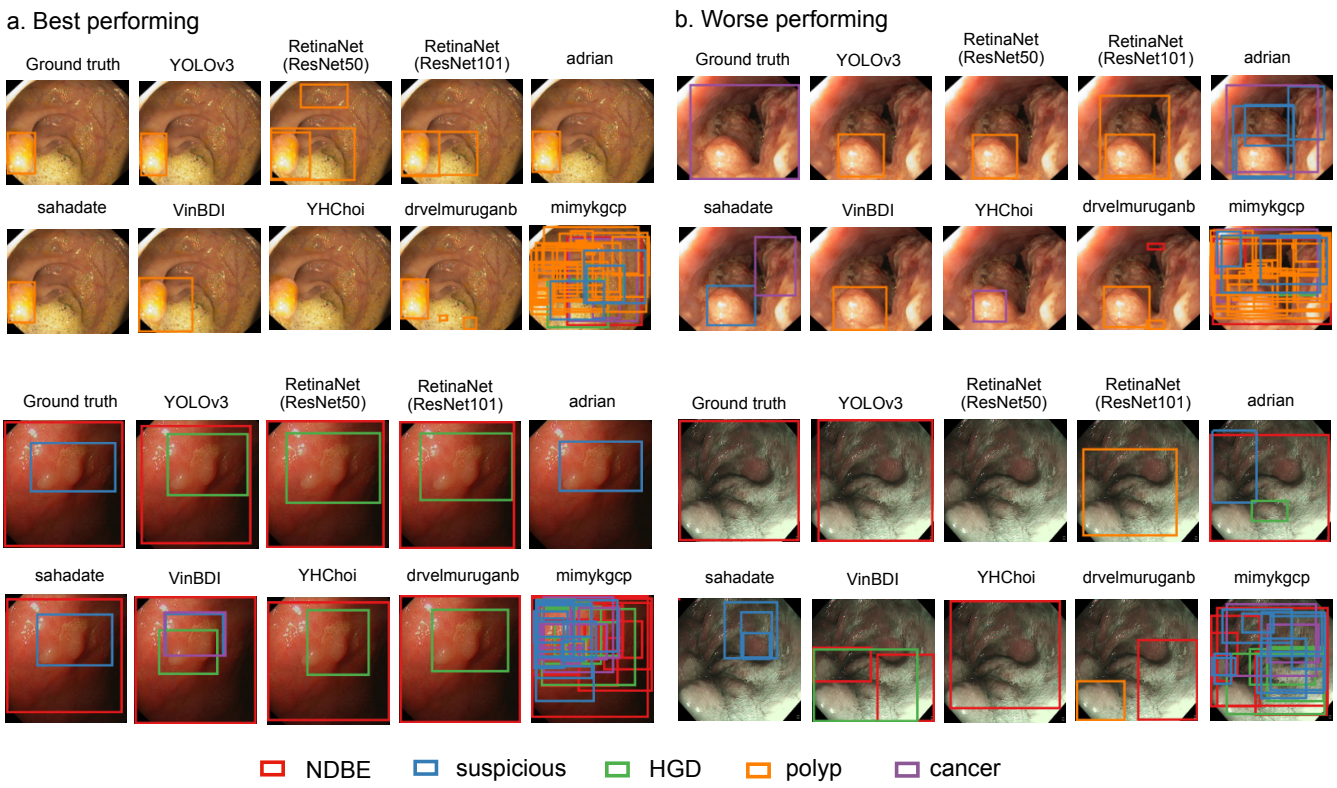

Figure 8. EDD2020 best and worse performing samples for the detection task. a) Best performing samples for 6 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

teams *VinBDI, higersky* and *StarStarG* produced multiple overlapping and different size bounding boxes. Qualitative results for the EDD2020 challenge is shown in Figure 8. The best performing samples in Figure 8 (a) shows polyp class (on top); non-dysplastic Barrett's Oesophagus (NDBE) and suspicious classes on the bottom. It can be observed that polyp class is detected and well localised by all the teams and baseline methods. However, for bottom row NDBE is detected by most of the methods while confusion is observed across the suspicious class with high-grade dysplasia (HGD) class. Team *mimykgcp* produced numerous bounding boxes failing to optimally localise adherent disease classes. For the worse performing frames (Figure 8 (b)), cancer class (top) in the ground truth is confused with the polyp class instance for most of the teams and the baseline methods. While, for the NDBE class in the bottom of Figure 8 (b), teams were either not able to detect the NDBE class (except team *adrian*, team *YHChoi* and YOLOv3) at all or partially detected the NDBE areas (e.g., teams *VinBDI* and *drvvelmuruganb*). Again, for the presented case, team *mimykgep* detected numerous bounding boxes.

Segmentation task

Endoscopic artefact segmentation samples representing best and worse performing teams is provided in Figure 9. For the sample with only the instrument class (see Figure 9 a, top panel) it can be observed that almost all the baseline and teams were able to predict precise delineation of the instrument class. Similarly, in the bottom panel of Figure 9 (a), specularity, saturation and artefact classes were segmented well by most of the teams and baseline methods. Even though, a single instrument class is present in the sample image in Figure 9 (b), none of the methods were able to segment the instrument. Also, for the bottom panel in the Figure 9 (b), specularity areas were segmented well by the teams *mouradaiox* and *mimykgcp*. However, saturation area was under segmented by most of the teams and baseline methods. Figure 10 (a) represents the polyp class (on top); NDBE and suspicious classes (on bottom). It can be observed that polyp is segmented well by all the baselines and most teams (except team *drvelmuruganb* who miss-classified the pixels to suspicious class). While, most teams and baselines were able to precisely delineate NDBE class for the frame in the bottom panel but missed suspicious area. In the worse performing sample (see Figure 10 (b)), most teams were able to segment NDBE area but large HGD area was missed by all the teams. Also, some teams confused HGD area with suspicious class. For the bottom panel in Figure 10 (b), instead of suspicious class present in the ground truth, almost all the teams detected this as polyp or cancer. However, the region delineation was close to the ground truth for most teams.

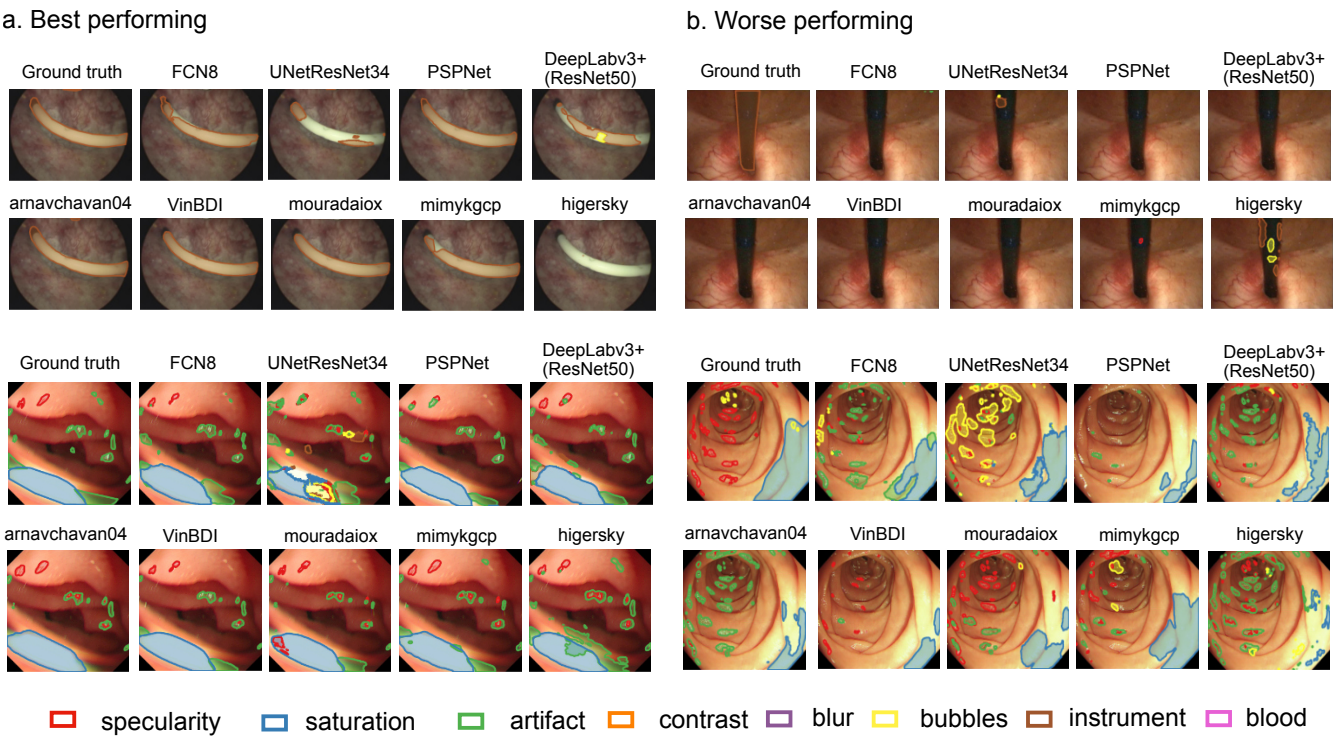

Figure 9. EAD2020 best and worse performing samples. a) Best performing samples for 5 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample (top). Single class samples are chosen at the top and multi-class samples are at the bottom in each category.

Discussion

Deep learning methods are rapidly being translated for the use of computer aided detection (CADe) and diagnosis (CADx) of diseases in complex clinical settings including endoscopy. However, the amount of data variability particularly in endoscopy is significantly higher than in natural scenes which possess a significant challenge in the process. It is therefore vital to determine an effective translational pathway in endoscopy. Majority of challenges in endoscopy are due to its complex surveillance that lead to severe artefacts that may confuse with disease. Similarly, a system designed for a particular organ may not generalise to be used in the other.

Most deep learning methods that were used in the EndoCV2020 challenge can be categorised into multiscale, symbiotic, ensemble, encoder-decoder and cascading nature, or a combination of these (see Table 2 and Table 3). Figure 11 presents the overview of the used methods for the detection (a) and segmentation (b) challenge tasks which is based on the architecture usage. It can be observed that the majority of detection methods used two-stage Faster-RCNN with 4/7 teams combining it with one-stage RetinaNet or YOLOv3 or a combination of all. Cascade R-CNN which is built upon Faster R-CNN cascaded architecture was exploited by 4 teams. Similarly, U-Net-based architectures were utilised by most teams for semantic segmentation task with 4 teams exploring pyramid module-based architectures and 2 teams used Deeplabv3+ architecture. Faster RCNN-based model was also explored with additional thresholding (e.g., team *adrian*) or per pixel prediction heads (e.g., team *sahadate*).

For the detection task, the top performing teams on the challenge metric in both EAD (team *polatgorkem*) and EDD (team *adrian*) were those using ensemble networks, i.e., maneuvering outputs from multiple architectures. However, these networks sacrifice the speed of detection which can be observed from the computational time which were significantly higher than teams that used a single architecture (see Table 6 and Table 8). Other teams that used such an approach included team *arnavchavan04* and *mimykgcp* who combined Faster R-CNN with RetinaNet but both teams were respectively on 10th and 11th ranking. Just using Faster R-CNN alone with ResNet101 backbone, teams *qzhang5* and *mathew666* were able to detect both small and large size bounding boxes with sub-optimal accuracy that put them at 2nd and 4th positions, respectively. Similarly, team *sahadate* claimed 2nd position on EDD detection task using Mask R-CNN which is based on the Faster R-CNN architecture.

An intelligent choice for improved speed and accuracy using a scalable network was presented by the teams *xiahong1* (used YOLACT) and *VinBDI* (used EfficientDet D0). Both the teams were placed 3rd and 5th, respectively, on the final detection score of the EAD2020. Team *VinBDI* was also ranked 3rd on the EDD detection task. Teams *higerssky*, *StarStarG* and *MXY* that used cascaded R-CNN were ranked respectively on 6th, 7th and 9th positions. Additionally, the team StarStarG was ranked

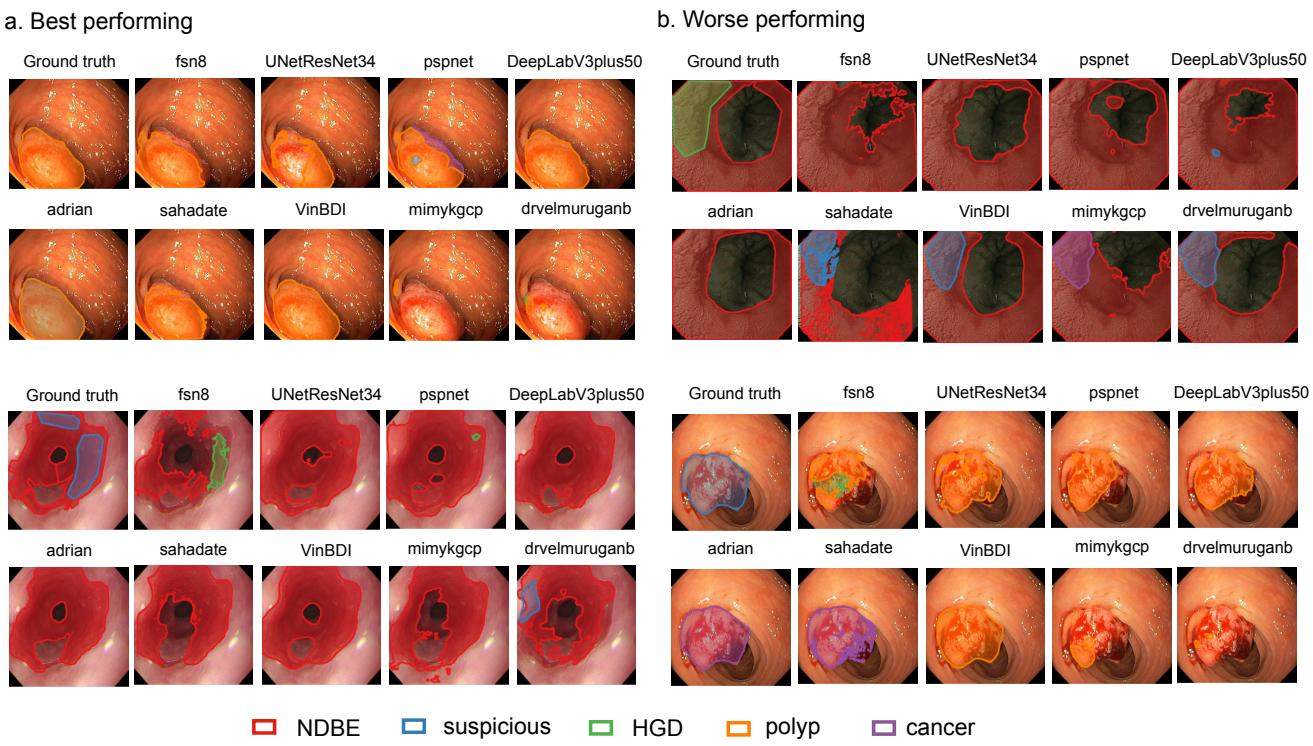

Figure 10. EDD2020 best and worse performing samples. a) Best performing samples for 5 top team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample (top).

1st and team higersky was ranked 2nd on the overall mAP. However, it is to be noted that taking only mAP scores into account for detection could lead to over detection of the bounding boxes that increases the chance of finding a particular class but at the same time weakens the localisation capability of the algorithm (see Figure 6). Similar observations were found for the EDD dataset where the team *mimykgcp* obtained an overall mAP of 20.742 but only 2.270 for the overall IoU (see Table 8). As a result, over detection of the bounding boxes can be seen in Figure 8. In order to deal with the over detection of the bounding boxes, YOLACT architecture used by *xiahong1* suppressed the duplicate detections using already-removed detections in parallel (*fast NMS*). Similarly, teams such as *polatgorkem* from the EAD and *adrian* from the EDD were able to eliminate the duplicate detections using ensemble network and a class agnostic NMS.

Hypothesis I: In the presence of multiple class objects, object detection methods may fail to precisely regress the bounding boxes. Methods need better penalisation on the bounding box regression or a technique to perform effective non-maximal suppression.

The choice of networks from each team depended on their ambition of either obtaining very high accuracy without focusing on speed or a trade-off between the speed and the accuracy or focusing on both and thinking out-of the box to use more recent developed methods which beats faster networks (such as YOLOv3) that included EfficientDet D0 architecture used by the team VinBDI (see Table 2). Due to the efficiency of the EfficientDet D0 network that used biFPN and efficientNet backbone, team *VinBDI* achieved second least deviation in mAP (i.e., $\text{dev}_g = 5.607$) with competitive mAP_g ($= 24.140$) and won the generalisation task together with the team *StarStarG* who had slightly higher mAP_g ($= 25.340$) but larger mAP deviation between detection and generalisation datasets. Most methods for the detection task on both the EAD and EDD dataset performed better than the baseline one-stage methods (YOLOv3 and RetinaNet). However, it was found that even though team *polatgorkem* won the detection task, the method failed on generalisation data where the team was ranked only last. The main reason behind this could be because the generalisation gap mAP_g was estimated between two mAP's (mAP_d and mAP_g) and not IoU. Also, the final ranking was done taking into account the rank of dev_g and mAP_g only. It can be observed in Figure 7 that the bounding box localisation of team *polatgorkem* is precise in (a) while it misses instrument area at the center in (b). However, the winning teams *VinBDI* and *StarStarG* both over detect the boxes. The generalisation ability of the methods were not explored for EDD dataset.

Hypothesis II: Metrics are critical but using a single metric does not always gives the right answer. Weighted metrics are desired in object detection task to establish a good trade-off between detection and precise localisation.

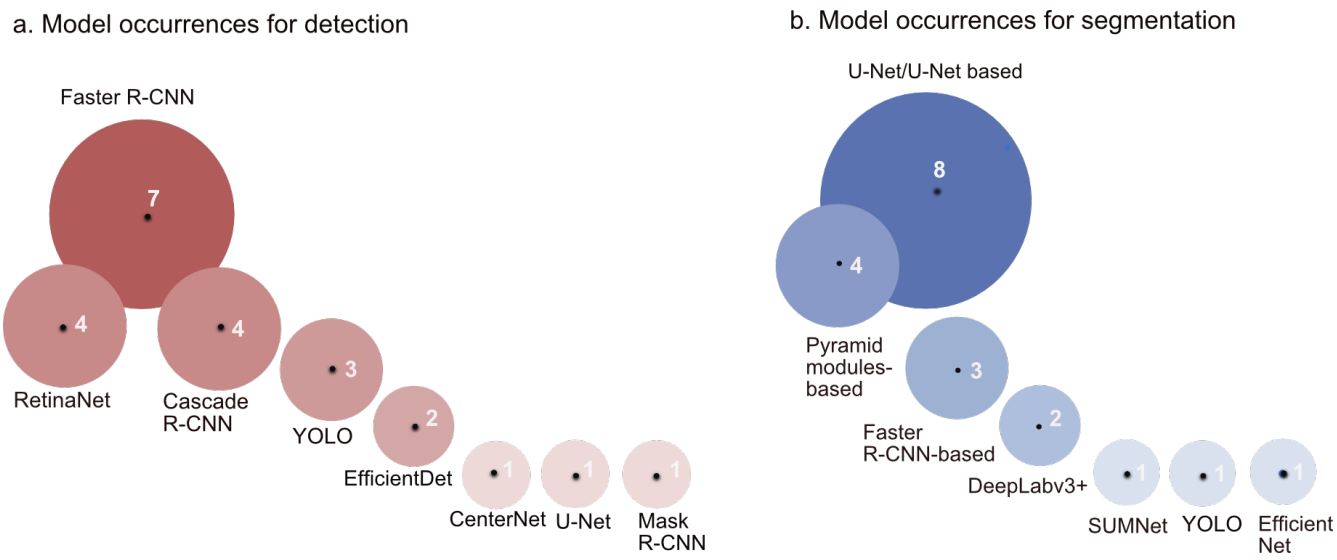

Figure 11. EndoCV2020 method categories in blob-representattion. Model occurences are presented for detection (a) and segmentation (b) tasks for both EAD202 and EDD2020 sub-challenges. The number of occurrences is numbered inside each blob.

A major problem in the detection of EDD dataset was class confusion mostly for suspicious, HGD and cancer classes. This could be because of smaller number of samples for each of these classes compared to NDBE and polyp (see Figure 2). While most methods were able to detect and localise NDBE and polyp class in general (3/7 teams with an overall mAP > 45 and 4/7 teams with > 50), all teams failed in suspicious class (overall mAP < 5.0) and most teams for cancer class (overall mAP < 15.0) (see Table 9). Figure 8 shows that polyp is detected and localised very well by most of teams (a, top). Similarly, NDBE is localised by most methods, however, in this case suspicious class is confused mostly with the HGD. Also, in Figure 8 (b, top), it can be observed that the cancer class instance is confused with mostly polyp class.

Hypothesis III: Detection bounding boxes confuse with classes that have similar morphology and smaller number of samples failing to learn the contextual features. To improve detection, such samples need to be identified and more data demonstrating such attributes need to be injected (both positive and negative samples).

Similar to the detection task, teams that used ensemble techniques were among the best performing teams for the segmentation task. Teams *arnavchavan04* and *VinBDI* secured first ($\textbf{score}_s = 0.731$) and second ($\textbf{score}_s = 0.730$) positions, respectively, on the EAD2020 segmentation task (see Table 7) and the team *adrian* won the EDD2020 segmentation task challenge with $\textbf{score}_s$ of 0.873 (see Table 10). The team *arnavchavan04* used multiple augmentation techniques including cutmix and a feature pyramid network with a combination of EfficientNet backbones from B3 to B5. Similarly, team *VinBDI* ensembled a U-Net architecture with EfficientNet B4 and BiFPN network with ResNet50 backbone. In the EDD2020 segmentation task, the team *adrian* combined predictions from three object detection architectures where the YOLOv3 and Faster R-CNN class predictions were used to correct the instance segmentation masks from Cascade R-CNN. A direct instance segmentation approach used by the team *sahadate* secured second position ($\textbf{score}_s = 0.856$) on the same while ensemble network of the team *VinBDI* secured the third position ($\textbf{score}_s = 0.847$). Direct usage of a single existing state-of-the-art methods utilising different augmentation techniques (e.g., *DuyHUYNH*) or different backbones (e.g., *mimykgcp*, *qzheng5*) resulted in improved results compared to the original baseline methods, however, much lower than the top performing methods (see Table 7 and Table 10).

Hypothesis IV: The choice of combinatorial networks that well synthesises width, depth and resolution to capture optimal receptive field, and a domain agnostic knowledge transfer mechanism are critical to tackle heterogeneous (multi-center and variable size) multi-class object segmentation task.

From Figure 5 it can be observed that the top three performing teams of the EAD2020 (*arnavchavan04*, *VinBDI*, *mouradai_ox*) has high DSC vaue (0.538, 0.548 and 0.492 respectively) compared to most methods for the specularity class instance. It is to be noted that the specularities are often confused with either artifact or bubbles which makes them hard to differentiate. For the instrument, saturation and bubbles class instances (see Figure 9 a.), most methods obtained high performance compared to other classes (e.g., the top three teams obtained 0.853, 0.844, 0.848 for the instrument; 0.722, 0.758, 0.703 for the saturation; and 0.738, 0.693, 0.693 for the bubbles class instance, respectively), artefact (DSC < 0.52) was among the worst class for most teams and for the baseline methods. This is mostly due to the variable size of artefacts; and the bubbles

class instance is predominantly confused with either artefact or the specularity class (see Figure 9 b.). Unlike the EAD2020, the EDD2020 segmentation task comprised of larger shaped regions and only a few classes confused (see 1 b.). Most methods scored comparably high DSC values with over 75% for most of the disease classes except for suspicious class by most of the team. However, Figure 10 (b) (top) shows that while majority of teams were able to segment NDBE class area, the teams either missed the HGD area or miss classified HGD as suspicious class instance. It is to be noted that there is a very subtle difference between the HGD and the suspicious region even for the expert endoscopists. Similar observation can be found for the segmentation of protruded structures (Figure 10 (b), bottom) where most methods confused the class with the polyp class and the top two teams (*adrian*, *sahadate*) classified it as cancer class. Looking up into our expert consensus notes we found that these samples had hard to reach agreement cases (i.e., suspicious and HGD classes; and cancer and polyp region).

Hypothesis V: Instead of hard scoring of predicted mask classes that penalises the method performance heavily in presence of marginal visual difference between classes and variability due to existing expert consensus in the dataset, probability maps can be used to mitigate such problem.

6 Conclusion

We provide a comprehensive analysis of the deep learning methods built to tackle two distinct challenges in the gastrointestinal endoscopy: a) artefact detection and segmentation and b) disease detection and segmentation. This has been possible by the crowd-sourcing initiative of the EndoCV2020 challenges. We have provided the summary of the methods developed by the top 17 participating teams and compared their methods with the state-of-the-art detection and segmentation methods. Additionally, the paper dissects different paradigms used by the teams and present a detailed analysis and discussion of the outcomes. We also suggest pathways to improve the methods for building reliable and clinically transferable methods.

Acknowledgments

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. S. Ali, B. Braden, A. Bailey and JE. East is supported by NIHR BRC and J. Rittscher by Ludwig Institute for Cancer Research and EPSRC Seebibyte Programme Grant (EP/M013774/1). We want to acknowledge Karl Storz for co-sponsoring the challenge workshop. We would also like to acknowledge the annotators and our EndoCV2020 challenge workshop proceedings reviewers and IEEE International Symposium on Biomedical Imaging 2020 challenge organizers Michal Kozubek and Hans Johnson.

Author contributions

S. Ali conceptualized the work, led the challenge and workshop, prepared the dataset, software and performed all analyses. M. Dmitrieva, N. Ghatwary, and S. Bano served as organising committee and participated in annotations. A. Bailey, B. Braden, J.E. East, R. Cannizzaro, D. Lamarque, S. Realdon were involved in the validation and quality checks of the annotations used in this challenge. A. Krenzer, A. Hekalo, YB. Guo, B. Matuszewski, M. Gridach, V. Yoganand assisted in compiling the related work. S. Ali wrote most of the manuscript with inputs from M. Dmitrieva, N. Ghatwary, S. Bano and all co-authors. All authors participated in the revision of this manuscript and provided substantial input.

Additional information

Competing interests The author(s) declare no competing interests.

References

1. Ali, S. *et al.* An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci. reports* **10**, 1–15 (2020).
2. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad gastroenterologiczny* **14**, 89–103 (2019).
3. Lee, J. *et al.* Risk factors of missed colorectal lesions after colonoscopy. *Medicine* **96**, e7468–e7468 (2017).
4. Zhang, X. *et al.* Real-time gastric polyp detection using convolutional neural networks. *PLOS ONE* **14**, 1–16, DOI: 10.1371/journal.pone.0214133 (2019).
5. Bernal, J. *et al.* Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Trans. Med. Imag* **36**, 1231–1249 (2017).

6. Jorge, B. & Aymeric, H. GastrointestinalG imageI analysisANA challenge. https://endovissub2017-giana.grand-challenge.org/ (2017).
7. Soberanis-Mukul, R. D. *et al.* A learning without forgetting approach to incorporate artifact knowledge in polyp localization tasks. *CoRR* **abs/2002.02883** (2020).
8. Ali, S. *et al.* A deep learning framework for quality assessment and restoration in video endoscopy. *CoRR* **abs/1904.07073** (2019).
9. Ali, S. *et al.* Endoscopy artifact detection (EAD 2019) challenge dataset. *arXiv preprint arXiv:1905.03209* (2019).
10. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).
11. Liu, W. *et al.* Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37 (Springer, 2016).
12. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).
13. Wang, P. *et al.* Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. biomedical engineering* **2**, 741–748 (2018).
14. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis machine intelligence* **39**, 2481–2495 (2017).
15. Urban, G. *et al.* Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology* **155**, 1069–1078 (2018).
16. Horie, Y. *et al.* Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest. endoscopy* **89**, 25–32 (2019).
17. Kayser, M., Soberanis-Mukul, R. D., Albarqouni, S. & Navab, N. Focal loss for artefact detection in medical endoscopy. In *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy* (2019).
18. Oksuz, I., Clough, J. R., King, A. P. & Schnabel, J. A. Artefact detection in video endoscopy using retinanet and focal loss function. In *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019* (2019).
19. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587 (2014).
20. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99 (2015).
21. Cai, Z. & Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162 (2018).
22. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (2017).
23. 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 (2012).
24. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).
25. Yamada, M. *et al.* Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. reports* **9**, 1–9 (2019).
26. Shin, Y., Qadir, H. A., Aabakken, L., Bergsland, J. & Balasingham, I. Automatic colon polyp detection using region based deep cnn and post learning approaches. *IEEE Access* **6**, 40950–40962 (2018).
27. Suhui Yang, G. C. Endoscopic artefact detection and segmentation with deep convolutional neural network. In Ali, S. & Zhou, F. (eds.) *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019* (2019).
28. Pengyi Zhang, Y. Z., Xiaoqiong Li. Ensemble mask-aided r-cnn. In Ali, S. & Zhou, F. (eds.) *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019* (2019).
29. Law, H. & Deng, J. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750 (2018).

30. Duan, K. *et al.* Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 6569–6578 (2019).

31. Zhou, X., Zhuo, J. & Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 850–859 (2019).

32. Wang, D. *et al.* Afp-net: Realtime anchor-free polyp detection in colonoscopy. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 636–643 (IEEE, 2019).

33. Liu, Y., Zhao, Z., Chang, F. & Hu, S. An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. *IEEE Access* **8**, 78193–78201 (2020).

34. 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).

35. Sun, C. *et al.* Automatic segmentation of liver tumors from multiphase contrast-enhanced ct images based on fcns. *Artif. intelligence medicine* **83**, 58–66 (2017).

36. Kaul, C., Manandhar, S. & Pears, N. Focusnet: an attention-based fully convolutional network for medical image segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 455–458 (IEEE, 2019).

37. Gao, J. *et al.* A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* 4037190 (2017).

38. Brandao, P. *et al.* Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 101 – 107, DOI: 10.1117/12.2254361. International Society for Optics and Photonics (SPIE, 2017).

39. Brandao, P. *et al.* Towards a computed-aided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks. *J. Med. Robotics Res.* **03**, 1840002, DOI: 10.1142/S2424905X18400020 (2018).

40. 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).

41. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70, DOI: 10.1038/s41592-018-0261-2 (2019).

42. Chlebus, G., Meine, H., Moltz, J. H. & Schenk, A. Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering. *arXiv preprint arXiv:1706.00842* (2017).

43. Sevastopolsky, A. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognit. Image Analysis* **27**, 618–624 (2017).

44. Norman, B., Pedoia, V. & Majumdar, S. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. *Radiology* **288**, 177–185 (2018).

45. Jha, D. *et al.* Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *OSF Preprints* (2020).

46. Khan, M. A. & Choo, J. Multi-class artefact detection in video endoscopy via convolution neural networks. In *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy* (2019).

47. Bano, S. *et al.* Deep placental vessel segmentation for fetoscopic mosaicking. *arXiv preprint arXiv:2007.04349* (2020).

48. Gupta, S., Ali, S., Goldsmith, L., Turney, B. & Rittscher, J. Mi-unet: Improved segmentation in ureteroscopy. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 212–216 (2020).

49. 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).

50. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).

51. Seferbekov, S. S., Iglovikov, V. I., Buslaev, A. V. & Shvets, A. A. Feature pyramid network for multi-class land segmentation. *CoRR* **abs/1806.03510** (2018). 1806.03510.

52. Guo, X. *et al.* Automated polyp segmentation for colonoscopy images: A method based on convolutional neural networks and ensemble learning. *Med. Phys.* **46**, 5666–5676, DOI: 10.1002/mp.13865 (2019).

53. Jia, X. *et al.* Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. *IEEE Transactions on Autom. Sci. Eng.* **17**, 1570–1584 (2020).

54. Zhang, Y.-y. & Xie, D. Detection and segmentation of multi-class artifacts in endoscopy. *J. Zhejiang Univ. B* **20**, 1014–1020 (2019).

55. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

56. 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).

57. Guo, Y., Bernal, J. & J Matuszewski, B. Polyp segmentation with fully convolutional deep neural networks—extended evaluation study. *J. Imaging* **6**, 69 (2020).

58. 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).

59. Ali, S. *et al.* Artificial intelligence-driven real-time 3D surface quantification of barrett's oesophagus for risk stratification and therapeutic response monitoring. *medRxiv* DOI: 10.1101/2020.10.04.20206482 (2020).

60. Yang, S. & Cheng, G. Endoscopic artefact detection and segmentation with deep convolutional neural network. In *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy* (2019).

61. Ali, S. *et al.* Endoscopy disease detection challenge 2020. *CoRR* **abs/2003.03376** (2020).

62. Polat, G., Sen, D., Inci, A. & Temizel, A. Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination. In *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*, 8–12 (CEUR-WS.org, 2020).

63. Guo, Y. B., Zheng, Q. & Matuszewski, B. J. Deep encoder-decoder networks for artefacts segmentation in endoscopy images. In *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*, 18–21 (CEUR-WS.org, 2020).

64. Gridach, M. & Voiculescu, I. OXENDONET: A dilated convolutional neural networks for endoscopic artefact segmentation. In *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*, 26–29 (CEUR-WS.org, 2020).

65. Y, V., Prakash, P. & Shivashankar, N. A submission note on EAD 2020: Deep learning based approach for detecting artefacts in endoscopy. In *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*, 30–36 (CEUR-WS.org, 2020).

66. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719* **2**, 7 (2019).

67. Huynh, L. D. & Boutry, N. A u-net++ with pre-trained efficientnet backbone for segmentation of diseases and artifacts in endoscopy images and videos. In *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*, 13–17 (CEUR-WS.org, 2020).

68. Hu, H. & Guo, Y. Endoscopic artefact detection in mmdetection. In *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*, 78–79.

69. Jadhav, S., Bamba, U., Chavan, A., Tiwari, R. & Raj, A. Multi-plateau ensemble for endoscopic artefact segmentation and detection. In *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*, 22–25 (CEUR-WS.org, 2020).

70. Yun, S. *et al.* Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, 6023–6032 (2019).

71. Subramanian, A. & Srivatsan, K. Exploring deep learning based approaches for endoscopic artefact detection and segmentation. In *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*, 51–56 (CEUR-WS.org, 2020).

72. Chen, H., Lian, C. & Wang, L. Endoscopy artefact detection and segmentation using deep convolutional neural network. In *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*, 37–41 (CEUR-WS.org, 2020).

73. Bodla, N., Singh, B., Chellappa, R. & Davis, L. S. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569 (2017).

74. Yu, Z. & Guo, Y. Endoscopic artefact detection using cascade R-CNN based model. In *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*, 42–46 (CEUR-WS.org, 2020).

75. Gao, X. W. & Braden, B. Artefact detection and segmentation based on a deep learning system. In *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*, 80–81 (2020).

76. Bolya, D., Zhou, C., Xiao, F. & Lee, Y. J. YOLACT: Real-time instance segmentation. In *ICCV* (2019).

77. Krenzer, A., Hekalo, A. & Puppe, F. Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks. In *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*, 58–63 (CEUR-WS.org, 2020).

78. Rezvy, S. *et al.* Transfer learning for endoscopy disease detection & segmentation with mask-rcnn benchmark architecture. In *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*, 68–72 (CEUR-WS.org, 2020).

79. Nguyen, N. T., Tran, D. Q. & Nguyen, D. B. Detection and segmentation of endoscopic artefacts and diseases using deep architectures. In *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*, 64–67 (CEUR-WS.org, 2020).

80. Choi, Y. H. *et al.* Centernet-based detection model and u-net-based multi-class segmentation model for gastrointestinal diseases. In *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*, 73–75 (CEUR-WS.org, 2020).

81. Balasubramanian, V., Kumar, R., Kamireddi, S. J., Sathish, R. & Sheet, D. Semantic segmentation, detection AND localisation of mucosal lesions from gastrointestinal endoscopic images using SUMNET. In *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*, 82–83 (2020).

Supplementary Material

Supplementary figure: Endoscopy computer vision EndoCV2020 challenge dataset details. (a) Multi-center, multi-modality and multi-organ dataset for EAD and EDD sub-challenges. For EAD2020, 2532 frames with 8 class bounding boxes for the detection task out-of which 573 included ground truth masks for segmentation task were provided. Participants were assessed on 317 frames for detection and 162 frames for segmentation tasks. An additional 99 frames were used to test out-of-sample generalisation task for EAD sub-challenge. While EDD2020 consisted of 384 train samples and 43 test samples for 5 disease classes. (b-c) The distribution of 8 artefact classes of EAD and 5 disease classes of EDD w.r.t. their size compared to their height and width of image is provided. Each class size variability is also shown on right as blobs with mean at center and radius as standard deviation.