# Endoscopic computer vision challenges 2.0

Sharib Ali[1,2], Noha Ghatwary[3]

[1] School of Computing, University of Leeds, Leeds, UK
[2] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, Oxford, UK
[3] Computer Engineering Department, Arab Academy for Science and Technology, 1029, Alexandria, Egypt

### Abstract

Accurate detection of artefacts is a core challenge in a wide-range of endoscopic applications addressing multiple different disease areas. The importance of precise detection of these artefacts is essential for high-quality endoscopic video acquisition crucial for realising reliable computer assisted endoscopy tools for improved patient care. In particular, colonoscopy requires colon preparation and cleaning to obtain improved adenoma detection rate. Computer aided systems can help to guide both expert and trainee endoscopists to obtain consistent high quality surveillance and detect, localize and segment widely known cancer precursor lesion, "polyps". While deep learning has been successfully applied in the medical imaging, generalization is still an open problem. Generalizability issue of deep learning models need to be clearly defined and tackled to build more reliable technology for clinical translation. Inspired by the enthusiasm of participants on our previous challenges, this year we put forward a 2.0 version of two sub-challenges (Endoscopy artefact detection) EAD 2.0 and (Polyp generalization) PolypGen 2.0. Both the sub-challenges consists of multi-center and diverse population datasets with tasks for both detection and segmentation but focus on assessing generalizability of algorithms. In this challenge, we aim to add more sequence/video data and multimodality data from different centers. The participants is aimed to be evaluated on both standard (some already present at leaderboard) and generalization metrics presented in our previous challenges. However, unlike previous challenges, in 2.0 we aimed to benchmark methods on larger test-set comprising of mostly video sequences as in the real-world clinical scenario.

### Keywords

Artefact, Polyp, Endoscopy, Deep learning, Generalization

## 1. Introduction

Endoscopy is a widely used clinical procedure for the early detection of numerous cancers (e.g., nasopharyngeal, oesophageal adenocarcinoma, gastric, colorectal cancers, bladder cancer etc.), therapeutic procedures and minimally invasive surgery (e.g., laparoscopy). A major drawback during endoscopic video surveillance is that they are heavily corrupted with multiple artefacts (e.g., pixel saturations, motion blur, defocus, specular reflections, bubbles, fluid, debris etc.). These artefacts not only present difficulty in visualizing the underlying tissue during diagnosis but also affect any post-analysis methods required for follow-ups. This is a huge problem during colonoscopy which is an endoscopic surveillance procedure widely done to identify colorectal cancer (CRC). CRC is the third most common cause of cancer mortality with about 1.3 million new cases worldwide [1]. Adenomas or serrated polyps to CRC are the main cause of CRC [2] and can be difficult to detect and remove because of their varying shape, size, appearances, locations and often occlusion with artefacts. Thus computer-aided detection, and segmentation methods can help improve colonoscopy procedures. Even though many methods have been built to tackle automatic detection and segmentation of polyps, benchmarking and development of computer vision methods still remains an open problem. This is mostly due to the lack of datasets or challenges that incorporate highly heterogeneous dataset appealing to participants to test for generalization abilities of the methods [3]. Polyps are usually protrusions (lumps) occurring as a single object or in groups, however, they also disguise themselves in different other appearances such as sessile or flat polyps or hidden behind other protruded mucosal structures [1]. In addition, during colonoscopy multiple artefacts can be present making the procedure more difficult and hard-to detect cancer precursor lesions such as polyp. Thus, this challenge aimed at tackling both of these existing problems using computer vision methods, in particular deep learning as sub-challenges: Endoscopy artefact detection (EAD 2.0) and polyp generalization (PolypGen 2.0). The aim of the sub-challenge EAD 2.0 is to localise bounding boxes, predict class labels and pixel-wise segmentation of 8 different artefact classes for given clinical endoscopy video clips. The 8 classes include specularity, bubbles, saturation, contrast, blood, instrument, blur and imaging artefacts. Similarly, PolypGen 2.0 aimed to benchmark methods on the basis of generalization capabilities to unseen colonoscopy video sequence data for both detection and segmentation deep learning methods. We challenged computer

*4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India*

ali.sharib2002@gmail.com (S. Ali)

vision and computational medical imaging community to participate and build methods that are generalizable in the different clinical settings that we believe provided the adaptability of built and trained methods on different population dataset without requiring them to train from scratch.

## 2. Dataset and challenge

Below we detail the datasets and challenge tasks that was used in each of our sub-challenge:

### 2.1. Datasets

We have already curated large multicenter dataset for both sub-challenges consisting of different endoscopy manufacturers, e.g. Olympus (mostly), Fujifilm, and Karl Storz. Heterogenous collection to reflect real clinical practices worldwide. This includes both standard definition, HD and Ultra HD. For EAD training dataset please refer to our data published at Mendeley[1] and discussed here [4]. A total of 280 patient videos from multiple organs and institutions were used for curating this dataset that led to over 45,478 annotations on both single frame and sequence video data. Training data for the detection task consisted of total 2531 frames with 31,069 bounding boxes while 643 frames with 7511 binary masks for the segmentation task (except for blur, blood and contrast). Sequences were required to mimic the change from large areas of artefacts to small or no artefact frames and vice versa similar to that in the natural occurrence in endoscopic procedures. A detailed overview is also presented in our EndoCV2020 joint paper [4]. A new set of test data were curated that include unique video sequences consisting of more than 500 frames of which 360 was used in leaderboard test assessment. While for "PolypGen 2.0" training data we refer to the newly curated dataset described in [3]. The dataset includes both single frame and sequence data with 3446 annotated polyp labels with precise delineation of polyp boundaries (pixel level for segmentation task and bounding boxes for detection task) verified by six senior gastroenterologists and consists of both small and large polyps including serrated and adenomas. Expert endoscopists (with 20+ years experience) were involved in acquiring all the data. These videos are obtained from routine clinical procedures. To our knowledge, this is the most comprehensive detection and segmentation dataset curated by a team of computational scientists and expert gastroenterologists. In addition to this dataset, we have curated additional 23 unique patient video clips (> 100 frames per video) making in total of 46 sequences for PoypGen2.0 and 24 sequences for EAD2.0. The test phase of this challenge that will make nearly 300-500 frames from multiple centers is the most comprehensive test set allowing for a robust generalizability test of algorithms. To make the competition relate to real-world scenarios we have picked our data centers for both of sub-challenges from different countries that includes Egypt, France, Italy, Norway, Sweden, and UK. The test splits will include - modality split, population split, endoscopy model or manufacturer split and polyp size split. All dataset (including test) will be released after a prospective joint-journal paper. That is, all the data used in the training and testing of the challenge can be used for research and educational purposes. Below we present ethics and annotation strategies involved in our data collection and curation: a) Ethical and privacy aspects of the data: Patient consenting procedure at each individual institution was performed prior to the collection. Additional review of the data collection plan by a local medical ethics committee or an institutional review board was also done in some centers [3, 5]. Challenge organisers performed all anonymisation of the video or image frames (including demographic information) prior to including them into any dataset. Future, build-up of new test samples presented here will follow these ethical procedures. b) Annotation strategy: First, a small subset of dataset will be annotated by all clinical experts and a joint consensus will be made available. Then, the remaining subset of dataset[2] was annotated by post-doctoral researchers (working on endoscopy) and validated by clinicians at two different centres (10-fold cross-validation). Finally, through a joint conference call all annotation validation will be achieved. We will use labelbox [3] for annotation processes. During the entire procedure we aim to produce an annotation protocol and document the entire phenomena which will be released publicly too. A statistical test on annotation variances between experts will also be observed and reported.

### 2.2. Challenge

Each sub-challenge will consist of two tasks:

1. Detection task: The aim of this task will be to test the performance of participants' methods for detection and localization task on our comprehensive and sorted multicenter datasets. The participants will be tested on both detection-based metric and localization metric. A weighted final metric will be used to evaluate for the best performing method.
2. Segmentation task: Similar to task 1, each participants methods will be evaluated on multicenter curated and sorted datasets. An ideal segmentation method will provide the top performance

[1] https://data.mendeley.com/datasets/c7fjbxcgj9/3

[2] https://doi.org/10.17632/c7fjbxcgj9.3

[3] https://labelbox.com

on all the variabilities in different splits and an unseen dataset.

Please note that generalizability assessment of each method will be conducted for both tasks and the winner will be based on this metric (for further details see Section III). Results should be submitted like the provided training ground truth annotations for each task category and as detailed below:

i Category 1 (artefact detection): csv file of bounding box coordinates corresponding to each class (e.g. label, confidence, x1, y1, x2, y2).

ii Category 2 (semantic segmentation): image label masks, integer valued for each image

iii Category 3 (generalization): csv file of bounding box coordinates corresponding to each class (e.g. label, confidence, x1, y1, x2, y2).

## 3. Evaluation metrics and baseline

**Detection task** For detection task we aimed to use widely accepted standard metrics and a generalization metric as detailed below:

- Standard computer vision metric: mean average precision (mAP, IoU interval [0.25:0.05:0.75]) (see PASCAL VOC3 and COCO4 detection challenges)
- Standard intersection over union (IoU, interval [0.25:0.05:0.75])
- Final detection score (trade-off between mAP and IoU): 0.6*mAP + 0.4*IoU (This metric have been used in our previous challenges. The standard metrics using only mAP can lead to very good detection but poor localisation. The penalisation proposed tackles such problem.)
- Generalization gap (Gerror): defined as the difference between detection score and the generalization score (on unseen data) [6]
- Centroid localisation error (Lerror): defined as the distance between centroids positions of detected boxes between the consecutive frames in a video (new)
- Clinical applicability metrics: runtime (to be used post challenge only)

**Segmentation task** For segmentation task we have taken into account widely used standard metrics and a generalization metric as detailed below:

- Standard segmentation metrics that include Dice coefficient (DSC or F1), F2-error, positive predictive value (PPV), Hausdorff distance (HD) and sensitivity (recall) will be used
- The ranking on leaderboard will be based on the highest mean value between DSC, PPV and sensitivity; and the least HD value
- Generalizability difference (Gerror): Difference between DSC on mixed sample data and DSC on unseen data will be key in deciding winner of this task
- Clinical applicability metrics: runtime (to be used post challenge only)

Most of the evaluation metrics are already available at our GitHub repositories(see EAD[4], polypGen6[5]).

**Baseline methods** Based on our previous challenges and current developments in deep learning methods for detection segmentation we have picked three baseline methods that will set the criteria for passing challenge threshold score. Test data was released in two sets. The first set determine which participants go to next round depending on their score threshold. RetinaNet and YOLO-v4 was used as the baseline for detection while UNet, PSPNet and DeepLabV3+ was used as baseline methods with ResNet50 backbone.

**Challenge leaderboard** The EndoCV2022 challenge leaderboard was splitted into two submissions. First submission (referred as round-I) included the results on 50% of the test data while the final submission (referred as round-II) included all 100% of test samples that were used to assess challenge participants methods. Please refer to https://endocv2022.grand-challenge.org/evaluation/round-i-det-gen/leaderboard/. Further, algorithmic details, assessment details, and insights of the developed methods are under compilation and will be published as a joint-journal.

## 4. Conclusion

This paper summarises the motivation of challenge, data collection and preparation, challenge tasks and evaluation metrics used in EndoCV2022 challenge. However, some of the evaluation metrics may have not been included in the leaderboard but is aimed at being used in the joint-journal paper for further analysis.

## References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. Siegel, L. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin. 68 (2018) 394–424.

[4] https://github.com/sharibox/EAD2019

[5] https://github.com/sharibox/EndoCV2021-polyp det seg gen

[2] F. Loeve, R. Boer, A. G. Zauber, M. Van Ballegooijen, G. J. Van Oortmarssen, S. J. Winawer, J. D. F. Habbema, National polyp study data: evidence for regression of adenomas, Int. J. Cancer 111 (2004) 633–639.

[3] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[4] S. Ali, F. Y. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. E. East, G. Wagnières, V. Loschenov, E. Grisan, W. Blondel, J. Rittscher, Endoscopy artifact detection (ead 2019) challenge dataset, ArXiv abs/1905.03209 (2019).

[5] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis 70 (2021) 102002. doi:10.1016/j.media.2021.102002.

[6] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, S. Albarqouni, X. Wang, C. Wang, S. Watanabe, I. Oksuz, Q. Ning, S. Yang, M. A. Khan, X. W. Gao, S. Realdon, M. Loshchenov, J. A. Schnabel, J. E. East, G. Wagnieres, V. B. Loschenov, E. Grisan, C. Daul, W. Blondel, J. Rittscher, An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific Reports 10 (2020). URL: https://doi.org/10.1038%2Fs41598-020-59413-5. doi:10.1038/s41598-020-59413-5.