Kvasir-Instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy

Debesh Jha[1,2], Sharib Ali[9], Krister Emanuelsen[3], Steven A. Hicks[1,5], Vajira Thambawita[1,5], Enrique Garcia-Ceja[10], Michael A. Riegler[1], Thomas de Lange[4,6,7], Peter T. Schmidt[8], Håvard D. Johansen[2], Dag Johansen[2], and Pål Halvorsen[1,5]

[1] SimulaMet, Norway
[2] UIT The Arctic University of Norway
[3] Simula Research Laboratory, Norway
[4] Augere Medical AS, Norway
[5] Oslo Metropolitan University, Norway
[6] Medical Department, Sahlgrenska University Hospital-Mölndal, Sweden
[7] Department of Medical Research, Bærum Hospital, Norway
[8] Karolinska University Hospital, Sweden
[9] Dept. of Engineering Science, University of Oxford, Oxford, UK
[10] Sintef Digital, Norway
debesh@simula.no

Abstract. Gastrointestinal (GI) pathologies are periodically screened, biopsied, and resected using surgical tools. Usually the procedures and the treated or resected areas are not specifically tracked or analysed during or after colonoscopies. Information regarding disease borders, development and amount and size of the resected area get lost. This can lead to poor follow-up and bothersome reassessment difficulties post-treatment. To improve the current standard and also to foster more research on the topic we have released the "Kvasir-Instrument" dataset which consists of 590 annotated frames containing GI procedure tools such as snares, balloons and biopsy forceps, etc. Beside of the images, the dataset includes ground truth masks and bounding boxes and has been verified by two expert GI endoscopists. Additionally, we provide a baseline for the segmentation of the GI tools to promote research and algorithm development. We obtained a dice coefficient score of 0.9158 and a Jaccard index of 0.8578 using a classical U-Net architecture. A similar dice coefficient score was observed for DoubleUNet. The qualitative results showed that the model did not work for the images with specularity and the frames with multiple instruments, while the best result for both methods was observed on all other types of images. Both, qualitative and quantitative results show that the model performs reasonably good, but there is a large potential for further improvements. Benchmarking using the dataset provides an opportunity for researchers to contribute to the field of automatic endoscopic diagnostic and therapeutic tool segmentation for gastrointestinal (GI) endoscopy.

Keywords: Gastrointestinal endoscopy · Tool segmentation · Endoscopic instrument · Convolutional Neural Network · Benchmarking

1 Introduction

Minimally Invasive Surgery (MIS) is a commonly used technique in surgical procedures. The advantage of MIS is that small surgical incisions are made in the patient for endoscopy that causes less pain, reduced time of the hospital stay, fast recovery, reduced blood loss, and less scaring process as compared to the traditional open surgery. The nature of the operation is complex, and the surgeons have to precisely tackle hand-eye coordination, which may lead to restricted mobility and a narrow field of view [5].

However, unlike the treatment of accessory organs such as liver and pancreas, no incision is required for GI tract organs (*oesophagus, stomach, duodenum, colon, and rectum*). GI procedures also includes both, minimally invasive surveillance and treatment (*including surgery*) procedures. A varied number of tools are used as per the requirement of these procedures. For example, balloon dilatation to help open the GI surface, biopsy forceps for tissue sample collection, polyp removal with snares and submucosal injections.

A computer and robotic-assisted surgical system can enhance the capability of the surgeons [9]. It can provide the opportunity to gain additional information about the patient, which can be useful for decision making during surgery [6]. However, it is difficult to understand the spatial relationship between surgical instruments, cameras, and anatomy for the patient [12]. In GI tract endoscopy, it is vital to track and guide surgeons during tumor resection or biopsy collection from a defined site, and help to correlate the biopsied samples and treatment locations post-diagnostic and therapeutic or surgical procedures. While most datasets and automated-algorithm developments for instrument segmentation are mostly focused on laparoscopy-based surgical removal, automatic guidance of tools for GI tract surgery has not been addressed before.

New developments in the area of robot-assisted systems show that there is potential for developing a fully automated robotic surgeon [15]. The da Vinci robot is a surgical system that is considered the de-facto standard-of-care for certain urological, gynecological, and general procedures [4]. Thus, it is critical to have information regarding the intra-operative guidance, which plays an essential role in decision making. However, there are specific challenges, such as limited field of view and difficulties with the surgeons handling the instruments during surgery [14]. Therefore, image-based instrument segmentation and tracking are gaining more and more attention in both robotic and non-robotic minimally invasive surgery. Previous work targeting instrument segmentation, detection, and tracking on endoscopic video images failed on challenging images such as images with blood, smoke, and motion artifacts [14]. Other reasons that make semantic segmentation of surgical instruments a challenging task are the presence of images containing shadows, specular reflections, blood, camera lens fogging, and the complex background tissue [15]. The segmentation masks of

these images can be useful for instrument detection and tracking. Similarly, in the GI tract procedures, from tissue sample collection to surgical removal of pathologies is performed in low field-of-view areas. Visual clutter such as artifacts, moving objects, and fluid, hinders the localisation of the target site during surgical procedures. Additionally, currently, there is no way of correlating the tissue sample collection with biopsied location and assessing surgical procedure effectiveness or even post-treatment recovery analysis. Automated localisation and tracking of instruments can help guide the endoscopists and surgeons to perform their tasks more effectively. Also, post-procedure video analysis can be done using these automated methods to track such tools, thus enabling improved surgical procedures or surveillance and their post-assessment. Currently, this is an open problem in the research community, where most procedures are not automated in GI tract endoscopy.

While there is an open research question for the automated tool detection and guidance in GI procedures, there is a lack of available public datasets. We aim to initiate the development of automated systems for the segmentation of GI tract diagnostic and therapeutic endoscopy tools. This research direction will enable tracking and localisation of essential tools used in endoscopy and help to improve targeted biopsies and surgeries in complex GI tract organs. To accomplish this, and to address the lack of publicly available labeled datasets, we have publicly released 590 pixel-level annotated frames that comprise of tools such as balloon dilation for facilitating opening of GI organs, biopsy forceps for tissue sample collection, polyp removal with snares, submucosal injections, radio-frequency ablation of dysplastic mucosa using probes and some other related surgical/diagnostic procedures. The released video frames will allow for building automated machine learning algorithms that can be applied during clinical procedures or post-analyses. To commence this effort, we provide a baseline benchmark on this dataset. U-Net [13] is a common semantic segmentation based architecture for medical image segmentation tasks. In this paper, we thus present results utilising two U-Net based architectures. The provided dataset is open and can be used for research and development, and we invite multimedia researchers to improve over the provided baseline methods. The main contributions of this paper are:

- Release of 590 annotated bounding box and segmentation masks of GI diagnostic and surgical tool dataset. To the best of our knowledge, this is the first dataset of segmented tools in the GI tract.
- Benchmark of the provided dataset using U-Net and DoubleUNet architectures for semantic segmentation. Standard computer vision metrics are used for a fair comparison of methods and possible future work.

2 Related Work

Surgical vision is evolving as a promising technique to segment and track instruments using endoscopic images [6]. To gather researchers on a single platform,

Table 1: Available instrument datasets

Dataset	Content	Task type	Procedure
Instrument segmentation and tracking (2015) [6]	Rigid and robotic instruments	Segmentation and tracking	Laparoscopy
Robotic Instrument Segmentation (2017) [4]	Robotic surgical instruments	Binary segmentation, part based segmentation, instrument segmentation	Abdominal porcine
Robotic Scene Segmentation (2018) [3]	Surgical instruments and other	Multi-instance segmentation	Robotic nephrectomy
Robust Medical instrument segmentation (2019) [14]	laparoscopic instrument	Binary segmentation, multiple instance detection, multiple instance segmentation	Laparoscopy
Kvasir-Instrument	Diagnostic and therapeutic tools in endoscopic images	Binary segmentation Detection and localization	Gastroscopy & colonoscopy

Endoscopic vision (EndoVis) challenge is being organized since 2015 at Medical Image Computing and Computer Assisted Intervention Society (MICCAI) with an exception in 2016. The Endovis challenge hosts different sub-challenges. The year-wise information about the hosted sub-challenge can be found on the challenge website[1].

Bodenstedt et al. [6] organized "EndoVis 2015 Instrument sub-challenge" for developing new techniques and benchmarking the Machine Learning (ML) algorithm for segmentation and tracking of the instruments on a common dataset. The organizers challenged on two different tasks, (1) Segmentation, (2) Tracking. The goal of the challenge was to address the problem related to segmentation and tracking of articulated instruments in both laparoscopic and robotic surgery[2]. A comprehensive evaluation of the methods used in instrument segmentation and tracking task for minimally invasive surgery is summarized in this work [6]. The extensive evaluation showed that deep learning works well for instrument segmentation and tracking tasks.

In 2017, a follow up to the previous 2015 challenge was organized called "Robotic Instrument Segmentation Sub-Challenge"[3]. The challenge was part

[1] `https://endovis.grand-challenge.org/`

[2] `https://endovissub-instrument.grand-challenge.org/EndoVisSub-Instrument/`

[3] `https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/`

of the Endoscopic vision challenge that was organized at MICCAI 2017. This challenge offered three tasks: (1) Binary segmentation task, (2) Parts based segmentation task, and (3) Instrument type segmentation task. The goal of the binary segmentation task was to separate the image into an instrument and background. Parts segmentation challenged the participants to divide the binary instrument into a shaft, wrist, and jaws. Type segmentation challenged the participants to identify different instrument types. A detailed description of the challenge tasks, dataset, methodologies used by ten participating teams in different tasks, challenge design, and limitation of the challenge can be found in the challenge summary paper [4].

In 2019, a similar challenge called "Robust Medical Instrument Segmentation Challenge 2019"[4] was organized by Roß et al. [14]. This challenge offered three tasks (1) Binary segmentation, (2) Multiple instance detection, and (3) Multiple instance segmentation. The challenge was focused on addressing two key issues in surgical instruments, *Robustness* and *Generalization*, and benchmark medical instrument segmentation and detection on the provided surgical instrument dataset. EAD2019 challenge focused on endoscopic artefact detection primarily, but also included instrument class in their detection, segmentation and "out-of-sample" generalisation tasks. The challenge outcome revealed that most methods performed well for instrument detection and segmentation class [2]. However, this dataset mostly consisted of large biopsy forceps.

In Table 1, we present available instrument datasets in the field. All of the datasets were designed for hosting challenges. The training dataset is released for all the datasets (except ROBUST-MIS); however, the test dataset is not provided by the challenge organizers. Thus, it makes it difficult to calculate and compare the results on the test dataset. However, experiments are still possible by splitting the training dataset into train, validation, and testing sets. The Robust Medical instrument segmentation dataset is yet not public. However, the participants who have participated in the challenge have the opportunity to download the training dataset. Usually, there are certain practicalities to download the dataset, such as signing the agreement and, getting permission from the owner, which takes time, and it is inconvenient. Moreover, to participate in the challenge, the participants have to signup in the particular year, and usually, the organizers do not make the dataset public unless they make a publication out of it, meaning it may take up to years. Thus, the significance of the datasets becomes less as the technology is changing rapidly. More information on available instrument datasets, contents, and offered tasks by the organizers and about the availability can be found from Table 1.

The literature review shows that there are only a few open-access datasets for MIS instrument segmentation. However, to the best of our knowledge, GI tract organ tools have never been explored. This is the first attempt to identify this avenue and provide the community with a curated and annotated public dataset that comprises of diagnostic and therapeutic tools in the GI tract. We believe that the presented dataset and the widely used U-Net based algorithm

[4] `https://robustmis2019.grand-challenge.org/`

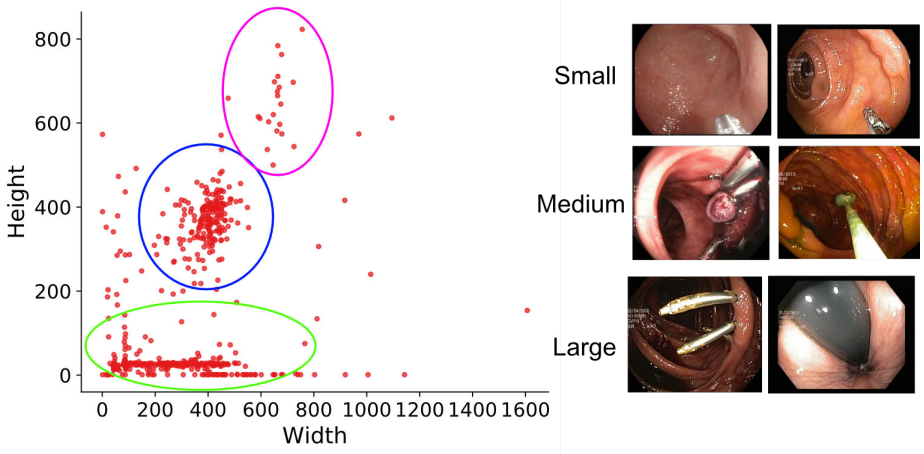

Fig. 1: Distribution of Kvasir-Instrument dataset. On left: Small (green), medium (blue) and large (pink) sized tool clusters. On right: sample images with variable tool size in images.

benchmark will encourage the multimedia researchers to develop a robust and efficient algorithm on the provided dataset that can help clinical procedures in endoscopy.

3 Kvasir-Instrument dataset

In this section, we introduce the Kvasir-Instrument dataset with details on how the data was collected, the annotation protocol, and the dataset's structure. The dataset was collected from endoscopic examinations performed at the Bærum Hospital in Norway. The unlabelled images' frames are selected from the HyperKvasir dataset [7]. HyperKvasir provides frame-level annotations for 10,662 frames for 23 different classes. However, the majority of the images (99,417 frames) are not labeled. We trained a model using the labeled samples of this dataset and tried to predict the classes of the unlabeled samples. Although our algorithm [16, 17] could not classify all the images correctly; however, we were able to classify the instrument class out of hundreds of thousands of provided image frames. Additionally, some images were extracted manually from the polyp class of the Kvasir-SEG [11] dataset. Below, we present the acquisition and annotation protocols used in the data preparation:

Data acquisition: The images and videos were collected using standard endoscopy equipment from Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany) at Vestre Viken Hospital Trust, Norway. All the data used in this study were obtained from videos for procedures that had followed the patient consenting protocol of Bærum Hospital. Additionally,

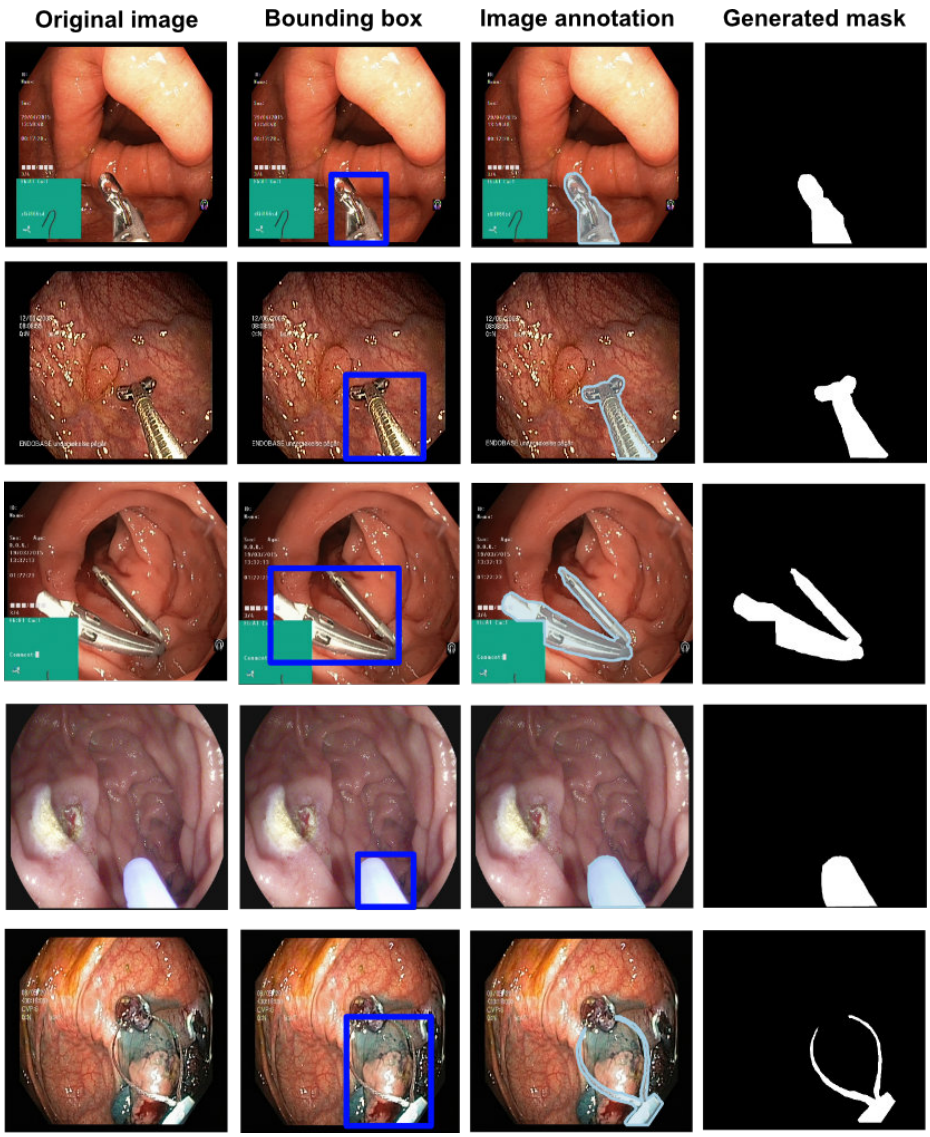

Fig. 2: Kvasir-Instrument dataset: First two rows represent frames with biopsy forceps, the middle row consist of metallic clip, the fourth row is a radio-frequency ablation probe and the last row depicts the crescent and hexagonal shaped snares for polyp removal.

no patient information was used for archiving. We have performed a random naming for each publicly released frame for effective annonymisation.

Annotation strategy: We have uploaded the Kvasir-Instrument dataset to labelbox[5] and labeled the Region of Interest (ROI) in the image frames, i.e., the ROI of diagnostic and therapeutic tools in our cases and generated all the ground truth masks. Figure 2 shows the example images, bounding box, image annotation, and generated masks for the Kvasir-Instrument dataset. All annotations were then exported in a JSON format which was used to generate masks for each of the annotations. Related codes and more information about the dataset can be found here[6].

The exported file contained the information of the images along with the coordinate points that were used for mask and bounding box generation. All annotations were performed using a three-step strategy:

- First, the selected samples were labeled by two experienced research assistants.
- The annotated samples where cross-validated for their delineation quality by two experienced GI experts (more than 10 years of work experience in colonoscopy).
- Finally, the suggested changes were incorporated using the comments from experts and were validated for only those samples.

The Kvasir-Instrument dataset includes 590 frames consisting of various GI endoscopy tools used during both, endoscopic surveillance and therapeutic or surgical procedures. A thorough annotation strategy (detailed above) was used to create bounding boxes and segmentation masks. The dataset consists of variable tool size with respect to image height and width as presented in Figure 1. The majority of the tools are small and medium-sized. The sample bounding box annotation, precise area delineation and extracted masks, are shown in Figure 2.

Our dataset is publicly available, and can be accessed at: `https://datasets.simula.no/kvasir-instrument/`. It consists of original image samples (in JPEG format), their corresponding masks (in PNG format), and bounding box information (in JSON format). A sample python script to help researchers visualise the data is also provided.

4 Benchmarking, results and discussion

In this section, we explore encoder-decoder based classical models for baseline algorithm benchmarking, their implementation details for reproducibility, details on evaluation metric used for quantitative analysis, and results and discussion.

4.1 Baseline methods

U-Net has been explored in the past through many biomedical segmentation challenges and has shown strength towards an effective supervised segmentation model. In this paper, we, therefore, use U-Net based architectures on our

[5] `https://www.labelbox.com/`

[6] `https://github.com/DebeshJha/Kvasir-Instrument`

Kvasir-Instrument dataset to provide a baseline result for future comparisons. U-Net uses an encoder-decoder architecture, that is, a contractive feature extraction path and expansive path with a classifier to perform binary classification of each image pixel in an upsampled feature map. In our previous work, we have shown that the strength of supervised classification can be amplified by using the output mask from one U-Net [13] architecture to the other by proposing DoubleUNet [10]. In addition, the DoubleUNet architecture uses VGG-19 pretrained on ImageNet as one of the encoder block, squeeze and excite block and Atrous spatial pyramid pooling (ASPP) block. All other components in the network remain the same as the U-Net. For both networks, dice loss gives an $1 - DSC$, where DSC is the dice similarity coefficient (see Eq. 1 below).

4.2 Implementation Details

We have implemented the U-Net-based and DoubleUNet based architectures using the Keras framework [8] with TensorFlow [1] as backend running on the Experimental Infrastructure for Exploration of Exascale Computing (eX3), NVIDIA DGX-2 machine. We have resized the training dataset into 512×512. We set the batch size of 8 for training. Both architectures are optimized by using the Adam optimizer. We have made use of dice loss as the loss function. We split the dataset using 80% of the dataset for training and the remaining 20% for the testing (evaluation). We performed basic augmentation, such as horizontal flip, vertical flip, and random rotation. Moreover, we have also provided the train-test split so that others can improve the methods on the same dataset.

4.3 Evaluation Metrics

In this medical image segmentation approach, each pixel of the diagnostic and therapeutic tool either belongs to a tool or non-tool region. Dice similarity coefficient (DSC) is the main evaluation metric used to evaluate this task. Additionally, we calculate other standard metrics such as Jaccard similarity coefficient (JC) or intersection over union (IoU), precision, recall, overall accuracy, F2, and frames per second (FPS) as it is a commonly used metric in biomedical image segmentation tasks. The mathematical expressions for them are as follows:

$$\text{DSC} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{1}$$

$$\text{JC or IoU} = \frac{tp}{tp + fp + fn} \tag{2}$$

$$\text{Recall } (r) = \frac{tp}{tp + fn} \tag{3}$$

$$\text{Precision } (p) = \frac{tp}{tp + fp} \tag{4}$$

$$\text{F2} = \frac{5p \times r}{4p + r} \tag{5}$$

Table 2: Baseline results for tool segmentation

Method	JC	DSC	F2-score	Precision	Recall	Acc.	FPS
U-Net [13]	**0.8578**	**0.9158**	**0.9320**	**0.8998**	**0.9487**	**0.9864**	**20.4636**
DoubleUNet [10]	0.8430	0.9038	0.9147	0.8966	0.9275	0.9838	10.0000

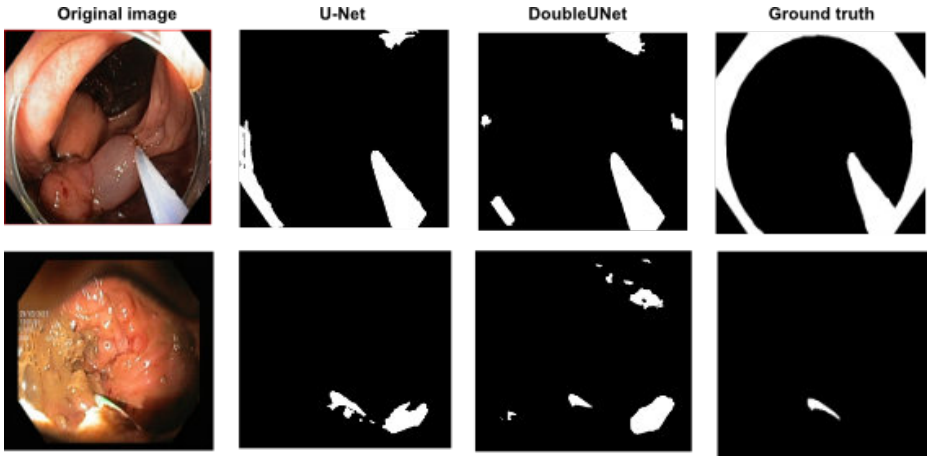

Fig. 3: Failed cases: Cap region (top) is under-segmented and small clip area is over-segmented and consist of large number of false positives (bottom).

$$\text{Overall accuracy } (Acc.) = \frac{tp + tn}{tp + tn + fp + fn} \tag{6}$$

$$\text{Frame Per Second } (FPS) = \frac{1}{sec/frame} \tag{7}$$

Here, tp, fp, tn, fn are the true positives, false positive, true negative, and false negative, respectively.

4.4 Quantitative and Qualitative results

Table 2 shows the results of the baseline methods for the tool segmentation on the Kvasir-Instrument dataset. From the table, we can observe that the UNet achieved a high JC of 0.8578 and DSC of 0.9158, which is slightly above than the DoubleUNet that yielded JC of 0.8430 and DSC of 0.9038. Also, UNet achieved a speed of 20.4636 FPS, whereas computational time is double for DoubleUNet with only 10 FPS. Similarly, both the recall and precision scores are very comparable for both U-Net ($p = 0.8998, r = 0.9487$) and DoubleUNet ($p = 0.8966, r = 0.9275$).

Figure 3 shows the qualitative result on the challenging images. It can be observed that that both UNet and DoubleUNet are under-segmenting the cap

region (top) and over-segmenting the small clip area (bottom). Some parts of these images are confused because of the presence of saturation areas. However, both models was able to segment well with most endoscopic tool samples in the dataset. This is also evident from the quantitative results.

4.5 Discussion

From the experimental results in Table 2, we can validate that the classical U-Net architecture outperforms DoubleUNet model. Additionally, U-Net is 2× faster than the DoubleUNet. This is because U-Net uses basic convolution blocks, whereas DoubleUNet uses pre-trained encoders, ASPP, squeeze and excite blocks, all of which increases the inference latency. Here, the UNet is optimized by dice loss instead of binary cross-entropy loss, which showed improved performance during our experiments.

Further, fine-tuning on other similar datasets, rigorous data augmentation and applying more advanced Deep learning (DL) techniques can improve the baseline results - eventually achieving the detection, localisation, and segmentation performance needed to make the technology useful in a clinical environment. Additionally, use of DL networks with less parameters could increase the computational efficiency thereby enabling real-time systems that can be used in clinical settings effectively.

5 Conclusion

We have curated, annotated, and publicly released a dataset that incorporates tools used in GI endoscopy screening and surgical procedures. The dataset consists of images, bounding boxes and segmentation masks of endoscopy tools used during different procedures in the GI tract. Additionally, we provided baseline segmentation methods for the automatic delineation of these tools and have compared them using standard computer vision metrics. In the future, we plan to continuously increase the amount of data and also call for multi-media challenges on using the presented dataset.

Acknowledgements

This work is funded in part by the Research Council of Norway, project number 263248 (Privaton) and project number 282315 (AutoCap). We performed all computations in this paper on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX^3), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] Abadi, M., et al.: Tensorflow: A system for large-scale machine learning. In: Proceeding of {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}). pp. 265–283 (2016)

[2] Ali, S., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Scientific reports 10(1), 1–15 (2020)
[3] Allan, M., Azizian, M.: Robotic scene segmentation sub-challenge. arXiv preprint arXiv:1902.06426 (2019)
[4] Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
[5] Bernhardt, S., Nicolau, S.A., Soler, L., Doignon, C.: The status of augmented reality in laparoscopic surgery as of 2016. Medical image analysis 37, 66–90 (2017)
[6] Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv preprint arXiv:1805.02475 (2018)
[7] Borgli, H., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Springer Nature Scientific Data (2020)
[8] Chollet, F., et al.: Keras (2015)
[9] Cleary, K., Peters, T.M.: Image-guided interventions: technology review and clinical applications. Annual review of biomedical engineering 12, 119–142 (2010)
[10] Jha, D., et al.: Doubleu-net: A deep convolutional neural network for medical image segmentation. arXiv preprint arXiv:2006.04868 (2020)
[11] Jha, D., et al.: Kvasir-seg: A segmented polyp dataset. In: Proc. of International Conference on Multimedia Modeling. pp. 451–462 (2020)
[12] Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N.: Deep residual learning for instrument segmentation in robotic surgery. In: Proc. of International Workshop on Machine Learning in Medical Imaging. pp. 566–573 (2019)
[13] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. of International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015)
[14] Ross, T., et al.: Robust medical instrument segmentation challenge 2019. arXiv preprint arXiv:2003.10299 (2020)
[15] Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: Proc. of International Conference on Machine Learning and Applications (ICMLA). pp. 624–628 (2018)
[16] Thambawita, V., et al.: The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning. arXiv preprint arXiv:1810.13278 (2018)
[17] Thambawita, V., et al.: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. arXiv preprint arXiv:2005.03912 (2020)