Supplemental Methods

Neural Network Architecture

The Neural Network architecture at the core of our algorithm is a DeepLabv3+ with a KSAC pooling layer [1] and a 101-layer ResNeSt backbone [2]. The network is trained for 72000 iterations with a batch size of 8 to minimize the cross-entropy loss with label smoothing. We set the initial learning rate for Stochastic Gradient-Descent to 1e-2 and polynomial decay over the training iterations. During training we randomly crop the input to equal height and width, apply horizontal and vertical flipping as well as slight alterations to brightness, hue, saturation, and contrast and add gaussian noise with a probability of 0.25.

Computer Vision Annotation Tool (CVAT)

In CVAT, each annotated instance represents a separate layer. These individual layers must be ordered from foreground to background, such that submucosal regions do not cover vessel annotations. Apparent conflicts in the annotations (Figure 1) are only present between the submucosal layer and the other annotated classes and are an artifact of the annotation process. It is more time-efficient to broadly annotate the submucosal region, omitting the subtle geometry of the knife or a small vessel and correct these conflicts with a post-processing by applying predefined ordering that always places the submucosal layer as background, in relation to the knife or vessel classes.

Image Annotation

Annotation of five categories within the training and test images was performed, including: 1) Submucosal vessels; 2) Submucosal layer; 3) Muscle layer, 4) Electrosurgical knife, 5) instrument shaft. Annotation was performed by expert endoscopists with an ESD experience of at least 500 procedures using the Computer Vision Annotation Tool (CVAT, doi: 10.5281/zenodo.4009388). The aim of annotation was to provide the ground truth for training and subsequent cross-validation or testing.

Training and Validation on still images

12 ESD- and 4 POEM-videos of about one hour duration per video were used for training and cross-validation. For the five-fold cross-validation, a total of 2012 frames were extracted from the videos. 453 further annotated frames from 9 ESD- and 2 POEM-videos were used for an additional performance test. These videos were not part of the training or cross-validation set. All images for training and validation were resized to a resolution of 512 x 640 pixels.

The individual folds are selected on a sequence level. Since all images from one sequence are only part of the validation set once, with varying amounts of images per sequence, each fold consists of a different number of training and validation data. Images were taken as screenshots from the ESD- and POEM-videos during the submucosal dissection stage and were selected to have a balanced distribution of the annotated classes. All procedures were performed at the University Hospital Augsburg using Olympus EVIS X1 CV-1500 series. Ethics approval for use of deidentified image and video material had been granted by the Ethics Committee of Ludwigs-Maximilians-Universität, Munich (Project Nr: 21-1216).

Validation on video

From three third-space endoscopic videos (1x rectal ESD; 1x esophageal ESD, 1x POEM) 31 video clips with special characteristics were extracted. Each video had to be 15 seconds to 100 seconds, within the first 5 seconds no vessel could be visible. To be included and regarded as relevant, a vessel had to have a diameter of at least 1mm [3] (reference: thickness of the electrosurgical knife shaft, Hook Knife J, Olympus, Tokyo, Japan). For two vessels to be counted separately within one clip, they had to have a space between each other of at least 3 mm (reference: thickness of the endoscopic instrument shaft). In Y-shaped vessels the same condition was applied for distance and length of the two arms. Vessels with a diameter of over 2mm were counted regardless of their distance to other vessels. From the 3 videos all vessels, which could be shown in clips according to these rules were extracted.

Hereby 27 videos containing a total of 52 predefined vessels were assembled. Four videos without vessels were also purposefully included in the test.

These videos were viewed frame by frame with AI overlay and for every positive measurement of a vessel it was determined visually if the measurement overlapped with a predefined vessel. For non-corresponding measurements it was determined visually, if a previously undetected vessel was visible, otherwise the measurement was counted as false positive. For analysis, false positive structures were counted.

Performance Measures

The algorithm's performance was evaluated by calculating the intersection over union (IoU) and Dice-Score. These metrics represent the percent overlap between expert annotation (ground truth) and the segmentation results of the algorithm. The IoU is the ratio between the correctly predicted area and the union of predicted and ground-truth regions. The Dice-Score is similar but puts a larger emphasis on the true positive region in the calculation. The pixel accuracy is computed for all classes at once and is the percentage of correct predictions among all predictions. All measures take values between 0 and 100 %. An IoU or Dice-Score of 0 % would mean no overlap between ground truth and AI prediction, while a Score of 100% would mean complete congruence between the two. If the prediction and ground truth have the same dimensions, but the prediction is shifted to the side such that only 50% of the prediction lies within the ground truth, the resulting IoU would be 33%. The degree of overlap that is satisfactory depends on the segmentation task in question, as in some circumstances, detection is more important than exact delineation.

$$IoU = TP / (TP + FP + FN)$$
$$Dice\ Score = 2\ TP / (2\ TP + FP + FN)$$
$$Pixel\ Accuracy = (TP + TN) / All$$

Abbreviations: TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives, All = all elements considered

To evaluate the model, we split the 16 video sequences into five cross-validation folds. The frames of a single video are either only present in the current training or the validation set. The presented validation metrics are calculated by accumulating the per-fold outputs in order to achieve one result for the whole validation set. The stated metrics are calculated from the fully trained model without early stopping on the best validation result.

In addition to the cross-validation results, we also demonstrate the performance on a separate test set that was strictly excluded during training. We applied the five previously trained fold-specific models as an ensemble to the test data, such that the segmentation of a single testing image is the average output of the five fully trained models.

The VDR was determined as the number of correctly detected vessels divided by the number of predetermined vessels.

References:

1 Ye Huang, Qingping Wang, Wenjing Jia, Lu Yue, Xiangjian He. See More Than Once - Kernel-Sharing Atrous Convolution for Semantic Segmentation. arXiv preprint arXiv:190809443 2019.

2 Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, Alexander Smola. ResNeSt: Split-Attention Networks. arXiv preprint arXiv:200408955 2020.

3 Yoshida N, Naito Y, Kugai M, Inoue K, Wakabayashi N, Yagi N*, et al.* Efficient hemostatic method for endoscopic submucosal dissection of colorectal tumors. World J Gastroenterol 2010;**16**:4180-6.