Paul Maria Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat P. Müller-Stich, Martin Wagner and Franziska Mathis-Ullrich*

Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery

https://doi.org/10.1515/cdbme-2020-0016

Abstract: Semantic segmentation of organs and tissue types is an important sub-problem in image based scene understanding for laparoscopic surgery and is a prerequisite for context-aware assistance and cognitive robotics. Deep Learning (DL) approaches are prominently applied to segmentation and tracking of laparoscopic instruments. This work compares different combinations of neural networks, loss functions, and training strategies in their application to semantic segmentation of different organs and tissue types in human laparoscopic images in order to investigate their applicability as components in cognitive systems. TernausNet-11 trained on Soft-Jaccard loss with a pretrained, trainable encoder performs best in regard to segmentation quality (78.31% mean Intersection over Union [IoU]) and inference time (28.07 ms) on a single GTX 1070 GPU.

Keywords: computer assisted surgery; endoscopy; minimally invasive interventions; surgical data science.

Problem

Computer assistance in laparoscopic surgery requires scene understanding from images to display critical areas to surgeons during manual navigation and planning, in augmented reality scenarios [1], and to generate safe trajectories for robot assisted surgeries [2]. Recognition and segmentation of different organs and tissue types in laparoscopic images are important sub-problems of image based scene understanding [3]. In laparoscopic cholecystectomy (i.e., gallbladder removal), for example, it is crucial to correctly discriminate between fat, liver, and gallbladder tissue. Deep Learning (DL) approaches are a promising technology for semantic segmentation of images [4]. However, different combinations of neural architectures and hyperparameters result in various outcomes across tasks [5]. For example, the selection of an applicable loss function for specific DL tasks is a challenging endeavor [6] itself. DL approaches for semantic segmentation in laparoscopic surgery are prominently applied to the segmentation of instruments [7] and *ex-vivo* datasets that often lack complexity due to the absence of fatty tissue occluding other tissue types [8]. As the visual features of tissues differ to those of instruments with clear edges and contours, it is challenging to predict which architecture-hyperparameter combinations perform best for semantic segmentation of different tissue types in laparoscopic images. Most research addressing semantic segmentation of organs and tissue types, as seen in Figure 1, is based on Support Vector Machines and Super Pixels [9], or segments only a small number of different tissues [3] (e.g., uterus and ovaries).

The aim of this work is to compare the segmentation performance of different state of the art neural networks when trained on different loss functions, each with frozen or trainable, pretrained encoders to investigate their applicability as components in cognitive systems for laparoscopic surgery.

Material and methods

Dataset

The data used for training, validation, and testing is composed of manually annotated frames of videos 1 and 2 of the Surgical Workflow and Skill Analysis of the Endoscopic Vision Challenge 2019 [10] dataset. The dataset contains 12 videos of laparoscopic cholecystectomies, recorded at a rate of 25 Hz and resolution of 960 × 540. Every 500th frame of the videos is extracted and segmented by medical students with a polygon annotation tool. The chosen classes for annotation are *out of image* (regions in the frame that appear black due to the current zoom of the endoscope), *liver*, *gallbladder*, *instrument* (regardless of the instrument type), *fat*, and *other* (e.g., abdominal wall, rescue bag etc.). To further increase the number of available samples, annotation conserving image augmentations are applied. The image augmentations are composed of Gaussian blurs with radii of 2 and

***Corresponding author: Franziska Mathis-Ullrich**, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany, E-mail: franziska.ullrich@kit.edu
Paul Maria Scheikl and Stefan Laschewski, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany
Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat P. Müller-Stich and Martin Wagner, Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany

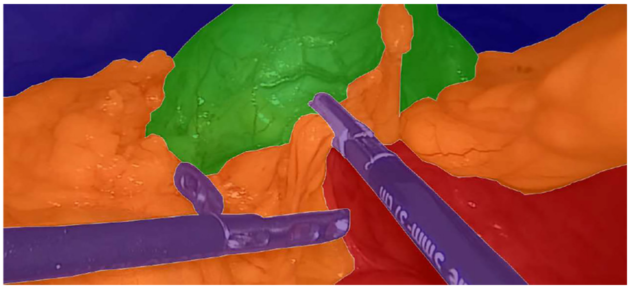

Figure 1: Laparoscopic image overlayed with its semantic segmentation annotations for classes: *fat* (orange), *liver* (red), *gallbladder* (green), *instrument* (purple), and *other* (blue).

5 pixels, centered circle crops with radii of 240 and 330 pixels, image rotations from −10 to 10° in steps of 5°, horizontal flips, and combinations of various image alterations. Figure 3 illustrates one original and two augmented image samples. The data is split by uniform random sampling into subsets for training, validation, and testing with 60, 20, and 20% of the samples, respectively.

Neural network architectures

The investigated neural network architectures in this work represent the state of the art in segmentation of biomedical images (U-Net and TernausNet), road scenes (LinkNet and SegNet), and general object segmentation (Fully Convolutional Network, FCN). The selected architectures and their respective variations are summarized in Table 1. Except for U-Net, all architectures use encoders that were pretrained as classifiers on the ImageNet Dataset [11].

In this work, Soft-Jaccard (SJ) [18], Generalized Dice (GD) [19], and Cross Entropy (CE) loss are explored for training the neural networks, whereas Intersection over Union (IoU) score is used to evaluate the semantic segmentation quality. SJ and GD loss are differentiable approximations of IoU score and Dice's coefficient, respectively.

Experiments

A consistent training procedure, as shown in Figure 2, is implemented to investigate the various combinations of hyperparameters and neural networks. The neural networks are implemented in Python utilizing the DL framework PyTorch. Adam is used as the optimizer for all experiments with its hyperparameters set to the default values (*learning rate* = 0.001, $\beta = (0.9, 0.999)$, $\varepsilon = 1e-08$, without weight decay). The images are scaled to a resolution of 128 × 128 and combined to batches of 20 samples for training.

The investigated parameters are (1) the neural network variation (e.g., TernausNet-16 or SegNet), (2) the loss function employed during the complete training, and (3) trainable or frozen encoder after five epochs of initial training. For the initial training the encoder is frozen to utilize the pretrained features of the encoder in early training stages (except for U-Net, as it does not have a pretrained encoder). After the initial training, the neural networks are trained until the loss function no longer decreases on the validation split over the period of 20 epochs (early stopping). The weights of the best performing epoch in regard to validation loss are used for final testing. Experiments are conducted for 12 different neural networks, three loss functions, and an either trainable or frozen encoder (except for U-Net) resulting in a total of 69 experiments. Each of the 69 experiments is conducted seven times to mitigate the effects of stochastic weight initialization.

Inference time on a single NVIDIA GeForce GTX 1070 Ti is measured and compared against a maximum execution time of 40 ms. This ensures that each frame of a laparoscope operating at 25 Hz can be processed.

Table 1: Investigated neural network architectures and variations. Fully Convolutional Network (FCN) and SegNet are built with VGG-16 encoders. U-Net does not use a pretrained encoder.

Architecture	Variations	Ref.
U-Net	Original	[12]
TernausNet	VGG-{11, 16} encoder	[13, 14]
FCN	Up-sampling sizes {32s, 16s, 8s}	[4]
LinkNet	ResNet-{18, 34, 50, 101, 152} encoder	[15, 16]
SegNet	Original	[17]

Results

Dataset

A total of 210 images were manually annotated by medical students with an approximate duration of 25 min per frame. In total 14 image augmentations were applied to the data, resulting in a total of 3,150 samples. After augmentation, the majority of pixels were part of classes *liver* (26.2%) and *out of image* (20.1%), followed by *other* (23.02%), *fat* (16%), and *instrument* (5.15%).

Semantic segmentation

The five best performing combinations in regard to maximum and mean achieved IoU testing scores over all

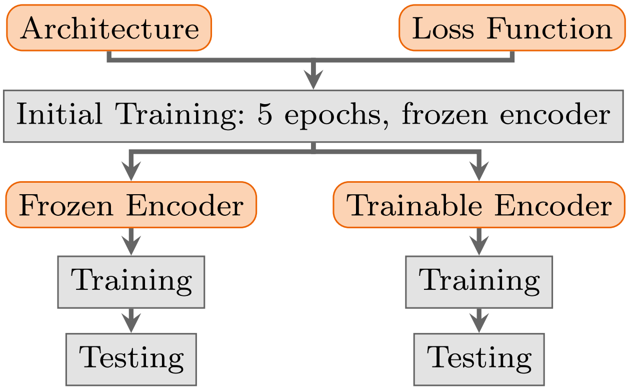

Figure 2: Training experiment pipeline. The experiments are characterized by network architecture, loss function, and whether the encoder is frozen or trainable after the initial training.

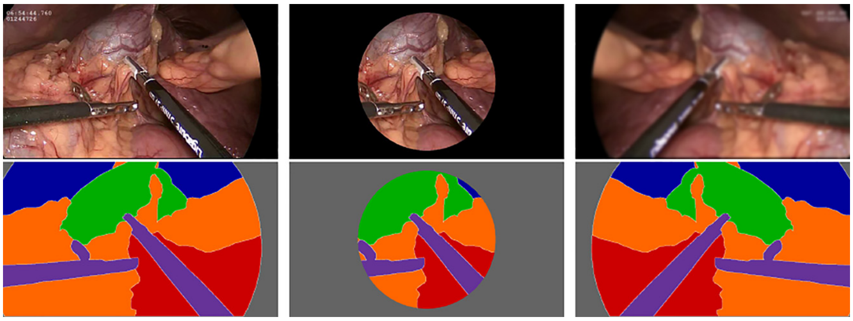

Figure 3: Original, cropped, and blurred & flipped sample images together with their annotation masks.

seven trials are listed in Table 2. The TernausNet architectures trained on the GD loss achieved the overall leading maximum IoU scores and all of the best performing combinations were achieved with a trainable encoder. In regard to mean and maximum IoU score, there is a noticeable difference between achieved IoU scores, with an exception of TernausNet-11 trained on the SJ loss, which shows good performance in both average and maximum IoU scores. Furthermore, when sorted in regard to the mean IoU score the combinations exhibit a standard deviation that is up to two orders of magnitude smaller than the best combinations when sorted in regard to the maximum IoU score.

The IoU scores per class for TernausNet-11 trained on SJ with trainable encoder were examined in more detail, because this combination is among the best in regard to both mean and maximum IoU scores. The highest IoU results were achieved for the class *out of image* (99.54%), followed by classes *instrument* (77.07%) and *fat* (76.8%). The lowest, but still reasonable high scores were achieved for classes *liver* (74.49%), *other* (71.59%), and *gallbladder* (70.35%). Sample images, their annotation, and predictions of the combination are shown in Figure 4.

Except for LinkNet-152 (42.51 ms), all architectures of the best performing networks in Table 2 require less than 40 ms for inference, rendering them utilizable for scenarios with endoscopes that operate at 25 Hz. The remaining networks all require less than 40 ms execution time on a single NVIDIA GeForce GTX 1070: TernausNet-11 (28.07 ms), TernausNet-16 (28.27 ms), FCN-8s (32.17 ms), LinkNet-50 (33.67 ms), and LinkNet-101 (38.13 ms).

Table 2: The five best training combinations in regard to (a) maximum and (b) mean IoU scores (n=7) on the test data.

(a)

Architecture	Loss	Enc.	IoU_{max}, %	IoU_{mean}, %
TernausNet-11	GD	Train.	79.74	55.81 ± 35.65
TernausNet-16	GD	Train.	79.71	28.04 ± 35.45
TernausNet-11	SJ	Train.	79.23	78.31 ± 01.11
TernausNet-16	SJ	Train.	79.22	24.53 ± 37.28
FCN-8s	GD	Train.	78.07	14.40 ± 28.07

(b)

Architecture	Loss	Enc.	IoU_{max}, %	IoU_{mean}, %
TernausNet-11	SJ	Train.	79.23	78.31 ± 01.11
LinkNet-50	SJ	Train.	77.33	76.45 ± 00.78
LinkNet-101	GD	Train.	77.18	76.30 ± 00.78
LinkNet-152	SJ	Train.	77.33	76.25 ± 00.58
TernausNet-11	CE	Train.	77.34	76.24 ± 00.83

Discussion and conclusion

The results suggest, that learned features from pretraining on the ImageNet Dataset do not exhibit sufficient similarity to the relevant features for semantic segmentation of laparoscopic images. Thus, implementing the encoder to be trainable seems to be the sensible recommendation when using pretrained weights from ImageNet. As an outlook, generalizability of the results and networks without pretrained weights may be investigated when increased data is available. Furthermore, data for testing may be drawn from a separate video to avoid possible data leakage effects. In regard to the loss function, the rankings in Table 2 support the findings of [20] who argue that models for semantic segmentation benefit from being trained on SJ and GD over CE.

The comparatively high IoU scores for classes *instrument* and *out of image* indicate that visual features for the respective classes, which are necessary for correct classification, are easier to learn compared to visual features for class *liver*. There are almost five times the number of annotated pixels for class *liver* compared to class *instrument*, however the achieved IoU score for class *instrument* is 2.58% higher than the respective IoU score for class *liver*. The comparatively low IoU score for class *other* is expected, as the class is composed of multiple visually dissimilar classes that share visual features with other classes. Not explicitly annotated classes, such as *cystic duct* in *other*

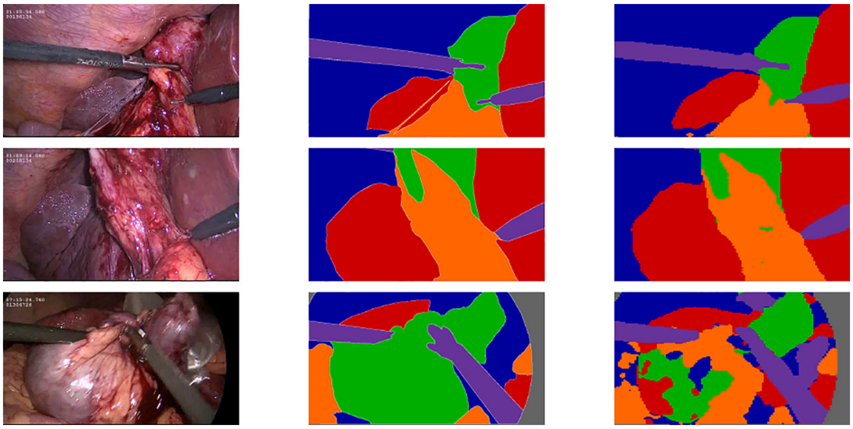

Figure 4: Original input images from the test data (left), their ground truth annotations (center), and respective predictions of TernausNet-11 trained on SJ loss with trainable encoder (right).

may thus be wrongly classified as a visually similar annotated class (e.g. *gallbladder*).

In conclusion, a total of 69 different combinations of neural networks, loss functions, and trainable or frozen encoder were trained on the task to semantically segment laparoscopic images. Based on the conducted experiments, the recommended combination is TernausNet-11 with a trainable encoder and the SJ loss function. The given combination exhibits high segmentation performance with low variance in the trials, as well as fast execution speed. However, it is to be noted, that the considered hyperparameters do not cover the complete search space of possible combinations in a DL training pipeline. Apart from taking more hyperparameters into account, such as utilized optimizer and learning rate, further research should investigate whether the achieved performance allows for integration into the vision pipeline of cognitive surgical robots or context-aware assistance systems.

Research funding: The author state no funding involved.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission
Conflict of interest: Authors state no conflict of interest.

References

1. Teatini A, Pelanis E, Aghayan DL, Kumar RP, Palomar R, Fretland ÅA, et al. The effect of intraoperative imaging on surgical navigation for laparoscopic liver resection surgery. Sci Rep 2019;9:1–11.
2. Hashizume M, Yasunaga T, Tanoue K, Ieiri S, Konishi K, Kishi K, et al. New real-time mr image-guided surgical robotic system for minimally invasive precision surgery. Int J Comput Assist Radiol Surg 2008;2:317–25.
3. Zadeh SM, Francois T, Calvet L, Chauvet P, Canis M, Bartoli A, et al. Surgai: deep learning for computerized laparoscopic image understanding in gynaecology. Surg Endosc 2020:1–7. https://doi.org/10.1007/s00464-019-07330-8.
4. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: IEEE CVPR; 2015: 3431–40 pp.
5. Isensee F. Petersen J, Kohl SAA, Jäger PF, Maier-Hein KH. nnu-net: breaking the spell on successful medical image segmentation. ArXiv 2019;abs/1904.08128.
6. De Brabandere B, Neven D, Van Gool L. Semantic Instance Segmentation with a Discriminative Loss Function. CoRR 2017; abs/1708.02551. http://arxiv.org/abs/1708.02551. arXiv: 1708.02551 [cs].
7. Bodenstedt S, Allan M, Agustinos A, Du X, Garcia-Peraza-Herrera L, Kenngott H, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. ArXiv 2018;abs/1805.02475.
8. Allan M, Kondo S, Bodenstedt S, Leger S, Kadkhodamohammadi R, Luengo I, et al. 2018 robotic scene segmentation challenge. ArXiv 2020; abs/2001.11190.
9. Moccia S, Wirkert SJ, Kenngott H, Vemuri AS, Apitz M, Mayer B, et al. Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Trans Biomed Eng 2018;65:2649–59.
10. EndoVis – Home [Online]. Available from: https://endovis.grand-challenge.org/ [Accessed 27 Oct 2019].
11. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211–52.
12. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. CoRR 2015; abs/1505:04597. https://doi.org/10.1007/978-3-319-24574-4_28.
13. Iglovikov V, Shvets A. TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation. CoRR 2018;abs/1801:05746.
14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR 2014;abs/1409:1556.
15. Chaurasia A, Culurciello E. LinkNet: exploiting encoder representations for efficient semantic segmentation. CoRR 2017;abs/1707:03718. https://doi.org/10.1109/VCIP.2017.8305148.
16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CoRR 2015;abs/1512:03385. https://doi.org/10.1109/CVPR.2016.90.

17. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. CoRR 2015;abs/1511:00561. https://doi.org/10.1109/TPAMI.2016.2644615.
18. Rahman MA, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In: ISVC; 2016:234–44 pp. https://doi.org/10.1007/978-3-319-50835-1_22.
19. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support; 2017:240–8 pp.
20. Bertels J, et al. Optimizing the dice score and jaccard index for medical image segmentation: theory and practice. In: MICCAI; 2019:92–100 pp.