Matthias Ivantsits*, Lennart Tautz, Simon Sündermann, Isaac Wamala,
Jörg Kempfert, Titus Kuehne, Volkmar Falk and Anja Hennemuth

DL-based segmentation of endoscopic scenes for mitral valve repair

https://doi.org/10.1515/cdbme-2020-0017

Abstract: Minimally invasive surgery is increasingly utilized for mitral valve repair and replacement. The intervention is performed with an endoscopic field of view on the arrested heart. Extracting the necessary information from the live endoscopic video stream is challenging due to the moving camera position, the high variability of defects, and occlusion of structures by instruments. During such minimally invasive interventions there is no time to segment regions of interest manually. We propose a real-time-capable deep-learning-based approach to detect and segment the relevant anatomical structures and instruments. For the universal deployment of the proposed solution, we evaluate them on pixel accuracy as well as distance measurements of the detected contours. The U-Net, Google's DeepLab v3, and the Obelisk-Net models are cross-validated, with DeepLab showing superior results in pixel accuracy and distance measurements.

Keywords: deep learning; detection; endoscopic; machine learning; mitral valve; mitral valve leaflet; segmentation; surgery.

***Corresponding author: Matthias Ivantsits**, Charité – Universitätsmedizin Berlin, Berlin, Germany, E-mail: matthiasivantsits@gmail.com
Lennart Tautz, Charité – Universitätsmedizin Berlin, Berlin, Germany; and Fraunhofer MEVIS, Bremen, Germany
Simon Sündermann, Charité – Universitätsmedizin Berlin, Berlin, Germany; and DZHK (German Centre for Cardiovascular Research), Berlin, Germany
Isaac Wamala, DZHK (German Centre for Cardiovascular Research), Berlin, Germany
Jörg Kempfert, , Titus Kuehne and Volkmar Falk, Charité – Universitätsmedizin Berlin, Berlin, Germany; German Heart Center Berlin, Berlin, Germany; and DZHK (German Centre for Cardiovascular Research), Berlin, Germany
Anja Hennemuth, Charité – Universitätsmedizin Berlin, Berlin, Germany; Fraunhofer MEVIS, Bremen, Germany; and DZHK (German Centre for Cardiovascular Research), Berlin, Germany

Introduction

Mitral valve regurgitation is a condition in which the valve does not close properly, sometimes prolapsing upwards into the atrium, and causing blood to leak back into the atrium from the ventricle during cardiac ejection. Consequentially leading to heart failure, mitral valve regurgitation has to be treated by surgery.

According to the German heart surgery report 2017 [1], more than 100,000 heart interventions were conducted in Germany in that year. About 7,100 of these operations were either mitral valve replacements or repairs. While the number of mitral valve repairs has risen approximately by 38% over the last decade, the number of replacements has only increased by around 16%, less than the overall growth of mitral valve interventions. Gillinov et al. [2] expounded the fact that repair is the better short-term and long-term option in these critical cases. 45% of all mitral valve interventions in 2017 were performed as a minimally invasive procedure. During surgery, live video endoscopy is used to visualize the operative field, the heart, and the valve apparatus.

Related work

We reviewed methods to extract anatomical regions of interest and instruments in surgical endoscopic images. Much effort has been put into the research of deep learning (DL) to segment regions of interest in the digestive system, mostly using wireless capsule endoscopes. Vemuri et al. [3] give an exhaustive overview of applications of computer vision and machine learning in gastrointestinal (GI) endoscopy. Shevets et al. [4], Sornapudi et al. [5] and Hajabdollahi et al. [6] contributed to segmentation of images acquired by wireless capsule endoscopy (WCE). The employed models range from standard multilayered perceptron (MLP) to more sophisticated architectures like TernausNet and region based CNNs.

A different field of interest in medical image processing is the segmentation of instruments, which is an essential component of any computer-assisted surgical system. An early implementation proposed by Haase et al. [7]

augments photometric information by range data, exploiting two modalities to create a 3D segmentation of the instruments. Other publications driving the research include Attia et al. [8], Pakhomov et al. [9], Garcia-Peraza-Herrera et al. [10], and Shvets et al. [11]. All implementations make use of CNN models.

The mentioned methodologies separate the tasks of extracting anatomical structures and instruments in surgical settings. Moreover, the proposed methods are applied mostly to the digestive system and liver, which are relatively rigid objects compared to the mitral valve. Therefore, we propose a multi-task architecture, combining the segmentation of highly non-rigid anatomical structures and rigid instruments. The applicability of the proposed method in real-time is essential. Training and validation data in medical settings are very scarce and expensive to acquire, hence the proposed learning methods need to be trainable with a limited dataset.

Approach

For the combined segmentation of anatomical structures and instruments in a given endoscope scene, we propose a CNN architecture. We compare three deep learning architectures. As a baseline, we select a variation of the U-Net architecture by Ronneberger et al. [11]. This model has shown to be successful in various domains, especially in medical image segmentation. Next, we propose Google's DeepLab v3 by Chen et al. [12], which is the benchmark for most RGB image segmentation tasks nowadays. The novel idea of DeepLab v3 is replacing the last ResNet block by an Atrous Spatial Pyramid Pooling (ASPP) block. The third architecture is the Obelisk-Net by Heinrich et al. [13], which employs sparse convolutions with arbitrary offsets to the kernel center. These sparse convolutions have the advantage of drastically decreasing the number of parameters in the model while keeping a large receptive field.

Data

As visual access to the heart, the stereo endoscope EinsteinVision 3.0 Aesculap system was used. The dataset was acquired as single loss-less RGB frames with a spatial resolution of 1920 × 1080 pixels, and a temporal resolution of 30 frames per second.

In total, 540 annotated RGB frames from nine interventions were available for training and validation. For validation, one patient with 60 frames was put aside, leaving 480 frames for training. We selected the frames for each patient based on the scene content, which we grouped into five classes. Each class has three sequences of four frames, resulting in 12 frames for each class, respectively. The five classes show, respectively, a clean shot of the anterior leaflet, a clean shot of the posterior leaflet, both leaflets visible, the left ventricle including the papillary muscle, and both leaflets during a so-called leakage test. The latter one is a test procedure during the intervention, where water is pumped into the left ventricle to test the closure of the valve (Figure 1). Lastly, all annotations were acquired by surgical experts.

Training

Since limited patient data was available for model training and testing, one patient was put aside as a test set; the remaining eight cases were used for 8-fold cross-validation. This pattern was used in all conducted experiments. A grid search with local refinements was performed to find optimal hyperparameters.

The hyperparameters considered include different forms of data augmentation. These involve grid distortion, horizontal flipping, scaling, shifting, rotation, random cutouts, elastic deformation [14], and optical distortion. The optical distortion augmentation adds some shearing to the input image, simulating a change in camera perspective. Furthermore, we tested transfer learning approaches to reuse weights learned on different image segmentation tasks, as well as different learning rates ranging from 5e-6 to 5e-5. The training epochs were fixed to 60. Furthermore, the images were resized to a 960 × 512 resolution to remove the squishing effect of channel interlacing and speed up model training and the final inference time by a factor of four.

Results

The presented experiments were conducted using two images per batch, on an Nvidia RTX 2048ti GPU with 10 GB memory. The baseline U-Net model, the DeepLab architecture and the Obelisk-Net require 120 ms, 90 ms and 20 ms, respectively, for segmentation inference. The U-net and DeepLab architectures achieve similar Dice distributions on both foreground classes (Figure 2), with a mean Dice score over 0.93. The Obelisk-Net is unable to robustly capture the relevant information.

The Hausdorff distance measures the largest minimum distance of two point sets. As it is susceptible to outliers, we extracted the largest connected component for each label and compared the resulting predicted contours to the ground truth labels (Figure 3).

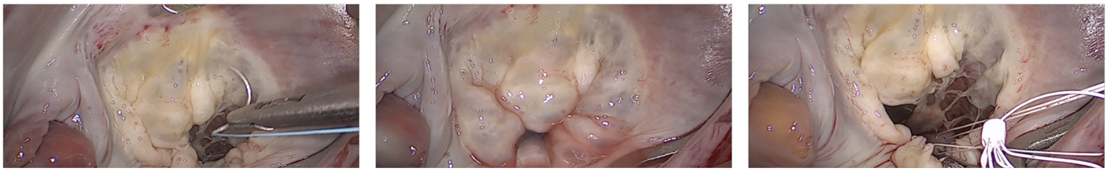

Figure 1: A visualization of a clearly visible anterior leftlet on the left. The mitral valve during a leakage test in the center, and a frame capturing the ventricle on the left.

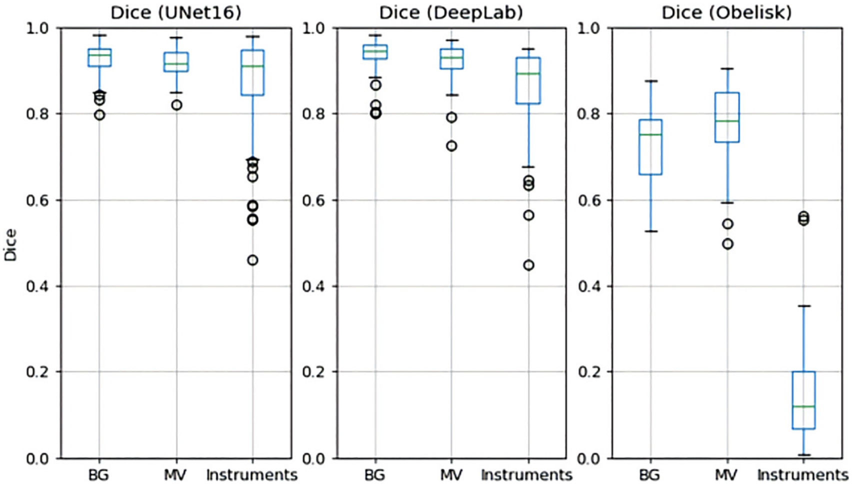

Figure 2: Dice score for all three tested models (left baseline U-net, center DeepLab v3, right Obelisk-Net).

The mean contour distance is roughly 15 pixels for the U-net and the DeepLab model on the mitral valve (Figure 4). Furthermore, these models show a similar mean distance for instruments, but with higher variance. Again, the Obelisk-Net infers segmentations with much higher distances.

Discussion

With respect to the Dice score, the U-Net and DeepLab architectures perform almost equally well (Figure 2). With a similar median in Dice, the DeepLab architecture has a slightly narrower distribution compared to the U-Net. The Obelisk-Net fails to achieve a Dice of 0.6 or more on average. Based on this metric alone it is hard to decide on a preferred architecture, partially due to the small test set.

The Obelisk-Net does not perform well with respect to the contour distance either. Comparing the U-Net and DeepLab distance distributions, the latter architecture is slightly favored, with a lower mean contour distance, and a smaller mean for the Hausdorff distance measurement. Furthermore, the distribution of the DeepLab is narrower

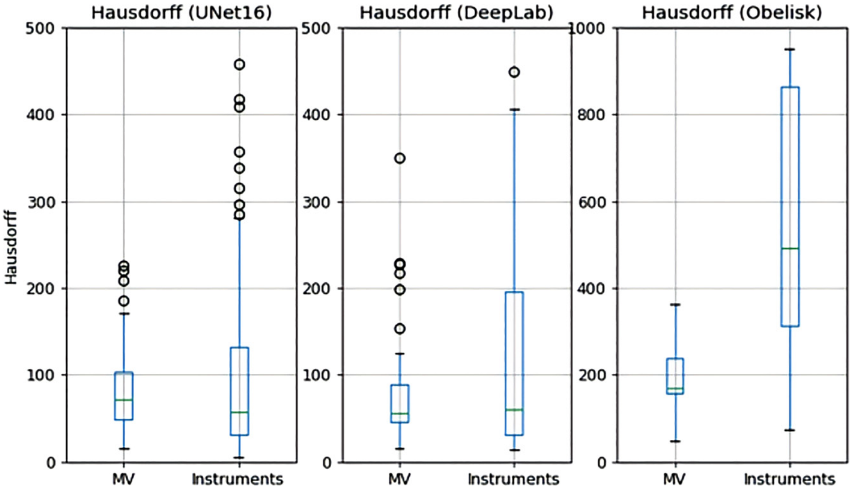

Figure 3: Hausdorff distance for all three tested models (left baseline U-net, center DeepLab v3, right Obelisk-Net). Note the different scaling of the Obelisk-Net plot.

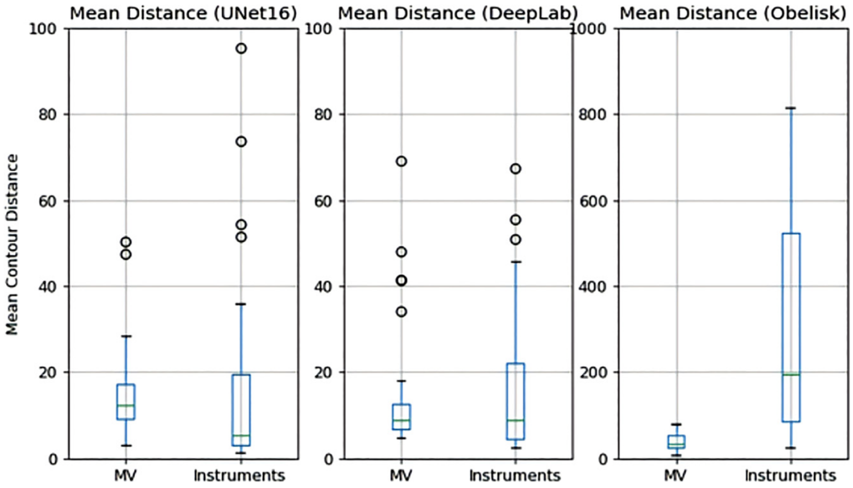

Figure 4: Mean contour distance for all three tested models, with the baseline U-Net onn the left, the DeepLb v3 in the center, and the Obelisk-Net on the right (note the different scaling).

compared to the U-Net, which indicates it is not only performing better on average but also in worst-case scenarios.

Due to real-time constraints, inference time is of particularly high importance to the clinical application developed in this work. The DeepLab network requires only 2/3 in inference time compared to the U-net architecture. When considering all metrics jointly, the DeepLab with superior contour distance, high Dice score and superior inference time outperforms the other architectures in this task.

A notable result is the poor performance of the Obelisk-Net architecture. There are two possible explanations for these results, demanding further research. Either the grid search on the hyperparameters must be improved, or the number of model parameters is too small to adapt to the heterogeneous training data.

Figure 5 shows a result where the network fails to segment the structures properly. The scene contains a lot of strings, which partially are classified as instruments. The boundary of the mitral valve is predicted inaccurately. This complex frame is also hard to label for an expert and leaves much room for discussion.

Lastly, an example where the network performed well is shown in Figure 6. The boundary of the predicted mitral valve seems to match the image better than the ground truth. The expert labels exhibit more curvature on the upper boundary of the valve than actually present in the image.

Generally, the results of these segmentation architectures can be further improved with a larger and more diversified dataset. Furthermore, a more detailed grid search, as well as applying more post-processing techniques such as the largest connected component analysis can further improve the performance. Sequential CNNs, like a recurrent neural network (RNN), or a long short-term

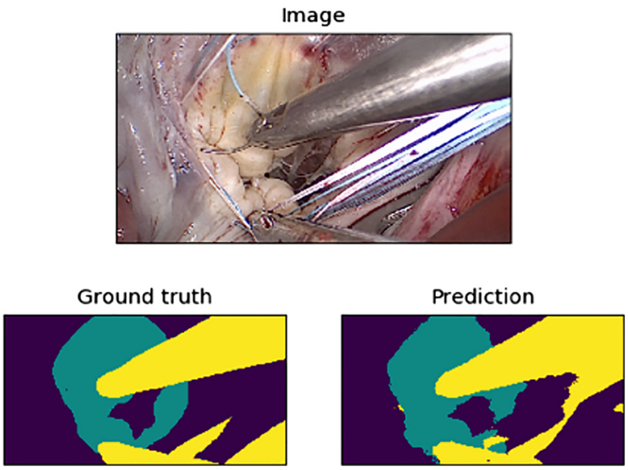

Figure 5: A prediction where the DeepLab architecture is unable to produce an accurate segmentation result.

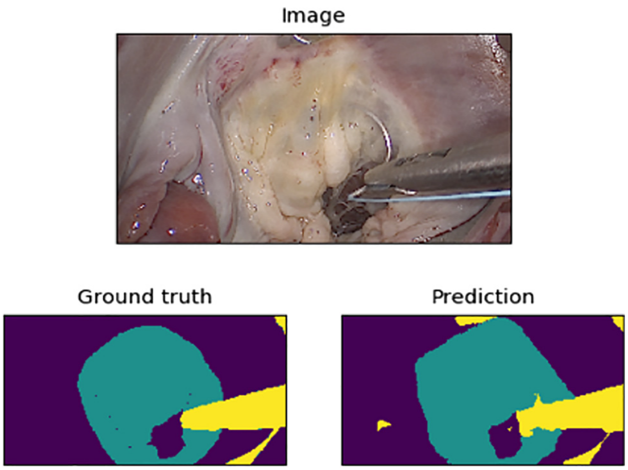

Figure 6: A good prediction by the DeepLab architecture, showing partially better results than the ground truth.

memory (LSTM), can probably enhance the accuracy of the model. Finally, utilizing both stereoscopic image channels can add valuable depth information.

Conclusion

We tested and cross-validated three distinct architectures for endoscopic scene segmentation. We compared the results based on Dice score, contour distances and inference times. The DeepLab model shows superior or comparable results with respect to all metrics, and we consider it the best choice for this task.

Research funding: This work was partially funded by BHI and the BMBF project BIFOLD (grant number 01IS18037E).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Informed consent has been obtained from all individuals included in this study.

Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

1. Beckmann A, Meyer R, Lewandowski J, Frie M, Markewitz A, Harringer W. German heart surgery report 2017. Thorac Cardiovas Surg 2018;66:608–21.
2. Gillinov AM, Wierup PN, Blackstone EH, Bishay ES, Cosgrove DM, White J. Is repair preferable to replacement for ischemic mitral regurgitation?. J Thoracic Cardiovas Surg 2001;122:1125–41.
3. Vemuri AS. Survey of computer vision and machine learning in gastrointestinal endoscopy 2019. https://arxiv.org/abs/1904.13307.
4. Shvets A, Iglovikov V, Rakhlin A, Kalinin AA. Angiodysplasia detection and localization using deep convolutional neural networks 2018. https://arxiv.org/abs/1804.08024.
5. Sornapudi S, Meng F, Yi S. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. Appl Sci 2019;9:2404.
6. Hajabdollahi M, Esfandiarpoor R, Soroushmehr SM, Karimi N, Samavi S, Najarian K. Segmentation of bleeding regions in wireless capsule endoscopy images an approach for inside capsule video summarization 2018. https://arxiv.org/abs/1802.07788.
7. Haase S, Köhler T, Kilgus T, Maier-Hein L, Hornegger J, Feußner H. Instrument segmentation in hybrid 3-D endoscopy using multi-sensor super-resolution 2013. https://www.researchgate.net/publication/267034230_Instrument_Segmentation_in_Hybrid_3-D_Endoscopy_using_Multi-Sensor_Super-Resolution.
8. Attia M, Hossny M, Nahavandi S, Asadi H. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In: 2017 IEEE International conference on systems, man, and cybernetics (SMC). IEEE, Banff, AB, Canada; 2017.
9. Pakhomov D, Premachandran V, Allan M, Azizian M, Navab N. Deep residual learning for instrument segmentation in robotic surgery 2017. https://arxiv.org/abs/1703.08580.
10. Garcia-Peraza-Herrera L, Li W, Gruijthuijsen C, Devreker A, Attilakos G, Deprest J. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking 2017. https://www.researchgate.net/publication/305770331_Real-Time_Segmentation_of_Non-Rigid_Surgical_Tools_based_on_Deep_Learning_and_Tracking.
11. 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. Springer, Cham; 2015.
12. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation 2017. https://arxiv.org/abs/1706.05587.
13. Heinrich MP, Oktay O, Bouteldja N OBELISK-net: fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. Med Image Anal 2019;54:1–9.
14. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation in breast cancer mass detection. In: 2018 IEEE EMBS International conference on biomedical & health informatics (BHI). IEEE, Las Vegas, NV, USA; 2018.