An Augmentation Strategy with Lightweight Network for Polyp Segmentation

Raman Ghimire[a], Sahadev Poudel[b] and Sang-Woong Lee[c]

[a]Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea
[b]Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea
[c]Department of Software, Gachon University, Seongnam 13120, South Korea

Abstract

Automatic segmentation of medical images is a difficult task in computer vision due to the various backgrounds, shapes, sizes, and colors of polyps or tumors. Despite the success of deep learning (DL)-based encoder-decoder architectures in medical image segmentation, it is not always suitable to implement in real-time clinical settings due to its high computation power and less speed. In this EndoCV2021 challenge, we focus on a light-weight deep learning-based algorithm for the polyp segmentation task. The network applies a low memory traffic CNN, i.e., HarDNet68, as a backbone and a decoder. The decoder block is based on a cascaded partial decoder famous for fast and accurate object detection. Further, to circumvent the issue of a small number of images while training, we propose a data augmentation strategy to increase the model's generalization by performing augmentation on the fly. Extensive experiments on the test set demonstrate that the proposed method produces outstanding segmentation accuracy.

Keywords

polyp segmentation, light-weight network, augmentation strategy

1. Introduction

The accurate segmentation of medical images is a deciding step in the diagnosis and treatment of several diseases under clinical settings. The automatic segmentation of diseases can assist doctors or medical professionals in predicting the size of a polyp or lesion and enables continuous monitoring, planning, and follow-up studies resulting in treatment without delay. However, background artifacts, noises, variations in shape and size of the polyps, and blurry boundaries are some of the main factors contributing to more complications for accurate segmentation.

In recent years, owing to the rapid progress in deep learning-based techniques such as convolutional neural networks (CNNs), it is now possible to segment medical images without human intervention. The robust, non-linear feature extraction capabilities of CNNs make it adaptable in other domains such as medical image classification [1, 2, 3], detection [4, 5, 6], image retrieval [7]. In particular, methods such as fully convolutional neural network [8] and encoder-decoder based architectures such as U-Net [9] and SegNet [10] have been widely applied for image segmentation tasks. These networks consist of a contraction path that captures the context in the image, and the symmetric expanding path consists of single or multiple upsampling

3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2021) in conjunction with the 18th IEEE International Symposium on Biomedical Imaging ISBI2021, April 13th, 2021, Nice, France
ghimirermn@gmail.com (R. Ghimire); sahadevp093@gmail.com (S. Poudel); slee@gachon.ac.kr (S. Lee)

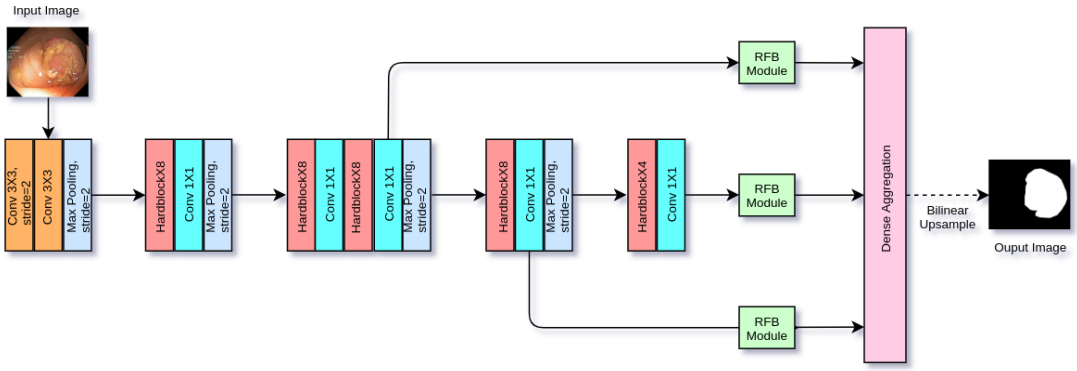

Figure 1: Overall framework of the HarDNet-MSEG [13]

techniques to enable a precise localization [8, 9, 10]. Further, skip-connection techniques have been effective in redeeming fine-grained details, enhancing the network's performance even on a complex dataset.

Inspired by these methods, many approaches have been presented to solve segmentation problems in a wide range of domains. However, the complex architectures, limiting resources, and the low frame per second (FPS) limit the practical implementation of U-Net variants in the clinical domain. Therefore, reducing model size by enhancing both energy and computation efficiency carries great importance. Usually, reduced model size indicates fewer FLOPs(floating-point operation per second) and lesser DRAM(dynamic random-access memory) traffic for reading and writing feature maps and network parameters. State-of-the-art networks like Residual Networks(ResNet) [11], and Densely Connected Networks(DenseNet) [12] have steered the research paradigm towards a compressed model with high parameter efficiency while maintaining high accuracy. When small training sets are available, the traditional deep learning model usually overfits; this lack of available data has been a significant bottleneck in the research field. Not only that, even when enough data is available, there is a high computational cost involved. Therefore, we leverage a lightweight network for accurate polyp segmentation, which provides good segmentation accuracy and speed in comparison to prior methods [13]. Besides decreasing the number of network parameters and the involved computational cost, the presented method also preserves the segmentation accuracy. Further, we propose an augmentation strategy for the polyp segmentation, which helps generalize the model in a complex environment. It encourages the model to learn the semantic features in different variations and scales.

This study's significant contributions can be summarized as follows: First, we leverage the high-speed HarDNet-MSEG model for accurate polyp segmentation. Second, we compares it with other existing architectures with EfficientNetB0 backbone [14] model. Third, we propose an augmentation strategy for the polyp segmentation so that the model can be generalized in a complex environment. Fourth, we evaluated the proposed methodology in different architectures, and the experimental results show the efficiency of our method. Overall, we show that the lightweight network with an improved augmentation strategy can be used in the real-time

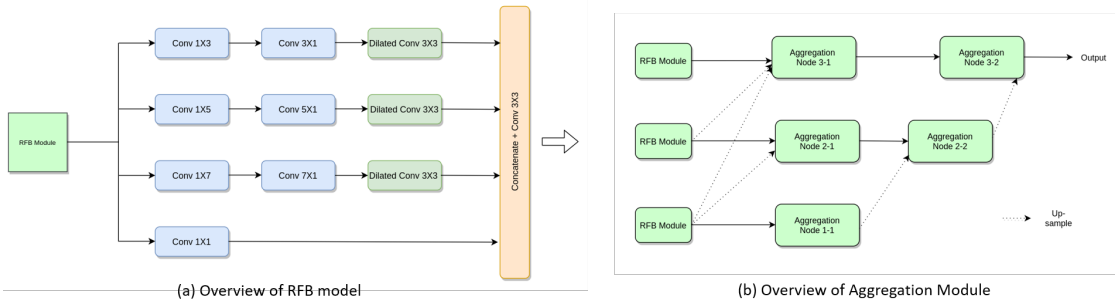

Figure 2: Overall view of each section of the HarDNet-MSEG.

clinical domain.

2. Methodology

Conventional architectures [15, 16] have achieved high accuracy over small model size counterparts but have low inference speed. HardNet [13], considering the influence of memory traffic on model design, achieves an increase in inference speed by reducing the shortcuts, and similarly to make up for the loss of accuracy, increases its channel's width for the key layer. 1x1 convolution is used to increase the computational density. With this, not only is the inference time reduced compared with DenseNet [12] and ResNet [15], the model also achieves higher accuracy on ImageNet [17].

The backbone of this model follows the Cascaded partial decoder [18]. In the cascaded partial decoder, the shallow features are discarded as the deeper layers' can represent the shallow information's spatial details comparably well. The addition of skip connections and appropriate convolution also helps in aggregating the feature maps at different scales. Fig. 3(a) shows a Receptive Field Block [19]. In RFB, varying convolutional and dilated convolutional layers generate features with diverse receptive fields. RFB block is used following the [18] to enlarge receptive fields in feature maps of different resolutions. As shown in Fig. 3(b), in dense aggregation, we upsample the lower scale features and do element-wise multiplication with another feature of the corresponding scale.

3. Experiments

3.0.1. Metrics

The most commonly used metrics for the medical image segmentation are the Dice coefficient and IOU- which can be defined as follows:

$$Dice \quad coefficient = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

Algorithm 1 Detailed augmentation strategy for the training process

1: p indicates the probability for the each augmentation performed on the image.
2: For each training image and corresponding mask:

- Crop the image size into 352 x 352 pixels.
- With probability of p=0.5, perform random rotation [0,90]
- With probability of p=0.5, perform horizontal flip.
- With probability of p=0.5, perform vertical flip.
- With p=0.3, apply one of:
  - random IAAAAdditive gaussian noise
  - gaussian noise
- With p=0.3, apply one of:
  - shiftScaleRotate With scale limit of 0.2 and rotate limit of 45.
  - random brightness shift within the range of -10 to +10 percent.
  - random contrast shift within the range of -10 to +10 percent.
- With p=0.3, apply one of:
  - motion blur.
  - median blur with blur limit of 2.
- With p=0.3, apply one of:
  - mask dropout in the RGB image.
  - gaussian noise
- With probability of p=0.3, perform color jitter of brightness (0.2), contrast (0.2), saturation (0.2) and hue (0.2).

3: Feed the transformed image into the network in each epoch.

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

where, TP represents true positive, FP represents false positive, and FN represents false negative. Both the Dice coefficient and IoU calculate the similarity between the predicted mask and the ground truth mask shown in Eqs. (1) and (2), respectively.

3.1. Implementation Details

We divided the whole EndoCV2021 dataset [20] into training and validation set with a ratio of 80:20 percent. Out of 1452 image, 1162 images are used for the training set and the remaining for the validation set from the challenge. All the images are resized to 352×352 and performed heavy data augmentation shown in Algorithm 1. We implement our model in Pytorch and conduct our experiments on GeForce RTX 2080 Ti. We use Adam optimizer with a learning rate of 0.00001 for all the experiments.

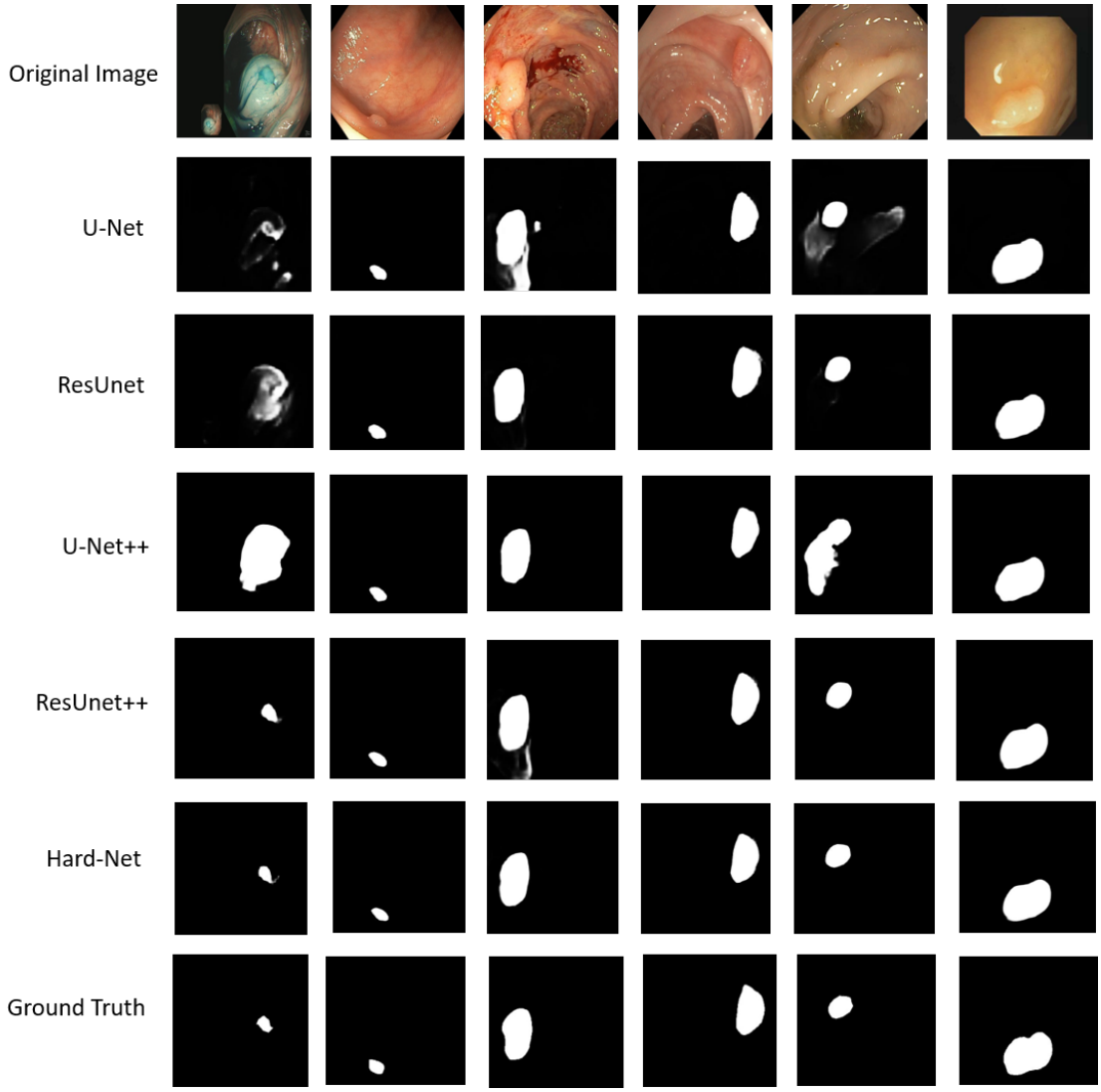

Figure 3: Qualitative analysis of the outputs of different architectures.

3.1.1. Baselines

We perform experiments on several state-of-the-architectures. We employ U-Net [9], U-Net++ [21], ResUNet [22] and ResUNet++ [23] with the EfficientNet-B0 [14] as a backbone to design the efficient and light-weight model. Further, we use rotation, flipping, scaling for normal augmentation and also perform heavy augmentation stated in Algorithm 1 and compares the performance gain in each architectures. With heavy augmentation, the model gets a different transformed image in each epoch and eventually helps in generalization.

Table 1
Experimental results of different architectures before and after heavy augmentation (H. aug) with frame per second (FPS).

Architecture	Jaccard		Dice coeff		FPS
	Before H.aug.	After H.Aug.	Before H.aug.	After H.Aug.	
U-Net [9]	0.7911	0.8226	0.8441	0.8643	65
ResUnet[22]	0.8016	0.8247	0.8539	0.8698	48
U-Net++[21]	0.8232	0.8453	0.8762	0.8859	44
ResUnet++[23]	0.8176	0.8392	0.8527	0.8753	42
Hard-Net[13]	0.8485	0.8649	0.8923	0.9124	88

Table 2
Various augmentation strategies for the polyp segmentation. Probability indicates the chances of applying each transformation on the image.

Augmentation type	Probabilities			
Rotation	0.5	0.5	0.5	0.5
Horizontal Flipping	0.5	0.5	0.5	0.5
Vertical Flipping	0.5	0.5	0.5	0.5
Random IAAAAdditive gaussian noise Gaussian noise	-	0.3	0.5	0.3
ShiftScaleRotate Random brightness shift Random contrast shift	-	0.5	0.5	0.3
Motion blur Median blur	-	0.5	0.5	0.3
Mask dropout	-	0.3	0.5	0.3
Color jitter	-	0.3	0.5	0.3
Dice coefficient	0.75	0.77	0.75	0.78

3.1.2. Experimental Results

Figure 3 displays the qualitative result for the polyp segmentation across several architectures on multi-institutional challenging images. It can be observed that the Hard-Net can segment polyp accurately, almost matching the ground truth of the original images, whereas other architectures like U-Net and ResUnet could not segment the polyps with higher confidence and misses the polyp in some images. Further, U-Net++ over segments the polyp part covering the unwanted parts (see in row 4).

From Table 1, it is observable that the HarDNet achieves higher segmentation accuracy in terms of both the Jaccard index and dice coefficient. The implementation of U-Net with efficientNet backbone obtained the least Jaccard index of 0.8226 after heavy augmentation. Similarly, ResUnet and U-Net++ achieved a Jaccard index of 0.8247 and 0.8453 under the same settings. The ResUnet++ obtains the second position with a Jaccard index of 0.8392 and a Dice coefficient score of 0.8753. Further, Hard-Net also has a higher frame per second (FPS)

than other existing SOTA methods. It obtained the highest speed of 88 FPS while the U-Net, ResUNet, U-Net++, and ResUNet++ have 65,48,44,42 FPS, respectively. Moreover, the augmentation strategy explained in Algorithm 1 also helps increase performance by at least 2 percent in each index. Usually, we started with simple augmentation techniques like rotation, flip (Horizontally and vertically) as a baseline augmentation, and then added other methods like Gaussian noise, blurring, masking, color jittering, etc. We carefully design the probabilities ratio during transformation because a substantial augmentation could lead the model not to learn anything from the input image. Therefore, we keep higher probabilities for the rotation and flipping and comparatively smaller probabilities to other techniques. According to our experiments (fourth column), a strong augmentation could not generalize the model well; instead obtained a similar accuracy as the baseline augmentations.

3.2. Discussion

In clinical settings, an expeditious deep learning method is much needed. Usually, it is found that there is always a tradeoff between the speed and the accuracy while applying a deep learning-based algorithm. However, in this case, Hard-Net surpassed other prior methods in terms of speed and accuracy, which is a good sign for clinical practice. From the extensive experimental results from Figure 3 and Table 1, we can observe that the Hard$-$Net shows improvement over all other existing methods in terms of Jaccard index, dice coefficient, and FPS. To decrease the network complexities, we utilized the EfficientNetB0 as an encoder backbone for all the architectures and tried to minimize the complication as far as possible. However, Hard-Net surpassed these architectures and achieved an unassailable lead over them in every index. Further, our augmentation strategies can achieve a significant performance gain, which helps the model generalize on the challenging validation set. The possible limitation of this study is setting the manual probabilities for the different augmentation techniques. Moreover, the current input resolution is 352×352 for the network, which can be increased more without reducing the speed. The results were also uploaded for round I and round II of the competition where we achieved third rank on round II based on the generalisability scores provided by the organisers similar to detection generalisation defined in [24].

4. Conclusion

This paper presented different methods for the accurate segmentation of polyps in GI tract diseases. We employ the EfficientNet model as an encoder backbone for all existing methods and compare it with the recently published HarDNet model. We conclude that the HarDNet took the unassailable lead over other methods in terms of segmentation accuracy and speed. Further, the augmentation strategies applied in the model increase the performance by 2 percent. In the future, we plan to continue researching more efficient tasks.

5. Acknowledgments

This work was supported by the GRRC program of Gyeonggi province. [GRRC-Gachon2020 (B02), AI-based Medical Information Analysis].

References

[1] M. Saha, C. Chakraborty, Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation, IEEE Transactions on Image Processing 27 (2018) 2189–2200.

[2] P. Nardelli, D. Jimenez-Carretero, D. Bermejo-Pelaez, G. R. Washko, F. N. Rahaghi, M. J. Ledesma-Carbayo, R. S. J. Estépar, Pulmonary artery–vein classification in ct images using deep learning, IEEE transactions on medical imaging 37 (2018) 2428–2440.

[3] S. Poudel, Y. Kim, D. Vo, S.-W. Lee, Colorectal disease classification using efficiently scaled dilation in convolutional neural network, IEEE Access PP (2020) 1–1. doi:10.1109/ACCESS.2020.2996770.

[4] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, IEEE transactions on medical imaging 35 (2016) 1285–1298.

[5] J. Zhang, M. Liu, D. Shen, Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks, IEEE Transactions on Image Processing 26 (2017) 4753–4764.

[6] L. Ding, M. H. Bawany, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, G. Sharma, A novel deep learning pipeline for retinal vessel detection in fluorescein angiography, IEEE Transactions on Image Processing (2020).

[7] H. Wang, Z. Li, Y. Li, B. Gupta, C. Choi, Visual saliency guided complex image retrieval, Pattern Recognition Letters 130 (2020) 64–72.

[8] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[9] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[10] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (2017) 2481–2495.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[13] C.-H. Huang, H.-Y. Wu, Y.-L. Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, arXiv preprint arXiv:2101.07172 (2021).

[14] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946 (2019).

[15] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3156–3164.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[18] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3907–3916.

[19] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.

[20] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, M. A. Riegler, P. Halvorsen, C. Daul, J. Rittscher, O. E. Salem, D. Lamarque, T. de Lange, J. E. East, Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv (2021).

[21] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[22] F. I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data, ISPRS Journal of Photogrammetry and Remote Sensing 162 (2020) 94–114.

[23] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 225–2255.

[24] 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) 2748. doi:10.1038/s41598-020-59413-5.