ResUNet++: An Advanced Architecture for Medical Image Segmentation

Debesh Jha*‡, Pia H. Smedsrud*†§, Michael A. Riegler*§, Dag Johansen‡,
Thomas de Lange†§, Pål Halvorsen*¶, Håvard D. Johansen‡

* SimulaMet, Norway   †Augere Medical AS, Norway
‡UiT The Arctic University of Norway, Norway   §University of Oslo, Norway
¶Oslo Metropolitan University, Norway
Email: debesh@simula.no

Abstract—Accurate computer-aided polyp detection and segmentation during colonoscopy examinations can help endoscopists resect abnormal tissue and thereby decrease chances of polyps growing into cancer. Towards developing a fully automated model for pixel-wise polyp segmentation, we propose ResUNet++, which is an improved ResUNet architecture for colonoscopic image segmentation. Our experimental evaluations show that the suggested architecture produces good segmentation results on publicly available datasets. Furthermore, ResUNet++ significantly outperforms U-Net and ResUNet, two key state-of-the-art deep learning architectures, by achieving high evaluation scores with a dice coefficient of 81.33%, and a mean Intersection over Union (mIoU) of 79.27% for the Kvasir-SEG dataset and a dice coefficient of 79.55%, and a mIoU of 79.62% with CVC-612 dataset.

Index Terms—Medical image analysis, semantic segmentation, colonoscopy, polyp segmentation, deep learning, health informatics.

I. INTRODUCTION

Colorectal Cancer (CRC) is one of the leading causes of cancer related deaths worldwide. Polyps are predecessors to this type of cancers and therefore important to discover early by clinicians through colonoscopy examinations. To reduce the occurrence of CRC, it is routine to resect the neoplastic lesions (for example, adenomatous polyps) [1]. Unfortunately, many adenomatous polyps are missed during the endoscopic examinations [2]. A Computer-Aided Detection (CAD) system that, in real-time, can highlight the locations of polyps in the video stream from the endoscope, can act as a second observer, potentially drawing the endoscopist's attention to the polyps displayed on the monitor. This can reduce the chance that some polyps are overlooked [3]. For this purpose, an important improvement of pure anomaly detection approaches, which only identify whether or not there is something abnormal in an image, we also want our CAD system to have pixel-wise segmentation capability so that the specific regions of interest within each abnormal image can be identified.

A key challenge for designing a precise CAD system for polyps is the high costs of collecting and labeling proper medical datasets for training and testing. Polyps come in a wide variety of shapes, sizes, colors, and appearances as shown in Figure 1. For the four main classes of polyps: adenoma,
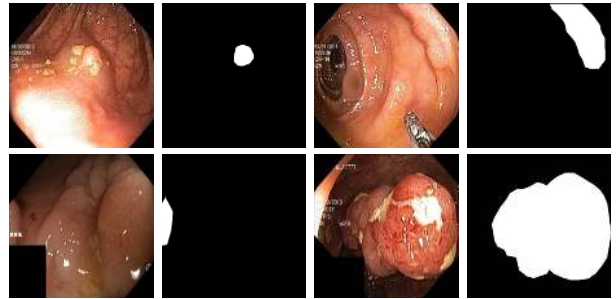

Fig. 1. Examples of polyp images and their corresponding masks from Kvasir-SEG dataset. The first and third column represents the original images, and the second column and fourth column represents their corresponding ground truth.

serrated, hyperplastic, and mixed (rare), there are high inter-class similarity and intra-class variation. There can also be high background object similarity, for instance, where parts of a polyp is covered with stool or when they blend into the background mucosa. Although these factors make our task challenging, we conjecture that there is still a high potential for designing a system with a performance acceptable for clinical use.

Motivated by the recent success of semantic segmentation-based approaches for medical image analysis [4]–[6], we explore how these methods can be used to improve the performance for automatic polyp segmentation and detection. A popular deep learning architecture in the field of semantic segmentation for biomedical application is U-Net [5], which have shown state-of-the-art performance at the 2015 ISBI cell tracking challenge [1]. The ResUNet [6] architecture, is a variant of U-Net architecture that has provided state-of-the-art results for the road image extraction. We therefore adapt this architecture as a basis for our work.

In this paper, we propose the ResUNet++ architecture for medical image segmentation. We have evaluated our model on two publicly available datasets. Our experimental results reveal that the improved model is efficient and achieved a performance boost compared to the popular U-Net [5] and ResUNet [6] architectures.

[1]http://brainiac2.mit.edu/isbi_challenge/.

In summary, the contributions of the paper are as follows:

1) We propose the novel ResUNet++ architecture, which is a semantic segmentation neural network that takes advantage of residual blocks, squeeze and excitation blocks, Atrous Spatial Pyramidal Pooling (ASPP), and attention blocks. ResUNet++ improved the segmentation results significantly for the colorectal polyps compared to other state-of-the-art methods. The proposed architecture works well with a smaller number of images.

2) We annotated the polyp class from the Kvasir dataset [7] with the help of an expert gastroenterologist to create the new Kvasir-SEG dataset [8]. We make this polyp segmentation dataset available to the research community to foster development of new methods and reproducible research.

II. RELATED WORK

Automatic gastrointestinal (GI) tract disease detection and classification in colonoscopic videos has been an active area of research for the past two decades. Polyp detection has in particular been given attention. The performance of the machine learning software has come close to the level of expert endoscopists [9]–[12].

Apart from work on algorithm development, researchers have also investigated complete CAD systems, from data annotation, analysis, and evaluation to visualization for the medical experts [13]–[15]. Thambawita et al. [16] explored various methods, ranging from Machine Learning (ML) to deep Convolutional Neural Network (CNN), and suggested five novel models as a potential solution for classifying GI tract findings into sixteen classes. Guo et al. [17] presented two variants of fully convolutional neural networks, which secured the first position at the 2017 Gastrointestinal Image ANAlysis (GIANA) challenge and second position at the 2018 GIANA challenge.

Long et al. [18] proposed a state-of-the-art semantic segmentation approach for image segmentation known as a Fully Convolutional Network (FCN). FCN are trained end-to-end, pixels-to-pixels, and outputs segmentation result without any additional post-processing steps. Ronneberger et al. [5] modified and extended the FCN architecture to U-Net architecture. There are various modification and extension based on U-Net architecture [4], [6], [17], [19]–[22] to achieve better segmentation results on both natural images and biomedical images.

Most of the published work in the field of polyp detection perform well on the specific datasets, and test scenarios often used small training and validation datasets [11], [23]. The model evaluated on the smaller dataset is neither generalizable nor robust. Moreover, some of the research work only focus on a specific type of polyps. Some of the current work also use non-publicly datasets, which makes it difficult to compare and reproduce results. Therefore, the goal of the ML models to reach a performance level similar to, or better than colonoscopists has not been achieved yet. There exists a potential for improvement in boosting the performance of the system.
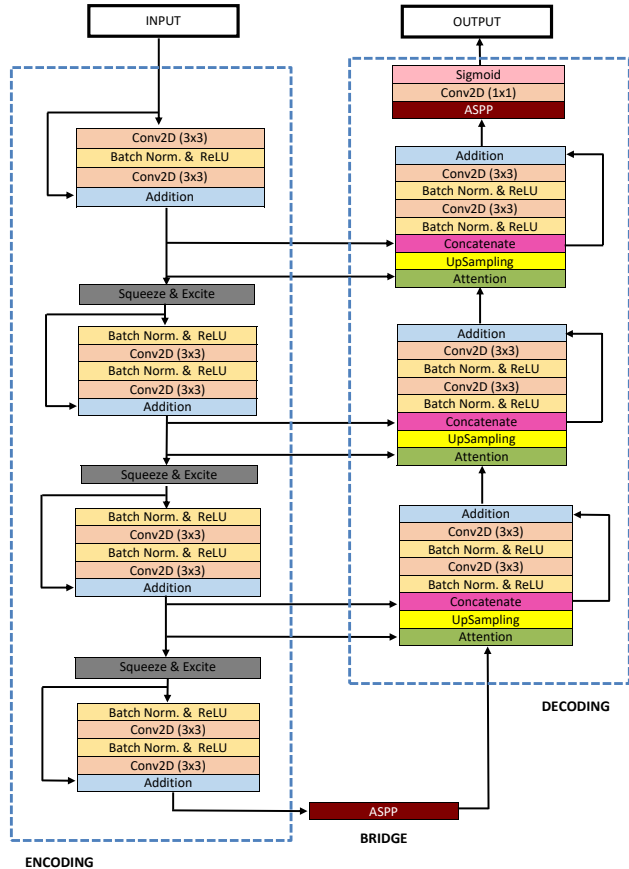

Fig. 2. Block diagram of the proposed ResUNet++ architecture.

III. RESUNET++

The ResUNet++ architecture is based on the Deep Residual U-Net (ResUNet) [6], which is an architecture that uses the strength of deep residual learning [24] and U-Net [5]. The proposed ResUNet++ architecture takes advantage of the residual blocks, the squeeze and excitation block, ASPP, and the attention block.

The residual block propagates information over layers, allowing to build a deeper neural network that could solve the degradation problem in each of the encoders. This improves the channel inter-dependencies, while at the same time reducing the computational cost. The proposed ResUNet++ architecture contains one stem block followed by three encoder blocks, ASPP, and three decoder blocks. The block diagram of the proposed ResUNet++ architecture is shown in Figure 2. In the block diagram, we can see that the residual unit is a combination of batch normalization, Rectified Linear Unit (ReLU) activation, and convolutional layers.

Each encoder block consists of two successive 3×3 convolutional block and an identity mapping. Each convolution block includes a batch normalization layer, a ReLU activation layer, and a convolutional layer. The identity mapping connects the input and output of the encoder block. A strided convolution layer is applied to reduce the spatial dimension

of the feature maps by half at the first convolutional layer of the encoder block. The output of encoder block is passed through the squeeze-and-excitation block. The ASPP acts as a bridge, enlarging the field-of-view of the filters to include a broader context. Correspondingly, the decoding path consists of residual units, too. Before each unit, the attention block increases the effectiveness of feature maps. This is followed by a nearest-neighbor up-sampling of feature maps from the lower level and the concatenation with feature maps from their corresponding encoding path.

The output of the decoder block is passed through ASPP, and finally, we apply a 1×1 convolution with sigmoid activation, that provides the segmentation map. The extension of the ResUNet++ is the squeeze-and-excitation blocks marked in light blue, the ASPP block marked in dark red, and attention block marked in light green. A brief explanation of each of the parts is given in the following subsections.

A. Residual Units

Deeper Neural Networks are comparatively challenging to train. Training a deep neural network with an increasing network depth can improve accuracy. However, it can hamper the training process and cause a degradation problem [6], [24]. He et al. [24] proposed a deep residual learning framework to facilitate the training process and address the problem of degradation. ResUNet [6] uses full pre-activation residual units. The deep residual unit makes the deep network easy to train and the skip connection within the networks helps to propagate information without degradation, improving the design of the neural network by decreasing the parameters along with comparable performance or boost in performance on semantic segmentation task [6], [24]. Because of these advantages, we use ResUNet as the backbone architecture.

B. Squeeze and Excitation Units

The squeeze-and-excitation network [25] boosts the representative power of the network by re-calibrating the features responses employing precise modeling inter-dependencies between the channels. The goal of the squeeze and excite block is to ensure that the network can increase its sensitivity to the relevant features and suppress the unnecessary features. This goal is achieved in two steps. The first one is squeeze (global information embedding), where each channel is squeezed by using global average pooling for generating channel-wise statistics. The second step is excitation (active calibration) that aims to capture the channel-wise dependencies fully [25]. In the proposed architecture, the squeeze and excitation block is stacked together with the residual block to increase effective generalization over different datasets and improve the performance of the network.

C. Atrous Spatial Pyramidal Pooling

The idea of ASPP comes from spatial pyramidal pooling [26] that was successful for re-sampling features at multiple scales. In ASPP, the contextual information are captured at various scales [27], [28] and many parallel atrous convolutions [29] with different rates in the input feature map are fused. Atrous convolution allows controlling the field-of-view for capturing multi-scale information precisely. In the proposed architecture, ASPP acts as a bridge between encoder and decoder in our architecture, as shown in Figure 2. The ASPP model has shown promising results on various segmentation tasks by providing multi-scale information. Therefore, we use ASPP to capture the useful multi-scale information for the semantic segmentation task.

D. Attention Units

The attention mechanism is mostly popular in Natural Language Processing (NLP) [30]. It gives attention to the subset of its input. Moreover, it has been employed in semantic segmentation tasks, for example, [31], for pixel-wise prediction. The attention mechanism determines which parts of the network require more attention in the neural network. The attention mechanism also burdens off the encoder to encode all the information of the polyp image into a vector of a fixed dimension. The main advantage of the attention mechanism is that they are simple, can be applied to any input size, enhance the quality of features that boosts the results.

In the previous two approaches, U-Net [5] and ResUNet [6], there exists a direct concatenation of the encoder feature maps with the decoder feature maps. Inspired by the success of attention mechanism, both in NLP and computer vision tasks, we implemented the attention block in the decoder part of our architecture to be able to focus on the essential areas of the feature maps.

IV. EXPERIMENTS

To evaluate the ResUNet++ architecture, we train, validate, and test models using two publicly available datasets. We compare the performance of our ResUNet++ models with ones trained using U-Net and ResUNet.

A. Datasets

For the task of polyp image segmentation, each pixel in the training images must be labeled as belonging to either the polyp class or the non-polyp class. For the evaluation of ResUNet++, we use the Kvasir-SEG dataset [8], which consists of 1,000 polyp images and their corresponding ground truth masks annotated by expert endoscopists from Oslo University Hospital (Norway). Example images and their corresponding masks from the Kvasir-SEG dataset are shown in Figure 1. The second dataset we have used is the CVC-ClinicDB database [32], which is an open-access dataset of 612 images with a resolution of 384×288 from 31 colonoscopy sequences.

B. Implementation details

All architectures were implemented using the Keras framework [33] with TensorFlow [34] as backend. We performed our experiment on a single Volta 100 GPU on a powerful Nvidia DGX-2 AI system capable of 2-petaFLOPS tensor performance. The system is part of Simula Research Laboratories

heterogeneous cluster and has dual Intel(R) Xeon(R) Platinum 8168 CPU@2.70GHz, 1.5TB of DDR4-2667MHz DRAM, 32TB of NVMe scratch space, and 16 of NVIDIAs latest Volta 100 GPGPUs interconnected using Nvidia's NVlink fully non-blocking crossbars switch capable of 2.4 TB/s of bisectional bandwidth. The system was running Ubuntu 18.04.3LTS OS and had the latest Cuda 10.1.243 installed. We start the training with a batch size of 16, and the proposed architecture is optimized by Adam optimizer. The learning rate of the algorithm is set to $1e-4$. A lower learning rate is preferred, although a lower learning rate slowed down convergence, and a larger learning rate often causes convergence failures.

The size of the image within the same dataset varies. Both the dataset used in the study consists of different resolution images. For efficient GPU utilization and to reduce the training time, we crop the images by putting a crop margin of 320×320 to increase the training dataset. Then, the images are resized to 256×256 pixels before feeding the images to the model. We have used the data augmentation technique such as center crop, random crop, horizontal flip, vertical flip, scale augmentation, random rotation, cutout, and brightness augmentation, etc., to increase the number of training samples. The rotation angle is randomly chosen from 0 to 90°. We have utilized 80% of the dataset for training, 10% for validation, and 10% for the testing. We trained all the models for 120 epochs with a lower learning rate so that a more generalized model can be built. The batch size, epoch, and learning rate were reset depending upon the need. There was an accuracy trade-off if we decrease the batch size; however, we preferred a larger batch size over accuracy because smaller batch size can lead to over-fitting. We also used the Stochastic Gradient Descent with Restart (SGDR) to improve the performance of the model.

V. RESULTS

To show the effectiveness of ResUNet++, we conducted two sets of experiments on Kvasir-SEG and CVC-612 datasets. For the model comparison, we compared the results of the proposed ResUNet++ with the original U-Net and original ResUNet architecture, as both of them are the common preference for the semantic segmentation task. The original implementation of ResUNet, which uses Mean Square Error (MSE) as the loss function, did not produce satisfactory results with Kvasir-SEG and CVC-612 datasets. Therefore, we replaced the MSE loss function with dice coefficient loss and did hyperparameter optimization to improve the results and named the architecture as ResUNet-mod. With this modification, we achieved a performance boost in ResUNet-mod architecture for both the datasets.

A. Results on the Kvasir-SEG dataset

We have tried different sets of hyperparameters (i.e., learning rate, number of epochs, optimizer, batch size, and filter size) for the optimization of ResUNet++ architecture. Hyperparameter tuning is done manually by training the models with different sets of hyperparameters and evaluating their results. The results of ResUNet++, ResUNet-mod, ResUNet [6], and

TABLE I
THE TABLE SHOWS THE EVALUATION RESULTS OF ALL THE MODELS ON KVASIR-SEG DATASET.

Method	Dice	mIoU	Recall	Precision
ResUNet++	**0.8133**	**0.7927**	**0.7064**	0.8774
ResUNet-mod	0.7909	0.4287	0.6909	0.8713
ResUNet	0.5144	0.4364	0.5041	0.7292
U-Net	0.7147	0.4334	0.6306	**0.9222**

U-Net [5] are presented in Table I. Table I shows that the proposed model achieved the highest dice coefficient, mIoU, recall, and competitive precision for the Kvasir-SEG dataset. U-Net achieved the highest precision. However, the dice coefficient and mIoU scores are not competitive, which is an important metric for semantic segmentation task. The proposed architecture has outperformed the baseline architectures by a significant margin in terms of mIoU.

B. Results on the CVC-612 dataset

We have performed additional experiments for in-depth performance analysis for automatic polyp segmentation. Therefore, we attempted for the generalization of the model to check the generalizability capability of the proposed architecture on a different dataset. Generalizability would be a further step toward building clinical acceptable model. Table II shows the results for all the architectures on CVC-612 datasets. The proposed model obtained highest dice coefficient, mIoU, and recall and competitive precision.

Figure 3 shows the qualitative results for all the models. From Table I, Table II, and Figure 3 we demonstrate the superiority of ResUNet++ over the baseline architectures. The quantitative and qualitative result shows that the ResUNet++ model trained on Kvasir-SEG and CVC-612 dataset performs well and outperforms all other models in terms of dice coefficient, mIoU, and recall. Therefore, the ResUNet++ architecture should be considered over these baselines architecture in the medical image segmentation task.

VI. DISCUSSION

The proposed ResUNet++ architecture produces satisfactory results on both Kvasir-SEG and CVC-612 datasets. From Figure 3, it is evident that the segmentation map produced

TABLE II
THE TABLE SHOWS THE EVALUATION RESULTS OF ALL THE MODELS ON CVC-612 DATASET.

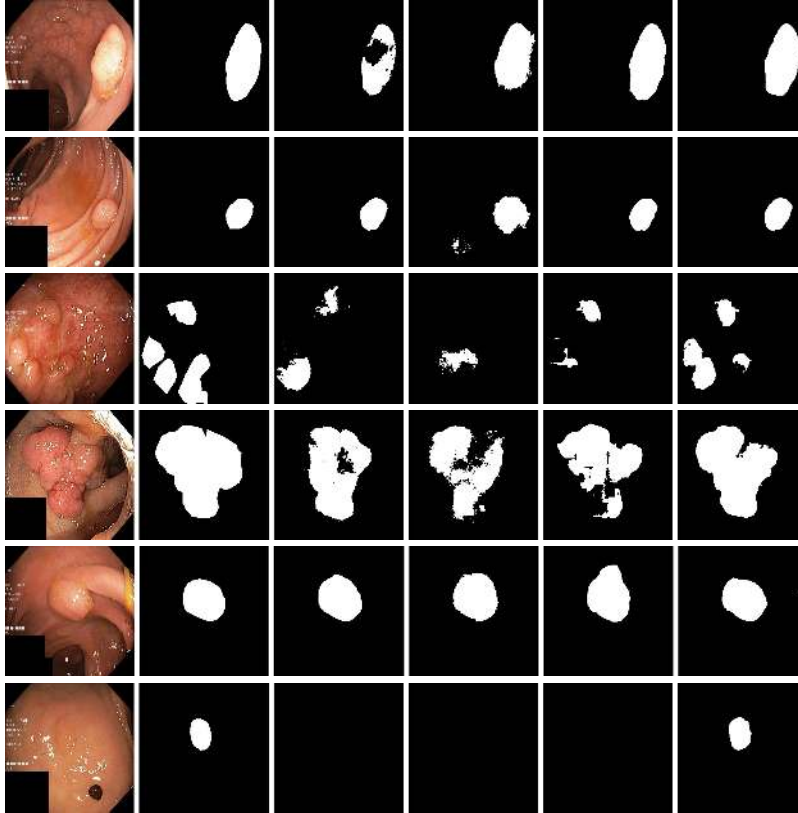| Method	Dice	mIoU	Recall	Precision
ResUNet++	**0.7955**	**0.7962**	**0.7022**	0.8785
ResUNet-mod	0.7788	0.4545	0.6683	**0.8877**
ResUNet	0.4510	0.4570	0.5775	0.5614
U-Net	0.6419	0.4711	0.6756	0.6868

Fig. 3. Qualitative results comparison on the Kvasir-SEG dataset. From the left: image (1), (2) Ground truth, (3) U-Net, (4) ResUNet, (5) ResUNet-mod, and (6) ResUNet++. From the experimental results, we can say that ResUNet++ produces better segmentation masks than other competitors.

by ResUNet++ outperforms other architectures in capturing shape information, in the Kvasir-SEG dataset. It means the generated segmentation mask in ResUNet++ has more similar ground truth then the presented state-of-the-art models. However, ResUNet-mod and U-Net also produced competitive segmentation masks.

We trained the model using different available loss functions, for example, binary cross-entropy, the combination of binary cross-entropy and dice loss, and mean square loss. We observed that the model achieved a higher dice coefficient value with all the loss function. However, mIoU were significantly lower with all other except dice coefficient loss function. We selected the dice coefficient loss function based on our empirical evaluation. Moreover, we also observed that the number of filters, batch size, optimizer, and loss function can influence the result.

We believe that the performance of the model can be further improved by increasing the dataset size, applying more augmentation techniques, and by applying some post-processing steps. Despite increased numbers of parameters with the proposed architecture, we trained the model to achieve higher performance. We conclude that the application of ResUNet++ should not only limited to biomedical image segmentation but could also be expanded to the natural image segmentation and other pixel-wise classification tasks, which

need further detailed validations. We have optimized the code as much as possible based on our knowledge and experience. However, there may exist further optimization, which may also influence the results of the architectures. We have run the code only on a Nvidia-DGX-2 machine, and the images were resized, which may have lead to the loss of some useful information. Additionally, ResUNet++ uses more parameters, which increases training time.

VII. CONCLUSION

In this paper, we presented ResUNet++, which is an architecture to address the need for more accurate segmentation of colorectal polyps found in colonoscopy examinations. The suggested architecture takes advantage of residual units, squeeze and excitation units, ASPP, and attention units. Comprehensive evaluation using different available datasets demonstrates that the proposed ResUNet++ architecture outperforms the state-of-the-art U-Net and ResUNet architectures in terms of producing semantically accurate predictions. Towards achieving the generalizability goal, the proposed architecture can be a strong baseline for further investigation in the direction of developing a clinically useful method. Post-processing techniques can potentially be applied to our model to achieve even better segmentation results.

REFERENCES

[1] A. G. Zauber, S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro, J. F. Panish *et al.*, "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths," *New England Journal of Medicine*, vol. 366, no. 8, pp. 687–696, 2012.

[2] J. C. Van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. Van Deventer, and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *The American journal of gastroenterology*, vol. 101, no. 2, p. 343, 2006.

[3] Y. Mori and S.-e. Kudo, "Detecting colorectal polyps via machine learning," *Nature biomedical engineering*, vol. 2, no. 10, p. 713, 2018.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceeding of International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

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

[6] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[7] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, june 2017, pp. 164–169.

[8] D. Jha, P. H. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020. [Online]. Available: https://datasets.simula.no/kvasir-seg/

[9] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[10] Y. Mori, S.-e. Kudo, T. M. Berzin, M. Misawa, and K. Takeda, "Computer-aided diagnosis for colonoscopy," *Endoscopy*, vol. 49, no. 8, pp. 813–819, 2017.

[11] P. Brandao, O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks," *Journal of Medical Robotics Research*, vol. 3, no. 2, p. 1840002, 2018.

[12] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature biomedical engineering*, vol. 2, no. 10, pp. 741–748, 2018.

[13] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *International Journal of Computer methods and programs in biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[14] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. Halvorsen, and T. D. Lange, "From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, p. 26, 2017.

[15] S. A. Hicks, S. Eskeland, M. Lux, T. de Lange, K. R. Randel, M. Jeppsson, K. Pogorelov, P. Halvorsen, and M. Riegler, "Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain," in *Proceedings of the ACM Multimedia Systems Conference*. ACM, 2018, pp. 369–374.

[16] V. Thambawita, D. Jha, M. Riegler, P. Halvorsen, H. L. Hammer, H. D. Johansen, and D. Johansen, "The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning," in *Working Notes Proceedings of the MediaEval Workshop*. CEUR Workshop Proceedings, 2018.

[17] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS-Science and Technology Publications, 2019, pp. 632–641.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 3431–3440.

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Proceeding of International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[20] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[21] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *arXiv preprint arXiv:1904.00592*, 2019.

[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and 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.

[23] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2013.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 7132–7141.

[26] K. He., X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[31] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[32] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[33] F. Chollet *et al.*, "Keras," 2015.

[34] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceeding of {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI})*, 2016, pp. 265–283.