

DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation

Nikhil Kumar Tomar¹, Debesh Jha^{1,3}, Sharib Ali⁴, Håvard D. Johansen³, Dag Johansen³, Michael A. Riegler¹, and Pål Halvorsen^{1,2}

¹ SimulaMet, Norway

² Oslo Metropolitan University, Norway

³ UIT The Arctic University of Norway

⁴ Department of Engineering Science, University of Oxford, Oxford, UK
debesh@simula.no

Abstract. Colonoscopy is the gold standard for examination and detection of colorectal polyps. Localization and delineation of polyps can play a vital role in treatment (e.g., surgical planning) and prognostic decision making. Polyp segmentation can provide detailed boundary information for clinical analysis. Convolutional neural networks have improved the performance in colonoscopy. However, polyps usually possess various challenges, such as intra-and inter-class variation and noise. While manual labeling for polyp assessment requires time from experts and is prone to human error (e.g., missed lesions), an automated, accurate, and fast segmentation can improve the quality of delineated lesion boundaries and reduce missed rate. The Endotect challenge provides an opportunity to benchmark computer vision methods by training on the publicly available Hyperkvasir and testing on a separate unseen dataset. In this paper, we propose a novel architecture called “DDANet” based on a dual decoder attention network. Our experiments demonstrate that the model trained on the Kvasir-SEG dataset and tested on an unseen dataset achieves a dice coefficient of 0.7874, mIoU of 0.7010, recall of 0.7987, and a precision of 0.8577, demonstrating the generalization ability of our model.

Keywords: Polyp segmentation, Deep Learning, Convolutional neural network, Benchmarking

1 Introduction

Colorectal cancer is one of the leading causes of cancer. Colonoscopy is a standard medical procedure for the surveillance examination and treatment. Regular screening and removal of pre-cancerous lesions through colonoscopy is essential for early cancer detection and prevention. Studies suggest that the miss-rate of adenoma is between 6 to 27% [1].

The automatic segmentation of the suspected areas with lesions in colonoscopy images can play a crucial role, and identifying each colon pixel can significantly impact clinical settings. With the increase of publicly available datasets, dominant methodology such as convolutional neural network, improved hardware,

and collaboration between computational and clinical communities to tackle the problems in endoscopic imaging through computer vision tasks is gaining momentum than ever before. An automatic polyp detection or surveillance system can help to achieve low-cost design solutions and save time of clinicians allowing them to use their time to look into more severe cases.

In this respect, the Endotect challenge [8] offers three tasks, namely, detection of Gastrointestinal (GI) tract images, efficient detection on the same images, and automatic polyp segmentation. The detection and efficient detection task are based on the HyperKvasir dataset [5], and the segmentation is based on the Kvasir-SEG dataset [12]. Out of these three tasks, we participated in the “segmentation task”, where the goal was to generate an automatic segmentation of the polyps for the unseen dataset.

In this paper, we propose a novel deep learning architecture, called Dual Decoder Attention Network (DDANet), for automatic polyp segmentation. It follows an encoder-decoder scheme and incorporates a single encoder that is shared by two parallel decoders, where the first decoder acts as a segmentation network and the second decoder acts as an autoencoder network. The autoencoder network helps to strengthen the feature maps in the encoder network. It is used as an auxiliary task training, which is used to generate an attention map. This attention map is used in each decoder to improve the semantic representation of the feature maps. This, in turn, helps to improve the performance of the entire network. The proposed DDANet is fed with an RGB input image, where it predicts the segmentation mask and the reconstructed grayscale image. The architecture is efficient in terms of Frame per Second (FPS) and also has a decent evaluation score. These metrics are the requirement for the real-world settings toward developing a Computer Aided Diagnosis (CADx) system.

2 Related Work

Automatic polyp segmentation task is a well-defined computer vision problem. Recently, there have been several competitions [4, 3, 2, 10] and individual efforts [7, 13, 11, 6, 9] toward building a CADx system for the polyp segmentation. With these competitions and individual efforts, polyp segmentation is becoming more and more mature. However, comparing models and results of the many individual approaches is difficult due to the use of diverse (often publicly non-available) datasets and different hardware. In this respect, competitions provide an opportunity to benchmark and compare the designed methods with other competitors’ on the same dataset. Moreover, the evaluation metrics are independently calculated by the organizers, including the ranking decision of each team.

The competitions can help us to define the strengths and weaknesses of each method. It also provides us with an opportunity to disseminate methods and discuss the results collectively in the same space. Through this year’s Endotect challenge, we provide a novel solution to develop more efficient algorithms that can be useful to build an automatic polyp segmentation system. Our architecture is composed of an autoencoder branch in addition to the segmentation

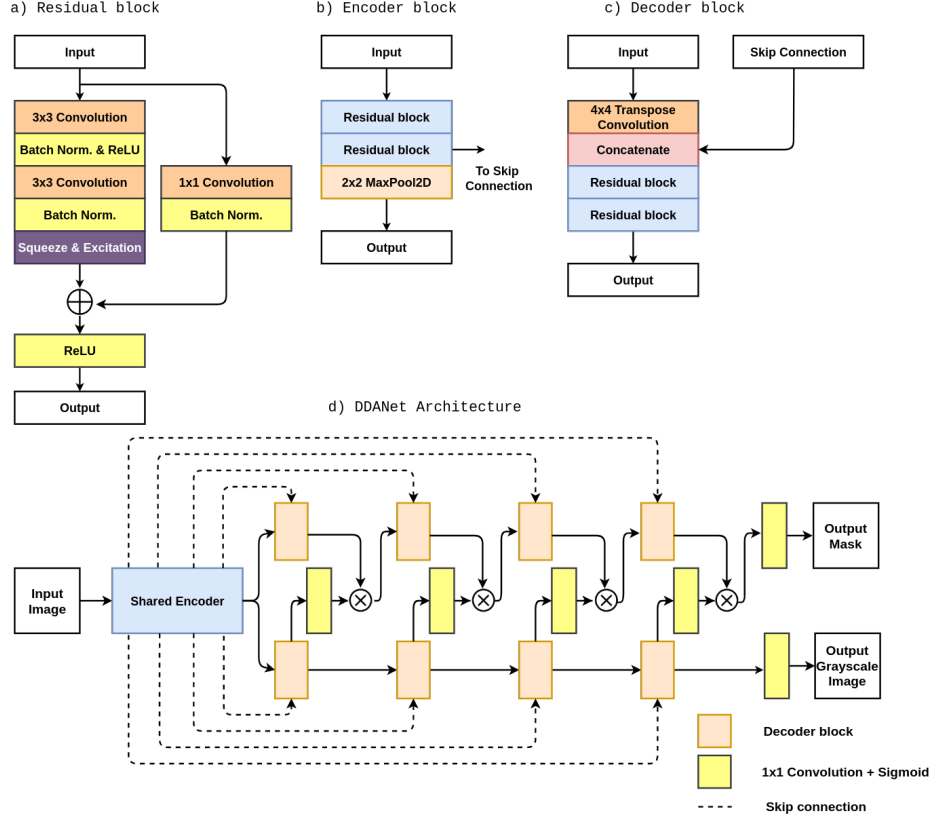


Fig. 1: DDANet architecture and its components.

branch, which is different from other encoder-decoder based network (for example, UNet [14], ResUNet++ [13], DoubleUNet [11]). The benefit of incorporating autoencoder in the network can be seen from the quantitative and qualitative results.

3 DDANet

In this section, we will first describe each component of our DDANet and then detail the overall proposed DDANet architecture.

3.1 Residual block

As the network depth increases, the performance also increases to a certain limit as the gradients can be effectively calculated. However, after a certain depth, the performance of the model may be impacted due to the vanishing or exploding

gradients as the gradients become either zero or too large. By introducing a skip-connection in residual learning, the problem of the vanishing or exploding gradients has been solved. Our residual block (see Figure 1a) consists of two 3×3 convolutions, each followed by a batch normalization and a Rectified linear unit (ReLU) activation function. The residual learning introduces a shortcut connection or identity mapping, which connects the input with the residual block's output. The identity mapping tries to learn an identity function since the input is directly passed to the output. It also helps in a better flow of the gradients during the backpropagation.

3.2 Squeeze and Excitation block

A Convolutional Neural Network (CNN) is used to extract features from an image and then transform the image into a feature map. A problem with CNNs is that they treat every feature channel as equally important. To overcome this problem, we introduce a squeeze and excitation layer, which acts as a channel-wise attention mechanism. It re-weights every feature channel accordingly to create a more accurate feature map. In this way, the overall network becomes more sensitive towards essential features that improve the network performance significantly. The squeeze and excitation network mainly consists of two steps. In the first step, the feature maps are compressed using the global average pooling function to generate a compressed representation for the feature maps. While, in the second step, a 2-layered neural network is used, where features are first reduced and then expanded. This generates a feature vector, which is used to scale the feature channels.

3.3 The DDANet architecture

The proposed architecture named DDANet follows an encoder-decoder design similar to ResUNet++ [13]. The DDANet combines the strength of the residual learning and the squeeze and excitation network. The proposed DDANet is a fully convolutional network that consists of a single encoder shared by dual decoders. The encoder network consists of a 4 encoder block, whereas each decoder network also consists of 4 decoder block (see Figure 1d).

The RGB input image is first fed into the encoder network (see Figure 1b), which encodes it into an abstract feature representation while gradually down-sampling it. The output of the encoder network is fed to both decoders (see Figure 1c), where it is followed by a 4×4 transpose convolution that doubles its spatial dimensions. After that, the image is concatenated with an appropriate feature maps from the encoder network using the skip connection. These skip connections fetch the features from earlier layers at their original resolution, which increases their feature representation strength. The skip connections also act as an alternative path for the gradient flow and are often beneficial for model convergence.

Two residual blocks are then used to learn the necessary feature required by the network during back-propagation. The output of the second decoder

block (autoencoder branch) follows a 1×1 convolution and a sigmoid activation function to generate an attention map. This attention map is multiplied by the output of the first decoder block (segmentation branch), which acts as an input for the next decoder block in the segmentation branch. The final decoder block’s output is passed through a 1×1 convolution and a sigmoid activation function, where the first decoder outputs a segmentation mask, and the second outputs the reconstructed grayscale image.

4 Experimental Setup

In this section, we present the implementation details and datasets used in this work.

4.1 Implementation Details

The proposed DDANet architecture is implemented in the PyTorch 1.6 framework¹. For training the DDANet, we used an NVIDIA DGX-2 machine that uses an Nvidia V100 Tensor Core GPUs. During training, we have used an input image resolution of 512×512 . We use a combination of binary cross-entropy and dice loss for calculating the loss between the predicted masks and the ground-truth masks. We have used binary cross-entropy in the case of predicting the grayscale image. An Adam optimizer was used with a learning rate of $1e^{-4}$. The models were trained for 200 epochs.

4.2 Datasets

The Kvasir-SEG [12] dataset was used for training the model. We have used 88% of the dataset for training and the remaining 12% images for development-test-set. Kvasir-SEG consists of 1000 polyp images, ground truth segmentation masks, and bounding boxes. A separate test dataset with 200 images was provided for prediction. However, the ground truth for this dataset was not provided by the organizers. The exact number of images used for the training and testing can also be found in our GitHub repository. More details about the dataset and the baseline results on it can be found in [12].

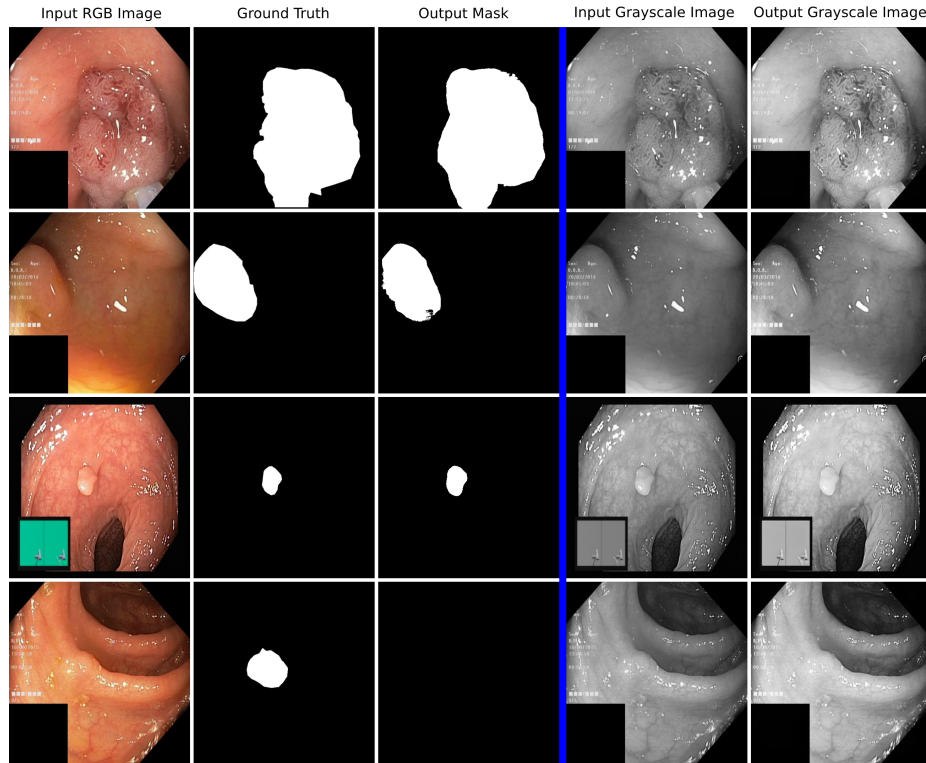
5 Results

Table 1 shows the results of the DDANet trained and validated on Kvasir-SEG. Additionally, evaluation scores on the test dataset can also be found here. The evaluation metrics for the challenge was Dice Coefficient (DSC). However, we have also calculated other commonly used metrics such as mean Intersection over Union (mIoU), recall, precision, and FPS. The DDANet obtained a DSC of 0.8576, a mIoU of 0.7800, a recall of 0.8880, and a precision of 0.8643. All the metrics suggest that our method performs quite well on the Kvasir-SEG dataset. When we compare the results with our previous results [13, 12], where

¹<https://github.com/nikhilroxtomar/DDANet>

Table 1: Quantitative results on Kvasir-SEG and unseen (Challenge) dataset.

Dataset	Method	DSC	mIOU	Recall	Precision	FPS
Kvasir-SEG	DDANet	0.8576	0.7800	0.8880	0.8643	69.59
Unseen (Challenge)	DDANet	0.7874	0.7010	0.7987	0.8577	70.23

**Fig. 2:** Qualitative results of the DDANet on the Kvasir-SEG test dataset. The blue line divides the segmentation and the reconstruction part. Columns 4 and 5 show the reconstruction part that was used in the DDANet as an auxiliary task.

the DSC values were 0.8133, and 0.7877, DDANet achieves a higher DSC of 0.8576. However, we can not compare directly with this work with our previous works as a different train-test split of the dataset is used.

Figure 2 shows the qualitative results of the DDANet on Kvasir-SEG. The figure shows that the proposed DDANet is able to segment both larger and smaller polyps. However, the figure also shows the challenges in identifying the flat polyps, which is one of the open issues in the field of development of CADx systems for colonoscopy. From the quantitative results on development and un-

seen test dataset, we can say that the proposed method is comprehensive in producing reliable segmentation output.

6 Discussion

The qualitative results (see Figure 2) show that the proposed model was able to segment polyps ranging from large to small (Figure 2), but still, challenges remain within some polyps (for example, flat or sessile). We can also see a nearly perfect reconstruction of the grayscale image. In the future, we would like to use image super-resolution instead of just a grayscale image reconstruction.

From all the results, we can see that our method achieves high precision and recall evaluation scores on both the Kvasir-SEG validation dataset and on the unseen test dataset (see Table 1). Additionally, we also achieved a DSC of 0.7874 on the unseen dataset. Thus, high DSC, recall, and precision results validate our proposed method. Moreover, our approach is quite fast with an average FPS of 70.23. Thus, the results show that our method can identify polyps in real-time.

7 Conclusion

The Endotect challenge [8] aims to benchmark various computer-vision approaches on the HyperKvasir dataset containing GI images and videos. Here, we have proposed the DDANet architecture for automatic polyp segmentation, and the proposed architecture provides good results in the segmentation task. We have obtained a high precision, recall, DSC, mIoU, and FPS. However, there are large rooms for improvements. We intend to further improve the architecture by applying post-processing and analyzing the optimal parameters in the future.

Acknowledgements

This work is funded in part by the Research Council of Norway, project number 263248 (Privaton) and project number 282315 (AutoCap). We performed all computations in this paper on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (*eX³*), which is financially supported by the Research Council of Norway under contract 270053.

References

- [1] Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. *Gut and liver* 6(1), 64 (2012)
- [2] Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al.: A translational pathway of deep learning methods in gastrointestinal endoscopy. *arXiv preprint arXiv:2010.06034* (2020)
- [3] Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific reports* 10(1), 1–15 (2020)

- [4] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al.: Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge. *IEEE transactions on medical imaging* 36(6), 1231–1249 (2017)
- [5] Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: HyperKvasir, a Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy. *Springer Nature Scientific Data* 7, Article no. 283 (2020)
- [6] Guo, Y.B., Matuszewski, B.: Giana polyp segmentation with fully convolutional dilation neural networks. In: *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. pp. 632–641 (2019)
- [7] Guo, Y., Bernal, J., J Matuszewski, B.: Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. *Journal of Imaging* 6(7), 69 (2020)
- [8] Hicks, S.A., Jha, D., Thambawita, V., Halvorsen, P., Hammer, H., Riegler, M.A.: EndoTect: A Competition on Automatic Disease Detection in the Gastrointestinal Tract. In: *ICPR 2020 Workshops and Challenges*. LNCS, Springer (2020)
- [9] Jha, D., Ali, S., Johansen, H.D., Johansen, D., Rittscher, J., Riegler, M.A., Halvorsen, P.: Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. *arXiv preprint arXiv:2006.11392* (2020)
- [10] Jha, D., Hicks, S.A., Emanuelsen, K., Johansen, H., Johansen, D., de Lange, T., Riegler, M., Halvorsen, P.: Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation. In: *Proceedings of MediaEval CEUR Workshop* (2020)
- [11] Jha, D., Riegler, M., Johansen, D., Halvorsen, P., Johansen, H.: DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In: *Proceedings of the International Symposium on Computer Based Medical Systems (CBMS)* (2020)
- [12] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A Segmented Polyp Dataset. In: *Proceedings of the International Conference on Multimedia Modeling (MMM)*. pp. 451–462 (2020)
- [13] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: *Proceedings of the International Symposium on Multimedia (ISM)*. pp. 225–230 (2019)
- [14] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *Proceedings of International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241 (2015)