Improving Generalizability in Polyp Segmentation using Ensemble Convolutional Neural Network

Nikhil Kumar Tomar[a], Nabil Ibtehaz[c], Debesh Jha[a,b], Pål Halvorsen[a] and Sharib Ali[d]

[a]*SimulaMet, Oslo, Norway*

[b]*UiT The Arctic University of Norway, Tromsø, Norway*

[c]*Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*

[d]*Department of Engineering Science, Big Data Institute, University of Oxford, Oxford, UK*

Abstract

Polyp segmentation is crucial for the diagnosis of colorectal cancer. Early detection and removal of polyps can prolong the life of patients and reduce the mortality rate. Despite near expert-label performance with applying the deep learning method in polyp segmentation tasks, the generalization of such models in the clinical environment remains a significant challenge. Transfer learning from a large medical dataset from the same domain is a common technique to address generalizability. However, it is difficult to find a similar large medical dataset. In this work, we investigate the feasibility of building a generalizable model for polyp segmentation using an ensemble of four MultiResUNet architectures, each trained on the combination of the different centered datasets provided by the challenge organizers. Our method achieved a decent performance of 0.6172 ± 0.0778 for the multi-centered dataset. Our findings show that significant work needs to be done to design a robust segmentation model for the development of a clinically acceptable system.

Keywords

Polyp segmentation, colonoscopy, generalization, deep learning

1. Introduction

The medical world concerned with the digestive system is currently in the midst of an uprising wave of increased adaption and technology usage for automatic analysis and decision support. With the increase of publicly available datasets, adapted methodologies such as convolutional neural networks, improved hardware, and increased collaboration of computer scientists and medical communities, this development is gaining more momentum than ever before. Global Cancer Statistics 2020 (GLOBOCAN 2020) estimated colorectal cancer as the third most frequently diagnosed cancer. Colorectal cancer accounts for 10.0% of total cancer, which is only 1% below to the most frequently caused cancer, i.e., female breast cancer (11.7%) and lung cancer (11.4%) [1]. Screening and removal of adenomatous polyps and other precancerous anomalies is one of the best working methods for the early detection and avoiding colorectal cancer-based mortality and incidence [2].

Deep learning-based methods have gained popularity in the development of the computer-aided diagnosis (CADx) system for detection of the colorectal polyps [3, 4, 5]. The successful deployment of a CAD system for polyp segmentation would require a trained model that achieves

high performance on unseen datasets irrespective of different hospitals, cohort populations, and imaging protocol. However, deep learning algorithms are data-driven. The desired generalizable algorithms would require large, high-quality, and diverse datasets samples to train algorithms. Creating such datasets requires expert endoscopists and computer scientists for labeling and pixel-wise annotations. In general, there are only a few publicly available datasets. Although some studies report high performance on a specific dataset, the dataset is not publicly released [5, 3]. Therefore, it is challenging to develop a generalizable polyp segmentation model with a limited or single-center dataset.

Challenges and competitions are a good technique to access and explore new datasets for experimentation. It is also a fair way to compare methods, analyze, and improve the results on provided dataset. Additionally, challenges provide a solution for the lack of dataset availability and help develop reliable and clinically applicable methods. We participated in the EndoCV2021 challenge[1] to explore a multi-center dataset and develop a generalizable polyp segmentation CADx system. Our goal is to investigate and develop a generalizable model, compare our results with the other participants in the challenge, and observe our model's behavior.

The EndoCV2021 challenge offered two different tasks, namely, Detection generalization challenge and Segmentation generalization challenge. We only participated in the **Segmentation generalization challenge**. We used an ensemble model as our solution for the segmentation generalization challenge. The main motivation behind using ensemble methods was that it showed winning results in the different challenges [6, 7]. For our solution, we made an ensemble of four MultiResUNet [8] model. In short, the main contribution of our work are as follows:

- We explore a convolutional neural network-based model for the generalizable polyp segmentation task with a multi-center dataset. In this study, the training dataset is collected from five different medical institutions from five different countries, and test data comes from independent institutions.
- Our work reveals that the proposed deep learning model has significant challenges with the images having bleeding, adenomas, and covered by dyed. The model mostly showed over-segmentation or failed miserably with such scenarios. We highlight these cases that are among the significant challenges for developing a generalizable algorithm for the polyp segmentation task.

The remainder of this paper is organized into five sections. Section 2 provides a short overview of the related work. Section 3 gives an overview Methodology, and Section 4 describes the experimental setup. Section 5 presents the results obtained using the challenge dataset. Finally, we summarize and conclude the paper in Section 6.

2. Related Work

CNN-based architectures for polyp segmentation have been a common strategy for the development of the CADx system. We briefly describe the work on polyp segmentation and generalizability in the below subsection.

---

[1]https://endocv2021.grand-challenge.org/

2.1. Polyp segmentation

There has been several study on colorectal polyp segmentation [5, 3, 8, 6, 9]. Most of the work have proposed an architecture based on U-Net [10]. There have been also work on improving the segmentation performance on the publicly available dataset [11, 4, 12] to the real-time performance [13, 14, 3]. Although mostly retrospective studies were conducted [5, 13], there has also been work that carried prospective randomized controlled studies [15, 16]. However, most of the studies were conducted on the dataset from a single center. The experiments on multi-center datasets have often been ignored.

2.2. Generalizibility

In medical image analysis, generalization refers to the ability of the machine learning algorithm that is trained on specific interventions in specific health centers should be able to perform well over other interventions or different health center [7]. Poor generalizability has become one of the major issues for the clinical translation of the deep learning methods into clinical practise [17]. Meta-learning under a few-shot setting has gained popularity in developing a generalizable deep learning model and resolve the issue of data scarcity [18, 19].

In our previous study [4], due to the lack of a publicly available multi-center dataset, we have used a trained dataset on one publicly available dataset [20] and tested it against another [21] to observe the generalization capability. Additionally, we have also mixed the datasets from two or more institutions to observe the model's generalization capability. This is our first work where we have the opportunity to train the model with a multi-center dataset (five different center datasets) and benchmark on the completely new dataset.

3. Methodology

To address the generalizability problem in polyp segmentation, we used an ensemble of the four MultiResUNet [8] models. As each folder of the dataset has images from a unique center, we use a different subset of the dataset to train each of the MultiResUNet models. The MultiResUNet is an encoder-decoder architecture, which is an improvement over the existing U-Net [10] architecture. It combines the strength of the U-Net and improving it by replacing the existing components with more effective components such as "MultiRes block" and "Res Path". The MultiResUNet consists of four encoder blocks, four decoder blocks, and a bridge connecting them. The encoder takes the input image, encodes it, and extracts more useful features from it. Later these features are passed to the decoder, where they are upsampled and concatenated with the feature maps from the skip connection. Finally, these features are used to generate a segmentation mask for the input image. The additional block to form MultiResUNet models is briefly described below.

3.1. MultiRes block

The MultiRes block is the major component used in the MultiResUNet [8] architecture. It is the replacement of the convolution block, i.e., two 3×3 convolution used in the U-Net.

The MultiRes block is inspired from the Inception architecture [22] which consists of multiple parallel convolutions with 3×3, 5×5, and 7×7 kernel size. These multiple parallel convolutions help in capturing objects with different shapes and sizes. Using the bigger 5×5 and 7×7 kernel size increases the memory requirement. Therefore, these bigger kernels are factorized and replaced by multiple 3×3 convolutions. The MultiRes block begins with a single 3×3 convolution, which is followed by two 3×3 convolutions which are combined together to get the resultant effect of a 5×5 convolution. Next again are the multiple 3×3 convolutions which are repeated to give the resultant effect of a 7×7 convolution. The outputs from these convolutional blocks are concatenated together to have different scale feature maps. A residual connection is also used, which connects the input to the concatenated output.

3.2. Res path

The introduction of the skip connection in the U-Net architecture proves to be a significant contribution towards improving semantic segmentation performance. These skip connections enable the flow of information from the encoder to the decoder that is lost during the pooling operation. The simple concatenation of the features from the encoders to the decoders is flawed. For example, the first skip connection contains the low-level features from the early layers, which are fused with high-level features in the decoder. Therefore, there is a semantic gap between the features that being merged. To resolve this semantic gap, some convolutional layers and shortcut connections are being introduced as the skip connection in the MultiResUNet, called the "Res path".

3.3. MultiResUNet Architecture

The MultiResUNet [8] architecture begins by feeding the input image to the first encoder, which consists of the MultiRes block, followed by a 2×2 max-pooling with a stride value of 2. The max-pooled feature maps are passed on to the next encoder, and this process is repeated four times. In each step, the number of filters doubles, and the spatial resolution reduces by half. The output of the MultiRes block acts as the skip-connection, which first passes through the Res path and joins the decoder block. Inside each Res path, the number of convolution blocks decreases from 4, 3, 2 to 1 respectively along the four Res paths. The decoder begins with a 2×2 transpose convolution, which doubles the feature maps' spatial dimensions. Next, the feature maps are concatenated with the output of the Res path. Subsequently, the MultiRes block is used to learn the semantic representation. Similarly, the network is followed by three more decoder blocks, where the number of filters decreases and the feature maps resolution increases. It is then followed by a 1×1 convolution with sigmoid activation to generate the binary segmentation mask.

4. Experiment

To evaluate the performance of the ensemble method, we have performed extensive experiments. This section describes the dataset, evaluation metrics, training strategy, and implementation details used in our experimentation. Figure 1 shows the block diagram of the proposed ensemble
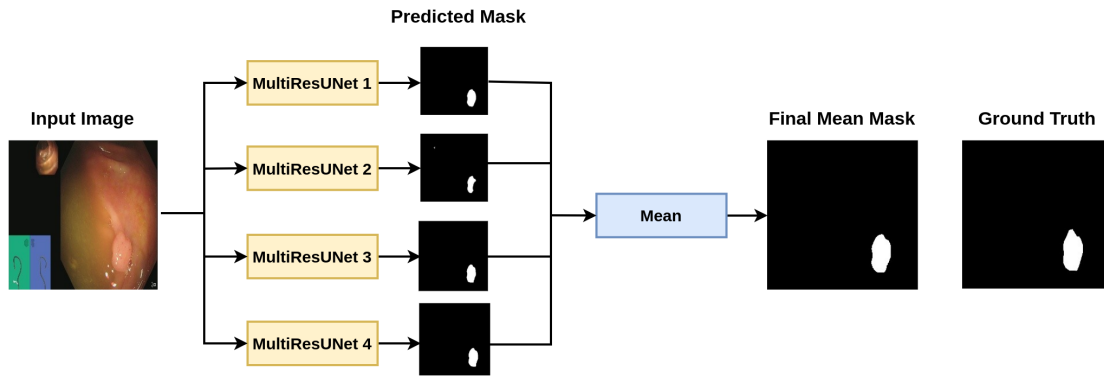
Figure 1: Block diagram of the proposed ensemble architecture

method. As explained in Section 3, the input image is fed to the different MultiResUNet models that produce different segmentation outputs. These predicted outputs from four distinct models are averaged to get the final mean mask.

4.1. Dataset

EndoCV2021 dataset [23] consists of both a single frame dataset and sequence dataset. The dataset is captured from five different institutes. Each center dataset is provided in a separate folder. The training dataset consists of 1452 single image frames. Additionally, the dataset also consists of 165 negative sequence frames and 490 positive sequence frames, in a total of 655 image sequences. The sequence frames are taken from videos. Both positive (polyp) and negative (normal) frames are provided. Each center dataset has a separate image, mask, image with the bounding box, and bounding box information. All the images and their corresponding masks are in jpeg format.

4.2. Evaluation Metrics

The evaluation metric for the detection task is the Average mean precision. Additionally, a mean deviation is also calculated. For the segmentation tasks, the evaluation metrics such as F1-score, mean Intersection over Union (mIoU), recall, precision, F2-score, and overall accuracy is calculated. The procedures for the calculation can be found at GitHub [2] and further details on generalisation metrics is provided in [24]. Out-of-sample distribution from multiple centers were compared among each other to assess the deviation in scores and provide a quantifiable generalisation score [24].

4.3. Training strategy

For training, the model1, i.e., MultiResUNet1, the subset from center1, center3, and center4 were used. Similarly, we used center2, center1, and center4 for training model2 (MultiResUNet2).

---

[2]https://github.com/sharibox/EndoCV2021-polyp_det_seg_gen
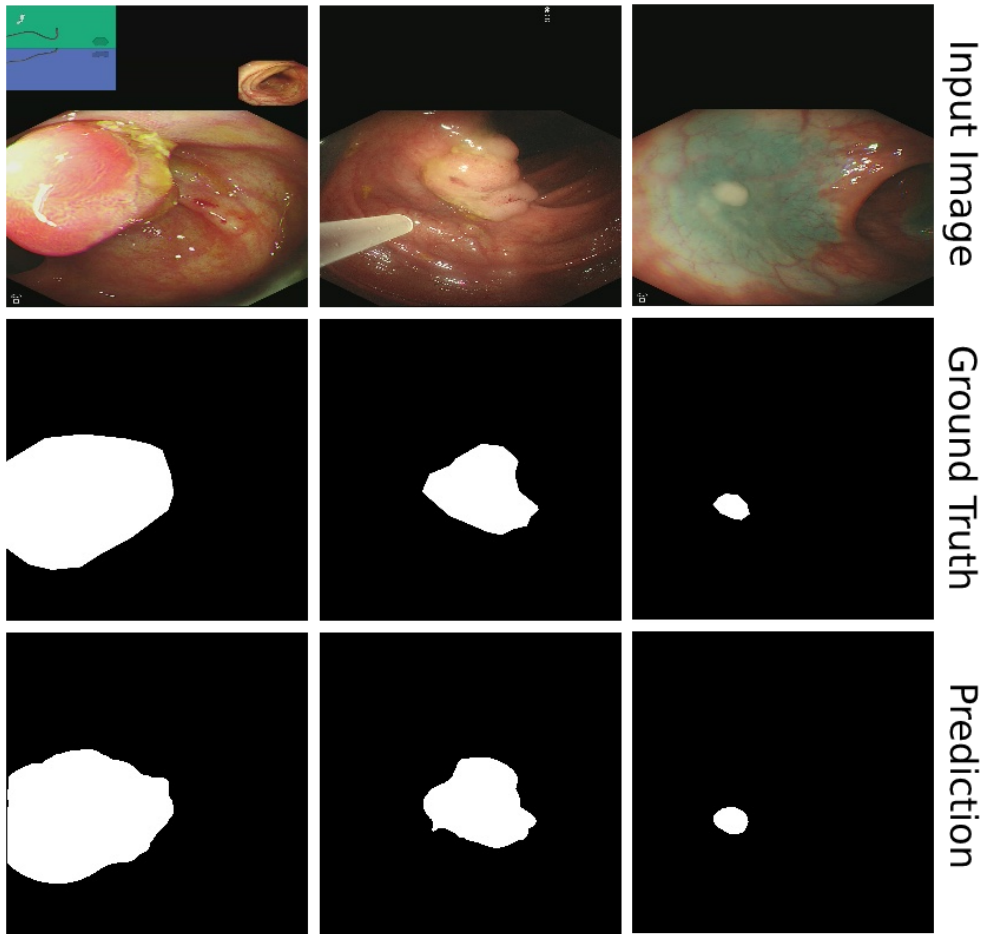
Figure 2: Qualitative results of the four ensemble MultiResUNet[8] models. The example images show that the ensemble models produce high-quality segmentation maps for different polyp shapes and sizes.

Likewise, we used center2, center3, and center1 for training model3. For training model4, we used the images from center2, center3, and center 4. We use the dataset from center5 as the validation set.

4.4. Implementation Details

We have implemented the MultiResUNet using the Keras with TensorFlow as a backend. The experiments were run on the Experimental Infrastructure for Exploration of Exascale Computing(eX3), NVIDIA DGX-2 machine. All four models are trained on 100 epochs using the same set of hyperparameters. Each model uses an image size of 256×256 pixels with a batch size of 8. The dice coefficient is used as the loss function with Adam optimizer. The default learning $1e-3$ is used to training the model. We also use the ReduceLROnPlateau callback to reduce further the learning rate for better generalization of the model.

5. Results and Discussion

On the test dataset, we achieved a score of 0.6172 ± 0.0778. Here, 0.6172 is the generalization score and 0.0778 is the generalization deviation. Figure 2 shows the qualitative results of the ensemble MultiResUNet model. The first, second and third column shows the input image, their corresponding ground truth, and the predictions. From the qualitative results, we can see that the model is performing well on polyp of different shapes and sizes (i.e., small, medium, and large-sized polyps).

However, a detailed dissection of the validation results shows that the models produce over-segmentation for the outputs when the input images have bleeding. The model also fails on challenging images such as flat polyps. The model also has a problem with detecting when the input images are covered with dyed. Mostly the models show over-segmentation, and sometimes the model completely fails to produce any segmentation masks. However, a more detailed conclusion can be made when we can visualize the qualitative results on the test dataset.

6. Conclusion

In this paper, we presented a cascaded MultiResUNet based solution for addressing the generalizability in polyp segmentation. The model can automatically segment polyp. The experimental results showed that the ensemble model obtained an evaluation score of 0.6172 ± 0.0778. The research results open a wide range of research directions to build generalizable model on new datasets. Moreover, we showed that ensemble models are not always the best choice for biomedical data science challenges. A deep analysis of the qualitative results showed that the model performs well on polyps of different shapes and sizes. In the future, we plan to explore the transfer learning from both large natural datasets and from biomedical imaging datasets (polyp or similar domain datasets) for improving the results on the polyp segmentation tasks.

Acknowledgment

References

[1] H. Sung, et al., Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians (2021).

[2] A. M. Wolf, E. T. Fontham, T. R. Church, C. R. Flowers, C. E. Guerra, S. J. LaMonte, R. Etzioni, M. T. McKenna, K. C. Oeffinger, Y.-C. T. Shih, et al., Colorectal cancer screening for average-

risk adults: 2018 guideline update from the american cancer society, CA: a cancer journal for clinicians 68 (2018) 250–281.

[3] J. Y. Lee, et al., Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets, Scientific reports 10 (2020) 1–9.

[4] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. Johansen, P. Halvorsen, M. Riegler, A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation, IEEE journal of biomedical and health informatics (2021).

[5] P. Wang, et al., Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy, Nature biomedical engineering 2 (2018) 741–748.

[6] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis (2021) 102002.

[7] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran, et al., Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the robust-mis 2019 challenge, Medical Image Analysis 70 (2021) 101920.

[8] N. Ibtehaz, M. S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Networks 121 (2020) 74–87.

[9] Y. Guo, J. Bernal, B. J Matuszewski, Polyp segmentation with fully convolutional deep neural networks—extended evaluation study, Journal of Imaging 6 (2020) 69.

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

[11] 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: Proc. of International Symposium on Multimedia (ISM), 2019, pp. 225–2255.

[12] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: Proc. of International Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 558–564.

[13] N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, P. Halvorsen, Ddanet: Dual decoder attention network for automatic polyp segmentation, in: Proc. of International Conference on Pattern Recognition (ICPR) Workshop, 2020.

[14] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, IEEE Access 9 (2021) 40496–40510.

[15] P. Wang, et al., Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study, Gut 68 (2019) 1813–1819.

[16] J.-R. Su, et al., Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos), Gastrointestinal endoscopy 91 (2020) 415–424.

[17] K. Yasaka, O. Abe, Deep learning and artificial intelligence in radiology: Current applications and future directions, PLoS medicine 15 (2018) e1002707.

[18] P. Zhang, J. Li, Y. Wang, J. Pan, Domain adaptation for medical image segmentation: A meta-learning method, Journal of Imaging 7 (2021) 31.

[19] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning (2016).

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: Proc. of International Conference on Multimedia Modeling (ISM), 2020, pp. 451–462.

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

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. of IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 1–9.

[23] 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).

[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.