COUNTERFACTUAL EXPLAINABLE GASTROINTESTINAL AND COLONOSCOPY IMAGE SEGMENTATION

Divij Singh[1]*, Ayush Somani*[2]*, Alexander Horsch*[2]*, Dilip K. Prasad*[2]

[1] Indian Institute of Technology (BHU) Varanasi, India
[2] Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

ABSTRACT

Segmenting medical images accurately and reliably is crucial for disease diagnosis and treatment. Due to the wide assortment of objects' sizes, shapes, and scanning modalities, it has become more challenging. Many convolutional neural networks (CNN) have recently been designed for segmentation tasks and achieved great success. This paper presents an optimized deep learning solution using DeepLabv3+ with ResNet-101 as its backbone. The proposed approach allows capturing variabilities of diverse objects. It provides improved and reliable quantitative and qualitative results in comparison to other state-of-the-art (SOTA) methods on two publicly available gastrointestinal and colonoscopy datasets. Few studies show the inadequacy of stable performance in varying object segmentation tasks, notwithstanding the sizes of objects. Our method has stable performance in the segmentation of large and small medical objects. The explainability of our robust model with benchmarking on SOTA approaches for both datasets will be fruitful for further research on biomedical image segmentation.

Index Terms— Augmentation, deep learning, explainability, image segmentaion, medical imaging.

1. INTRODUCTION

Medical image segmentation and identification is an essential task in clinical diagnosis. The semantic segmentation results can help identify regions of interest, such as polyps and instruments. Polyps identified in the colon can help examine potential cancerous cells, while identified instruments can segregate the remaining area of interest for detection. Thus, the segmentation can help detect missed lesions, prevent diseases, and improve therapy planning and medication.

The significant challenge in medical imaging is the requirement of large high-quality annotated and labeled datasets, which is critical in achieving the desired algorithmic goal of automated medical image segmentation. Manual annotation of biomedical datasets consists of very comprehensive guidelines and protocols defined by experts in the field. Annotating sizeable datasets is a time-consuming and expensive process that requires efforts from diverse, skilled medical experts. In addition, several imaging modalities frequently lack standard annotation protocols. This situation often confuses the experts when identifying a particular area in the lesion as cancerous or non-cancerous. Additionally, the low image quality also sometimes influences the quality of the annotation. Hence, an automated computer-aided segmentation can provide a more accurate, faster, and reliable solution to transform clinical procedures. It shall reduce human error, expert's workload, improve patient care, and reduce the overall treatment cost. Convolutional neural networks (CNNs) have achieved state-of-the-art (SOTA) performance for automatic medical image segmentation. However, they have not demonstrated sufficiently accurate and robust results for clinical use [1]. Furthermore, they are limited due to a lack of image-specific adaptation.

We propose an effective deep learning model for robust segmentation to withstand the diversity of the available biomedical datasets. The approach overcomes the inadequacy of small object detection with good mean Intersection over Union (mIoU) and dice scores for many methods. Early detection of small objects is crucial to reduce mortality, demonstrated with validation on colonoscopy medical imaging.

2. BACKGROUND AND MOTIVATION

A fully convolutional network (FCN), which included only convolutional layers for semantic segmentation, was first proposed in 2014 [2]. Subsequently, for segmentation of HeLa cells and neuronal structures of electron microscopic stacks, Ronneberger [3] modified the FCN with an encoder-decoder U-Net architecture. The low- and high-level feature maps are combined through skip connections in the U-Net [3] architecture. The low-level features are propagated from the initial layers of the network, whereas deeper layers of the encoder process the high-level feature maps before passing through the decoder. Hence, a semantic gap between the high- and low-level features [4, 5] is created, leading to multiple other proposed extensions of the U-Net. Chen et al. [6] proposed the atrous spatial pyramid pooling (ASPP) to aggregate the global features used in the DeepLabv3+ architecture that employs skip connections between the encoder and decoder.

Pranet [7], Caranet [8], UACANet [9], NanoNet [10],

HarDNet-MSEG [11], and MSRF-Net [12] are namely few good performing recent architectures for medical image segmentation. In our work, we prioritized explainability and focused on optimizing already developed models efficiently. We adopted DeepLabv3+ with ResNet-101 [6] as its backbone to produce appealing results for Kvasir-seg [13] (segmented polyp dataset) and Kvasir-instrument [14] (segmented instrument dataset) as opposed to other models developed so far on all the metrics.

To summarize, the paper makes the following contribution: a) Optimization of a single robust deep learning model to outperform all SOTA models in both polyp and instrument segmentation tasks; b) Implementation of heuristic augmentation pipeline to adapt to low illumination and intensity variation in the colonoscopy images; and c) Explainability of the model's performance using counterfactuals and qualitative comparison with other models.

3. MATERIALS AND METHODS

3.1. Dataset and Evaluation Metrics

Two publicly open biomedical datasets on gastrointestinal and colonoscopy imaging have been used to validate the effectiveness of our model. Table 1 summarizes the availability of the training dataset with ground truth mask and testing samples for the experiment. The accuracy evaluation on this dataset is the task from MedAI Challenge 2021 [15].

Dataset	*Images*	*Input size*	*Train*	*Test*
Kvasir-seg (Polyp) [13]	1000	variable	800	200
Kvasir-instrument [14]	590	variable	500	90

Table 1. Colonoscopy datasets used in the experiment.

Standard computer vision metrics for medical image segmentation such as dice coefficient (DSC), mean Intersection over Union (mIoU), recall (r), and precision (p) are used for the evaluation of our experiments.

3.2. Implementation details

We introduce a novel deep learning-based Pytorch framework by incorporating CNNs into a heuristic-based segmentation pipeline. The deep neural networks, namely spatial pyramid pooling [16] module and encoder-decoder structure, are used for semantic segmentation tasks. Spatial pyramid pooling modules can encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and multiple effective fields-of-view. At the same time, the encoder-decoder structure can capture sharper object boundaries by gradually recovering spatial information. Combining them gives DeepLabv3+ [6] architecture. This paper exhibits the hybrid optimized DeepLabv3+ with ResNet-101 as its backbone, which uses an Adam optimizer with a learning rate of $1e^{-4}$. The images were resized to 400×400 resolution, fed in a batch size of 12 to the model, and trained with binary cross-entropy loss for 50 epochs for both datasets. The training resulted in the early stopping of the model at 30 epochs for the instrument dataset and 25 epochs for the polyp detection dataset. This model has a great real-time segmentation efficiency of 78 fps (frames per second). The training time for both datasets was less than 30 minutes. All experiments were majorly carried on Windows 10, Xenon Gold 5218 CPU at 2.30 GHz (Intel), with 96 GB RAM and 11 GB Nvidia RTX 2080 Ti GPU.

For Kvasir-seg, 5-fold cross-validation with 20% data for testing is used to report the average confidence interval in Table 4. Kvasir-instrument dataset is evaluated only once pertaining to the dataset size.

3.3. Data Augmentations

For data augmentations, the cv2 library is incorporated over transforms from torch-vision to retain the real image and produce new augmented images. Statistically devised augmentation algorithm of horizontal and vertical flips with brightness (color jitter) variation of 30 levels is followed by cropping the image size to $9/10^{th}$ of its sides. The strategy increases the randomness of the augmented images with underlying robustness for the specified segmentation tasks.

4. RESULTS AND DISCUSSION

4.1. Ablation Study

The ablation study for data augmentation is shown in Table 2, we compared the results for different types of augmentations for kvasir-seg [13] and kvasir-instrument [14]. We have done without our augmentation (W/O), W_{Jitter}, W_{Flip}, W_{crop}, and rightmost column as $W_{Proposed}$ (flipping + color jitter + cropping). The proposed method increases the mIoU score by 2.46% for kvasir-seg [13] and 1.6% for kvasir-intrument [14] compared to without augmentation training. We also analyzed sensitivity for batch sizes of 2,4,8,12,16 in Table 3 with BS_{12} being a more optimized batch size. It is a clear pick of our proposed augmentation heuristic devised for the varying image illumination colonoscopy task having the highest average mIoU metric for both the datasets consistently. Refraining from using image normalization in training has also improved the mIoU score by around 1% and DSC score by around 1.5% for both datasets.

Augmentation	*W/O*	W_{Jitter}	W_{Flip}	W_{Crop}	$W_{Proposed}$
Kvasir-seg [13]	0.8771	0.8892	0.8857	0.8824	**0.9017**
Kvasir-instr. [14]	0.9351	0.9398	0.9421	0.9375	**0.9515**

Table 2. mIoU metric result for ablation study of the proposed model on data augmentation.

Batch Size	BS_2	BS_4	BS_8	BS_{12}	BS_{16}
Kvasir-seg [13]	0.8749	0.8881	0.8968	**0.9017**	0.8996
Kvasir-instr. [14]	0.9195	0.9397	0.9480	**0.9515**	0.9509

Table 3. mIoU metric result for ablation study of the proposed model on batch size (BS).

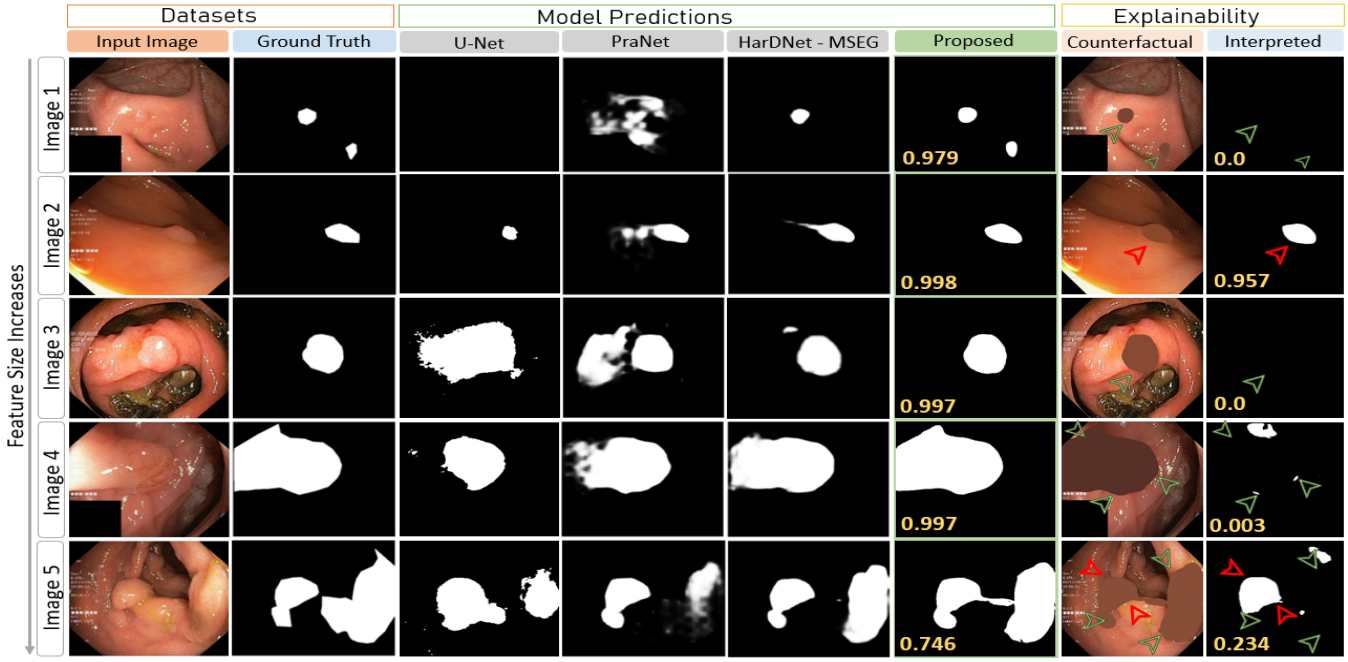

Fig. 1. Qualitative comparison of model's performance with counterfactual explanation against Kvasir-seg dataset [15]. The region pointed by the red mark on interpreted mask represents the failed case, while the green mark symbolises the effectiveness of the corresponding feature learning in the input image. The class prediction interpretability score is overlayed in yellow text on the proposed and interpreted images. Look for significant differences between the proposed (column 6) and interpreted (column 8) prediction scores. The lower the score of an interpreted image after counterfactual manipulation of an input image with the predicted mask implies the higher importance of the feature for an input image.

4.2. Kvasir-seg

The quantitative results in Table 4 confirm that our method secures 94.02% DSC score, 90.017% mIoU score, 98.24% recall score, and 98.13% precision score. It outperforms SOTA methods on all metrics with a margin of 1.85% improvement on DSC, 1.03% on mIoU, 1.47% on precision, and 6.26% increase on the recall compared to MSRF-Net [12].

The network's ability to segment polyps can be observed by comparison of the predicted mask against the ground truth given in Fig. 1. It confirms the ability to detect polyp features with varying increasing sizes taken at random from the validation data for representation. In Fig. 1, we observe evident bleeding around specularity with illumination variation, especially in brighter regions, even for a few high accuracy model predictions like PraNet [7], HarDNet-MSEG [11]. Our proposed method resolves the challenges mentioned above to a great extent. The precision of the model's performance is explained with counterfactual interpretation to detect polyp in varied illumination environments. An overall lower interpretability score for the predicted segmentation mask gives an understanding of the underlying feature of interest learned by the model. Explanation unveils the likelihood of smaller high-contrast regions correctly interpreted in rows 2 and 4 of the Fig. 1 is due to the spatial context learning of the model with textural property and feature shape.

Method	DSC	mIoU	Recall	Precision
SFA(MICC 19)	0.7230	0.6110	-	-
NanoNet-C [10]	0.7494	0.6360	0.8081	0.7738
NanoNet-B [10]	0.7860	0.6799	0.8392	0.8004
ResUNet-mod [12]	0.7909	0.4287	0.6909	0.8713
ResUNet++ [4]	0.8133	0.7927	0.8774	0.7064
FANet [15]	-	0.8153	0.9058	0.9005
HRNetV2-W18-Smallv2 [12]	0.8179	0.7470	0.8016	0.8696
U-Net [12]	0.8180	0.7460	0.6306	0.9222
ColonSegNet [17]	0.8206	0.7239	0.8496	0.8435
U-Net++ [12]	0.8210	0.7430	-	-
NanoNet-A [10]	0.8227	0.7282	0.8588	0.8367
ResUNet+ + TTA + CRF [18]	0.8508	0.8329	0.8756	0.8228
DDANet [19]	0.8576	0.7800	0.8880	0.8643
DeepLabv3+(Mobilenet)[12]	0.8656	0.8186	0.8808	0.9205
HRNetV2-W48 [12]	0.8896	0.8262	0.8973	0.9056
DeepLabv3+(Xception) [12]	0.8965	0.8575	0.8984	.9496
PraNet [7]	0.8980	0.8400	-	-
AG-CUResNeSt-101 [20]	0.902	0.845	-	-
UACANet-S [9]	0.905	0.852	-	-
UACANet-L [9]	0.912	0.859	-	-
HarDNet-MSEG [11]	0.912	0.857	-	-
Polyp-PVT [21]	0.917	0.864	-	-
CaraNet [8]	0.918	0.865	-	-
TransFuse-S [22]	0.918	0.868	-	-
TransFuse-L [22]	0.920	0.870	-	-
MSRF-Net [12]	0.9217	0.8914	0.9198	0.9666
Proposed Method (avg)	**0.9402**	**0.9017**	**0.9824**	**0.9813**
5-fold cross-Validation (95% confidence interval)	± 0.46%	± 0.63%	± 0.36%	± 0.30%

Table 4. Standard metric evaluation for Kvasir-seg [13].

4.3. Kvasir-instrument

The quantitative results in Table 5 confirm that our method secures 97.25% DSC score, 95.15% mIoU score, 99.79% recall score and 99.34% precision score and outperforms SOTA methods on all metrics with a margin of 4.41% improvement on DSC, 7.25% on mIoU, 4.52% on precision, and 9.42% on recall as compared to SOTA NanoNet-B [10]. Qualitative results for instrument segmentation are reported in Fig. 2. The capacity of the network to segment instruments successfully using the same model opens a new horizon of the optimization need in deep learning approaches with explainability.

In Fig. 2, all three instruments with distinct textures, shapes, and sizes are segmented accurately compared with the ground truth. An observed failure marked in red in counterfactual explainability of highly contrasting instruments owes to the textural learning ability of the model with shape attention. Fig. 3 exhibits the potential fusion of the single architecture into an ensemble model with transfer learning for medical diagnosis with efficacy.

Method	DSC	mIoU	Recall	Precision
DoubleUNet [15]	0.9038	0.8430	0.9275	0.8966
NanoNet-C [10]	0.9139	0.8600	0.9037	0.9452
ResUNet++ (ISM'19) [4]	0.9140	0.8635	0.9103	0.9348
U-Net [15]	0.9158	0.8578	0.9487	0.8998
NanoNet-A [10]	0.9251	0.8768	0.9142	0.9540
NanoNet-B [10]	0.9284	0.8790	0.9205	0.9482
Proposed Method	**0.9725**	**0.9515**	**0.9979**	**0.9934**

Table 5. Metric evaluation for Kvasir-instrument [14]

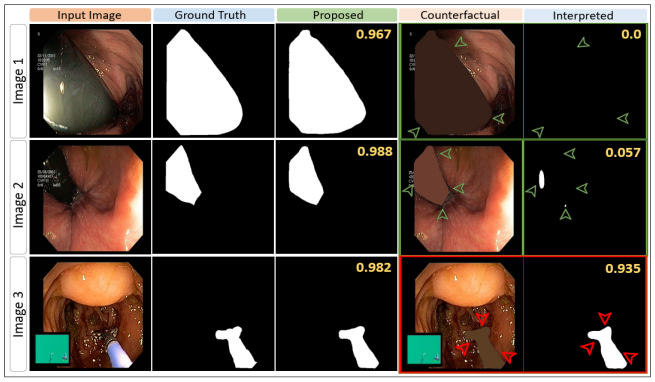

Fig. 2. The qualitative comparison with counterfactual explanation for the Kvasir-instrument dataset [14]. The interpretability score for instrument segmentation is overlayed in yellow text. The larger difference between proposed and interpreted prediction scores accentuates the higher learning quality of the feature space.

4.4. Interpretability

The attention map gives a good understanding of the model's feature learning ability apt for structural learning with a smaller field of view. The preceding heuristics is somewhat convincing but lacks the structure of interpretation needed in the high-risk medical domain. Thereby, an attempt is made to understand the difference in extracted features from the model and trace the potential red flags perceived as a challenge in learning. The certainty in the health field is necessary, but even a diminutive visual interpretation can reveal a rich story contrasted to just statistical figures.

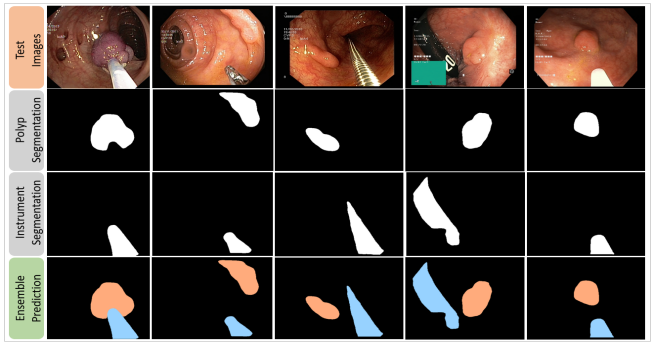

Fig. 3. Model evaluation on test inputs shows the precision of segmenting both polyp and instruments by the proposed method.

In Fig. 1 and Fig. 2, counterfactual images shown in the second last column are produced by replacing the segmented area of interest with the average pixel value of the remaining image. This counterfactual image is passed through the model, and the newly generated image is the interpreted image shown in the rightmost column.

The counterfactual interpretation concludes weak learning of high structural contrast images like image 2 in Fig. 1 and image 3 in Fig. 2. The differences appear because of the pixel-wise learning challenge. Variability in the interpreted images shows that the model is not only learning the textures and spatial context but also the shapes of polyps or instruments. The explicit representation for both kinds of datasets helps even a medical layman conceive and interpret the efficacy of the automatic segmentation task.

5. CONCLUSION

This paper proposes an effective use of data augmentation to an optimized DeepLabv3+ model that can help retain the pre-trained model with a small number of training data. Our experiments confirm that DeepLabv3+ with our proposed method outperforms several SOTA methods on two independent biomedical datasets. We also demonstrate explainability of our approach and feature importance using the counterfactual interpretation method.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [15]. Ethical approval was not required, as confirmed by the license attached with the open-access data.

7. ACKNOWLEDGMENTS

We acknowledge UiT's thematic funding project VirtualStain for registration charges with Cristin Project ID 2061348 (to A.H. and D.K.P.) and D.S. acknowledges research internship funding support from Research Council Norway's INTPART grant no. 309802.

8. REFERENCES

[1] G. Wang et al., "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1562–1573, 2018.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[3] O. Ronneberger, P. Fischer, and 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.

[4] D. Jha et al., "Resunet++: An advanced architecture for medical image segmentation," in *IEEE International Symposium on Multimedia*. IEEE, 2019, pp. 225–2255.

[5] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "Saunet: Shape attentive u-net for interpretable medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 797–806.

[6] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[7] D.P. Fan et al., "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

[8] A. Lou, S. Guan, and M. Loew, "Caranet: Context axial reverse attention network for segmentation of small medical objects," *arXiv preprint arXiv:2108.07368*, 2021.

[9] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp semgnetaion," *arXiv preprint arXiv:2107.02368*, 2021.

[10] D. Jha et al., "Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy," *arXiv preprint arXiv:2104.11138*, 2021.

[11] C.H. Huang, H.Y. Wu, and 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.

[12] A. Srivastava et al., "Msrf-net: A multi-scale residual fusion network for biomedical image segmentation," *arXiv preprint arXiv:2105.07451*, 2021.

[13] D. Jha et al., "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[14] D. Jha et al., "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 218–229.

[15] S. Hicks et al., "MedAI: Transparency in Medical Image Segmentation," *Nordic Machine Intelligence*, 2021.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 2, pp. 2169–2178.

[17] D. Jha et al., "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[18] D. Jha et al., "A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2029–2040, 2021.

[19] N.K. Tomar et al., "Ddanet: Dual decoder attention network for automatic polyp segmentation," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 307–314.

[20] D.V. Sang, T.Q. Chung, P.N. Lan, D.V. Hang, D.V. Long, and N.T. Thuy, "Ag-curesnest: A novel method for colon polyp segmentation," *arXiv preprint arXiv:2105.00402*, 2021.

[21] B. Dong, W. Wang, D.P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[22] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *arXiv preprint arXiv:2102.08005*, 2021.