Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418037

CNN based U-Net with Modified Skip Connections for Colon Polyp Segmentation

Sushma B
Department of ECE
C M R Institute of Technology
Bengaluru
sushma.b@cmrit.ac.in

Raghavendra C K
Department of CSE
B N M Institute of Technology
Bengaluru
raghav.ck.clk@gmail.com

Prashanth J
Department of CSE
B N M Institute of Technology
Bengaluru
Jayaramprash@gmail.com

***Abstract*—Colonoscopy is a medical procedure performed to detect the anomalies in the colon and rectum. A thin flexible wire embedded with the camera is inserted to directly visualize the colon. Direct visualization enables early detection and removal of the polyps in the colon. Polyps exist in different shapes and sizes. The physicians find it very challenging to diagnose small polyps in colonoscopy video. Delay in polyp removal leads to colorectal cancer and leads to cancer-related death. In this work UNET architecture with spatial attention layer is proposed to improve the precision of segmenting polyp regions in colonoscopy video. The CNN models proposed in the literature for polyp segmentation are basically trained using common loss functions such as dice and binary cross entropy loss. Under the training of these loss functions the model learns poorly in segmenting small sized polyps. This can be solved by using focal Tversky loss. Experiments are conducted on a publicly available dataset. Results show that UNET with spatial attention layer trained with focal Tversky loss performs better compared to standard UNet model trained with common loss functions.**

***Keywords*—Colonoscopy, CNN, Image Segmentation, Polyps, Loss function, UNet**

I. Introduction

Colonoscopy is considered as an important medical procedure performed to diagnose colon related abnormalities such as polyps, colorectal cancer etc. Colorectal cancer is one of the prime cause for the increase in death rate, especially in men. Colon polyp is considered as an early indication of colorectal cancer. If polyps are not detected and removed, they can turn out into cancer and can spread into other organs. Chances of survival can be increased, if the polyps are detected and removed when they are very small [1]. Therefore it is advised by many doctors to have colonoscopy regularly.

In recent days, colonoscopy procedures performed are painless and many people are advised to undergo this procedure on a regular basis for early detection of colon polyps. When the polyps are detected they are removed by resection. Endoscopists while performing the procedure find it very hard to detect polyps in colonoscopy video. Even experienced endoscopists sometimes fail to detect the flat natured polyps. Polyp detection is a challenging task for endoscopists due to the variation in size, shape and texture. The manual detection of small sized flat polyps is missed because of difficulty in identifying them in colonoscopy video. The missed polyps can lead to cancer if the patient delays in undergoing colonoscopy second time. Computer aided polyp detection tool acts as second detector and can reduce the missing rate by detecting the overlooked polyps. Also, accurate and precise segmentation of polyps can reduce the missing rate significantly and acts as a constructive automated clinical tool which enables faster procedural screening. Accurate segmentation of polyps requires effective extraction of the various multi-formed polyp features which lies in various shape, size and texture. Computer aided segmentation tool can help in precise detection of the polyps and can contribute significantly in early detection of polyps. With the help of the computer aided tool, even an inexpert endoscopist can detect the polyps and can perform resection.

Feature extraction play a major role in segmenting the objects. Many polyp segmentation methods proposed in literature consider texture, color and shape features [2]–[4]. All these features are considered as low level features and not sufficient for precise segmentation of polyp regions. In recent literature convolutional neural network (CNN) architectures are proven to be efficient in automated extraction of features for various computer vision applications. Even for segmentation of polyps in colonoscopy video many CNN architectures are proposed. Fully CNN models are used for polyp region classification which lacks in capturing contextual information [5]. Mask-RCNN is adopted for better segmentation of polyps []. For more accurate segmentation of the polyps with various sizes and shapes. Image segmentation tasks can be done effectively by using encoder-decoder based CNN architectures. Single encoder based CNN with dilated kernels is used for pixel-level polyp segmentation. UNet, a popular encoder-decoder based CNN architecture with skip connections is the accepted choice for image segmentation [6]. Due to continuous down-sampling of features at different levels on the encoder, side informations associated with edges and small objects is lost during up-sampling in the base-line UNet architecture.

To overcome the above problem, techniques such as pyramid pooling [7] and attention mechanism [8] are introduced in encoder decoder based architectures. The inclusion of the above techniques results in developing efficient segmentation architectures like ResUNet [9], Twin UNet [10], UNet with dilated convolutions [11] etc. The above architectures deals in segmenting polyps with diversified features. In all these architectures the low-level features extracted at the encoder are directly concatenated with corresponding high-level decoder

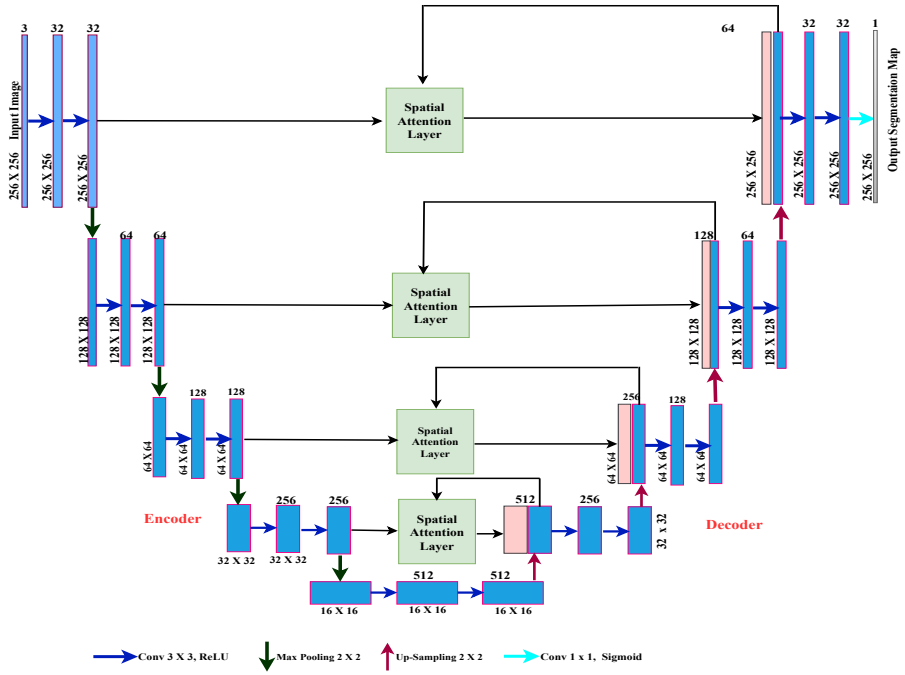

Fig. 1: UNet with modified skip connections

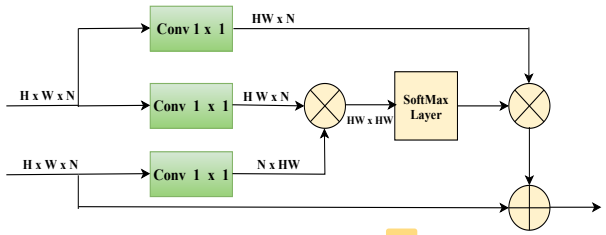

Fig. 2: Spatial attention layer, where $\oplus$ is a matrix addition and $\otimes$ is a matrix multiplication operation.

features. Because of the incompatibility between low and high level encoder-decoder features, polyp pixels are poorly predicted. Other set of CNN based polyp segmentation models proposed adopts spatial attention mechanisms and gives good segmentation results. But these models architectures are highly complex and require a large number of model parameters and reduce the speed of segmenting.

Based on the above studies, in this work UNet with modified skip connections is proposed. In the proposed method a spatial attention layer [12] is inserted between every encoder-decoder skip connections. Spatial attention layer merges the features extracted by the encoder at different levels. Including spatial attention layers between encoder-decoder gives more precise results compared to CNN models with direct encoder-decoder skip connections. In this work, the proposed UNet architecture is trained with different losses and observed it performs better when trained with focal-Tversky loss [13]. The performance of the proposed method is better compared to the state of the art CNN based polyp segmentation methods.

II. Methodology

A. Network Architecture

Proposed UNet architecture with modified skip connections is shown in Fig.1. The modified UNet model with spatial attention layer is end-end trainable. Every level in the encoder includes a sequence of convolution layers followed by batch normalization layer and ReLU activations. Max-pooling operations are utilized to down-sample the features size extracted from each level. After each level the number of features are doubled. On the decoder side each level consists of the transposed convolution layer which doubles the feature size by 2. Number of features are reduced by half after each level. Features extracted at each level from the encoder are concatenated with the corresponding level decoder features through skip connections. Instead of direct concatenation of features, modified skip connections are used in the proposed work. The skip connections are modified by including spatial attention layer (SAL). SAL is used to generate the spatial relationship between the features produced at each stage of the encoder [14]. The different operations associated with SAL is shown in Fig. 2. The final level of the decoder side consists of a 1 x 1 convolution layer and sigmoid activation layer. All the convolution layers in terms of kernel-size, number of kernels and output is described in Table I.

High level local features are concatenated with low level global features using skip connections. However, spatial details are lost in high level feature maps due to series of convolutions and non-linear transformations. Due to this it is difficult to reduce false detections, mainly for small polyp regions. To overcome this problem skip connections are modified by using SAL. SAL extracts the required spatial information from low level features and concatenates it with corresponding decoder level features.

TABLE I: Network parameters of UNet with modified skip connections

Layer	Number of kernels	Kernel Size	Output Size
Input	-	-	256 x 256 x 3
Encoder Parameters			
Conv-1	32	3x3	256x256x32
Conv-2	32	3x3	256x256x32
Conv-3	64	3x3	128x128x64
Conv-4	64	3x3	128x128x64
Conv-5	64	3x3	128x128x64
Conv-6	128	3x3	64x64x128
Conv-7	128	3x3	64x 64x128
Conv-8	128	3x3	64x64x128
Conv-9	256	3x3	32x32x256
Conv-10	256	3x3	32x32x256
Conv-11	256	3x3	32x32x256
Encoder-Decoder Connection Layer			
Conv-12	512	3x3	16x16x512
Conv-13	512	3x3	16x16x512
Conv-14	512	3x3	16x16x512
Decoder Parameters			
T-Conv-15	256	3x3	32x32x256
Conv-16	256	3x3	32x32x256
Conv-17	256	3x3	32x32x256
T-Conv-18	128	3x3	64x64x128
Conv-19	128	3x3	64x64x128
Conv-20	128	3x3	64x64x128
T-Conv-21	64	3x3	128x128x64
Conv-22	64	3x3	128x128x64
Conv-23	64	3x3	128x128x64
T-Conv-24	32	3x3	256x256x32
Conv-25	32	3x3	256x256x32
Conv-26	32	3x3	256x256x32
Conv-27	1	3x3	256x256x1
Spatial attention layer at level1			
Conv-28	32	1x1	256x 256x32
Conv-29	32	1x1	256x256x32
Conv-30	32	1x1	256x256x32
Spatial attention layer at level2			
Conv-31	64	1x1	128x128x64
Conv-32	64	1x1	128x128x64
Conv-33	64	1x1	128x128x64
Spatial attention layer at level3			
Conv-34	128	1x1	64x64x128
Conv-35	128	1x1	64x64x128
Conv-36	128	1x1	64x64x128
Spatial attention layer at level4			
Conv-37	256	1x1	32x32x256
Conv-38	256	1x1	32x32x256
Conv-39	256	1x1	32x32x256

B. Loss Functions

When image segmentation related tasks are performed using deep learning models, loss functions plays an important role. The selection of loss function while training the deep learning model is very important as it incites the learning process of the model. In the proposed work, UNet and UNet with modified skip connections is trained with various segmentation loss functions. The important loss functions considered in this work are binary cross entropy (BCE), Dice loss (DL), mean squared error (MSE) and Tversky Loss (TL). These losses fail to attain proper segmentation of small polyps because false-positives and false-negatives are weighted equally. These results in high precision and low recall. Recall rate can be improved by high rate detection of false-negatives compared to false-positives. The balance between the both can be achieved by using Focal Tversky loss (FTL). Deep segmentation models trained with FTL make an effort to learn segmenting small region of interests using γ coefficient. Computation of FTL is given in (1).

$$FTL = \sum_{c}(1 - TI_c)^{\frac{1}{\gamma}} \tag{1}$$

where, TI is the Tversky similarity index given in (2), γ value is set to 1.3 at which it gives the best performance when different experiments are conducted.

$$TI_c = \frac{\sum_{x=1}^{N} p_{xc} g_{xc} + 10^{-5}}{\sum_{x=1}^{N} p_{xc} g_{xc} + \alpha \sum_{x=1}^{N} p_{i\bar{c}} g_{xc} + \beta \sum_{x=1}^{N} p_{xc} g_{x\bar{c}} + 10^{-5}} \tag{2}$$

Here, p_{xc} is the probability of pixel belonging to polyp region and $p_{x\bar{c}}$ is the probability of the pixel belonging to non-polyp region in the predicted segmentation map. Similarly g_{xc} and $g_{x\bar{c}}$ are the probabilities of the ground-truth mask. α and β are the hyper-parameters set to 0.75 and 0.25 respectively. Model convergence will improve when higher α is selected by minimizing false negatives.

III. Experiments and Results

A. Dataset Details

To train the model a popular polyp segmentation dataset KVASIR-SEG [15] is used. Dataset consists of around 2134 images with polyps and corresponding segmentation masks. The images are having different resolutions. For training purpose all the images are resized to the size 256 X 256. Dataset is divided into train, validation and test data with 70-15-15 split.

B. Training Details

The proposed UNet model is implemented in Keras by using Tensorflow as a backend. The model is trained on NVIDIA-P100 GPU with 16GB GPU memory available on Kaggle. Adam optimizer with 0.0001 learning rate is used to optimize the model parameters. The entire network is trained end to end for over 50 epochs with batch size of 8.

C. Evaluation Metrics

The model trained for different loss functions is evaluated using Dice score coefficient (DSC), mean-intersection over union (mean-IoU), precision and recall computed by using (3), (4), (5) and (6) respectively. Here GSM is the ground truth segmentation mask and PSM is the predicted segmentation mask.

$$DSC = \frac{2\ (pixels\ in\ GSM)\ \cap\ (pixels\ in\ PSM)}{(pixels\ in\ GSM)\ +\ (pixels\ in\ PSM)} \tag{3}$$

$$mean - IoU = \frac{(pixels\ in\ GSM)\ \cap\ (pixels\ in\ PSM)}{(pixels\ in\ GSM)\ \cup\ (pixels\ in\ PSM)} \tag{4}$$

$$Precision = \frac{Correctly\ predicted\ polyp\ pixels}{Total\ number\ of\ polyp\ pixels\ in\ PSM} \tag{5}$$

$$Recall = \frac{Correctly\ predicted\ polyp\ pixels}{Total\ number\ of\ polyp\ pixels\ in\ GSM} \tag{6}$$

TABLE II: Performance comparison of standard UNet and UNet with modified skip connections for different losses

Model	Loss	DSC	mean-IoU	Precision	Recall
UNet	BCE	0.621	0.434	0.703	0.578
	DL	0.547	0.406	0.653	0.558
	MSE	0.677	0.421	0.664	0.598
	TL	0.657	0.569	0.732	0.623
	FTL	0.669	0.607	0.775	0.715
UNet-MSC	BCE	0.703	0.651	0.813	0.624
	DL	0.810	0.791	0.874	0.627
	MSE	0.819	0.746	0.849	0.762
	TL	0.828	0.755	0.819	0.715
	FTL	0.911	0.808	0.894	0.930

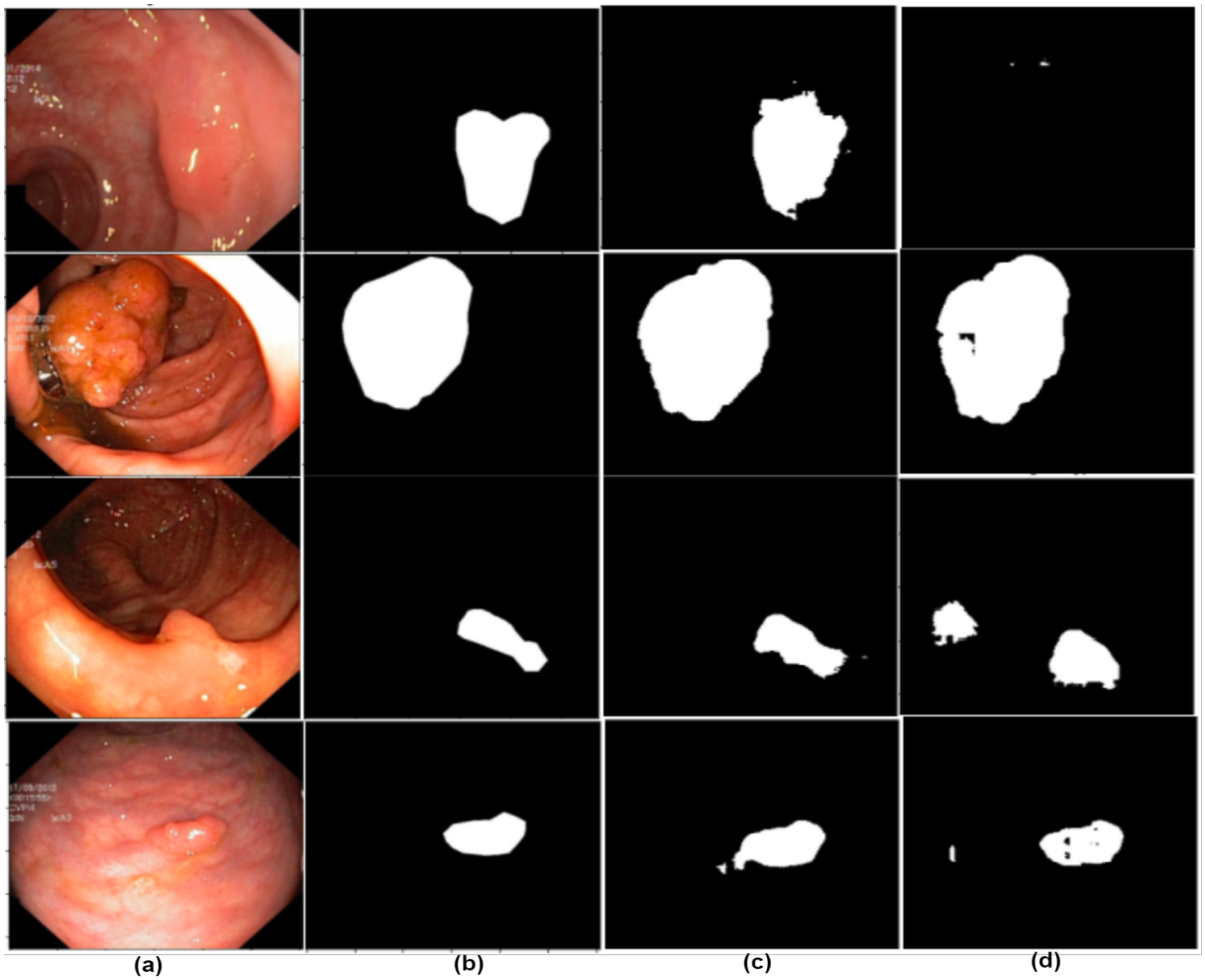

Fig. 3: Polyp segmentation results obtained by modified UNet and standard UNet for FTL. (a) Input Image; (b) Ground-truth Segmentation map; (c) Modified UNet; (d) Standard UNet

D. Results

UNet with modified skip connections (UNet-MSC) is compared with UNet with plain skip connections. Both the models are evaluated for considered metrics trained under different losses. Comparative analysis is given in Table II. It is desired to obtain low standard deviation in precision and recall. High standard deviation results in unstable learning. From the results it can be observed that the standard deviation between precision and recall is very less when the model is trained with FTL. Predicted polyp regions along with ground truth segmentation maps are given in Fig. 3. From the segmentation maps it can be observed that for small and flat polyps modified UNet gives better segmentation performance than standard UNet trained with the same loss function.

IV. Conclusion

In this work, a modified UNet architecture for polyp segmentation in colonoscopy videos is proposed. The skip connections of the standard UNet are modified by a spatial attention layer. This work demonstrates the importance of the spatial attention layer in segmenting small and flat polyp regions. Importance in selection of loss function for segmentation is also shown. Modified UNet model trained with focal-Tversky loss performs better in DSC by 26% and mean-IoU by 34% with stabilized precision and recall scores compare to standard UNet.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[2] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE transactions on medical imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[3] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[4] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE transactions on medical imaging*, vol. 34, no. 11, pp. 2379–2393, 2015.

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

[6] 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.

[7] 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.

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[9] D. Jha, S. Ali, H. D. Johansen, D. D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning," *arXiv preprint arXiv:2011.07631*, 2020.

[10] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 558–564.

[11] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS-Science and Technology Publications, 2019, pp. 632–641.

[12] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[13] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[15] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.