# THE IMPORTANCE OF SKIP CONNECTIONS IN ENCODER-DECODER ARCHITECTURES FOR COLORECTAL POLYP DETECTION

*Nita Mulliqi*[1], *Sule Yildirim*[1], *Ahmed Mohammed*[1], *Lule Ahmedi*[2], *Hao Wang*[1], *Ogerta Elezaj*[1], *Øistein Hovde*[3]

[1] Department of Computer Science, NTNU, Gjøvik, Norway
[2] Department of Computer Engineering, University of Prishtina, Kosovo
[3] Department of Gastroenterology, University of Oslo, Oslo, Norway

## ABSTRACT

Accurate polyp detection during the colonoscopy procedure impacts colorectal cancer prevention and early detection. In this paper, we investigate the influence of skip connections as the main component of encoder-decoder based convolutional neural network (CNN) architectures for colorectal polyp detection. We conduct experiments on long and short skip connections and further extend the existing architecture by introducing dense lateral skip connections. The proposed segmentation architecture utilizes short skip connections in the contracting path, moreover it utilizes dense long and lateral skip connections in between the contracting and expanding path. Results obtained from the MICCAI 2015 Challenge dataset show progressive improvement of the segmentation result with expanded utilization of skip connections. Our proposed colorectal polyp segmentation architecture achieves near state-of-the-art results with significantly reduced number of model parameters.

***Index Terms—*** Colorectal polyps, encoder-decoder methods, polyp segmentation, skip connections, U-Net++.

## 1. INTRODUCTION

Colorectal cancer is the third most common cancer in the world [1]. Patients with a localized stage of the *in vivo* colon cancer have survival rate up to 90%, though patients with distant stages have survival rate up to 14% [2]. Therefore, detection at early stages is of high importance. Colonoscopy is considered as the main screening procedure by allowing complete colon examination as well as removing precancerous polyps. It was demonstrated that there is a considerable miss-rate of non-neoplastic polyps and adenomas as 28 % and 20 % respectively [3]. Multiple factors contribute to missed polyps during colonoscopy, such as endoscopist experience and bowel preparation [4]. Furthermore, endoscopists over or under-estimate size of polyps, causing inappropriate surveillance time in 10% of the cases [5]. Different methods are dedicated to automated polyp detection [6], however encoder-decoder based methods have recently shown superior performance [7] [8] [9] [10]. In encoder-decoder based convolutional neural network (CNN) architectures colorectal polyp segmentation is conducted by employing traditional CNNs structures in the contracting path with repeated down-sampling layers (such as pooling layers) [11] [12]. Down-sampling layers provide local translation invariance, reduce spatial size of the representation and extract low-level feature maps. This squanders important localization information crucial to construct pixel-wise predictions in the expanding path. To address this issue, skip connections are introduced to recover the full spatial resolution by consolidating fine-grain, low-level feature maps from the contracting path with semantic, coarse-grain feature maps in the expanding path [11] [12]. In the Fully Connected Network (FCN) [11] skip connections are utilized to build high-resolution feature maps by merging contracting and expanding path feature maps on each scale. In Seg-Net [13] they are utilized to up-sample low-level feature maps by carrying pooling indices to the expanding path. U-Net [12] utilizes skip connections to recover spatial information similar as in [11], but employs convolutions and non-linearities after merging. This results in better information retain, making U-Net [12] widely used for biomedical image segmentation. In Mask-RCNN [14] skip connections are utilized in a similar approach, enabling segmentation of occluded objects. In Dense-Net [15] skip connections merge feature maps from each layer to every other

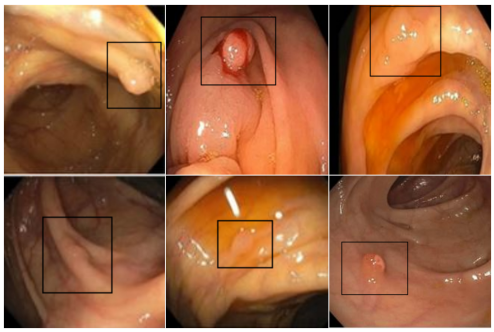


**Fig. 1**: Variability in appearance of polyps. Polyps appear in different size, shape, color, texture and location within colon and rectum walls.

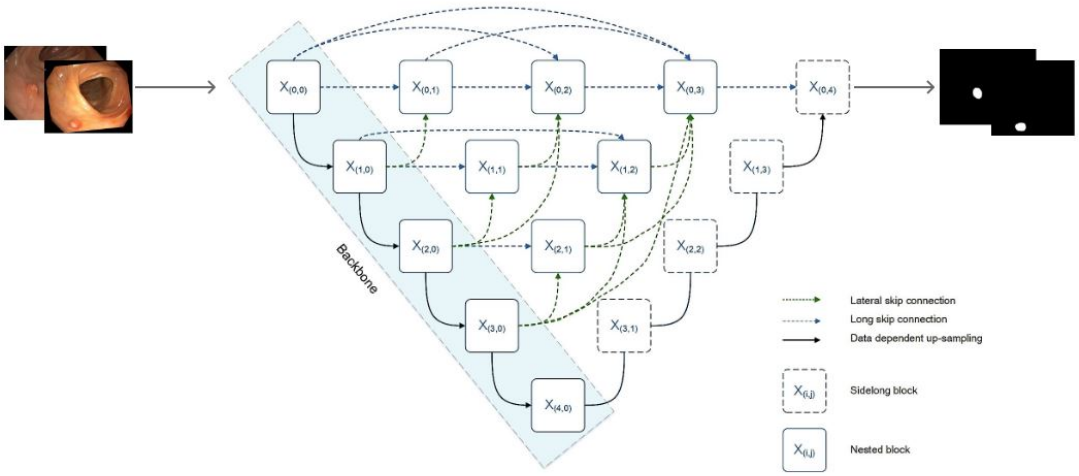



**Fig. 2**: Our proposed segmentation architecture: Dense U-Net++. It consists of contracting and expanding path filled with dense skip pathways, blocks of convolutions and activation functions in between. Three types of skip connections are utilized, long skip connections (shown in blue), lateral skip connections (shown in green) and short skip connections comprehended in the backbone structure.

layer in a feed-forward fashion. Similarly, [16] utilizes hybrid densely connected U-Net for the segmentation of tumor and liver. U-Net++ [10] utilizes dense long skip connections along with lateral skip connections in a deeply-supervised encoder-decoder network. It uses blocks of convolutions on each scale in between contracting and expanding path, hence combining semantically similar feature maps and improving performance [10]. To all the above varieties, we will refer to as long (lateral) skip connections [17]. As CNN architectures become deeper with increasingly more layers, the problem of vanishing gradients appears when training with back-propagation. This issue is treated by various architectures through skip connections [18] [19] [20]. Employing the so-called identity connections in parallel with the activation function blocks [18] [19] enables the gradient to pass through layers uninterruptedly, resulting in higher derivative of the block. In [17] short skip connections were employed in the contracting path of the FCN architecture. We will refer to this variety of skip connections as short skip connections [17]. Generally all these approaches resulted in high performance in natural images segmentation, however their utilization in medical domain still requires further assessments. Given that polyps have large appearance variability, causes more challenge in redeeming fine-grain information. We have depicted colorectal polyp appearance varieties in Figure 1. Skip connections directly influence colorectal polyp segmentation result, hence further investigating their impact is essential. The contributions of the paper can be summarized as: (1) we propose a new colorectal polyp segmentation architecture: Dense U-Net++ introducing dense lateral skip connections in between the contracting and expanding path, (2) we show that the new segmentation architecture achieves performance comparable to the state-of-the-art on the MICCAI Challenge 2015 dataset under significantly reduced number of parameters, (3) we demonstrate the substantial influence of skip connections in the encoder-decoder CNN architectures for colorectal polyp segmentation.

## 2. OUR METHOD

As demonstrated in U-Net++ [10] and Dense-Net [15], skip connections improve the segmentation and classification performance. In this work, we further enhance the colorectal polyp segmentation result by introducing the *dense lateral skip connections*. Figure 2 shows a high-level overview of the proposed architecture. Sidelong blocks in the contracting path are gained from the Res-Net [19] backbone structure. Sidelong blocks in the expanding path contain layers of CONV-BatchNorm(BN)-RELU-Dropout. Analogously, nested blocks contain two sub-blocks with CONV-BatchNorm(BN)-RELU and a dropout layer (only in the second sub-block). Concatenation is performed to all feature maps gathered from skip connections to each nested block. To improve the localization accuracy in the expanding path, we perform data depended up-sampling [21] at the final scale. A composite function $H(\cdot)$ represents a block with layers of CONV-BatchNorm(BN)-RELU-Dropout. For any composite function $H(\cdot)$, [ ] concatenation operation and $U(\ )$ up-sampling operation, based in Figure 3, the input-output correlation $x_{i,\,j}$ in $i$ scale and $j$

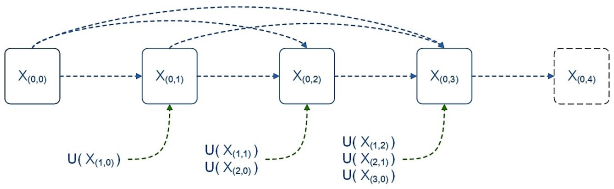



**Fig. 3**: First scale of the proposed segmentation architecture. It contains three nested blocks $(X_{0,1}, X_{0,2}, X_{0,3})$ with dense skip connections along the path and two side-long blocks $(X_{0,0}, X_{0,4})$. The three bottom arrows depict the information conveyed from dense lateral skip connections.

block can be denoted as: $X_{(0,1)} = H[X_{(0,0)},\ U(X_{(1,0)})]$, $X_{(0,2)} = H[X_{(0,0)},\ X_{(0,1)},\ U(X_{(1,1)}),\ U(X_{(2,0)})]$ and $X_{(0,3)} = H[X_{(0,0)},\ X_{(0,1)},\ X_{(0,2)},\ U(X_{(1,2)}),\ U(X_{(2,1)}),\ U(X_{(3,0)})]$. By employing short skip connections [19], each block has direct access to gradients of the original input providing an implicit deep supervision in the contracting path. This way, eluding blocks of $H(\cdot)$ with the identity functions, the output is gained by summing the weights of two previous paths as $x_i = H_i(x_{i-1}) + x_{i-1}$. The expanding path does not utilize skip connections, hence the output of the $i^{th}$ layer in the expanding path is actually the input of the $(i+1)^{th}$ layer as $x_i = H_i(x_{i-1})$. Finally, we can depict the output $x_{lo(i,j)}$ in $i$ scale and $j$ block acquiring as input all feature maps from long skip connections as $x_{lo(i,j)} = H([x_{i,l}]_{l=0}^{j-1}), j > 0$, where $l$ is the positioning index. All feature maps from the preceding blocks are first merged with the concatenation operation, then passed into the $H(\cdot)$ block. Through long skip connections every block is connected to every other block on same scale in horizontal orientation.

### 2.1. Dense lateral skip connections

Lateral skip connections are similar with long skip connections in terms of connectivity pattern, however fundamentally different in position as they fuse semantically similar feature maps in a diagonal approach. By introducing dense lateral skip connections we employ diagonal association of every sidelong block in the contracting path with every nested block. In this regard, for each block feature maps of all preceding blocks are used as inputs and its own feature maps are used as inputs to all consequent blocks in every scale of the encoder-decoder structure. This approach improves the information flow until it reaches the $X_{(0,4)}$ block in the expanding path. Furthermore, it re-utilizes feature maps of all scales, increasing variability in the information that is transferred to the expanding path. Through scales with distinct resolution levels, feature maps have different spatial dimensions. Hence, before the concatenating operation we first up-sample feature maps using bilinear up-sampling, then we pass them into the $H(\cdot)$ block. The output $x_{la(i,j)}$ in $i$ scale and $j$ block acquiring as input all the feature maps from lateral skip connections can be defined as $x_{la(i,j)} = H([U(x_{k,l})]_{k=i+1}^{n}|_{l=j-1}^{0}), i < n$, where $k$ and $l$ are positioning indexes. Each block acquiring as input feature maps from dense long and lateral skip connections can be defined as $x_{i,j} = [x_{lo(i,j)}, x_{la(i,j)}]$. The main characteristic of the three types of skip connections is that they create shortcuts from early layers to later layers of the architecture. With dense long and lateral skip connections merging feature maps by concatenating, we provide great distinction of the information that is being added with the already existing information in the proposed architecture.

## 3. EXPERIMENTS

Experiments are conducted on the ASU-Mayo clinic polyp database [22] of MICCAI 2015 Challenge on polyp detection. Dataset comprises of 20 training and 18 testing colonoscopy videos. Each frame extracted from particular videos has pixel level annotated mask. There are in total 4300 frames in test set and 4278 frames containing polyps in the training set. Dataset colonoscopy videos represent practical approach of the procedure, as some of them contain biopsy elements, some of them represent rapid inspection and others more gradual examination. To increase robustness and reduce over-fitting, we perform affine and perspective data augmentation techniques, before and after training respectively. Initially, frames with polyp we employ rotation (10 to 350 degrees), zoom (1 to 1.3), translation in x,y (-10 to 10) and shear (-25 to 25). Secondly, we apply random horizontal (P=0.3) and vertical flip (P=0.4). Before feeding images to the contracting path, we resize into fixed dimensions with spatial size of 224x224 and normalize them to [0, 1]. Our model is implemented on PyTorch library with single NVIDIA GeForce GTX 1080 GPU. We use Adam optimizer with batch size 8 and learning rate $\eta$ set to $10^{-4}$. We monitor Dice coefficient and use early-stop criteria on the validation set error. We employ Dice loss function combined with weighted binary cross entropy to tackle the imbalanced dataset problem. We employ the Sigmoid activation function in the segmentation result to form the error loss. Each pixel is categorized as positive, corresponding to foreground if above the 0.5 threshold. For all positive pixels as the bounding box, x denotes the ground truth bounding box and y denotes segmentation result bounding box, N denotes number of batches, $pc$ denotes the positive weight value and $\sigma$ denotes the Sigmoid function, hence the error loss function is formulated as $Loss(x, y) = \frac{1}{N}(\sum_{i=1}^{N} pc \cdot x \cdot log\sigma(y)) + (1 - \frac{2 \cdot \sum_{i=1}^{N} x \cdot y + \epsilon}{\sum_{i=1}^{N} x + \sum_{i=1}^{N} y + \epsilon})$. Model performance is based on overlap based metric as detection evaluation. In this regard, if the intersection over union (IoU) between $x$ and $y$ is greater than zero, the detection is considered as a true positive, otherwise detection is considered as false positive. IoU can be formulated as $IoU = \frac{y \cdot x}{\sum y + \sum x + \epsilon}$. We use the metrics of precision, recall, F1-Score and F2-Score to evaluate our model performance.

### 3.1. Ablation Study

This section depicts the results of the ablation study conducted, whereat different architectures with distinct approaches of skip connections are investigated. The results shown in Table 1 demonstrate progressive improvement of the segmentation result with expanded utilization of skip connections. The proposed segmentation architecture is highly parameter efficient as it reduces the number of model parameters in comparison with the state-of-the-art (SOTA) shown in Table 2. Furthermore, our proposed segmentation architecture outperforms previous methods for colorectal polyp detection, based on hand-crafted features or 2-D/3-D CNNs. With the re-designed skip connections, our model skips entire blocks, hence does not need to learn redundant parameters. Requiring lower memory, parameter efficiency is a substantial advantage which makes our model more practical in real time inference. In addition, the proposed segmentation architecture can be utilized for different image segmentation purposes, besides the colorectal use case.

| **Architecture** | **Acc.** | **Prec.** | **Rec.** | **F1** | **F2** |
|---|---|---|---|---|---|
| **U-Net [12]** | 83.54 | 89.85 | 32.23 | 47.44 | 36.97 |
| **U-Net++ [10]** | 86.44 | 90.97 | 49.55 | 64.14 | 54.51 |
| **Dense U-Net++ (Ours)** | 90.42 | 87.46 | 70.99 | 78.37 | 73.77 |

**Table 1**: Results of our proposed architecture Dense U-Net++ in comparison with the U-Net [12] and U-Net++ [10]. All the experiments were conducted under common hyper parameter settings.

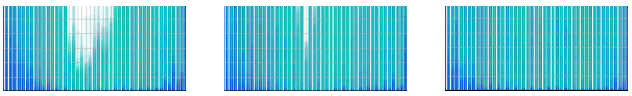
| **Architecture** | **Params** | **TP** | **FP** | **FN** | **Prec.** | **Rec.** | **F1** | **F2** |
|---|---|---|---|---|---|---|---|---|
| **PLS** | | 1594 | 10103 | 2719 | 13.6 | 36.9 | 19.9 | 27.5 |
| **CVC-CLINIC** | | 1578 | 3456 | 2735 | 31.3 | 36.6 | 33.8 | 35.4 |
| **OUS** | | 2222 | 229 | 2091 | 90.6 | 51.5 | 65.7 | 56.4 |
| **ASU** | | 2636 | 184 | 1677 | 93.5 | 61.1 | 73.9 | 65.7 |
| **CUMED** | | 3081 | 769 | 1232 | 80.0 | 71.4 | 75.5 | 73.0 |
| **Fusion** | | 3062 | 414 | 1251 | 88.1 | 71.0 | 78.6 | 73.9 |
| **Y-Net [8] (SOTA)** | **75M** | 3582 | 513 | 662 | 87.40 | 84.40 | 85.90 | 85.00 |
| **Dense U-Net++ (Ours)** | **52M** | 3062 | 439 | 1251 | 87.46 | 70.99 | 78.37 | 73.77 |

**Table 2**: Results of our proposed architecture Dense U-Net++ in comparison with different polyp detection methods on ASU-MAYO dataset. All the methods are published in [6], model parameters were not available.

By employing dense long and lateral skip connections, we fuse semantically similar feature maps in two different orientations (horizontal and diagonal), thus we further decrease the semantic gap between feature maps, consistent with the literature [10]. Besides the proposed skip pathways, additional difference with the U-Net++ [10] is that we up-sample feature maps with data depended up-sampling [21] only in the last scale of the expanding path. In this regard, from bottom to top, sidelong blocks in the expanding path decrease the number of channels, but keep the same spatial dimension. The stack of feature maps gathered in block $x_{(0,4)}$ (see Figure 2) is collected by incoming feature maps gathered from dense skip pathways and up-sampled feature maps from sidelong blocks of the expanding path. Thereby, fine-grain information is better preserved while assisting in better localization accuracy. We investigate the learning behaviour of the model with emphasis to parameter update, by monitoring the gradient flow through layers during first few epochs of training. We empirically validate that layers closer to the center of the model can not be effectively updated with the U-Net in Figure 5(a) and U-Net++ in Figure 5(b). With our proposed segmentation architecture, Figure 5(c), deep center parts of the network get updated better than with preserved usage of skip connections as in U-Net. In the three architectures early shallow layers are not updated maximally (marked with dark blue color).

(a) (b) (c) (d)

**Fig. 4**: Figure displays visual colorectal polyp segmentation result: (a) original image, (b) label, (c) U-Net++ result and (d) Dense U-Net++ result. We can observe that the number of true positives detected in the figures, increases along with the recall rate.



(a) (b) (c)

**Fig. 5**: Gradient flow visualization: (a) U-Net [12] utilizing only long skip connections, (b) U-Net++ [10] utilizing short, lateral skip connections and dense long skip connections, (c) Dense U-Net++ (Ours) utilizing short skip connections, dense long and dense lateral skip connections.

## 4. CONCLUSIONS

Skip connections demonstrated to be the key factor for enhanced colorectal polyp segmentation performance in encoder-decoder CNN architectures. They encourage feature re-use, improve information and gradient flow. Dense skip connections strongly reduce the number of model parameters, achieving near state-of-the-art results for colorectal polyp detection.

## 5. REFERENCES


[1] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D.m. Parkin, M. Piñeros, A. Znaor, and F. Bray, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International Journal of Cancer*, vol. 144, pp. 1941–1953, 2019.

[2] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. Meester, A. Barzi, and A. Jemal, "Colorectal cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, pp. 177–193, 2017.

[3] D. Heresbach, T. Barrioz, M. G. Lapalus, D. Coumaros, P. Bauret, P. Potier, D. Sautereau, C. Boustière, J.C. Grimaud, C. Barthélémy, J. Sée, I. Serraj, PN. D'Halluin, B. Branger, and T. Ponchon, "Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies," *Endoscopy*, vol. 40, pp. 284–290, 2008.

[4] S. N. Bonnington and M. D. Rutter, "Surveillance of colonic polyps: Are we getting it right?," *World Journal of Gastroenterology*, vol. 22, pp. 1925–1934, 2016.

[5] L. Chaptini, A. Chaaya, F. Depalma, K. Hunter, S. Peikin, and L. Laine, "Variation in polyp size estimation among endoscopists and impact on surveillance intervals," *Gastrointestinal Endoscopy*, vol. 80, pp. 652–659, 2014.

[6] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Córdova, C. Sánchez-Montes, S. R. Gurudu, G. Fernández-Esparrach, X. Dray, J. Liang, and A. Histace, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 1231–1249, 2017.

[7] Y. B. Guo and B. J. Matuszewski, "GIANA polyp segmentation with fully convolutional dilation neural networks," *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, pp. 632–641, 2019.

[8] A. K. Mohammed, S. Y. Yayilgan, I. Farup, M. Pedersen, and Ø. Hovde, "Y-Net: A deep convolutional neural network for polyp detection," *British Machine Vision Conference*, p. 308, 2018.

[9] X. Sun, P. Zhang, D. Wang, Y. Cao, and L. Benyuan, "Colorectal polyp segmentation by U-Net with dilation convolution," *IEEE International Conference On Machine Learning and Applications*, 2019.

[10] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045, pp. 3–11, 2018.

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

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention MICCAI*, vol. 9351, pp. 234–241, 2015.

[13] B. Vijay, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, 2017.

[14] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision ICCV*, pp. 2980–2988, 2017.

[15] G. Huang, Zh. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," pp. 2261–2269, 2017.

[16] X. Li, H. Chen, X. Qi, Q. Dou, Ch. Fu, and P. Heng, "H-DenseUnet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2663–2674, 2018.

[17] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," *Deep Learning in Medical Image Analysis DLMIA*, vol. 10008, pp. 179–187, 2016.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *International Conference on Machine Learning ICML*, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 770–778, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *European Conference on Computer Vision ECCV*, vol. 9908, pp. 630–645, 2016.

[21] Zh. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 3126–3135, 2019.

[22] 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, pp. 630–644, 2016.