DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation

Debesh Jha*†, Michael A. Riegler*, Dag Johansen†, Pål Halvorsen*‡, Håvard D. Johansen†

*SimulaMet, Norway †UiT The Arctic University of Norway, Norway ‡Oslo Metropolitan University, Norway
Email: debesh@simula.no

***Abstract*—Semantic image segmentation is the process of labeling each pixel of an image with its corresponding class. An encoder-decoder based approach, like U-Net and its variants, is a popular strategy for solving medical image segmentation tasks. To improve the performance of U-Net on various segmentation tasks, we propose a novel architecture called DoubleU-Net, which is a combination of two U-Net architectures stacked on top of each other. The first U-Net uses a pre-trained VGG-19 as the encoder, which has already learned features from ImageNet and can be transferred to another task easily. To capture more semantic information efficiently, we added another U-Net at the bottom. We also adopt Atrous Spatial Pyramid Pooling (ASPP) to capture contextual information within the network. We have evaluated DoubleU-Net using four medical segmentation datasets, covering various imaging modalities such as colonoscopy, dermoscopy, and microscopy. Experiments on the 2015 MICCAI sub-challenge on automatic polyp detection dataset, the CVC-ClinicDB, the 2018 Data Science Bowl challenge, and the Lesion boundary segmentation datasets demonstrate that the DoubleU-Net outperforms U-Net and the baseline models. Moreover, DoubleU-Net produces more accurate segmentation masks, especially in the case of the CVC-ClinicDB and 2015 MICCAI sub-challenge on automatic polyp detection dataset, which have challenging images such as smaller and flat polyps. These results show the improvement over the existing U-Net model. The encouraging results, produced on various medical image segmentation datasets, show that DoubleU-Net can be used as a strong baseline for both medical image segmentation and cross-dataset evaluation testing to measure the generalizability of Deep Learning (DL) models.**

***Index Terms*—semantic segmentation, convolutional neural network, U-Net, DoubleU-Net, CVC-ClinicDB, ETIS-Larib, ASPP, 2015 MICCAI sub-challenge on automatic polyp detection, 2018 Data Science Bowl, Lesion Boundary Segmentation challenge**

I. Introduction

Medical image segmentation is the task of labeling each pixel of an object of interest in medical images. It is often a key task for clinical applications, varying from Computer Aided Diagnosis (CADx) for lesions detection to therapy planning and guidance [1]. Medical image segmentation helps clinicians focus on a particular area of the disease and extract detailed information for a more accurate diagnosis. The key challenges associated with medical image segmentation are the unavailability of a large number of annotated, lack of high-quality labeled images for training [2], low image quality, lack of a standard segmentation protocol, and a large variations of images among patients [3]. The quantification of segmentation accuracy and uncertainty is essential to estimate the performance on other applications [1]. This indicates the requirement for an automatic, generalizable, and efficient semantic image segmentation approach.

Convolutional Neural Networks (CNNs) have shown state-of-the-art performance for automated medical image segmentation [4]. For semantic segmentation tasks, one of the earlier Deep Learning (DL) architecture trained end-to-end for pixel-wise prediction is a Fully Convolutional Network (FCN). U-Net [5] is another popular image segmentation architecture trained end-to-end for pixel-wise prediction. The U-Net architecture consists of two parts, namely, analysis path and synthesis path. In the analysis path, deep features are learned, and in the synthesis path, segmentation is performed based on the learned features. Additionally, U-Net uses skip connections operation. The skip connection allows propagating dense feature maps from the analysis path to the corresponding layers in the synthesis part. In this way, the spatial information is applied to the deeper layer, which significantly produces a more accurate output segmentation map. Thus, adding more layers to the U-Net will allow the network to learn more representative features leading to better output segmentation masks.

Generalization, i.e., the ability of the model to perform in an independent dataset, and robustness, i.e., the ability of the model to perform on challenging images, are keys for the development of Artificial Intelligence (AI) system to be used in clinical trials [6]. Therefore, it is essential to design a powerful architecture that is robust and generalizable across different biomedical applications. Pre-trained ImageNet [7] models have significantly improved the performance of the CNN architectures. One of the examples of such models trained on ImageNet is VGG19 [8]. Inspired by the success of U-Net and its variants for medical image segmentation, we propose an architecture that uses modified U-Net and VGG-19 in the encoder part of the network. Because we use two U-Net architectures in the network, we term the architecture as **DoubleU-Net**. The main reasons for using the VGG network are: (1) VGG-19 is a lightweight model as compared to other pre-trained models, (2) the architecture of VGG-19 is similar to U-Net, making it easy to concatenate with U-Net, and (3) it will allow much deeper networks for producing better output segmentation mask. Thus, we aim to improve the overall segmentation performance of the network by enabling this architectural changes.

The main contributions of this work are:

- We propose a novel architecture, DoubleU-Net, for semantic image segmentation. The proposed architecture uses two U-Net architecture in sequence, with two encoders and two decoders. The first encoder used in the network is pre-trained VGG-19 [8], which is trained on ImageNet [7]. Additionally, we use Atrous Spatial Pyramid Pooling (ASPP) [9]. The rest of the architecture is built from scratch.
- Experiments on multiple datasets are prerequisites for showing the enhancement of the proposed algorithm over other algorithms. In this respect, we have experimented on four different medical imaging datasets, two different datasets from colonoscopy, one from dermoscopy, and one from microscopy. DoubleU-Net shows better segmentation performance as compared to baseline algorithms on 2015 MICCAI sub-challenge on automatic polyp detection dataset, CVC-ClinicDB dataset, Lesion Boundary Segmentation challenge from ISIC-2018, and 2018 Data Science Bowl challenge dataset.
- An extensive evaluation of DoubleU-Net across four dataset shows a significant improvement over U-Net. Therefore, DoubleU-Net can be a new baseline for medical image segmentation task.

The paper is organized into seven sections. Section II provides an overview of the related work in the field of medical image segmentation. In Section III, we describe the proposed architecture. Section IV describes the experiments. Section V presents the results obtained from the experimental evaluation on different datasets. A discussion of the work is provided in Section VI. Finally, we summarize the paper and discuss future work and limitations in Section VII.

II. Related Work

Among different CNN architectures, an encoder-decoder network like FCN [10] and its extension U-Net [5] have gained significant popularity among semantic segmentation approach for 2D images. Badrinarayan et al. [11] proposed a deep fully CNN for semantic pixel-wise segmentation that has significantly fewer parameters and produces good segmentation maps. Yu et al. [12] proposed a new convolutional network module that particularly targeted dense prediction problems. The proposed module used dilated convolutions for systematically aggregating multi-scale contextual information, and the presented context module improved the accuracy for state-of-the-art semantic image segmentation systems.

Chen et al. [13] proposed DeepLab to solve segmentation problem. Later, DeeplabV3 [9] significantly improved over their previous DeepLab versions without DenseCRF post-processing. The DeepLabV3 architecture uses a synthesis path that contains the fewer number of convolutional layers that are different from the synthesis path of FCN and U-Net. DeepLabV3 uses skip connection between analysis path and synthesis path similar to U-Net architecture. Zhao et al. [14] proposed effective scenes parsing network for complex scene understanding, where global pyramidal features provide an opportunity to capture additional contextual information. Zhang et al. [15] proposed Deep Residual U-Net, which uses residual connections better output segmentation map. Chen et al. [16] proposed Dense-Res-Inception Net (DRINET) for medical image segmentation and compared their results with FCN, U-Net, and ResUNet. Ibtehaz et al. [17] modified U-Net and proposed an improved MultiResUNet architecture for medical image segmentation where they compared their results with U-Net on various medical image segmentation datasets and showed superior accuracy than U-Net.

Jha et al. [18] proposed ResUNet++, which is an enhanced version of standard ResUNet by integrating an additional layer such as squeeze-and-excite block, ASPP, and attention block to the network. The proposed architecture uses dice loss as the loss function and produces an improved output segmentation maps as compared to U-Net and ResUNet on the Kvasir-SEG [2] and CVC-ClinicDB [19] datasets. Zhou et al. [20] proposed UNet++, a neural network architectures for semantic and instance segmentation tasks. They improved the performance of UNet++ by alleviating the unknown network depth, redesigning the skip connections, and devising a pruning scheme to the architecture.

From the above-related work, we can observe that there has been substantial efforts toward developing deep CNN architectures for the segmentation of both natural and medical images. Recently, more works are focused on developing generalizable models, which is why most of the researchers test their algorithms on different datasets [17], [18], [20]. The accuracy achieved is now is high for both natural imaging [13] and medical imaging [17], [18], [20]. However, AI in medicine is still an emerging field. One of the significant challenges in the medical domain is the lack of test datasets. Moreover, the obtained datasets are often imbalanced. To some extent, we can say that the performance is acceptable in the case of natural images. In the medical imaging, there are many challenging images (for example, flat polyps in colonoscopy), which are usually missed out during colonoscopy examination and can develop into cancer if early detection is not performed. Therefore, there is a need for a more accurate medical image segmentation approach to deal with the challenging images. Toward addressing this need, we have proposed DoubleU-Net architecture that produces efficient output segmentation masks with the challenging images.

III. The DoubleU-Net Architecture

Figure 1 shows an overview of the proposed architecture. As seen from the figure, DoubleU-Net starts with a VGG-19 as encoder sub-network, which is followed by decoder sub-network. What distinguishes DoubleU-Net from U-Net in the first network (NETWORK 1) is the use of VGG-19 marked in yellow, ASPP marked in blue, and decoder block marked in light green. The squeeze-and-excite block [21] is used in the encoder of NETWORK 1 and decoder blocks of NETWORK 1 and NETWORK 2. An element-wise multiplication is performed between the output of NETWORK 1 with the input of the same network. The difference between DoubleU-Net and

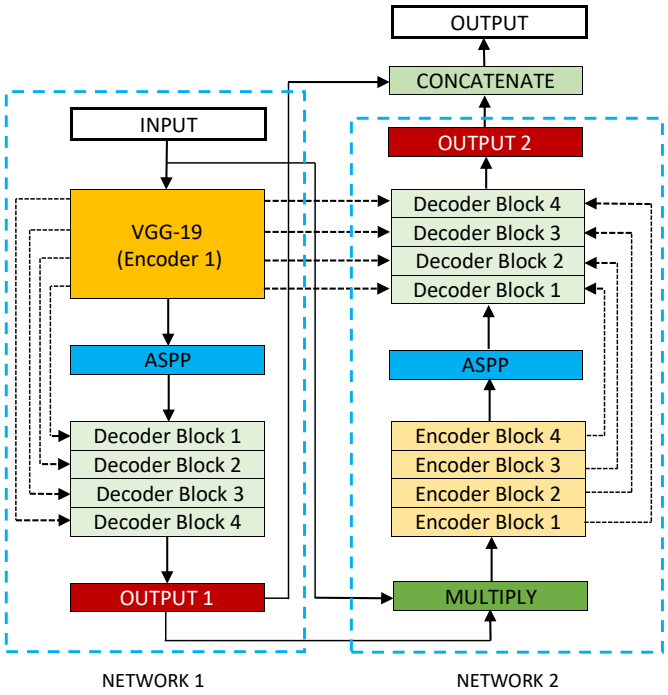

Fig. 1: Block diagram of the proposed DoubleU-Net architecture

U-Net in the second network (NETWORK 2) is only the use of ASPP and squeeze-and-excite block. All other components remain the same.

In the NETWORK 1, the input image is fed to the modified U-Net, which generates a predicted mask ($Output1$). We then multiply the input image and the produced mask ($Output1$), which acts as an input for the second modified U-Net that produces another mask ($Output2$). Finally, we concatenate both the masks ($Output1$ and $Output2$) to see the qualitative difference between the intermediate mask ($Output1$) and final predicted mask ($Output2$).

We assume that the produced output feature map from NETWORK 1 can still be improved by fetching the input image and its corresponding mask again, and concatenating with $Output2$ will produce a better segmentation mask than the previous one. This is the main motivation behind using two U-Net architectures in the proposed architecture. The squeeze-and-excite block in the proposed networks reduces the redundant information and passes the most relevant information. ASPP has been a popular choice for modern segmentation architecture because it helps to extract high-resolution feature maps that lead to superior performance [18].

A. Encoder Explanation

The first encoder in DoubleU-Net ($encoder1$) uses pre-trained VGG-19, whereas the second encoder ($encoder2$), is built from scratch. Each encoder tries to encode the information contained in the input image. Each encoder block in the $encoder2$ performs two 3×3 convolution operation, each followed by a batch normalization. The batch normalization reduces the internal co-variant shift and also regularizes the model. A Rectified Linear Unit (ReLU) activation function is applied, which introduces non-linearity into the model. This is followed by a squeeze-and- excitation block, which enhances the quality of the feature maps. After that, max-pooling is performed with a 2×2 window and stride 2 to reduce the spatial dimension of the feature maps.

B. Decoder Explanation

As shown in Figure 1, we use two decoders in the entire network, with small modifications on the decoder as compared with that of the original U-Net. Each block in the decoder performs a 2×2 bi-linear up-sampling on the input feature, which doubles the dimension of the input feature maps. Now, we concatenate the appropriate skip connections feature maps from the encoder to the output feature maps. In the first decoder, we only use skip connection from the first encoder, but in the second decoder, we use skip connection from both the encoders, which maintains the spatial resolution and enhance the quality of the output feature maps. After concatenation, we again perform two 3×3 convolution operation, each of which is followed by batch normalization and then by a ReLU activation function. After that, we use a squeeze and excitation block. At last, we apply a convolution layer with a sigmoid activation function, which is used to generate the mask for the corresponding modified U-Net.

IV. Experiments

In this section, we present datasets, evaluation metrics, experiment setup and configuration, and data augmentation techniques used in all the experiments to validate the proposed framework.

A. Datasets

To evaluate the effectiveness of the DoubleU-Net, we have used four publicly available datasets from medical domain.

- The 2015 MICCAI sub-challenge on automatic polyp detection [22] used the CVC-ClinicDB [19] for training and ETIS-Larib [23] for testing in the case of polyp detection task. The 2015 MICCAI sub-challenge on automatic polyp detection dataset is the first dataset used in our study.
- Similarly, CVC-ClinicDB has been a common choice for polyp segmentation. Therefore, we use this dataset for comparison.
- The third dataset used in our experiment is from the ISIC-2018 challenge, namely, Lesion Boundary Segmentation dataset [24], [25]. The dataset contains skin lesions and their corresponding annotations.
- The fourth dataset used in this study is nuclei segmentation, from the 2018 Data Science Bowl challenge[1]. This dataset is publicly available at Broad Bioimage Benchmark Collection[2].

[1]https://www.kaggle.com/c/data-science-bowl-2018
[2]https://data.broadinstitute.org/bbbc/BBBC038/

TABLE I: Summary of biomedical segmentation dataset used in our experiments

Dataset	No. of Images	Input size	Application
2015 MICCAI sub-challenge on automatic polyp detection dataset	808	384×288	Colonoscopy
CVC-ClinicDB	612	384×288	Colonoscopy
Lesion Boundary Segmentation challenge	2594	Variable	Dermoscopy
2018 Data Science Bowl Challenge	670	256×256	Nuclei

More information about the datasets are presented in Table I. All of the datasets are clinically relevant during diagnosis, and therefore, their segmentation can be crucial for patient outcome.

B. Evaluation metrics

DoubleU-Net is evaluated on the basis of Sørensendice coefficient (DSC), mean Intersection over Union (mIoU), Precision, and Recall. We evaluate all of these metrics for all four datasets. However, we compare and emphasize more on the official evaluation metrics that were used in the challenge. For example, the official evaluation metrics for the Lesion Boundary Segmentation challenge is mIoU.

C. Experiment setup and configuration

All models are implemented using Keras framework [26] with Tensorflow 2.1.0 [27] as backend. The implementation can be found at our GitHub repository[3]. We ran our experiments on a Volta 100 GPU and an Nvidia DGX-2 AI system. In all of the datasets, we used 80% of dataset for training, 10% for validation, and 10% for testing. During training, we used the original image size for the smaller dataset, such as CVC-ClinicDB and Nuclei segmentation dataset, and resized the images to 384×512 for the Lesion Boundary segmentation challenge dataset to balance between training time and complexity. The size of ETIS-Larib was adjusted similarly to that of CVC-ClinicDB. We use binary cross-entropy as the loss function for all the networks and the Nadam optimizer with its default parameters. For the lesion boundary segmentation dataset and the Nuclei segmentation dataset, where dice loss and Adam optimizer performed slightly higher, the batch size is set to 16 and the learning rate to $1e{-}5$. All models are trained for 300 epochs. Early stopping and ReduceLROnPlateau is also used.

D. Data augmentation techniques

Medical datasets are challenging to obtain and annotate [2]. Most existing datasets have only a few samples, which makes training DL models on these datasets challenging. One potential solution to the challenge of data insufficiency, is to use data augmentation techniques that increase the number of samples during training. For this, we first split the dataset into training, validation, and testing sets. We then apply different data augmentation methods to each set, including center crop, random rotation, transpose, elastic transform, etc. More details about the augmentation techniques we used can be found in our GitHub repository. A single image was converted into 25 different images; thus, in total, 26 images including the original image. The same augmentation techniques were applied to all four datasets.

[3]https://github.com/DebeshJha/2020-CBMS-DoubleU-Net

TABLE II: Experimental results using the 2015 MICCAI sub-challenge on automatic polyp detection dataset

Method	DSC	mIoU	Recall	Precision
FCN-VGG [28]	0.7023	0.5420	-	-
Mask R-CNN with Resnet101 [29]	0.7042	0.6124	-	-
U-Net	0.2920	0.1759	0.5930	0.2021
DoubleU-Net	**0.7649**	**0.6255**	**0.7156**	**0.8007**

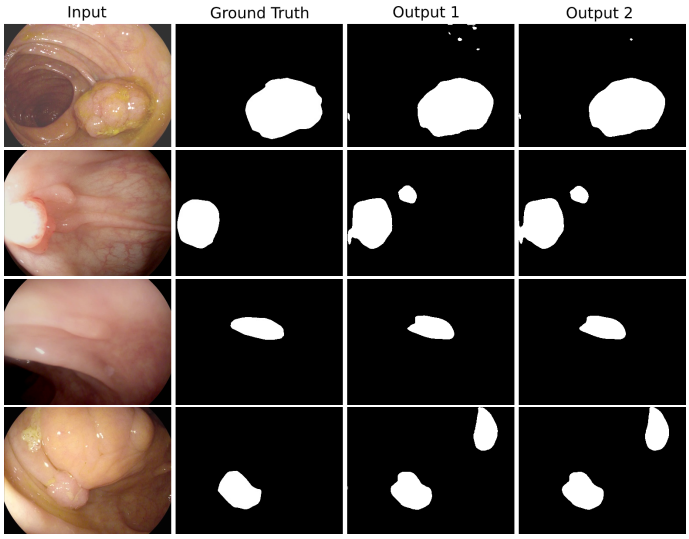

Fig. 2: Qualitative result of DoubleU-Net on large, medium, and flat polyps from 2015 MICCAI sub-challenge on automatic polyp detection dataset

V. Results

In this section, we present the results and compare them with the baselines on the respective datasets. U-Net is still considered as the baseline for various medical image segmentation tasks. Therefore, we compare the proposed model with U-Net by using the same data augmentation techniques as described above to demonstrate its effectiveness. We also report the results on four datasets and show the qualitative results to prove the usefulness of DoubleU-Net. In all of the figures demonstrating the qualitative results, the sequence of input, ground truth, $Output1$, and $Output2$ are followed, where $Output1$ and $Output2$ are the intermediate and final output respectively.

A. Comparison on 2015 MICCAI sub-challenge on automatic polyp detection dataset

Our quantitative results on the 2015 MICCAI sub-challenge on automatic polyp detection dataset are summarized in Table II. The experimental results shows that DoubleU-Net achieved a DSC of 0.7649 and a mIoU of 0.6255. From

TABLE III: Result comparison on CVC-ClinicDB

Method	DSC	mIoU	Recall	Precision
Fully Convoutional Network [30]	-	-	0.7732	0.8999
CNN [31]	(0.62-0.87)	-	-	-
SegNet [32]	-	-	**0.8824**	-
Multi-scale patch-based CNN [33]	0.8130	-	0.7860	0.8090
MultiResUNet with data augmentation [17]	-	0.8497	-	-
Conditional generative adversarial network [34]	0.8848	0.8127	-	-
U-Net	0.8781	0.7881	0.7865	0.9329
DoubleU-Net	**0.9239**	**0.8611**	0.8457	**0.9592**

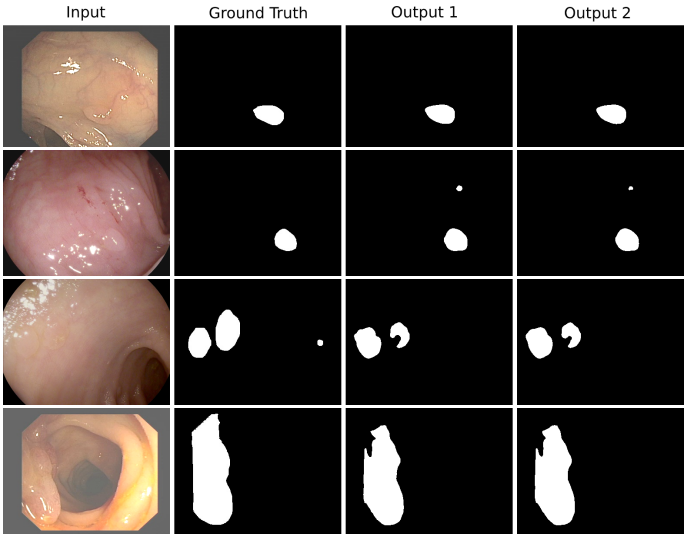

Fig. 3: Qualitative result of DoubleU-Net on challenging images from CVC-ClinicDB

Table II, we can see that DoubleU-Net outperforms the baseline [29] by 6.07% in terms of DSC and 1.31% in mIoU. From the above table, we can also observe that the model that uses a pre-trained ImageNet network (for instance, Resnet101 or VGG-16) as a backbone achieves a higher score on cross-dataset evaluation as compared to that of training a network from scratch (see Table II). The visual results of the proposed model can be seen in Figure 2. From the visual analysis, we can observe that the segmentation mask produced by $Output2$ is better than that of $Output1$. This also justifies the significance of the proposed model over U-Net.

B. Comparison on CVC-ClinicDB

DoubleU-Net is compared with U-Net and the recent works that used the same dataset for evaluation. Table III shows the results on CVC-ClinicDB dataset. The evaluation results shows that DoubleU-Net achieve a DSC of 0.9239 which is 3.91% higher than [34] and mIoU of 0.8611, which is 1.14% higher than [17]. A careful visual analysis of the result shows that DoubleU-Net produces better segmentation masks as compared to the intermediate network. The model performs reasonably well on the challenging images such as flat and small polyps, which are usually missed-out during colonoscopy examinations (see Figure 3).

TABLE IV: Result on Lesion boundary segmentation dataset from ISIC-2018

Method	DSC	mIoU	Recall	Precision
U-Net [17]	-	0.7642 ± 0.4518	-	-
Multi-ResUNet [17]	-	0.8029 ± 0.3717	-	-
DoubleU-Net	**0.8962**	**0.8212**	**0.8780**	**0.9459**

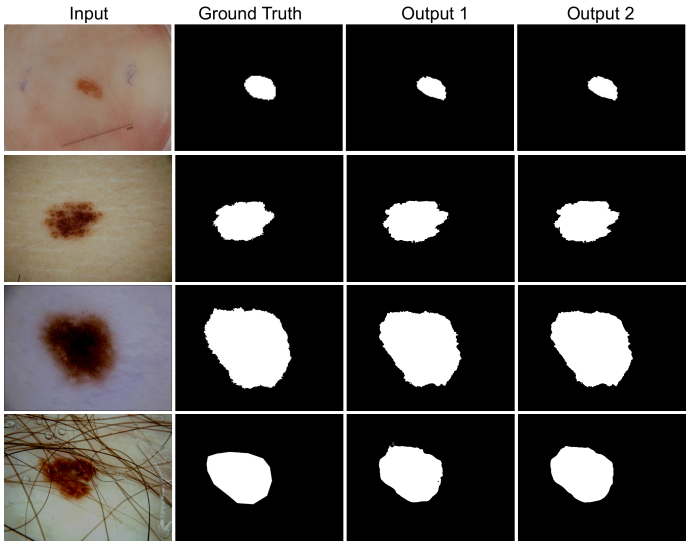

Fig. 4: Qualitative result of DoubleU-Net on small, medium and large size skin lesions from Lesion Boundary segmentation challenge

C. Comparison on Lesion Boundary segmentation challenge dataset

The official evaluation metric for the challenge was mIoU. DoubleU-Net achieve a DSC of 0.8962 and mIoU of 0.8212 on this challenge dataset. From the quantitative results comparison (see Table IV), we can see that the DoubleU-Net outperforms U-Net [17] by an approximate margin of 5.7%, and Multi-ResUNet [17] by an approximate margin of 1.83% in terms of mIoU on Lesion boundary segmentation challenge dataset from ISIC-2018. Figure 4 shows the qualitative results. From the figure, we can see that both intermediate output and the final output produced by the network perform well on all types of lesions ranging from small to medium to large lesions. However, a careful analysis shows that the final output produced by the network is better than the intermediate one.

D. Comparison on 2018 Data Science Bowl challenge dataset

Table V and Figure 5 presents the quantitative and qualitative results on 2018 Data Science Bowl challenge dataset. We have compared our work with U-Net++ [20]. Our method produced a DSC of 0.9133, which is 1.59% higher than the method proposed by Zhou et. al [20], and comparable mIoU with U-Net and UNet++ that uses Resnet101 as the backbone model. UNet++ has been used as a strong baseline for result comparison over various image segmentation tasks.

TABLE V: Result on Nuclei segmentation from 2018 Data Science Bowl challenge

Method	Pre-trained network	DSC	mIoU	Recall	Precision
U-Net [20]	Resnet101	0.7573	0.9103	-	-
UNet++ [20]	Resnet101	0.8974	**0.9255**	-	-
DoubleU-Net	VGG-19	**0.9133**	0.8407	**0.6407**	**0.9496**

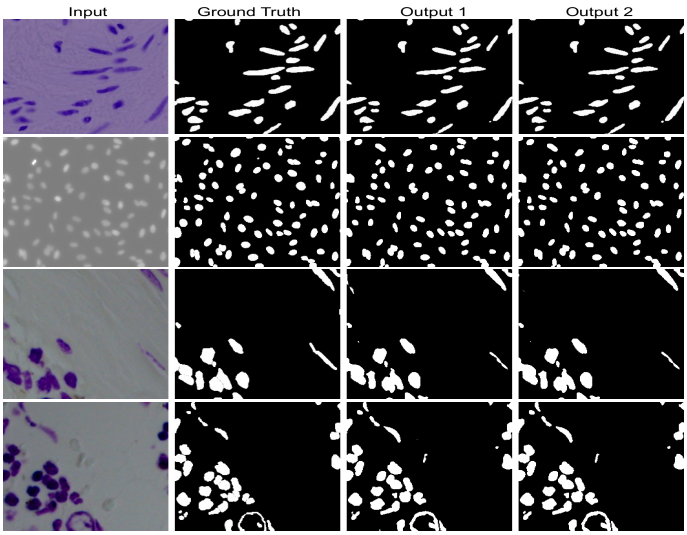

Fig. 5: Qualitative result of DoubleU-Net on Nuclei images from 2018 Data Science Bowl challenge dataset

Therefore, the DoubleU-Net set a new baseline for semantic image segmentation task.

VI. Discussion

Table VI shows the DSC comparison of U-Net and DoubleU-Net. From the above table, we can see that DoubleU-Net performs reasonably well as compared to U-Net for all the presented datasets. For the CVC-ClinicDB dataset, the performance of U-Net is competitive. However, for 2015 MICCAI sub-challenge on automatic polyp detection dataset and the 2018 Data Science Bowl, DoubleU-Net has a significant DSC improvement of 0.4729% and 15.60% respectively. Additionally, the 2015 MICCAI sub-challenge on automatic polyp detection dataset provides us the opportunity to study the cross-data generalizability, which is critical in the medical domain [35]. The generalization test showed that DoubleU-Net outperforms its competitors (see Table II). From the Table, we observe that the model trained on pre-trained ImageNet [7] performs much better on the cross-dataset test than that of the model trained from scratch. We have trained U-Net on the CVC-ClinicDB dataset, which is competitive with DoubleU-Net when tested on the same dataset (see Table III). The same model was used to test against the ETIS-Larib dataset. However, the performance of the U-Net was poor as compared to that of DoubleU-Net (see Table II). This fact suggests that DoubleU-Net is more generalizable and can be used for the cross-dataset test across the different domains.

TABLE VI: Relative improvement of DoubleU-Net on U-Net

Modality	U-Net (DSC)	DoubleU-Net (DSC)	Overall Improvement
Colonoscopy (MICCAI 2015)	0.2920	0.7649	0.4729
Colonoscopy (CVC-ClinicDB)	0.8781	0.9239	0.0458
Dermoscopy (ISIC-2018)	–	0.8962	–
Microscopy (2018 Data Science Bowl)	0.7573	0.9133	0.1560

From the qualitative results, we can see that DoubleU-Net is capable of producing better segmentation mask even for the challenging images. This can be observed from Figure 2 and Figure 3. Moreover, Figure 4 and Figure 5 show that the model produces high-quality segmentation masks for Lesion Boundary Segmentation challenge dataset and 2018 Data Science Bowl challenge dataset. The overall qualitative result shows that the model performs well for different multi-organ and multi-centered medical image segmentation datasets. Thus, the above results suggest that the robustness of the proposed model.

From the above experiments, we observed that the transfer learning from a pre-trained ImageNet network significantly improves the results on every dataset, which tries to compensate for the lack of enough training data. The qualitative and quantitative results suggest using DoubleU-Net as a baseline for result comparisons over four medical image segmentation datasets.

VII. Conclusion

In this paper, we have proposed a novel CNN architecture called DoubleU-Net. The DoubleU-Net has five main components, namely two U-Net networks, VGG-19, a squeeze-and-excite block and ASPP. The performance of DoubleU-Net is significantly better when compared with the baselines and U-Net on all four datasets.

Moreover, the proposed architecture is flexible, and that makes it possible to integrate other CNN blocks into DoubleU-Net. We believe that the segmentation results can be improved by further integrating different CNN blocks and by the use of post-processing techniques such as conditional random field and Otsu threshold.

In the future, we plan to research building one model for different medical image segmentation tasks and focus on simplifying the architecture while retaining its ability to produce high segmentation masks. A limitation of the DoubleU-Net is that it uses more parameters as compared to U-Net, which leads to an increase in the training time. In the future, the research should focus more on designing simplified architectures with fewer parameters while maintaining its ability.

Acknowledgement

This work is funded in part by Research Council of Norway project number 263248 (Privaton). The computations in this paper were performed 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] M. Lê, J. Unkelbach, N. Ayache, and H. Delingette, "Gpssi: Gaussian process for sampling segmentations of images," in *Proceeding of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 38–46.

[2] 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 (MMM)*, 2020, pp. 451–462.

[3] F. Zhao and X. Xie, "An overview of interactive medical image segmentation," *Annals of the BMVA*, vol. 2013, no. 7, pp. 1–22, 2013.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis (MedIA)*, vol. 42, pp. 60–88, 2017.

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

[6] T. Ross, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. M. Filimon, P. Scholz, T. N. Tran *et al.*, "Robust medical instrument segmentation challenge 2019," *arXiv preprint arXiv:2003.10299*, 2020.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009, pp. 248–255.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[10] 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 (CVPR)*, 2015, pp. 3431–3440.

[11] V. Badrinarayanan, 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, no. 12, pp. 2481–2495, 2017.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 2881–2890.

[15] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[16] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Drinet for medical image segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2453–2462, 2018.

[17] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[18] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *Proceeding of IEEE International Symposium on Multimedia (ISM)*, 2019, pp. 225–2255.

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

[20] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, 2019.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of computer vision and pattern recognition (CVPR)*, 2018, pp. 7132–7141.

[22] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[23] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[24] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 168–172.

[25] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[26] F. Chollet *et al.*, "Keras," 2015.

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceeding of {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI})*, 2016, pp. 265–283.

[28] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 2017, pp. 101 340F1 – 101 340F1.

[29] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better?" in *Proceeding of International Symposium on Medical Information and Communication Technology (ISMICT)*, 2019, pp. 1–6.

[30] Q. Li, G. Yang, Z. Chen, B. Huang, L. Chen, D. Xu, X. Zhou, S. Zhong, H. Zhang, and T. Wang, "Colorectal polyp segmentation using a fully convolutional neural network," in *Proceeding of International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2017, pp. 1–5.

[31] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *Proceeding of International Conference on Artificial Intelligence and Knowledge Engineering*, 2018, pp. 208–211.

[32] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature biomedical engineering*, vol. 2, no. 10, pp. 741–748, 2018.

[33] D. Banik, D. Bhattacharjee, and M. Nasipuri, "A multi-scale patch-based deep learning system for polyp segmentation," in *Advanced Computing and Systems for Security*, 2020, pp. 109–119.

[34] J. Poomeshwaran, K. S. Santhosh, K. Ram, J. Joseph, and M. Sivaprakasam, "Polyp segmentation using generative adversarial network," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019, pp. 7201–7204.

[35] V. Thambawita, D. Jha, H. L. Hammer, H. D. Johansen, D. Johansen, P. Halvorsen, and M. A. Riegler, "An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 3, 2020.