Retinal OCT Image Segmentation: How Well do Algorithms Generalize or How Transferable are the Data?

M. Melinščak[1,2] and S. Lončarić [2]
[1] Karlovac University of Applied Sciences, Karlovac, Croatia
[2] Faculty of Electrical Engineering and Computing, Zagreb, Croatia
martina.melinscak@gmail.com

***Abstract* - The success of deep learning depends, among other things, on a large amount of labeled data. However, in medical applications, large labeled datasets are the exception, rather than the rule. Manual image labeling is time-consuming and is generally performed only with the purpose of developing algorithms, and not as a part of standard clinical practice. The goal of this study is twofold. Since there is always a trade-off between the ability to collect data and achieve the best possible performance, we wanted to explore how performance depends on the amount of data. For this purpose, a database of manually annotated OCT images was collected. Also, we wanted to see how much transfer learning can help. Retinal OCT images vary depending on the type of device, therefore developed methods should be as robust as possible. Transfer learning was performed so that the model was trained with similar OCT images and then fine-tuned with images from the collected database. It has been shown that transfer learning helps in terms of generalization and better prediction in case the source database is similar to the target database. We can also assume that further improvement can probably be achieved by adding images from another distribution (medical or nonmedical).**

Keywords* – deep learning*; retinal OCT images; image segmentation;

I. INTRODUCTION

Age-related macular degeneration (AMD) is one of the most common diseases leading to vision loss in the elderly population. Currently, 170 million people worldwide are suffering from it and, by some estimates, 288 million will be affected by 2040 [1]. Due to the disease, pathological biomarkers such as intraretinal fluids appear and there is a significant disturbance in the distribution and thickness of the retinal layers [2]. Since the introduction of optical coherence tomography (OCT), imaging has been done using OCT devices. Diagnostics and treatment in clinical practice rely only on the qualitative analysis and assessment of the ophthalmologist. For more successful diagnostics and therapy, computer segmentation is therefore necessary in quantitative analysis. It should also contribute to a better understanding of the disease that has not yet been fully explored.

Recently, the segmentation of medical and thus ophthalmic images is performed almost exclusively using deep learning methods. One of the preconditions for the success of deep learning algorithms is a large amount of data available. But in medicine, this is rarely the case. In the case of supervised learning, images manually annotated by an expert are required, which is very time-consuming. Even more, image labeling is not part of clinical practice, and labeling is done solely to develop machine learning methods. As images from different OCT devices and different device generations vary, it is required that the segmentation algorithms be as robust as possible. This means that in case of migration to a new device it is not necessary to re-annotate a larger number of images, and existing algorithms will be upgraded with a smaller number of new images.

Although transfer learning is often applied in medicine, by using the known architectures of neural networks (Xception, Inception V3, ResNet50, VGG16, VGG19, MobileNet) pretrained on the ImageNet dataset, in ophthalmology this is rarely the case [3]. Several papers [4]–[6] examine the adequacy of such a method since ImageNet dataset is a database of natural images that bears no resemblance to medical ones. In a recent paper [7], a hybrid architecture was proposed in which only the first few layers of networks such as the ResNet architecture (pretrained on ImageNet) would be used and then some lighter versions of CNN would be applied (and fine-tuned) in the end.

The aim of this study was twofold. First, to see how the performance of the algorithm depends on the amount of data. To begin with, the network was trained from scratch with a different amount of data to see how performance varies. Segmentation was done using the U-net architecture [8], which is almost standard in the segmentation of medical images. In order to perform the segmentation of retinal structures and pathological biomarkers, a database of images with manually labeled features was collected. And second, we explored how many additional labeled images are needed and how much transfer learning can help in the case the network was trained on one dataset (source data) and then migrated to another (target data). As a source dataset, the image database from the RETOUCH challenge (The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge) was used [9]. Transfer learning was then performed so that the network was trained with part of the

images from RETOUCH challenge and then fine-tuned with the images of the target database: again, with a different number of images.

II. MATERIAL AND METHODS

A. *Database*

For the research, a database of manually annotated images was collected. For convenience, let it be called the target dataset. In collaboration with Sisters of Charity Hospital (KBC Sestre milosrdnice), Zagreb, images were annotated for 25 patients with neovascular age-related macular degeneration (nAMD). Macular SD-OCT volumes were recorded with the Zeiss Cirrus HD OCT 4000 device. Each OCT volume consisted of 128 B-scans with a resolution of 1024 x 512 pixels (pixel size 1.96 x 11.74 μm). Retinal fluids were annotated for 1224 B-scans. Of the intraretinal fluids, the following were annotated: pigment epithelial detachment (PED), subretinal fluid (SRF), and intraretinal fluid (IRF) (Fig. 1).

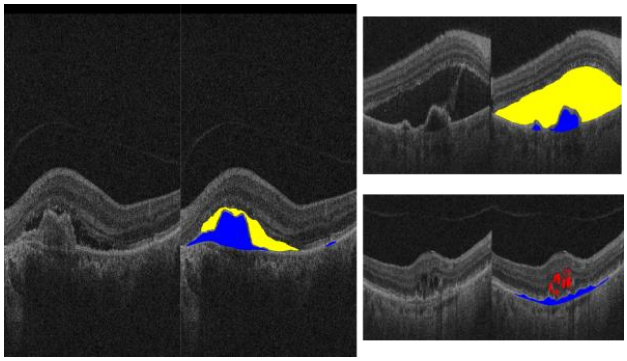

Figure 1. Examples of B-scans from a dataset with manual annotations of retinal fluids. PED is colored with blue, SRF with yellow, and IRF with red. On the left there is an example of an entire B-scan (1024x512 pixels). It is visible that there is a large amount of background and thus a large class imbalance. Examples cropped to the region of interest (ROI) are on the right.

Data collection adhered to the tenets of the Declaration of Helsinki and the standards of Good Scientific Practice of Sisters of Charity Hospital. The presented study was approved by the Ethics Committee of the Sisters of Charity Hospital and Faculty of Electrical Engineering and Computing, Zagreb, Croatia.

To estimate human error, for 75 B-scans, the same expert made a re-annotation with a time delay (and without insight into former annotations). Also, annotations were made by another expert. An intra-observer and an inter-observer error were calculated. Dice scores for inter-observer and intra-observer errors are 0.822 and 0.889, respectively. Fig. 2 shows two examples. It became evident that there are some even major differences in the opinions of experts. The pathology of the disease is not sufficiently known, and it is often difficult to decide what type of fluid was present, especially since in addition to intraretinal fluids there are other pathological phenomena (hyperreflective foci, druse, cysts, etc.). Due to the poor quality of images (a large amount of speckle-noise), and the necessary prior knowledge of retinal anatomy and pathological changes, it is difficult to expect great accuracy (especially in the regime of small data) as it is a demanding task even for humans.

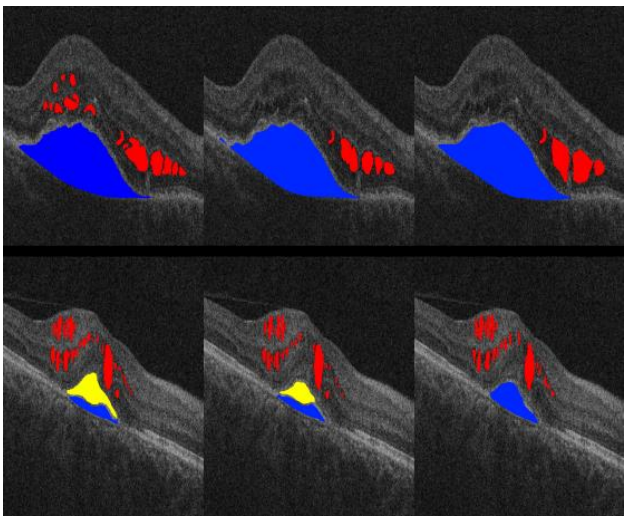

Figure 2. Two examples of B-scans from a dataset with manual annotations (from left to right): annotation from the 1st expert (considered as ground truth); annotation form the 1st expert with time delay and with no insight into previous annotation (used to calculate intra-observer error) and annotation from the 2nd expert (used to calculate inter-observer error). All images are cropped to show just ROI.

The other dataset we used is from the RETOUCH challenge [9] – let it be called source dataset. The database consists of images from three types of devices (Topcon, Spectralis, and Cirrus). For each type of device, all B-scans for 24 patients were labeled. We used only part of the images from the Cirrus device (1000 randomly picked images) to train the network. Unlike the target dataset, where all B-scans contained intraretinal fluids, many B-scans did not contain pathological changes (whole B-scans belonged to the background class). It resulted in the even greater disproportion of background pixels and pixels belonging to fluids. The database included images of patients with AMD, retinal vein occlusion (RVO), and diabetic macular edema (DME). More details can be found on the website [9].

B. *Segmentation approach*

For segmentation we used the U-net architecture [8]. The U-net architecture, its various modifications [10], stacking of multiple U-net architectures, and combining the U-net architecture with other machine learning methods [11], [12] have become standard methods for segmenting OCT images in recent years. Eight teams participated in the RETOUCH challenge in 2017, and the results and algorithms of all participants were presented in [13].

To make it easier to see the dependence of the algorithm's performance on the amount of data, the problem was reduced to binary segmentation (all three fluids belonged to the same class). Data augmentation, which is usually used in the case of a small amount of data, hadn't been performed, in order to better isolate the dependence of network performance on the amount of data.

In the first case, the target dataset was divided into a training set (1000 images) and a test set (224 images). Furthermore, the training set was divided into three groups with a different number of images: 256, 512, and 1000. The network was trained from scratch. Dice loss was used to train the model, since the dice score is usually used as a metric for evaluation, and the goal was to maximize it. We tested the model training with binary cross-entropy loss, but worse results were obtained. The batch size was set to 4. Adam optimizer with learning rate 1e-5 was used. The model had 7759521 trainable parameters. Training was performed during 70 epochs, without early stopping. Dropout was used to prevent overfitting. The validation dataset was 10% of the training dataset.

In the second case, the network training was performed with the source dataset. The same hyperparameters were used. The network had only been trained for 20 epochs, as the purpose was only to have a pretrained network. Fine-tuning of all layers with the same three data groups from the target dataset (256, 512, and 1000) was then performed. Training was executed during 50 epochs with all the same parameters as in the first case.

The model was trained on Google Colab [14] with a GPU. The code was written in Keras with the TensorFlow backend. Code is available on GitHub: https://github.com/mmelinscak/MIPRO_2020

III. RESULTS

In the first case, when the model is trained from scratch, as expected, the performance is better due to the larger number of images (it is well known that deep learning algorithms are "data-hungry"). Obtained dice scores (mean ± SEM) on a test dataset (224 images) for 256, 512, and 1000 images are: 0.579±0.272, 0.632±0.279, and 0.731±0.268, respectively.

Fig. 3 shows examples for one raw image and ground truth mask, followed by model predictions on the test set in case the model is trained with 256, 512, and 1000 training images. As it is as expected, segmentations are better with a larger amount of data. With an experiment like this, we can get a better insight into how much data we need, since there is always a trade-off between the amount of manually annotated data we can/want to collect and the performance of algorithms. The second row shows segmentation predictions for the second case (with the model pretrained on source dataset and then fine-tuned on target dataset). In this case, obtained dice scores (mean ± SEM) on the test dataset were 0.612±0.284, 0.677±0.272, and 0.826±0.270 respectively. It is evident that predictions after transfer learning have more "serrated" shapes while predictions from a model trained from scratch are more oval. As shape types are one of the key features in distinguishing different kinds of intraretinal fluids, this is a promising result.

Fig. 4 shows a graph with mean dice scores for all cases. It is visible that the mean value of the dice score is higher in the second case (with transfer learning).

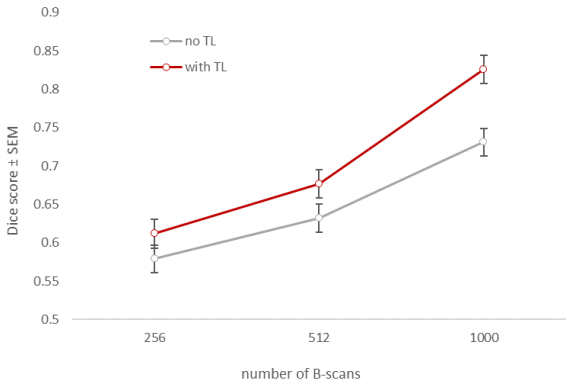

Figure 4. Mean dice score with error bars (standard error of the mean – SEM) for training model from scratch (no TL) and in case of fine-tuning the pretrained model (with TL).

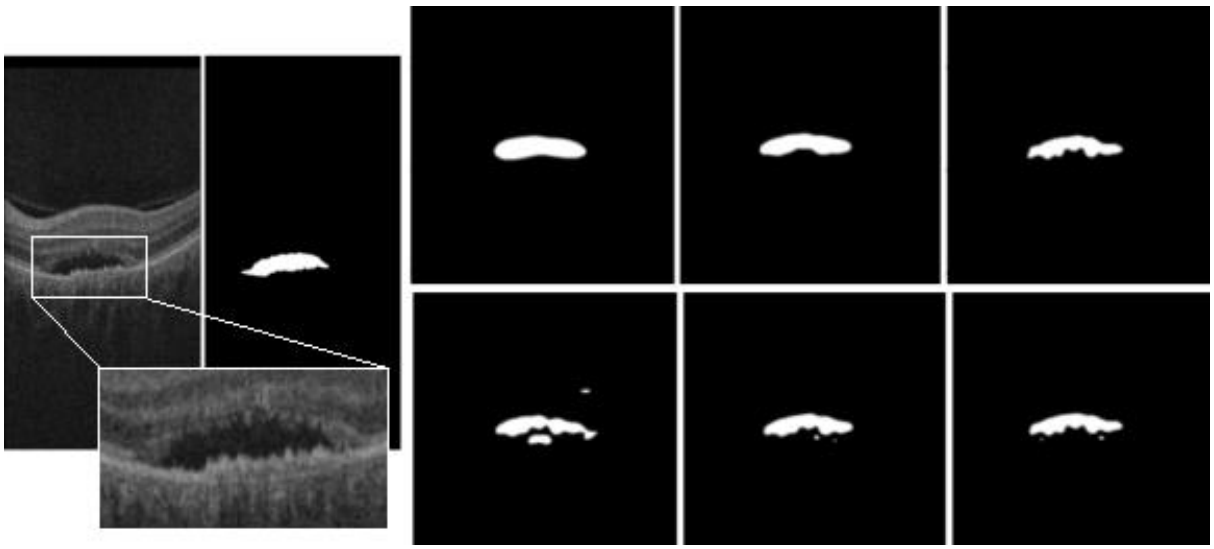

Figure 3. First row (from left to right): raw image from test dataset, ground truth (manual annotation), prediction from a network trained with 256 images, 512 images, and 1000 images. Second row (from left to right): predictions after training the network with 1000 images from source dataset and fine-tuning pretrained network with 256 images, 512 images, 1000 images (from left to right) from target data.

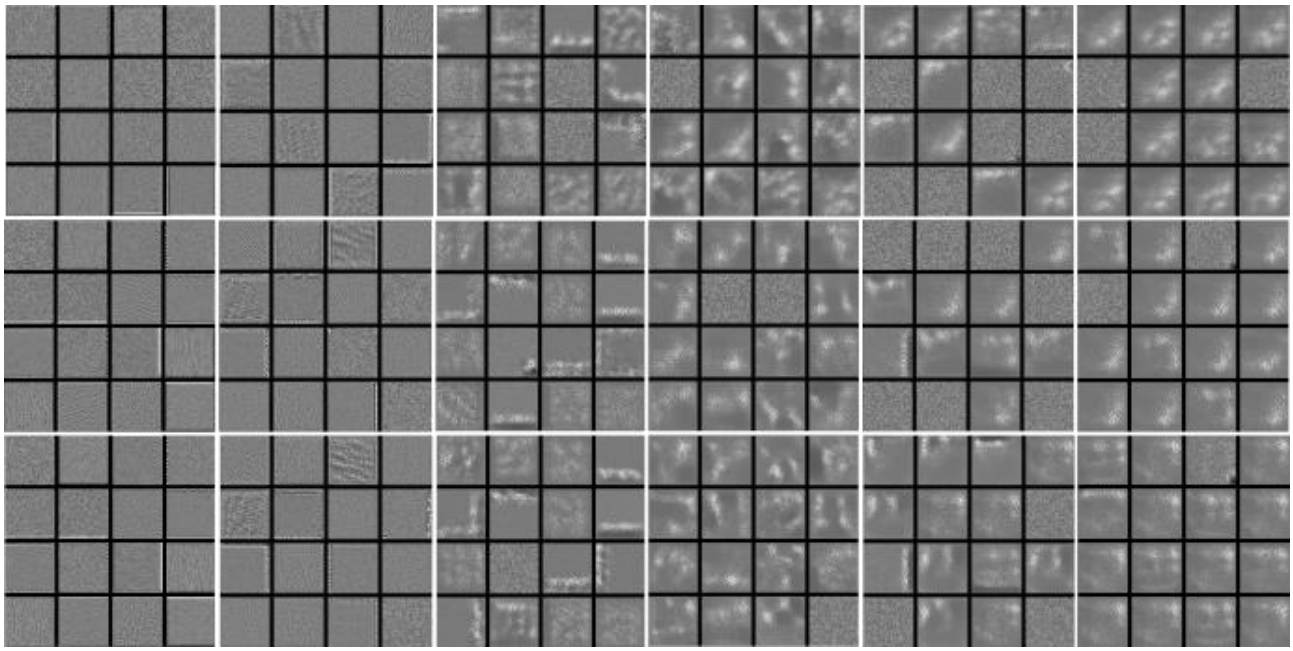

Figure 5. Visualization of convolution layers 2, 4, 7, 10, 13, 16 (from left to right). Layers 2, 4, and 7 belong to the encoder part of U-net architecture, and 10, 13, and 16 to the decoder part. First row: network trained from a scratch with 256 images (target data). Second row: network trained with 1000 images (source data). Third row: network fine-tuned with target data (256 images) after learning with source data (1000 images).

Although the model trained with 1000 images from the source dataset achieved a high value for dice score on training test (0.717), the value of the dice score is very low for the test set (0.348). The reason is probably that both datasets cannot be considered to fit the same distributions despite the same type of OCT device being used in both. One of the reasons are the differences in annotations (we have seen that there is quite a significant inconsistency between expert opinions). Another reason may be that the source dataset contains images of patients with three types of disease, and the target dataset only contains patients with AMD. Therefore, we have a case of a model trained on data from one distribution and tested on a different image distribution.

Fig. 5 shows a visualization of convolution layers in case of a training model from scratch with 256 images (first row). Then for the pretrained model (second row), and after fine-tuning the pretrained model with 256 images (third row). First three layers belong to the encoder and the last three layers to the decoder. It is visible that in the third row, convolution layers appear to be a combination from convolution layers shown in the first two rows, as might be expected. It is evident that convolution layers are clearer and more completed after fine-tuning with additional 256 target images then after only training with 1000 source images. It can be concluded that using images from different distributions leads to better results than increasing the number of images from the same distribution. However, the generality of this finding requires further investigation. Also, it is visible that the first layers of the encoder which should look like Gabor filters are not very clear, and maybe it would help to initialize the first few convolution layers with Gabor filters, and then to see if freezing or fine-tuning them leads to better results.

IV. CONLUSION

In this study, we wanted to examine how the performance of deep learning segmentation approaches depends on the amount of data. Since medical labeled images are difficult to collect, we found it useful to get an insight into the performance dependence of the algorithm on the number of images. Furthermore, we investigated how transfer learning affects the robustness of the algorithm and how much new data needs to be collected if the model is not trained from scratch but fine-tuned after it was pretrained on existing data.

Segmentation was performed with the basic U-net architecture. No data augmentation or other common ways of increasing performance were used to get a better result, since in the case of more complex architecture and changing multiple parameters it would be difficult to assess the dependence only on the amount of data.

It was demonstrated that training the model on very similar OCT images had a low dice score on the test set despite the relatively high dice score on the training dataset. This means that in the case of migration to a new OCT device, it would still be necessary to additionally train the model, but with a smaller number of images than in case of training the model from scratch.

Transfer learning is often applied in medicine, by using the known architectures of neural networks pretrained on the ImageNet dataset. ImageNet is a database with over 14 million natural images. Intuitively, it is not clear how would transfer learning help in case of medical images which are often grayscale and dissimilar to natural images. Optimizing transfer learning (with model pretrained on ImageNet) is still an active area of research in the case of medical images.

By visualizing convolutional filters, it becomes visible what the network is actually learning. In the study, it has been demonstrated that the initial layers, that are usually similar to Gabor filters, are not clear even after transfer learning when the source database is similar. Our assumption is that better generalization and prediction is possible by adding images from another distribution (medical and/or non-medical images). In addition to adding images from another distribution, our assumption is that incorporating some prior knowledge would help. Mapping from human to computer methods is not always completely identical but can help and serve as a useful analogy. Ophthalmologists use some general features of vision developed during life, some specific features learned about ophthalmic images, and use knowledge of retinal anatomy and pathological changes. We surmise that optimal results might be achieved by similar combinations: learning on a large number of images such as ImageNet, learning on the ophthalmic images themselves, and finally by adding some prior knowledge. Further research will test these assumptions.

ACKNOWLEDGMENT

We would like to thank the Sisters of Charity Hospital, the Clinic for eye diseases (Zagreb, Croatia). Special thanks to Prof. Zoran Vatavuk, MD, and to Marin Radmilović, MD who did image labeling.

REFERENCES

[1] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," Prog. Retin. Eye Res., Aug. 2018, doi: 10.1016/j.preteyeres.2018.07.004.

[2] P. Mitchell, G. Liew, B. Gopinath, and T. Y. Wong, "Age-related macular degeneration," *The Lancet*, vol. 392, no. 10153, pp. 1147–1159, Sep. 2018, doi: 10.1016/S0140-6736(18)31550-2.

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[4] J. Lee, P. Sattigeri, and G. Wornell, "Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks," p. 11.

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *ArXiv14111792 Cs*, Nov. 2014, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1411.1792.

[6] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet Pre-training," *ArXiv181108883 Cs*, Nov. 2018, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1811.08883.

[7] M. Raghu, J. Kleinberg, C. Zhang, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," p. 11.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *ArXiv150504597 Cs*, May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597.

[9] "RETOUCH - MICCAI 2017 Workshop," Dec. 13, 2017. https://retouch.grand-challenge.org/workshop/ (accessed Dec. 13, 2017).

[10] A. G. Roy *et al.*, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Express*, vol. 8, no. 8, p. 3627, Aug. 2017, doi: 10.1364/BOE.8.003627.

[11] D. Lu, M. Heisler, S. Lee, G. Ding, M. V. Sarunic, and M. F. Beg, "Retinal Fluid Segmentation and Detection in Optical Coherence Tomography Images using Fully Convolutional Neural Network," *ArXiv171004778 Cs*, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.04778.

[12] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, "Pathological OCT Retinal Layer Segmentation Using Branch Residual U-Shape Networks," in Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017, vol. 10435, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 294–301.

[13] H. Bogunovic *et al.*, "RETOUCH -The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge," *IEEE Trans. Med. Imaging*, pp. 1–1, 2019, doi: 10.1109/TMI.2019.2901398.

[14] "Google Colaboratory." https://colab.research.google.com (accessed May 17, 2020).