Pathological OCT Retinal Layer Segmentation using Branch Residual U-shape Networks

Stefanos Apostolopoulos[1], Sandro De Zanet[2], Carlos Ciller[2], Sebastian Wolf[3], and Raphael Sznitman[1]

[1]University of Bern, Switzerland
[2]RetinAI Medical GmbH, Switzerland
[3]Bern University Hospital, Switzerland

Abstract

The automatic segmentation of retinal layer structures enables clinically-relevant quantification and monitoring of eye disorders over time in OCT imaging. Eyes with late-stage diseases are particularly challenging to segment, as their shape is highly warped due to pathological biomarkers. In this context, we propose a novel fully-Convolutional Neural Network (CNN) architecture which combines dilated residual blocks in an asymmetric U-shape configuration, and can segment multiple layers of highly pathological eyes in one shot. We validate our approach on a dataset of late-stage AMD patients and demonstrate lower computational costs and higher performance compared to other state-of-the-art methods.

1 Introduction

Optical Coherence Tomography (OCT) is a non-invasive medical imaging modality that provides micrometer-resolution volumetric scans of biological tissue [8]. Since its introduction in 1991, OCT has seen widespread use in the field of ophthalmology, as it enables direct, non-invasive imaging of the retinal layers. As shown in Fig. 1, OCT allows for the visualization of both healthy tissue and pathological biomarkers such as drusen, cysts and fluid pockets within and underneath the retinal layers. Critically, these have been linked to diseases such as Age-related Macular Degeneration (AMD), Diabetic Retinopathy (DR) and Central Serous Chorioretinopathy (CSC) [1, 13].

Given the widespread occurrence of these diseases, which is estimated at over 300 million people worldwide, medical image analysis methods for OCT imaging have gained popularity in recent years. The automatic segmentation of retinal layer structures is of particular interest as it allows for the quantification, characterization and monitoring of retinal disorders over time. This remains a challenging task, as retinal layers can be heavily distorted in the presence of pathological biomarkers. In this context, the present paper focuses on providing more accurate retinal layer segmentations in pathological eyes, at clinically-relevant speeds.

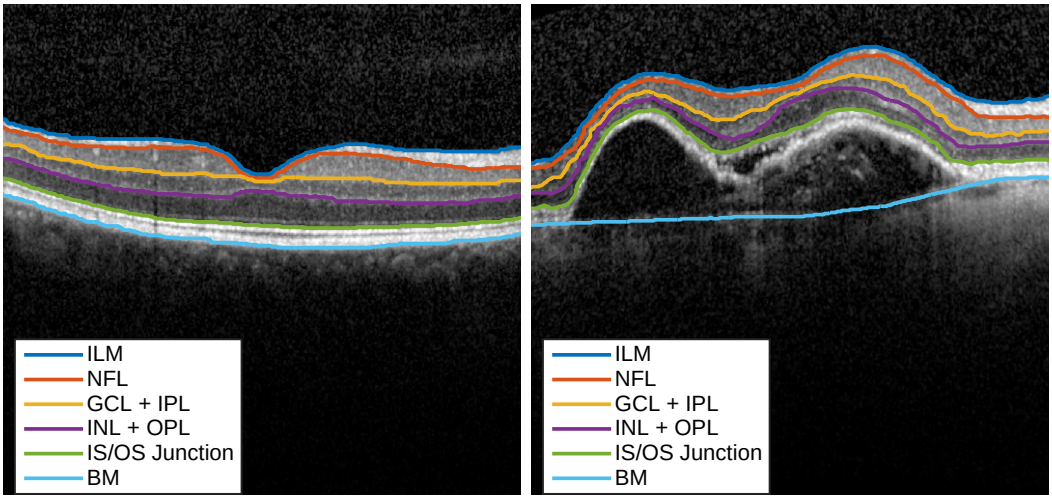

Figure 1: Example of OCT cross-sections with retinal layer boundaries highlighted for (left) healthy subject and (right) late-stage AMD patient. The images were manually annotated by an expert ophthalmologist.

A number of relevant methods on this topic can be found in the literature. Mayer et al. [12] propose the use of a series of edge filters and denoising steps to extract layers in OCT cross-sections. In [5], a Markov Random Field (MRF)-based optimization with soft constraints is proposed to segment 7 retinal layers using volumetric information. Chen et al. [4] use a constrained graph-cut approach to segment layers and quantify fluid pockets in pathological OCTs. Overall, most of these methods face difficulties in segmenting all retinal layers accurately for subjects with pathological eyes.

To this end, we present a novel strategy to overcome the above limitations and provide accurate results in a wider range of cases. Inspired by recent CNN approaches for semantic segmentation [14] and image classification [6], we introduce a novel CNN architecture that learns to segment retinal layers as a supervised regression problem. Our proposed network combines residual building blocks with dilated convolutions into an asymmetric U-shape configuration, and can segment multiple layers of highly pathological eyes in one shot. Using lower computational resources, our strategy achieves superior segmentation performance compared to both state-of-the-art deep learning architectures and other OCT segmentation methods.

2 Methods

Our goal is to segment retinal cell layers in OCT images. The main challenge in this task stems primarily from the highly variable and irregular shape of pathological eyes, and secondarily from the variable image quality (i. e., signal strength and speckle noise) of clinical OCT scans. Due to the image acquisition process, wherein each cross-section, or *Bscan* is acquired separately without a guaranteed global alignment, we opt to segment retinal layers at the Bscan level. This avoids the need for computationally intensive 3-dimensional convolutions [3] and volumetric pre-processing (i. e., registration and alignment).

In our approach, we treat the task of segmenting retinal layers as a regression problem.

Given a Bscan image, $\mathcal{I}$, we wish to find a function $T : \mathcal{I} \rightarrow \mathcal{L}$, that maps each pixel in $\mathcal{I}$ to a label $\mathcal{L} \in \{0, 1, 2, 3, 4, 5, 6\}$ corresponding to an anatomical retinal cell layer region. As in [5], we consider the following six retinal layers: (1) Internal Limiting Membrane (ILM) to Nerve Fibre Layer (NFL), (2) NFL to Ganglion Cell Layer (GCL), (3) GCL and Inner Plexiform Layer (IPL), (4) Inner Nuclear Layer (INL) and Outer Plexiform Layer (OPL), (5) OPL to Inner Segment/Outer Segment (IS/OS) Junction and (6) IS/OS Junction to Bruch's Membrane (BM).

2.1 Branch Residual U-Network

Fully convolutional U-net style networks have established themselves as the state-of-the-art for binary segmentation and have been successfully used in a variety of biomedical applications [14]. In such architectures, input images are convolved and downsampled level by level with exponentially increasing numbers of filters up to a predefined depth (*descending branch*), from which they are subsequently upsampled and convolved to the original size (*ascending branch*). Skip connections from corresponding levels transfer information from the descending to the ascending branch.

A number of important limits arise from this architecture. First, the largest possible object that can be segmented is defined by the cumulative receptive field of the network. According to our experiments, a regular U-net with $3\,\text{px} \times 3\,\text{px}$ convolutions and a depth of 5 layers [14], will start exhibiting holes when segmenting objects with discontinuities wider than $3 * 2^5 = 96\,\text{px}$. Second, due to the exponential growth of trainable parameters, the maximum depth of such a network is limited to 5-7 layers before the computational demands become intractable. Third, the convergence rate of a U-net tends to decrease as the network grows in depth. We attribute this to the vanishing gradient problem that affects deeper networks.

We have designed our network to address each of these problems:

1. We use a building block based on dilated convolutions with dilation rates of $\{1, 3, 5\}$ to increase the effective receptive field of each network level and without increasing the number of trainable parameters. We enhance this block with residual connections [7, 17] and batch normalization [10], which are summed together with the dilated convolutions. Depending on the branch direction, each block ends with a max-pooling or upsampling operation. We denote those blocks as $Block_D$ and $Block_U$, respectively.

2. We insert bottleneck connections between blocks to control the number of trainable parameters [15, 16, 9]. Furthermore, we increase the number of filters based on a capped Fibonacci sequence. We chose this sequence after experimenting with zero, constant and quadratic growth, as a good trade-off between network capacity and segmentation performance.

3. Finally, we add connections from the input image, downscaled to the appropriate size, to all levels in the ascending and descending branches.

Combined, these result in a significant increase in the learning rate, segmentation accuracy and, due to the reduced number of parameters, processing speed. We name the

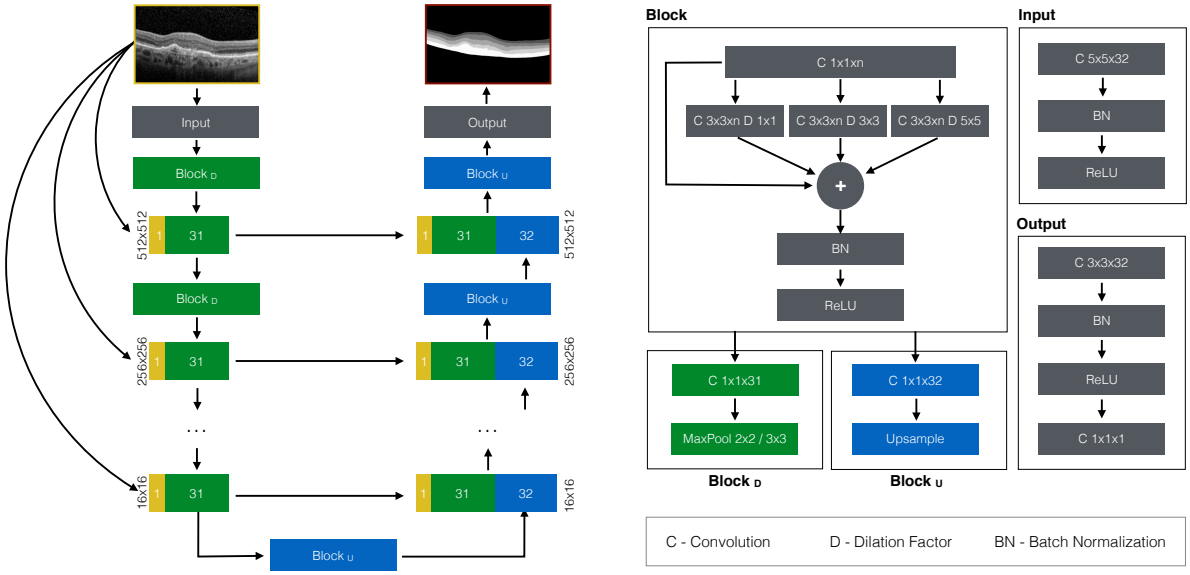

Figure 2: (left) Branch Residual U-Network (BRU-net). The descending branch takes a single Bscan as input and performs consecutive Block$_D$ operations. The ascending branch receives the output of the descending branch and performs consecutive Block$_U$ operations. The numbers indicate the number of filters output by each block. Skip connections connect each descending to each ascending level, while the original Bscan is provided to each level for context. The final output is a regressed layer class for each pixel of the input image.
(right) Block$_D$, Block$_U$, input and output blocks. The rectangles illustrate computations.

resulting architecture *Branch Residual U-shape Network* (BRU-net). The precise architecture and building blocks are illustrated in Fig. 2.

Throughout our network, we employ 3×3 convolutional kernels with n filters where n increases according to the Fibonacci sequence $\{32, 64, 96, 160, 256, 416\}$, capped to a maximum of 416 parameters per level. This avoids the larger growth of parameters encountered in traditional U-networks and allows for deeper networks. More specifically, our network requires 21 million parameters for a depth of 5 levels and grows to 55 million parameters for a depth of 6 levels. The corresponding U-net requires 44 million and 176 million parameters for the same depths, an increase of $2\times$ and $3\times$, respectively.

2.2 Training

The block layout has been optimized using an evolutionary grid search strategy, by training two variants in parallel and selecting the best performer. To keep training time reasonable, the grid search is performed on a $4\times$ subsampled dataset. This process is repeated 50 times, each one taking up to 30 minutes.

To increase convergence rate and reduce training time, we pre-initialize our network by training it as an autoencoder for 10 epochs, using a small set of 50 OCT volumes of healthy eyes, acquired from the same OCT device. This set is distinct from the volumes we use for segmentation. To avoid learning the identity function, we disable the skip connections of the network during this process.

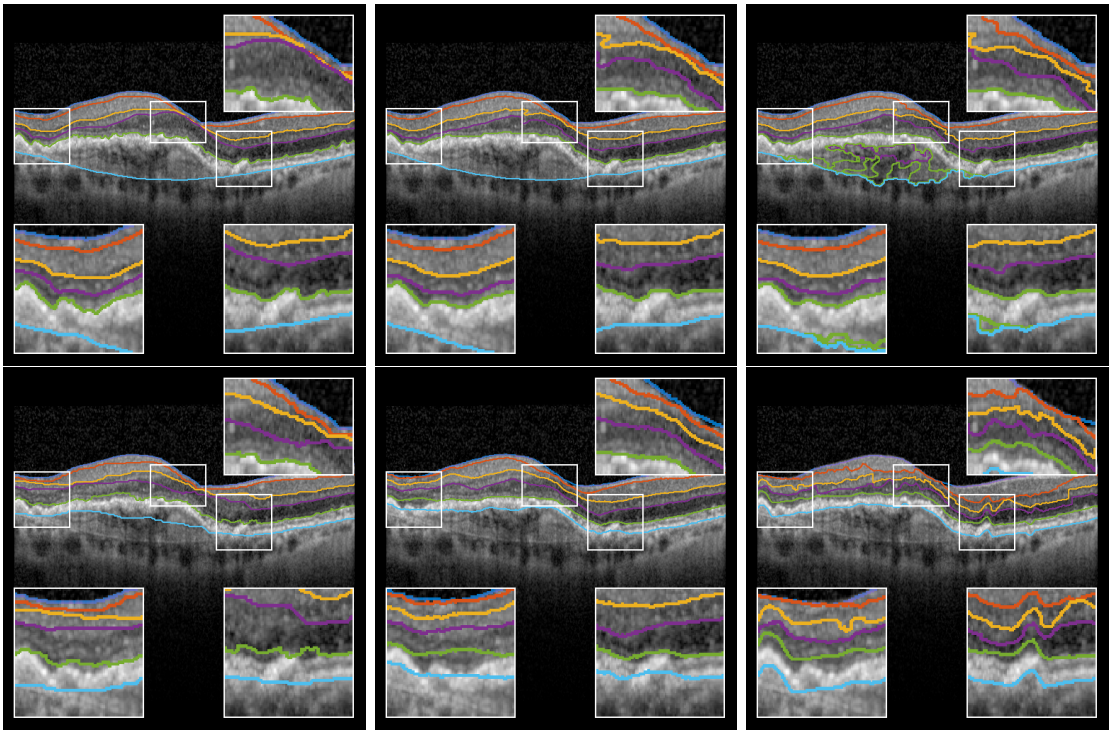

Figure 3: Qualitative comparison of each segmentation approach. Top row, left to right: ground truth, BRU-net, U-net; bottom row, left-to-right: Dufour et al., Chen et al., Mayer et al. Only BRU-net is able to segment the BM layer under the pathological region. The smaller receptive field of U-net results in discontinuities. Further qualitative results are provided in the supplementary material.

The output of the network is an image with the same size as the input Bscan. Each pixel of the output image is assigned a value between 0 and 6 which corresponds to the identity of its corresponding retinal layer. We train the network to minimize the pixel-wise Mean Square Error (MSE) loss between the predicted segmentation and the ground truth. This loss penalizes anatomically implausible segmentations (e.g. class 6 next to 0) more than plausible ones (e.g. class 1 next to 0). We rely this asymmetry to ensure segmentation continuity.i. e., The network parameters are updated via back-propagation and the Adam optimization process with the infinity norm. [11].

Each fold is trained for a maximum of 150 epochs. We start training with an initial learning rate of 10^{-3} and reduce it by a factor of 2 if the MSE loss does not improve for 5 consecutive epochs, down to a minimum of 10^{-7}. We interrupt the training early if the MSE loss stops improving for 25 consecutive epochs. Using a dedicated validation set, comprising 10% of the training set, we evaluate the MSE loss to adaptively set the learning rate and perform early stopping. At the end of the training procedure, we use the network weights of the epoch with the lowest validation loss to evaluate images in the test set.

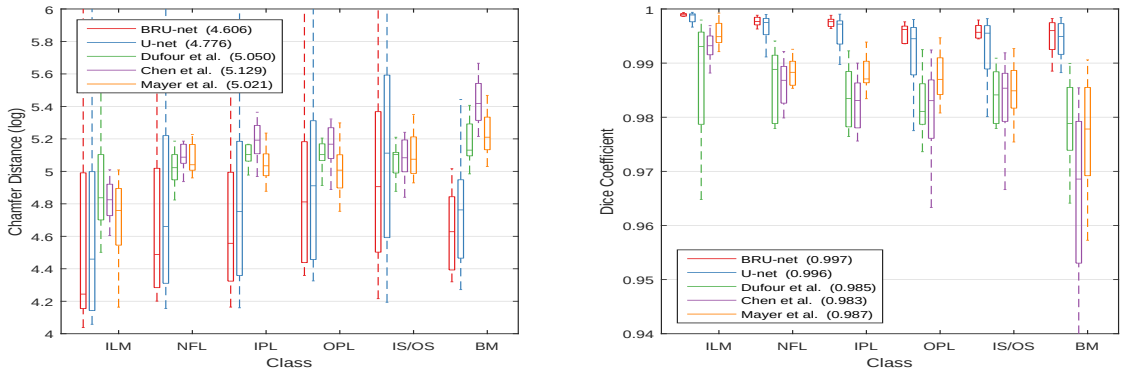

Figure 4: Quantitative comparison of segmentation accuracy per layer, using (left) the Chamfer distance error and (right) the Dice score.

3 Experimental Results

A trained ophthalmologist collected 20 macular OCT volumes from pathological subjects using a Heidelberg Spectralis OCT device (Heidelberg Engineering AG, Heidelberg, Germany). Each volume comprises 49, 512×496 Bscans, with a lateral (x-y) resolution of $15\,\mu m$ and an axial (z-) resolution of $3.9\,\mu m$. No volumes or Bscans were removed from our initial acquisition, to maintain the complete range of image quality observed in the clinic. For each Bscan in each volume, manually segmented ground truth layers were provided by the ophthalmologist.

We split our dataset into 5 equally sized subsets, each using 16 patients for training and 4 for testing. We repeat this process for each of those subsets for a 5-fold cross-validation. In each fold, the training set contains 784 training samples (Bscans), which we double to 1568 by flipping horizontally, taking advantage of the bilateral symmetry of the eye. The Bscans are first padded with a black border to a size of 512×512 pixels and then augmented with affine transformations, additive noise, Gaussian blur and gamma adjustments. Training is performed on batches of 8 Bscans at a time. Finally, the output image is quantized to integer values (0 to 6) without further post-processing.

To evaluate BRU-net, we compare it with the 3D methods of Dufour [5], Chen et. al. [4], and the 2D method of Mayer et. al. [12] on the same dataset. Additionally, we train a traditional U-net configuration [14] using the procedure described above. Fig. 3 provides a qualitative comparison of the results.

To quantify those results, we make use of two metrics: (1) the Chamfer distance [2] between each ground truth layer boundary and the boundary produced by a given method and (2) the Dice score of each predicted layer surface. Note that BRU-net is not constrained to convex shapes. Since pathological retinal layers may be non-convex, other metrics that rely on pixel distances are ill-suited for this problem. Fig. 4 demonstrates the performance of each of the evaluated methods.

Fig. 5 displays the mean training and validation loss of the 5-folds over time for both BRU-net and U-net. In both the training and validation sets, BRU-net achieves faster convergence and slightly better MSE loss.

We evaluated the statistical significance of those results using paired t-tests between BRU-net and each baseline. The resulting p-values indicate statistically significant results

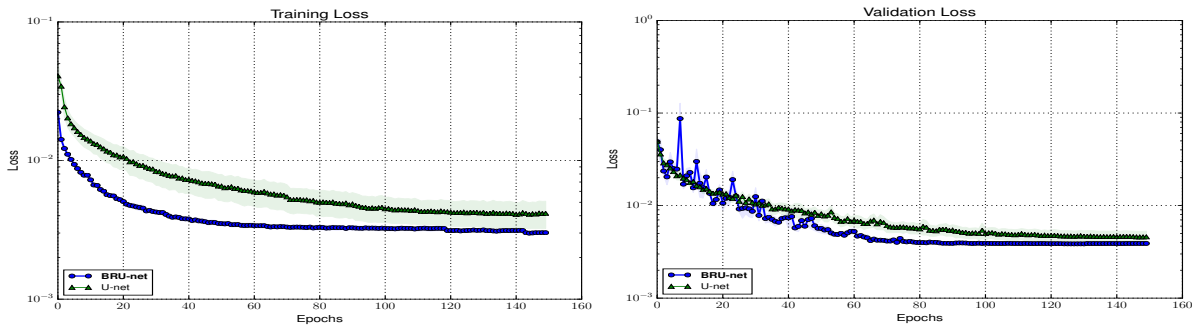

Figure 5: Training loss (left) and validation loss (right) comparison between BRU-net and U-net. BRU-net exhibits faster convergence speed and lower loss compared to U-net.

between BRU-net and every other baseline except U-net:

	U-net	Dufour et. al.	Chen et. al.	Mayer et. al.
p (Dice)	1.05e-01	2.03e-04	3.19e-05	1.08e-05
p (Chamfer)	1.09e-01	2.70e-04	3.18e-03	9.49e-03

Finally, we estimated the total runtime for each method. To process a single volume, BRU-net requires 5 s (Python), compared to 7 s for U-net (Python), 85 s for Mayer et al. (Matlab), 150 s for Dufour et al. (C++), and and 216 s for Chen et al (C++). The results were calculated on the same system using a 3.9 GHz Intel 6600K processor and a Nvidia 1080GTX GPU.

4 Conclusion

We have presented a method for performing layer segmentation on OCT scans of highly pathological retinas. Inspired by recent advances in computer vision, we have designed a novel fully-convolutional CNN architecture that can segment multiple layers in one shot. We have compared our method to several baselines and demonstrated qualitative and quantitative improvements in both segmentation accuracy and computational time on a dataset of late-stage AMD patients. Given the robustness of this approach on pathological cases, we plan to investigate how retinal layers change over time in the presence of specific diseases.

References

[1] M.D. Abramoff, M.K. Garvin, and M. Sonka. Retinal Imaging and Image Analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010.

[2] M. Butt and P. Maragos. Optimum design of chamfer distance transforms. *IEEE Transactions on Image Processing*, 7(10):1477–1484, Oct 1998.

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9901 LNCS, pages 424–432, 2016.

[4] X. Chen, M. Niemeijer, L. Zhang, K. Lee, M. D. Abramoff, and M. Sonka. Three-dimensional segmentation of fluid-associated abnormalities in retinal oct: Probability constrained graph-search-graph-cut. *IEEE Transactions on Medical Imaging*, 31(8):1521–1531, Aug 2012.

[5] Pascal A. Dufour, Lala Ceklic, Hannan Abdillahi, Simon Schroder, Sandro De Zanet, Ute Wolf-Schnurrbusch, and Jens Kowal. Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints. *IEEE Transactions on Medical Imaging*, 32(3):531–543, 2013.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Arxiv.Org*, 7(3):171–180, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, and James G Fujimoto. Optical Coherence Tomography HHS Public Access. *Science. November*, 22(2545035):1178–1181, 1991.

[9] Gao Huang, Zhuang Liu, and Kilian Q Weinberger. Densely Connected Convolutional Networks. *ArXiv preprint*, pages 1–12, 2016.

[10] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Arxiv*, pages 1–11, 2015.

[11] Diederik P Kingma and Jimmy Ba. Adam: {A} Method for Stochastic Optimization. *CoRR*, abs/1412.6980, 2014.

[12] Markus A. Mayer, Joachim Hornegger, Christian Y. Mardin, and Ralf P. Tornow. Retinal nerve fiber layer segmentation on fd-oct scans of normal subjects and glaucoma patients. *Biomed. Opt. Express*, 1(5):1358–1383, Dec 2010.

[13] Jessica I W Morgan. The fundus photo has met its match: Optical coherence tomography and adaptive optics ophthalmoscopy are here to stay, 2016.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015.

[15] C Szegedy, Wei Liu, Yangqing Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, jun 2015.

[16] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Arxiv*, page 12, 2016.

[17] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. 2016.