Dense Segmentation in Selected Dimensions: Application to Retinal Optical Coherence Tomography

Bart Liefers[1,2] BART.LIEFERS@RADBOUDUMC.NL
Cristina González-Gonzalo[1,2]
Caroline Klaver[3,4]
Bram van Ginneken[2]
Clara I. Sánchez[1,2,3]
[1] *A-Eye Research Group, Radboudumc, Nijmegen, The Netherlands*
[2] *Diagnostic Image Analysis Group, Radboudumc, Nijmegen, The Netherlands*
[3] *Department of Ophthalmology, Radboudumc, Nijmegen, The Netherlands*
[4] *Ophthalmology & Epidemiology, Erasmus MC, Rotterdam, The Netherlands*

Abstract

We present a novel convolutional neural network architecture designed for dense segmentation in a subset of the dimensions of the input data. The architecture takes an N-dimensional image as input, and produces a label for every pixel in M output dimensions, where $0 < M < N$. Large context is incorporated by an encoder-decoder structure, while funneling shortcut subnetworks provide precise localization. We demonstrate applicability of the architecture on two problems in retinal optical coherence tomography: segmentation of geographic atrophy and segmentation of retinal layers. Performance is compared against two baseline methods, that leave out either the encoder-decoder structure or the shortcut subnetworks. For segmentation of geographic atrophy, an average Dice score of 0.49 ± 0.21 was obtained, compared to 0.46 ± 0.22 and 0.28 ± 0.19 for the baseline methods, respectively. For the layer-segmentation task, the proposed architecture achieved a mean absolute error of 1.305 ± 0.547 pixels compared to 1.967 ± 0.841 and 2.166 ± 0.886 for the baseline methods.

Keywords: Segmentation, Retina, OCT

1. Introduction

Many applications of deep convolutional neural networks (CNNs) in medical imaging can be formulated as either a classification or a segmentation problem (Litjens et al., 2017). Classification problems can be defined as a mapping from an input of N spatial dimensions (usually N is 2 or 3, for 2D or 3D images) to an output without any spatial dimension, while for segmentation problems the N-dimensional input is mapped to an N-dimensional output. However, some medical applications require the projection of N-dimensional images to M-dimensional output, where $0 < M < N$. Problems that can be formulated as extracting a 2D manifold from a 3D volume include the detection of boundary surfaces or fissures, such as separating the cerebral hemispheres in MRI (Liang et al., 2007), segmenting pulmonary fissures in CT (Wang et al., 2006), or delineating the diaphragm in CT (Rangayyan et al., 2008). Similarly, a 1D line can be extracted from a 2D image (e.g. retinal layers (Fang et al., 2017)). Direct application of well-known classification or segmentation network architectures to this class of problems comes with limitations. Classification networks do not main-

tain any spatial information in their output values, so adapting them to produce a label for each pixel in the desired output dimensions is not feasible in practice. Segmentation networks, on the other hand, produce a dense N-dimensional output for all pixels, but there is no natural way to enforce output in just M-dimensions. This is problematic in case we do not have a label for all pixels.

In this paper we present a novel CNN architecture that produces dense predictions only in the desired output dimensions, maintaining spatial correlation. To the best of our knowledge, this is the first CNN architecture that is specifically designed for this task. Large context from N input dimensions is incorporated by an encoding network. A selective decoding is applied only in the required M output dimensions.

The encoder-decoder paradigm has been applied successfully to many segmentation problems in medical imaging, where U-Net (Ronneberger et al., 2015) (or a 3D variation (Çiçek et al., 2016)) is especially popular. In this network several layers of downsampling are applied to the input image in a contracting or encoding path. Subsequently, the original resolution is reconstructed in an expanding or decoding path that is connected via shortcut connections to the feature maps of the corresponding resolution in the contracting path. These shortcut connections are important for precise localization (Drozdzal et al., 2016).

In our network architecture, direct shortcut connections between the encoding and decoding blocks are not possible, because of a mismatch in the dimensions of the feature maps. This issue is resolved by adding funneling shortcut subnetworks between the encoder and decoder that contract only in the dimensions that are spliced out.

Although the proposed general formulation of the network architecture is applicable to any N and M ($0 < M < N$), we will focus in this paper on the case where $N = 2$ and $M = 1$. More specifically, we focus on application to retinal optical coherence tomography (OCT), where we predict a label for every vertical column (A-scan) in each slice of the OCT volume (B-scan). We demonstrate its applicability to two problems: segmentation of retinal layers, and segmentation of geographic atrophy (GA), and compare performance against two baseline models.

2. Background

The retina is a layered structure in the back of the eye that converts light into a neural response that provides vision. This layered structure can be visualized non-invasively using OCT, an imaging technique based on low coherence interferometry. OCT is commonly used to generate stacks of cross-sectional 2D images (B-scans), where every column in the image (an A-scan) represents a single acquisition entity. Often, the distance between slices is much larger than the distance between A-scans, so the acquired volume is highly anisotropic.

The proposed network architecture is particularly suitable for application to OCT, due to the nature of the image acquisition and the horizontally-layered structure of the retina. The pixel intensities within A-scans are strongly correlated, and the presence or absence of certain features can be identified for each of them, without specifying their location within the A-scan. In order to accurately classify each A-scan, contextual information from surrounding A-scans is usually required. Some examples in this category include segmentation of atrophic areas (Hu et al., 2013; Niu et al., 2016), localization of anatomical landmarks such as the fovea (Liefers et al., 2017) or segmentation of vessels (Tan et al., 2018). Another application typical for retinal OCT is the segmentation of retinal layers (Fang et al., 2017; Kugelman et al., 2018). These layers are horizontally stacked, so the boundary between them can be encoded as the vertical pixel index of the transition between two

layers. Hence, these problems can all be formulated as taking a 2D image as input and generating a label or a regression output for each column in the image, yielding a 1D output.

The two applications we focus on in this paper are segmentation of GA and retinal layers. GA occurs as an end stage in age-related macular degeneration (AMD) and is characterized by loss of retinal tissue and pigment. Next to absence of certain layers, this can be identified on OCT as a hyper-reflective area below the retina (refered to as hypertransmission). Segmentation of the individual layers of the retina allows to measure their thickness. Both applications can help ophthalmologists to diagnose disease status and decide treatment options.

3. Method

3.1. Data

Data for the application of the proposed model to segmentation of GA was collected from the Rotterdam Study (Hofman et al., 2007). This data set contains 55 OCT volumes acquired with a Topcon system, each containing 128 B-scans of either 512×885 pixels or 512×650 pixels (width $\times$ height). Manual annotations were made by four to five experienced graders in consensus using a multimodal annotation workstation. The graders did not directly annotate the OCT volume, but delineated the area of GA on 2D en-face retinal images (color fundus, and for some cases also autofluorescence or infrared). Image registration of these en-face images to the OCT volume was used to efficiently obtain a label (presence or absence of GA) for every A-scan in the volume. The data was split in a training set of 25 volumes, a validation set of 10 volumes, and a test set of 20 volumes.

For the layer-segmentation problem we used the publicly-available data set (Farsiu et al., 2014), containing annotations for three retinal layers: inner limiting membrane (ILM), retinal pigment epithelium drusen complex (RPEDC) and Bruch's membrane (BM). This data set contains OCT volumes for 269 AMD patients and 115 normal subjects. It was split in a training set (159 AMD, 5 normal), a validation set (10 AMD, 10 normal) and a test set (100 AMD, 100 normal). OCT volumes in this data set were acquired using a Bioptigen system, and contain 100 B-scans of 1000×512 pixels each.

3.2. Network architecture

The proposed network architecture follows an encoder-decoder structure, providing a large contextual window, with funneling subnetworks that provide local context. In this section we will focus on the case where the model takes a 2-dimensional input, and produces a label or a regression output for every vertical column in the image. This specific instance of the proposed generic network architecture is visualized in Figure 1.

The proposed architecture applies 9 levels of downsampling, which effectively reduces the input image from 512×512 pixels to 1×1 pixel in the deepest layer. This allows the model to capture information from a large contextual window. Residual connections are used to facilitate effective training of the deep network architecture (He et al., 2015; De Fauw et al., 2018).

Next to a path with 2×2 downsampling operations, separate downsampling paths are added to the encoder path at every resolution. In these downsampling paths 2×1 (height $\times$ width) operations are used, therefore only reducing the vertically resolution. These paths constitute the funneling

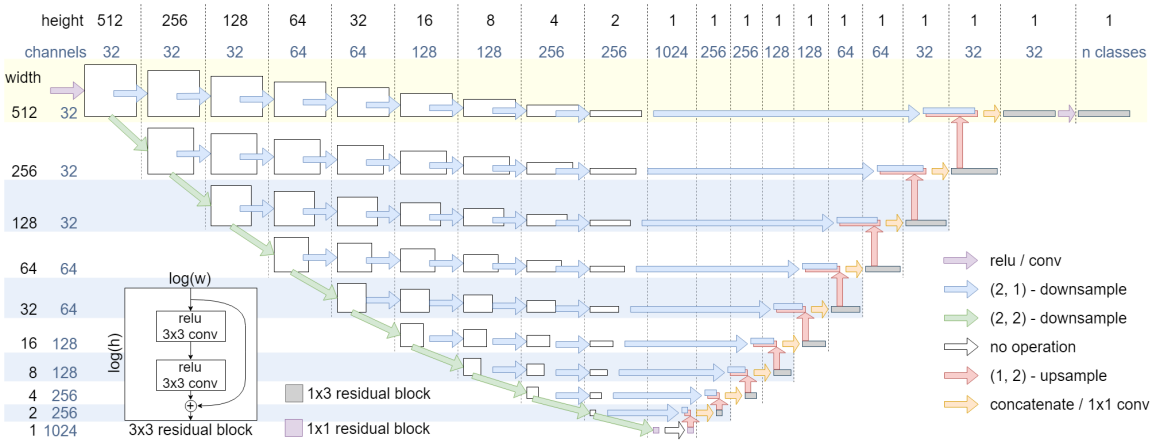

Figure 1: Proposed network architecture for 2D images to 1D segmentation. The white squares represent the size (log-scale) of the feature maps. Within each square, two convolutions are applied in a residual manner. For visualization purposes, the feature maps with width/height 1, have the same size as the feature maps with size 2. In the upsampling path, the residual blocks have height 1, and the convolutional kernels are of size 1×3.

shortcut subnetworks that should provide the model with better localization abilities. Downsampling is performed by either 2×2 convolutions with stride 2×2 or 2×1 convolutions with stride 2×1.

At the bottom of the network, 2 residual blocks with 1×1 convolutions are applied. The upsampling path consists of 1×2 upsampling operations, followed by 1×2 convolutions with stride 1×2. Before entering the first residual block, a 3×3 convolution with 32 filters is applied to match the required number of features for the summation operation in the first residual block. After concatenating the feature maps of the upsampling path and the shortcut networks, a 1×1 convolution is applied to match the required number of features for the next residual block. The kernel size of the convolutions in the downsampling path is 3×3; at the bottom of the network, 1×1; and in the upsampling path, 1×3.

Although the height of the input images is fixed by the network architecture, their width can be any multiple of 512. This is exploited in the application to layer segmentation, where we apply the same architecture to images that are 1024 pixels wide.

3.3. Experimental design

We compare our proposed network architecture to two baseline models, that are constructed by leaving out either the encoder-decoder structure, or the shortcut subnetworks.

The first baseline model, referred to as *base 1*, applies downsampling only in the dimensions that will be spliced out, until the feature maps are sufficiently compressed and can be passed to a classification layer. This architecture corresponds to part of the proposed architecture highlighted in the top yellow row in Figure 1. For this model, before the output layer, we also added a classification

part that consists of two 1×1 residual blocks with 1024 filters each, in order to mimic classification networks more closely.

The second baseline model, referred to as *base 2*, leaves out the shortcut subnetworks. This architecture can be created by following the green and red arrows in Figure 1, ignoring the blue and orange arrows.

3.4. Training procedure

3.4.1. GA SEGMENTATION

Images were cropped to 512×512 pixels. The vertical position was centered on the row in the original image with the highest cumulative pixel intensity, to ensure the retina was visible in the output image. The images were augmented using random horizontal and vertical translations and horizontal flipping. Additionally, pixel intensities were modified using gamma corrections. A sigmoid function was applied to the output image of 512×1 pixels, to obtain a normalized binary label for every A-scan. Mean log loss (binary cross-entropy) was used a loss function. The models were trained for 2×10^4 iterations using the Adam optimizer, with a learning rate of 1×10^{-5}, divided by 2 after 10^4 iterations, on batches of 4 images.

3.4.2. LAYER SEGMENTATION

Input images were padded to 1024×512 pixels. Random horizontal and vertical translation, and random horizontal flips were used as data augmentation. The output layer has three channels, each producing an image of size 1024×1, representing the normalized pixel index of the three layers (ILM, RPEDC, BM). An output of 0 for a specific layer in an A-scan is interpreted as the layer being completely at the top of the image (pixel 0), while an output of 1 is interpreted as the layer being completely at the bottom (pixel 512). No non-linearity was applied to the output layer.

Mean squared error was used as loss function. However, to calculate the loss for a given layer, we only took into account those A-scans where the squared error was larger than the mean squared error of all A-scans for that layer. This modification was made because the contribution to the loss of many small errors may dominate some of the larger, more localized errors. Additionally, images were more frequently sampled for training if their loss was larger than the median loss of the last 100 iterations. The model was trained for 4×10^5 iterations, one image per batch, using the Adam optimizer and a learning rate of 2×10^{-5}, divided by 2 every 10^5 iterations.

4. Results

On the GA data set we calculated a single Dice score per OCT-volume. The proposed model obtained a mean ($\pm$ std) score of 0.49 ± 0.21, compared to 0.46 ± 0.22 for base 1 and 0.28 ± 0.19 for base 2. The proposed model performed significantly better than the two baseline models ($p < 0.01$, paired t-test). Examples of the obtained GA segmentations per B-scan can be found in Figure 2. A selected set of predictions for full volumes can be found in Figure 3.

On the layer-segmentation task, the proposed model obtained a mean absolute difference to the reference standard (averaged over all layers) of 1.305 ± 0.547 pixels, compared to 1.967 ± 0.841 and 2.166 ± 0.886 for base 1 and base 2, respectively. Table 1 summarizes the performance of the models in more detail. An example with predictions for the different models on a single B-scan can be found in Figure 4.

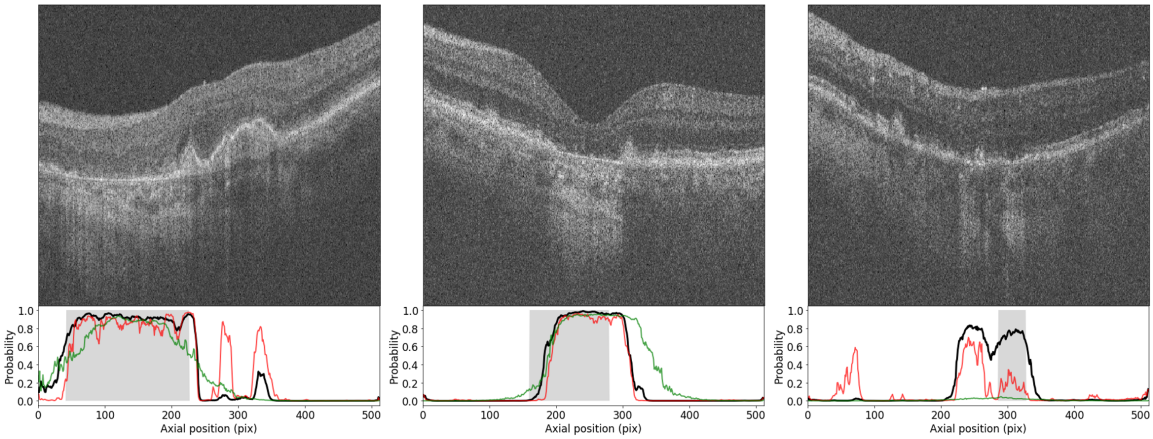

Figure 2: Predictions for GA for 3 different B-scans. The graphs show the reference standard (gray area), and predicted GA probabilities. The black line represents the proposed model; the red line, base 1; and the green line, base 2.

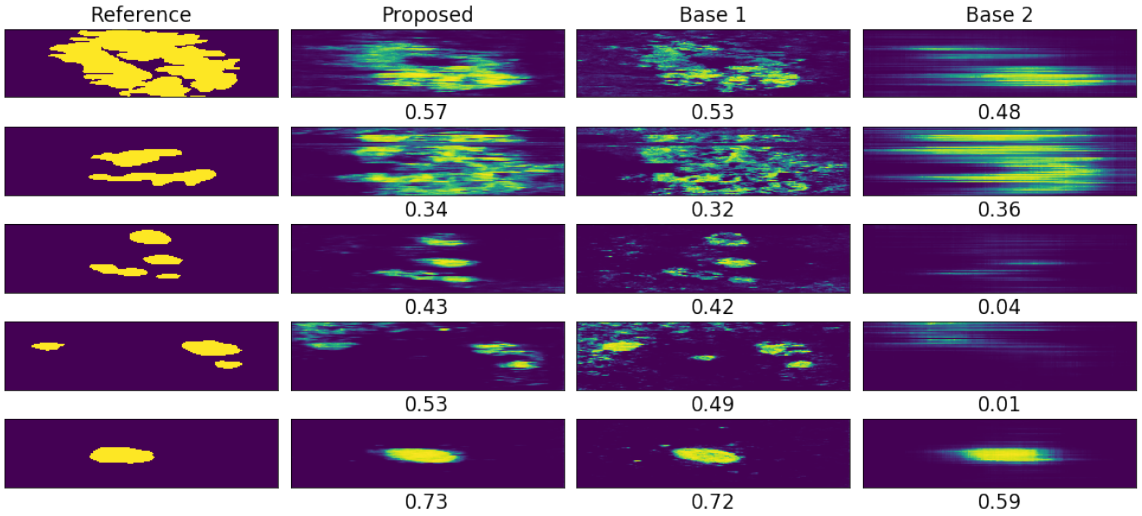

Figure 3: Selection of predictions of GA area for full OCT-volumes. Each horizontal line in the images represents a B-scan. The left column represents the reference standard, the heatmaps in the second to fourth column represent the probabilities generated by the proposed model, base 1, and base 2, respectively. Numbers below the images are the Dice scores.

Table 1: Results on the layer-segmentation task. Values represent mean absolute difference between prediction and annotation in pixels.

	Control			AMD			
	ILM	RPEDC	BM	ILM	RPEDC	BM	Overall
Proposed	0.840	1.280	1.227	1.055	1.568	1.858	1.305
Base 1	1.031	1.507	2.246	1.200	2.267	3.553	1.967
Base 2	1.880	1.762	1.673	2.203	2.785	2.693	2.166

5. Discussion

We proposed a novel CNN architecture for dense segmentation in selected output dimensions. The proposed architecture was validated on two applications in retinal OCT, where it achieved superior performance compared to two baseline models on both tasks. We attribute this to the ability of the proposed architecture to simultaneously make use of a large contextual window, and local context through funneling shortcut connections. This was demonstrated on the layer-segmentation task, where base 1 performed worse especially in localization of BM in AMD, for which a large contextual window is required in presence of abnormalities. Base model 2 can make use of a large context, but appears to be less accurate in general. This may be due to missing shortcut connections, which hampers its ability to perform exact localization. This observation is illustrated by the example in Figure 4. For the GA segmentation we observe similar results. For example, in the left image of Figure 2, the misclassifications in the center of the image of base 1 may be due to lack of context, while base 2 is unable to accurately delineate the borders of the atrophic region.

The Dice scores for the GA segmentation task are relatively low. This is partly due to the inherent difficulty of the task, which may have led to inaccuracies in the reference standard. Moreover, the actual GA area sometimes does not align perfectly with the GA that is visible in the B-scan due to registration errors. This is demonstrated in Figure 2 in the center and right images, where the reference grading does not align well with the observed hypertransmissive region. Visually judging the predictions of the proposed model in Figure 3, however, seems to indicate that even with relatively low Dice scores, the predicted GA areas are plausible.

A limitation of this work is that the proposed general architecture for problems that require mapping an N-dimensional input to an M-dimensional output, has only been validated for the 2D to 1D mapping of retinal OCT data. In future work, we hope to apply the proposed architecture also for the 3D to 2D case. Although the anisotropic resolution of OCT volumes is a hurdle in direct application of 3D convolutions, we do see potential applications of the proposed architecture for isotropic images in OCT angiography. Here it could be used, for example, in the generation of 2D images representing perfusion density in different vascular plexi from 3D volumetric data.

6. Conclusion

We demonstrated the applicability of a new CNN architecture to two problems in retinal OCT, where its value was warranted by the unique properties of the retina (its horizontally-layered structure), and the image acquisition of OCT (where each A-scan represents a column of information on a retinal

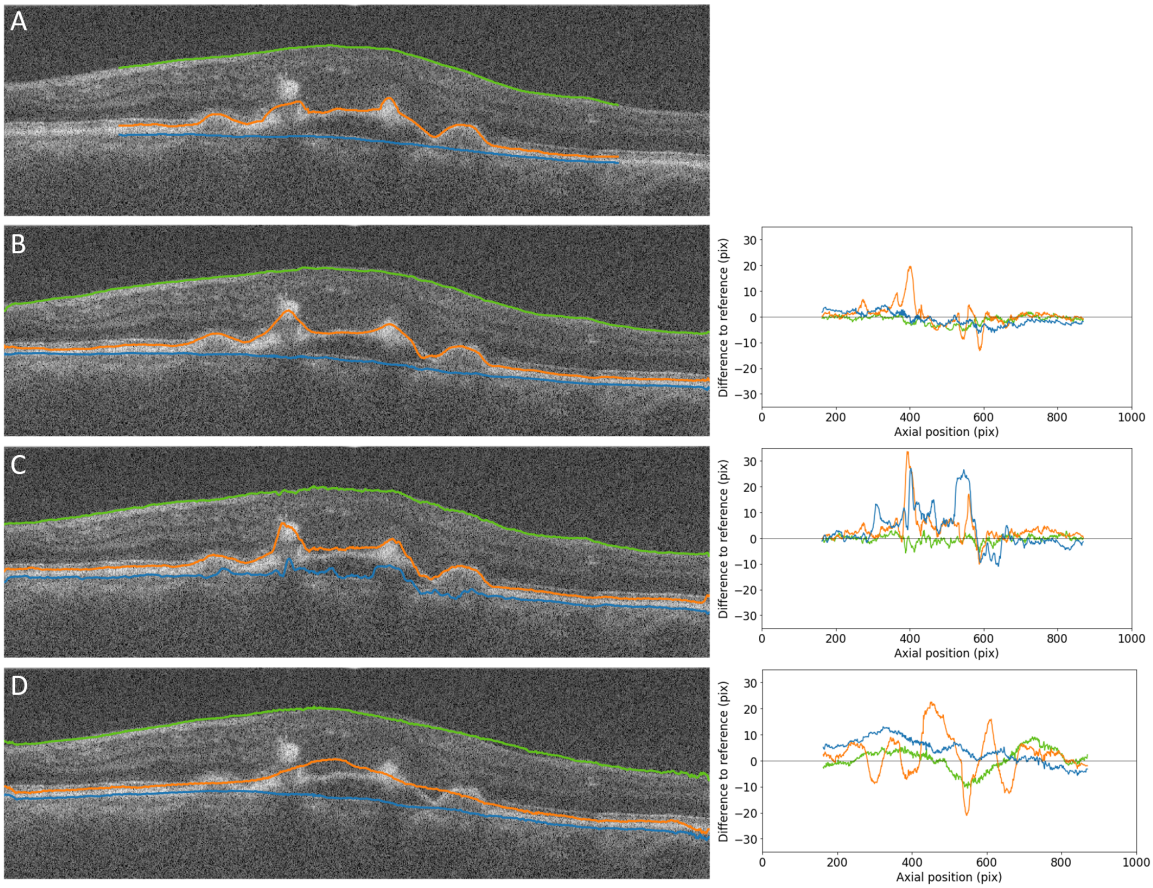

Figure 4: Example result for the layer-segmentation task. Image A shows the reference standard, which is only provided for the central part of the image. Images B, C and D represent the proposed model, base 1, and base 2, respectively. The graphs on the right depict the difference in pixels with the reference for this example for the ILM, RPEDC and BM, in green, orange and blue, respectively.

surface). The proposed model consistently outperformed two baseline models, which indicates that the combination of a large contextual window, provided by an encoder-decoder structure, and accurate localization, provided by funneling shortcut subnetworks, is beneficial.

References

Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al. Clinically applicable deep learning for diagnosis

and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical Optics Express*, 8(5):2732–2744, 2017.

S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

A. Hofman, M. M. B. Breteler, C. M. van Duijn, G. P. Krestin, H. A. Pols, B. H. C. Stricker, Henning T., A. G. Uitterlinden, J. R. Vingerling, and J. C. M. Witteman. The Rotterdam Study: objectives and design update. *European Journal of Epidemiology*, 22(11):819–829, 2007.

Z. Hu, G. G. Medioni, M. Hernandez, A. Hariri, X. Wu, and S. R. Sadda. Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. *Investigative Ophthalmology & Visual Science*, 54(13):8375, 2013.

J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins. Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search. *Biomedical Optics Express*, 9(11):5759–5777, 2018.

L. Liang, K. Rehm, R. P. Woods, and D. A. Rottenberg. Automatic segmentation of left and right cerebral hemispheres from mri brain volumes using the graph cuts algorithm. *NeuroImage*, 34 (3):1160–1170, 2007.

B. Liefers, F. G. Venhuizen, V. Schreur, B. van Ginneken, C. Hoyng, S. Fauser, T. Theelen, and C. I. Sánchez. Automatic detection of the foveal center in optical coherence tomography. *Biomedical Optics Express*, 8(11):5160–5178, 2017.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

S. Niu, L. de Sisternes, Q. Chen, T. Leng, and D. L. Rubin. Automated geographic atrophy segmentation for SD-OCT images using region-based cv model via local similarity factor. *Biomedical Optics Express*, 7(2):581–600, 2016.

R. M. Rangayyan, R. H. Vu, and G. S. Boag. Automatic delineation of the diaphragm in computed tomographic images. *Journal of Digital Imaging*, 21(1):134–147, 2008.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015.

B. Tan, A. Wong, and K. Bizheva. Enhancement of morphological and vascular features in OCT images using a modified Bayesian residual transform. *Biomedical Optics Express*, 9(5):2394–2406, 2018.

J. Wang, M. Betke, and J. P. Ko. Pulmonary fissure segmentation on CT. *Medical Image Analysis*, 10(4):530–547, 2006.