Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography

FREERK G. VENHUIZEN,[1,2,*] BRAM VAN GINNEKEN,[1]
BART LIEFERS,[1,2] FREEKJE VAN ASTEN,[2] VIVIAN SCHREUR,[2]
SASCHA FAUSER,[3,4] CAREL HOYNG,[2] THOMAS THEELEN,[1,2]
AND CLARA I. SÁNCHEZ[1,2]

[1]Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands
[2]Department of Ophthalmology, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands
[3]Roche Pharma Research and Early Development, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[4]Cologne University Eye Clinic, Cologne, Germany
*freerk.venhuizen@radboudumc.nl

Abstract: We developed a deep learning algorithm for the automatic segmentation and quantification of intraretinal cystoid fluid (IRC) in spectral domain optical coherence tomography (SD-OCT) volumes independent of the device used for acquisition. A cascade of neural networks was introduced to include prior information on the retinal anatomy, boosting performance significantly. The proposed algorithm approached human performance reaching an overall Dice coefficient of 0.754 ± 0.136 and an intraclass correlation coefficient of 0.936, for the task of IRC segmentation and quantification, respectively. The proposed method allows for fast quantitative IRC volume measurements that can be used to improve patient care, reduce costs, and allow fast and reliable analysis in large population studies.

References and links

1. P. Enders, P. Scholz, P. S. Muether, and S. Fauser, "Variability of disease activity in patients treated with ranibizumab for neovascular age-related macular degeneration," Eye (Lond) **30**, 1072–1076 (2016).
2. L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," Lancet **379**, 1728–1738 (2012).
3. F. van Asten, M. M. Rovers, Y. T. Lechanteur, D. Smailhodzic, P. S. Muether, J. Chen, A. I. den Hollander, S. Fauser, C. B. Hoyng, G. J. van der Wilt, and B. J. Klevering, "Predicting non-response to ranibizumab in patients with neovascular age-related macular degeneration," Ophthalmic Epidemiol. **21**, 347–355 (2014).
4. CATT Research Group, D. F. Martin, M. G. Maguire, G. S. Ying, J. E. Grunwald, S. L. Fine, and G. J. Jaffe, "Ranibizumab and bevacizumab for neovascular age-related macular degeneration," N. Engl. J. Med. **364**, 1897–1908 (2011).
5. S. M. Waldstein, J. Wright, J. Warburton, P. Margaron, C. Simader, and U. Schmidt-Erfurth, "Predictive Value of Retinal Morphology for Visual Acuity Outcomes of Different Ranibizumab Treatment Regimens for Neovascular AMD," Ophthalmology **123**, 60–69 (2016).
6. S. M. Waldstein, A. M. Philip, R. Leitner, C. Simader, G. Langs, B. S. Gerendas, and U. Schmidt-Erfurth, "Correlation of 3-Dimensionally Quantified Intraretinal and Subretinal Fluid With Visual Acuity in Neovascular Age-Related Macular Degeneration," JAMA Ophthalmol **134**, 182–190 (2016).
7. U. Schmidt-Erfurth and S. M. Waldstein, "A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration," Prog Retin Eye Res **50**, 1–24 (2016).
8. U. Schmidt-Erfurth, V. Chong, A. Loewenstein, M. Larsen, E. Souied, R. Schlingemann, B. Eldem, J. Mones, G. Richard, and F. Bandello, "Guidelines for the management of neovascular age-related macular degeneration by the European Society of Retina Specialists (EURETINA)," Br. J. Ophthalmol. **98**, 1144–1167 (2014).
9. M. W. M. Wintergerst, T. Schultz, J. Birtel, A. K. Schuster, N. Pfeiffer, S. Schmitz-Valckenberg, F. G. Holz, and R. P. Finger, "Algorithms for the Automated Analysis of Age-Related Macular Degeneration Biomarkers on Optical Coherence Tomography: A Systematic Review," Transl. Vis. Sci. Technol. **6**, 10 (2017).

10. M. K. Garvin, M. D. Abramoff, R. Kardon, S. R. Russell, X. Wu, and M. Sonka, "Intraretinal layer segmentation of macular optical coherence tomography images using optimal 3-D graph search," IEEE Trans. Med. Imaging **27**, 1495–1505 (2008).
11. R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map," Med. Image Anal. **17**, 907–928 (2013).
12. F. G. Venhuizen, B. van Ginneken, B. Liefers, M. J. J. P. van Grinsven, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sanchez, "Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks," Biomed. Opt. Express **8**, 3292–3316 (2017).
13. S. J. Chiu, J. A. Izatt, R. V. O'Connell, K. P. Winter, C. A. Toth, and S. Farsiu, "Validated automatic segmentation of AMD pathology including drusen and geographic atrophy in SD-OCT images," Invest. Ophthalmol. Vis. Sci. **53**, 53–61 (2012).
14. A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunovic, "Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context," Biomed Opt Express **8**, 1874–1888 (2017).
15. A. Lang, A. Carass, A. K. Bittner, H. S. Ying, and J. L. Prince, "Improving graph-based OCT segmentation for severe pathology in Retinitis Pigmentosa patients," Proc SPIE Int Soc Opt Eng **10137** (2017).
16. 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," Biomed Opt Express **8**, 2732–2744 (2017).
17. L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation Based Sparse Reconstruction of Optical Coherence Tomography Images," IEEE Trans. Med. Imaging **36**, 407–421 (2017).
18. L. Fang, S. Li, R. P. McNabb, Q. Nie, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," IEEE Trans. Med. Imaging **32**, 2034–2049 (2013).
19. H. M. Salinas and D. C. Fernandez, "Comparison of PDE-based nonlinear diffusion approaches for image enhancement and denoising in optical coherence tomography," IEEE Trans Med Imaging **26**, 761–771 (2007).
20. A. Wong, A. Mishra, K. Bizheva, and D. A. Clausi, "General Bayesian estimation for speckle noise reduction in optical coherence tomography retinal imagery," Opt. Express **18**, 8338–8352 (2010).
21. M. Wu, Q. Chen, X. He, P. Li, W. Fan, S. Yuan, and H. Park, "Automatic Subretinal Fluid Segmentation of Retinal SD-OCT Images with Neurosensory Retinal Detachment Guided by Enface Fundus Imaging," IEEE Trans. Biomed. Eng. **65**, 87–95 (2017).
22. Y. Zheng, J. Sahni, C. Campa, A. N. Stangos, A. Raj, and S. P. Harding, "Computerized assessment of intraretinal and subretinal fluid regions in spectral-domain optical coherence tomography images of the retina," Am. J. Ophthalmol. **155**, 277–286 (2013).
23. W. Ding, M. Young, S. Bourgault, S. Lee, D. A. Albiani, A. W. Kirker, F. Forooghian, M. V. Sarunic, A. B. Merkur, and M. F. Beg, "Automatic detection of subretinal fluid and sub-retinal pigment epithelium fluid in optical coherence tomography images," Conf Proc IEEE Eng Med Biol Soc **2013**, 7388–7391 (2013).
24. S. Niu, L. de Sisternes, Q. Chen, T. Leng, and D. L. Rubin, "Automated geographic atrophy segmentation for SD-OCT images using region-based C-V model via local similarity factor," Biomed Opt Express **7**, 581–600 (2016).
25. 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," Invest. Ophthalmol. Vis. Sci. **54**, 8375–8383 (2013).
26. Z. Ji, Q. Chen, S. Niu, T. Leng, and D. L. Rubin, "Beyond Retinal Layers: A Deep Voting Model for Automated Geographic Atrophy Segmentation in SD-OCT Images," Transl. Vis. Sci. Technol. **7**, 1 (2018).
27. L. de Sisternes, G. Jonna, M. A. Greven, Q. Chen, T. Leng, and D. L. Rubin, "Individual Drusen Segmentation and Repeatability and Reproducibility of Their Automated Quantification in Optical Coherence Tomography Images," Transl. Vis. Sci. Technol. **6**, 12 (2017).
28. Q. Chen, T. Leng, L. Zheng, L. Kutzscher, J. Ma, L. de Sisternes, and D. L. Rubin, "Automated drusen segmentation and quantification in SD-OCT images," Med. Image Anal. **17**, 1058–1072 (2013).
29. S. Farsiu, S. J. Chiu, J. Izatt, and C. Toth, "Fast detection and segmentation of drusen in retinal optical coherence tomography images", Proc. SPIE **6844**, 6844OD (2008).
30. D. Iwama, M. Hangai, S. Ooto, A. Sakamoto, H. Nakanishi, T. Fujimura, A. Domalpally, R. P. Danis, and N. Yoshimura, "Automated assessment of drusen using three-dimensional spectral-domain optical coherence tomography," Invest. Ophthalmol. Vis. Sci. **53**, 1576–1583 (2012).
31. S. Khalid, M. U. Akram, T. Hassan, A. Jameel, and T. Khalil, "Automated Segmentation and Quantification of Drusen in Fundus and Optical Coherence Tomography Images for Detection of ARMD," J. Digi.t Imaging (2017).
32. M. Esmaeili, A. M. Dehnavi, H. Rabbani, and F. Hajizadeh, "Three-dimensional Segmentation of Retinal Cysts from Spectral-domain Optical Coherence Tomography Images by the Use of Three-dimensional Curvelet Based K-SVD," J. Med. Signals Sens. **6**, 166–171 (2016).
33. G. R. Wilkins, O. M. Houghton, and A. L. Oldenburg, "Automated segmentation of intraretinal cystoid fluid in optical coherence tomography," IEEE Trans. Biomed. Eng. **59**, 1109–1114 (2012).
34. T. Schlegl, S. M. Waldstein, W. D. Vogl, U. Schmidt-Erfurth, and G. Langs, "Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks," Inf. Process Med. Imaging **24**, 437–448 (2015).

35. A. Rashno, D. D. Koozekanani, P. M. Drayna, B. Nazari, S. Sadri, H. Rabbani, and K. K. Parhi, "Fully-Automated Segmentation of Fluid/Cyst Regions in Optical Coherence Tomography Images with Diabetic Macular Edema using Neutrosophic Sets and Graph Algorithms," IEEE Trans. Biomed. Eng. **99**, 1 (2017).

36. A. Rashno, B. Nazari, D. D. Koozekanani, P. M. Drayna, S. Sadri, H. Rabbani, and K. K. Parhi, "Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: Kernel graph cut in neutrosophic domain," PLoS ONE **12**, e0186949 (2017).

37. S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," Biomed. Opt. Express **6**, 1172–1194 (2015).

38. 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 Trans. Med. Imaging **31**, 1521–1531 (2012).

39. C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," Biomed. Opt. Express **8**, 3440–3448 (2017).

40. J. Wang, M. Zhang, A. D. Pechauer, L. Liu, T. S. Hwang, D. J. Wilson, D. Li, and Y. Jia, "Automated volumetric segmentation of retinal fluid on optical coherence tomography," Biomed Opt Express **7**, 1577–1589 (2016).

41. D. C. Fernandez, "Delineating fluid-filled region boundaries in optical coherence tomography images of the retina," IEEE Trans. Med. Imaging **24**, 929–945 (2005).

42. M. Pilch, K. Stieger, Y. Wenner, M. N. Preising, C. Friedburg, E. Meyer zu Bexten, and B. Lorenz, "Automated segmentation of pathological cavities in optical coherence tomography scans," Invest. Ophthalmol. Vis. Sci. **54**, 4385–4393 (2013).

43. G. N. Girish, A. R. Kothari, and J. Rajan, "Automated segmentation of intra-retinal cysts from optical coherence tomography scans using marker controlled watershed transform," Conf. Proc. IEEE Eng. Med. Biol. Soc. **2016**, 1292–1295 (2016).

44. T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstrasser, A. Sadeghipour, A. M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning," Ophthalmology (Epub ahead of print) (2017).

45. A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," Biomed Opt Express **8**, 3627–3642 (2017).

46. G. M. GN Girish, V. A. Anima, A. K. Kothari, P. V. Sudeep, S. Roychowdhury, and J. Rajan, "A benchmark study of automated intra-retinal cyst segmentation algorithms using optical coherence tomography B-scans," Comput. Methods Programs Biomed. **153**, 105–114 (2018).

47. S. Fauser, D. Smailhodzic, A. Caramoy, J. P. van de Ven, B. Kirchhof, C. B. Hoyng, B. J. Klevering, S. Liakopoulos, and A. I. den Hollander, "Evaluation of serum lipid concentrations and genetic variants at high-density lipoprotein metabolism loci and TIMP3 in age-related macular degeneration," Invest. Ophthalmol. Vis. Sci. **52**, 5525–5528 (2011).

48. J. P. van de Ven, D. Smailhodzic, C. J. Boon, S. Fauser, J. M. Groenewoud, N. V. Chong, C. B. Hoyng, B. J. Klevering, and A. I. den Hollander, "Association analysis of genetic and environmental risk factors in the cuticular drusen subtype of age-related macular degeneration," Mol. Vis. **18**, 2271–2278 (2012).

49. J. Wu, A. M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, C. Simader, S. M. Waldstein, and U. M. Schmidt-Erfurth, "Multivendor Spectral-Domain Optical Coherence Tomography Dataset, Observer Annotation Performance Evaluation, and Standardized Evaluation Framework for Intraretinal Cystoid Fluid Segmentation," J. Ophthalmol. **2016**, 3898750 (2016).

50. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," MICCAI 2015: 18th International Conference **9351**, 234–241 (2015).

51. Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints **abs/1605.02688** (2016).

52. S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sonderby, D. Nouri, "Lasagne: First release." (2015).

53. S. Zheng, Y. Song, T. Leung, and I. J. Goodfellow, "Improving the robustness of deep neural networks via stability training," CoRR **abs/1604.04326** (2016).

1. Introduction

Age-related macular degeneration (AMD) is a complex multi-factorial retinal disease where genetic and environmental factors play a large role in the development of the disease. Since the disease is influenced by a wide array of intrinsic and extrinsic risk factors and protective factors, a universal one-fits-all treatment is arguably not the optimal solution as shown by the large percentage of non-responders to available therapies [1]. A common treatment option for exudative AMD is the injection of intravitreal anti-vascular endothelial growth factor (anti-VEGF). This treatment stops the growth of abnormal blood vessels that cause leakage of vascular fluid, blood
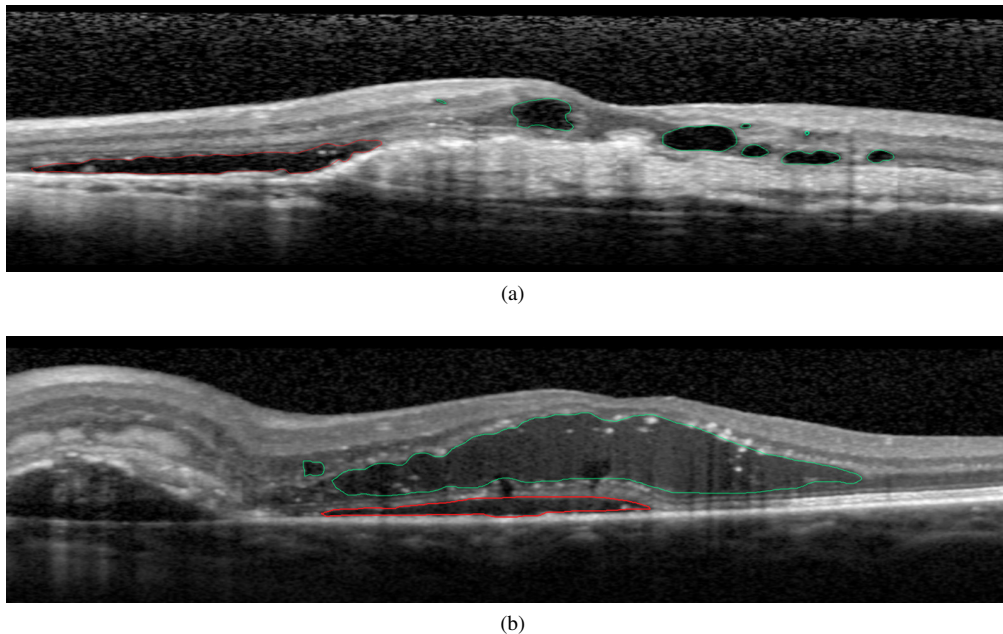
(a)

(b)

Fig. 1. Examples of B-scans showing the different types of retinal fluid: Subretinal fluid (SRF) is indicated in red, while intraretinal fluid (IRF) is indicated in green. Intraretinal fluid can be subdivided in intraretinal cystoid fluid (IRC) shown in (a), and diffuse non-cystic IRF shown in (b).

and lipids in the retina [1, 2]. Anti-VEGF treatment is typically administered monthly for an extended period of time, causing a high burden on the patient while not always resulting in improved vision [1, 3]. Moreover, as anti-VEGF treatment is expensive, treatment on a monthly basis also causes a significant financial burden to the healthcare system.

To improve patient care, and reduce cost, other treatment regimens such as the Pro Re Nata (PRN or "as needed") and the Treat-and-Extend (TE) regimes have been proposed, where re-treatment is only applied in case of re-occurrence of retinal bleeding or fluid accumulation [4]. However, these regimens require a constant assessment of the presence of fluid and its changes over time in order to provide personalized care and reduce unnecessary injections.

Fluid accumulation is best visualized on spectral domain optical coherence tomography (SD-OCT) imaging, an almost indispensable tool in the assessment of AMD treatment success. SD-OCT imaging provides a noninvasive, high resolution, three-dimensional visualization of the retina, where fluid is visible as a hyporeflective area. A clear distinction can be made on SD-OCT imaging between intraretinal fluid (IRF) and subretinal fluid (SRF) based on the relative location in the retina, i.e., inside the sensory retina or below it, respectively. For IRF a distinction can be made between intraretinal cystoid fluid (IRC) or cysts, and diffuse non-cystic IRF. Examples of SRF and both types of IRF are shown in Fig. 1. Detecting areas of fluid accumulation, and the specific type of fluid, in SD-OCT, is important in determining the best treatment option and treatment efficacy [5]. Especially the presence of IRC at baseline has been found to be an important factor related to decreased treatment response [6, 7]. During treatment, detection of changes in IRC helps assessing treatment efficacy, especially in the PRN and TE treatment regimens, where retreatment depends on monitoring fluid changes. Accurate analysis of these changes by visual assessment is difficult and subjective, especially if they are small. Manual delineation of IRC does allow for an exact and reproducible objective comparison, but it is extremely time-consuming and

often infeasible, especially in large scale population studies or time-constrained clinics. For this reason, computer-aided detection systems capable of automatically detecting and quantifying IRC are highly sought after [8]. In the last years, computer-based algorithms have shown their potential in the automatic analysis of retinal images and, particularly, in SD-OCT volumes [9]. Previously proposed works for automated analysis of SD-OCT volumes have reported good results in the automated segmentation of intraretinal layers, either for healthy or mildly affected retinas [10, 11] or retinas with more severe pathology [12–16]. As the image quality and the amount of noise in OCT images can vary strongly, a significant amount of research has also been spent on the development of effective denoising algorithms, either to improve viewing quality or to improve performance of a consecutive segmentation algorithm [17–20]. Detection and segmentation of specific retinal lesions has also been proposed, e.g., segmentation of SRF [21–23], segmentation of geographic atrophy [24–26] and automated detection of drusen [27–31]. For the segmentation of IRF a distinction can be made in the previously published methods: while some methods focused on IRC [22, 32–36], other methods extend the segmentation to also include diffuse non-cystic IRF [37–41], making a direct comparison between methods difficult. One of the first attempts to automatically segment IRC was approached using a deformable model reporting good results in a small dataset, but requiring manual initialization for each lesion separately [41]. Segmentation of IRC was addressed in a previous work using a three-dimensional curvelet based k-singular value decomposition (K-SVD) and validated in four SD-OCT volumes acquired by a Heidelberg SD-OCT scanner, reaching a Dice overlap coefficient of 0.65 [32]. A method relying upon a transformation to the neutrosophic domain followed by a shortest path algorithm was evaluated on the same four Heidelberg SD-OCT volumes used by the previously stated method, reaching a Dice coefficient of 0.705 [35]. Recently, a modified version of this algorithm was published improving upon their previous results, achieving a Dice coefficient of 0.812 [36]. Unfortunately, evaluation was only performed on the same four Heidelberg SD-OCT volumes. Other authors proposed a filtering and thresholding based algorithm for IRC segmentation and evaluated it in a small dataset of 16 SD-OCT volumes obtained by a Cirrus SD-OCT scanner, showing good results but failing to detect small cysts [33]. A method based on k-means cluster analysis showed good performance, but was only evaluated in a small set of 130 lesions [42]. Other authors proposed a method based on the marker controlled watershed transform, showing good performance in a small set of 4 OCT volumes acquired by a single OCT device [43]. A semi-automatic variational segmentation algorithm has also been proposed for the segmentation of IRC achieving good performance [22]. While the authors claimed the method can be extended to other SD-OCT vendors, the algorithm was only validated on SD-OCT volumes obtained by a Heidelberg SD-OCT scanner. A convolutional neural network (CNN) was proposed for the segmentation of IRC using a multiscale patch based approach [34].

The method was validated in a large dataset of 157 SD-OCT volumes from a single vendor with good results. Recently, this method was extended and evaluated in a large dataset containing cases of AMD, diabetic macular edema (DME), and retinal vein occlusion (RVO) and reported good performance for segmentation and quantification of IRC [44]. Another deep learning based method obtained a Dice coefficient of 0.729, but instead of evaluating on SD-OCT volumes, the analysis was performed in a set of 934 B-scans with varying pathology acquired with a single SD-OCT device [39]. A CNN developed for the segmentation of 7 retinal layers and fluid masses showed good performance [45]. The method was, however, developed and evaluated in a small dataset containing 110 B-scans from 10 patients acquired by a single SD-OCT device. For the undifferentiated segmentation of IRC and diffuse IRF, a method based on kernel regression was proposed and validated in a dataset containing cases with diabetic macular edema (DME) acquired with a Heidelberg SD-OCT scanner [37]. The algorithm showed segmentation performance close to the interobserver variability. Fuzzy level-sets were applied for the segmentation of IRF obtaining good agreement with manual segmentations, but were evaluated in a small set of

SD-OCT volumes acquired by a single SD-OCT device [40]. Other authors described a method using probability constrained graph cuts to segment so-called 'symptomatic exudate-associated derangement' or SEADs, which includes IRF, both cystoid and diffuse [38]. While the method showed reasonable performance, it relied heavily on a good initialization of the method. Recently a method for the simultaneous segmentation of retinal fluid and retinal layers was proposed using an auto-context approach [14], obtaining an undifferentiated segmentation of both diffuse and cystoid IRF. Although relatively poor results were reported in a private dataset containing Cirrus SD-OCT volumes, the method achieved comparable results in the DME dataset used by a previously discussed method [37]. Finally, in a recent benchmark study, several methods for IRC segmentation were compared and evaluated in a publicly available dataset [46]. Unfortunately, only data from two of the four SD-OCT devices present in the dataset were included for analysis. As described, the majority of the previously proposed methods were evaluated on SD-OCT data from a single SD-OCT vendor, limiting the assessment and the widespread use of these algorithms in data with different imaging characteristics.

In this work we propose a deep learning algorithm for the detection, segmentation and quantification of IRC in SD-OCT volumes acquired with different SD-OCT devices. The proposed algorithm is based on a fully convolutional neural network (FCNN), where IRC is segmented by performing a semantic segmentation of the SD-OCT volume, i.e., every pixel in an SD-OCT volume is analyzed and given a probability of belonging to IRC. The proposed FCNN implements a multi-scale analysis, providing a large contextual window that allows for accurate segmentation of a wide range of cysts, ranging from small micro-cysts to large intraretinal cysts spanning over a wide area of a B-scan. The large modeling capacity of the proposed FCNN allows it to learn a wide range of different complex features, capable of capturing the large variability in IRC appearance and vendor-dependent SD-OCT characteristics. Segmentation performance is further enhanced by the inclusion of an additional step for retina segmentation as a means to constrain the search space during training.

The algorithm performance is evaluated in 1) a large private database of SD-OCT volumes from AMD patients with advanced AMD with IRC; 2) a dataset of healthy controls; and 3) a publicly available dataset containing SD-OCT volumes with IRC from four different vendors.

2. Methods

2.1. Data

For this study a total of 221 SD-OCT volumes from 151 patients (6158 B-scans) with varying presence of IRC were randomly selected from the European Genetic Database (EUGENDA, http://eugenda.org), a large multi-center database for clinical and molecular analysis of AMD [47, 48]. Written informed consent was obtained before enrolling patients in EUGENDA. The EUGENDA study was performed according to the tenets set forth in the Declaration of Helsinki, and Institutional Review Board approval was obtained.

SD-OCT volumes were acquired using a Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany) at a wavelength of 870 nm, a transversal resolution ranging from 6 μm to 14 μm and an axial resolution of up to 3.9 μm. The dimension in the axial resolution was 496 pixels, in the transversal direction the dimensions varied between 512 and 1536 pixels. The number of slices, i.e. the number of B-scans, varied from 19 to 37, corresponding to a B-scan spacing ranging from 240 μm up to 120 μm, respectively. Before processing, to remove the variability in resolution, all B-scans from an SD-OCT volume were resampled to a constant pixel size of 11.5 μm x 3.9 μm.

The data were randomly divided into three sets (approximately 40/10/50 split on patient level): a **training set**, consisting of 103 SD-OCT volumes (3131 B-scans) from 60 patients, for the development and optimization of the algorithm; a **validation set** of 19 SD-OCT volumes (540 B-scans) from 16 patients, for monitoring the algorithm training process; and an independent
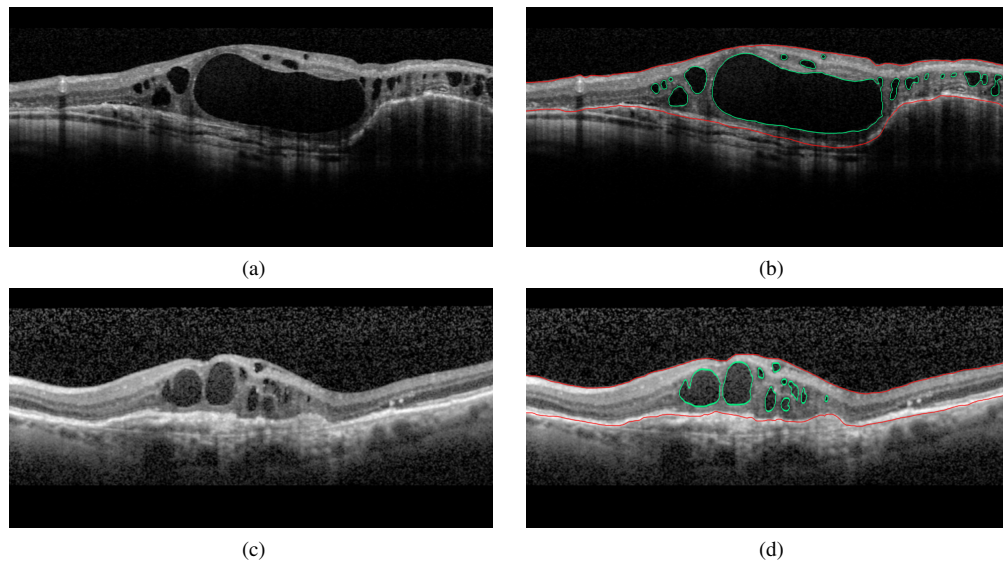
Fig. 2. Examples of B-scans from the EUGENDA database and the corresponding manual annotations: (a) and (c) original image, (b) and (d) corresponding retina and IRC annotations. IRC annotations are indicated in green, retina annotations are indicated in red.

test set consisting of 99 SD-OCT volumes (2487 B-scans) from 75 patients for the evaluation of the algorithm. SD-OCT volumes from the same patient were kept in the same subset. The test set was further split in three subsets, a **first test set** consisting of 53 SD-OCT volumes (1480 B-scans) from 32 patients, a **second test set** consisting of 10 SD-OCT volumes (324 B-scans) from 10 patients, and a **control test set** consisting of 33 SD-OCT volumes (683 B-scans) from 33 healthy controls.

In the **training set** full volume annotation of IRC and the total retina, i.e., the region between the inner limiting membrane (ILM) and the outer boundary of the retinal pigment epithelium (RPE), was performed. Annotation was performed by an experienced grader using a computer-assisted annotation platform which allows manual pixel level annotation and correction. In order to reduce annotation time, annotations were made using a semi-automated approach. An initial retina and IRC segmentation produced by a preliminary segmentation algorithm was proposed to the human grader, who was provided with tools for manual correction if errors were present in the initial annotation proposed by the system. Examples of annotated B-scans from the EUGENDA database are shown in Fig. 2.

In the **first test set** annotation of IRC was performed by the human grader using the same computer-assisted annotation platform as used for the **training set**.

In order to assess the inter-observer variability 50 B-scans containing IRC were randomly selected from the **second test set** for annotation. These 50 B-scans were annotated by three independent observers, of which one was selected as reference standard. Annotations were made without the support of the computer assisted annotation platform.

To assess the generalizability of the proposed algorithm to multiple vendors an **external set** was created from a publicly available database (OPTIMA) containing 30 SD-OCT volumes acquired by four different SD-OCT devices [49], namely:

- Heidelberg Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany)

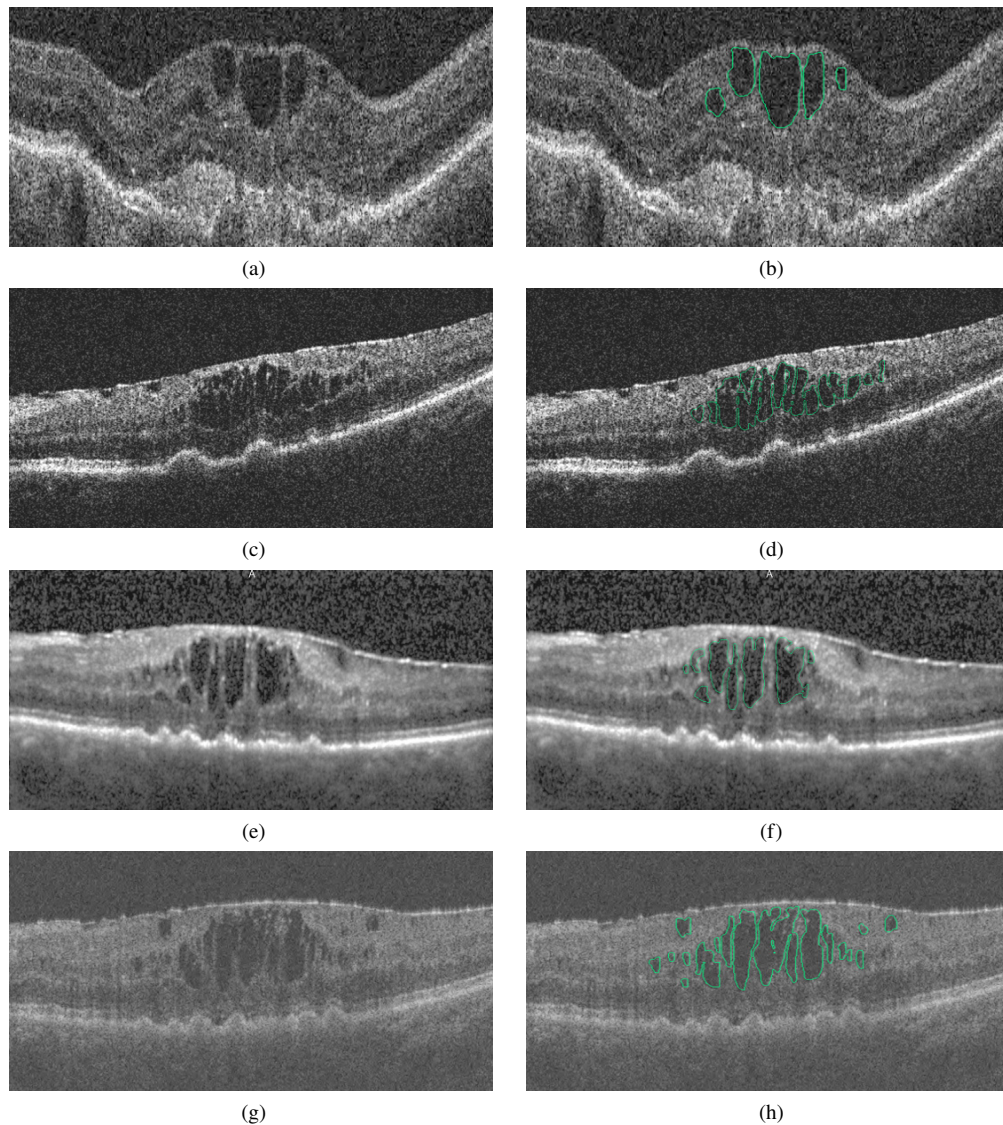- Zeiss Cirrus (Carl Zeiss Meditec, Dublin, CA, USA)

Fig. 3. Examples of B-scans from the OPTIMA database and the corresponding manual annotations: (a,b) Example Cirrus B-scan and the corresponding IRC annotation, (c,d) Example Nidek B-scan and IRC annotation, (e,f) Example Spectralis B-scan and IRC annotation, (g,h) Example Topcon B-scan and IRC annotation. IRC annotations are indicated in green.

- Topcon 3D 2000 (Topcon Medical Systems, Paramus, NJ, USA)

- Nidek RS3000 (Nidek, Aichi, Japan)

Examples of annotated B-scans for each of the four devices are shown in Fig. 3.

This **external dataset** is comprised of a training set of 15 SD-OCT volumes, and a test set also containing 15 SD-OCT volumes. Both the training and test subsets are comprised of four volumes per vendor aside from Nidek with three. For further details and information concerning
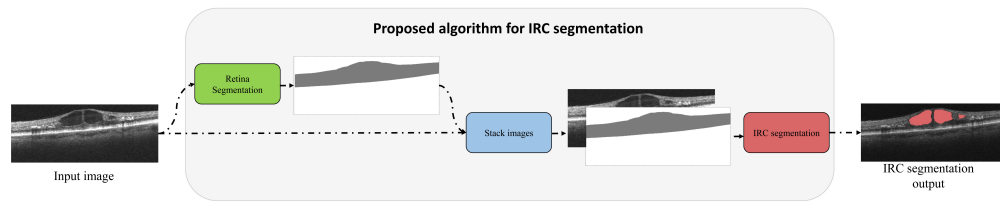
Fig. 4. Overview of the proposed algorithm for IRC segmentation. The first FCNNs responsible for retina segmentation is visualized in green, while the second FCNN responsible for IRC segmentation is shown in red. The retina segmentation produced by the first FCNN is stacked together with the input B-scan to form a two-channel input for the IRC segmentation.

the inclusion criteria we refer to the paper describing the data set [49]. Manual pixel level IRC annotations made by two independent observers were provided for both the training and test set in the **external dataset**. For this study we used the intersection of the two observers, i.e., where both observers agree on the presence of IRC, as the reference standard for training the algorithm.

2.2. Deep learning algorithm

The proposed deep learning algorithm automatically detects, segments and quantifies IRC in the entire SD-OCT volume. The algorithm produces an IRC volume segmentation by processing every B-scan in an SD-OCT volume individually. After processing all B-scans in a volume the resulting segmentations are combined to form the output volume segmentation. The algorithm consists of a cascade of two FCNNs with two complementary tasks: the first FCNN aims at delimiting the total retinal area; and the second FCNN aims at segmenting IRC by integrating the output of the first FCNN in the optimization process. This allows for the inclusion of a priori knowledge of the retinal anatomy and improves segmentation performance. Figure 4 gives a schematic overview of the entire processing pipeline with the two cascaded FCNNs indicated in green and red. The specifics for each task will be explained in more detail in the two following subsections. The FCNN architecture used for these tasks and their training procedures will be explained in subsections 2.2.3 and 2.2.4, respectively.

2.2.1. First task: Retina segmentation

As IRC is restricted to the neurosensory retina, information about the relative location in the retina can be beneficial in determining if a pixel is part of IRC. For example, fluid closer to the RPE is more likely to be SRF than IRC. This a priori information about the retina anatomy can be extracted by identifying and delimiting the retina, i.e. by segmenting the retina. Besides location information, retina segmentation also allows to focus the learning process within the limits of the retina area, ignoring other areas present in the scan, such as the choroid or the vitreous.

Building upon our previous work [12], we develop an FCNN to obtain a semantic segmentation of the retina, i.e. every pixel in an SD-OCT volume is analyzed and given a probability of belonging to the retina. The architecture of this FCNN is specifically developed for robust segmentation of the retina in the presence of disruptive pathology such as IRC, explained in section 2.2.3. The output of this first step, after thresholding, is a binary image indicating all the pixels between the ILM and the RPE. Figure 5(b) shows an example of the output of this task. Further details can be found in section 2.2.3 describing the network architecture, and section 2.2.4 describing the training procedure. Qualitative and quantitative retina segmentation results can be found in our previously published paper introducing the retina segmentation FCNN. [12]
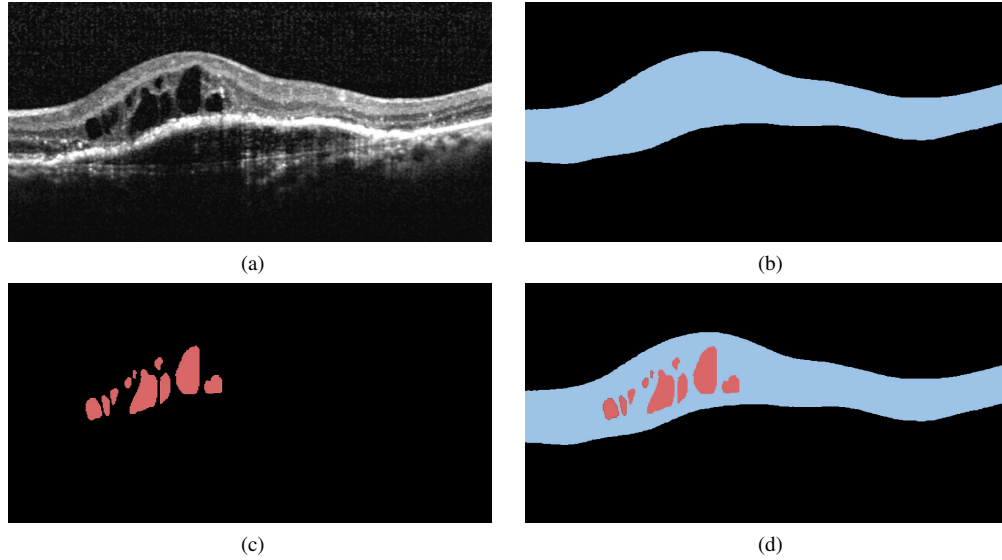
(a)

(b)

(c)

(d)

Fig. 5. Example data used for training the proposed algorithm: (a) Input B-scan, (b) the derived retina segmentation (blue), (c) the corresponding IRC annotations (red), and (d) the weight map calculated from the IRC annotation and retina segmentation. The weight for background pixels (black) is set to 0, retina pixels (blue) get a weight of 1, and IRC pixels (red) get a weight between 0 and 5.

2.2.2. Second task: Intraretinal cystoid fluid segmentation

For the second task a FCNN is developed to obtain a semantic segmentation of IRC, i.e. a pixelwise classification of the full SD-OCT scan indicating the areas of IRC. In order to introduce a priori information about the retina anatomy, this FCNN integrates the output produced by the previous step using two different approaches: 1) as an additional input, and 2) as a constraint of the network training process. For the first approach, additional input, the binary retina segmentation produced by the retina segmentation step is stacked together with the input B-scan and provided as a two-channel input to the FCNN. This addition allows the network to learn location specific features that can improve segmentation performance. For the second approach, the retina segmentation output is used to create a weight map that give locations within the retina more importance during training, ignoring areas outside the retina, i.e., choroidal tissue and vitreous fluid. Given a B-scan B_i from the training set, its obtained retina segmentation R_i and the manually annotated IRC S_i considered as reference standard (see Fig. 5), the weight map for each location $\mathbf{x}$ in B_i is defined as:

$$w_i(\mathbf{x}) = \begin{cases} min(5, \frac{N_{retina}}{N_{IRC}}) & \text{if } S_i(\mathbf{x}) = 1, \text{ i.e., } \mathbf{x} \in \text{IRC} \\ 1 & \text{if } S_i(\mathbf{x}) = 0 \text{ and } R_i(\mathbf{x}) = 1 \text{, i.e., } \mathbf{x} \in \text{retina} \\ 0 & \text{if } S_i(\mathbf{x}) = 0 \text{ and } R_i(\mathbf{x}) = 0 \text{, i.e., } \mathbf{x} \in \text{other areas} \end{cases} \tag{1}$$

With N_{retina} the total number of pixels that are part of the retina (excluding IRC) in B_i, and N_{IRC} the number of pixels that are part of IRC in B_i. This weight map is then introduced in the loss function of the developed FCNN in order to focus the training process on those areas with higher weight, such as IRC, and disregard less important parts of a B-scan such as the vitreous fluid or choroidal tissue. An example of the weight map is shown in Fig. 5(d), where the weight
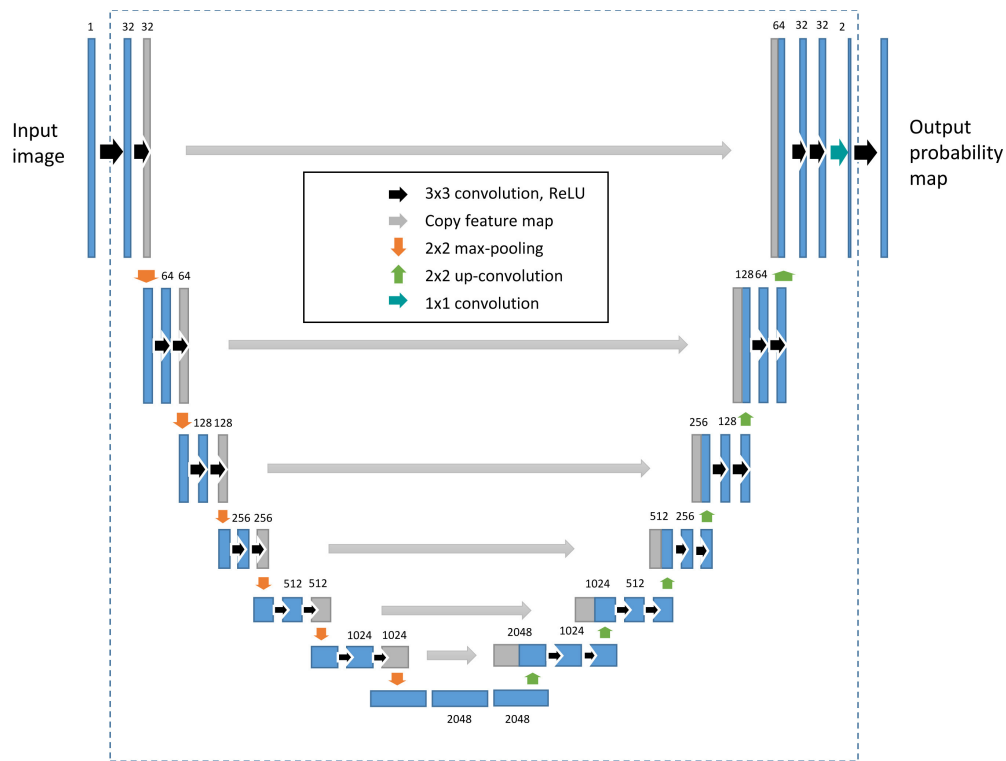
Fig. 6. Schematic overview of the proposed neural network architecture consisting of a total of 27 convolutional layers, 6 max pooling operations (orange arrows) and 6 upsample operations (green arrows), providing a receptive field of 572 x 572 pixels.

for background pixels is set to 0, retina pixels get a weight of 1, and IRC pixels get a weight between 0 and 5.

2.2.3. FCNN architecture

The FCNN architecture used in both tasks is based on U-net, a fully convolutional deep learning approach specifically designed for dense segmentation tasks [50]. The U-net architecture designed for this problem is visualized in Fig. 6. U-net can be subdivided in a left and right side, where the left side is responsible for obtaining contextual information, i.e., what, while the right side is responsible for accurate localization, i.e., where. Due to the successive application of max pooling operations in the left side of the network, indicated by the orange downward arrows in Fig. 6, the spatial context, or receptive field, increases, allowing the input image to be analyzed at multiple scales. At every consecutive scale the spatial context is doubled while the resolution is halved. To recover a segmentation at full resolution, the right side of the "U" performs a series of upsampling operations indicated by the green upward arrows in Fig. 6. Finally, shortcut connections going from the left part of the "U" to the right part are added to integrate high resolution information from the left side with low resolution information from the right side.

The standard U-net architecture has a receptive field of 140x140 pixels, corresponding to the largest red square in Fig. 7. An increase in contextual information is required to account for large deformations in the retina, and to allow correct segmentation of large IRC lesions. We therefore increase the receptive field by adding two additional downsampling steps. The receptive field size increases to 572x572 pixels, corresponding to the largest green square in Fig. 7, i.e., the
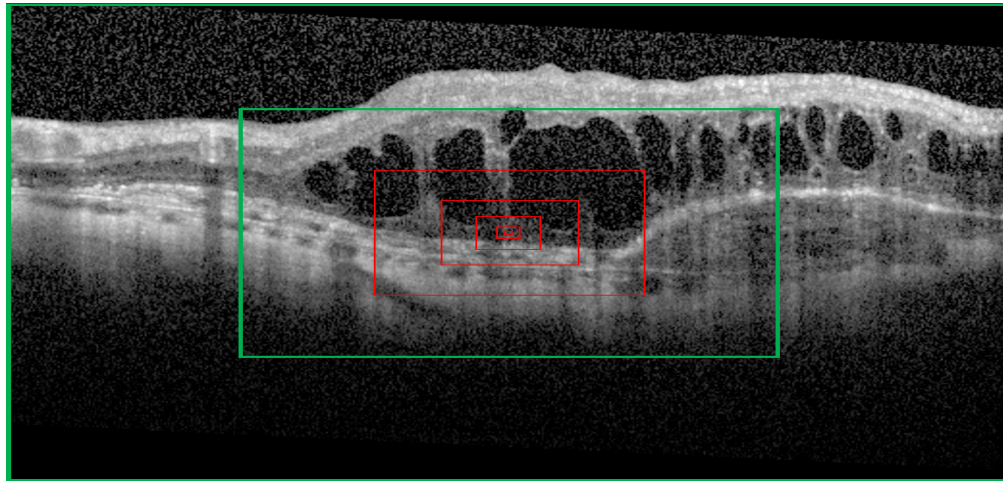
Fig. 7. Receptive field (RF) sizes after a max pooling operation. With every consecutive max pooling operation the RF increases in size, allowing the image to be analyzed at multiple scales. The RFs at each level for the original U-net architecture are shown in red, while the RFs for the proposed method are shown in green, i.e., By adding two *additional* max pooling operations the RF for the original U-net (the red squares) can be increased to include the entire image at the highest scale. Due to the anisotropic pixel size, the receptive field covers a larger area in the transversal direction.

context is as large as the entire B-scan.

2.2.4. FCNN training procedure

The network parameters of the defined FCNNs were optimized by training on randomly selected samples from the **training data**. The two models were individually trained until convergence was reached. Convergence was determined by calculating the performance on the independent **validation set** at regular intervals during training. Since the proposed network belongs to the subclass of FCNNs, i.e., does not contain any fully connected layers, learning and inference are performed using whole images by dense feedforward computation and backpropagation, increasing processing speed compared to patch-based networks. At every iteration, a small subset of B-scans (mini-batch) and the corresponding IRC annotations were randomly selected from the **training data** and used to optimize the system parameters. To maximize the throughput of the GPU the mini-batch size was set to process six B-scans at the same time. To increase robustness of the system, we increase the variance in the training dataset by applying an extensive data augmentation strategy. The following data augmentations are randomly applied to every B-scan selected for training:

- Random cropping of a 512x512 subimage from a B-scan (additional zero padding if necessary)

- Random rotation between -15 and +15 degrees

- Random mirroring

- Random multiplicative speckle noise with a magnitude between 0 and 0.4

The limits have been selected in such a way that after augmentation the resulting B-scan is a plausible example observable in clinical practice. The data augmentation process is visualized in Fig. 8.
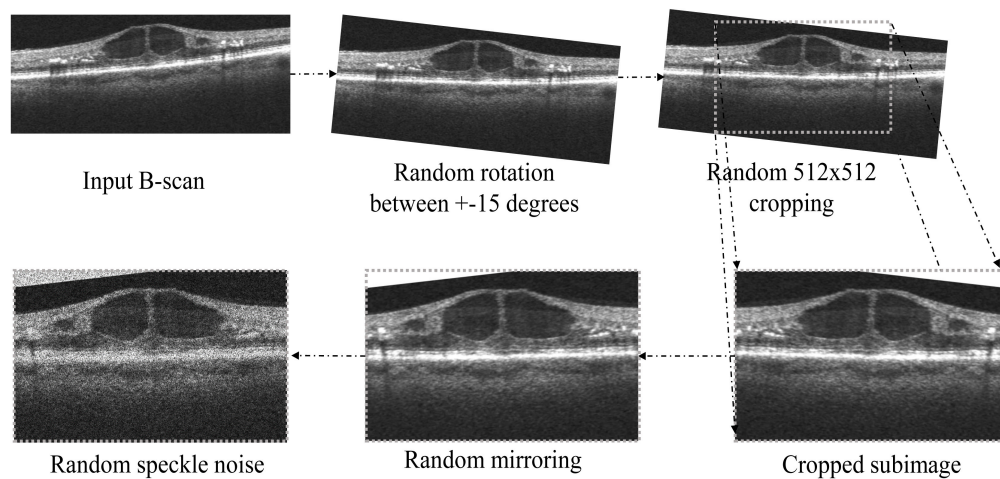
Fig. 8. Data augmentation strategy. The training dataset is synthetically increased by data augmentation. The following augmentations are successively applied: random rotation, random cropping, random mirroring and speckle noise addition.

The proposed CNN was trained using stochastic gradient descent with a learning rate starting at 10^{-3} up to 10^{-6}. The learning rate was manually decreased by a factor of 10 whenever a plateau was reached in the learning curve. Momentum was set to 0.99 in order to include a large amount of previously seen samples in the current update step. He-normal initialization was used for weight initialization of the network. The network architecture and training procedures were implemented in Python 2.7 using the Theano [51] and Lasagne [52] packages for the deep learning framework. Training of the network was performed on a Nvidia titan X GPU using CUDA 7.5 with cuDNN v4 and CNMeM optimizations enabled. Convergence of the network was reached in about 36 hours. When the training procedure is finished, the time required to produce a IRC segmentation is about 5 seconds for a complete OCT volume.

2.3. Evaluation design

We perform three different experiments to evaluate the performance of the proposed deep learning algorithm in three different scenarios.

2.3.1. Segmentation of IRC

To assess whether we can accurately segment IRC we compare the segmentation performance of our proposed algorithm to manual annotations by human observers. Segmentation performance is quantified using two different metrics: 1) the Dice similarity coefficient (DC), a statistic that quantifies the overlap between two binary images, e.g., ground truth and segmentation output, and 2) the area segmentation error (ASE), i.e., the difference between the area annotated in the ground truth and the segmented area by the proposed method in mm^2. The DC is assessed on the total volume, while the area segmentation error is assessed on a B-scan level.

To assess whether the addition of the retina segmentation as a means to include prior information is effective, we perform four different experiments: 1) using the retina segmentation as a weight map during training; 2) using the retina segmentation as an additional input to the system; 3) using both methods to include prior information simultaneously; and 4) without using the retina segmentation. These experiments are performed in the **first test set** containing 53 SD-OCT volumes, the **second test set** containing 10 SD-OCT volumes, and the **third test set** containing 33 OCT volumes of healthy controls.
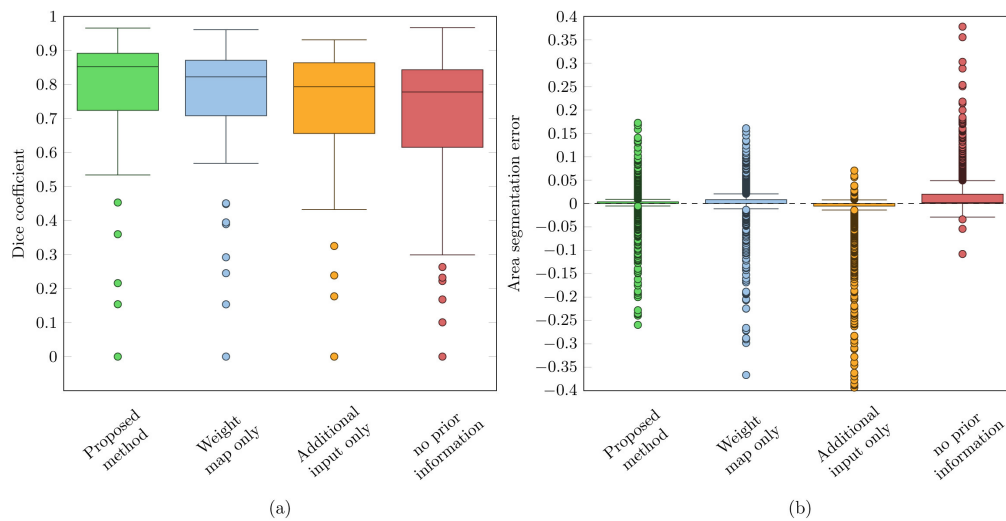
Fig. 9. Boxplots showing the distribution of (a) the dice coefficients and (b) the area segmentation errors in **test set 1**. Four different approaches to include prior information are compared, i.e., without using prior information (red), using the retina segmentation as an additional input (yellow), using the retina segmentation as a weight map during training (blue), and finally the proposed method where both techniques to include prior information are used together (green).

2.3.2. Quantification of IRC volume

To investigate the ability of the proposed method to make an accurate quantification of the volume of IRC in the retina, we compare the total segmented volume in µl to the volume annotated by independent human observers. This experiment is performed in the **second test set** consisting of 50 B-scans from 10 independent SD-OCT volumes. Volume quantification performance is assessed by the intraclass correlation coefficient (ICC) and Pearson's correlation coefficient (ρ), two commonly used summarizing measures to indicate how well the proposed method and the two observers resemble the ground truth IRC volume. Finally, the volume differences are visualized using Bland-Altman plots comparing the proposed method and the two observers to the ground truth. This experiment is performed for all four approaches to include prior information.

2.3.3. Generalization to multiple OCT vendors

To assess the performance of the system to segment SD-OCT volumes acquired by different SD-OCT devices, we compare the output of the system to manual annotations by two human observers in the **external dataset** consisting of SD-OCT volumes acquired by four different SD-OCT devices. The proposed method was retrained on the EUGENDA dataset extended with the training data from the **external dataset** (15 OCT volumes, 3-4 cases per vendor) Finally, the analysis is performed on the 15 OCT volumes from the test set in the **external dataset**.

The output of the proposed system is compared to reference standard 1 (i.e., the intersection of both observers), and reference standard 2 (i.e., the intersection of both observers where pixels from which the observers disagreed were not considered for the evaluation). This means that IRC segmented by the proposed method that overlaps with a region annotated by only a single observer does not contribute to oversegmentation, i.e., are considered as not segmented. This reference standard allows evaluation of the method on a more certain ground truth, and does not penalize segmented IRC in regions for which observers are uncertain.

Using these two reference standards, we report the average and median DC. Additionally, we

Table 1. Dice coefficients, i.e., mean ± standard deviation (median) and area segmentation error obtained on the **first test set** using four different approaches to include prior information, i.e., 1) by using the retina segmentation as a weight map during training; 2) using the retina segmentation as an additional input to the system ; 3) both methods to include prior information used simultaneously; and 4) without using the retina segmentation.

Strategy	Dice coefficient	Area segmentation error
No prior information	0.698 ± 0.237 (0.777)	$0.0187 \, \text{mm}^2 \pm 0.0389 \, \text{mm}^2$
Additional input only	0.727 ± 0.196 (0.793)	$-0.0296 \, \text{mm}^2 \pm 0.1070 \, \text{mm}^2$
Weight map only	0.748 ± 0.214 (0.822)	$0.002 \, 89 \, \text{mm}^2 \pm 0.0372 \, \text{mm}^2$
Proposed method	0.775 ± 0.208 (0.852)	$0.000 \, 812 \, \text{mm}^2 \pm 0.0320 \, \text{mm}^2$

include the inter-observer variability between the two observers to provide an indication of the maximum achievable human performance.

3. Results

3.1. Segmentation of IRC

The boxplots visualizing the DC and the ASE obtained when comparing the output of the system to the ground truth for all 53 SD-OCT volumes in the **first test set** are shown in Fig. 9. The proposed method achieves an average and median DC of 0.775 ± 0.208 and 0.852. When the prior information from the retina segmentation is not used an average and median DC obtained of 0.698 ± 0.237 and 0.777 is obtained, respectively.

Similarly for the ASE, the proposed method achieves an ASE of $0.000 \, 812 \, \text{mm}^2 \pm 0.032 \, \text{mm}^2$. When the prior information from the retina segmentation is not used an ASE of $0.0187 \, \text{mm}^2 \pm 0.0389 \, \text{mm}^2$ is obtained. Results obtained when each approach to include prior information is used individually are shown in Table 1.

The DC and segmentation error obtained on the **second test set** for the proposed method and the human observers are shown in Fig. 10. The proposed method achieves an average and median DC of 0.754 ± 0.136 and 0.778, respectively. The average and median DC obtained without using prior information are 0.701 ± 0.184 and 0.755, respectively. Observer 1 obtained an average and median dice of 0.783 ± 0.103 and 0.805, respectively, while observer 2 obtained an average and median dice of 0.783 ± 0.097 and 0.786, respectively.

When considering the ASE, the proposed method achieves an value of $0.022 \, \text{mm}^2 \pm 0.040 \, \text{mm}^2$ compared to the observers with an ASE of $0.0046 \, \text{mm}^2 \pm 0.030 \, \text{mm}^2$ and $-0.0185 \, \text{mm}^2 \pm 0.020 \, \text{mm}^2$ for observer 1 and observer 2 respectively. The ASE obtained without using prior information is $0.061 \, \text{mm}^2 \pm 0.066 \, \text{mm}^2$. Results obtained when each approach to include prior information is used individually are shown in Table 2.

Finally, the boxplots visualizing the segmentation error obtained in the **control test set** are shown in Fig. 11. As no IRC is present in this set, the DC can not be defined. For the ASE, the proposed method, which included both approaches to include prior information, achieves an ASE of $0.000 \, 17 \, \text{mm}^2 \pm 0.0011 \, \text{mm}^2$. When no prior information is exploited, an ASE of $0.000 \, 36 \, \text{mm}^2 \pm 0.0015 \, \text{mm}^2$ is obtained. Results obtained when each approach to include prior information is used individually are shown in Table 3.

3.2. Quantification of IRC volume

Figure 12 shows the Bland-Altman plots visualizing the volume difference on the **second test set** when comparing the output segmentation to the manual ground truth for each of the four different approaches to include prior information. The quantification performance of the two human observers is also included. Furthermore, in Table 4 the performance for each approach to include prior information is shown.
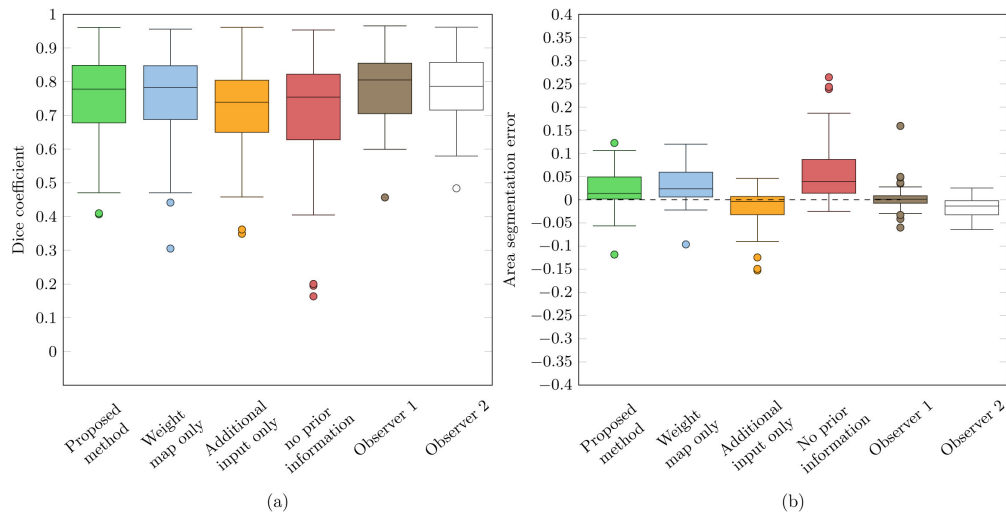
Fig. 10. Boxplots showing the distribution of (a) the dice coefficients and (b) the area segmentation errors in **test set 2** Four different approaches to include prior information are compared, i.e., without using prior information (red), using the retina segmentation as an additional input (yellow), using the retina segmentation as a weight map during training (blue), and finally the proposed method where both techniques to include prior information are used together (green). The performance of the two human observers is also visualizes (brown, white)

Table 2. Dice coefficients, i.e., mean ± standard deviation (median) and area segmentation error obtained on the **second test set** using four different approaches to include prior information, i.e., 1) by using the retina segmentation as a weight map during training; 2) using the retina segmentation as an additional input to the system ; 3) both methods to include prior information used simultaneously; and 4) without using the retina segmentation.

Strategy	Dice coefficient	Area segmentation error
No prior information	0.701 ± 0.184 (0.755)	0.061 mm^2 ± 0.066 mm^2
Additional input only	0.719 ± 0.135 (0.739)	−0.027 mm^2 ± 0.087 mm^2
Weight map only	0.754 ± 0.137 (0.783)	0.031 mm^2 ± 0.040 mm^2
Proposed method	0.754 ± 0.136 (0.778)	0.022 mm^2 ± 0.040 mm^2
Observer 1	0.783 ± 0.103 (0.805)	0.0046 mm^2 ± 0.030 mm^2
Observer 2	0.783 ± 0.097 (0.786)	−0.0185 mm^2 ± 0.020 mm^2

The proposed method obtained an average absolute volume difference (AVD) of 0.0334 µl ± 0.0324 µl, where observer 1 and observer 2 obtained an average AVD of 0.0175 µl ± 0.0253 µl and 0.0202 µl ± 0.0181 µl, respectively. The systematic difference for the proposed method is 0.0221 µl with the 95% limits of agreement at −0.0581 µl and 0.1022 µl, for the lower and upper bound, respectively. The systematic difference for observer 1 and observer 2 was −0.0186 µl and 0.0047 µl, respectively. The lower and upper bound of the 95% limits of agreement for observer 1 were at −0.0572 µl and 0.0200 µl, and at −0.0549 µl and 0.0642 µl for observer 2.

Overall agreement with the ground truth reference volume was established using the ICC and Pearson's correlation coefficient (ρ). The proposed method achieved an ICC (ρ) of 0.936 (0.949), where observer 1 and observer 2 obtained an ICC (ρ) of 0.978 (0.990), and 0.975 (0.978), respectively. Finally, in Fig. 13 qualitative results and the corresponding annotations by the two graders are shown for two representative cases.
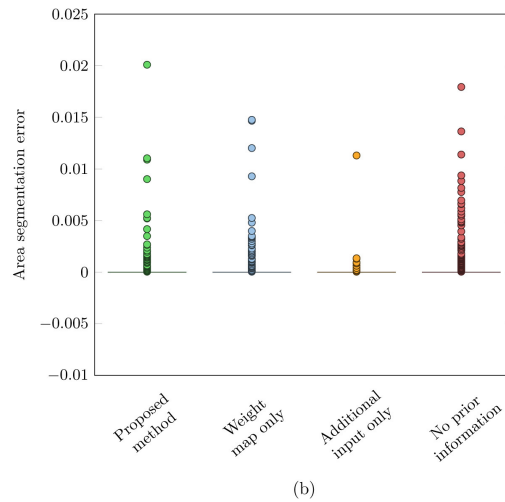
(b)

Fig. 11. Boxplots showing the distribution of the area segmentation errors in the control test set. Four different approaches to include prior information are compared, i.e., without using prior information (red), using the retina segmentation as an additional input (yellow), using the retina segmentation as a weight map during training (blue), and finally the proposed method where both techniques to include prior information are used together (green).

Table 3. Area segmentation error obtained on the **control test set** using four different approaches to include prior information, i.e., 1) by using the retina segmentation as a weight map during training; 2) using the retina segmentation as an additional input to the system ; 3) both methods to include prior information used simultaneously; and 4) without using the retina segmentation.

Strategy	Area segmentation error
No prior information	$0.000\,36\,\text{mm}^2 \pm 0.001\,50\,\text{mm}^2$
Additional input only	$0.000\,025\,\text{mm}^2 \pm 0.000\,44\,\text{mm}^2$
Weight map only	$0.000\,19\,\text{mm}^2 \pm 0.001\,10\,\text{mm}^2$
Proposed method	$0.000\,17\,\text{mm}^2 \pm 0.001\,10\,\text{mm}^2$

3.3. Generalization to multiple OCT vendors

When applying the proposed method trained on EUGENDA data, extended with additional vendor specific training data from the **external dataset**, an average Dice coefficient of 0.738 ± 0.159 is obtained when comparing against the intersection of the two observers (reference standard 1). When excluding the regions where the two observers disagree from the evaluation (reference standard 2) an average dice coefficient of 0.786 ± 0.174 is obtained. The dice coefficients obtained for the subgroups of vendor specific data are shown in Table 5. Finally, example segmentation results and the corresponding annotations by the two graders are shown for B-scans acquired by a Cirrus, Nidek, Spectralis and Topcon scanner, in Fig. 14, Fig. 15, Fig. 16 and Fig. 17, respectively.

4. Discussion

In this study we assessed the performance of a deep learning algorithm for the automated detection, segmentation and quantification of IRC in SD-OCT imaging. A novelty of the developed system is the applicability in SD-OCT volumes acquired by SD-OCT devices from different vendors with widely varying imaging quality and image characteristics. Furthermore, the proposed method implements a multiscale analysis of an input B-scan, allowing accurate segmentation of both
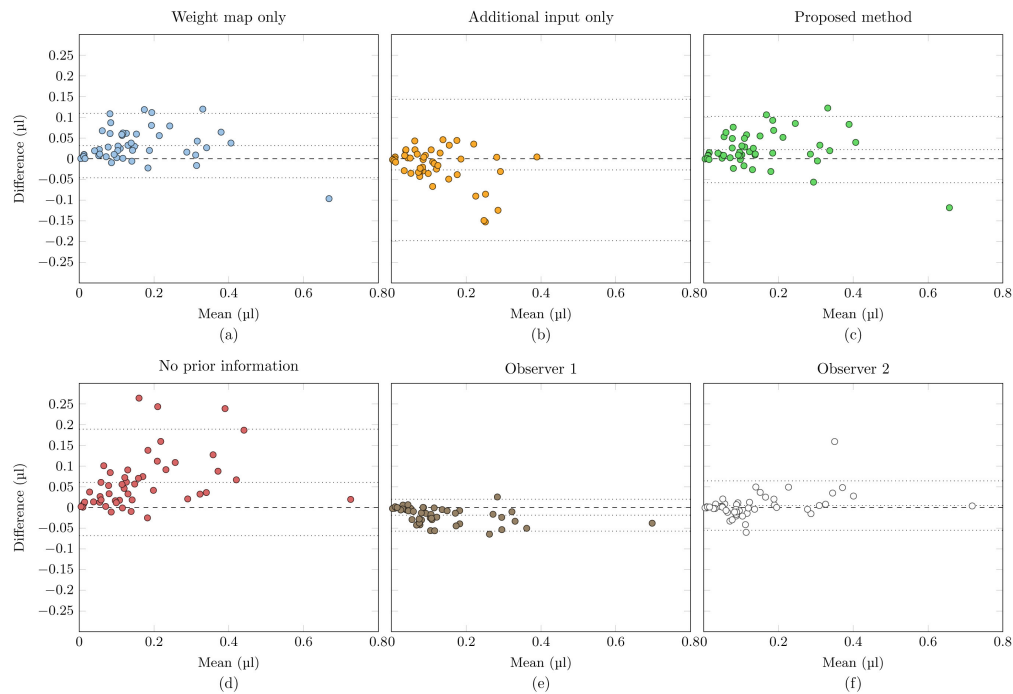
Fig. 12. Bland altman plots visualizing the volume difference on the **second test set** using four different approaches to include prior information, i.e., (a) by using the retina segmentation as a weight map during training; (b) by using the retina segmentation as an additional input to the system; (c) by using both methods to include prior information simultaneously; and (d) without using the retina segmentation. The performance of the two human observers is shown in (e) and (f), respectively.

Table 4. Absolute volume difference, intraclass correlation coefficient (ICC) an Pearson's correlation coefficient (ρ) on the **second test set** using four different approaches to include prior information, i.e., 1) by using the retina segmentation as a weight map during training; 2) by using the retina segmentation as an additional input to the system; 3) by using both methods to include prior information simultaneously; and 4) without using the retina segmentation. The performance of the two human observers is also included.

Strategy	Average AVD	ICC	ρ
No prior information	0.0624 µl ± 0.0638 µl	0.822	0.906
Additional input only	0.0411 µl ± 0.0815 µl	0.663	0.748
Weight map only	0.0624 µl ± 0.0638 µl	0.925	0.952
Proposed method	0.0334 µl ± 0.0324 µl	0.936	0.949
Observer 1	0.0175 µl ± 0.0253 µl	0.978	0.990
Observer 2	0.0202 µl ± 0.0181 µl	0.975	0.978

small and large intraretinal cysts. Finally, segmentation performance is improved by embedding a priori information about retinal anatomy in the algorithm.

The proposed algorithm was trained and validated in a set of 122 SD-OCT volumes, and finally evaluated in three independent test sets of which the first contains 53 semi-automatically annotated SD-OCT volumes, the second contains 10 manually annotated SD-OCT volumes, and the third contains 33 SD-OCT volumes acquired from healthy controls. The SD-OCT volumes were all acquired using a Heidelberg Spectralis SD-OCT device. The proposed algorithm was applied in all test sets using the Dice similarity coefficient and the area segmentation error as

Table 5. Dice coefficients, i.e., mean ± standard deviation (median), of the proposed method compared against reference standard 1 and reference standard 2 as defined in section 2.3.3

	Cirrus	Nidek	Spectralis	Topcon	Overall
Reference standard 1	0.703 ± 0.220 (0.807)	0.659 ± 0.188 (0.790)	0.775 ± 0.067 (0.747)	0.796 ± 0.072 (0.817)	0.738 ± 0.159 (0.791)
Reference standard 2	0.748 ± 0.238 (0.865)	0.730 ± 0.235 (0.895)	0.804 ± 0.067 (0.779)	0.849 ± 0.056 (0.865)	0.786 ± 0.174 (0.864)
Interobserver variability	0.838 ± 0.022 (0.832)	0.840 ± 0.032 (0.818)	0.861 ± 0.020 (0.862)	0.843 ± 0.083 (0.864)	0.846 ± 0.049 (0.839)

metrics for performance evaluation. The boxplots visualizing the distributions of obtained Dice coefficients for the **first test set** and the **second test set** are shown in Fig. 9(a) and Fig. 10(a), respectively. Similarly, the area segmentation error is visualized in Figs. 9(b) and 10(b). To assess whether the added retina segmentation step in the proposed algorithm provides additional information to improve segmentation performance, four different experiments were performed: 1) the retina segmentation was used as a weight map during training; 2) the retina segmentation was used as an additional input to the system; 3) both methods to include prior information were used simultaneously; and 4) without using the retina segmentation.

Without including the retina segmentation an average Dice coefficient of 0.698 ± 0.237 was obtained in the **first test set**, whereas the Dice coefficient increased to 0.775 ± 0.208 when both methods to include prior information are used. Similarly, for the **second test set**, the Dice coefficient increased from 0.701 ± 0.184 to 0.754 ± 0.136, when using both approaches to include prior information. These result supports the hypothesis that the information contained in the retina segmentation can be exploited to improve segmentation performance.

This claim is further supported by the observed decrease in the area segmentation error when including the retina segmentation. In the **first test set**, the average ASE drops from $0.0187 \, \text{mm}^2 \pm 0.0389 \, \text{mm}^2$ to $0.000\,812 \, \text{mm}^2 \pm 0.032 \, \text{mm}^2$ when using both methods to include prior information. The same effect can also be observed in the **second test set**, where the ASE improves from $0.061 \, \text{mm}^2 \pm 0.066 \, \text{mm}^2$ to $0.022 \, \text{mm}^2 \pm 0.040 \, \text{mm}^2$.

When looking at the performance gain for each of the two approaches to include prior information individually, it can be concluded that using the retina segmentation as a weight map during training has a bigger impact on performance compared to adding the retina segmentation as an additional input. However, the best performance is still obtained when both approaches are used simultaneously.

Several outliers with low Dice coefficients are observer in the **first test** set as indicated in Fig. 9. While the use of the retina segmentation offers some improvement, there are still cases for which the Dice coefficient is unsatisfactory. After visual inspection of the outliers it was found that these cases typically only contained small cysts that were either missed by the proposed method, or for which the output segmentation was slightly different from the ground truth annotation. The Dice coefficient is known to be very sensitive when the object to be segmented is small, strongly penalizing small errors resulting in a low dice coefficient for these specific cases. An interesting outlier is the case for which a dice coefficient of zero was obtained, this case is shown in Fig. 18. The proposed method detected a IRC like structure that was not present in the ground truth annotation. After closer inspection, the method mistakenly segmented a small outer retinal tubulation (ORT), retinal pathology with visual characteristics similar to IRC. This type of error can be explained by the underrepresentation of ORT in the training set. Adding cases with IRC confounders like ORT will likely resolve this issue.

In the **second test**, in addition to a single reference standard, two additional human graders manually annotated the cases, both achieving an average Dice coefficient of 0.783 ± 0.103 and 0.783 ± 0.097 when comparing them to the reference standard. While human performance is
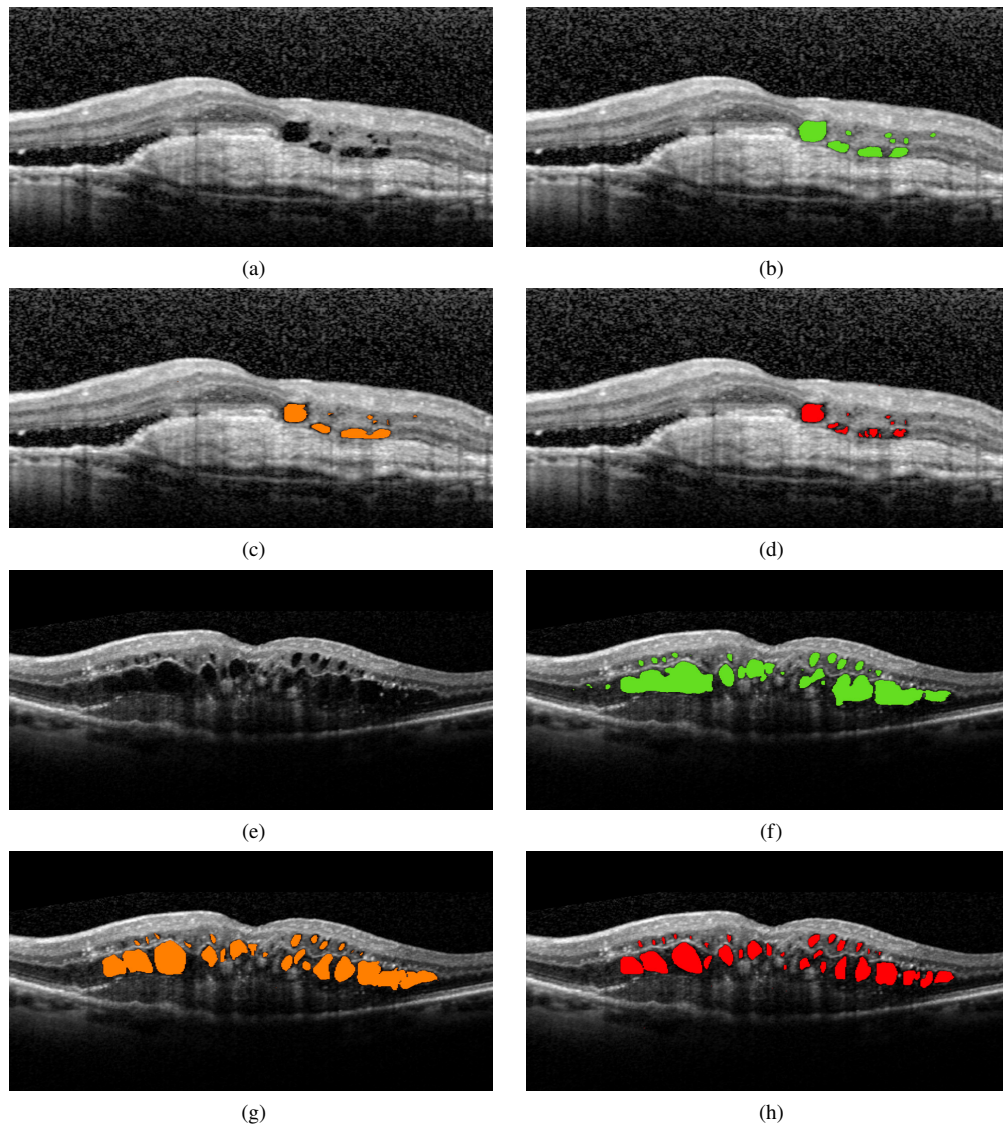
(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 13. Example segmentation results of the proposed algorithm applied to B-scans from the **second test** set acquired with a **Spectralis** device: (a, e) Original input image, (b, f) segmentation output from the proposed algorithm (green), (c, g) manual annotations performed by human observer 1 in orange and (d, h) manual annotations performed by human observer 2 in red.

still higher with respect to the proposed method 0.754 ± 0.136, the performance gap is small. Qualitative results showing the segmentation output from the proposed method and the manual annotations by the two human observers are shown in Fig. 13.

While segmentation in itself is an important task, fluid quantification might arguably be more clinically relevant as the total volume can be directly used as a (prognostic) biomarker. We therefore compare the total segmented IRC volume for the proposed method and both observers to the annotated ground truth volume. The proposed method achieves an average absolute volume
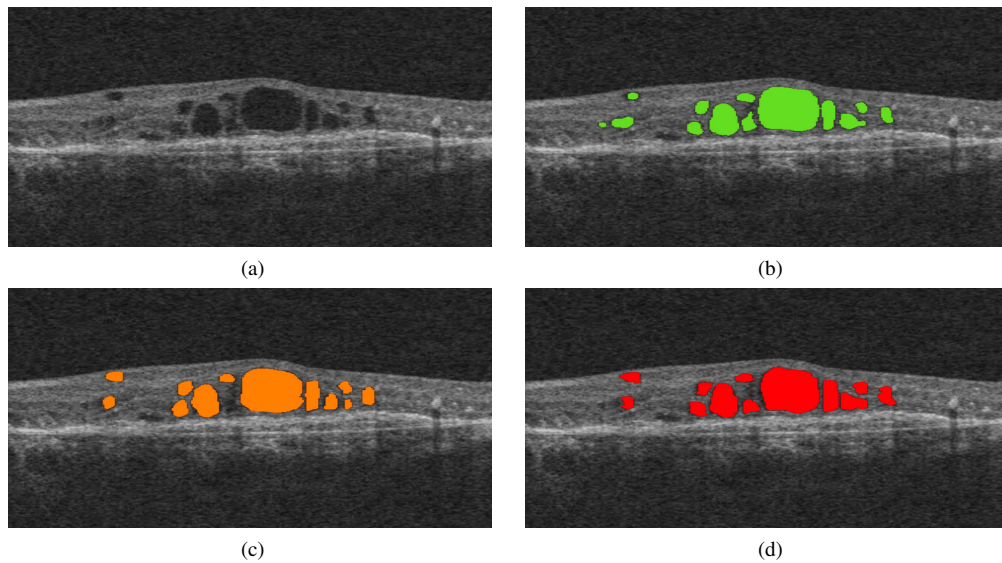
(a)

(b)

(c)

(d)

Fig. 14. Example segmentation results of the proposed algorithm applied to a B-scan acquired with a **Cirrus** device. (a) Original input image, (b) segmentation output from the proposed algorithm in green, (c) manual annotations performed by human observer 1 in orange, (d) manual annotations performed by human observer 2 in red.
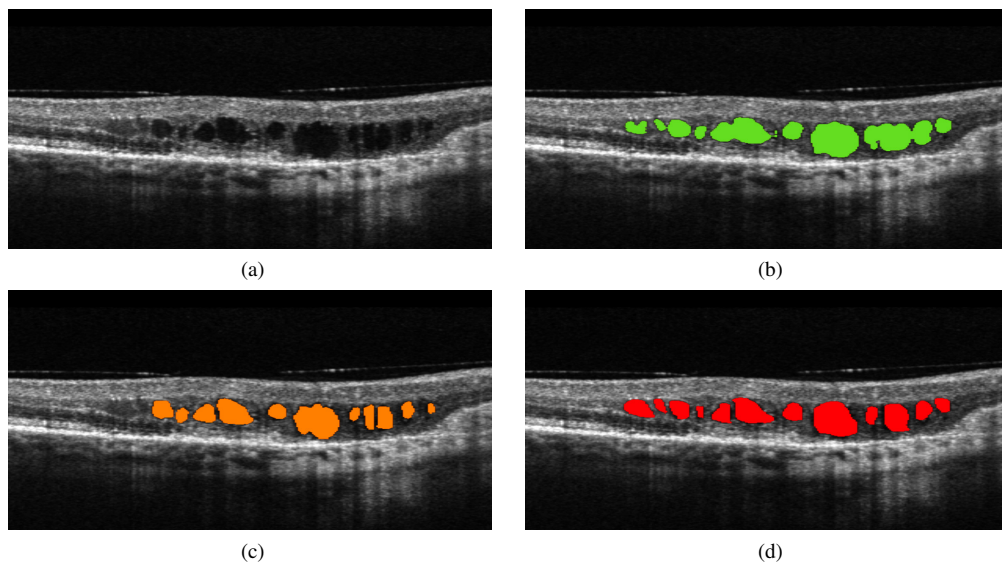

(a)

(b)

(c)

(d)

Fig. 15. Example segmentation results of the proposed algorithm applied to a B-scan acquired with a **Nidek** device. (a) Original input image, (b) segmentation output from the proposed algorithm in green, (c) manual annotations performed by human observer 1 in orange, (d) manual annotations performed by human observer 2 in red.

difference (AVD) of $0.0334\,\mu l \pm 0.0324\,\mu l$, where observer 1 and observer 2 obtained an AVD of $0.0175\,\mu l \pm 0.0253\,\mu l$ and $0.0202\,\mu l \pm 0.0181\,\mu l$, respectively. Although the automated volume
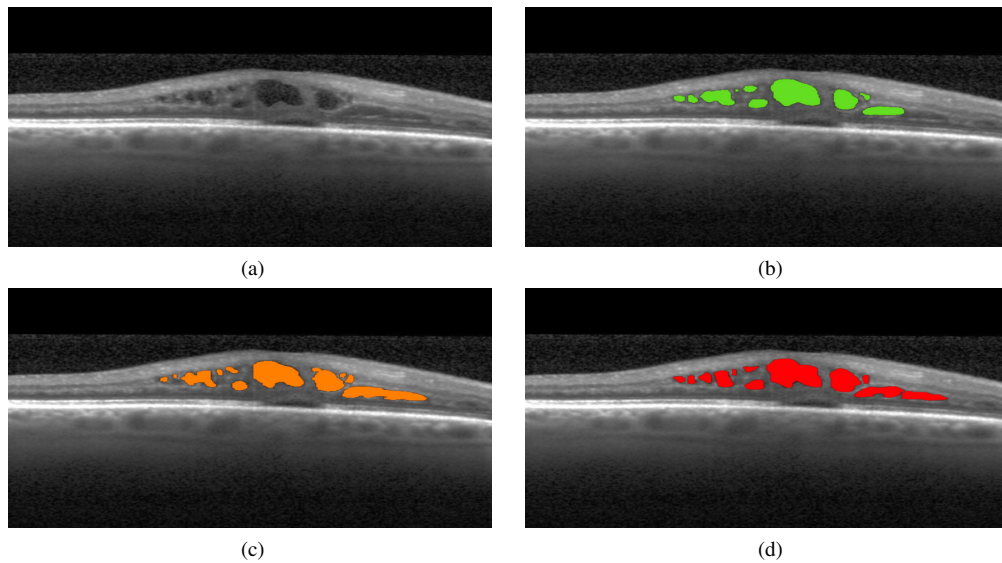
Fig. 16. Example segmentation results of the proposed algorithm applied to a B-scan acquired with a **Spectralis** device. (a) Original input image, (b) segmentation output from the proposed algorithm in green, (c) manual annotations performed by human observer 1 in orange, (d) manual annotations performed by human observer 2 in red.
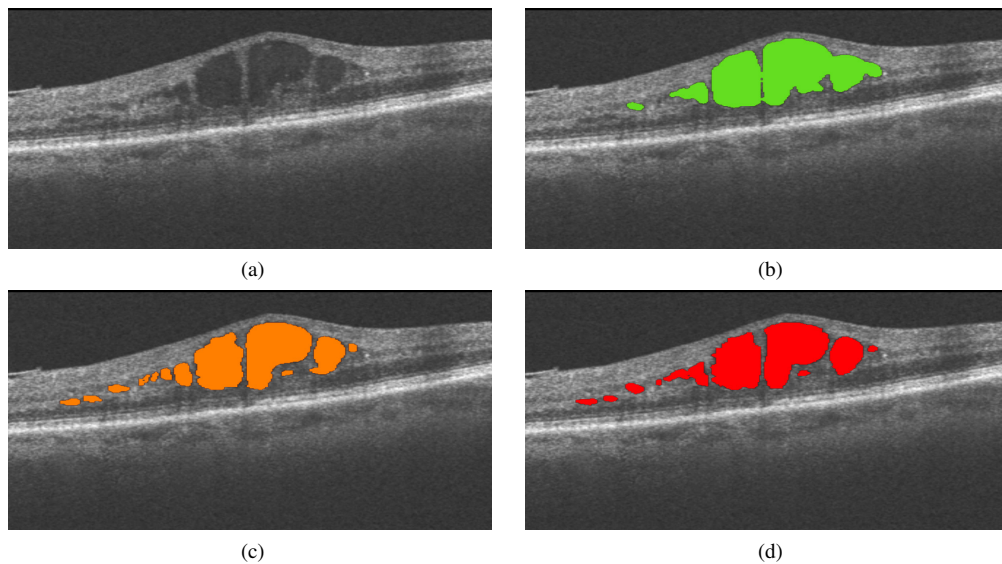

Fig. 17. Example segmentation results of the proposed algorithm applied to a B-scan acquired with a **Topcon** device. (a) Original input image, (b) segmentation output from the proposed algorithm in green, (c) manual annotations performed by human observer 1 on orange, (d) manual annotations performed by human observer 2 in red.

quantification is not at human performance yet, the ICC and Pearson's correlation coefficient (ρ) of 0.936 (0.949) for the proposed method shows that a strong correlation is found with the
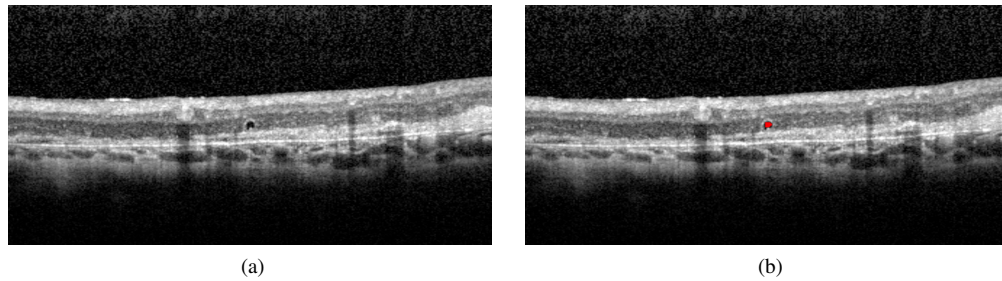
<div align="center">(a)   (b)</div>

Fig. 18. Example case where the proposed method segmented an outer retinal tubulation, a retinal pathology with a strong similarity to IRC, resulting in a dice coefficient of zero. (a) Ground truth image, (b) segmentation output by the proposed algorithm.

manually annotated ground truth volume. This ICC is comparable to the two observers with ICC (ρ) values of 0.975 (0.978) and 0.978 (0.990), respectively. This suggests that our proposed method can serve as a reliable tool that will produce a fluid quantification similar to human performance in a fraction of the time. This statement is further supported by the Bland-Altman plots in Fig. 12, showing quantification performance that is highly correlated to both human observers. Moreover, the obtained Pearson's correlation coefficient obtained by our proposed method (0.949) compares favorably to the performance reported on IRC quantification in a recently published method [44] (0.86) using a similar deep learning approach and evaluated on data acquired using the same OCT device (Heidelberg Spectralis).

In addition to the two private test sets, the performance of the proposed algorithm was also assessed in a publicly available **external dataset** consisting of SD-OCT volumes acquired by four different SD-OCT devices. When applying the proposed method, which was retrained on EUGENDA data extended with a small amount of vendor specific data (3-4 cases per vendor), an average Dice coefficient of 0.786 ± 0.174 is obtained based on the analysis of the regions for which both observers are in agreement. The results for the individual SD-OCT devices are shown in Table 1. The level of performance obtained is approaching the interobserver variability between the two observers of 0.846 ± 0.049.

It is interesting to note that *without* retraining, i.e., without adding vendor specific data to the training set, the proposed method already generalizes well across different datasets reaching an average Dice coefficient of 0.721 ± 0.196. It is surprising that without specifically optimizing for data acquired by different SD-OCT devices, the proposed method already produces adequate results. This indicates that without optimizing for data acquired by a specific SD-OCT device the proposed method already generalizes well across different datasets. Moreover, by adding only a few vendor specific SD-OCT volumes the proposed method is able to transfer and apply the knowledge obtained from one vendor to data from another vendor efficiently.

In Table 1, a difference in performance can be observed when assessing the performance of the proposed system on each vendor independently. Especially the large variance in segmentation performance for the Cirrus and Nidek device is apparent, both having cases for which the dice coefficient is substantially lower compared to the average performance. This variability in performance can largely be attributed to the varying imaging quality obtained by the Cirrus and Nidek devices. High noise and low image contrast makes IRC segmentation substantially more difficult. An example Cirrus case is shown in Fig. 19, where poor image contrast caused the proposed algorithm to fail in detecting IRC. However, when image quality is sufficient, the performance is similar or even higher compared to the performance on data from the other SD-OCT devices as indicated by the median Dice coefficients of 0.865 and 0.895 in Table 1 for the Cirrus and Nidek device, respectively. Qualitative results shown in Fig. 14 and Fig. 15 also
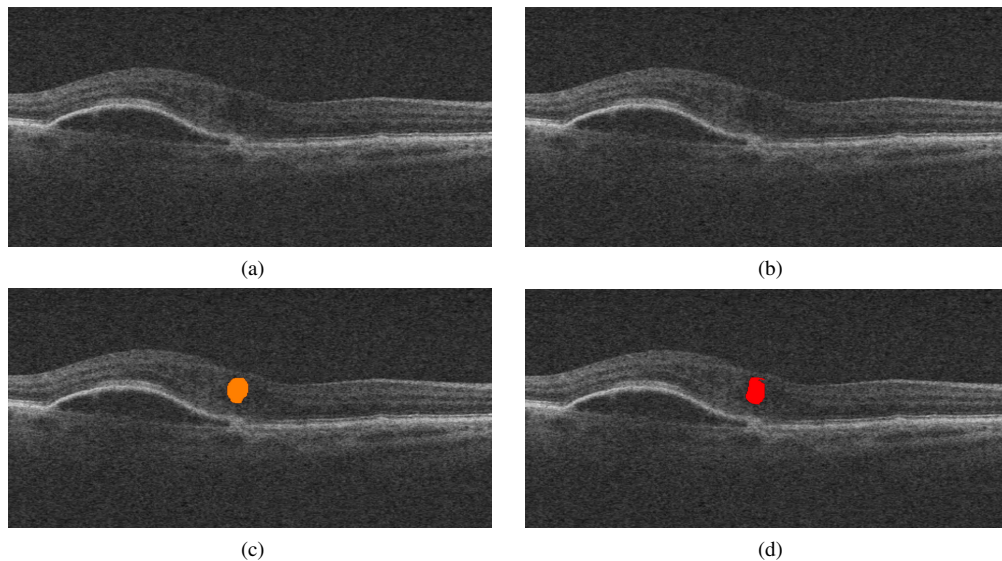
Fig. 19. Example case where the proposed method missed an annotated cyst due to very poor image contrast in a **Cirrus** images from the **external dataset**. (a) Original input image, (b) segmentation output from the proposed algorithm (no IRC detected), (c) ground truth annotations performed by human observer 1 in orange, (d) ground truth annotations performed by human observer 2 in red.

show a good correspondence to both of the human observers for the Cirrus and Nidek scanner, respectively. It has to be noted that preprocessing steps to remove image noise and improve image quality are typically used to improve performance. Instead of removing noise from the images, we chose to apply noise augmentation during the training phase of the system, i.e., random noise is randomly added to an input image to make the system more robust to variations in noise [53]. A more advanced denoising step might however give rise to an increase in performance.

As shown in Table 1 the performance on the Spectralis and Topcon scanners is more consistent, with a lower variance in the obtained dice coefficients, and approaches the interobserver variability. The average Dice coefficient compared to the intersection of both observers in the subset op Topcon cases reached a value of 0.796 ± 0.072, where an average Dice coefficients of 0.775 ± 0.067 was obtained for the Spectralis cases. This compares favorably to the results by two previously published methods that were evaluated on the same external dataset, i.e., a three-dimensional curvelet based K-SVD approach that obtained an average Dice coefficient of 0.652 [32], and a method built around a transformation to the neutrosophic domain with a Dice coefficient of 0.705 [35]. Furthermore, these other methods were not evaluated in SD-OCT volumes from vendors other than the Heidelberg Spectralis. Finally, visual examples of our performance in the Spectralis and Topcon dataset are shown in Fig. 16 and Fig. 17, respectively.

It has to be noted that the performance of the proposed method is only evaluated in a dataset containing IRC as a result of AMD. In future work a more extensive evaluation of the method on IRC as a result of DME and RVO will be included. Nonetheless, considering the performance of the proposed method, which approaches the interobserver variability, a few clinical applications are to be considered or are within reach. One such application would be the automated tracking of fluid volume changes in consecutive SD-OCT volumes. Quantitative volume measures can help guide PRN and TE treatment regimens, where fluid change is the main determining factor in the decision for retreatment. Volume quantification can detect non-response in an early stage, allowing

for improved personalized patient care, resulting in better patient outcome and a reduction of the financial burden anti-VEGF treatment places on the healthcare system. Another possible application of the proposed method is in the analysis of genotype-phenotype correlations in large population studies. Selecting homogeneous subgroups based on quantification of IRC, is time-consuming or even infeasible. The proposed method with performance in the range of human graders allows for fast processing of such large datasets.

In conclusion, we developed a fully automated system to detect and segment IRC in SD-OCT volumes independent of the SD-OCT device used for acquisition. The system proved to have excellent performance compared to that of two expert human observers on data from four different SD-OCT vendors. The proposed method allows for fast quantitative volume measurements that can be used to improve patient care, reduce costs, and allow fast and reliable analysis in large population studies. Our automatic approach may be considered to be applied as a fast, reliable and cost-effective method in retinal research and clinical practice.

Disclosures

The authors declare that there are no conflicts of interest related to this article.