Medical Image Analysis 91 (2024) 103019

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Simultaneous alignment and surface regression using hybrid 2D–3D networks for 3D coherent layer segmentation of retinal OCT images with full and sparse annotations

Hong Liu [a,b,c,1], Dong Wei [c,1], Donghuan Lu [c], Xiaoying Tang [d], Liansheng Wang [a,*], Yefeng Zheng [c]

[a] School of Informatics, Xiamen University, Xiamen 361005, China
[b] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China
[c] Jarvis Research Center, Tencent YouTu Lab, Shenzhen 518075, China
[d] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China

ARTICLE INFO

Keywords:
Optical coherence tomography
3D coherent layer segmentation
B-scan alignment
2D–3D hybrid network
Semi-supervised learning with sparse annotation

ABSTRACT

Layer segmentation is important to quantitative analysis of retinal optical coherence tomography (OCT). Recently, deep learning based methods have been developed to automate this task and yield remarkable performance. However, due to the large spatial gap and potential mismatch between the B-scans of an OCT volume, all of them were based on 2D segmentation of individual B-scans, which may lose the continuity and diagnostic information of the retinal layers in 3D space. Besides, most of these methods required dense annotation of the OCT volumes, which is labor-intensive and expertise-demanding. This work presents a novel framework based on hybrid 2D–3D convolutional neural networks (CNNs) to obtain continuous 3D retinal layer surfaces from OCT volumes, which works well with both full and sparse annotations. The 2D features of individual B-scans are extracted by an encoder consisting of 2D convolutions. These 2D features are then used to produce the alignment displacement vectors and layer segmentation by two 3D decoders coupled via a spatial transformer module. Two losses are proposed to utilize the retinal layers' natural property of being smooth for B-scan alignment and layer segmentation, respectively, and are the key to the semi-supervised learning with sparse annotation. The entire framework is trained end-to-end. To the best of our knowledge, this is the first work that attempts 3D retinal layer segmentation in volumetric OCT images based on CNNs. Experiments on a synthetic dataset and three public clinical datasets show that our framework can effectively align the B-scans for potential motion correction, and achieves superior performance to state-of-the-art 2D deep learning methods in terms of both layer segmentation accuracy and cross-B-scan 3D continuity in both fully and semi-supervised settings, thus offering more clinical values than previous works.

1. Introduction

Optical coherence tomography (OCT) – a non-invasive imaging technique based on the principle of low-coherence interferometry – can acquire 3D cross-section images of human tissue at micron resolutions (Huang et al., 1991). Due to its micron-level axial resolution, non-invasiveness, and fast speed, OCT is commonly used in eye clinics for diagnosis and management of retinal diseases (Abràmoff et al., 2010). Notably, OCT provides a unique capability to directly visualize the stratified structure of the retina of cell layers, whose statuses are biomarkers of presence, severity, and prognosis for a variety of retinal and neurodegenerative diseases, including age-related macular degeneration (Keane et al., 2009), diabetic retinopathy (Bavinger et al., 2016), glaucoma (Kansal et al., 2018), Alzheimer's disease (Knoll et al., 2016), and multiple sclerosis (Saidha et al., 2011). Usually, layer segmentation is the first step in quantitative analysis of retinal OCT images, yet can be considerably labor-intensive, time-consuming, and subjective when done manually. Therefore, computerized tools for automated, prompt, objective, and accurate retinal layer segmentation in OCT images are highly desired.

Automated layer segmentation in retinal OCT images has long been explored. Earlier explorations (Garvin et al., 2009; Yazdanpanah et al., 2009; Lang et al., 2013) relied on empirical rules and/or hand-crafted features, which may be difficult to generalize. Recently, researchers started to implement deep convolutional neural networks (CNNs) for

[*] Corresponding author.
E-mail address: lswang@xmu.edu.cn (L. Wang).
[1] Hong Liu and Dong Wei contributed equally. Hong Liu contributed to this work during an internship at Tencent.

https://doi.org/10.1016/j.media.2023.103019
Received 11 July 2022; Received in revised form 28 August 2023; Accepted 31 October 2023
Available online 4 November 2023

retinal layer segmentation in OCT images and achieved superior performance to classical methods (He et al., 2019a; Xie et al., 2022a). However, most previous methods (both classical and CNNs) segmented each OCT slice (called a B-scan) separately given the relatively big inter-B-scan distance, despite the fact that a modern OCT sequence actually consists of many B-scans covering a volumetric area of the eye (Drexler and Fujimoto, 2008). Correspondingly, these methods failed to utilize the anatomical prior that the retinal layers are generally smooth surfaces (instead of independent curves in each B-scan), and may be subject to discontinuity in the segmented layers between adjacent B-scans, potentially affecting volumetric analysis following layer segmentation. Although some works (Antony et al., 2013; Carass et al., 2014; Chen et al., 2018; Garvin et al., 2009; Lang et al., 2013; Novosel et al., 2017) attempted 3D OCT segmentation, all of them belong to the classical methods that yielded inferior performance to the CNN-based ones, and overlooked the misalignment artifact of the B-scans in an OCT volume.

The inter-B-scan misalignment happens unavoidably mainly because of the involuntary eye movements during acquisition time (Sánchez Brea et al., 2019).[2] The motion artifacts may adversely affect qualitative interpretation and quantitative analysis of the images. For example, they may be mistaken for pathologies distorting the retinal pigment epithelium, and affect clinical decisions and the tracking of fine-grained features such as cysts or vessels between B-scans (Montuoro et al., 2014). In addition, the artifacts may distort the underlying 3D structures. Such distortion may pose challenges in some applications, e.g., multi-modality registration where an OCT en face image and a color fundus image need to be aligned (Cheng et al., 2016), 3D reconstruction and analysis of layer surface/thickness maps (Hood and Raza, 2014; Jáñez-Escalada et al., 2019), and 3D OCT segmentation where raw 3D operations may be invalidated in the presence of misalignment between B-scans.

Besides the motion artifact, another obvious obstacle to developing a CNN-based method for 3D OCT segmentation is the apparent anisotropy in resolution (Shah et al., 2018). For example, the physical resolutions of one of the datasets employed in this work are 3.24 μm (within A-scan, which is a column in a B-scan image), 6.7 μm (cross-A-scan), and 67 μm (cross-B-scan). Given the void of any existing CNN-based 3D OCT segmentation method, it is therefore not strange that the anisotropy problem has not been considered in such a context before.

In this work, we propose a novel CNN-based 2D–3D hybrid framework for simultaneous B-scan alignment and 3D surface regression for coherent retinal layer segmentation across-B-scan in OCT images. This framework consists of a shared 2D encoder followed by two 3D decoders (the alignment branch and the segmentation branch), and a spatial transformer module (STM; Balakrishnan et al., 2019) inserted to the shortcuts (Ronneberger et al., 2015) between the encoder and the segmentation branch. Given a B-scan volume as input, we employ per B-scan 2D operations for the encoder for two reasons. First, as suggested by previous studies (Zhang et al., 2019; Wang et al., 2020), intra-slice feature extraction followed by inter-slice (2.5D or 3D) aggregation is an effective strategy against anisotropic resolution, thus we propose a similar 2D–3D hybrid structure for the anisotropic OCT data. Second, the B-scans in the input volume are subject to misalignment, thus 3D operations across-B-scan prior to proper realignment may be invalid. Following the encoder, the alignment branch employs 3D operations to aggregate features across-B-scan to align them properly. Then, the resulting displacement vectors are employed to align the 2D features at different scales and compose well-aligned 3D features by the STM. These 3D features are passed to the segmentation branch for 3D surface regression. Noteworthily, the alignment only ensures validity of subsequent 3D operations, but not necessarily the cross-B-scan coherence of the regressed layer surfaces. Hence, we further impose a gradient-based, 3D regulative loss (Wei et al., 2018) on the regressed surfaces to encourage surface smoothness, which is an intrinsic property of many biological layers; we refer to this loss as the *global coherence loss*. While it is straightforward to implement the global coherence loss within our surface regression framework and comes for free (no manual annotation is needed), it proves effective in our experiments. The entire framework is trained end-to-end.

Last but not least, we are delighted to discover that our proposed 2D–3D framework can naturally handle a practical scenario of semi-supervised learning for OCT layer segmentation, where only a subset of B-scans in each OCT volume is manually annotated (i.e., sparse annotation). Owing to the introduction of the global coherence loss, layers segmented in non-annotated B-scans can be optimized according to their coherence with these layers in a neighborhood of B-scans—no matter annotated or not. In such scenario, the segmentation and alignment branches are tangled even more closely than in the fully supervised setting, mutually benefiting each other. Considering the labor-intensive and time-consuming nature of the manual layer annotation, effective semi-supervised learning is especially valuable. It should be noted that although few other works also attempted semi-supervised OCT layer segmentation (Liu et al., 2018a; Sedai et al., 2019), they all treated the B-scans as independent 2D images and relied on the notion of uncertainty/confidence. So far as we are aware of, our work is the first that bases semi-supervised OCT layer segmentation on 3D coherence of the layers.

In summary, our contributions are as follows:

- First, we propose a new framework for simultaneous B-scan alignment and 3D layer segmentation of retinal OCT images. This framework features a hybrid 2D–3D structure comprising a shared 2D encoder, a 3D alignment branch, a 3D surface regression branch, and an STM to allow for simultaneous alignment and 3D segmentation of anisotropic OCT data.
- Second, we further incorporate a conceptually straightforward and easy-to-implement regulating loss, the global coherence loss, to encourage the regressed layer surfaces to be coherent—not only within but also across-B-scan. Jointly, the first two contributions enable our framework to produce *more coherent layer surfaces in 3D* than existing state-of-the-art (He et al., 2019a, 2021; Xie et al., 2022a,b), as validated by our experiments. This advantage makes our framework preferred in applications where 3D fidelity of the segmented structures is crucial, e.g., 3D reconstruction and analysis of layer surface/thickness maps.
- Third, we extend the framework for semi-supervised learning where only a subset of B-scans in each OCT volume is annotated. Thanks to our novel design of coupled B-scan alignment and 3D layer segmentation, and the global coherence loss, the extension is straightforward yet remarkably effective. Experiments show that the performance advantages of our framework over other methods become more prominent with decreasing number of B-scans annotated and demonstrate its practical usability given sparse annotations. This capability of semi-supervised learning is valuable considering the effort and difficulty of manual labeling.

We conduct thorough experiments on three public OCT datasets as well as synthetic data, to evaluate effectiveness of the proposed framework, validate its design, and demonstrate its superiority toward existing methods in terms of both B-scan alignment and fully/semi-supervised segmentation.

This work is a comprehensive extension to our proof-of-concept exploration (Liu et al., 2021) in three main aspects, i.e., we (1) extend the framework to support semi-supervised learning, (2) additionally use synthetic data to quantify the B-scan alignment performance, and (3) employ two more public datasets to evaluate the generalization of the proposed framework.

[2] Intra-B-scan misalignment often can be ignored given the fast A-scan acquisition rate of modern spectral domain OCT systems (McNabb et al., 2012).

2. Related work

2.1. Retinal OCT segmentation

Earlier attempts at automated retinal layer segmentation in OCT images included graph based (Antony et al., 2013; Garvin et al., 2009; Lang et al., 2013), contour modeling (Carass et al., 2014; Novosel et al., 2017; Yazdanpanah et al., 2009), and machine learning (Antony et al., 2013; Lang et al., 2013) methods. For example, the graph theory and dynamic programming framework (Chiu et al., 2010), an inferential classic approach, modeled the layer segmentation problem as finding the shortest path in a graph representing the OCT image. Although greatly advanced the field, most of these classical methods relied on empirical rules and/or hand-crafted features which may be difficult to generalize. Motivated by the success of deep convolutional neural networks (CNNs) in a wide variety of medical image analysis tasks (Ker et al., 2017; Litjens et al., 2017; Shen et al., 2017), researchers also implemented CNNs for retina OCT segmentation and achieved superior performance to classical methods, mainly attributed to the data-driven automatic extraction of task-appropriate features. Fang et al. (2017) and Kugelman et al. (2018) conducted graph search on CNN-based probability maps. Liu et al. (2018b) trained a structured random forest classifier on integrated deep CNN and hand-crafted features. These works relied on patch-based classification of (the local neighborhood of) each pixel of interest. More recently, fully convolutional networks (FCNs) (Long et al., 2015) were applied to retina OCT segmentation, achieving great improvement in both efficiency and accuracy (Roy et al., 2017; Shah et al., 2018; He et al., 2019a, 2021). Li et al. (2021) employed graph based representations (Atif et al., 2007) to assist FCNs in exploiting anatomical prior knowledge and performing spatial reasoning. Xie et al. (2022a) proposed to explicitly enforce mutual surface interaction constraints with a graph model and realize simultaneous total surface cost minimization and surface order constraints with a primal–dual interior-point method (IPM). Xie et al. (2022b) proposed to integrate a constrained differentiable dynamic programming (DDP) module in end-to-end training to enforce surface smoothness. Our method also belongs to the FCN genre. However, distinct from all the FCN-based methods above which segmented individual B-scans as independent 2D images separately, our method segments all B-scans in the same OCT volume together in 3D, after aligning them properly within the same framework.

2.2. OCT motion correction

Correction of involuntary eye motion in retinal OCT can be accomplished with either a hardware or software solution (Baghaie et al., 2017). Hardware correction can be performed in an online or off-line manner, and the correction effects are promising (Ferguson et al., 2004; Vienola et al., 2012; Kocaoglu et al., 2014). However, it requires special hardware not broadly available in current clinic practice, and cannot be applied to legacy data.

Alternatively, software-based postprocessing provides an economic solution. Capps et al. (2011), Xu et al. (2009), and Ricco et al. (2009) used scanning laser ophthalmoscopy or color fundus images acquired along with the OCT scans for correcting the transverse motion, and Kraus et al. (2014), Antony et al. (2011) used additional orthogonal scans to help reconstruct true curvature of the retina. Needing extra scans as alignment reference, however, these methods added complexity to the imaging process and still had a limited applicability to a large amount of legacy data. Montuoro et al. (2014) assumed local symmetry for the shape of the retina and eliminated the need for any auxiliary scan, yet the assumption can be violated in pathological areas and in the proximity of the fovea. More recently, segmentation—e.g., of background, retinal layers, or vessels—guided motion correction was proposed (Montuoro et al., 2014; Fu et al., 2016; Lezama et al., 2016). However, the segmentations in the previous works were often coarse, independent of the motion correction, and 2D in nature (as 3D segmentation cannot be done with validity prior to proper B-scan alignment). In contrast, we couple precise 3D layer segmentation with motion correction for effective mutual performance boosting.

2.3. Hybrid 2D-3D networks

To leverage the strengths of both 2D and 3D networks for volumetric image analysis, i.e., parameter (and computation) efficiency and inter-slice correlation, respectively, hybrid 2D–3D networks were proposed (Li et al., 2018; Tran et al., 2018; Wang et al., 2019; Xie et al., 2018; Zhang et al., 2019). Besides, the hybrid architecture allowed the use of existing pretrained 2D network parameters for effective transfer learning (Li et al., 2018; Wang et al., 2020). Several studies showed that hybrid 2D–3D networks were also suitable for volume segmentation of anisotropic resolutions (Li et al., 2018; Wang et al., 2019; Zhang et al., 2019), where individual slices were first processed by 2D operations to yield features appropriate for subsequent 3D operations. However, none of them considered inter-slice motion artifacts and thus could not be directly applied to 3D segmentation of OCT volumes. In this work, besides employing a hybrid 2D–3D architecture to deal with the data anisotropy, we additionally rely on the same architecture to correct the inter-B-scan misalignment simultaneously.

2.4. Semi-supervised medical image segmentation with sparse annotation

To ease the heavy burden of manual volumetric segmentation labeling, many methods have been proposed for semi-supervised medical image segmentation with sparse annotation. Some of these methods employed non-rigid registration for label progoration (Bai et al., 2018), conducted self-training for pseudo label generation (Zheng et al., 2020), or combined both (Bitarafan et al., 2020). A central idea of these semi-supervised methods was the uncertainty estimation, based on which the pseudo labels were filtered or weighted. Several methods were also proposed for semi-supervised segmentation of retinal layers in OCT images. Sedai et al. (2019) proposed uncertainty guided semi-supervised learning based on a teacher–student approach, whereas (Liu et al., 2018a) proposed to estimate the uncertainty with a discriminator based on adversarial learning. Unlike these works relying on pseudo label generation and uncertainty estimation, our method makes use of a physical property of the segmentation target (i.e., 3D coherence of the retinal layers) and also benefits from the entangled 3D surface regression and B-scan alignment. Thanks to the unique problem formulation, it can accommodate both fully and sparsely annotated segmentation with the same framework and are competent in both cases.

3. Problem formulation

Let $\Omega \subset \mathbb{R}^3$, then a 3D OCT volume can be written as a real-valued function $V(x, y, z) : \Omega \rightarrow \mathbb{R}$, where the x and z axes are the row and column directions of a B-scan image, and y axis is orthogonal to the B-scan image (see the illustration on the far left of Fig. 1).[3] Alternatively, V can be considered as an ordered collection of all its B-scans: $V = \{I_b\}$, where $I_b : \Phi \rightarrow \mathbb{R}$ is the b^{th} B-scan image, $\Phi \subset \mathbb{R}^2$, $b \in [1, N_B]$, and N_B is the number of B-scans. Then, a retinal layer surface can be expressed by $S(b, a) = r_{b,a} : \Psi \rightarrow \mathbb{R}$, where $\Psi \subset \mathbb{R}^2$, $a \in [1, N_A]$, N_A is the number of A-scans, and $r_{b,a}$ is the row index indicating the surface location in the a^{th} A-scan of the b^{th} B-scan. That is, the surface intersects with each A-scan exactly once. Due to the image acquisition process wherein each B-scan is acquired separately without a guaranteed global alignment and the inevitable eye movement, consecutive B-scans in an

[3] Note the definition of axes is different from what is commonly used in axial CT and MRI volumes, where the z axis is orthogonal to the imaging planes.

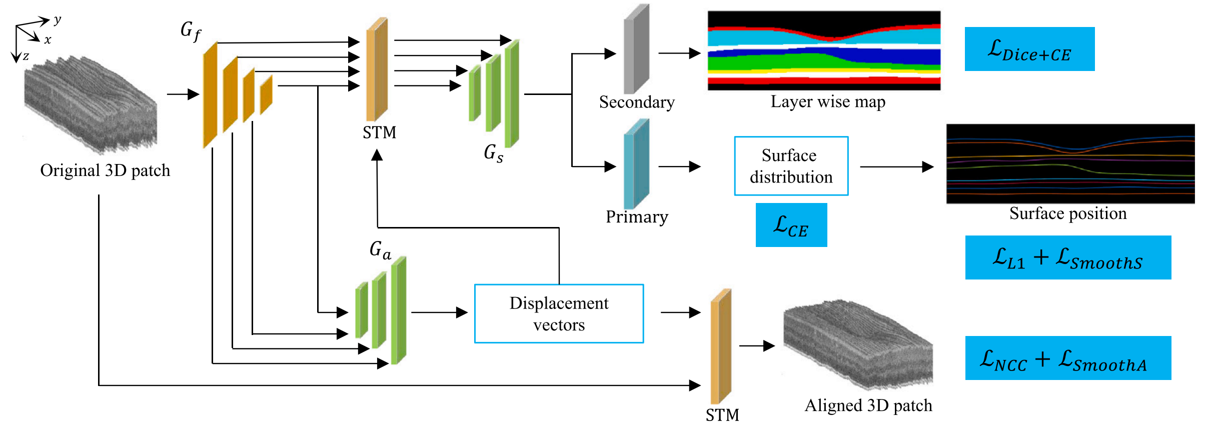


Fig. 1. Overview of the proposed framework.

OCT volume may be subject to misalignment (Cheng et al., 2016). The misalignments along the x and z axes are also known as the transverse and axial motion artifacts, respectively. Therefore, the goal of this work is to locate a set of retinal layer surfaces of interest $\{S\}$—preferably being smooth—for accurate segmentation of the layers, while at the same time re-aligning the set of B-scans $\{I_b\}$ in V. It is worth mentioning that as the transverse motion only occurs on several B-scans (meaning most of the B-scans have no x movement) (Fu et al., 2016), we only correct the axial motion within the proposed simultaneous alignment and surface regression framework, but (optionally) correct the transverse with simple post-processing.

4. Method

4.1. Overview

The overview of our framework is shown in Fig. 1. The framework comprises three major components: a contracting path G_f (the shared encoder) consisting of 2D CNN layers and two expansive paths consisting of 3D CNN layers G_a (the alignment branch) and G_s (the segmentation branch), and a functional module: the spatial transformer module (STM). First, 2D features of separate B-scans in an OCT volume are extracted by G_f. These features are then used to generate B-scan alignment displacement by G_a, which is used by the STM in turn to align the 2D features. After that, the well-aligned features are fed to G_s to yield final segmentation. Each of G_a and G_s forms a hybrid 2D–3D residual U-Net (Ronneberger et al., 2015) with G_f. The difference with the general U-Net is that our U-Net consists of both 2D and 3D CNN networks. The entire framework is trained end-to-end. As G_f is implemented as a routine 2D CNN feature extractor, below we focus on describing our novel G_a, G_s, and STM.

4.2. B-scan alignment branch

Although it is possible to add an alignment step while preprocessing, a comprehensive framework that couples the B-scan alignment and layer segmentation would mutually benefit each other (supported by our experimental results), besides being more integrated. To this end, we introduce a B-scan alignment branch consisting of an expansive path into our framework, which takes 2D features extracted from a set of B-scans by G_f and outputs a displacement vector $\triangle \boldsymbol{d} = [d_1, \ldots, d_{N_B}]$, with each element d_b indicating the displacement for a B-scan in the z direction. As smoothness is one of the intrinsic properties of the retinal layers, if the B-scans are aligned properly, ground truth surface positions of the same layer should be close at nearby locations of adjacent B-scans. To model this prior, we propose a supervised loss function to help with the alignment:

$$\mathcal{L}_{\text{SmoothA}} = \sum\nolimits_{b=1}^{N_B-1} \sum\nolimits_{a=1}^{N_A} \left((r^g_{b,a} - d_b) - (r^g_{b+1,a} - d_{b+1}) \right)^2, \tag{1}$$

where r^g is the ground truth surface location.

Meanwhile, we also use the local normalized cross-correlation (NCC) (Balakrishnan et al., 2019) of adjacent B-scans as the unsupervised optimization objective of G_a:

$$\mathcal{L}_{\text{NCC}} = \sum_{b=1}^{N_B-1} \sum_{p\in\Phi} \frac{\left[\sum_{p_k}\left(\hat{I}_b(\boldsymbol{p}_k) - \bar{I}_b(\boldsymbol{p})\right)\left(\hat{I}_{b+1}(\boldsymbol{p}_k) - \bar{I}_{b+1}(\boldsymbol{p})\right)\right]^2}{\left[\sum_{p_k}(\hat{I}_b(\boldsymbol{p}_k) - \bar{I}_b(\boldsymbol{p}))^2\right]\left[\sum_{p_k}(\hat{I}_{b+1}(\boldsymbol{p}_k) - \bar{I}_{b+1}(\boldsymbol{p}))^2\right]}, \tag{2}$$

where $\boldsymbol{p}$ iterates over all pixels in the image space Φ, $\hat{I}_b$ is the b^{th} B-scan image displaced according to the corresponding d_b (described in the following section), and $\bar{I}$ denotes images with local mean intensities subtracted out: $\bar{I}(\boldsymbol{p}) = \hat{I}(\boldsymbol{p}) - \frac{1}{n^2}\sum_{p_k}\hat{I}(\boldsymbol{p}_k)$, where $\boldsymbol{p}_k$ iterates over an $n \times n$ region around $\boldsymbol{p}$. We follow Balakrishnan et al. (2019) to set $n = 9$. The final optimization objective of the alignment branch is:

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{NCC}} + \mathcal{L}_{\text{SmoothA}}. \tag{3}$$

4.3. Spatial transformer module

Besides being used to align the input B-scan images, the displacement vector $\triangle \boldsymbol{d}$ output by the alignment branch G_a is also used to align the 2D features extracted by G_f, such that subsequent 3D operations of the segmentation branch G_s are valid. To do so, we propose to add a spatial transformer module (STM) (Balakrishnan et al., 2019) to the shortcuts between G_f and G_s. It is worth noting that the STM adaptively rescales $\triangle\ \boldsymbol{d}$ to suit the size of the features at different scales. Without loss of generality, we use the input B-scan images for explanation. Specifically, for each pixel $\boldsymbol{p} = (p_x, p_z)$ in the relocated B-scan image $\hat{I}_b$, we compute a (sub-)pixel location $\boldsymbol{p}' = \boldsymbol{p} + (0, d_b)$ (recall that we only consider axial motion in the networks) in the original image I_b. Then, we linearly interpolate the values at neighboring pixels of $\boldsymbol{p}'$ as the value for $\hat{I}_b(\boldsymbol{p})$:

$$\hat{I}_b(\boldsymbol{p}) = \sum\nolimits_{q\in\mathcal{Z}(p')} I_b(\boldsymbol{q})\left(1 - |p'_x - q_x|\right)\left(1 - |p'_z - q_z|\right), \tag{4}$$

where $\mathcal{Z}(\boldsymbol{p}')$ are the pixel neighbors of $\boldsymbol{p}'$. The STM allows back progation during optimization (Balakrishnan et al., 2019). The application of $\triangle\ \boldsymbol{d}$ to the 2D features is mostly the same, except for that $\triangle\ \boldsymbol{d}$ is rescaled to suit the downsampling factors of the features at different scales. In this way, we couple the B-scan alignment and retinal layer segmentation in our framework for an integrative end-to-end training, which not only simplifies the entire pipeline but also boosts the segmentation performance as validated by our experiments.

4.4. Layer segmentation branch

Our layer segmentation branch substantially extends the fully convolutional boundary regression (FCBR) framework proposed by He et al. (2019a). Above all, we replace the purely 2D FCBR framework by a hybrid 2D–3D framework, to perform 3D surface regression in an OCT volume instead of independent 2D boundary regression in individual B-scans. On top of that, we propose a global smoothness guarantee loss to encourage coherent surfaces both within and across-B-scan, whereas FCBR only enforces intra-B-scan smoothness. Third, our segmentation branch is coupled with the B-scan alignment branch, which boosts the performance of each other.

The segmentation branch has two heads sharing the same decoder: the primary head outputs the surface position distribution for each A-scan, and the secondary head outputs pixel-wise semantic labels. The secondary head is used only to provide an additional task for training the network, especially considering its pixel-wise dense supervision. Eventually the output of the secondary head is ignored during testing. We follow He et al. to use a combined Dice and cross entropy loss (Roy et al., 2017) $\mathcal{L}_{\text{Dice+CE}}$ for training the secondary head, and refer interested reader to He et al. (2019a) for more details.

Surface distribution head. This primary head generates an independent surface position distribution $q_{b,a}(r|V;\theta)$ for each A-scan, where θ denotes the network parameters, and a higher value indicates a higher possibility that the surface is on the r^{th} row. Like in He et al. (2019a), a cross entropy loss is used to train the primary head:

$$\mathcal{L}_{\text{CE}} = -\sum_{b=1}^{N_B}\sum_{a=1}^{N_A}\sum_{r=1}^{R}\mathbb{1}(r^g_{b,a}=r)\log q_{b,a}(r^g_{b,a}|V,\theta), \tag{5}$$

where R is the number of rows of an A scan, and $\mathbb{1}(x)$ is the indicator function where $\mathbb{1}(x)=1$ if x is evaluated to be true and zero otherwise. Further, a smooth L1 loss is adopted to directly guide the predicted surface location $\hat{r}$ to be the ground truth:

$$\mathcal{L}_{\text{L1}} = \sum_{b=1}^{N_B}\sum_{a=1}^{N_A} 0.5t^2_{b,a}\mathbb{1}(|t_{b,a}|<1) + (|t_{b,a}|-0.5)\mathbb{1}(|t_{b,a}|\geq 1), \tag{6}$$

where $t_{b,a} = \hat{r}_{b,a} - r^g_{b,a}$, and $\hat{r}_{b,a}$ is obtained via the soft-argmax $\hat{r}_{b,a} = \sum_{r=1}^{R} r q_{b,a}(r|V,\theta)$.

Global coherence loss. Previous studies have demonstrated the effectiveness of modeling prior knowledge that reflects anatomical properties such as the structural smoothness (Wei et al., 2018) in medical image segmentation. Following this line, we also employ a global smoothness loss to encourage the detected retinal surface $\hat{S}$ to be coherent both within and across-B-scan based on its gradients:

$$\mathcal{L}_{\text{SmoothS}} = \sum_{b=1}^{N_B}\sum_{a=1}^{N_A}\left\|\nabla\hat{S}(b,a)\right\|^2. \tag{7}$$

Finally, the overall optimization objective of the segmentation branch is

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{Dice+CE}} + \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{L1}} + \lambda\mathcal{L}_{\text{SmoothS}}, \tag{8}$$

where λ is a hyperparameter controlling the influence of the global coherence loss.

4.5. Semi-supervised learning with sparse annotations

Next, we extend our framework for semi-supervised learning with sparse annotations. To reduce the human efforts for labeling whole OCT volumes, we propose to further leverage the smoothness property of the retinal layers, so that the segmentation model can be effectively trained with only a fraction of the B-scans annotated in each given OCT volume. Thanks to our unique problem formulation of coupled 3D surface regression and B-scan alignment, the adaptation for the semi-supervised setting is straightforward with only minor alterations to the loss functions. For the B-scan alignment branch, we adapt the supervised loss $\mathcal{L}_{\text{SmoothA}}$ (Eq. (1)) by using the surface locations predicted by G_s for unannotated B-scans:

$$\mathcal{L}^{\text{Semi}}_{\text{SmoothA}} = \sum_{b=1}^{N_B}\sum_{a=1}^{N_A}\left((r_{b,a}-d_b)-(r_{b+1,a}-d_{b+1})\right)^2, \tag{9}$$

where $r_{b,a} = \hat{r}_{b,a}$ for unannotated B-scans and $r^g_{b,a}$ otherwise, while the unsupervised alignment loss $\mathcal{L}_{\text{NCC}}$ remains unchanged. To this end, $\mathcal{L}^{\text{Semi}}_{\text{SmoothA}}$ couples G_s and G_a in a more delicate way, where the segmentation and alignment results interactively influence each other. As to the layer segmentation branch, we now calculate the supervised losses $\mathcal{L}_{\text{Dice+CE}}$, $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{CE}}$ only on the annotated B-scans, with the regulating global coherence loss $\mathcal{L}_{\text{SmoothS}}$ intact.

It is worth mentioning that as $\mathcal{L}^{\text{Semi}}_{\text{SmoothA}}$ (Eq. (9)) relies on the model-predicted surface location $\hat{r}_{b,a}$, we notice that bad quality of $\hat{r}_{b,a}$ would impede the training or even cause a collapse. Therefore, we first warm up the model without $\mathcal{L}^{\text{Semi}}_{\text{SmoothA}}$ for five epochs, such that the predicted $\hat{r}_{b,a}$ is reasonable when adding $\mathcal{L}^{\text{Semi}}_{\text{SmoothA}}$ back afterwards.

4.6. Transverse alignment by post-processing

The transverse motions are much less frequent than the axial for the OCT B-scans (Fu et al., 2016). Notwithstanding, for completeness, we still propose a simple yet effective post-processing transverse alignment for optional use when necessary. Specifically, we first average the A-scans of each B-scan image to turn the latter into a strip of mean intensity projections. We then align adjacent B-scans by shifting them in the x direction to minimize the mean squared error between their projections. Notably, as we already have the retinal layers segmented, we eliminate the interference of the background by computing the mean projection only in the retinal layers. The entire OCT volume is aligned by repeating the pairwise alignment.

5. Experiments

5.1. Datasets and preprocessing

The proposed framework is validated on three publicly available SD-OCT datasets. In addition, synthetic images are utilized for quantitative evaluation of the inter-B-scan motion correction. Below we describe the public datasets first, and defer the description of the synthetic data to the corresponding experiments section.

A2A SD-OCT study dataset. The Age-Related Eye Disease Study 2 (AREDS2) Ancillary SD-OCT (A2 A SD-OCT) Study dataset (Farsiu et al., 2014) includes both normal (115) and age-related macular degeneration (AMD) (269) cases. The images were acquired using the Bioptigen Tabletop SD-OCT system (Bioptigen, Inc., Research Triangle Park, NC). The physical resolutions are 3.24 μm (within A-scan), 6.7 μm (cross-A-scan), and 67 μm (cross-B-scan). Since the manual annotations are only available for a region centered at the fovea, subvolumes of size 400 × 41 × 512 (N_A, N_B, and R) voxels are extracted around the fovea. We train the model on 263 subjects and test on the other 72 subjects (49 cases are eliminated from analysis as the competing alignment algorithm (Pnevmatikakis and Giovannucci, 2017) fails to handle them), which are randomly split with the proportion of AMD cases unchanged. The inner aspect of the inner limiting membrane (ILM), inner aspect of the retinal pigment epithelium drusen complex (IRPE), and outer aspect of Bruch's membrane (OBM) were manually traced.

JHH SD-OCT dataset. The Johns Hopkins Hospital (JHH) dataset (He et al., 2019b) contains 14 healthy controls (HC) and 21 cases with multiple sclerosis (MS) which exhibit mild thinning of retinal layers. The data was acquired with a Spectralis OCT system (Heidelberg Engineering, Heidelberg, Germany), which has its own motion correction, registration, and averaging algorithms during image acquisition. Each of the 35 cases includes 49 B-scans of 1024 × 496 (N_A×R) pixels in size. The physical resolutions are 3.87 μm (within A-scan), 5.8 μm (cross-A-scan), 123.6 μm (cross-B-scan). Following the train/test split in He et al. (2019a), we use the last six HCs and last nine MS cases for training and the other 20 subjects for testing. Nine surfaces were manually delineated in each B-scan, separating the following retinal layers: the retinal nerve fiber layer (RNFL); the ganglion cell layer (GCL) combined with the inner plexiform layer (IPL), denoted as GCIP; the inner nuclear layer (INL); the outer plexiform layer (OPL); the outer nuclear layer (ONL); the inner segment (IS); the outer segment (OS); and the retinal pigment epithelium (RPE). Surfaces between these layers are denoted by hyphenating their acronyms. Three other named surfaces are: the inner limiting membrane (ILM); the external limiting membrane (ELM); and Bruch's membrane (BM).

Duke DME dataset. The Duke Eye Center dataset (Chiu et al., 2015) contains 10 diabetic macular edema (DME) patients (one OCT volume per patient), each with 61 B-scans of size 768 × 496 ($N_A \times R$) pixels. The first five patients were rated as having severe macular edema with damaged retinal structures. The data was acquired with a standard Spectralis (Heidelberg Engineering, Heidelberg, Germany) 61-line volume scan protocol. The physical resolutions are 3.87 μm (within A-scan), 10.94–11.98 μm (cross-A-scan), and 118–128 μm (cross-B-scan). Eight retinal surfaces (the same ones as those delineated on the JHH dataset except for ELM) were manually delineated for 11 B-scans per patient. We follow the 50%:50% train/test split in previous works (Chiu et al., 2015; He et al., 2021; Karri et al., 2016; Rathke et al., 2017) to use the last five patients for training and the challenging first five patients for testing. As our framework takes 3D volumes as input, we use the partially annotated OCT volumes (i.e., 11 of 61 B-scans annotated per volume) in this dataset as semi-supervised learning with sparse annotations (cf. Section 4.5).

Preprocessing. An intensity gradient method (Lang et al., 2013) is employed to flatten the B-Scan images to the estimated Bruch's membrane. After that, B-scan images in the A2 A, JHH, and DME datasets are cropped to 400 × 320, 1024 × 128 and 768 × 224 pixels, respectively, to exclude background while ensuring inclusion of retinal tissue (He et al., 2021). The preprocessing effectively reduces memory consumption for model training.

5.2. Evaluation metrics

For B-scan alignment, we adopt the mean absolute distance (MAD) of the same surface and the NCC between two adjacent B-scans for quantitative evaluation on the A2 A dataset. In addition, on the synthetic dataset we directly calculate the mean absolute difference between the estimated and ground truth motions (note that the motions are simulated in this setting thus the ground truth is available) for evaluation. For retinal layer segmentation, the MAD and 95th percentile of the Hausdorff distance (HD95) between predicted and ground truth surface positions are used. To quantify the cross-B-scan continuity of the segmented surfaces, inspired by He et al. (2021), we compute the surface distances between adjacent B-Scans as the statistics of smoothness and plot the histogram for visual analysis. We compute the mean metrics and standard deviations (std.) per volume, per the volumetric nature of our proposed framework. Especially for MAD, all A-scan-wise differences of an OCT volume are first averaged to yield a volume-wise metric. Then, a mean metric and std. are estimated from the volume-wise metrics of all test volumes. Note that for the DME dataset, we follow He et al. (2019a, 2021) to ignore the positions where Chiu et al. (2015)'s result or the manual delineation are missing for evaluation.

5.3. Implementation

The PyTorch framework (1.4.0) is used for all experiments. For the network design, we mainly follow the architecture proposed in Model Genesis (Zhou et al., 2019) with necessary adaptations: (i) the feature extractor G_f is constructed by replacing 3D operations of the Model Genesis's encoder with 2D counterparts, (ii) the alignment and segmentation branches G_a and G_s are mostly the same as the decoder in Model Genesis, except that we keep the dimension corresponding to the number of B-scans unchanged throughout, and (iii) we halve the number of channels in each CNN block to reduce the number of network parameters. All networks are trained from scratch. Due to the GPU memory constraint, the strategy of patch-wise training is employed: the OCT volumes are cut by planes perpendicular to the B-scan planes into subvolumes, which are input to the networks. The sizes of the subvolumes (N_A, N_B, and R) are 48 × 41 × 320, 48 × 49 × 128 and 48 × 41 × 224 voxels for the A2 A, JHH and DME datasets, respectively. For the A2 A dataset, we train the networks on three 2080 Ti GPUs with a mini-batch size of 9. As to the JHH and DME datasets, we train the networks on one 2080 Ti GPU with the mini-batch sizes of 6 and 4, respectively. The networks are trained for 80, 100 and 100 epochs for the A2 A, JHH and DME datasets, respectively, with the Adam optimizer (Kingma and Ba, 2014). The learning rate is set to 0.001 for the A2 A dataset; for the relatively smaller JHH and DME datasets, the learning rate is initialized to 0.003 and adjusted by a cosine annealing scheduler with the half period and minimum value set to 40 epochs and 3×10^{-7}, respectively.

For multi-layer segmentation, there are two practical considerations. First, the set of surface locations $\{\hat{r}_{b,a,l}\}_{l=1}^{L}$ (where L is the total number of surfaces) predicted for an A-scan are not guaranteed to follow the strict anatomical order. Therefore, we implement the iterative surface swap trick (He et al., 2019a), where the locations of two predicted neighboring surfaces are swapped if they do not obey the correct anatomical order. Second, the extents of smoothness of different surfaces vary naturally, and the weight λ for the global coherence loss $\mathcal{L}_{\text{SmoothS}}$ in Eq. (8) should also vary accordingly to accommodate the natural variation. Empirically, for the lth surface, we compute $\lambda_l = \lambda_b / \left(\sum_{b=1}^{N_B} \sum_{a=1}^{N_A} \left\| \nabla S_l^g(b,a) \right\| \right)$, where λ_b is the base weight for a specific dataset and set to 0.1 for the A2 A and JHH datasets and 0.03 for the DME dataset, and S_l^g is the ground truth surface.[4] Intuitively, the smoother the surface naturally is, the more we penalize its global coherence loss. In practice, we use the arithmetic mean of λ_l's of all the training OCT volumes for the l^{th} surface.

For reproducible research, our implementation and trained models are available at: https://github.com/ccarliu/Retinal-OCT-LayerSeg/tree/following-work.

5.4. Motion correction results

We compare our proposed approach with Montuoro et al. (2014), NoRMCorre (Pnevmatikakis and Giovannucci, 2017) and Fu et al. (2016) on both real clinical (the A2 A) and synthetic OCT data. The method proposed by Montuoro et al. (2014) is based on the hypothesis that a motion-free SD-OCT volume of a healthy person is predominantly locally symmetric along the axial scan direction (z axis), and sequentially corrects the motions in the z and x directions. NoRMCorre is a template matching based algorithm originally proposed for fast and robust motion correction of calcium imaging data, a similar scenario to the B-scan alignment. Since the B-scans are mostly motion artifact free internally (i.e., no misalignment among the A-scans within a B-scan), we only need to operate NoRMCorre in a rigid fashion. In addition, given the fast cross-B-scan change in image content due to the low y axial resolution, each B-scan image directly uses its immediate predecessor as the template to match, rather than the median of the buffer as in Pnevmatikakis and Giovannucci (2017). The method proposed by Fu et al. (2016) is based on the retinal layer saliency map and center bias constraint, to alleviate performance degradation caused by background noise and strong vessels, respectively.

5.4.1. Results on synthetic data

For quantitative validation against ground truth motions, we create a synthetic dataset of 20 SD-OCT volumes from the A2 A clinical

[4] Note that a quick motion correction with NoRMCorre (Pnevmatikakis and Giovannucci, 2017) has to be applied to the A2 A dataset beforehand to ensure a valid estimate of λ_l.

Table 1
Mean absolute differences (in pixels) between recovered and ground truth motion vectors on the synthetic SD-OCT dataset. Results of NoRMCorre (Pnevmatikakis and Giovannucci, 2017), Montuoro et al. (2014) and Fu et al. (2016) are presented for comparison. Format: mean (std.).

Motion	Montuoro et al.	NoRMCorre	Fu et al.	No_layer	Ours
Axial	7.06 (0.25)	2.11 (0.53)	2.88 (1.32)	**1.76** (0.60)	**1.76** (0.60)
Transverse	8.92 (6.26)	8.64 (5.99)	5.91 (2.61)	4.91 (2.26)	**4.80** (1.64)

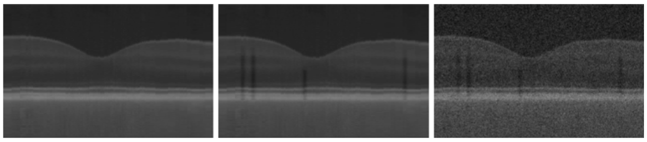

(a) (b) (c)

Fig. 2. A synthetic B-scan image of an OCT volume created adopting the procedures described in Montuoro et al. (2014). (a) First synthesized without vessel shadow or noise. (b) Vessel shadows overlaid. (c) Noise added.

dataset adopting the procedures described in Montuoro et al. (2014) (Fig. 2(a)–(c)).[5] Then, we simulate motion artifacts on the synthesized volumes as follows: (i) for the axial motion in the z axis, each B-scan image is moved by a random displacement uniformly sampled from $[-15, 15]$ pixels; and (ii) for the transverse motion in the x axis, we group all the B-scans into three to five consecutive groups and apply a random displacement uniformly sampled from $[-15, 15]$ pixels to each of the groups as a whole, to simulate the micro-saccades of the eye (Fu et al., 2016). We repeat the above procedures five times on each of the synthesized volumes, resulting a total of 100 misaligned synthetic volumes.

We then apply various motion correction methods to the purposely misaligned synthetic OCT volumes, to recover the applied motion vectors. The results are shown in Table 1. We can see that NoRMCorre substantially outperforms (Montuoro et al., 2014) in axial motion correction by ∼70%, but remains at the same level for transverse motion correction (8.92 versus 8.64 pixels). Compared with NoRMCorre, (Fu et al., 2016) substantially improves the transverse motion by ∼32% (5.91 pixels), while slightly increasing the axial residual error. In contrast, our method not only further reduces the axial residual errors by apparent advantages (2.11 versus 1.76 pixels), but also substantially reduces the transverse residual errors from 5.91 to 4.80 pixels. We conjecture this is because our method not only relies on the grayscale image information for alignment but additionally fully utilizes the layer segmentation—produced by the coupled segmentation branch. We implement a variant of our method (No_layer) where the layer segmentation is not utilized to exclude background for transverse alignment. As expected, the transverse correction performance drops notably, again emphasizing the value of making use of the segmentation. In addition, as we conduct our transverse motion correction as post-processing, it is likely that our method also benefits from its superior performance on axial motion correction for transverse motion correction. In conclusion, our proposed method yields the best motion correction performance on the synthetic data.

5.4.2. *Results on the A2A dataset*

In addition to the synthetic data, we also evaluate the motion correction performance on the A2 A real clinical data. As the underlying ground truth motions are unknown for the clinical dataset, we instead compute the MAD between the locations of the same surface in adjacent B-scans (note the ground truth surface locations are used here), and the NCC between adjacent B-scans, for quantitative evaluation. The lower the MAD and the higher the NCC, the better the two B-scans are matched. The results are shown in Table 2. As we can see, all the evaluated motion correction methods improve both metrics upon the baseline before correction. While the improvements by Montuoro et al. (2014) are minor, those by NoRMCorre (Pnevmatikakis and Giovannucci, 2017) are more substantial. Compared to NoRMCorre, (Fu et al., 2016) is better in NCC but worse in MADs. Our proposed method further improves upon both NoRMCorre and Fu et al. (2016), achieving the lowest MADs for all three evaluated surfaces and their average and comparable NCC to Fu et al. (2016). Fig. 3 visualizes motion correction results of an A2 A OCT volume by these methods. We can observe obvious mis-alignment between the B-scans before correction. While Montuoro et al. (2014) hardly realigns the B-scans properly, NoRMCorre, Fu et al. (2016) and our method make the B-scans more aligned, and our results are visually better as highlighted by the red arrows.

5.5. *Layer segmentation results*

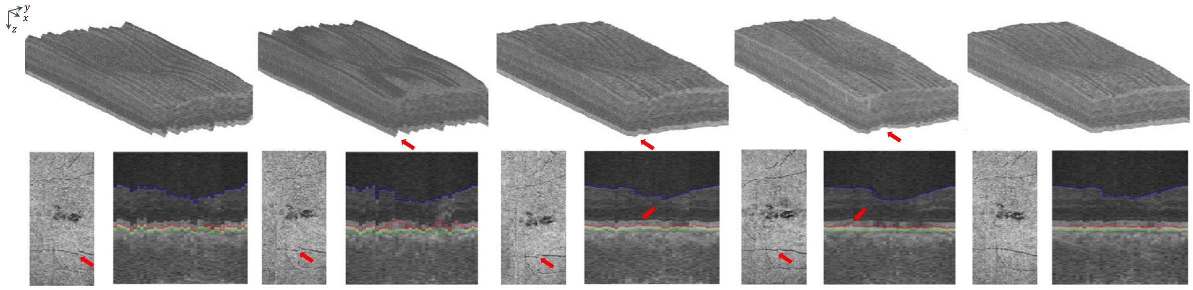
5.5.1. *Comparison with state-of-the-art (SOTA) methods*

We compare our proposed method with several up-to-date baselines: ReLayNet, MGU-Net, FCBR, IPM, and DDP. ReLayNet (Roy et al., 2017) is a U-Net based method which outputs the layer maps. MGU-Net (Li et al., 2021) employs graph convolutional networks to simultaneously label the retinal layers and optic disc. FCBR (He et al., 2019a, 2021) is a SOTA method implementing 2D surface regression, thus can directly output surface locations like ours. IPM and DDP further employ the primal–dual interior-point method (IPM) and differentiable dynamic programming (DDP) to explicitly enforce surface interaction and smoothness constraints, respectively. For the methods that only output layer maps (ReLayNet and MGU-Net), we obtain the surface locations by summing up the output layer maps in each A-scan as done by He et al. (2021). We use the official implementation of MGU-Net and RelayNet, and implement and empirically optimize FCBR, IPM, and DDP, to get their results.

The MADs between the predicted surface locations and manual delineation on the A2 A and JHH SD-OCT datasets are charted in Table 3. On the A2 A dataset, our method achieves a significantly lower overall MAD with a smaller standard deviation (2.68 ± 1.39 μm) than the previous best performing methods: FCBR (2.74 ± 1.87 μm), IPM (2.75 ± 1.66 μm), and DDP (2.73 ± 1.48 μm), all with $p < 0.05$. Meanwhile, our method, FCBR, IPM, and DDP are substantially better than the other two compared methods ReLayNet (7.64 ± 13.68 μm) and MGU-Net (3.45 ± 4.00 μm) by large margins. On the JHH dataset, we note that the performance variations between the methods largely decrease, probably because this dataset does not present severe pathologies and is better in image quality, thus is easier to segment than the A2 A dataset. Our method achieves the lowest overall MAD of 2.77 μm, marginally lower than that of DDP (2.79 μm), IPM (2.81 μm) and FCBR (2.82 μm) with no statistical significance. Notwithstanding, our overall MAD is still significantly better than that of ReLayNet (3.23 μm) and MGU-Net (3.01 μm), both with $p < 0.001$.

The mean HD95 values between the predicted surfaces and manual delineation on the A2 A and JHH datasets are presented in Table 4. The general trends are the same as the MADs in Table 3. On the A2 A dataset, our method is significantly better than all others in term of the overall mean HD95, although the performance gaps between the methods become more obvious using HD95 as the evaluation metric. On the JHH dataset, the performance of our method (overall mean HD95:

[5] Although the JHH dataset was motion corrected by its provider, we prefer synthetic data here considering potential residual motions of the JHH data.



(a) No correction (b) Montuoro et al. (2014) (c) NoRMCorre (d) Fu et al. (2016) (e) Ours

Fig. 3. Visualization of motion correction results of an OCT volume in the A2 A dataset, including: 3D visualization, OCT fundus image obtained by averaging intensity values between the OBM and IRPE surfaces, and yz-plane cross-section image (with ground truth layer surfaces overlaid: blue: ILM, red: IRPE, and green: OBM). NoRMCorre was proposed by Pnevmatikakis and Giovannucci (2017). Red arrows highlight places where our results are visually better.

Table 2
Motion correction results on the A2A clinical SD-OCT dataset, evaluated with the mean absolute distance (MAD; in pixel) of the same surface, and the normalized cross-correlation (NCC) between adjacent B-scans. Results of NoRMCorre (Pnevmatikakis and Giovannucci, 2017), Montuoro et al. (2014) and Fu et al. (2016) are included for comparison. Format: mean (std.).

Metric	No correction	Montuoro et al.	NoRMCorre	Fu et al.	Ours
MAD					
ILM	3.92 (1.57)	3.42 (1.53)	1.74 (0.52)	2.94 (2.15)	**1.58** (0.49)
IRPE	4.17 (1.64)	3.68 (1.69)	2.19 (0.93)	3.31 (2.22)	**2.12** (0.89)
OBM	3.93 (1.59)	3.43 (1.55)	1.87 (0.65)	3.06 (2.09)	**1.81** (0.64)
Average	4.00 (1.59)	3.51 (1.57)	1.93 (0.62)	3.10 (1.58)	**1.83** (0.59)
NCC	0.0456 (0.0049)	0.0455 (0.0050)	0.0470 (0.0067)	0.0483 (0.0068)	**0.0481** (0.0062)

Table 3
Layer segmentation results evaluated by the mean absolute distance (μm) between the predicted and ground truth surface locations (std. in parentheses). Results of ReLayNet (Roy et al., 2017), MGU-Net (Li et al., 2021), FCBR (He et al., 2019a, 2021), IPM (Xie et al., 2022a), and DDP (Xie et al., 2022b) are included for comparison. The asterisks denote statistically significant differences from our proposed method with the Wilcoxon signed-rank test (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

A2A dataset							
Method	ILM (AMD)	ILM (Normal)	IRPE (AMD)	IRPE (Normal)	OBM (AMD)	OBM (Normal)	Overall
ReLayNet	5.23 (10.65)***	2.44 (4.08)***	12.45 (24.95)***	3.80 (3.59)***	11.55 (19.76)***	3.23 (1.83)***	7.64 (13.68)***
MGU-Net	2.48 (4.71)***	1.49 (0.45)***	4.51 (7.50)***	2.39 (1.38)***	5.22 (4.19)***	2.47 (0.34)*	3.45 (4.00)***
FCBR	1.83 (2.74)***	**1.22** (0.46)***	3.09 (2.29)	2.15 (1.38)	4.51 (3.28)	**2.28** (0.34)***	2.74 (1.87)*
IPM	**1.80** (1.83)	1.28 (0.41)*	3.15 (1.87)***	2.18 (1.26)*	4.46 (2.46)	2.31 (0.34)***	2.75 (1.66)*
DDP	1.81 (2.38)	1.23 (0.46)*	3.11 (2.09)*	2.12 (1.25)	4.46 (3.63)	2.31 (0.38)***	2.73 (1.48)*
Proposed	**1.80** (1.97)	1.30 (0.52)	**2.91** (1.61)	**2.10** (1.35)	**4.34** (2.55)	2.40 (0.38)	**2.68** (1.39)

JHH dataset									
Method	ILM	RNFL-GCL	IPL-INL	INL-OPL	OPL-ONL	ELM	IS-OS	OS-RPE	BM	Overall
ReLayNet	2.67 (0.46)***	3.58 (1.11)***	3.41 (0.81)***	3.42 (0.67)***	3.11 (0.82)*	2.96 (0.74)**	2.43 (0.95)**	4.02 (1.15)**	3.45 (2.05)**	3.23 (0.71)***
MGU-Net	2.58 (0.26)***	3.16 (0.64)***	3.02 (0.38)***	3.29 (0.53)	2.90 (0.58)***	2.82 (0.69)*	2.15 (0.59)	3.60 (0.62)	3.59 (2.34)**	3.01 (0.42)***
FCBR	2.39 (0.30)*	3.01 (0.70)***	2.96 (0.38)**	3.24 (0.51)	2.86 (0.55)***	2.71 (0.86)	1.98 (0.75)	3.52 (0.94)	**2.74** (1.67)*	2.82 (0.40)
IPM	2.32 (0.48)*	2.94 (0.68)***	2.92 (0.38)**	**3.15** (0.36)	2.77 (0.57)**	2.73 (0.87)	2.01 (0.76)	3.46 (0.92)	3.04 (2.13)	2.81 (0.48)
DDP	2.32 (0.26)*	3.10 (0.65)***	2.94 (0.36)**	3.17 (0.49)	2.74 (0.55)**	**2.61** (0.63)	**1.94** (0.66)	**3.31** (0.80)*	2.95 (1.97)	2.79 (0.41)
Proposed	**2.21** (0.35)	**2.73** (0.61)	**2.79** (0.42)	3.18 (0.33)	**2.62** (0.58)	2.65 (0.52)	2.04 (0.73)	3.56 (1.04)	3.19 (2.02)	**2.77** (0.51)

6.75 μm) is comparable to that of FCBR and DDP (6.78 and 6.73 μm, respectively, no statistical significance), slightly better than that of IPM (6.84 μm, no statistical significance), while significantly better than that of ReLayNet and MGU-Net (8.41 and 7.23 μm, respectively, $p < 0.001$).

Figs. 4 and 6 show example segmentation by our framework and FCBR on the JHH and A2 A datasets, respectively. In most cases the segmentation by both methods looks comparable, yet in difficult situations (pointed by red arrows) our framework produces layer boundaries closer to the manual segmentation.

5.5.2. B-scan connectivity analysis

A hypothesized advantage of our proposed 3D OCT layer segmentation over 2D counter-methods such as FCBR (He et al., 2019a, 2021) is the cross-B-scan surface smoothness, i.e., 3D continuity beyond the 2D B-scan image planes. To test the hypothesis, we compute the surface distance between adjacent B-scans by $\left|r_{b+1,a} - r_{b,a}\right|$ to quantify the cross-B-scan (dis)continuity. The surface distance histograms for the A2 A and JHH datasets are shown in Fig. 5. On the A2 A dataset, the surfaces segmented by our framework have better cross-B-scan connectivity than those by FCBR, as indicated by the more conspicuous spikes clustered around 0 of our framework. After B-scan pre-alignment by NoRMCorre (Pnevmatikakis and Giovannucci, 2017), the connectivity of FCBR improves, yet is still inferior to that of our framework. For intuitive perception, we visualize the ILM layer segmented by FCBR (with NoRMCorre pre-alignment) and our framework in Fig. 7. It can be observed that our segmentation is visually smoother. Similarly, IPM (Xie et al., 2022a) and DDP (Xie et al., 2022b) with pre-alignment still lag behind our framework on connectivity. This suggests that merely conducting 3D alignment does not guarantee 3D continuity of the segmentation results, as long as the B-scans are handled separately. It is worth noting that our method also achieves better cross-B-scan connectivity than the ground truth after alignment, likely due to the same reason (i.e., human annotators work with one B-scan at a time). On the JHH dataset, the surfaces segmented by our framework again show the best cross-B-scan connectivity, as expected (note that the JHH data are already motion-corrected by the provider, thus we do

Table 4
Layer segmentation results evaluated by the mean 95th percentile of the Hausdorff distance (μm) between the predicted and ground truth surface locations (std. in parentheses). Results of ReLayNet (Roy et al., 2017), MGU-Net (Li et al., 2021), FCBR (He et al., 2019a), IPM (Xie et al., 2022a), and DDP (Xie et al., 2022b) are included for comparison. The asterisks denote statistically significant differences from our proposed method with the Wilcoxon signed-rank test (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

A2A dataset

Method	ILM (AMD)	ILM (Normal)	IRPE (AMD)	IRPE (Normal)	OBM (AMD)	OBM (Normal)	Overall
ReLayNet	19.31 (39.37)***	8.29 (19.52)***	42.90 (61.92)***	13.97 (18.57)***	34.98 (45.63)***	10.07 (10.37)***	25.49 (37.79)***
MGU-Net	6.96 (12.59)***	3.98(2.64)***	15.97 (26.35)*	5.86 (2.81)**	15.32 (14.30)*	5.95 (0.89)*	10.36 (13.22)**
FCBR	4.32 (7.71)***	**2.68** (1.34)***	8.55 (8.29)	4.75 (2.91)	11.43 (9.56)	**5.00** (0.93)***	6.82 (5.62)**
IPM	**4.12** (4.90)**	3.02 (1.43)***	9.05 (7.51)***	4.87 (2.54)*	11.20 (7.93)	5.04 (0.87)***	6.91 (4.82)***
DDP	4.45 (6.22)***	2.95 (1.36)***	8.92 (7.54)***	4.97 (2.64)***	12.83 (15.87)	5.09 (0.89)**	7.33 (7.02)*
Proposed	4.37 (5.88)	2.89 (1.54)	**7.92** (5.80)	**4.68** (2.86)	**10.72** (7.56)	5.28 (1.00)	**6.58** (4.21)

JHH dataset

Method	ILM	RNFL-GCL	IPL-INL	INL-OPL	OPL-ONL	ELM	IS-OS	OS-RPE	BM	Overall
ReLayNet	6.24 (0.95)***	11.02 (4.79)***	9.34 (3.14)***	9.15 (3.05)*	9.18 (3.62)***	7.54 (2.92)**	5.46 (1.53)***	9.65(3.83)**	8.08 (3.14)***	8.41 (2.44)***
MGU-Net	6.03 (0.64)***	8.96 (2.44)***	7.86 (1.13)**	7.98 (1.22)	7.96 (1.94)***	6.45 (1.42)**	4.90 (1.02)*	7.98 (1.18)	6.90 (3.16)**	7.23 (0.98)***
FCBR	5.52 (0.66)	8.38 (2.26)***	7.68 (1.06)**	7.90 (1.18)	7.70 (1.81)	6.09 (1.59)**	4.50 (1.48)	7.61 (1.64)	**5.61** (2.55)**	6.78 (0.97)
IPM	5.49 (0.67)	8.34 (2.01)***	7.66 (0.80)**	7.95 (0.98)	7.71 (1.70)*	6.15 (1.19)	4.59 (1.50)	7.69 (1.71)	6.07 (3.00)	6.84 (0.95)
DDP	5.43 (0.74)	8.66 (2.17)***	7.47 (0.90)	**7.80** (1.08)	7.63 (1.84)*	**5.94** (1.31)	**4.35** (1.12)	**7.37** (1.54)*	5.96 (2.87)	**6.73** (0.92)
Proposed	**5.32** (0.74)	**7.64** (1.94)	**7.30** (0.93)	7.86 (0.84)	**7.39** (1.78)	6.00 (1.17)	4.61 (1.25)	8.18 (2.23)	6.45 (2.98)	6.75 (1.11)

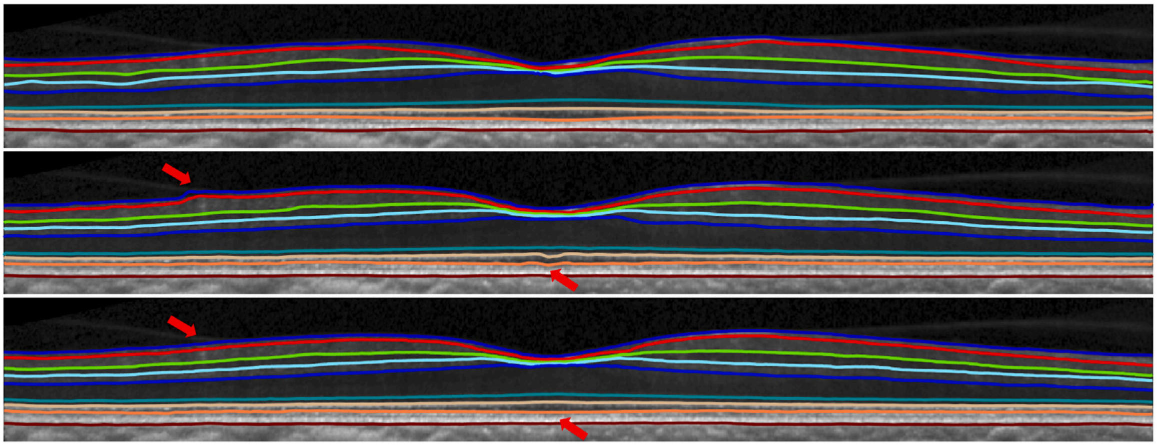


Fig. 4. Visualization of the manual segmentation (top), and segmentations by FCBR (He et al., 2019a) (middle) and our framework (bottom) of a B-scan (an MS case) in the JHH dataset. Layer boundaries from top to bottom: ILM, RNFL-GCL, IPL-INL, INL-OPL, OPL-ONL, ELM, IS-OS, OS-RPE, and BM surfaces. The red arrows indicate where our segmentation is better. Note that our framework correctly segments the foveal pit region.

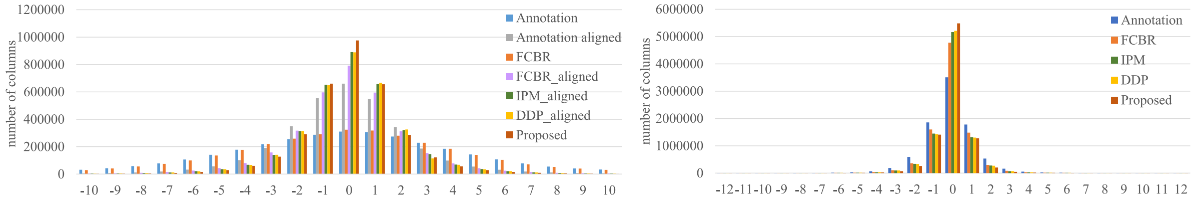


Fig. 5. Histograms of the surface distances (x axis; in pixels) between adjacent B-Scans. Left: A2 A dataset, and right: JHH dataset.

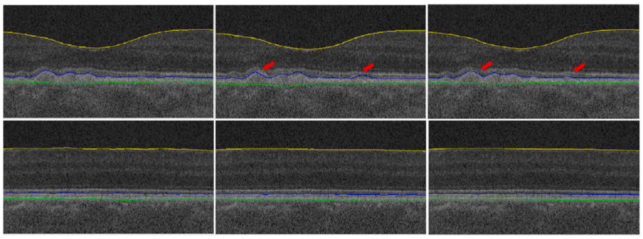


Fig. 6. Example manual segmentation (left), and segmentation by FCBR (He et al., 2019a) (middle) and our framework (right) on the A2 A dataset. Top: an AMD case; bottom: a normal control. The yellow, blue, and green curves indicate the ILM, IRPE, and OBM boundaries, respectively. The red arrows indicate where our segmentation is better. Note that our framework correctly segments the AMD case in the presence of drusen.

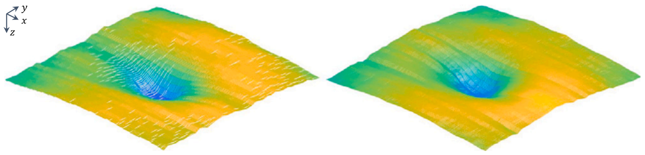


Fig. 7. 3D surface visualization of the segmented ILM layer of an OCT volume in the A2 A dataset. Left: FCBR with B-scan pre-alignment by NoRMCorre, and right: our method.

not preprocess them with motion correction for compared methods or manual segmentation).

Table 5

Ablation study results evaluated by the mean absolute distance (μm) between the predicted and ground truth surface locations (std. in parentheses). The asterisks denote statistically significant differences from our proposed full model with the Wilcoxon signed-rank test (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

A2A dataset

Ablation	ILM (AMD)	ILM (Normal)	IRPE (AMD)	IRPE (Normal)	OBM (AMD)	OBM (Normal)	Overall
no_align	2.25 (3.77)***	1.40 (0.42)	3.14 (1.72)***	2.18 (1.37)*	4.96 (3.26)***	2.49 (0.40)**	3.00 (1.78)***
pre_align	1.80 (2.36)	1.30 (0.49)	3.09 (1.79)***	**2.05** (1.40)	4.75 (3.61)*	**2.34** (0.37)*	2.77 (1.79)*
cascade	1.73 (2.13)**	**1.26** (0.45)**	3.08 (1.98)*	2.16 (1.35)**	4.53 (3.02)	2.45 (0.40)**	2.74 (1.56)*
no_smooth	**1.68** (1.84)	1.27 (0.47)	3.10 (1.97)***	2.13 (1.45)*	4.84 (3.43)***	2.45 (0.41)*	2.81 (1.53)
3D-3D	1.87 (2.19)*	1.31 (0.46)*	3.12 (1.74)*	2.13 (1.45)***	4.78 (2.99)***	2.43 (0.40)*	2.85 (1.82)**
Proposed	1.80 (1.97)	1.30 (0.52)	**2.91** (1.61)	2.10 (1.35)	**4.34** (2.55)	2.40 (0.38)	**2.68** (1.39)

JHH dataset

Ablation	ILM	RNFL-GCL	IPL-INL	INL-OPL	OPL-ONL	ELM	IS-OS	OS-RPE	BM	Overall
no_align	2.29 (0.47)	2.75 (0.64)	2.92 (0.37)	3.20 (0.45)	2.84 (0.68)	**2.65** (0.76)	**1.94** (1.02)	**3.39** (0.82)	3.09 (2.07)	2.79 (0.57)
no_smooth	2.31 (0.36)*	2.85 (0.70)*	2.91 (0.56)	3.26 (0.39)**	2.70 (0.59)**	2.73 (0.61)	1.99 (0.55)	3.54 (1.07)	**2.90** (2.02)	2.80 (0.49)
3D-3D	2.34 (0.35)*	2.88 (0.73)**	3.03 (0.56)**	3.50 (0.52)**	2.80 (0.63)**	**2.65** (0.58)	1.99 (0.95)	3.59 (0.90)	3.11 (1.58)	2.88 (0.51)*
Proposed	**2.21** (0.35)	**2.73** (0.61)	**2.79** (0.42)	**3.18** (0.33)	**2.62** (0.58)	**2.65** (0.52)	2.04 (0.73)	3.56 (1.04)	3.19 (2.02)	**2.77** (0.51)

Table 6

Layer segmentation results evaluated by the mean absolute distance (μm) between the predicted and ground truth surface locations on the Duke DME dataset (std. in parentheses). Results of FCBR (He et al., 2019a, 2021) and three SOTA graph-based methods, i.e., Chiu et al. (2015), Karri et al. (2016), and Rathke et al. (2017), are included for comparison. *Note:* results of comparing methods are directly cited from He et al. (2019a, 2021) based on the same data split and evaluation protocol, with no standard deviation reported though.

Method	ILM	RNFL-GCL	IPL-INL	INL-OPL	OPL-ONL	IS-OS	OS-RPE	BM	Overall
Chiu et al. (2015)	6.59	8.38	9.04	11.02	11.01	4.84	5.74	5.91	7.82
Karri et al. (2016)	**4.47**	11.77	11.12	17.54	16.74	4.99	5.35	4.30	9.54
Rathke et al. (2017)	4.66	6.78	8.87	11.02	13.60	4.61	8.06	5.11	7.71
FCBR (He et al., 2019a, 2021)	4.51	**6.71**	8.29	10.71	9.88	**4.41**	4.52	4.61	6.70
Ours	4.62 (0.63)	7.51 (2.13)	**7.52 (1.73)**	**8.49 (2.09)**	**9.40 (2.49)**	4.75 (1.16)	**4.50 (0.50)**	**4.16 (0.50)**	**6.37 (1.01)**

5.5.3. Ablation study

Next, we conduct ablation experiments to verify the effectiveness of the design and each module of the proposed framework. Specifically, we evaluate several variants of our model: no_align (without the alignment branch or pre-alignment), pre_align (without the alignment branch but pre-aligned by NoRMCorre (Pnevmatikakis and Giovannucci, 2017)), cascade (cascading an alignment network and a pure 3D encoder–decoder segmentation network; or in other words, breaking our model into two cascading, exclusive networks for B-scan alignment and 3D segmentation, respectively), no_smooth (without the global coherence loss $\mathcal{L}_{\text{SmoothS}}$), and 3D-3D (replacing the encoder G_f with 3D CNNs). The results on the A2 A dataset are presented in Table 5 top, from which several conclusions can be drawn. First, the variant without any alignment (no_align) yields the worst results, suggesting that the mismatch between B-scans does have a negative impact on 3D analysis of OCT data such as the 3D surface segmentation. Second, our full model integrating the alignment branch improves over both pre_align and cascade. We speculate this is because the alignment branch can produce better alignment results than pre_align, and more importantly, it produces a slightly different alignment each time, serving as a kind of data and feature augmentation for enhanced diversity for the segmentation decoder G_s. Third, removing $\mathcal{L}_{\text{SmoothS}}$ (no_smooth) apparently decreases the performance, demonstrating its effectiveness in exploiting the anatomical prior of smoothness. Lastly, our hybrid 2D–3D framework outperforms its counterpart 3D-3D network, indicating that the 2D CNNs can better deal with the mismatched B-scans prior to proper realignment.

The ablation results on the JHH dataset are shown in Table 5 bottom. Since this dataset was acquired on a scanner with built-in motion correction, we do not evaluate the pre_align or cascading variant on it. As can be seen, our full model still slightly outperforms the no_align variant even on the (theoretically) motion-free data, which again may be attributed to the side benefit of data and feature augmentation of the alignment branch G_a. Meanwhile, the full model also slightly outperforms the no_smooth variant. Last but not least, our hybrid 2D–3D architecture significantly outperforms the 3D-3D counterpart with appreciable margins, suggesting that the former can better handle the anisotropic OCT volumes even in the absence of obvious motion artifact.

5.5.4. Performance on data with severe pathology

The Duke DME dataset contains patients with severe DME pathology, especially for the ones in the testing split with damaged retinal structures by large pathological regions. Therefore, we use the DME dataset to assess the applicability of our method to the group of data where large variations can be caused by severe pathologies. In addition, we compare the performance of our method to that of several exiting methods which were also evaluated on the dataset, including FCBR (He et al., 2019a, 2021) and three SOTA graph-based methods: (Chiu et al., 2015), Karri et al. (2016), and Rathke et al. (2017) (results of comparing methods are from He et al. (2019a, 2021) based on the same data split and evaluation protocol). The results are shown in Table 6. As we can see, our method achieves the lowest overall MAD averaged over eight surfaces, apparently outperforming FCBR and other methods by modest (with a 0.33 μm advantage) and substantial (with 1.34–3.17 μm advantages) margins, respectively. It also yields the lowest MADs for five surfaces. Notably, for the three surfaces commonly disrupted by DME (IPL-INL, INL-OPL, and OPL-ONL), e.g., in disorganization of the retinal inner layers (Sun et al., 2014), our method demonstrates apparent improvements over the existing SOTA. These results validate our method's applicability and efficacy on OCT data with severe pathology, too. Fig. 8 shows example segmentations by our method of two B-scans in the Duke DME dataset.

5.6. Semi-supervised results with sparse annotation

To simulate sparse annotation at different degrees of sparseness, we sample the original slice-wise annotation evenly with varying fraction factors. For example, a fraction factor of 1/8 means we take only one B-scan's annotation for every eight B-scans. Note that to avoid unwanted boundary effect, annotations of the first and last B-scans of an OCT volume are always included. In the extreme case, only three B-scans are annotated for an OCT volume, i.e., the first, last,

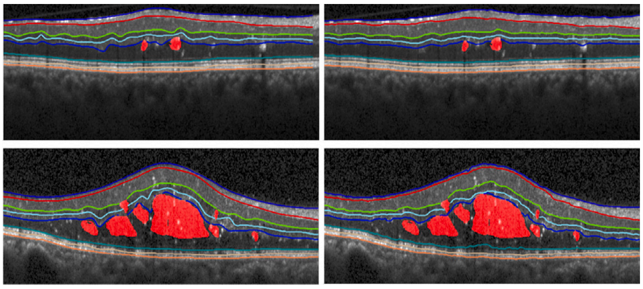


Fig. 8. Example manual segmentation (left) and segmentation by our framework (right) of two B-scans in the Duke DME dataset. Layer boundaries from top to bottom: ILM, RNFL-GCL, IPL-INL, INL-OPL, OPL-ONL, IS-OS, OS-RPE, and BM surfaces. The red area indicates edema. Note that our framework correctly segments the layers in the presence of modest and severe pathologies.

and middle ones. As it is not straightforward to optimally extend the SOTA fully supervised methods (i.e., RelayNet, MGU-Net, FCBR, IPM, and DDP) for the semi-supervised settings, we only use the annotated B-scans for their training. In addition, we further include an uncertainty-guided semi-supervised method exclusively developed for OCT layer segmentation, namely U-SLS (Sedai et al., 2019), for comparison in the semi-supervised settings. We implement and empirically optimize U-SLS.

The results on the A2 A and JHH datasets are shown in Fig. 9. The performance of all methods degrades with the decreasing number of annotated B-scans, as expected. Yet the extent of degradation varies, and for all the evaluated fraction factors (from 1/4 to 1/24) our method maintains the best performance of all methods on both datasets. Specifically, the performance of our method is relatively stable for fraction factors down to 1/12, and its advantage over other methods becomes most prominent in the extreme case of three annotations (1/20 and 1/24 faction factors on the A2 A and JHH datasets, respectively). With equal and less than 1/8 of the B-scans annotated, our method is significantly better than all other ones on both datasets (except for DDP in the 1/8 setting on the JHH dataset), as indicated by the Wilcoxon signed-rank test. Notably, the performance of our method with only three annotations is equal to or better than that of ReLayNet (Roy et al., 2017), MGU-net (Li et al., 2021) and U-SLS (Sedai et al., 2019) with full annotations on both datasets, suggesting its practical usability with sparse annotation. We attribute the superior performance of our method to the effective use of unannotated B-scan images by enforcing 3D surface coherence of the retinal layers, and the coupling of B-scan layer segmentation and motion correction.

6. Discussion and conclusion

This work presented a novel hybrid 2D–3D framework for simultaneous B-scan alignment and retinal surface regression of volumetric OCT data, which was applicable and proved effective to both fully and semi-supervised settings. The core idea behind our framework was the global coherence of the retinal layer surfaces both within and across-B-scan. Experimental results on three public clinic datasets and a synthetic dataset showed that our framework could effectively align the B-scans for motion correction and that it was superior to existing state-of-the-art methods for retinal layer segmentation in both fully and semi-supervised settings. Also, the ablative experiments verified the efficacy of the design and newly proposed modules of our framework.

The core motivation of this work was that smoothness was an intrinsic property of the retinal layers. Correspondingly, we proposed two losses to make use of the natural smoothness: the supervised B-scan alignment loss $\mathcal{L}_{\mathrm{SmoothA}}$ and the regulating global coherence loss $\mathcal{L}_{\mathrm{SmoothS}}$. The efficacy of these two losses was validated by the ablation experiments, contributing to the superior performance of our method. Further, these two losses also enabled our framework to use B-scans of sparsely annotated OCT volumes for effective semi-supervised segmentation. In contrast to FCBR (He et al., 2019a), which was among the previous best-performing fully supervised methods and only paid attention to the intra-B-scan layer coherence, our framework comprehensively took into account the complete 3D layer coherence, both intra- and inter-B-scan.

Our global smoothness loss $\mathcal{L}_{\mathrm{SmoothS}}$ took the form $\sum_{b=1}^{N_B}\sum_{a=1}^{N_A}\left\|\nabla\hat{S}(b,a)\right\|^2$, which can be dated to the classical Mumford-Shah functional (Mumford and Shah, 1989). When functioning alone, it preferred a flat surface $\hat{S}=c$, where c is a constant. In practice, such loss is almost always used with other loss function(s) as a regulating term, e.g., our overall optimization objective in Eq. (8). Appropriately weighed in this case, $\mathcal{L}_{\mathrm{SmoothS}}$ encouraged locally constant and slowly varying surfaces as a compromise, which was also our desirable notion of “smoothness” in this work. Given this notion, most foveal pit and pathology regions can be considered smooth for their slowly happening transitions and relative local constancy. Note that this notion of smoothness could also handle sudden jumps between normal tissue and severe pathology (i.e., edges), where piece-wise smooth surfaces on both sides of the edges were preferred to a single flat surface due to the balanced effects of the various loss terms. This was because, with proper weights, the decreases in other losses outweighed the increase in $\mathcal{L}_{\mathrm{SmoothS}}$ due to the jump on the edge. Fig. 4, Fig. 6, and Fig. 8 showed examples of successful segmentation by our framework in (1) the foveal pit region, (2) an AMD case with drusen, and (3) DME cases with minor to severe pathology regions, respectively. Also, the quantitative evaluation results in Table 3, Table 4, and Table 6 showed that the performance of our framework was not appreciably affected by these regions. Therefore, our framework generally worked well in the presence of 3D incoherence/discontinuity due to the natural structure and pathology of the retinal layers.

Meanwhile, small, early disruptions in the layer boundaries that are in the scale of artifacts may exhibit differently. On the one hand, the possibilities were low for such disruptions to be mistaken for misalignment artifacts by our framework. As we only considered B-scan-wise realignment, the impact of local disruptions of a few layers could be effectively mitigated by most other layers. On the other hand, it was possible that such subtle disruptions might be smoothed out if the smoothness constraint was overly emphasized by an improper

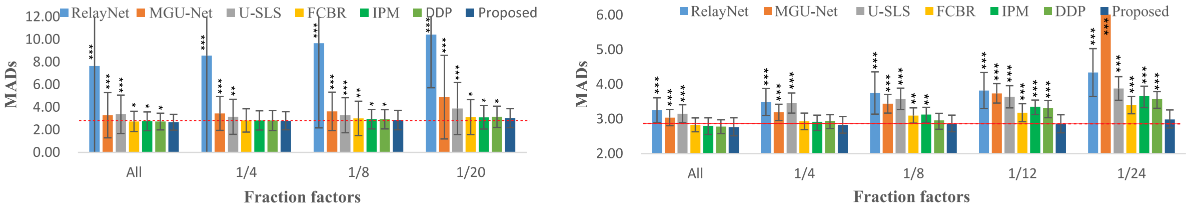


Fig. 9. Semi-supervised layer segmentation results in overall MAD (μm) on the A2 A (left) and JHH (right) datasets with different fraction factors (standard deviation overlaid). The dashed horizontal lines indicate the performance of our proposed method with the 1/8 fraction factor. The asterisks denote statistically significant differences from our proposed method with the Wilcoxon signed-rank test (*: $p<0.05$; **: $p<0.01$; ***: $p<0.001$).

Table 7
Comparison of model complexity and inference time with three existing best performing methods: FCBR (He et al., 2019a, 2021), IPM (Xie et al., 2022a), and DDP (Xie et al., 2022b). The inference time is evaluated by aligning and segmenting the OCT volumes from the A2A dataset using an NVIDIA 2080 Ti GPU. For fair comparison, the inference times of comparing methods include pre-alignment by NoRMCorre (Pnevmatikakis and Giovannucci, 2017).

	FCBR	IPM	DDP	Proposed
Num. parameters (million)	1.07	134.49	134.50	4.28
Inference time (second)	1.81	1.76	1.92	2.75

weight and thus harmful. In this work, we empirically found that setting the weights of $\mathcal{L}_{\text{SmoothS}}$ for different layer surfaces according to their extents of natural smoothness estimated from ground truth segmentation worked well for all three evaluated datasets. For potential application to datasets of primarily more subtle, early pathology, we caution that it may be necessary to identify optimal weights with more rigorous/advanced techniques such as grid search. In addition, it would be interesting to explore piece-wise smoothness constraint, which explicitly accommodates edges.

Coupling the B-scan alignment and layer segmentation also contributed to the superior performance of our method. On the A2 A dataset, which was subject to appreciable motion artifact (Table 5 top), pre-aligning the OCT volumes was 0.1 μm short compared to the proposed framework in overall MAD. We conjecture this was due to the side benefit of data and feature augmentation of the alignment branch G_a. Meanwhile, on theoretically motion-free data (the JHH dataset; Table 5 bottom), incorporating the alignment branch did not harm the segmentation accuracy, but improved it slightly. This is desirable as we can employ a unified framework for volumetric OCT data with and without motion artifact. Lastly, our method also outperformed previous, dedicated motion correction methods.

Considering that our hybrid 2D–3D framework included two 3D encoders—one for segmentation and the other for B-scan alignment, we investigate its model complexity and inference efficiency compared to existing best performing methods which all employed pure 2D networks: FCBR (He et al., 2019a, 2021), IPM (Xie et al., 2022a), and DDP (Xie et al., 2022b). As shown in Table 7, our model had more parameters than FCBR (4.28 versus 1.07 millions)—as expected, but the difference (in theory about 12 MiB memory for commonly used single-precision floats) was negligible on most modern hardware. Meanwhile, IPM and DDP employed deeper networks of significantly larger scales with more layers and feature channels, both having ∼134.5 million parameters. As to the inference efficiency, our method was less than a second slower than the other methods (2.75 versus 1.81, 1.76, and 1.92 s). We regard the marginally slower speed of our method as an acceptable trade-off in practice, especially to applications where 3D continuity or annotation cost is crucial.

This work also had some limitations. First, like the existing approaches to B-scan motion correction which did not require any additional reference image, our motion correction could not guarantee restoration of the true retinal curvature. However, not depending on any extra image acquisition, our method can be readily applied to existing archive data. Second, this work implemented a primitive method for transverse motion correction, although it empirically worked well on the specific data used in this study. In the future, we may need to develop more sophisticated, integrated techniques for potential data unlike those in this work.

Finally, we note from our literature search that compared with the active research on pushing the frontier of 2D layer segmentation in retinal OCT images with the significant development of deep neural networks, research on motion artifact correction has lagged behind. Therefore, we advocate more attention to the latter to facilitate effective 3D analysis of retinal OCT images to the community and consider this work a step forward.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We use public data and refer to the sources properly in this work; our code is released on GitHub with this revision.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology of the People's Republic of China (2021ZD0201900, 2021ZD0201903) and in part by the Shenzhen Basic Research Program, China (JCYJ20190809120205578).

References

Abràmoff, M.D., Garvin, M.K., Sonka, M., 2010. Retinal imaging and image analysis. IEEE Rev. Biomed. Eng. 3, 169–208.

Antony, B.J., Abràmoff, M.D., Harper, M.M., Jeong, W., Sohn, E.H., Kwon, Y.H., Kardon, R., Garvin, M.K., 2013. A combined machine-learning and graph-based framework for the segmentation of retinal surfaces in SD-OCT volumes. Biomed. Opt. Expr. 4 (12), 2712–2728.

Antony, B., Abràmoff, M.D., Tang, L., Ramdas, W.D., Vingerling, J.R., Jansonius, N.M., Lee, K., Kwon, Y.H., Sonka, M., Garvin, M.K., 2011. Automated 3-D method for the correction of axial artifacts in spectral-domain optical coherence tomography images. Biomed. Opt. Expr. 2 (8), 2403–2416.

Atif, J., Hudelot, C., Fouquier, G., Bloch, I., Angelini, E.D., 2007. From generic knowledge to specific reasoning for medical image interpretation using graph based representations. In: International Joint Conference on Artificial Intelligence. pp. 224–229.

Baghaie, A., Yu, Z., D'Souza, R.M., 2017. Involuntary eye motion correction in retinal optical coherence tomography: Hardware or software solution? Med. Image Anal. 37, 129–145.

Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P.M., Rueckert, D., 2018. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 586–594.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38 (8), 1788–1800.

Bavinger, J.C., Dunbar, G.E., Stem, M.S., Blachley, T.S., Kwark, L., Farsiu, S., Jackson, G.R., Gardner, T.W., 2016. The effects of diabetic retinopathy and pan-retinal photocoagulation on photoreceptor cell function as assessed by dark adaptometry. Invest. Ophthalmol. Vis. Sci. 57 (1), 208–217.

Bitarafan, A., Nikdan, M., Baghshah, M.S., 2020. 3D image segmentation with sparse annotation by self-training and internal registration. IEEE J. Biomed. Health Inf. 25 (7), 2665–2672.

Capps, A.G., Zawadzki, R.J., Yang, Q., Arathorn, D.W., Vogel, C.R., Hamann, B., Werner, J.S., 2011. Correction of eye-motion artifacts in AO-OCT data sets. In: Ophthalmic Technologies XXI, Vol. 7885. International Society for Optics and Photonics, p. 78850D.

Carass, A., Lang, A., Hauser, M., Calabresi, P.A., Ying, H.S., Prince, J.L., 2014. Multiple-object geometric deformable model for segmentation of macular OCT. Biomed. Opt. Expr. 5 (4), 1062–1074.

Chen, Z.-l., Wei, H., Shen, H.-l., Peng, P., Yue, K.-j., Li, J.-f., Zou, B.-j., 2018. Intraretinal layer segmentation and parameter measurement in optic nerve head region through energy function of spatial-gradient continuity constraint. J. Central South Univ. 25 (8), 1938–1947.

Cheng, J., Lee, J.A., Xu, G., Quan, Y., Ong, E.P., Kee Wong, D.W., 2016. Motion correction in optical coherence tomography for multi-modality retinal image registration. In: International Workshop on E. University of Iowa, pp. 65–72.

Chiu, S.J., Allingham, M.J., Mettu, P.S., Cousins, S.W., Izatt, J.A., Farsiu, S., 2015. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed. Opt. Expr. 6 (4), 1172–1194.

Chiu, S.J., Li, X.T., Nicholas, P., Toth, C.A., Izatt, J.A., Farsiu, S., 2010. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. Opt. Express 18 (18), 19413–19428.

Drexler, W., Fujimoto, J.G., 2008. State-of-the-art retinal optical coherence tomography. Progr. Retinal Eye Res. 27 (1), 45–88.

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

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

Ferguson, R.D., Hammer, D.X., Paunescu, L.A., Beaton, S., Schuman, J.S., 2004. Tracking optical coherence tomography. Opt. Lett. 29 (18), 2139–2141.

Fu, H., Xu, Y., Wong, D.W.K., Liu, J., 2016. Eye movement correction for 3D OCT volume by using saliency and center bias constraint. In: IEEE Region 10 Conference. IEEE, pp. 1536–1539.

Garvin, M.K., Abramoff, M.D., Wu, X., Russell, S.R., Burns, T.L., Sonka, M., 2009. Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. IEEE Trans. Med. Imaging 28 (9), 1436–1447.

He, Y., Carass, A., Liu, Y., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2019a. Fully convolutional boundary regression for retina OCT segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 120–128.

He, Y., Carass, A., Liu, Y., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2021. Structured layer surface segmentation for retina OCT using fully convolutional regression networks. Med. Image Anal. 68, 101856.

He, Y., Carass, A., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2019b. Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls. Data Brief 22, 601–604.

Hood, D.C., Raza, A.S., 2014. On improving the use of OCT imaging for detecting glaucomatous damage. Br. J. Ophthalmol. 98 (Suppl 2), ii1–ii9.

Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al., 1991. Optical coherence tomography. Science 254 (5035), 1178–1181.

Jáñez-Escalada, L., Jáñez-García, L., Salobrar-García, E., Santos-Mayo, A., de Hoz, R., Yubero, R., Gil, P., Ramírez, J.M., 2019. Spatial analysis of thickness changes in ten retinal layers of Alzheimer's disease patients based on optical coherence tomography. Sci. Rep. 9 (1), 1–14.

Kansal, V., Armstrong, J.J., Pintwala, R., Hutnik, C., 2018. Optical coherence tomography for glaucoma diagnosis: An evidence based meta-analysis. PLoS One 13 (1), e0190621.

Karri, S., Chakraborthi, D., Chatterjee, J., 2016. Learning layer-specific edges for segmenting retinal layers with large deformations. Biomed. Opt. Expr. 7 (7), 2888–2901.

Keane, P.A., Liakopoulos, S., Jivrajka, R.V., Chang, K.T., Alasil, T., Walsh, A.C., Sadda, S.R., 2009. Evaluation of optical coherence tomography retinal thickness parameters for use in clinical trials for neovascular age-related macular degeneration. Invest. Ophthalmol. Vis. Sci. 50 (7), 3378–3385.

Ker, J., Wang, L., Rao, J., Lim, T., 2017. Deep learning applications in medical image analysis. IEEE Access 6, 9375–9389.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knoll, B., Simonett, J., Volpe, N.J., Farsiu, S., Ward, M., Rademaker, A., Weintraub, S., Fawzi, A.A., 2016. Retinal nerve fiber layer thickness in amnestic mild cognitive impairment: Case-control study and meta-analysis. Alzheimer's Dementia: Diagn. Assess. Disease Monitor. 4, 85–93.

Kocaoglu, O.P., Ferguson, R.D., Jonnal, R.S., Liu, Z., Wang, Q., Hammer, D.X., Miller, D.T., 2014. Adaptive optics optical coherence tomography with dynamic retinal tracking. Biomed. Opt. Expr. 5 (7), 2262–2284.

Kraus, M.F., Liu, J.J., Schottenhamml, J., Chen, C.-L., Budai, A., Branchini, L., Ko, T., Ishikawa, H., Wollstein, G., Schuman, J., et al., 2014. Quantitative 3D-OCT motion correction with tilt and illumination correction, robust similarity measure and regularization. Biomed. Opt. Expr. 5 (8), 2591–2613.

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

Lang, A., Carass, A., Hauser, M., Sotirchos, E.S., Calabresi, P.A., Ying, H.S., Prince, J.L., 2013. Retinal layer segmentation of macular OCT images using boundary classification. Biomed. Opt. Expr. 4 (7), 1133–1152.

Lezama, J., Mukherjee, D., McNabb, R.P., Sapiro, G., Kuo, A.N., Farsiu, S., 2016. Segmentation guided registration of wide field-of-view retinal optical coherence tomography volumes. Biomed. Opt. Expr. 7 (12), 4827–4846.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging 37 (12), 2663–2674.

Li, J., Jin, P., Zhu, J., Zou, H., Xu, X., Tang, M., Zhou, M., Gan, Y., He, J., Ling, Y., et al., 2021. Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images. Biomed. Opt. Expr. 12 (4), 2204–2220.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, X., Fu, T., Pan, Z., Liu, D., Hu, W., Li, B., 2018a. Semi-supervised automatic layer and fluid region segmentation of retinal optical coherence tomography images using adversarial learning. In: 25th IEEE International Conference on Image Processing. IEEE, pp. 2780–2784.

Liu, X., Fu, T., Pan, Z., Liu, D., Hu, W., Liu, J., Zhang, K., 2018b. Automated layer segmentation of retinal optical coherence tomography images using a deep feature enhanced structured random forests classifier. IEEE J. Biomed. Health Inf. 23 (4), 1404–1416.

Liu, H., Wei, D., Lu, D., Li, Y., Ma, K., Wang, L., Zheng, Y., 2021. Simultaneous alignment and surface regression using hybrid 2D-3D networks for 3D coherent layer segmentation of retina OCT images. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 108–118.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

McNabb, R.P., LaRocca, F., Farsiu, S., Kuo, A.N., Izatt, J.A., 2012. Distributed scanning volumetric SDOCT for motion corrected corneal biometry. Biomed. Opt. Expr. 3 (9), 2050–2065.

Montuoro, A., Wu, J., Waldstein, S., Gerendas, B., Langs, G., Simader, C., Schmidt-Erfurth, U., 2014. Motion artefact correction in retinal optical coherence tomography using local symmetry. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 130–137.

Mumford, D.B., Shah, J., 1989. Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Appl. Math..

Novosel, J., Vermeer, K.A., De Jong, J.H., Wang, Z., Van Vliet, L.J., 2017. Joint segmentation of retinal layers and focal lesions in 3-D OCT data of topologically disrupted retinas. IEEE Trans. Med. Imaging 36 (6), 1276–1286.

Pnevmatikakis, E.A., Giovannucci, A., 2017. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. J. Neurosci. Methods 291, 83–94.

Rathke, F., Desana, M., Schnörr, C., 2017. Locally adaptive probabilistic models for global segmentation of pathological OCT scans. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 177–184.

Ricco, S., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J., 2009. Correcting motion artifacts in retinal spectral domain optical coherence tomography via image registration. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 100–107.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 234–241.

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

Saidha, S., Syc, S.B., Ibrahim, M.A., Eckstein, C., Warner, C.V., Farrell, S.K., Oakley, J.D., Durbin, M.K., Meyer, S.A., Balcer, L.J., et al., 2011. Primary retinal pathology in multiple sclerosis as detected by optical coherence tomography. Brain 134 (2), 518–533.

Sánchez Brea, L., Andrade De Jesus, D., Shirazi, M.F., Pircher, M., van Walsum, T., Klein, S., 2019. Review on retrospective procedures to correct retinal motion artefacts in OCT imaging. Appl. Sci. 9 (13), 2700.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., Garnavi, R., 2019. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 282–290.

Shah, A., Zhou, L., Abrámoff, M.D., Wu, X., 2018. Multiple surface segmentation using convolution neural nets: Application to retinal layer segmentation in OCT images. Biomed. Opt. Expr. 9 (9), 4509–4526.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Sun, J.K., Lin, M.M., Lammer, J., Prager, S., Sarangi, R., Silva, P.S., Aiello, L.P., 2014. Disorganization of the retinal inner layers as a predictor of visual acuity in eyes with center-involved diabetic macular edema. JAMA Ophthalmol. 132 (11), 1309–1316.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6450–6459.

Vienola, K.V., Braaf, B., Sheehy, C.K., Yang, Q., Tiruveedhula, P., Arathorn, D.W., de Boer, J.F., Roorda, A., 2012. Real-time eye motion compensation for OCT imaging with tracking SLO. Biomed. Opt. Expr. 3 (11), 2950–2963.

Wang, S., Cao, S., Chai, Z., Wei, D., Ma, K., Wang, L., Zheng, Y., 2020. Conquering data variations in resolution: A slice-aware multi-branch decoder network. IEEE Trans. Med. Imaging 39 (12), 4174–4185.

Wang, G., Shapey, J., Li, W., Dorent, R., Demitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., et al., 2019. Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 264–272.

Wei, D., Weinstein, S., Hsieh, M.-K., Pantalone, L., Kontos, D., 2018. Three-dimensional whole breast segmentation in sagittal and axial breast MRI with dense depth field modeling and localized self-adaptation for chest-wall line detection. IEEE Trans. Biomed. Eng. 66 (6), 1567–1579.

Xie, H., Pan, Z., Zhou, L., Zaman, F.A., Chen, D.Z., Jonas, J.B., Xu, W., Wang, Y.X., Wu, X., 2022a. Globally optimal OCT surface segmentation using a constrained IPM optimization. Opt. Express 30 (2), 2453–2471.

Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K., 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision. pp. 305–321.

Xie, H., Xu, W., Wu, X., 2022b. A deep learning network with differentiable dynamic programming for retina OCT surface segmentation. arXiv preprint arXiv:2210.06335.

Xu, J., Ishikawa, H., Tolliver, D., Wollstein, G., Miller, G., Bilonick, R., Kagemann, L., Schuman, J., 2009. Shape context algorithm applied to correct eye movement artifacts on three-dimensional (3D) spectral domain optical coherence tomography (SD-OCT). Invest. Ophthalmol. Vis. Sci. 50 (13), 1104.

Yazdanpanah, A., Hamarneh, G., Smith, B., Sarunic, M., 2009. Intra-retinal layer segmentation in optical coherence tomography using an active contour approach. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 649–656.

Zhang, J., Xie, Y., Zhang, P., Chen, H., Xia, Y., Shen, C., 2019. Light-weight hybrid convolutional network for liver tumor segmentation. In: International Joint Conference on Artificial Intelligence. pp. 4271–4277.

Zheng, H., Perrine, S.M.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 802–812.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3D medical image analysis. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 384–393.