Medical Image Analysis (2020)

Contents lists available at ScienceDirect

Medical Image Analysis

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

Improving Statistical Power Of Glaucoma Clinical Trials Using An Ensemble Of Cyclical Generative Adversarial Networks

Georgios Lazaridis[a,c,d,*], Marco Lorenzi[b], Sebastien Ourselin[c], David Garway-Heath[d]

[a] *Centre for Medical Image Computing, University College London, London, United Kingdom*
[b] *Université Côte d'Azur, Inria, Epione Team, 06902 Sophia Antipolis, France*
[c] *School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom*
[d] *NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and the Institute of Ophthalmology, University College London, London, United Kingdom*

ARTICLE INFO

Article history:
Received 2020
Received in final form 2020
Accepted 2020
Available online 2020

Communicated by 2020

Keywords: clinical trials, glaucoma, optical coherence tomography, deep learning, perceptual loss, GAN, label fusion, statistical power

ABSTRACT

Albeit spectral-domain OCT (SDOCT) is now in clinical use for glaucoma management, published clinical trials relied on time-domain OCT (TDOCT) which is characterized by low signal-to-noise ratio, leading to low statistical power. For this reason, such trials require large numbers of patients observed over long intervals and become more costly. We propose a probabilistic ensemble model and a cycle-consistent perceptual loss for improving the statistical power of trials utilizing TDOCT. TDOCT are converted to synthesized SDOCT and segmented via Bayesian fusion of an ensemble of GANs. The final retinal nerve fibre layer segmentation is obtained automatically on an averaged synthesized image using label fusion. We benchmark different networks using i) GAN, ii) Wasserstein GAN (WGAN) (iii) GAN + perceptual loss and iv) WGAN + perceptual loss. For training and validation, an independent dataset is used, while testing is performed on the UK Glaucoma Treatment Study (UKGTS), i.e. a TDOCT-based trial. We quantify the statistical power of the measurements obtained with our method, as compared with those derived from the original TDOCT. The results provide new insights into the UKGTS, showing a significantly better separation between treatment arms, while improving the statistical power of TDOCT on par with visual field measurements.

1. Introduction

Glaucoma is the leading cause of irreversible blindness worldwide and the second major cause for blind registration in the UK (Bunce and Wormald, 2008). It is a progressive optic neuropathy in which retinal ganglion cell (RGC) axon loss, probably as a consequence of damage at the optic disc, causes a loss of vision, predominantly affecting the mid-peripheral visual field (VF) and in the "macula vulnerability zone" (Hood et al., 2013). Evaluating the progression rate of the pathology is crucial in order to assess the risk of functional impairment and to establish sound treatment strategies. Clinically, optical coherence tomography (OCT) is used as a surrogate measure to evaluate retinal ganglion cell loss by measuring retinal nerve fibre layer (RNFL) thickness around the optic nerve head (ONH), whereas standard automated perimetry (SAP) is employed to assess the status of the VF (Garway-Heath et al., 2015).

Glaucoma research has produced several clinical trials to monitor disease progression and the efficacy of disease-modifying drugs (Wormald et al., 2020). Typically, the observation period for trials of VF preservation in patients with

*Corresponding author
e-mail: g.lazaridis@cs.ucl.ac.uk (Georgios Lazaridis)

open-angle glaucoma has been > 5 years (Musch et al., 2009) with the shortest observation period lasting 30 months (Krupin et al., 2011). One of the disadvantages of long trial duration is that assessment of new interventions to prevent vision loss is not efficient and cost-effective in terms of drug development cost. For this reason, the likelihood of new treatments being made available for patient benefit is reduced. Published clinical trials with imaging outcomes preceded the introduction of high-resolution spectral-domain OCT (SDOCT) and relied on time-domain OCT (TDOCT), which is characterized by lower quality images. Thus, structural measurements in past studies provided low statistical power in detecting significant treatment effects. Such an example is the UK Glaucoma Treatment Study (UKGTS) (Garway-Heath et al., 2015). The UKGTS is the only glaucoma study to assess the vision-preserving efficacy of one disease-modifying drug, i.e. Latanoprost, with both VF and OCT outcomes. Nonetheless, the combination of TDOCT information with VF outcomes did not improve detection of a treatment effect. Improving the quality of image-related anatomical measurements is therefore of high priority for increasing statistical power in clinical trials which should lead to a shorter trial duration with cost-effective interventions.

Extensive efforts have been made to develop better image acquisition, reconstruction and processing methods for medical images. Specifically for OCT images, several methods attempt to decrease the noise and artifacts which can compromise diagnostic information (van Velthoven et al., 2007). However, despite advancements in OCT technology, B-scans are still contaminated by speckle noise (Du et al., 2014) low signal strength (Hardin et al., 2015) and motion artefacts (Asrani et al., 2014). Speckle noise, specifically, significantly deteriorates image contrast, preventing small and low-intensity structures to be detected, i.e. intra-retinal structures (Du et al., 2014; Bashkansky and Reintjes, 2000), compromising the clinical interpretation of OCT data. Consequently, automated segmentation algorithms of retinal layers may fail systematically (Asrani et al., 2014), leading to incorrect tissue thickness estimation that can potentially affect clinical decisions or trial outcomes. Classical schemes to denoise OCT B-scans, rely on software or hardware implementations. Software methods to denoise OCT images can employ filtering (Dabov et al., 2006), wavelet transform (Rabbani et al., 2013; Chang et al., 2000; Mayer et al., 2012), low-rank decomposition (Chang et al., 2000) or diffusion-based techniques (Bernardes et al., 2010), whereas hardware approaches use frequency (Pircher et al., 2003), angular (Desjardins et al., 2007) or spatial compounding (multi-frame averaging) (Bashkansky and Reintjes, 2000) to suppress noise. Although these methods have been shown to enhance image quality, they are limited by registration errors or longer acquisition times (Wu et al., 2013), computational complexity (Rabbani et al., 2013) and sensitivity with respect to the choice of parameters (Mayer et al., 2012). Moreover, knowledge about the underlying OCT generative process and the structures of the eye is not incorporated. This knowledge, however, is highly relevant to this task, given the complex and sample-dependent nature of noise. Thus, clinical usage of these algorithms is limited.

Meanwhile, deep learning algorithms based on Convolutional Neural Networks (CNNs) have been shown particularly efficient at extracting relevant image features from 2D and 3D images (LeCun et al., 2015). Recently, it was also shown that deep learning can provide previously unimaginable insights into images, as, for example predicting the sex of a person from a snapshot of their ocular fundus (Poplin et al., 2018). Even though this particular application is not clinically relevant, as sex can be readily known, it showcases that deep learning can identify links between quantities that may have been considered as disconnected. Therefore, deep learning networks are promising modeling methodologies when quantities that do not have a foreseen mathematical or even direct physical relationship, are considered. Based on this rationale, various methods for image super resolution (SR) using CNNs, such as GANs (Goodfellow et al., 2014), have been proposed to perform noise reduction or to transform image quality and appearance learning the semantic characteristics of their input domains (Nie et al., 2017; Wolterink et al., 2017; Ben-Cohen et al., 2017; Wang et al., 2018; Zhu et al., 2017; Isola et al., 2017; Halupka et al., 2018; Huang et al., 2019; Dong et al., 2014; Gondara, 2016; Yang et al., 2018; Chen et al., 2017; Chen et al., 2017; Li et al., 2018; Devalla et al., 2019; Shi et al., 2018; Fei et al., 2017; Johnson et al., 2016; Ledig et al., 2016).

In medical imaging, GANs have been successfully employed to address the ill-posed nature of cross-modal synthesis. For example, in (Nie et al., 2017; Wolterink et al., 2017; Ben-Cohen et al., 2017), GANs have been proposed to predict computed tomography (CT) and positron emission tomography (PET) images from magnetic resonance imaging (MRI) with positive results. Concerning image denoising and signal enhancement, GAN-based approaches have been adopted with significant performance gains (Halupka et al., 2018; Huang et al., 2019; Dong et al., 2014; Gondara, 2016; Yang et al., 2018; Chen et al., 2017; Chen et al., 2017; Li et al., 2018; Devalla et al., 2019; Shi et al., 2018; Fei et al., 2017; Wolterink et al., 2017). These works, however, may present important limitations for SR in OCT imaging. First, due to the restricted view of GANs spatial window, preservation of spatial smoothness and anatomical features in predictions is not always guaranteed. Second, the use of standard metrics, such as per-pixel mean-squared error (MSE), to assess joint statistics of results, may fail in properly quantifying spatial coherence of the predicted signal. Finally, single GAN predictions are characterized by spatial and intensity variability regardless of the loss function used. Therefore, to extract robust anatomical quantifications from the output of GANs, principled schemes accounting for prediction uncertainty must be developed. This requires, for instance, probabilistic modeling of the uncertainty of the underlying signal distributions on distinct image parts, to preserve anatomical structures and account for spatial coherency. For example, in (Wang et al., 2018), synthesis was achieved at different resolution scales, albeit not focusing on medical applications.

This paper presents a novel ensemble method to improve the signal-to-noise ratio of TDOCT imaging and subsequent image segmentation, thereby leading to improved statistical power with low quality images. Our methodology leverages Bayesian

fusion of modified GANs to infer morphological descriptors from low to high quality anatomical information. The transfer mapping is learned in one dataset and the proposed method is tested in an independent dataset, i.e. the UKGTS data, enhancing the power of TDOCT via quality transfer from SDOCT. As a result, RNFL segmentations are improved and further refined via the effective label-propagation of multi-atlas segmentation (MAS) inheriting the ability to preserve anatomical shape, including faint or invisible boundaries, e.g., between layers or layer regions. In particular, to preserve anatomical structures and account for spatial coherency, we require to learn a range of possible distributions on distinct image parts and propagate anatomical information to provide robust morphological assessment of the underlying anatomy. Generally, GANs are not stable (Goodfellow et al., 2014; Arjovsky et al., 2017), and their objective function depends on a pixel-wise loss function, e.g. based on L1 or L2 metrics, to make the generated output image closer to the ground-truth image. Although in (Isola et al., 2017), the authors use L1 instead of L2 loss to avoid oversmoothed edges and loss of details (Johnson et al., 2016; Ledig et al., 2016), sometimes results still suffer from blurring effects. For this reason, we propose a cycle-consistent perceptual loss, which itself is a deep CNN and we further explore the Wasserstein distance (Arjovsky et al., 2017) as an alternative metric between distributions in our modified cyclical GAN. As loss network we employ the VGG-19 network (Simonyan and Zisserman, 2015) pretrained on ImageNet (Russakovsky et al., 2014) and compute the difference between images in a standard feature space. Note that the target domain in our setting is still noisy, but to a far lesser degree. Furthermore, in order to improve synthesis and training, we separate actual layer signal from background information before training and stitch them back together during inference. Results illustrate that using the ensemble of modified GANs with different field of views, and with this separation taking place, as well as including the proposed cycle-consistent perceptual loss, does improve synthesis in all scenarios. Results on the UKGTS clinical trial further show a significantly better separation between treatment arms than conventional segmentation of TDOCT, and that imaging and visual field (VF) measurements have similar power to distinguish treatment groups. The paper is structured as follows. In Section 2, we present the studies used in our research. In Section 3, we introduce our proposed framework. Section 4 describes our experiments and results. Finally, Section 5 concludes the paper providing discussion and future perspectives. This work extends our previous research (Lazaridis et al., 2019), in which we introduced the basic ensemble cycleGAN framework using GAN loss. Here, we extend our previous publication by i) introducing the cycle-consistent perceptual loss as a novel optimization objective, ii) by benchmarking different networks, and iii) by adding novel experiments and improving the clinical validation with respect to the face validity of visual field measurements.

2. Data

We used two studies to validate and test our proposed methodology. For training and validation, we used the RAPID study (Garway-Heath et al., 2017). For testing, we consider the UKGTS trial (Garway-Heath et al., 2015). **Note that there are no common participants between the two datasets.**

2.1. RAPID Test-Retest Dataset

The RAPID dataset was acquired from volunteer patients attending the glaucoma clinics at Moorfields Eye Hospital NHS Foundation Trust, which functions as a district general and teaching hospital and a tertiary referral centre; VF testing and imaging was undertaken in the National Institute for Health Research Clinical Research Facility. Eighty-two stable glaucoma patients under standard treatment were recruited to a test–retest study. Seventy seven (148 eyes) of the participants recruited attended for up to 10 visits within a 3-month period, for a total of 1256 patient-eye visits. This data set was taken to represent a 'stable glaucoma' cohort. The RAPID study consists of 4,902 TDOCT (Carl Zeiss Meditec Inc., Dublin, CA, USA) and 1,789 SDOCT (SpectralisOCT, Heidelberg Engineering) images. **For SDOCT, a 3.5 mm-diameter scan circle centred on the optic disc with the eye-tracking system activated with Spectralis SDOCT Heidelberg Eye Explorer (Heidelberg Engineering, Heidelberg, Germany) (software version 5.2.4) was used. Automatic real-time (ART) function was activated, thereby allowing multiple frames, i.e. B-scans, to be averaged for speckle noise reduction. For TDOCT, the fast RNFL 3.4 scan protocol was used with TD Stratus OCT™ (Carl Zeiss Meditec Inc., Dublin, CA, USA) (software version 5.0). A scan circle of 3.4 mm in diameter consisting of 256 A-scans was positioned manually at the centre of the optic disc.** More details can be found in Garway-Heath et al. (2017).

2.2. UKGTS

The UKGTS is a multicentre, randomized, triple-masked, placebo-controlled trial assessing visual function preservation in newly diagnosed open-angle glaucoma (OAG) patients (trial registration number, ISRCTN96423140). 516 newly-diagnosed (previously untreated) participants with OAG were prospectively recruited at 10 UK centres between 2007 and 2010. The observation period was 2 years, with subjects monitored by VF testing, quantitative imaging, optic disc photography and tonometry at 11 scheduled visits. ONH structure was monitored with the Heidelberg Retina Tomograph at all study sites and with TDOCT and GDxECC Nerve Fiber Analyzer (Carl Zeiss Meditec Inc., Dublin, CA, USA) at study sites with those devices. The participants were allocated randomly to receive the IOP-reducing prostaglandin analog latanoprost (0.005%) or placebo eye drops. For testing, we consider the subset of UKGTS participants who had TDOCT imaging available (Garway-Heath et al., 2015) consisting of 373 glaucoma patients. The UGKTS dataset consists of 78,415 TDOCT images. More details can be found in Garway-Heath et al. (2015).

3. Methods

Fig.1 illustrates the flow diagram of the proposed architecture: Firstly, the training data is created by generating the maximum number of suitable TDOCT and SDOCT image pairs.

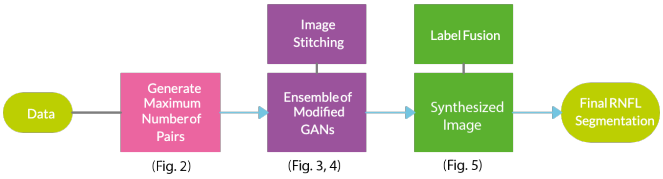


Fig. 1. Flow diagram of proposed training architecture.

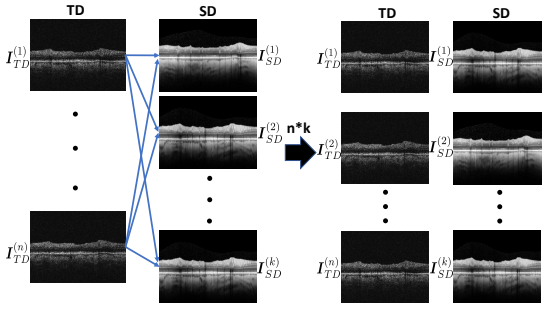


Fig. 2. A patient with n TDOCT images and k SDOCT images can theoretically produce a maximum of n×k images.

Secondly, after separating their background from the actual layer signal, we model them with our modified ensemble of cycleGANs using cycle-consistent perceptual loss. Thirdly, we stitch predictions with a painted-black background and average the three candidate super-resolved images. Finally, we propagate RNFL labels to this test image and obtain the final RNFL segmentation.

The definition of our framework requires a number of challenges to be addressed. Firstly, due to different acquisition protocols, the pairing between target SDOCT and predictor TDOCT training images is ill-defined. To solve this issue, we propose an automated method for target-predictor image pairing, i.e. Section 3.1. Secondly, extending our previous work (Lazaridis et al., 2019), in 3.2 - 3.7, we describe the proposed image synthesis model and objectives, including the cycle-consistent perceptual loss. Thirdly, we introduce our networks **(Section 3.8)** and in Section 3.9, **we present** our method to obtain representations accounting for the different spatial coherence of OCT images. This is a critical requirement as OCT signal is characterized by diverse degrees of noise and spatial information, whereas RFNL segmentation is subject to variability due to the different attributes of the synthesized images. Finally, in Section 3.11 we identify a probabilistic consensus strategy for RNFL segmentations.

3.1. Training Pairs Generation

Despite the fact that TDOCT and SDOCT images were obtained at each visit, across patients, there is not a mapping between the two sets of target and predictor acquisitions, respectively. The acquisitions are at different visits, but we can make the assumption that the underlying anatomy is comparable given the nature of the human anatomy and the strict inclusion criteria of the test-retest study (Garway-Heath et al., 2017). Moreover, a spatial matching can be estimated up to some noise level that will be subsequently accounted by the model. To generate a valid set of paired TD- and SD-OCT images, we establish a pairing based on local and global image descriptors given by (i) the vessel profile represented by the average pixel intensity of the retinal pigment epithelium (RPE), (ii) the contour of the internal limiting membrane (ILM) and (iii) the average norm of the deformation fields between the patient's test-retest TDOCT and SDOCT acquisitions. First, given the fact that the topography around the ONH undulates, we flatten TDOCT and SDOCT images based on a pilot estimate of the RPE, which is the most hyper-reflective layer. As a result, using this fixed vertical RPE offset, we align all images accordingly. Furthermore, we detect the vessels, using the estimation of the RPE, since they appear as shade-like bands in the RPE. We then use the dark-to-bright gradient image to determine the upper high-contrast boundary. This boundary is the contour of the ILM and we use Gaussian Process interpolation to further smooth it. To evaluate the matching between the descriptors in (i) and (ii), we employ the iterative closest point algorithm and to assess the image registration in (iii), we use mutual information. The robustness of our pairing methodology is evaluated on a dataset of synthetic images with various degrees of noise and spatial variability. We achieve 100% sensitivity in finding the right pair for each image (see Supplementary material). We note that a patient with n TDOCT and k SDOCT can theoretically produce a maximum of n×k images (Fig. 2). For instance, at a visit, a patient can have 9 left eye TDOCT acquisitions but 3 left eye SDOCT **acquisitions and thus, the pairing method results to** $9 \times 3 = 27$ **pairs**. Application to the RAPID dataset lead to 24,792 TDOCT and SDOCT pairs.

3.2. Image Synthesis Model

Typically, speckle noise in TDOCT acquisitions is multiplicative and drastically reduces the already low resolution. On the contrary, in SDOCT images, the noise model is still defined by speckle noise, but in a far lesser degree. Thus there is no clear way that indicates how data distributions of TDOCT and SDOCT images are related to each other. This makes it difficult to translate TDOCT to SDOCT images and more importantly to evaluate the resulting synthesized image. However, uncertainty in noise modeling can be ignored in adversarial denoising as the underlying OCT generative process and the structures of the human eye can be efficiently learnt. Given the complex and sample-dependent nature of noise, the model should efficiently learn high-level features and a representation of the data distribution from modest sized image patches. To this end, we propose the following image synthesis model. Let $\boldsymbol{I}_{TD} \in \mathcal{R}^{N \times N}$ be a TDOCT image and $\boldsymbol{I}_{SD} \in \mathcal{R}^{N \times N}$ be the corresponding SDOCT image. We seek to learn a mapping from the observed TDOCT image $\boldsymbol{I}_{TD}$ to the target SDOCT image $\boldsymbol{I}_{SD}$, $G: \boldsymbol{I}_{TD} \rightarrow \boldsymbol{I}_{SD}$. CycleGANs allow bidirectional synthesis between the source and the target domain. Thus, two mapping functions are incorporated: $G_1: \boldsymbol{I}_{SD} \rightarrow \boldsymbol{I}_{TD}$ and $G_2: \boldsymbol{I}_{TD} \rightarrow \boldsymbol{I}_{SD}$ where G_1 and G_2 are two generator CNNs. Each of the generator networks is trained adversarially using a

corresponding discriminator network, D_1 and D_2. The first generator network G_1 receives a source domain TDOCT image, as an input, $x \in \boldsymbol{I}_{TD}$, and outputs a synthetic target SDOCT image, $\hat{y} = G_1(x)$. D_1 receives as input both the synthetic output $\hat{y}$ and a paired image sampled from the desired target domain, $y \in \boldsymbol{I}_{SD}$. The two networks, G_1 and D_1, compete against each other, where D_1 acts as a binary classifier attempting to distinguish between the translated samples and the target domain samples. On the other hand, G_1 attempts to improve the quality of the translated output, thus deceiving the discriminator. A typical CycleGAN uses a combination of adversarial losses and the pixel-wise cycle-consistency loss. Here, we propose an extra cycle-consistent perceptual loss and further examine the use of Wasserstein loss on top of our proposed ensemble framework. Although Pix2Pix (Isola et al., 2017) was investigated in our first experiments, we noticed repeated vessel filling, leading to artificial information in the area of vessels and elsewhere. For this reason, we focus on modifying cycleGANs for 'almost-paired' image synthesis to obtain better representations of anatomical structures. In what follows, we introduce the losses and optimization tasks which incorporate our proposed perceptual loss.

3.3. Adversarial Loss

This training procedure is formulated as a min-max optimization task over the adversarial loss function:

$$\mathcal{L}_{GAN}(G, D) = \min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where x is a real image from the true data distribution p_{data}, and z is a noise vector sampled from the prior distribution p_z (e.g., uniform or Gaussian distribution). In practice, the generator G is modified to maximize $log(D(G(z)))$ instead of minimizing $\log(1 - D(G(z)))$ to mitigate the problem of gradient vanishing (Goodfellow et al., 2014). We use this modified non-saturating objective in all our experiments. A similar adversarial loss for the opposite mapping function $F : \boldsymbol{I}_{SD} \rightarrow \boldsymbol{I}_{TD}$ and its discriminator is used as well.

3.4. Cycle Consistency

In some cases, training GANs solely with adversarial losses is not sufficient since it may lead to mode collapse, where a set of different input images is mapped into a single image in the target domain (Zhu et al., 2017). Therefore, an additional constraint to regularize the mapping functions, i.e. reduce the mapping dimensions, is necessary, exploiting the property that synthesis should be cycle consistent. This is achieved by enforcing cycle consistency between the two mapping functions, G_1 and G_2. As a result, the two generator networks should satisfy the inversion $\hat{x} = G_2(G_1(x)) \approx x$ and $\hat{y} = G_1(G_2(y)) \approx y$.

$$\begin{aligned}\mathcal{L}_{cyc}(G_1, G_2) =& \mathbb{E}_{x\sim p_{data}(x)}[\|x - G_2(G_1(x))\|_1] \\ & \mathbb{E}_{y\sim p_{data}(y)}[\|y - G_1(G_2(y))\|_1]\end{aligned} \quad (2)$$

3.5. Wasserstein GAN Gradient Penalty

In adversarial training (Eq.1), the GAN-loss attempts to minimize the KL-divergence between the generated distribution and the true data distribution. In a Wasserstein GAN (WGAN) setting, the minimization search is equivalent to minimizing the Jensen-Shannon (JS) divergence between the generated and the real sample data distributions. Instead of computing a probability of the sample being real or fake, the discriminator instead evaluates an unbounded score of sample realism. The WGAN loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{WGAN}(G) &= -\mathbb{E}_{x\sim p_g(x)}[D(x)] \\ \mathcal{L}_{WGAN}(G) &= \mathbb{E}_{x\sim p_g(x)}[D(x)] + -\mathbb{E}_{x\sim p_r(x)}[D(x)]\end{aligned} \quad (3)$$

and solves the following minmax problem:

$$\begin{aligned}\mathcal{L}_{WGAN}(G, D) &= \min_G \max_D V(D, G) \\ &= -\mathbb{E}_{x\sim p_r(x)}[D(x)] + \mathbb{E}_z[D(G(z)]\end{aligned} \quad (4)$$

To accelerate convergence, Arjovsky et al. (2017) propose to clip the weights of the discriminator which, nevertheless, leads to a vanishing gradient, exploding gradients, or weights being pushed towards the extremes of the clipping range (Gulrajani et al., 2017). Hence, we impose a gradient penalty method (Gulrajani et al., 2017) and solve the following minmax problem:

$$\begin{aligned}\mathcal{L}_{WGAN}(G, D) &= \min_G \max_D V(D, G) \\ &= -\mathbb{E}_{x\sim p_r(x)}[D(x)] + \mathbb{E}_z[D(G(z)] \\ &+ \lambda \mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]\end{aligned} \quad (5)$$

where the first two terms perform a Wasserstein distance estimation, the last term is the gradient penalty term for network regularization and λ is a constant weighting parameter. Compared to the original GAN, WGAN: (a) does not use the log function in the losses and (b) removes the sigmoid activation from the final layer of the discriminator, so predictions are no longer constrained to fall in the range [0,1] but (c) clamps the weights to a small fixed range after every gradient update on the discriminator function.

3.6. Proposed Perceptual Loss

To keep image details or information content, perceptual loss functions are often used in cross-domain synthesis by extracting representations of a feature map. Typically, a pure MSE-based loss function is used, which tries to minimize the pixel-wise error between source and target image patches. Nevertheless, the MSE loss sometimes still suffers from the blurring effect and can potentially cause distortion or loss of details (Johnson et al., 2016). Here, instead of solely applying a MSE measure (see paragraph 3.4), a perceptual loss is additionally utilized for cyclical consistency. Fig. 3 illustrates the proposed cyclical perceptual loss. Thus, to minimize the difference of content representation between the source and target images, we use the following perceptual loss function defined in feature space:

$$\mathcal{L}_{Perc}(G) = \mathbb{E}_{(x,y)}[\frac{1}{h_i w_i d_i}\|\phi_i(G(z)) - \phi_i(x)\|_F^2] \quad (6)$$

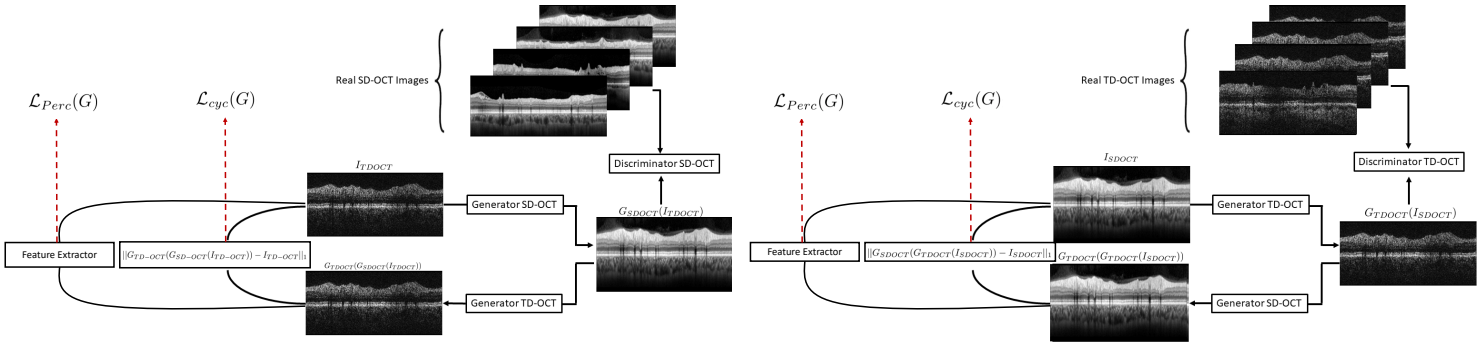


Fig. 3. Modified CycleGAN architecture with the proposed perceptual cyclical loss calculated using VGG-19 as a feature extractor.

where h_i, w_i, d_i, are the spatial height, weight and depth of the extracted feature map of the i^{th} layer of the feature extractor network, respectively. Specifically, the perceptual loss function extracts feature responses in different layers of a CNN. The deeper the network, the more the input image is represented by features instead of pixel values; higher layer features have larger receptive fields, representing actual image content and spatial structure (Johnson et al., 2016). For this reason, we use a deep VGG-19 network pretrained on the ImageNet classification task (Russakovsky et al., 2014) as a feature extractor; the VGG-19 contains 16 convolutional layers followed by 3 fully-connected layers. The output of the 16th layer is the feature map extracted by the VGG network and used in the perceptual loss function. We duplicate the OCT images to make them RGB-compatible before feeding them to the VGG network as the pretrained VGG network takes color images, while OCT images are grayscale.

3.7. Objectives

We formulate the four different objectives we compare in our proposed ensemble methodology.

- cycleGAN

$$\min_G \max_D \mathcal{L} = \mathcal{L}_{GAN} + \lambda_{cyc}\mathcal{L}_{cyc} \tag{7}$$

- cycleWGAN

$$\min_G \max_D \mathcal{L} = \mathcal{L}_{WGAN} + \lambda_{cyc}\mathcal{L}_{cyc} \tag{8}$$

- cycleGAN-Perceptual

$$\min_G \max_D \mathcal{L} = \mathcal{L}_{GAN} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{perc}\mathcal{L}_{perc} \tag{9}$$

- cycleWGAN-Perceptual

$$\min_G \max_D \mathcal{L} = \mathcal{L}_{WGAN} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{perc}\mathcal{L}_{perc} \tag{10}$$

where where λ_{cyc} and λ_{perc} are the weighting parameters for the cycle-consistency and perceptual losses, respectively, i.e. they control the trade-off between the GAN/WGAN adversarial loss and the VGG perceptual loss. For each case, we aim to solve:

$$\mathcal{G}^*, \mathcal{F}^* = arg \min_G \max_D \mathcal{L} \tag{11}$$

3.8. Networks

The generator part of the network contains two stride-2 convolutions, 9 residual blocks, and two fractionally strided convolutions with stride $\frac{1}{2}$. Similar to Johnson et al. (2016), we use instance normalization. To tackle the blurring effect, we further add skip connections to both generators; concatenate the output of each down-sampling layer to the input of the corresponding up-sampling layer.

The proposed perceptual calculator part of our network is the cycle-consistent perceptual loss network, which is the pre-trained VGG network (Simonyan and Zisserman, 2015). Fig. 3 illustrates the cycle consistent perceptual loss path through the feature extractor. The calculation of such losses does not require any explicit pairing of the input datasets during training although paired inputs increase consistency. The output images from generators G_1, G_2 and the reconstructed images $\hat{x}$, $\hat{y}$ are fed into the pre-trained VGG network for feature extraction. Then, the objective loss is computed using the extracted features from the 16th VGG-layer according to Eq. 6. The perceptual reconstruction error updates only the weights of the generators, while keeping the VGG parameters intact.

For the discriminator we use PatchGANs, i.e. ConvNets. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion (Isola et al., 2017). We note that a $N \times N$ PatchGAN is mathematically equivalent to manually chopping up the image into $N \times N$ overlapping patches, running a regular discriminator over each patch and averaging the results.

3.9. Ensemble GANs

The specific signal properties and anatomical geometry found in OCT images need to be addressed. As a result, to improve the accuracy and robustness of the modality transfer, we propose an ensemble of our perceptually modified cycleGANs (Zhu et al., 2017), i.e. Fig. 4. Geometry in OCT images is very specific, where the vitreous cavity, i.e. background, is clearly distinct with respect to the layers at the ILM border. As a result, we identify and separate layer signal from background using image stitching, exploiting the identification of the ILM before feeding images into our network. Moreover, while learning of mappings and spatial filters is usually performed using a fixed window in cycleGANs, a fixed spatial window modality

transfer method might not be adequate enough to capture all the spatial information necessary for synthesis as noise and signal properties are defined by different spatial scales. Thus, the probability for cross-modal distributions to share supports in latent space is reduced. To address this problem, we propose an ensemble of spatially coherent cycleGANs (Fig. 4) to learn TDOCT-to-SDOCT image mapping and to translate TDOCT images into synthesized SDOCT ones.

3.10. Training

Inspired by spatial compounding which is the most commonly adopted denoising scheme by OCT instrumentation, we propose the ensemble of GANs. The scheme is the following. To avoid producing scores for each image region where **minimal or not relevant** information is present, i.e. vitreous cavity (top **dark** background), we separate background from layer signal. Images are of size 512×512 and after background separation are 256×512. We load them as rectangular images, and modify receptive fields of different size, i.e. 128×128, 256×256, while in the latter case we use the full resized 512×512 image, i.e. ImageGAN. As a result, each GAN is trained by employing a different spatial window size on the pure signal: we use $\text{Patch}_{128\times128}$, $\text{Patch}_{256\times256}$ and $\text{Image}_{512\times512}$, learning a mapping from the observed TDOCT image $\boldsymbol{I}_{TD}$ and random noise vector $\boldsymbol{z}$, to the target SDOCT image $\boldsymbol{I}_{SD}$, $G\colon \{\boldsymbol{I}_{TD},\ \boldsymbol{z}\} \rightarrow \boldsymbol{I}_{SD}$. As a result, we train three GANs with layer pairs, whereas the backgrounds are painted black and are stitched back with the corresponding layers according to size. We produce predictions of size 512×512. A $N \times N$ PatchGAN is mathematically equivalent to manually chopping up the image into $N \times N$ overlapping patches, running a regular discriminator over each patch and averaging the results. Since our network is fully convolutional, the resulting learnt spatial filters should be in principle independent from image and patch size. However, in our experiments, we show that there is a significant difference between networks associated with different patch sizes, and combining them provides an optimal representation of RNFL appearance. Finally, we average the three synthesized candidates and the average synthesized stitched image $\bar{\boldsymbol{I}}$ is obtained. Note that to further preserve the morphological relationship between training pairs, cycleGANs were trained with windows centered at the same geometrical location in both pairs, i.e. we modify the training window to look at the same region in both input images. Albeit cycleGANs are used in the absence of paired aligned examples, we implicitly also show that having paired examples with input windows looking at the same location (instead of random ones) in input pairs improves prediction. Note that for TDOCT, the circular scan is centered on the ONH with a diameter of 3.4 mm, whereas for SDOCT, with a diameter of 3.5mm. Thus, really small alignment issues between pairs do exist by acquisition. Although preliminary experiments naturally took place using Pix2Pix (Isola et al., 2017) (used for aligned input training pairs), results indeed showed improved image quality, but falsely generated artificial information and repeated blood vessel 'filling-in'. As a result, we focused on modifying cycleGANs in order to bring them up to the task of training almost-aligned input pairs and getting the best out of them. Fig. 4 shows the proposed framework for OCT synthesis via the ensemble of perceptually modified GANs. For convenience, we use the notation $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{256\times256}$, $\boldsymbol{I}_{512\times512}$ for the previously mentioned generated images.

3.11. Label Fusion

After obtaining the average synthesized stitched image $\bar{\boldsymbol{I}}$, we need to find a sound RNFL segmentation taking into account the variability modelled during synthesis. Thus, we consider synthesized images as being in a theoretical stack of images: we use $\bar{\boldsymbol{I}}$ as test image, while, we use $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{256\times256}$, $\boldsymbol{I}_{512\times512}$, and the original $\boldsymbol{I}_{\text{TDOCT}}$ as atlases, here denoted by $\{I_n(\boldsymbol{x})\}_{n=1,\dots,4}$ (Fig. 5a). We need to propagate RNFL labels from the atlas image to the novel coordinates of the test image, where the label of each pixel is selected through a fusion scheme. To take into account the variability across atlases, we employ a Bayesian averaging technique, the graphical model of which is shown in Fig. 5b. Let $\{L_n(\boldsymbol{x})\}_{n=1,\dots,4}$ be the segmentations of atlases $\{I_n(\boldsymbol{x})\}$ and these atlases which are assumed to be registered to the test image $\bar{I}(\boldsymbol{x})$, with unknown labels $L(\boldsymbol{x})$. A label fusion approach tries to estimate the label map $\boldsymbol{L}$ associated with $\bar{\boldsymbol{I}}$, given the registered atlases. We assume that the posterior segmentation probability p factorizes over pixels:

$$p(\boldsymbol{L}|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) = \prod_{\boldsymbol{x}\in\Omega} p_x(L(\boldsymbol{x})|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) \tag{12}$$

The local label fusion model from Sabuncu et al. (2010) is chosen for inference of the labels, i.e. to model p_x. The model is based on a latent discrete field $\boldsymbol{M}(\boldsymbol{x})$ that indexes the segmentation of the test image at each location and which atlas generates it. Moreover, the image intensities $\bar{\boldsymbol{I}}$ and labels $\mathbf{L}$ are assumed to be conditionally independent given the field $\boldsymbol{M}$. Following Sabuncu et al. (2010), a Gaussian likelihood term for the intensities of the images and a LogOdds model which relies on the labels signed distance transform are used. To reflect the lower reliability for the atlases with lower registration accuracy, we choose a prior for the field $\boldsymbol{M}$ (Atzeni et al., 2018). The prior, for each 2D location $\boldsymbol{x}$ takes the form $p(M(\boldsymbol{x}) = n) \propto \exp(-k_n\alpha)$, where the coefficients k_n, $n = 1,2,3,4$, are the distances between the test image $\bar{\boldsymbol{I}}$ and the atlases, while α is a parameter that allows us to control the sharpness of the prior. Based on the measured agreement when evaluating GANs performance individually, decaying weights were set accordingly. We set the lowest distance value, $k_1 = 1$, for atlas $\boldsymbol{I}_{256\times256}$, and increasing ones, $k_i = i$, for atlases $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{512\times512}$ and $\boldsymbol{I}_{\text{TDOCT}}$, respectively based on our experimental results (see Supplementary Material). The labels posterior probability finally is:

$$p(L(\boldsymbol{x})|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) = \frac{\sum_{n=1}^{N} \mathcal{N}\left(\bar{I}(\boldsymbol{x}); I_n(\boldsymbol{x}), \sigma^2\right) e^{\rho D_x[L(x);\boldsymbol{L}_n]} e^{-k_n\alpha}}{\sum_{n=1}^{N} e^{-k_n\alpha} \mathcal{N}\left(\bar{I}(\boldsymbol{x}); I_n(\boldsymbol{x}), \sigma^2\right)} \tag{13}$$

where D_x is the signed distance transform evaluated at location x; $\mathcal{N}$ is the Gaussian probability density function; and ρ and σ^2 are the likelihood parameters.

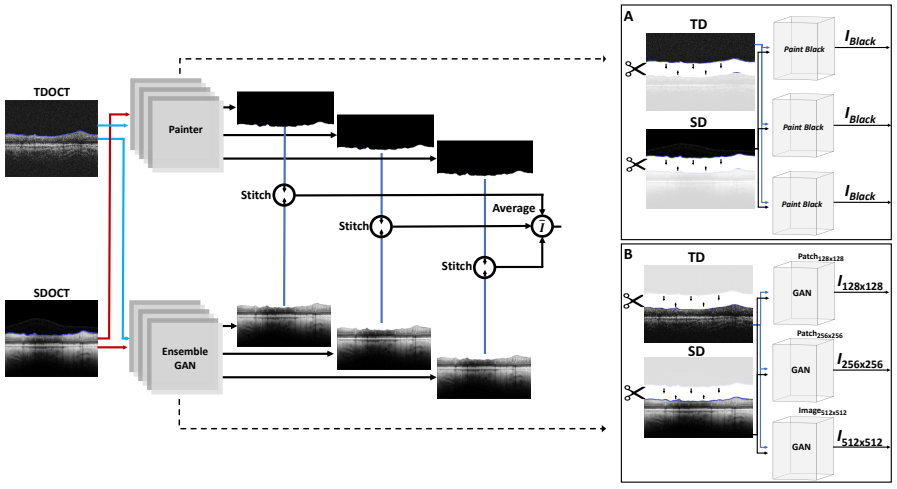
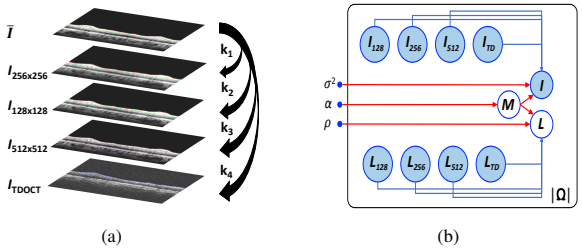


Fig. 4. SDOCT synthesis via ensemble of GANs. Box A: Backgrounds are painted black. Box B: Three GANs are trained with layer pairs. Synthesized images are stitched back with the backgrounds and the average synthesized stitched image is obtained. Separation of layers and background is illustrated with scissors.

4. Experiments and Results

4.1. Experimental Setup

For quantitative analysis in RAPID, we compare our proposed ensemble method to the original TDOCT and the ground truth SDOCT images. The proposed method, i.e. label fusion strategy on the GANs output with cycle-consistent perceptual loss plus image stitching is further compared to the results obtained using different optimization objectives. We calculate the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). We note that these metrics may not capture fine details in the image, and thus give high scores to images with unsatisfying quality. For this reason, to further quantify the anatomical plausibility of the synthesized SDOCT images, we segment their RNFL and compare the resulting average RNFL thickness with the original SDOCT average RNFL thickness. The intuition is that if we can produce realistic SDOCT images, an off-the-shelf segmentation model should estimate the same RNFL thickness as that obtained with the original data. In what follows, we adopt the layer segmentation model of Mayer et al. (2010). For label fusion, as atlases, we used the segmented RNFL sections of the synthesized SDOCT and the original TDOCT RNFL segmentation. For the test image, we used the average synthesized stitched image in which we register the retinal layers of the atlases. We use the method from Du et al. (2017) for non-rigid registration of OCT layers, and compute predictions for the final RNFL labels with Eq.13. The parameters were kept constant for all experiments: $\sigma^2 = 625$, $\rho = 30\mu m^{-1}$, $\alpha = 1mm^{-1}$. Testing on UKGTS was instead performed by quantifying the statistical power of the trial using the measurements obtained with our method compared with those derived from the Stratus TDOCT. All experiments were performed on a NVIDIA Titan X (12GB) GPU using PyTorch.



(a) (b)

Fig. 5. (a) Stack of images, where k_1, k_2, k_3, k_4 are the distances between $\bar{I}$ and $I_{256\times256}$, $I_{128\times128}$, $I_{512\times512}$ and I_{TDOCT}. (b) Graphical model representing the relationship between the model variables in MAS. Replications are illustrated with plates. Shaded variables are observed.

4.2. Validation on RAPID Test-Retest Dataset

We illustrate the SDOCT synthesis results on a randomly selected B-scan as shown in Fig. 7. Significant synthesis results are observed in all networks; the synthesized images are very similar to real SDOCT images. We do notice, however, differences in the vessel locations and how they appear in the outer RNFL between the different ensemble methods; this is largely due to the fact that the cyclical perceptual loss used in cycleGAN-Perceptual and cycleWGAN-Perceptual is computed in a feature space trained on a very large natural image dataset. By using the VGG loss, the knowledge of human perception that is embedded in VGG network is transferred to OCT image quality evaluation. Nevertheless, the performance of using an ensemble cycleWGAN or an ensemble cycleGAN alone is still highly acceptable despite the fact that these networks solely model the original data distribution from TDOCT to SDOCT. As far as the RNFL is concerned, it can be seen that all networks enhance the layer's visibility compared to the original noisy TDOCT images. In all our experiments, we did not observe deformations nor substantial blurs and distortions in the synthesized images. Moreover, all images have global structure which closely matches that of the target ground-truth images. It can be seen that the proposed ensemble methodologies all lead to recovery of information that cannot be seen in TDOCT images. This is not only observed in the retinal layers, but more importantly, in the vessels. In Table 2, PSNR and SSIM metrics for all ensemble methods are summarized. Although the proposed ensemble cycleGAN with perceptual loss appears to rank first in terms of PSNR and SSIM, we note that these metrics do

Table 1. Limits of agreement and mean difference of all methods versus ground truth *thickness*, and mean *thickness* SD of the first three test-retest visits for both eyes.

Method	GAN	WGAN	WGAN + Perceptual	GAN + Perceptual	TDOCT	SpectralisOCT
95% LOA (μm)	[8.11, -6.73]	[7.97, -6.65]	[7.20, -6.27]	**[6.57, -5.79]**	[26.64, -22.95]	[4.16, -4.04]
Mean Diff. (μm)	0.69	0.71	0.46	**0.39**	1.84	0.06
Mean SD (μm)	1.29	1.27	1.13	**1.06**	2.76	0.77

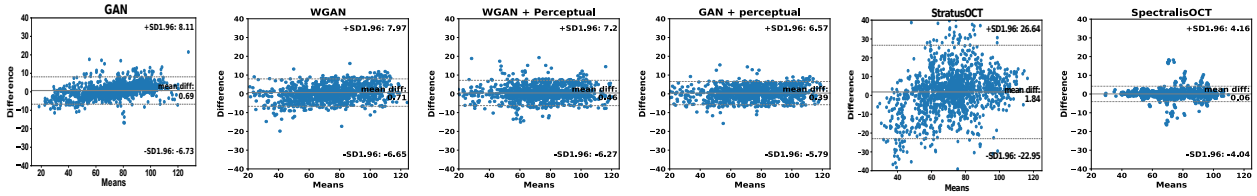
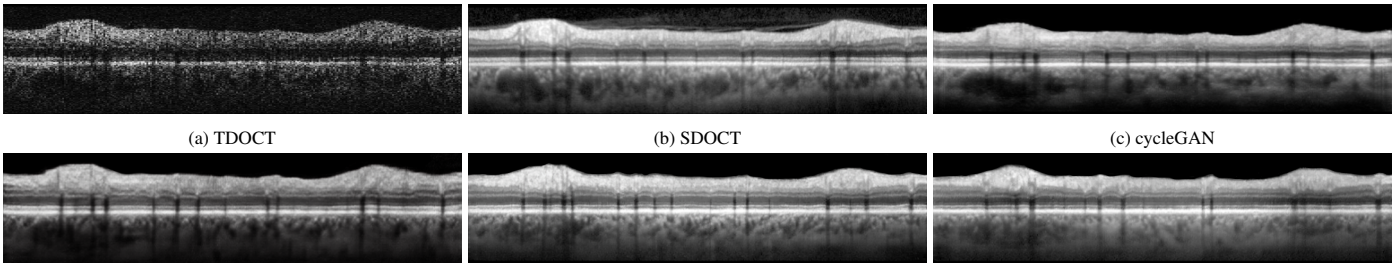


Fig. 6. Bland-Altman plots on the *thickness* agreement between all ensemble methods versus ground truth on RAPID. The proposed ensemble method with s cycle-consistent perceptual loss leads to significantly better agreement (lower spread on the y-axis). *Units in μm*.



(a) TDOCT (b) SDOCT (c) cycleGAN

(d) cycleWGAN (e) cycleWGAN-Perceptual (f) cycleGAN-Perceptual

Fig. 7. OCT synthesis results via fusion of GANs. (a) and (b) illustrate a pair of TDOCT and SDOCT images. (a) - (f) Synthesized SDOCT from (a) using proposed ensemble methodology with different objectives.

Table 2. PSNR and SSIM for each ensemble method.

	PSNR (SD)	SSIM (SD)
GAN	20.9418 (±3.45)	0.7990 (±0.16)
WGAN	20.8327 (±3.66)	0.7874 (±0.13)
WGAN+Perceptual	22.9919 (±3.17)	0.7981 (±0.11)
GAN+Perceptual	**23.4185** (±3.08)	**0.8030** (±0.14)

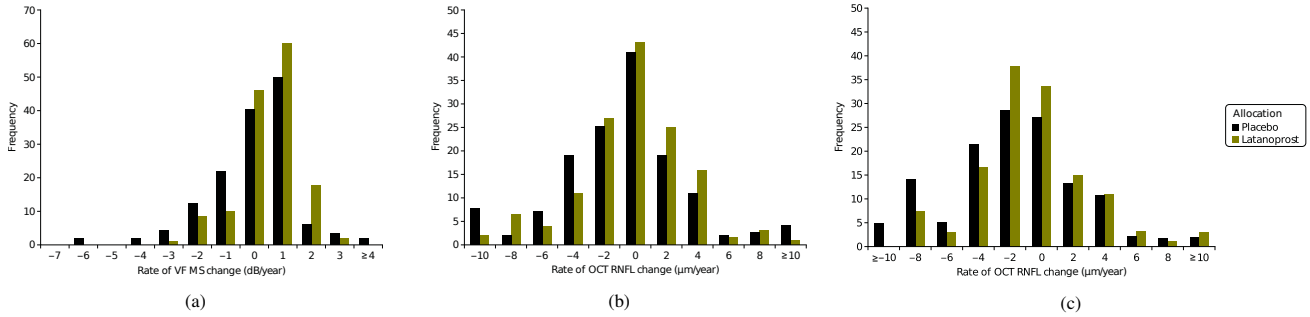
not allow to properly quantify anatomical plausibility of the estimated images (Yang et al., 2018). This indicates that PSNR and SSIM may not be sufficient in evaluating image quality despite our PSNR and SSIM validation results being consistent with literature (Yang et al., 2018; Wolterink et al., 2017). For this reason, we extend validation by re-segmenting the RNFL of all synthesized images and comparing the resulting average RNFL thickness with that of the original ground truth SDOCT. We previously reported results with respect to the ensemble cycleGAN methodology (Lazaridis et al., 2019): we illustrated that individual $GAN_{256\times256}$ yields better scores compared to $GAN_{128\times128}$ and $GAN_{512\times512}$. Also, label fusion, without image stitching, on the average synthesized image outperforms the individual output of GANs, while a further improvement is obtained by integrating image stitching (see Supplementary Material). These results suggest that combining the synthesized images of each individual GAN enables us to take advantage of the strengths of all architectures. Fig. 16 further illustrates the compatibility of the measurements in Bland-Altman plots between the proposed ensemble methodologies and the ground truth SDOCT segmentations. Our approach not only manages to produce a RNFL segmentation closer to the ground truth, but also reduces variability in the measurements in all cases. Table 4 illustrates the 95% limits of agreement (LOA), mean difference and the mean standard deviation (SD) of the difference for three visits across all subjects in RAPID. We observe that the proposed method leads to significantly better agreement and less variability than the TDOCT images. For reference, we present the limits of agreement and SD for repeat SDOCT images to illustrate the best possible performance if synthesized images were identical to SDOCT images.

4.3. Results on the UKGTS

For the UKGTS TDOCT images, all raw intensity OCT data were used, including each of the three individual sequential "fast" circular scans; and images with any signal quality were accepted for analysis. A total of 36,169 (31.6%) TDOCT individual scans failed our RPE-vessel detection detection algorithm and were excluded from further analysis. **In the original investigation (Garway-Heath et al., 2013), a) the averaged**

Table 3. Comparison of rate of RNFL change between our method and Stratus TDOCT in the UKGTS. Significance between group progression rates ($p < 0.05$, Mann–Whitney U test) is indicated with (*). Sample size for 90% power with $p = 0.05$.

Method	TDOCT		Proposed	
	Treatment	Placebo	Treatment	Placebo
Mean (SD) (μm/visit)	0.034 (1.964)	-0.073 (2.066)	-0.069 (1.204)	-0.352 (1.231)
Diff. in mean rate (95% CI)	0.107 (-0.358 to 0.574)		0.282* (0.0003 to 0.5654)	
Sample size	7356		578	



(a) (b) (c)

Fig. 8. (a) Distribution of the rate of VF mean sensitivity (MS) change in decibels per year for the subset of UKGTS participants with OCT images (placebo, n = 131 participants; latanoprost, n = 127 participants). Bottom: Distribution of the rate of OCT RNFL thickness change for the subset of UKGTS participants with OCT images. (b) Original UKGTS TDOCT data (placebo, n = 131 participants; latanoprost, n = 127 participants). (c) Synthesized UKGTS SDOCT data (placebo, n = 131 participants; latanoprost, n = 127 participants).

measurement from three images acquired in quick succession and b) a signal strength of 7 or more were required for structural imaging of participants, resulting in 10,633 (21.3%) TDOCT scans to be excluded. In our study, each of the raw images prior averaging was was used and participants were not excluded because of poor scan quality since those scans could theoretically become scans with good quality after image enhancement. As a result, the total ratio of excluded TDOCT scans in our analysis is similar to Garway-Heath et al. (2013) and does not have any analytical implications as analysis was based on participants who did have five or more time points with both VF tests and OCT images available (Garway-Heath et al., 2015) in both datasets. We applied the proposed ensemble methodology with the newly proposed cycle-consistent perceptual loss function to the TDOCT images available from the UKGTS and subsequently segmented the newly synthesized SDOCT images. We further calculate a rate of change of the segmented RNFL thickness over time and compare the rates of RNFL loss in the two treatment arms of the UKGTS to calculate a sample size for a new trial. Thus, the methodology is tested by quantifying the sample size required for 90% statistical power (Type I error 5%) when using the original TDOCT measurements and the measurements obtained with our method. Table 3 shows the results of our method compared to the original Stratus TDOCT; the mean and the range of RNFL loss rates for TDOCT and synthesized SDOCT images and the respective sample size calculations for a study to distinguish the UKGTS treatments groups are presented. Note that for the sample size calculation we assume that all patients would have usable SDOCT images. We appreciate a statistically significant improvement in the separation between treatment and placebo groups ($p = 0.0017$), leading to a markedly lower sample size in power analysis. These results are a further improvement in regards to those we previously reported (Lazaridis et al., 2019) using an ensemble of cycleGANs. Fig.8 illustrates the VF mean sensitivity (MS) change in decibels per year and the distribution of rate of RNFL thickness change for the subset of UKGTS participants with usable OCT images. Fig.8b is generated from the original TDOCT whereas Fig.8c from the synthesized SDOCT data. The rate of loss was taken from the eye with the worse baseline MD. Our method allowed the detection faster rates of deterioration in the placebo than the latanoprost group, which the original TDOCT was unable to do. This effect can be qualitatively appreciated from the shift towards the left of the placebo RNFL rate histogram (Fig.8). For the original TDOCT UKGTS data, the difference in distribution of slopes was not statistically significant (Mann-Whitney U Test, p = 0.18). For the synthesized SDOCT UKGTS data, the difference was statistically significant (Mann-Whitney U Test, p = 0.04).

5. Discussion and Conclusion

We previously reported (Garway-Heath et al., 2017) that the rate of RNFL loss from TDOCT measurements was not able to distinguish the treatment arms in the UKGTS. In this work, we demonstrate that the proposed ensemble methodology with further adoption of the proposed cycle-consistent perceptual loss applied to TDOCT images significantly improves the agreement of segmented RNFL thickness measurements with SDOCT measurements and significantly reduces the test-retest variability. When the rate of RNFL loss in the UKGTS data set is calculated from the synthesized SDOCT images, the difference in RNFL slopes is able to distinguish the treatment groups.

The analysis of the capability of TDOCT images to distinguish the UKGTS treatment arms shows that, although the rate of RNFL loss was faster in the placebo-treated eyes, the difference from the latanoprost-treated eyes did not reach statistical significance. In contrast, the same analysis with the synthesized SDOCT images demonstrated a statistically significant difference between treatment and placebo progression rates. The difference between treatment groups in the rate of RNFL thinning was similar to the difference between groups for the rate of VF MD deterioration. Moreover, there is an appreciable reduction in the sample size required to distinguish the treatment arms in the UKGTS if SDOCT RNFL thickness were to be the primary outcome, compared to TDOCT RNFL thickness. Therefore, we have shown that the rate of RNFL thinning responds in a similar manner to treatment as does the rate of VF deterioration. Our contributions extend current literature on image synthesis, as we propose to learn the image distribution by probabilistic fusion of several generative models and using a novel cycle-consistent loss. Our approach is based on semi-automated segmentation of synthesized SDOCT images and image stitching is shown to further improve statistical separation between treatment groups (see Supplementary Material). The proposed methodology appears robust and flexible both in terms of architecture and label fusion. Since the training dataset is large and of high resolution, training of each individual GAN model is computationally expensive. This however a negligible problem in practice as the model can be run offline. Future work will focus on regularized attention schemes to improve conditioning on the RNFL.

Acknowledgment

The research was supported by the EPSRC (CDT in Medical Imaging, EP/L016478/1), the International Glaucoma Association, Santen Pharmaceutical Co., Ltd., the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. ML is supported by the French government, through the 3IA Côte d'Azur Investments in the Future project (ANR-19-P3IA-0002) managed by the National Research Agency. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein Generative Adversarial Networks, in: 34th International Conference on Machine Learning, pp. 214–223.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95 – 113. doi:https://doi.org/10.1016/j.neuroimage.2007.07.007.

Asrani, S., Essaid, L., Alder, B.D., Santiago-Turla, C., 2014. Artifacts in Spectral-Domain Optical Coherence Tomography Measurements in Glaucoma. JAMA Ophthalmology 132, 396–402.

Atzeni, A., Jansen, M., Ourselin, S., Iglesias, J.E., 2018. A Probabilistic Model Combining Deep Learning and Multi-atlas Segmentation for Semi-automated Labelling of Histology, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 219–227.

Bashkansky, M., Reintjes, J., 2000. Statistics and Reduction of Speckle in Optical Coherence Tomography. Opt. Lett. 25, 545–547.

Ben-Cohen, A., Klang, E., Raskin, S.P., Amitai, M.M., Greenspan, H., 2017. Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results, in: Simulation and Synthesis in Medical Imaging, pp. 49––57.

Bernardes, R., Maduro, C., Serranho, P., Araújo, A., Barbeiro, S., Cunha-Vaz, J., 2010. Improved Adaptive Complex Diffusion Despeckling Filter. Opt. Express 18, 24048–24059.

Bunce, C., Wormald, R., 2008. Causes of Blind Certifications in England and Wales: April 1999–March 2000. Eye 22, 905–911.

Chang, S.G., Bin Yu, Vetterli, M., 2000. Adaptive Wavelet Thresholding for Image Denoising and Compression. IEEE Transactions on Image Processing 9, 1532–1546. doi:10.1109/83.862633.

Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., et al., 2017. Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. IEEE Transactions on Medical Imaging 36, 2524–2535. doi:10.1109/TMI.2017.2715284.

Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G., 2017. Low-Dose CT via Convolutional Neural Network. Biomedical optics express 8, 679–694. doi:10.1364/BOE.8.000679.

Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K., 2006. Image Denoising with Block-matching and 3D Filtering, in: Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, pp. 354–365.

Desjardins, A.E., Vakoc, B.J., Oh, W.Y., Motaghiannezam, S.M., Tearney, G.J., Bouma, B.E., 2007. Angle-Resolved Optical Coherence Tomography with Sequential Angular Selectivity for Speckle Reduction. Optics Express 15, 6200–6209.

Devalla, S.K., Subramanian, G., Pham, T.H., Wang, X., Perera, S., Tun, T.A., et al., 2019. A Deep Learning Approach to Denoise Optical Coherence Tomography Images of the Optic Nerve Head. Scientific Reports 9, 14454. doi:10.1038/s41598-019-51062-7.

Dong, C., Loy, C.C., He, K., Tang, X., 2014. Image Super-Resolution Using Deep Convolutional Networks. arXiv:1501.00092.

Du, X., Gong, L., et al., 2017. Non-rigid Registration of Retinal OCT Images Using Conditional Correlation Ratio, in: Fetal, Infant and Ophthalmic Medical Image Analysis, pp. 159–167.

Du, Y., Liu, G., Feng, G., Chen, Z., 2014. Speckle reduction in optical coherence tomography images based on wave atoms. Journal of Biomedical Optics 19, 1 – 7. doi:10.1117/1.JBO.19.5.056009.

Fei, X., Zhao, J., Zhao, H., Yun, D., Zhang, Y., 2017. Deblurring Adaptive Optics Retinal Images Using Deep Convolutional Neural Networks. Biomedical optics express 8, 5675–5687. doi:10.1364/BOE.8.005675.

Garway-Heath, D.F., Crabb, D.P., Bunce, C., Lascaratos, G., Amalfitano, F., Anand, N., et al., 2015. Latanoprost for Open-Angle Glaucoma (UKGTS): A Randomised, Multicentre, Placebo-Controlled Trial. The Lancet 385, 1295–1304.

Garway-Heath, D.F., Lascaratos, G., Bunce, C., Crabb, D.P., Russell, R.A., Shah, A., 2013. The united kingdom glaucoma treatment study: A multicenter, randomized, placebo-controlled clinical trial: Design and methodology. Ophthalmology 120, 68 – 76. doi:https://doi.org/10.1016/j.ophtha.2012.07.028.

Garway-Heath, D.F., Quartilho, A., Prah, P., Crabb, D.P., Cheng, Q., Haogang, Z., 2017. Evaluation of Visual Field and Imaging Outcomes for Glaucoma Clinical Trials (An American Ophthalomological Society Thesis). Transactions of the American Ophthalmological Society 115, T4.

Gondara, L., 2016. Medical image denoising using convolutional denoising autoencoders, in: 2016 IEEE 16th International Conference on Data Mining Workshops, pp. 241–246.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., et al., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems 27, pp. 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved Training of Wasserstein GANs. CoRR abs/1704.00028. arXiv:1704.00028.

Halupka, K.J., Antony, B.J., Lee, M.H., Lucy, K.A., Rai, R.S., Ishikawa, H., et al., 2018. Retinal Optical Coherence Tomography Image Enhancement via Deep Learning. Biomed. Opt. Express 9, 6205–6221.

Hardin, J.S., Taibbi, G., Nelson, S.C., Chao, D., Vizzeri, G., 2015. Factors Affecting Cirrus-HD OCT Optic Disc Scan Quality: A Review with Case Examples. Journal of Ophthalmology 2015, 746150. doi:10.1155/2015/746150.

Hood, D.C., Raza, A.S., de Moraes, C.G.V., Liebmann, J.M., Ritch, R., 2013. Glaucomatous Damage of the Macula. Progress in Retinal and Eye Research 32, 1–21.

Huang, Y., Lu, Z., Shao, Z., Ran, M., Zhou, J., Fang, L., Zhang, Y., 2019. Simultaneous Denoising and Super-Resolution of Optical Coherence Tomography Images Based on Generative Adversarial Network. Opt. Express 27, 12289–12307.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-Image Translation with Conditional Adversarial Networks, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976.

Johnson, J., Alahi, A., Li, F., 2016. Perceptual losses For Real-Time Style Transfer and Super-Resolution. CoRR abs/1603.08155. arXiv:1603.08155.

Krupin, T., Liebmann, J.M., Greenfield, D.S., Ritch, R., Gardiner, S., 2011. A Randomized Trial of Brimonidine Versus Timolol in Preserving Visual Function: Results From the Low-pressure Glaucoma Treatment Study. American Journal of Ophthalmology 151, 671–681. doi:10.1016/j.ajo.2010.09.026.

Lazaridis, G., Lorenzi, M., Ourselin, S., Garway-Heath, D., 2019. Enhancing OCT Signal by Fusion of GANs: Improving Statistical Power of Glaucoma Clinical Trials, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 3–11.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep Learning. Nature 521, 436.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A.P., Tejani, A., et al., 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. CoRR abs/1609.04802. arXiv:1609.04802.

Li, M., Shen, S., Gao, W., Hsu, W., Cong, J., 2018. Computed Tomography Image Enhancement Using 3d Convolutional Neural Network. CoRR abs/1807.06821. arXiv:1807.06821.

Mayer, M.A., Borsdorf, A., Wagner, M., Hornegger, J., Mardin, C.Y., Tornow, R.P., 2012. Wavelet Denoising of Multiframe Optical Coherence Tomography Data. Biomed. Opt. Express 3, 572–589.

Mayer, M.A., Hornegger, J., Mardin, C.Y., Tornow, R.P., 2010. Retinal Nerve Fiber Layer Segmentation on FD-OCT Scans of Normal Subjects and Glaucoma Patients. Biomed. Opt. Express 1, 1358–1383. doi:10.1364/BOE.1.001358.

Musch, D.C., Gillespie, B.W., Lichter, P.R., Niziol, L.M., Janz, N.K., 2009. Visual Field Progression in the Collaborative Initial Glaucoma Treatment Study: The Impact of Treatment and Other Baseline Factors. Ophthalmology 116, 200–207. doi:10.1016/j.ophtha.2008.08.051.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., et al., 2017. Medical Iage Synthesis with Context-Aware Generative Adversarial Networks, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, pp. 417–425.

Pircher, M., Götzinger, E., Leitgeb, R.A., Fercher, A.F., Hitzenberger, C.K., 2003. Speckle Reduction in Optical Coherence Tomography by Frequency Compounding. Journal of Biomedical Optics 8, 565 – 569.

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R., 2018. Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning. Nature Biomedical Engineering 2, 158–164.

Rabbani, H., Sonka, M., Abramoff, M.D., 2013. Optical Coherence Tomography Noise Reduction Using Anisotropic Local Bivariate Gaussian Mixture Prior in 3D Complex Wavelet Domain. International Journal of Biomedical Imaging 2013, 417–491.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al., 2014. ImageNet Large Scale Visual Recognition Challenge. CoRR abs/1409.0575. arXiv:1409.0575.

Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P., 2010. A Generative Model for Image Segmentation Based on Label Fusion. IEEE Trans. Med. Imaging 29, 1714–1729.

Shi, J., Liu, Q., Wang, C., Zhang, Q., Ying, S., Xu, H., 2018. Super-Resolution Reconstruction of MR Image With a Novel Residual Learning Network Algorithm. Physics in Medicine & Biology 63, 085011.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition, in: International Conference on Learning Representations.

van Velthoven, M.E., Faber, D.J., Verbraak, F.D., van Leeuwen, T.G., de Smet, M.D., 2007. Recent Developments in Optical Coherence Tomography for Imaging the Retina. Progress in Retinal and Eye Research 26, 57 – 77.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I., 2017. Deep MR to CT Synthesis Using Unpaired Data, in: Simulation and Synthesis in Medical Imaging, pp. 14–23.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2017. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. IEEE Transactions on Medical Imaging 36, 2536–2545.

Wormald, R., Virgili, G., Azuara-Blanco, A., 2020. Systematic Reviews and Randomised Controlled Trials on Open Angle Glaucoma. Eye 34, 161–167.

Wu, W., Tan, O., Pappuru, R.R., Duan, H., Huang, D., 2013. Assessment of Frame-Averaging Algorithms in OCT Image Analysis. Ophthalmic Surgery, Lasers & Imaging Retina 44, 168–175.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., et al., 2018. Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. IEEE Transactions on Medical Imaging 37, 1348–1357.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in: IEEE International Conference on Computer Vision, pp. 2242–2251.

Supplementary Material

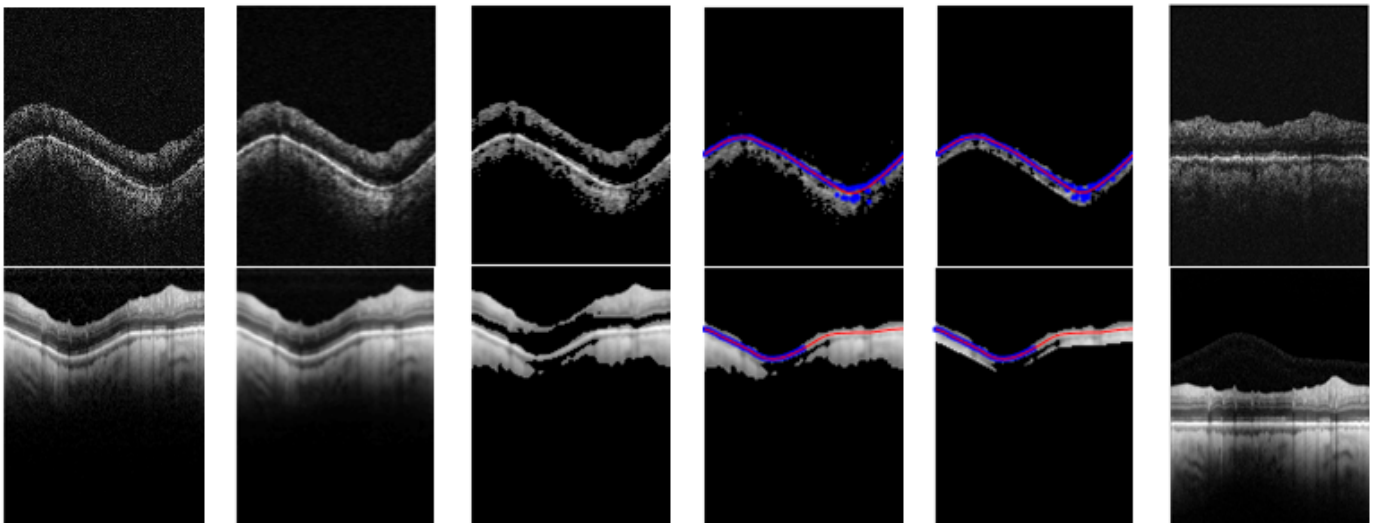


Fig. 9. Flattening process for TDOCT (top row) and SDOCT (bottom row). Interpretation from left (input image) to right (output image). First column: Input TDOCT and SDOCT images, respectively. Second column: Filtered images after application of diffusion filter to denoise images. Third column: The brightest pixel in each column is assigned as an estimate of the retinal pigment epithelium (RPE). Using this initial estimate, pixels lying in columns that present a significantly lower signal-to-noise ratio (SNR) than RPE pixels are removed. Forth column: Discontinuities greater than 50 pixels which are often associated with the nerve fibre layer (NFL) are identified and removed. The remaining RPE points are fit with a second order polynomial. Fifth column: Remaining layer pixels are further identified and removed heuristically based on distances from polynomial RPE fit, ensuring polynomial continuity. Sixth column: Each column is shifted up or down in order to force the RPE points to lie on a flat line, with fixed y-axis offset, guided by the polynomial RPE fit.

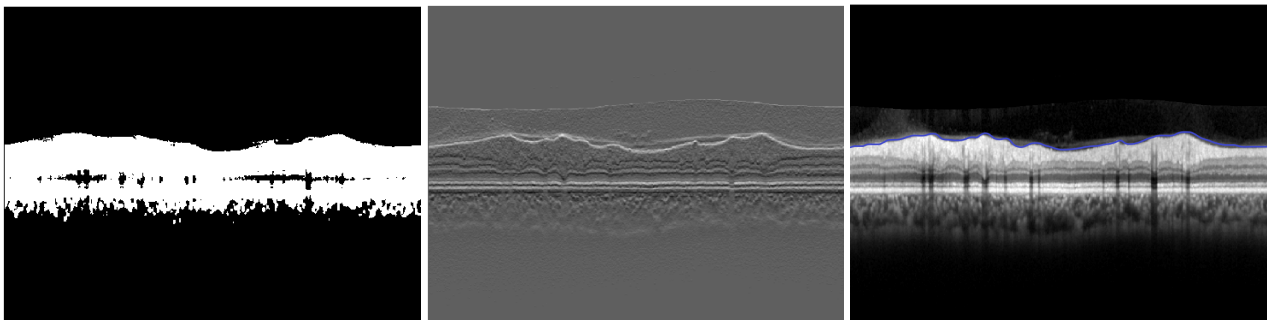


Fig. 10. ILM contour segmentation in a flattened image. Both the thresholded image (left) and the gradient image (right) yield internal limiting membrane (ILM) contour pixel candidates. Final contour segmentation (bottom) is obtained by local averaging of proposed pixel candidates. Left: Thresholded image using Otsu's method. Right: Gradient image. Bottom: SDOCT image with obtained ILM contour.

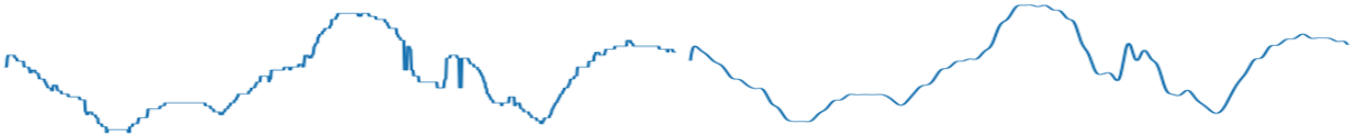


Fig. 11. ILM contour interpolation using Gaussian Process(GP) regression. Left: Before GP interpolation. Right: After GP interpolation.

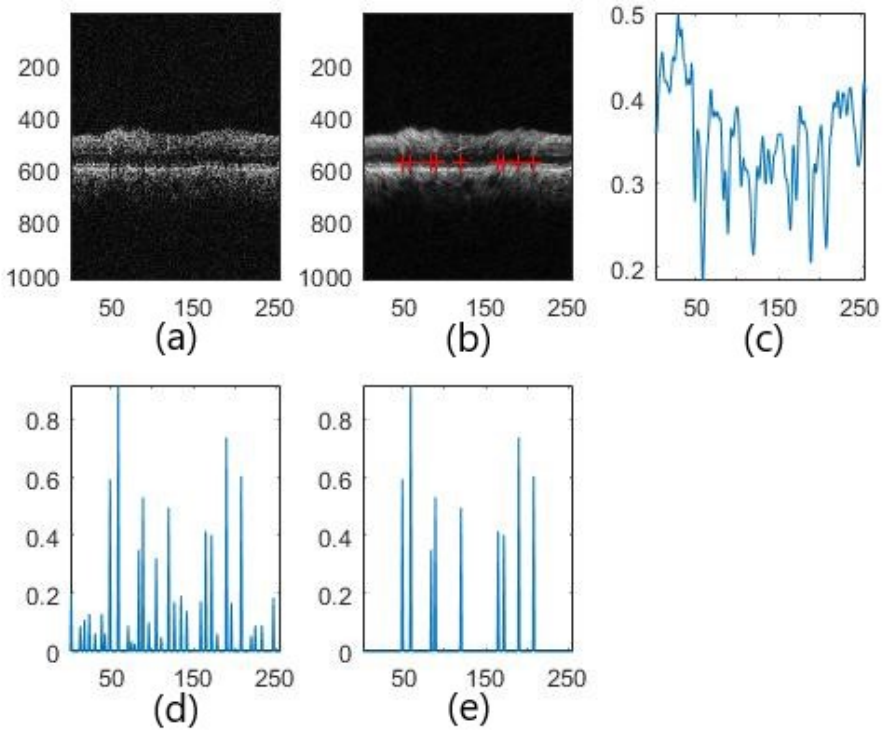


Fig. 12. Vessel profile and vessel detection. (a) TDOCT flattened image. (b) TDOCT flattened image with detected vessels marked with red crosses. (c) Vessel profile given by the average retinal pigment epithelium (RPE) pixel intensity. (d) Local minima on the vessel profile. (e) Local minima on the vessel profile after removal of spurious peaks, thresholded based on standard deviation. These minima indicate vessel locations.

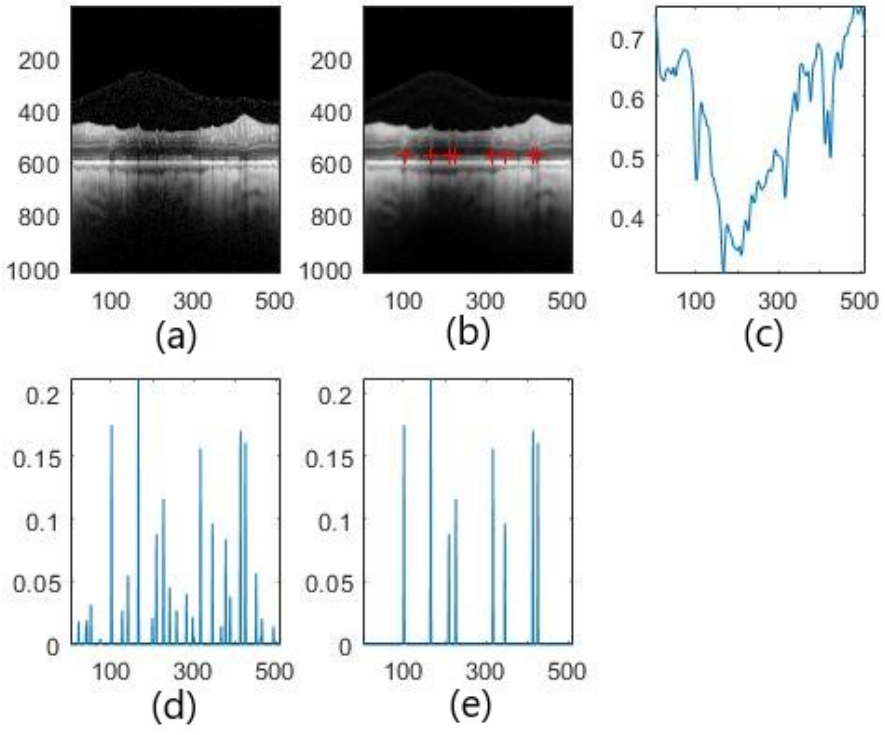


Fig. 13. Vessel profile and vessel detection. (a) SDOCT flattened image. (b) SDOCT flattened image with detected vessels marked with red crosses. (c) Vessel profile given by the average retinal pigment epithelium (RPE) pixel intensity. (d) Local minima on the vessel profile. (e) Local minima on the vessel profile after removal of spurious peaks, thresholded based on standard deviation. These minima indicate vessel locations.

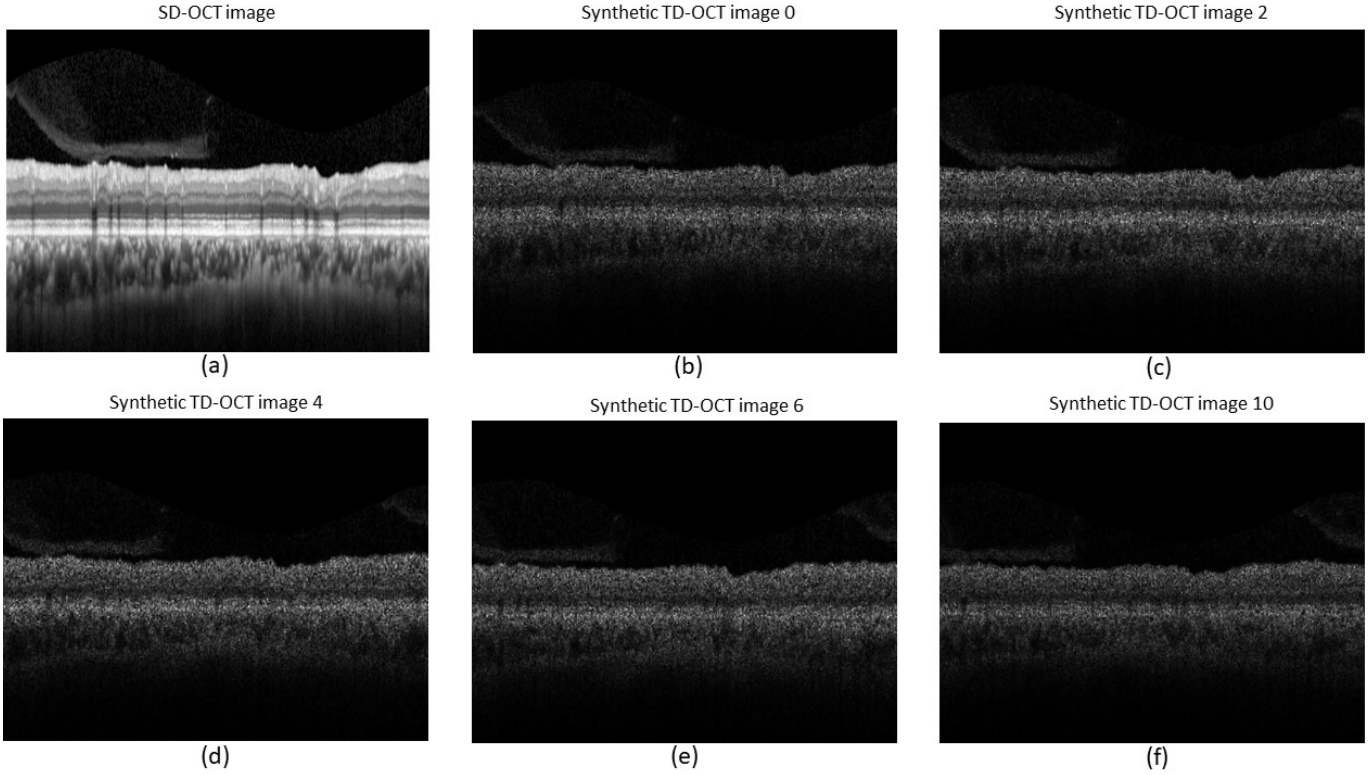


Fig. 14. Illustration of an image subgroup of one of ten groups of the synthetic dataset. Each group contains one SDOCT image and ten synthetic TDOCT images with spatial variability generated from each group's SDOCT image. Each synthetic TDOCT image is deformed, except one, which is identified as the best match by our pairing method. SDOCT images are downsampled, contaminated with speckle noise and spatially deformed using a random stationary velocity field in order to generate synthetic TDOCT images with spatial variability (Ashburner, 2007).

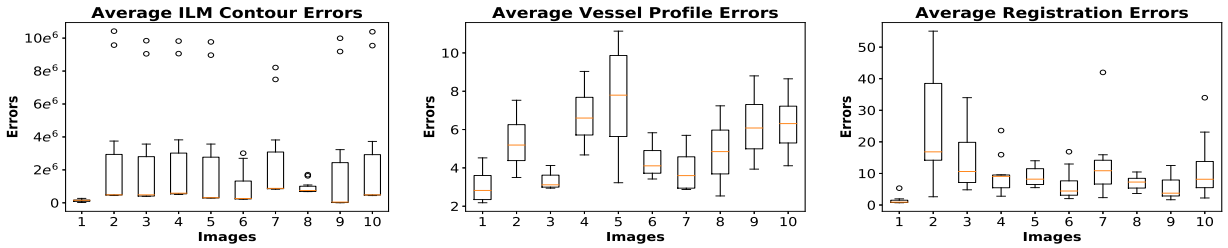


Fig. 15. Averaged group errors of each method. The synthetic TDOCT image with no deformations, i.e. Image 1, is correctly identified as the best pair for each SDOCT which generate it (in each group). Average ILM contour errors and average registration errors yield 100% sensitivity in identifying the correct synthetic TDOCT pair for each SDOCT, whereas the average vessel profile errors yield 89% sensitivity for the same identification. Validation was achieved with 100% pairing sensitivity on synthetic data with voting using all three methods.

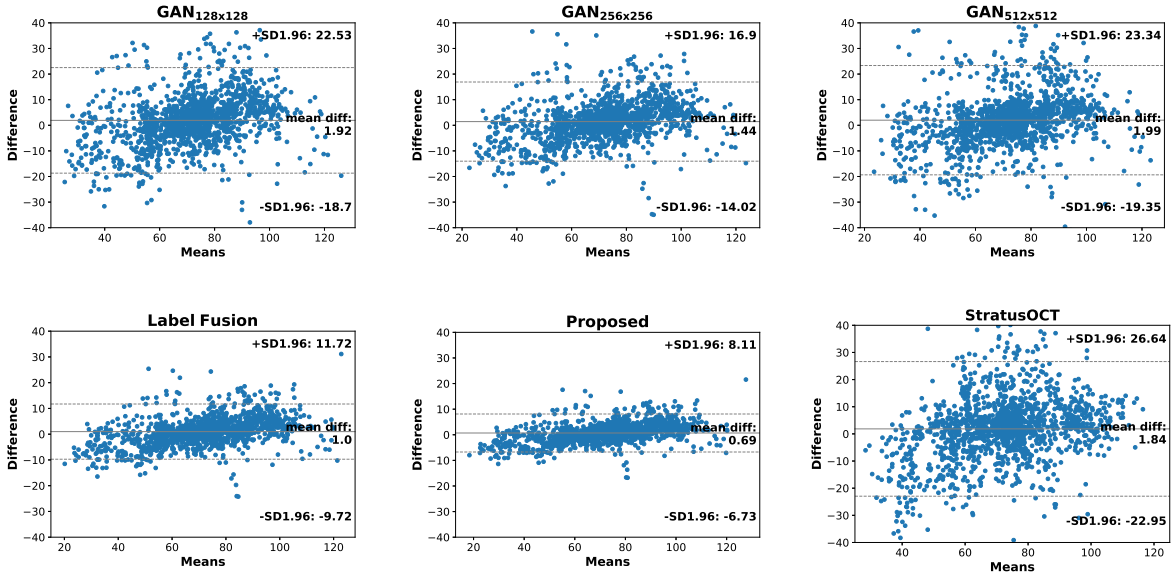


Fig. 16. Bland-Altman plots on the agreement between all methods versus ground truth on RAPID. The proposed method leads to significantly better agreement.

Table 4. Limits of agreement, mean difference, correlation of all methods versus ground truth, and mean SD of the first three test-retest visits for both eyes.

Method	GAN			Label Fusion		StratusOCT
	128x128	256x256	512x512	Direct	Proposed	
95% LOA	[22.53, -18.7]	[16.9, -14.2]	[23.34, -19.35]	[11.72, -9.72]	[**8.11**, **-6.73**]	[26.64, -22.95]
Mean Diff.	1.92	1.44	1.99	1.00	**0.69**	1.84
Pearson r	0.79	0.85	0.71	0.89	**0.92**	0.76
Mean SD	2.27	1.87	3.01	1.33	**1.29**	2.67