Research Article | Vol. 11, No. 1 / 1 January 2020 / Biomedical Optics Express 346
Biomedical Optics EXPRESS

Reducing image variability across OCT devices with unsupervised unpaired learning for improved segmentation of retina

DAVID ROMO-BUCHELI,[1,2] PHILIPP SEEBÖCK,[1,2] JOSÉ IGNACIO ORLANDO,[1] BIANCA S. GERENDAS,[1] SEBASTIAN M. WALDSTEIN,[1] URSULA SCHMIDT-ERFURTH,[1,*] AND HRVOJE BOGUNOVIĆ[1]

[1] *Christian Doppler Laboratory for Ophthalmic Image Analysis (OPTIMA), Department of Ophthalmology, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[2] *Contributed equally*
[*] *ursula.schmidt-erfurth@meduniwien.ac.at*

Abstract: Diagnosis and treatment in ophthalmology depend on modern retinal imaging by optical coherence tomography (OCT). The recent staggering results of machine learning in medical imaging have inspired the development of automated segmentation methods to identify and quantify pathological features in OCT scans. These models need to be sensitive to image features defining patterns of interest, while remaining robust to differences in imaging protocols. A dominant factor for such image differences is the type of OCT acquisition device. In this paper, we analyze the ability of recently developed unsupervised unpaired image translations based on cycle consistency losses (cycleGANs) to deal with image variability across different OCT devices (Spectralis and Cirrus). This evaluation was performed on two clinically relevant segmentation tasks in retinal OCT imaging: fluid and photoreceptor layer segmentation. Additionally, a visual Turing test designed to assess the quality of the learned translation models was carried out by a group of 18 participants with different background expertise. Results show that the learned translation models improve the generalization ability of segmentation models to other OCT-vendors/domains not seen during training. Moreover, relationships between model hyper-parameters and the realism as well as the morphological consistency of the generated images could be identified.

1. Introduction

Optical coherence tomography (OCT) is a non-invasive technique that provides 3D volumes of the retina at a micrometric resolution [1]. Each OCT volume comprises multiple cross-sectional 2D images, or B-scans, each of them composed of 1D columns, or A-scans. By means of the OCT imaging modality, clinicians are allowed to perform detailed ophthalmic examinations for disease diagnosis, assessment and treatment planning. Standard treatment and diagnosis protocols nowadays rely heavily on B-scan images to inform clinical decisions [2].

Several imaging tasks such as segmentation of anatomical structures or classification of pathological cases are successfully addressed by automated image analysis methods. These approaches are commonly based on machine learning (ML) models, which are trained on manually annotated datasets in a supervised setting. Among the existing ML tools, deep learning (DL) techniques based on convolutional neural networks have been remarkably successful in different domains [3], including automated OCT image analysis [4].

However, these models are usually prone to errors when deployed on real clinical scenarios. This is partially due to the differences in the data distributions of the training sets and the real world data sets. This phenomenon is known as *covariate shift*, as formally defined in [5]. An important example of this phenomenon is observed in practice when the training and deployment

#379978 https://doi.org/10.1364/BOE.379978
Journal © 2020 Received 10 Oct 2019; revised 10 Dec 2019; accepted 11 Dec 2019; published 20 Dec 2019

Biomedical Optics EXPRESS

sets come from different acquisition devices. In the daily ophthalmological routine, several OCT devices from different vendors with varying acquisition protocols are used. As a consequence, regardless of the scanned area, the resulting OCT volumes present different image characteristics and patterns (Fig. 2). This covariate shift has been observed to cause remarkable drops in the performance of different DL models for retinal fluid and layer segmentation [6–8].

Related Work Image translation has been applied in cross-modality medical image analysis. Early approaches, requiring paired training data across the datasets, are frequently based on patch-based learning [9] and convolutional neural networks [10]. Recently, strategies based on generative adversarial networks (GAN) [11,12] allow a more flexible model training compared to the former since the training examples across multiple modalities do not need to be paired.

Another application for image translation in medical imaging is reducing the image differences within a medical image modality. Lately, the covariate shift phenomenon due to different acquisition devices has received a lot of attention across several imaging modalities in medical image analysis. For instance, [13] proposed a siamese neural network that generates device independent features and improves the generalization ability of brain tissue segmentation models in magnetic resonance imaging (MRI). The whole process requires a limited amount of fully annotated MRI images describing the tissue class in the target device. A cycle-consistency based loss was used in [14] to learn a stain-translating model for histology images across two datasets from different centers. The authors report that the stain-translation model improved the segmentation model's performance in renal tissue slides. Cheng *et al.* [15] proposed a domain adaptation method for chest X-ray imaging based on unsupervised cycle consistency adversarial and semantic aware loss, improving the performance of lung segmentation models on different public datasets. In general, several of these recently proposed techniques are based on cycle-consistency adversarial losses [12,14–20] which were firstly introduced in the cycleGAN unpaired unsupervised learning algorithm [21]. Unpaired unsupervised learning techniques for alleviating the performance drop of retinal and choroidal layer segmentation models across OCT acquisition devices is also being proposed in non-peer reviewed pre-print [22].

Currently, the most straightforward approach to deal with image variability across OCT devices is to train vendor-specific models for each task. An example of this approach was shown in De Fauw *et al.* [8], where a two-stage DL approach was used to diagnose retinal disease. The first stage generates segmentation maps associated with retinal morphology and other OCT imaging related concepts. Then, the second stage generates referral predictions based on the segmented feature maps and is consequently device independent. However, time-consuming manual annotations needed to be collected for each vendor to adequately (re-)train the retinal segmentation models of the first stage, and hence assure device independence of the second stage. In contrast, using unpaired unsupervised algorithms to adjust the image properties of the B-scans would not require additional manual annotations. This remarkable advantage would likely have an important impact when OCT low-cost devices [23] become available, as they could benefit from the already pretrained state-of-the-art retinal segmentation DL models.

Contribution In this work, we evaluated the ability of the translations obtained by the cycleGAN algorithm [21] to reduce the covariate shift in DL segmentation models across different OCT acquisition devices. This paper is an extension of our previous work, where an unpaired unsupervised algorithm was evaluated on retinal fluid segmentation performance [24]. In that work, we observed that covariate shift in OCT imaging may be successfully alleviated for fluid segmentation task by means of cycleGANs. Such an approach allows to translate images acquired using an OCT device *B* to resemble images produced by a different device *A* (Fig. 1). As a result, the new image can then be processed using a segmentation model trained on images from *A* without a significant drop in performance. In this paper, we extended the evaluation by analyzing the applicability of cycleGANs also in the context of an anatomical layer segmentation task. Furthermore, a visual Turing test was performed by 18 participants with

Biomedical Optics EXPRESS

different background expertise to comprehensively study the appearance of the resulting synthetic images. To the best of our knowledge, this is the first work that comprehensively explores cycleGAN algorithms in OCT images to reduce image variability across OCT acquisition devices.

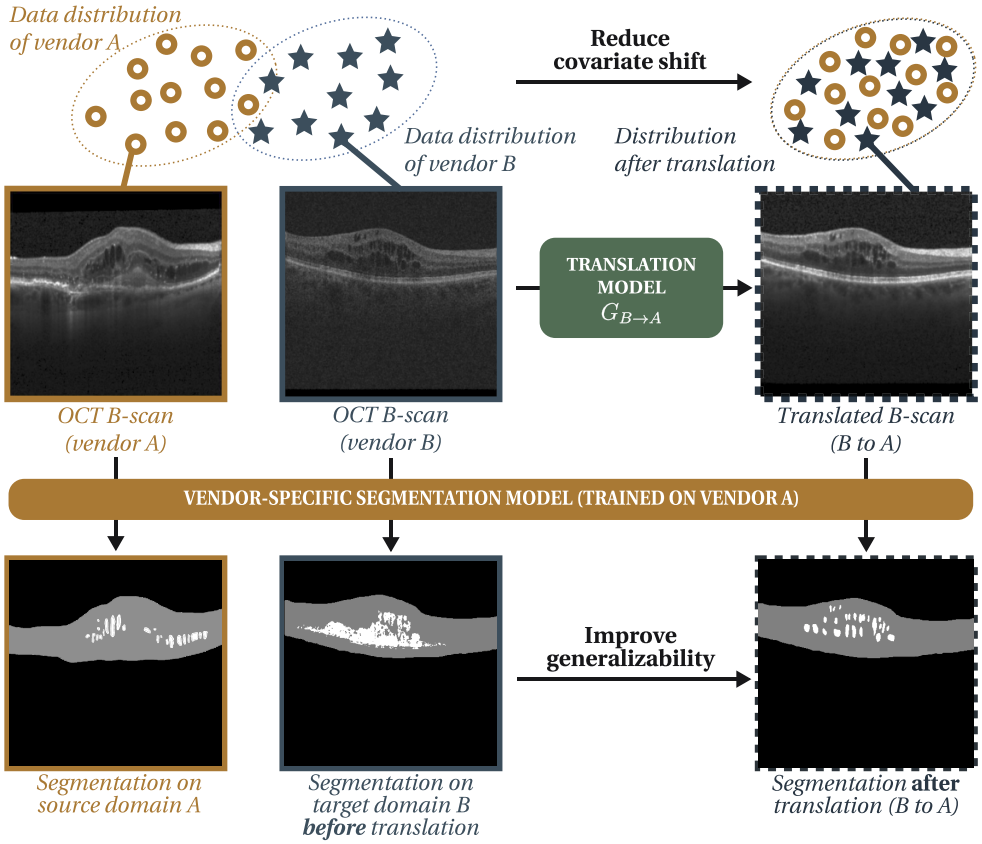


Fig. 1. Schematic overview demonstrating how unpaired translation models reduce performance drop due to covariate shift in models trained with different OCT acquisition vendors.

The main contributions of this paper can be summarised by the following three points:

- We empirically demonstrated that unsupervised unpaired algorithms are able to reduce the covariate shift between different OCT acquisition devices. Segmentation performance in particular cases of retinal layer and fluid segmentation showed significant quantitative improvements. These tasks are clinically relevant, and possibly the most active research topics in automated OCT image segmentation.
- We performed an extensive analysis of the effect of the unsupervised unpaired algorithm training patch size in the cross-vendor segmentation tasks. Particularly, we found that optimal patch size might depend on each task, with larger patch sizes more suitable for tasks in which larger structures need to be segmented.
- We performed a visual Turing test with a large set of participants (n=18) with different background expertise. The results show that the image patch size at training stage is an important parameter, controlling the trade-off between “realism” and ”morphological fidelity” of the images generated by the translation model. The test demonstrated that

Biomedical Optics EXPRESS

translation models trained with larger patch sizes induce a more realistic appearance but were prone to generate artifacts distorting morphological features in the retina. On the other hand, models trained with smaller patch sizes rarely present such distortions, but are easily identified as "fake" images.

This paper is organized as follows. First, Section 2 describes the methods used in this work: the deep learning segmentation model, the cycleGAN and baseline translation algorithms. In Section 3 we describe the used datasets, training details of the models and the experimental evaluation setup. In Section 4, qualitative and quantitative results of the segmentation tasks and the visual Turing test are given. Finally, a discussion of these results and conclusions of the work are presented in Section 5 and Section 6.

2. Methods

This section presents a summary of the methodologies used in our study. In particular, Section 2.1 describes the unsupervised unpaired translation algorithm based on cycleGANs. Subsequently, Section 2.2 presents a brief description of alternative filtering-based translations that are used as baselines to compare with our proposed approach. Finally, Section 2.3 formalizes the segmentation tasks and describes the neural network architectures used for fluid and photoreceptor layer segmentation.

2.1. Unsupervised unpaired translation using cycleGANs

We propose to apply unsupervised unpaired generative adversarial networks to automatically translate OCT images from one vendor to another. Formally, let A and B be two different image domains, where any image $a \in A$ has different visual characteristics compared to any other image $b \in B$. Cycle generative adversarial networks (cycleGANs) allow to learn a suitable translation function between the image domains A and B without requiring paired samples, allowing to tackle an ill-posed translation problem (i.e. *unpaired* translation) [21].

A cycleGAN uses two generator/discriminator pairs ($G_{A\to B}/D_B$, $G_{B\to A}/D_A$), which are implemented as deep neural networks. $G_{A\to B}$ is supplied with an image from the source domain A and translates it into the target domain B. Analogously, $G_{B\to A}$ translates the image back from the target B to the source domain A. D_A (D_B) is trained to distinguish between real samples from the source (target) domain and the translated images, being associated to the likelihood that a certain image is sampled from domain A (B). The objective for the mapping function G_{A-B} can be expressed as:

$$\begin{aligned}\arg \min_{G_{A\to B}, G_{B\to A}} \max_{D_A, D_B} \mathcal{L}(G_{A\to B}, G_{B\to A}, D_A, D_B) = \mathcal{L}_{\text{GAN}}(G_{A\to B}, D_B) \\ + \mathcal{L}_{\text{GAN}}(G_{B\to A}, D_A) \\ + \eta_1 L_{\text{cyc}} \\ + \eta_2 L_{\text{identity}}(G_{A\to B}, G_{B\to A}),\end{aligned} \tag{1}$$

where the last two terms are the cycle consistency L_{cyc} and the identity mapping loss L_{identity}, as defined in [21], with weights η_1 and η_2 that control their relevance in the overall loss. Both L_{cyc} and L_{identity} serve as useful regularization terms improving the obtained translation functions $G_{A\to B}$ and $G_{B\to A}$. Minimizing L_{cyc} constrains the mentioned translations to be reversible, so successive application of $G_{A\to B}$ and $G_{B\to A}$ ($G_{B\to A}$ and $G_{A\to B}$) to an image from domain A (B) generates an image that matches the original source image. In contrast, minimizing L_{identity} regularizes the resulting translation $G_{A\to B}$ ($G_{B\to A}$) to generate a target image that is close to the source image, if the source image already have a target B (A) domain appearance. Intuitively speaking, this means that the translation $G_{A\to B}$ ($G_{B\to A}$) is constrained so that changes of source images that are not needed are avoided.

Biomedical Optics EXPRESS

The first two terms L_{GAN} correspond to the least square generative adversarial loss terms [25]:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G_{A\to B}, D_B) &= \mathcal{L}_{\text{GAN}}(D_B) + \mathcal{L}_{\text{GAN}}(G_{A\to B}), \text{ with}\\ \mathcal{L}_{\text{GAN}}(D_B) &= \frac{1}{2}\mathbb{E}_{b\sim B}[(D_B(b)-1)^2] + \frac{1}{2}\mathbb{E}_{a\sim A}[D_B(G_{A\to B}(a))^2]\mathcal{L}_{\text{GAN}}(G_{A\to B})\\ &= \frac{1}{2}\mathbb{E}_{a\sim A}[(D_B(G_{A\to B}(a))-1)^2]\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{GAN}}(G_{B\to A}, D_A)$, is defined analogously. Both the generator and discriminator were implemented as deep neural networks following the ResNet based architecture presented in [21].

2.2. Baseline translation algorithms

Domain translation across images of different acquisition devices is not a common operation in OCT imaging. As such, no baselines are currently available for comparison purposes. However, some preprocessing pipelines including denoising and histogram matching algorithms have been used to standardize the appearance of OCT B-scans [26]. In this work, we followed the methodology applied in [27] to define two suitable translations to approximate the appearance of Spectralis OCTs from Cirrus B-scans. The first translation strategy (T_1) consist of an initial B-scan level median filtering operation (with 3×3 kernel size) followed by a second median filtering operation across neighboring B-scans, with a $1 \times 1 \times 3$ kernel size. The second translation (T_2) consists of an initial histogram matching step using a random Spectralis OCT volume as a template, followed by the same filtering operations as described for T_1. These strategies were applied as baseline translations when converting images from the Cirrus to Spectralis domain. Translations from Spectralis to Cirrus are not commonly employed in the literature, since Cirrus B-scans have a lower signal-to-noise ratio (SNR).

2.3. Segmentation models

A segmentation model f_θ with parameters θ aims to output a label $y \in \{1, \ldots, k\}$ for each pixel x in an input image $I \in \mathbb{R}^{h\times w}$ (i.e. a B-scan), with k being the maximum number of classes and h and w the height and width of the images in pixels. In general, f is modeled using convolutional neural networks. In that case, the parameters θ are learned in a supervised way from a training set $S = \{(I^{(i)}, Y^{(i)})\}, 1 \le i \le n$ with n pairs of training images $I^{(i)}$ and their corresponding manual annotations $Y^{(i)}$. This is done by minimizing a pixel-wise loss function $J(f_\theta(I^{(i)}), Y^{(i)})$ that penalizes the differences between the prediction $f_\theta(I^{(i)})$ and its associated ground truth labeling $Y^{(i)}$.

In the fluid and photoreceptor segmentation tasks analyzed in this paper, f is a fully convolutional neural network with an encoder-decoder architecture, inspired by the U-Net [28] architecture. The encoder path consists of convolution blocks followed by max-pooling layers, which contracts the input and uses the context for segmentation. The decoder counterpart, on the other hand, performs up-sampling operations followed by convolution blocks, enabling precise localization in combination with skip-connections. For fluid segmentation, a standard U-Net architecture was applied. For photoreceptor segmentation, we took advantage of the recently proposed U2-Net [29], which is to the best of our knowledge the only existing deep learning approach for this task.

3. Experimental setup

We evaluated the ability of cycleGANs for translating OCT images from one vendor to another in a twofold basis. First, we estimated the generalization ability of models for fluid (i.e. pathology) and photoreceptor layer (i.e. regular anatomy) segmentation on a new unseen domain when translating the input images to resemble those used for training. Secondly, we performed a visual Turing test in which different expert groups analyzed the images produced by the

Biomedical Optics EXPRESS

cycleGAN. We additionally asked the experts to identify morphological changes introduced by the cycleGAN-based translations.

The datasets (Section 3.1), deep learning training setups (Section 3.2) and the evaluation of the translation models (Section 3.3) are described next.

3.1. Materials

All the OCT volumes used in our experiments were acquired either with Cirrus HD-OCT 400/4000 (Carl Zeiss Meditec, Dublin, CA, USA) or Spectralis (Heidelberg Engineering, GER) devices. Spectralis (Cirrus) devices utilize a scanning superluminescence diode to emit a light beam with a center wavelength of 870nm (840nm), an optical axial resolution of 7μm (5μm) in tissue, and an optical transversal resolution of 14μm (10μm) in tissue. All the scans were centered at the fovea and covered approximately the same physical volume of $2\mu m \times 6\mu m \times 6\mu m$. Cirrus images had a voxel dimension of $1024 \times 512 \times 128$ or $1024 \times 200 \times 200$, and Spectralis volumes had a voxel dimension of $496 \times 512 \times 49$. All Cirrus volumes were resampled using nearest-neighbor interpolation to match the resolution of Spectralis volumes ($496 \times 512 \times 49$). As observed in Fig. 2, the B-scans produced by each device differ substantially from each other. In addition to the differences mentioned earlier with respect to the scanning light source, axial and transversal OCT resolution, the Spectralis devices apply a B-scan averaging procedure (by default 16 frames per B-scan). This procedure leads to an improved signal-to-noise ratio (SNR) compared to Cirrus scans. Image data was anonymized and ethics approval was obtained for research use from the ethics committee at the Medical University of Vienna (Vienna, Austria).

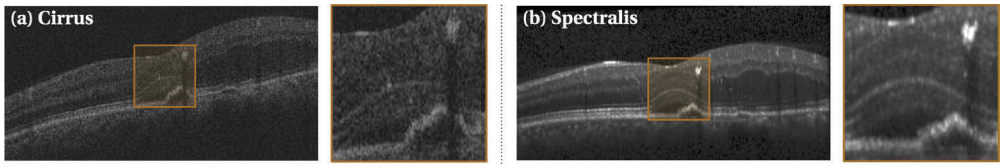


Fig. 2. Cirrus **(a)** and Spectralis **(b)** B-scans with a corresponding close-up of the retinal layers. Both B-scans were acquired from the same patient at approximately the same time and retinal location, to illustrate their difference in appearance.

Four different datasets were used for our study, each of them for (1) training the *unpaired translation* model, (2) train and evaluate the *fluid segmentation* method, (3) train and evaluate the *photoreceptor layer segmentation* method and (4) for the *visual Turing test*. All the datasets and their corresponding partitions into training, validation and test are described in the sequel.

Unpaired Training Dataset This dataset comprises a total of 1,179 OCT volumes (57,771 B-scans after resampling Cirrus OCTs). 592 (587) OCT volumes were acquired with the Spectralis (Cirrus) device. The Spectralis (Cirrus) OCT volumes correspond to pathological retinas of patients suffering from different diseases, with 195 (192) samples with age-related macular degeneration (AMD), 197 (196) with retinal vein occlusion (RVO) and 200 (199) with diabetic macular edema (DME). For each vendor, the data was randomly split into training (51,695) and validation sets (6,076), with no patient overlap between these two sets.

Fluid Segmentation Dataset A total amount of 228 OCT volumes (66 Spectralis, 162 Cirrus) were selected for the fluid segmentation experiment, comprising a total number of 11,172 B-scans. The Spectralis (Cirrus) set contained 59(62) pathological retinas with AMD and 7(100) with RVO. Experienced graders of the Vienna Reading Center performed manual pixel-wise annotations of intra-retinal cystoid fluid (IRC) and sub-retinal fluid (SRF) on those OCT volumes. The annotations were supervised by retinal expert ophthalmologists following a standardized annotation protocol. The datasets were randomly divided on patient distinct basis into 2254

Biomedical Optics EXPRESS

(5586), 294 (441), and 686 (1911) B-scans used for training, validation and testing, respectively, on the Spectralis (Cirrus) set.

Photoreceptor Layer Segmentation Dataset Photoreceptor segmentation experiments were performed using two data sets of Cirrus and Spectralis scans, comprising 43 and 50 volumes, respectively. Each Cirrus (Spectralis) volume is composed of 128 (49) B-scans with a resolution of 496×512 pixels. All the volumes were acquired from diseased patients. In particular, the Spectralis subset comprised 16 images with DME, 24 with RVO and 10 with intermediate AMD. The distribution of diseases in the Cirrus subset was approximately uniform, with 16 DME, 17 RVO and 10 early AMD cases. Each volume was manually delineated by trained readers and supervised by a retina expert that corrected the segmentations when needed. The datasets were randomly divided on patient-basis into 1519 (1323), 196 (147) and 739 (637) B-scans used for training, validation and test, respectively, on the Spectralis (Cirrus) set.

Visual Turing Test Dataset For the visual Turing test, we randomly selected a subset of B-scans from the *fluid segmentation test set*. In particular, B-scans were sampled from randomly selected OCT volumes following a Gaussian distribution, having a mean on the central B-scan (#25) and a standard deviation of 8 B-scans. A total of 90 Spectralis and 90 Cirrus B-scans were selected.

3.2. Training setup

The network architectures, training setups and configurations of each deep learning model are summarized in the sequel for each stage of our evaluation pipeline.

Unsupervised Image Translation Model Four different cycleGAN models were trained on the *unpaired training dataset* using squared patches of sizes of 64×64, 128×128, 256×256, and 460×460, respectively finally resulting in four different models (*CycleGAN*64, *CycleGAN*128, *CycleGAN*256, *CycleGAN*460). Each training phase consisted of 20 epochs using a mini batch size of 1. At each epoch, the deep learning model processes a patch randomly extracted from each B-scan in the training set. Each pair of generator/discriminator was saved after each epoch to subsequently select the best performing model. To address the fact that adversarial losses are unstable during training, we applied the following model selection strategy based on the validation set. First, all 20 generators were used to translate the corresponding (Cirrus or Spectralis) central B-scans of the validation set. Then, the average $L_{\text{GAN}}(D)$ term in Eq. 2 was computed for all associated discriminators using the same image sampling order for the target and source image sets. The maximum adversarial loss of the 20 discriminators was then used as the final selection score for each generator. Finally, the generator with the minimum score was selected for a specific patch size configuration. This procedure allowed us to select pairs of generators that were not necessarily paired at the training stage.

The generator is an encoder/decoder ResNet based architecture with the following structure. The first layer was a 7×7 Convolution-InstanceNorm-ReLU block with 32 filters and a stride of 1, followed by two 3×3 Convolution-InstanceNorm-ReLU blocks with 64 and 128 filters, respectively. Afterwards, a sequence of 9 residual blocks were stacked, each of them comprising two 3×3 convolutional layers with 128 filters. Subsequently, two 3×3 Transposed-Convolution-InstanceNorm-ReLU blocks with 64 and 32 filters (stride $\frac{1}{2}$) were followed by a final 7×7 Convolution-InstanceNorm-ReLU block with 3 filters. On the other hand, the discriminator consisted of four 4×4 Convolution-InstanceNorm-LeakyReLU blocks with 64, 128, 256 and 512 filters.

Fluid Segmentation Model The fluid segmentation model is an encoder/decoder network inspired by the U-Net architecture. We used five levels of depth, with the number of output channels going from 64 in the first to $1,024$ in the bottleneck layer, in powers of 2. Each convolutional block consisted of two 3×3 convolutions, each followed by a batch-normalization

Biomedical Optics EXPRESS

layer and a rectified linear unit (ReLU). While 2×2 max-pooling was used for downsampling, upsampling was performed using nearest-neighbor interpolation.

We used the negative log-likelihood loss in all our segmentation experiments, Kaiming initialization [30], Adam optimization [31], and a learning rate of $1e^{-3}$, which was decreased by half every 15 epochs. We trained our networks for 80 epochs and selected the model with the best average F_1-score on the validation set.

Photoreceptor Layer Segmentation Model Photoreceptor segmentation was performed by means of the U2-Net approach described in [29]. Such an architecture allows to retrieve probabilistic segmentations of the region of interest and uncertainty maps highlighting potential areas of pathological morphology and/or errors in the prediction. The core model is inspired by the U-Net, while incorporating dropout with a rate of 0.2 after several convolutional layers. By using dropout in test time, $T = 10$ Monte Carlo samples were obtained, retrieving the final segmentation as the pixel-wise average of the resulting samples.

3.3. Evaluation of the translation model

We evaluated the quality of the translated images obtained by the cycleGAN translation models in two different scenarios. On one hand, we measured the ability of the cycleGAN algorithm to reduce the covariate shift between images from different OCT vendors in automated retinal segmentation tasks. On the other hand, we carried out a visual Turing test to evaluate both the "realism" of the generated images and to identify potential morphological artifacts introduced in the translated version of the scans.

Evaluation via Segmentation Tasks The performance of the segmentation models was assessed in several versions of the B-scans in the corresponding test set. These B-scan versions were obtained by applying one of the following processes: (1) No translation, four different versions of cycleGAN models trained with different image patch size: (2) *CycleGAN*64, (3) *CycleGAN*128, (4) *CycleGAN*256 and (5) *CycleGAN*460. Additionally, the (6) $T1$ and (7) $T2$ translations were only applied as baseline for the Cirrus-to-Spectralis translation (Section 2.2). We evaluated the performance of the segmentation models for the different image versions by computing the Dice score, precision and recall on the test set at voxel level. One-sided Wilcoxon signed-rank tests were performed to test for statistically significant differences.

Evaluation via Visual Turing Test The perceptual evaluation was carried out by a group of 18 participants with a different professional background (6 computer scientists, 6 OCT readers and 6 ophthalmologist), all of them experienced working with OCT images. The graphical user interface was implemented using the jsPsych library [32]. In the first task, the participants were asked to identify the "fake" translated image of a shown pair of original/translated B-scans. For this part of the visual Turing test, one-sided Wilcoxon signed-rank tests were performed to check for statistically significant differences.

In the second task, the participants were asked to identify morphological changes in the retina introduced by the cycleGAN-based translations. The evaluation resulted in a *morphology preservation score (MPS)* for each original/translated image pair, ranging from 1 (morphological changes detected) to 5 (no morphological changes detected). In particular, the participants were asked to rate their level of agreement (from 1 'strong disagreement' to 5 'strong agreement') with the statement "There are no morphological differences between the original/translated image", meaning that MPS $\in \{1, 2, 3, 4, 5\}$. If the MPS was 1 or 2, the participants were further asked to select the portions of the image in which they found differences. This second task was carried out only by the ophthalmologist group.

Each participant observed 60 translated/original B-scans pairs (30 original Spectralis and 30 original Cirrus B-scans), randomly sampled from the *Visual Turing Test Dataset* (Section 3.1). For each sample, randomly either the *CycleGAN*128 or *CycleGAN*460 model was used to generate

Biomedical Optics EXPRESS

the translated version of the image. For the second task, one-sided Mann-Whitney-U-Tests were performed to determine statistical significance.

4. Results

While general qualitative results for the Cirrus-to-Spectralis translation are illustrated in Fig. 3, qualitative segmentation results are shown in Figs. 4–5 for both translation directions. Retinal fluid segmentation results are presented in Section 4.1 (Figs. 6–7) and results for retinal photoreceptor layer segmentation are provided in Section 4.2 (Fig. 8). Results of the Visual Turing test are covered in Section 4.3 (Figs. 9–10).

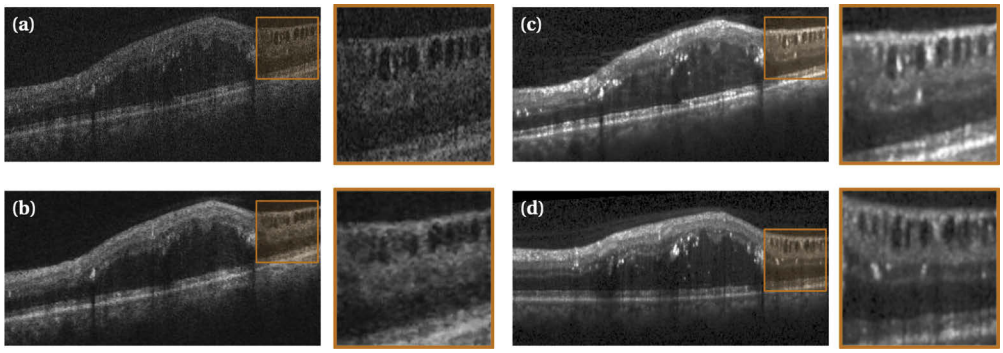


Fig. 3. Qualitative results of the image translation algorithms. An original Cirrus OCT B-scan **(a)** was translated to the Spectralis domain using T_2 **(b)**, and *CycleGAN*460 **(c)**. The corresponding original Spectralis B-scan **(d)** acquired from the same patient at approximately the same time and retinal location is also shown for reference. The image generated in **(c)** has intensity values and an image noise level similar to those observed in the original Spectralis image **(d)**.

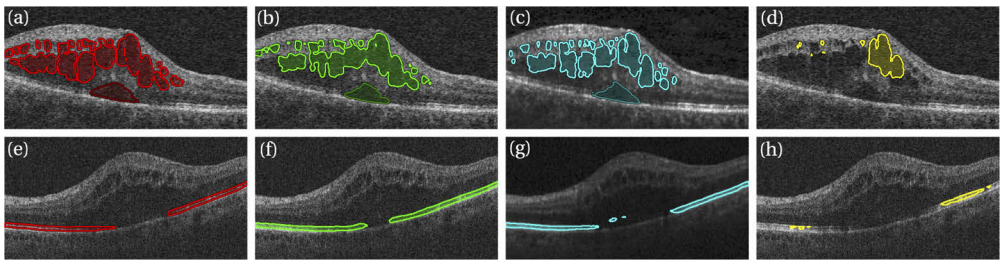


Fig. 4. Qualitative results in the Cirrus test set, shown for fluid (top row: bright=IRC, dark=SRF) and photoreceptor layer segmentation tasks (bottom row). Manual annotations are shown in red (a,e). Segmentations of the upper-bound model (*Cirrus-on-Cirrus*) are highlighted in green (b,f). Results of the best cycleGAN-models are illustrated in cyan (c: CycleGAN256, g: CycleGAN128). Predictions of the Spectralis model on the original Cirrus-scans (no translation) are denoted in yellow (d,h).

4.1. Fluid segmentation results

Initial evaluation of the Spectralis segmentation model on the Spectralis test set showed a Precision, Recall and Dice for IRC (SRF) of 0.78 (0.86), 0.75 (0.74) and 0.76 (0.76). The evaluation of the Cirrus segmentation model on the Cirrus test set yielded a Precision, Recall and Dice for IRC (SRF) of 0.74 (0.87), 0.67 (0.45), 0.66 (0.56). Notably, the performance of the

Biomedical Optics EXPRESS

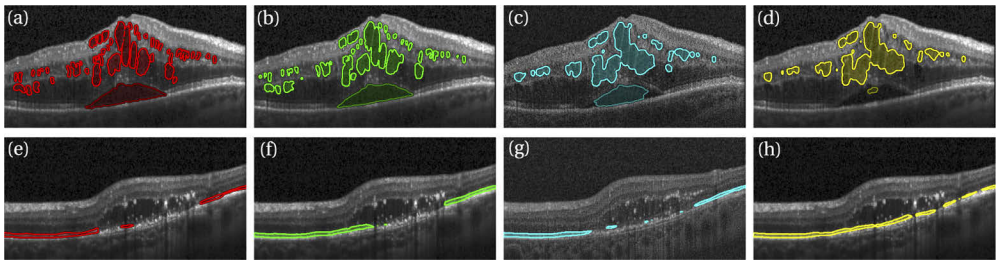
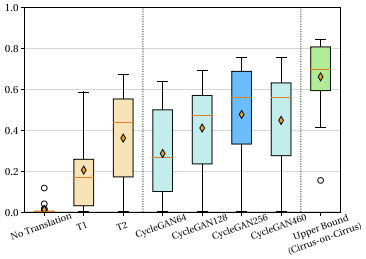


Fig. 5. Qualitative results in the Spectralis test set, shown for fluid (top row: bright=IRC, dark=SRF) and photoreceptor layer segmentation tasks (bottom row). Manual annotations are shown in red (a,e). Segmentations of the upper-bound model (*Spectralis-on-Spectralis*) are highlighted in green (b,f). Results of the best cycleGAN-models are illustrated in cyan (c: CycleGAN256, g: CycleGAN64). Predictions of the Cirrus model on the original Spectralis-scans (no translation) are denoted in yellow (d,h).

(a)

Translation	Precision	Recall	Dice
None	0.85±0.18	0.01±0.02	0.01±0.03
*T*1	**0.87±0.12**	0.13±0.14	0.21±0.19
*T*2	0.52±0.31	0.31±0.24	0.36±0.24
CycleGAN64	0.57±0.30	0.21±0.18	0.29±0.22
CycleGAN128	0.73±0.30	0.31±0.18	0.41±0.22
CycleGAN256	0.65±0.31	**0.42±0.23**	**0.48±0.25**
CycleGAN460	0.70±0.33	0.36±0.23	0.45±0.25
**Cirrus-on-Cirrus*	*0.74±0.23*	*0.67±0.15*	*0.66±0.18*

(b)



(c)

Translation	Precision	Recall	Dice
None	0.64±0.20	0.61±0.20	0.59±0.15
CycleGAN64	0.73±0.20	0.38±0.22	0.44±0.21
CycleGAN128	**0.73±0.22**	0.39±0.19	0.47±0.19
CycleGAN256	0.59±0.19	**0.67±0.21**	0.58±0.15
CycleGAN460	0.62±0.21	0.66±0.20	**0.59±0.16**
**Spec.-on-Spec.*	*0.78±0.13*	*0.75±0.11*	*0.76±0.11*

(d)

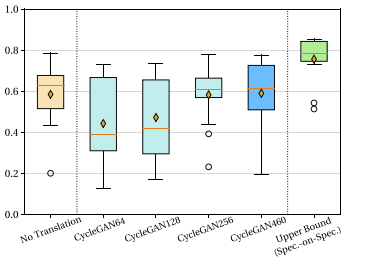


Fig. 6. Quantitative results of intra-retinal cyst (IRC) segmentation, obtained on (a-b) Cirrus and (c-d) Spectralis scans. For different translation strategies Precision, Recall and Dice are shown together with the corresponding box-plots of Dice values. Native upper-bound models are indicated by * in (a-b). The CycleGAN model with highest mean Dice is highlighted in dark blue (b,d).

native Cirrus models was lower than of their Spectralis counterparts, with the larger difference for the SRF class.

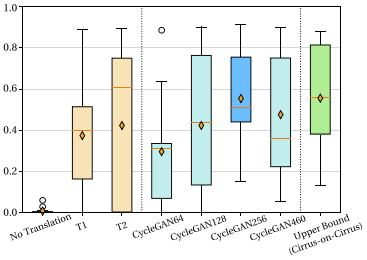
For the IRC class, the cross-vendor evaluation of both segmentation models (Spectralis and Cirrus) on the non-translated datasets showed a clear performance drop (Fig. 6), especially on the Cirrus test set (Fig. 6(a-b)). There, all applied translation strategies significantly improved the cross-vendor segmentation model performance with respect to the 'no translation' scenario ($p << 0.05$). The approach with the highest Dice was the cycleGAN model trained with image patches of 256×256 (*CycleGAN256*). This model showed a significantly better performance both compared to the T_1 and T_2 baseline approaches ($p < 0.05$). Finally, the tests did not yield a significant difference between the best translation model (*CycleGAN460*)) and the 'no translation' scenario in the Spectralis test set (Fig. 6(c-d).

Biomedical Optics EXPRESS

(a)

Translation	Precision	Recall	Dice
None	**0.97±0.02**	0.01±0.02	0.01±0.02
*T*1	0.80±0.26	0.29±0.27	0.37±0.28
*T*2	0.53±0.37	0.40±0.37	0.42±0.37
CycleGAN64	0.77±0.23	0.22±0.24	0.30±0.25
CycleGAN128	0.75±0.32	0.34±0.29	0.42±0.31
CycleGAN256	0.75±0.23	**0.49±0.28**	**0.55±0.25**
CycleGAN460	0.68±0.26	0.43±0.30	0.48±0.31
**Cirrus-on-Cirrus*	*0.87±0.12*	*0.45±0.24*	*0.56±0.23*

(b)



(c)

Translation	Precision	Recall	Dice
None	0.88±0.21	0.48±0.32	0.55±0.30
CycleGAN64	**0.89±0.20**	0.27±0.29	0.34±0.32
CycleGAN128	0.86±0.20	0.23±0.23	0.31±0.27
CycleGAN256	0.80±0.21	**0.61±0.23**	**0.66±0.22**
CycleGAN460	0.75±0.24	0.49±0.18	0.56±0.18
**Spec.-on-Spec.*	*0.86±0.19*	*0.74±0.26*	*0.76±0.23*

(d)

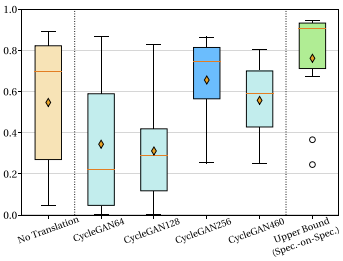
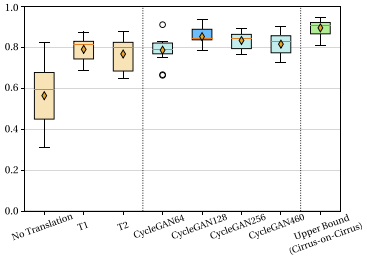


Fig. 7. Quantitative results of sub-retinal fluid (SRF) segmentation, obtained on (a-b) Cirrus and (c-d) Spectralis scans. For different translation strategies Precision, Recall and Dice are shown together with the corresponding box-plots of Dice values. Native upper-bound models are indicated by * in (a-b). The CycleGAN model with highest mean Dice is highlighted in dark blue (b,d).

(a)

Translation	Precision	Recall	Dice
None	**0.91±0.05**	0.43±0.17	0.56±0.15
T1	0.86±0.06	0.74±0.10	0.79±0.07
T2	0.87±0.05	0.70±0.13	0.77±0.08
CycleGAN64	0.86±0.05	0.73±0.11	0.79±0.06
CycleGAN128	0.87±0.06	**0.84±0.05**	**0.85±0.04**
CycleGAN256	0.83±0.05	0.83±0.04	0.83±0.05
CycleGAN460	0.80±0.07	0.83±0.06	0.81±0.05
**Cirrus-on-Cirrus*	*0.88±0.05*	*0.91±0.04*	*0.89±0.04*

(b)



(c)

Translation	Precision	Recall	Dice
None	0.85±0.07	0.84±0.16	0.84±0.11
CycleGAN64	0.86±0.05	**0.90±0.07**	**0.88±0.05**
CycleGAN128	0.86±0.06	0.90±0.08	0.88±0.06
CycleGAN256	**0.87±0.06**	0.85±0.06	0.86±0.05
CycleGAN460	0.86±0.06	0.77±0.07	0.81±0.05
**Spec.-on-Spec.*	*0.89±0.04*	*0.89±0.06*	*0.90±0.05*

(d)

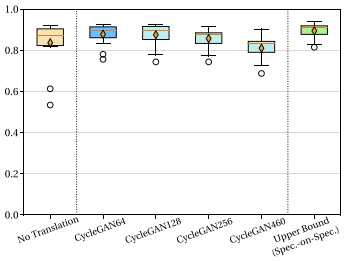


Fig. 8. Quantitative results of layer segmentation, obtained on (a-b) Cirrus and (c-d) Spectralis scans. For different translation strategies Precision, Recall and Dice are shown together with the corresponding box-plots of Dice values. Native upper-bound models are indicated by * in (a-b). The CycleGAN model with highest mean Dice is highlighted in dark blue (b,d).

We also observed a performance drop in the cross-vendor evaluation of the SRF class when no translation was applied, where this drop was more prominent in the Cirrus test set (Fig. 7(a-b)). There, all the translation models showed an improvement with respect to the scenario without

Biomedical Optics EXPRESS

translation ($p << 0.05$). The best performing *cycleGAN* model was trained with images patches of 256×256 (*CycleGAN256*), showing a significantly better performance than the $T1$ and $T2$ baseline translation algorithms ($p < 0.05$). Notably, the *CycleGAN*256 model performed on a par with the upper-bound *Cirrus-on-Cirrus* model ($p = 0.37$). In the Spectralis test set (Fig. 7(c-d)), the *CycleGAN256* model achieved a higher mean Dice and a lower variance compared to the "no translation" scenario. However, a significant difference between the distributions was not found ($p = 0.18$).

4.2. Photoreceptor layer segmentation results

Results for the photoreceptor layer segmentation task are summarized in Fig. 8. The upper-bound models obtained a Precision, Recall and Dice in the Cirrus (Spectralis) test set of 0.88 (0.89), 0.91 (0.89) and 0.89 (0.90). A drop in Dice was observed for both segmentation models when no translation was used for both translation directions. The decrease was larger in the Cirrus test set (0.90 to 0.56). Moreover, the cross-vendor evaluation showed that all applied

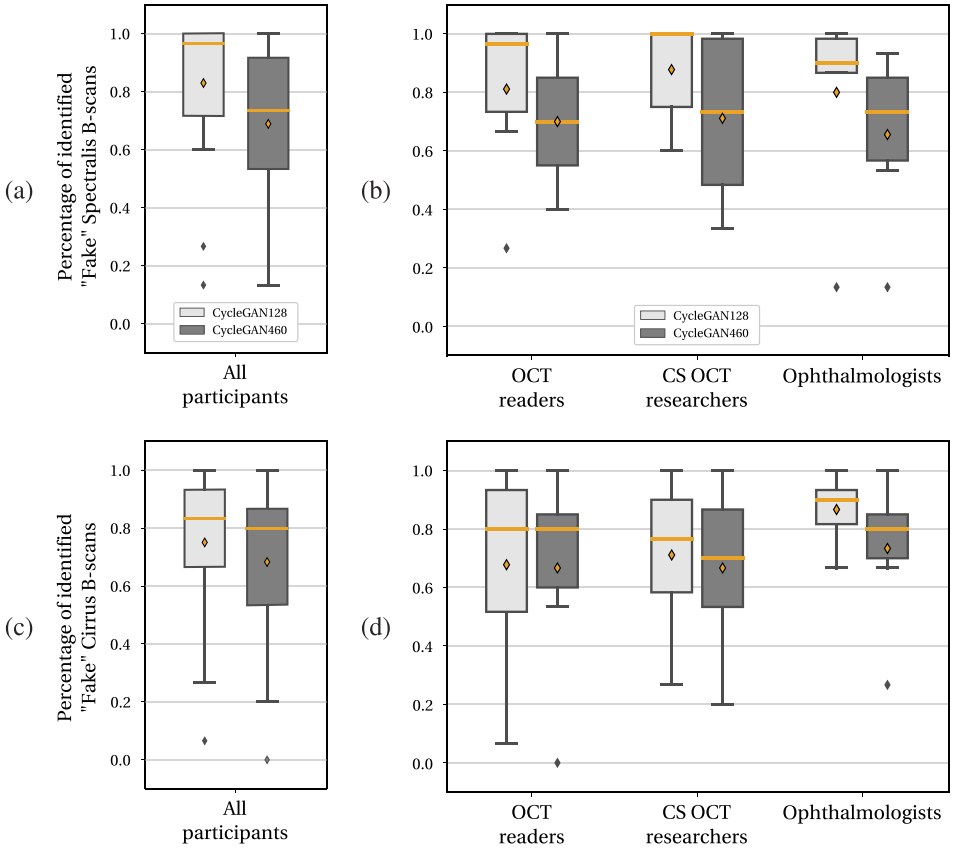


Fig. 9. Box plots illustrating the amount of correctly identified "fake" B-scans, when showing the original/translated pairs to the participants (OCT readers ($n = 6$), computer scientists ($n = 6$), ophthalmologists ($n = 6$)). The amount of identified "fake" Spectralis B-scans for the Cirrus-To-Spectralis translation is shown in the top row for (a) all participants and (b) stratified by expertise background. The amount of identified "fake" Cirrus B-scans for the Spectralis-To-Cirrus translation is shown in the bottom row for (c) all participants and (d) stratified by expertise background.

Biomedical Optics EXPRESS

translation strategies significantly outperformed the "no translation" scenario in the Cirrus test set ($p << 0.05$, Fig. 8(a-b)). The best result in terms of Dice was obtained by *CycleGAN*128, showing a significantly better performance than $T1$ and $T2$ ($p << 0.05$). In the Spectralis test set (Fig. 8(c-d)), the *CycleGAN*64 based translation achieved the highest Dice (0.88), significantly outperforming the "no translation" scenario ($p < 0.05$).

4.3. Visual Turing test results

The results of the first visual Turing task involving an original/translated image pair (see Section 3.3) are summarized in Fig. 9, showing the percentage of identified "fake" Spectralis B-scans for the Cirrus-to-Spectralis translation in Fig. 9(a-b) and the percentage of identified "fake" Cirrus B-scans for the Spectralis-to-Cirrus translation in Fig. 9(c-d). Note that a value equal to 0.5 would correspond to a scenario in which the participant randomly select any of the B-scans as fake, meaning that the transformed images could not be distinguished from the original scans by the participants.

When evaluating the median of the amount of identified "fake" images in the (*CycleGAN*128 / *CycleGAN*460) models across all participants, a difference was found in both the median identification rate of the Cirrus-to-Spectralis (0.97 / 0.73) and the Spectralis-to-Cirrus (0.83 / 0.8) translations. This means that the images that were generated by the *CycleGAN*460 model were harder to identify than the images generated with the *CycleGAN*128 model. A paired one-sided Wilcoxon signed-rank significance test found that the difference in both directions was significant ($p << 0.05$ in the Cirrus-to-Spectralis and $p < 0.05$ in the Spectralis-to-Cirrus direction). This is also reflected in Fig. 9(b), where for each expert group the generated "fake" Spectralis B-scans of the *CycleGAN*460 model were harder to identify than generated B-scans of the *CycleGAN*128 model. However, this effect was less pronounced in the other translation direction (Spectralis-to-Cirrus, Fig. 9(d)). Finally, we can observe that the ophthalmologists showed a lower variance compared to the other expert groups.

Quantitative results of the second visual Turing task (Fig. 10) are showing the distribution of the MPS for both the *CycleGAN*128 and *CycleGAN*460 model as well as for both translation directions. For the Cirrus-to-Spectralis translation, the median MPS was significantly higher for *CycleGAN*128 (median MPS=5) than for *CycleGAN*460 (median MPS=4), with $p << 0.05$.

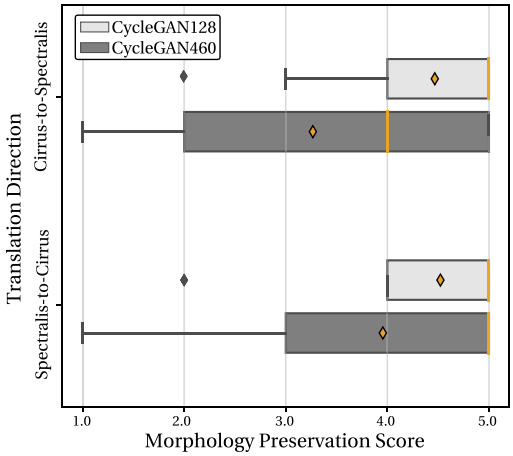


Fig. 10. Boxplot of the morphology preservation scores (MPS) assigned by the ophthalmologist group to the original/translated B-scan pair. *Top:* Cirrus-to-Spectralis translation ($N = 180$). *Bottom:* Spectralis-to-Cirrus translation ($N = 180$).

Biomedical Optics EXPRESS

The same trend can be observed for the Spectralis-to-Cirrus translation ($p < 0.05$), although the median MPS was the same for both cycle-GAN models (median MPS=5). In summary, these results clearly indicate that the *CycleGAN*128 model introduced fewer morphological changes during the translation compared with the *CycleGAN*460 model, for both translation directions. Exemplary qualitative results for introduced morphological changes are shown in Fig. 11.

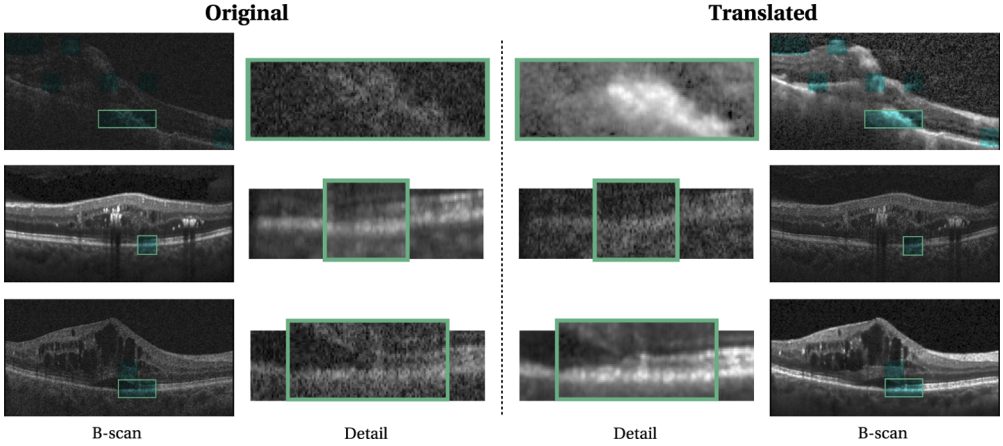


Fig. 11. *Left:* Original B-scan and a zoomed-in region of interest (6× original size). *Right:* Translated B-scan with the same zoomed-in region of interest. Regions that were identified as morphological differences by the experts are highlighted in cyan. *First row:* Spectralis-to-Cirrus translation. *Second and third row:* Cirrus-to-Spectralis translation.

5. Discussion

The main hypothesis of our work was that unsupervised cycleGAN based translation algorithms would allow to significantly reduce the covariate shift phenomenon in automated retinal segmentation models across different OCT acquisition devices. Our results show that these translation approaches indeed allowed deep learning models to improve its generalization ability in cross-vendor OCT images unseen at training stage. In all the evaluated segmentation tasks the effect of the covariate shift phenomenon was larger when using Spectralis models on the Cirrus test set than when using Cirrus models on the Spectralis test set, without any translations (Figs. 6(a-b), 7(a-b) and 8(a-b)). A possible explanation may be that the Spectralis segmentation model could not deal with the lower SNR in Cirrus scans, e.g. due to the noise in the fluid regions appearing much brighter than in Spectralis images. Cirrus models seemed to be more resilient against the covariate shift phenomenon when applied on Spectralis scans, showing a smaller but still significant performance drop compared to the upper bound (Figs. 6(c-d), 7(c-d) and 8(c-d)).

In most of the segmentation tasks and translation directions (Cirrus-to-Spectralis or Spectralis-to-Cirrus), we found that the cycleGAN-based translation models trained with image patch size of 256×256 improved significantly or slightly – depending on the translation direction and segmentation task – the performance of the segmentation models, with respect to a scenario without any translation. However, each task and translation direction had a different optimal training image patch size. In the fluid segmentation task, in which the objective is to identify hypo-reflective black regions, the best performance was obtained with larger image patch sizes (256×256, 460×460). In contrast, for the photoreceptor layer segmentation task, in which the aim is to detect a thin layered region, the best performance was obtained with smaller image patch sizes (64×64, 128×128). One factor explaining these results may be the size of the structures that are segmented. Larger structures seem to require larger training image patches for

Biomedical Optics EXPRESS

the cycleGAN to capture the needed contextual information. Thus, the segmentation models may focus more on the global appearance (context) when segmenting larger structures, meaning that in this case it may be more important to reduce the covariate shift effect on a "global appearance level" rather than on a local one. Additionally, using larger patches in the training stage allowed the translation models to generate more realistic images at test time. However, those models were also more likely to introduce image artifacts in the translated B-scans. This was empirically demonstrated in the visual Turing test, as discussed in the next paragraphs.

The first part of the visual Turing test was conducted to evaluate the realism of the generated B-scans. This part of the visual Turing test was inspired by [33], but our set-up constituted a more stringent assessment, since it not only required the participants to judge the "realism" of the generated images but also enabled a direct comparison with the original image. The main finding of this test was that using larger image patch sizes during training of the cycleGAN-based models resulted in more "realistic" translated B-scans (Section 3.3). The results of the visual Turing test showed that it was harder to identify the *CycleGAN*460 generated "fake" images both in Spectralis-to-Cirrus and Cirrus-to-Spectralis directions (Figs. 9(a,c)). This may be related with the above mentioned theory that larger patches during training allow to learn more complex/realistic translations and reduce the covariate shift effect also on a "global appearance level". When observing the percentage of identified "fake" B-scans stratified by expertise background (Figs. 9(b,d)), we found that the ophthalmologist group seemed to perform consistently in identifying "fake" images generated by the *CycleGAN*128. This result might indicate that their knowledge about retinal structures and B-scan appearance in general helped them to robustly identify "fake" images. However, this advantage was no longer evident for the *CycleGAN*460 model. Some of the reported cues the ophthalmologists and OCT readers used for identifying fake B-scans were the quality of choroid tissue, the vitreous border or the smoothness of retinal layer borders. Conversely, participants of the CS OCT researchers group reported relying more on visual cues based on differences in pre-processing operations used by the OCT acquisition devices (i.e quality of the B-scan filtering, B-scan tilt correction.

The results of the second part of the visual Turing test illustrate that translation models trained with smaller image patch sizes generate a smaller amount of retinal morphological differences in the translated images. An explanation of this phenomenon might be that training translation models with a smaller patch size results in more conservative (simpler) translations less likely to introduce artifacts (but also less realistic). In particular, the MPS distribution for the *CycleGAN*460 models (in both directions) had a significantly larger number of decisions rating the image pairs as having strong morphological differences (1 – 3) than that observed for the *CycleGAN*128 model. This indicates that models trained with larger patch sizes are prone to induce artifacts in the generated images. Additionally, the amount of perceived retinal anatomical differences also depended on the direction of the translation. When translating from Cirrus-to-Spectralis a slightly larger number of morphological differences were identified than in the B-scans translated from Spectralis-to-Cirrus.

We conjecture that translating from a low-SNR to a high-SNR B-scan required the translation models to "invent" information that was not present in the source image, meaning that the lower the signal-to-noise ratio the more likely it is that image artifacts are introduced. Moreover, we empirically observed that the quality of the input images was also a factor for the quality of the translated images. For instance, an extremely low contrast B-scan from any of the OCT vendors would likely generate a low quality B-scan after the translation. An example of such a case is presented in the first row of Fig. 11. We also found qualitatively that a substantial amount of regions highlighted by the ophthalmologists as artifacts were related to the border of the layers delimiting the retina. For instance, changes in the appearance of the bottom layers were observed in the two lower rows of Fig. 11. In the second row, the generated Cirrus B-scan attenuates/removes one of the layers observed in the original Spectralis B-scan. In the third row,

Biomedical Optics EXPRESS

the appearance of the retinal pigment epithelial (RPE) layer seemed to be altered in the translated Spectralis B-scan compared to the Cirrus original counterpart.

6. Conclusion

In this work, we presented an unsupervised unpaired learning strategy using cycleGAN to reduce the image variability across OCT acquisition devices. The method was extensively evaluated in multiple different retinal OCT image segmentation tasks (IRC, SRF, photoreceptor layer) and visual Turing tests. The results show that the translation algorithms improve the performance of the segmentation models on the target datasets coming from a different vendor than the training set, thus effectively reducing the covariate shift (the difference between the target and source domain). This demonstrates the potential of the presented approach to overcome the limitation of existing methods, whose applicability is usually limited to samples that match the training data distribution. In other words, the proposed translation strategy allows to improve the generalizability of segmentation models in OCT imaging. Since automated segmentation methods are expected to be part of routine diagnostic workflows [8] and could affect the therapy of millions of patients, this finding is of particular relevance. Specifically, the presented approach could help to reduce device-specific dependency of DL algorithms and therefore facilitate their deployment on a larger set of OCT devices.

Furthermore, results indicate that the training image patch size was an important factor for the performance of the cycleGAN-based translation model. Larger training image patch sizes usually resulted in models whose generated B-scans were more realistic, i.e., more difficult to identify as "fake" by human observers. However, such images were more likely to contain morphological differences in comparison with the source images. These results indicate that special care should be taken to reduce the likelihood of introducing morphological artifacts. Besides matching the pathological distributions across the domains in the cycleGAN training stage [34], a good practice may be to reduce the complexity of the learned translation by using smaller patch sizes, as there seems to be a trade-off between the "realism" and the quality of the translation models. In this context, future work should be focused on evaluating different architectures and/or loss-functions to address the problem of cross-vendor translation while preserving retinal anatomical features in the OCT image.

Funding

Christian Doppler Research Association; Austrian Federal Ministry for Digital and Economic Affairs; National Foundation for Research, Technology and Development.

Acknowledgments

We thank the NVIDIA corporation for GPU donation.

Disclosures

DRB, PS, JIO and HB declare no conflicts of interest. SMW: Bayer (C,F), Novartis (C) and Genentech (F). BSG: Roche (C), Novartis (C,F), Kinarus (F) and IDx (F). US-E: Böhringer Ingelheim (C), Genentech (C), Novartis (C) and Roche (C).

References

1. D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography," Science **254**(5035), 1178–1181 (1991).
2. J. Fujimoto and E. Swanson, "The development, commercialization, and impact of optical coherence tomography," Invest. Ophthalmol. Visual Sci. **57**(9), OCT1–OCT13 (2016).
3. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal. **42**, 60–88 (2017).

Biomedical Optics EXPRESS

4. U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," PROG RETIN EYE RES (2018).
5. A. J. Storkey, "When training and test sets are different: characterising learning transfer," in *Dataset Shift in Machine Learning* (The MIT Press, 2009), , pp. 3–28.
6. T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully automated detection and quantification of macular fluid in oct using deep learning," Ophthalmology **125**(4), 549–558 (2018).
7. L. Terry, N. Cassels, K. Lu, J. H. Acton, T. H. Margrain, R. V. North, J. Fergusson, N. White, and A. Wood, "Automated retinal layer segmentation using spectral domain optical coherence tomography: evaluation of inter-session repeatability and agreement between devices," PLoS One **11**(9), e0162001 (2016).
8. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. Van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nat. Med. **24**(9), 1342–1350 (2018).
9. A. Torrado-Carvajal, J. L. Herraiz, E. Alcain, A. S. Montemayor, L. Garcia-Cañamaque, J. A. Hernandez-Tamames, Y. Rozenholc, and N. Malpica, "Fast patch-based pseudo-ct synthesis from t1-weighted mr images for pet/mr attenuation correction in brain studies," J. Nucl. Med. **57**(1), 136–143 (2016).
10. C. Zhao, A. Carass, J. Lee, Y. He, and J. L. Prince, "Whole brain segmentation and labeling from ct using synthetic mr images," in *International Workshop on Machine Learning in Medical Imaging*, (Springer, 2017), pp. 291–298.
11. K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. of IPMI*, (Springer, 2017), pp. 597–609.
12. Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-modality image synthesis from unpaired data using cyclegan," in *International Workshop on Simulation and Synthesis in Medical Imaging*, (Springer, 2018), pp. 31–41.
13. W. M. Kouw, M. Loog, W. Bartels, and A. M. Mendrik, "Learning an mr acquisition-invariant representation using siamese neural networks," arXiv preprint arXiv:1810.07430 (2018).
14. T. de Bel, M. Hermsen, J. Kers, J. van der Laak, and G. Litjens, "Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology," in *Proc. of MIDL*, (2019), pp. 151–163.
15. C. Chen, Q. Dou, H. Chen, and P.-A. Heng, "Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation," in *International Workshop on Machine Learning in Medical Imaging*, (Springer, 2018), pp. 143–151.
16. Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," IEEE Trans. Med. Imaging **38**(4), 1016–1025 (2019).
17. J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan, "Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *Proc. of MICCAI*, (Springer, 2018), pp. 777–785.
18. J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Deep mr to ct synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging*, (Springer, 2017), pp. 14–23.
19. Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation," in *Proc. of MICCAI*, (Springer, 2018), pp. 599–607.
20. Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proc. of IEEE CVPR*, (2018), pp. 9242–9251.
21. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of IEEE ICCV*, (2017), pp. 2242–2251.
22. J. Wang, C. Bian, M. Li, X. Yang, K. Ma, W. Ma, J. Yuan, X. Ding, and Y. Zheng, "Uncertainty-guided domain alignment for layer segmentation in oct images," arXiv preprint arXiv:1908.08242 (2019).
23. G. Song, K. K. Chu, S. Kim, M. Crose, B. Cox, E. T. Jelly, J. N. Ulrich, and A. Wax, "First Clinical Application of Low-Cost OCT," Trans. Vis. Sci. Tech. **8**(3), 61 (2019).
24. P. Seeböck, D. Romo-Bucheli, S. Waldstein, H. Bogunovic, J. I. Orlando, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Using cyclegans for effectively reducing image variability across oct devices and improving retinal fluid segmentation," in *Proc. of IEEE ISBI*, (2019), pp. 605–609.
25. X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. of IEEE ICCV*, (2017).
26. X. Xu, K. Lee, L. Zhang, M. Sonka, and M. D. Abramoff, "Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical OCT data," IEEE Trans. Med. Imaging **34**(7), 1616–1623 (2015).
27. R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, and A. Bab-Hadiashar, "Retinal fluid segmentation in oct images using adversarial loss based convolutional neural networks," in *Proc. of IEEE ISBI*, (IEEE, 2018), pp. 1436–1440.

Biomedical Optics EXPRESS

28. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, (Springer, 2015), pp. 234–241.
29. J. I. Orlando, P. Seeböck, H. Bogunović, S. Klimscha, C. Grechenig, S. Waldstein, B. S. Gerendas, and U. Schmidt-Erfurth, "U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans," arXiv preprint arXiv:1901.07929 (2019).
30. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of IEEE ICCV*, (2015), pp. 1026–1034.
31. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
32. J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," Behav. Res. **47**(1), 1–12 (2015).
33. M. J. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis," in *Proc. of IEEE ISBI*, (IEEE, 2018), pp. 240–244.
34. J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *Proc. of MICCAI*, (2018), pp. 529–536.