Dual image and mask synthesis with GANs for semantic segmentation in optical coherence tomography

Jason Kugelman[1], David Alonso-Caneiro[1,2], Scott A. Read[1], Stephen J. Vincent[1], Fred K. Chen[2,3,4], and Michael J. Collins[1]

[1] Queensland University of Technology (QUT), Contact Lens and Visual Optics Laboratory, Centre for Vision and Eye Research, School of Optometry and Vision Science, Kelvin Grove, Qld 4059, Australia

[2]Centre for Ophthalmology and Visual Science (incorporating Lions Eye Institute), The University of Western Australia, Perth, Western Australia, Australia

[3]Department of Ophthalmology, Royal Perth Hospital, Perth, Western Australia, Australia

[4]Department of Ophthalmology, Perth Children's Hospital, Nedlands, Western Australia, Australia

{j.kugelman, d.alonsocaneiro, sa.read, sj.vincent, m.collins}@qut.edu.au, fredchen@lei.org.au

***Abstract*— In recent years, deep learning-based OCT segmentation methods have addressed many of the limitations of traditional segmentation approaches and are capable of performing rapid, consistent and accurate segmentation of the chorio-retinal layers. However, robust deep learning methods require a sufficiently large and diverse dataset for training, which is not always feasible in many biomedical applications. Generative adversarial networks (GANs) have demonstrated the capability of producing realistic and diverse high-resolution images for a range of modalities and datasets, including for data augmentation, a powerful application of GAN methods. In this study we propose the use of a StyleGAN inspired approach to generate chorio-retinal optical coherence tomography (OCT) images with a high degree of realism and diversity. We utilize the method to synthesize image and segmentation mask pairs that can be used to train a deep learning semantic segmentation method for subsequent boundary delineation of three chorio-retinal layer boundaries. By pursuing a dual output solution rather than a mask-to-image translation solution, we remove an unnecessary constraint on the generated images and enable the synthesis of new unseen area mask labels. The results are encouraging with near comparable performance observed when training using purely synthetic data, compared to the real data. Moreover, training using a combination of real and synthetic data results in zero measurable performance loss, further demonstrating the reliability of this technique and feasibility for data augmentation in future work.**

Keywords—generative adversarial networks, deep learning, OCT, neural networks

I. INTRODUCTION

In recent years, there has been increasing interest in generative adversarial networks (GANs) [1] within the rapidly progressing deep learning research field. Indeed, state-of-the-art GAN approaches are now capable of generating remarkably realistic and diverse high-resolution images for a range of modalities and datasets. The original GAN approach [1] proposed the idea of training two networks against each other: a generator and a discriminator. The generator's aim is to produce synthetic images that are realistic such that the discriminator cannot distinguish them from real images. The discriminator's aim is to differentiate between the two sets of images (real/synthetic). By training these two networks in an adversarial manner, each network learns and becomes stronger by attempting to maximise the error of the other. Unfortunately, the traditional GAN design suffers from a range of problems including vanishing and exploding gradients, lack of diversity in the generated images, limited resolution, training instability, sensitivity to network architecture choices, slow convergence, long training times, and mode collapse.

However, different GAN architectures and methods have been developed, which have significantly improved performance. Notably, the introduction of the Wasserstein GAN (WGAN) [2,3] has allowed for greater training stability, improved robustness to the choice of network architecture, and increased diversity within the generated images. An alternative approach, the least squares GAN (LSGAN) [4] can also stabilise training and prevent vanishing gradient problems. More recently, the progressive growing of GANs (PGAN) [5] methodology has been instrumental for the synthesis of higher resolution images. This involves beginning training at a lower resolution and gradually integrating additional layers throughout the training process to increase the resolution. This approach frames the problem of image generation in a way that is notably easier and faster to solve. A notable extension of the PGAN approach is the StyleGAN [6], which has demonstrated the ability to not only produce high resolution images that are both realistic and diverse but also allows for finer control over the style and features at different resolution levels.

Optical coherence tomography (OCT) has become a widely adopted imaging modality that can capture high-resolution cross-sectional images of living tissue in a non-invasive manner. OCT is particularly useful for acquiring in-vivo scans of the tissue layers at the back of the eye (retina and choroid), allowing for the clear visualisation of the layered retinal structure. The high level of detail within these scans enables clinicians and researchers to measure the thickness of the posterior eye tissues including, the internal retinal layers and choroid, on a micron scale. These measurements are critical for tracking thickness changes associated with normal ocular development and the progression and monitoring of eye diseases. Segmenting these images manually involves the tedious and slow process of an expert human grader marking the position of each individual layer boundary. To overcome

Funding: Telethon-Perth Children's Hospital Research Fund (FKC, DAC), National Health & Medical Research Council Ideas Grant (APP1186915, DAC), Rebecca L. Cooper 2018 Project Grant (DAC).

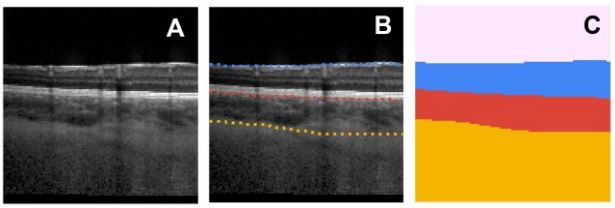

Fig. 1. A: Example of a down-sampled OCT image slice (128 x 128 pixels). B: The boundaries of interest marked with dotted lines, blue: ILM surface, red: RPE base, yellow: CSI. C: Corresponding area mask (128 x 128 pixels) with each region displayed in a different colour, pink: vitreous, blue: retina, red: choroid, yellow: sclera.

this inefficiency, several algorithms have been developed to automate this process. In particular, algorithms based on machine learning methods have demonstrated the ability to perform segmentation quickly, accurately, and consistently. A significant number of these methods [7-12], are based on neural networks and deep learning techniques. However, deep learning generally requires a sufficiently large and diverse set of labelled images for training, which may not always be available or feasible to obtain in many biomedical applications. One option to address a paucity of data is to employ data augmentation techniques.

Data augmentation using GANs has been explored for several image modalities. Wei et al. [13] utilised cycle-consistent GANs (CycleGANs) to generate synthetic colorectal histopathology images of less common polyp classes. Similarly, Sandfort et al. [14] used CycleGANs to transform contrast computed tomography (CT) images into their non-contrast variants. These generated images were then used for augmentation in several CT segmentation tasks, improving performance. Frid-Adar et al. [15] proposed the use of deep convolutional GANs (DCGANs) to generate computed tomography (CT) images of liver lesions that were used to augment the original dataset. Another approach by Yamaguchi et al. [16] proposed the use of multi-domain learning GANs in a method called Domain Fusion, which aims to generate new samples for the target dataset based on knowledge from an outer dataset.

There has been a variety of applications of GANs to OCT images not involving data augmentation. To detect anomalous images and image segments that can serve as imaging biomarker candidates, Schlegl et al. [17] proposed the Fast AnoGAN (f-AnoGAN) for anomaly detection in retinal OCT images. To reduce the variability of OCT images produced by two different instruments, Seeböck et al. [18] utilised CycleGANs to perform unsupervised unpaired image transformation while Cheong et al. [19] proposed their DeshadowGAN to remove shadows caused by blood vessel obstruction. Denoising using GANs has also been investigated [20-24] while Huang et al. [21] performed both simultaneous denoising and super-resolution using a GAN-based approach named SDSR-OCT.

Few studies have investigated OCT data augmentation using GANs. Kugelman et al. [25] explored the use of conditional GANs with a range of techniques to construct OCT image patches for patch-based OCT segmentation methods and noted the potential of this approach for data augmentation. Zheng et al. [26] assessed the ability of a PGAN to generate realistic synthetic OCT images of retinal disorders. They gauged the quality and realism of their images in a few ways, most notably with two retinal specialists who judged the quality of both the real and synthetic images without any prior knowledge of whether the image was real or fake (synthetic). They note the potential use of a GAN approach to generate synthetic images for both educational and training purposes as well as for use in developing deep learning algorithms for classifying retinal disorders.

In this study, we propose the use of GANs to generate chorio-retinal OCT data for training deep learning based semantic segmentation methods. One option is to pose this as an image-to-image translation problem to map an area mask label to a corresponding OCT image. Instead we generate not only the OCT image but its corresponding area mask label together as a dual output. This approach removes a constraint on the generated images and enables the generation of new, previously unseen area mask labels. Synthesising new labels has the potential to enhance the ability of the method as a tool for data augmentation. The main contributions of this paper are as follows:

- We demonstrate the feasibility of GANs to generate synthetic OCT data for semantic segmentation of chorio-retinal layers, setting a strong basis for future extensions such as data augmentation.
- We present an empirical analysis to better understand a number of relevant GAN parameters.
- We illustrate the ability of style localisation to be a powerful tool to control different levels of features in OCT.

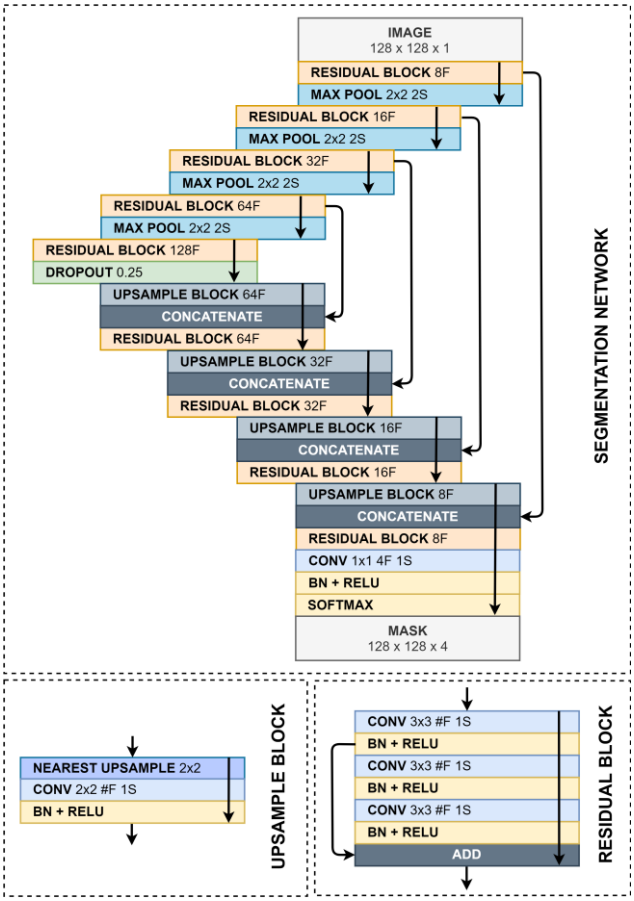

Fig. 2. Network used for semantic segmentation of the OCT image slices. #F: filters, #S stride, BN: batch normlization. Arrows show direction of information flow.

TABLE I. DATASET SUMMARY.

Set	Subjects	Total Scans	Total Slices
Training	40	240	4,080
Validation	10	60	1,020
Testing	49	294	4,998

TABLE II. PROGRESSIVE GROWING TRAINING STEPS.

Phase	4^2	8^2	16^2	32^2	64^2	128^2
Fade-In	-	10k	10k	10k	25k	50k
Consolidation	25k	25k	25k	25k	50k	100k

II. METHOD

A. Data

The images used throughout this study consists of a set of chorio-retinal spectral-domain OCT scans acquired from 99 healthy paediatric participants. This data, obtained over four visits, is taken from a previous study [27,28]. However, in our work, only scans from each participant's initial study visit are included. All scans are centred about the fovea and measure 1536 pixels wide and 496 pixels high (approximately 8.8 x 1.9 mm). For each scan, three boundary positions (in pixels) (annotated through manual segmentation by an expert human grader) are available. These include: the inner aspect (surface) of the inner limiting membrane (ILM), the outer aspect (base) of the retinal pigment epithelium (RPE) and the chorio-scleral interface (CSI). According to these boundary positions, area masks are constructed for each image. This is achieved by setting each pixel of the mask in each column to one of four classes: 1) vitreous (top of image to ILM surface), 2) retina (ILM surface to RPE base), 3) choroid (RPE base to CSI) or 4) sclera (CSI to bottom of image).

For computational reasons with regards to GAN training, we utilise down-sampled slices in place of the original full-size scans for this study. These slices are obtained through the following two steps: 1) 17 slices (496 x 496 pixels) are obtained from each full-size OCT scan by sampling overlapping slices using a stride of 64 pixels across the width of the scan, 2) Each slice is down-sampled (using nearest neighbour interpolation) to a size of 128 x 128 pixels. The boundary positions and masks are similarly sliced and down-sampled to match. An example of an OCT image slice, as well as the corresponding boundary positions and area mask (4 classes) are all illustrated in Fig. 1. The data was divided into 3 subsets including training, validation, and testing respectively with participants in each subset assigned randomly. Care was taken to ensure that slices from a single participant do not overlap between sets. See Table I for a detailed breakdown and summary of the three sets of images.

B. Semantic segmentation model and boundary delineation

To segment the OCT slices, we employed an encoder-decoder style semantic segmentation network based on U-Net [29] that has been presented in one of our previous studies for a similar application and has demonstrated competitive performance [12]. The OCT image slice is passed as the input

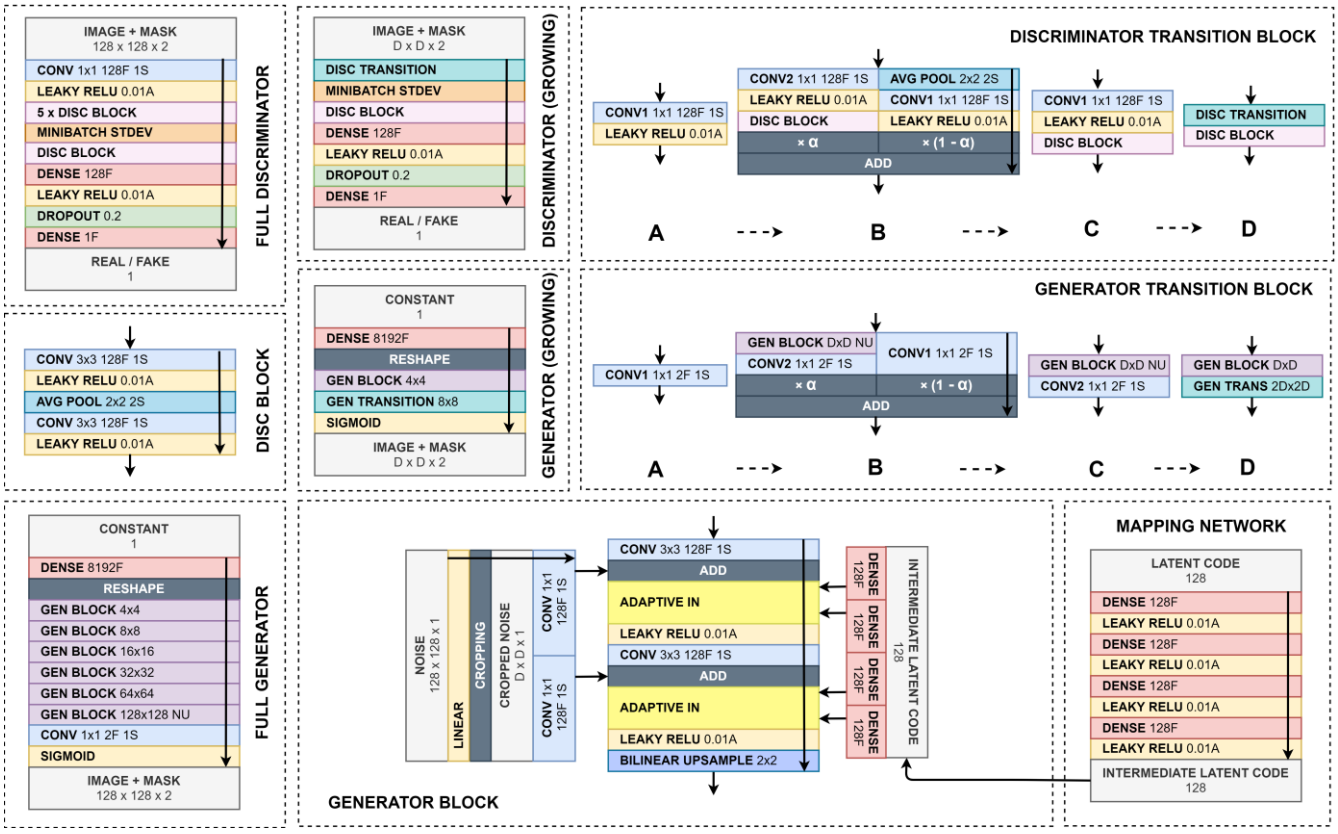

Fig. 3. Illustration of the GAN model used including details for the generator, discriminator and mapping network as well as the progressive growing strategy with each of the transition blocks showing the phases and the architecture progression for each resolution (A: initial, B: fade-in, C: consolidation, D: [if full resolution has not been reached] adding new layers at double resolution). #F: filters, #S: stride, #D image resolution, #A: alpha (leaky relu negative slope coefficient), NU: no upsampling, α: fade-in weight. Note that all convolutional layers throughout are padded such that the input spatial size is equal to the output spatial size. Arrows show direction of information flow. Light grey boxes correspond to inputs and outputs. Note that different latent codes may be used for individual generator blocks (i.e. for mixing regularisation).

and segmented, with the output area mask classifying each image pixel into one of the four areas: vitreous, retina, choroid or sclera. The performance of this network can be evaluated by computing the Dice coefficient between the predicted and true mask. A detailed summary of the architecture and layer parameters is provided in Fig. 2. Throughout this study, this network was trained using images paired with one-hot encoded masks in a consistent manner using the following setup: 1) Adam optimizer [30] with default parameters, 2) batch size of three, 3) minimising Dice loss, 4) training for 100 epochs while performing model selection based on the maximum Dice coefficient, 5) shuffling the order of presentation of samples for each new epoch.

The method for boundary delineation has been used in several previous studies [7-9,12] so only a brief description is presented here. Given an area mask prediction output from the segmentation network, the method for delineating boundary positions involves two steps. First, a boundary probability map is constructed for each boundary using the Sobel filter to detect the edge between each pair of adjacent areas in the predicted mask. Each probability map then forms the basis of a graph from which a shortest path graph search (utilising Dijkstra's algorithm) is performed across the width of the scan. This traversed path is taken as the predicted boundary location. The mean absolute error (MAE) between the predicted location and the ground truth can be used to gauge the accuracy for each boundary of interest.

C. Generative adversarial network model

For our GAN model, we employ a StyleGAN [6] inspired approach utilising the least squares GAN (LSGAN) [4] training regime. Here, a generator network and a discriminator network are trained in a cycle on a batch by batch basis. The generator is tasked to produce an image and mask output pair of size 128x128x2 pixels where the first channel contains the grayscale OCT image and the second channel contains the normalised (0-1) area mask. The discriminator is tasked to identify whether a provided image and mask pair is real or fake. Here, we utilise values of 1 and -1 for real and fake images, respectively. For generator training, the discriminator's weights are frozen, and the generator and discriminator are trained end-to-end with the generated images labelled as real and the goal of minimising the mean squared error at the output of the discriminator. For discriminator training, two batches, one containing real images, and the other containing fake images are provided to the discriminator. Once again, the goal is to minimise the mean squared error at the output of the discriminator where the images are labelled correctly (i.e. real: 1, fake: -1). Rather than directly providing a latent code (sampled from a unit Gaussian) to the generator, a "mapping" network is utilised to map the latent code to an intermediate latent space. This helps to encourage disentanglement of the latent space and prevent the generation of invalid samples. This intermediate latent code is then transformed to styles by learned affine transformations (using a pair of dense layers) which subsequently control the adaptive instance normalisation (AdaIN) [31] operations in the generator. Further stochastic variation of the generated images is introduced using a unit Gaussian noise input (128x128x1 pixels) which is cropped to match the resolution for each layer of the generator. Further details of the overall architecture of the discriminator, generator and mapping network are illustrated in Fig 3.

With the goal of accelerating training convergence and improving stability, we utilise a progressive growing of GANs (PGAN) [5] approach. Here, rather than training the entire network at the full 128x128 pixel resolution from the beginning, we instead start at a 4x4 resolution and gradually add new layers throughout the training process. Each set of new layers doubles the resolution of the image with these layers gradually integrated across a transition or "fade-in" period to prevent sudden shocks to the network weights. This transition is controlled by a parameter α (initialised to zero), which weights the contribution of the new layers accordingly and is linearly increased to a value of one across the period. The network then undergoes a "consolidation" phase at this resolution, which is then followed by one of the following two steps: 1) the next set of layers are added and subsequently faded-in or, 2) the network has already reached full-resolution and training is complete. Further architectural details associated with the progressive growing approach are demonstrated in Fig. 3. At each resolution level, images and masks are down-sampled as required using nearest neighbour interpolation. Note that the real images and masks during each fade-in period are taken as the linear interpolation between the relevant lower and higher resolution versions based on the parameter α, effectively mimicking the synthetic images output by the transitioning generator. The number of training steps for each fade-in and consolidation phase are summarised in Table II. In total, across all phases, we train each network for 355,000 steps. The Adam optimizer [30] (with a learning rate of 5×10^{-4}, decay of 1×10^{-5}, $\beta_1 = 0$ and $\beta_2 = 0.99$) was used for training in all cases. Note that this learning rate is reset at the beginning of each new phase. For regularisation purposes, a gradient penalty loss [32], with weight of 50 was utilised throughout training. To encourage styles to better localise to the individual resolution levels, we employed mixing regularisation [6] whereby two latent codes are provided to the network rather than one. During training, we alternate between using a single latent code and mixing regularisation for five batches each. A batch size of 16 is used throughout. To further encourage diversity of the generated images, we also made use of a minibatch standard deviation layer [5], which is inserted towards the end of the discriminator. We implemented all GAN models in Keras 2.2.4 with Tensorflow backend in Python 3.6.4 with our specific implementation inspired by the work in [33]. This model was utilised throughout this study to generate synthetic training datasets noting that we always keep the validation set fixed with real images.

III. Experiments and Results

A. Effect of network capacity

The capacity of the discriminator and generator networks are crucial for the network's ability to generate sufficiently diverse and realistic images. It can be noted that the OCT images used here exhibit greater levels of noise and a greater variety of finer resolution details than that of other datasets used with similar GAN methods (i.e. human faces or landscape photos). We started with a baseline network architecture which utilised (lowest resolution layer first): 16, 32, 64, 128, 192, and 256 filters respectively for each block of both the generator and the discriminator. We also utilise 128 filters for the final dense layer in the discriminator. We then performed an experiment by modifying the number of filters in the network. We considered five other network variants each with a uniform number of filters in each layer. These

TABLE III. EFFECT OF NETWORK CAPACITY. BASELINE NETWORK (16/32/64/128/192/256 FILTERS), THE REST HAVE UNIFORM NUMBER OF FILTERS.

Network capacity	Validation Dice [%]	ILM MAE (SD) [px]	RPE MAE (SD) [px]	CSI MAE (SD) [px]	Validation Dice inter-run SD [%]	ILM MAE inter-run SD [px]	RPE MAE inter-run SD [px]	CSI MAE inter-run SD [px]
256	**99.03**	**0.15** (**0.12**)	**0.16** (**0.15**)	**0.94** (**0.99**)	0.11	0.01	**0.01**	0.15
128	99.00	**0.15** (**0.12**)	**0.16** (0.19)	1.00 (1.01)	**0.04**	0.01	**0.01**	**0.04**
64	98.82	0.16 (0.14)	0.17 (0.19)	1.04 (1.07)	**0.04**	0.01	**0.01**	0.08
32	98.67	0.18 (0.20)	0.18 (0.23)	1.22 (1.28)	0.13	0.01	0.02	0.12
Baseline	98.30	0.18 (0.16)	0.19 (0.21)	1.41 (1.31)	0.21	**0.00**	0.03	0.27
16	98.05	0.19 (0.17)	0.21 (0.22)	1.83 (1.67)	0.16	0.03	0.03	0.41

variants included 16, 32, 64, 128 and 256 filters. For each variant as well as the baseline, we trained three GANs identically except for their initial weights that are randomly initialized. For each, we then generate a synthetic training dataset with the same size as the real training set. Each of the three datasets was then used to separately train three semantic segmentation networks (nine total) and subsequently perform boundary delineation. For all GANs we use a latent code of 128 dimensions. The median validation Dice overlap coefficient (%) as well as the median mean absolute boundary error results are summarised in Table III with the baseline network capacity variant denoted "Baseline". We also report the inter-run variability between each of the nine networks for each variant of network capacity. It is apparent that there are notable performance gains evident as the network capacity is increased compared to the baseline. Additionally, the variability between training runs generally decreases as the number of filters increases. A notable exception to this is the somewhat higher level of variability of the Dice coefficient and CSI boundary error using 256 filters rather than 128 for all layers. It should be noted that, although the boundary errors appear low, the differences between the methods (particularly for the CSI) are clinically significant for two reasons: 1) down-sampled images are used here which inherently decrease the magnitude of the differences and, 2) the temporal changes when analysing the layers for clinical purposes can be subtle (less than the differences between methods).

B. Effect of latent code dimensionality

Another parameter of interest is the dimensionality of the latent code. Indeed, effective values for this parameter are likely to be dataset and modality dependent. Here, we performed a simple comparison between 32, 128 and 512 dimensions. We used the 128-filter variant from the previous network capacity experiment. Once again, we trained three identical GANs for each, and then trained (using the synthetic data) a total of nine separate segmentation networks. The median and best validation Dice overlap coefficient (%) as well as the median and best mean absolute boundary errors are summarised in Table IV. There appears to be little overall difference between the median performance of all sizes with no clear trend as the size of latent code is increased or decreased. There are some differences to note, particularly the lowest boundary errors on the CSI, however it is likely that these are the result of training variability.

C. Synthetic vs. real

To evaluate the quality of the synthetic data for segmentation purposes we compare it to a baseline using real

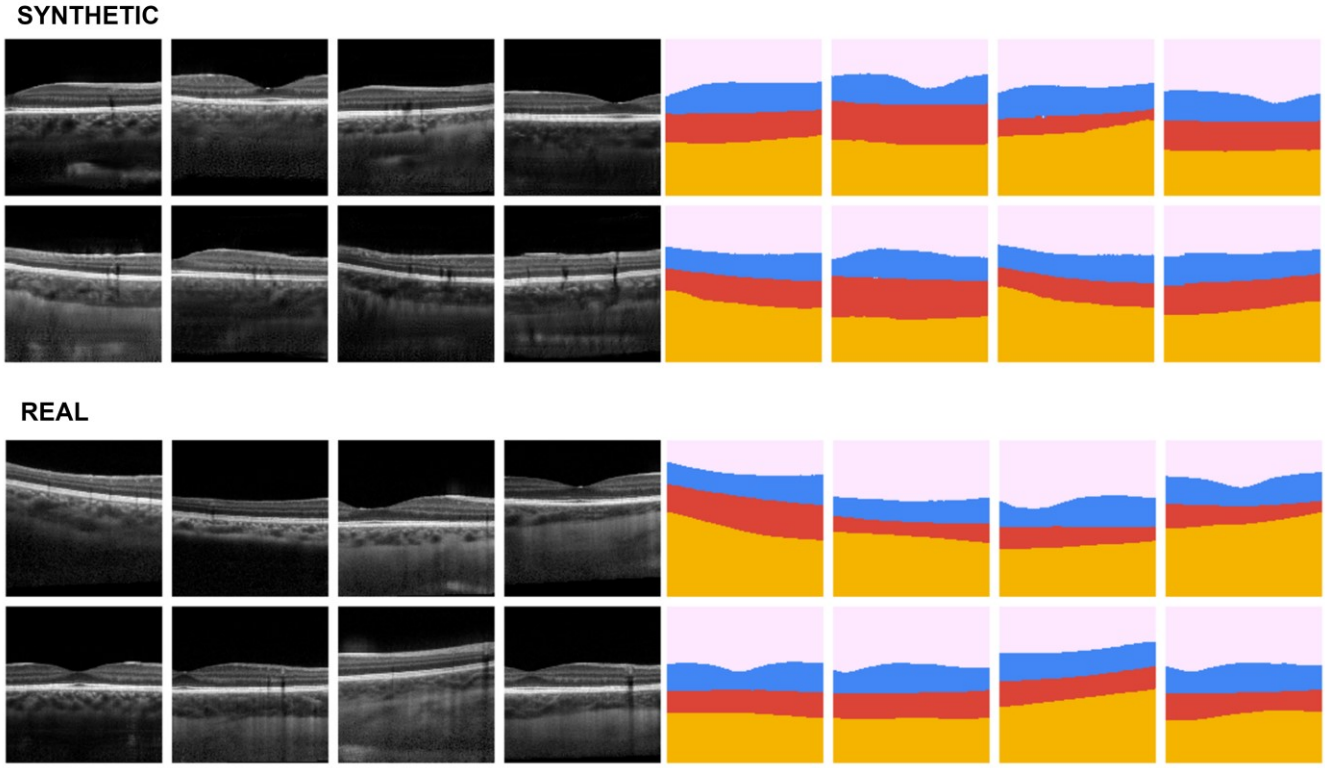

Fig. 4. A set of example real image slices and associated masks (top 2 rows) and a set of example synthetic image slices and masks (bottom 2 rows).

TABLE IV. EFFECT OF LATENT CODE DIMENSIONALITY.

# dimensions	Validation Dice [%]	ILM MAE (SD) [px]	RPE MAE (SD) [px]	CSI MAE (SD) [px]
512 (median)	98.98	0.16 (0.12)	0.16 (0.15)	0.97 (1.00)
512 (best)	99.01	0.14 (0.09)	0.15 (0.13)	0.85 (0.81)
128 (median)	99.00	0.15 (0.12)	0.16 (0.19)	1.00 (1.01)
128 (best)	99.03	0.14 (0.09)	0.15 (0.14)	0.96 (0.92)
32 (median)	98.99	0.15 (0.12)	0.16 (0.16)	0.95 (0.90)
32 (best)	99.07	0.15 (0.07)	0.15 (0.13)	0.89 (0.80)

TABLE V. REAL, SYNTHETIC AND COMBINED PERFORMANCE.

Training Data	Validation Dice [%]	ILM MAE (SD) [px]	RPE MAE (SD) [px]	CSI MAE (SD) [px]
Synthetic (median)	98.98	0.16 (0.12)	0.16 (0.15)	0.97 (1.00)
Real (median)	99.19	0.14 (0.08)	0.14 (0.13)	0.81 (0.79)
Combined (median)	99.19	0.15 (0.08)	0.14 (0.13)	0.80 (0.67)
Synthetic (best)	99.01	0.14 (0.09)	0.15 (0.13)	0.85 (0.81)
Real (best)	99.19	0.14 (0.08)	0.14 (0.13)	0.74 (0.61)
Combined (best)	99.20	0.14 (0.08)	0.14 (0.11)	0.73 (0.63)

data. We utilise the GAN model with 512 latent dimensions and 128 filters used in the previous experiments. Using the training dataset containing real images, we trained three segmentation networks and performed boundary delineation. In the interest of assessing the suitability of this technique for data augmentation, the performance of the method can also be analysed when combining the real and synthetic datasets. For this, we trained three additional segmentation networks using the combination of real and synthetic data (one for each trained GAN). The results for the median and best validation Dice coefficient (%) as well as the median and best mean absolute boundary errors are summarised in Table V. A visual comparison between example real and synthetic image and masks pairs are provided in Fig. 4. These examples suggest that the synthetic images and masks exhibit a promising level of realism and diversity in line with the real images. We can also visually inspect the nature of the latent space from which the images are generated. This is performed by linearly interpolating between two randomly chosen latent codes and generating images for each latent code along this path. Two such examples are provided in Fig. 5. Part of the motivation of using a StyleGAN based approach relates to the ability to control the style of the images at different resolution levels. Therefore, we can visualise the effect of varying the latent code that is input to some layers of the generator while keeping the latent code for others fixed. An example is provided in Fig. 6 by illustrating this effect by separating the six layers into two groups: 1) the lowest resolution layer, and 2) the five other higher resolution layers.

IV. DISCUSSION AND CONCLUSION

In this study, GAN-based synthesis of chorio-retinal OCT image and segmentation mask pairs has been demonstrated along with the use of such data for semantic segmentation methods. Visual inspection shows that the synthetic images exhibit a high degree of realism and diversity, comparable to the real data. This is further justified by using such data to directly train a machine learning-based semantic segmentation method. The results of the Dice overlap coefficient (%) and mean absolute boundary error (px) are very encouraging with the synthetic performance nearly comparable to that of the real data. For the synthetic data alone, the median mean absolute error on the three boundaries were 0.16, 0.16, and 0.97 pixels respectively compared to 0.14, 0.14, and 0.81 pixels respectively for the real data. The median validation Dice coefficient (%) was 98.98 and 99.19 for the synthetic and real data, respectively. Importantly, combining the real and synthetic data results in no measurable loss in performance, an encouraging sign for applying this method as a data augmentation technique in the future. Indeed, demonstrating that performance does not deteriorate when combining the two datasets is a crucial step as it indicates that the generated images and corresponding masks are sufficiently realistic and do not inherently contribute error to the training process.

Exploring and visualising the latent space through interpolation demonstrates that the latent space exhibits two key properties: 1) Completeness: generated samples are meaningful and not erroneous, and 2) Continuity: samples close together in the latent space do not differ significantly in appearance. In the examples in Fig. 5, these properties are clearly observable as the images undergo a gradual transition along the interpolated path. In the first example, it is a particularly interesting and promising result to observe how the foveal pit position in adjacent slices undergoes a gradual even translation across the width of the image. Similarly, the second example illustrates the foveal pit translation towards the left across the first half of the interpolation before disappearing off the edge of the image. The second half of this interpolation path shows a different transition with an increase in curvature to the global structure of the image. Changes in other aspects of the image such as the choroidal structure and shadow intensity and position can also be observed in these examples.

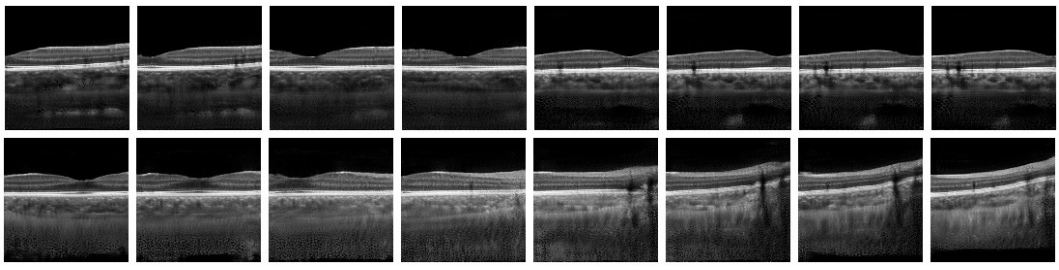

Fig. 5. Two example sets of image from interpolating through the latent space. Example 1 (top row) and example 2 (bottom row) illustrate the changes in the image as the latent code gradually changes.

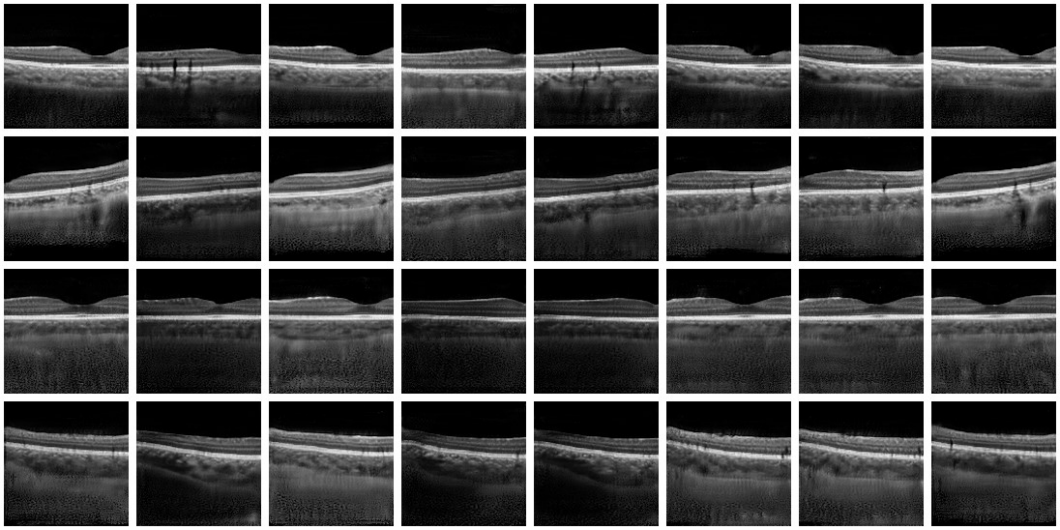

Fig. 6. Visualisation of style localisation. Horizontal direction depicts changes in the latent code input for the 5 highest resolution layers of the generator (while keeping the lowest resolution layer fixed) whereas the vertical direction depicts changes in the latent code for the lowest resolution generator layer.

The GAN model and overall method used possess several parameters that can affect performance. In this study, we examined the effect of two of these. First, we identified the network capacity to be a potential performance bottleneck. This was a particular risk given the type of images used here as these contain significantly higher noise levels and higher frequency details compared to other datasets (e.g. human face or landscape photos) used with a similar setup. Therefore, the importance of network capacity should not be overlooked. Compared to a baseline capacity, significant performance improvements were realised when increasing the number of filters within both the generator and the discriminator. The selected network capacity, utilising 128 filters in all blocks of the generator and discriminator provides a 0.7% Dice coefficient improvement compared to the baseline while the mean absolute boundary errors and standard deviations also show notable improvements. We note that allocating the filters across the layers in a non-uniform and more optimal way (i.e. not using an identical number for each) is likely to help reduce computational complexity and training time.

The dimensionality of the latent code was the other parameter investigated in this study. For image generation, it is necessary to have a latent code that is sufficiently large enough to capture the diversity of features across the dataset. A good choice for this parameter is likely to be dataset and modality dependent so an experiment was performed to investigate the effect of decreasing the number of dimensions of the latent code compared to the baseline (decreasing from 512 to 128 to 32 dimensions). The results suggested that there is little difference in performance between the three latent code sizes. With the lack of a clear trend, we note that variability in the training process is the most likely explanation for any observed differences. However, such an outcome may differ for other datasets where a higher level of dimensionality in the latent space may be required to fully capture the diversity.

A StyleGAN based approach was selected here not only for its documented capacity to generate realistic and diverse images but also for the ability to control the image style at specific resolutions on a layer-by-layer basis. This functionality may be particularly valuable for a range of OCT image generation tasks in cases where only particular levels of detail need to be changed while fixing the style at other levels. Indeed, generating a range of images with differing high frequency detail and structure while retaining a fixed global structure (i.e. overall shape and curvature of the retinal tissue) may be desired. Fig. 6 demonstrates the ability of the GAN in this situation, maintaining the overall global structure by fixing the latent code to the lowest resolution layer (4x4) of the generator. In the horizontal direction, there is a high level of diversity with differences in contrast, shadow intensity/position, choroidal and internal retinal layer structure as well as noise and other artifacts. In the vertical direction, it is evident that modifying the latent code of the lowest resolution layer does indeed alter the overall global structure as expected. One limitation we observe is that there does appear to be a small amount of feature entanglement, a common problem with GANs. A clear example of this relates to the shadow positions across images. For instance, the second column of Fig. 6 shows several shadows in the image in the first row with almost no shadows in the corresponding image in the second row after modifying the lowest resolution layer's latent code. This indicates that the style associated with these shadows is somewhat entangled with the overall global structure. In future, more precise control and a greater level of image diversity could be possible with a higher level of feature disentanglement.

Due to the computational complexity of these methods and in the interest of running a range of experiments, the method here only operates on sliced, down-sampled versions of the OCT images (128x128 pixels). Future work should investigate the extensibility of this method to larger resolutions. However, the results demonstrate encouraging performance for this application and the findings presented here inherently allow for future expansion of this method to high-resolution OCT images. Indeed, the motivation to use a progressive growing GAN (PGAN) approach was two-fold: 1) The problem of image generation is divided up and becomes simpler to solve resulting in faster convergence and overall shorter training times, and 2) This approach can be easily scaled up to support larger resolution images without any further modifications to the training process.

The segmentation performance combining the real and synthetic data was not inhibited, indicating that the generated data is sufficiently realistic, diverse and free of error. However, only marginal improvements in performance were

observed when combining the data. Future work may extend this and investigate this application for data augmentation purposes where improvements to the segmentation performance are likely to be achievable by further refining the method. Analysing this method for data augmentation using different datasets may also prove beneficial, particularly for datasets that exhibit a lower degree of homogeneity (e.g. pathological or sparse datasets). Such datasets are likely to be missing a larger number of intermediate modes that may be more readily generated by a GAN which may be more useful from a data augmentation perspective. Additionally, although the generated segmentation masks exhibit a high level of quality, another interesting avenue to investigate may involve a method to constrain these outputs from a layer topology perspective (i.e. removal of any possible holes/artifacts and constraining the vertical ordering of the layers to be correct), to further improve performance.

Overall, these results are very encouraging and demonstrate the application of GANs to generate synthetic data for semantic segmentation in OCT images. Such data may then be used to train a deep learning segmentation method and subsequently obtain boundary positions for chorio-retinal layer boundaries. By generating a combined image and segmentation mask pair as the output, we remove a constraint on the generated images while also allowing for previously unseen masks to be produced. The findings presented here are likely to be beneficial for future work involving synthetic OCT data generation. In particular, we provide a strong basis for the further development of OCT data augmentation methods using GANs.

ACKNOWLEDGMENT

Computational resources and services used in this work were provided in part by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia. We gratefully acknowledge support from the NVIDIA Corporation for the donation of GPUs used in this work.

REFERENCES

[1] I. J. Goodfellow et al. “Generative Adversarial Networks,” arXiv:1406.2661, 2014.

[2] M. Arjovksy, S. Chintala, and L. Bottou, “Wasserstein GAN,” arXiv:1701.07875, 2016.

[3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of Wasserstein GANs,” arXiv:1704.00028, 2017.

[4] X. Mao et al. “Least Squares Generative Adversarial Networks,” arXiv:1611.04076, 2016.

[5] T. Karras, T. Aila, S. Laine, J. Lehtinen. “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” arXiv:1710.10196, 2017.

[6] T. Karras, S. Laine, and T. Aila. “A style-based generator architecture for generative adversarial networks,” arXiv:1812.04948, 2018.

[7] L. Fang et al. “Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search,” Biomed. Opt. Express, vol. 8, no. 5, pp. 2732–2744, 2017.

[8] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,” Biomed. Opt. Express, vol. 9, no. 7, pp. 3049–3066, 2018.

[9] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent & M. J. Collins, “Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search,” Biomed. Opt. Express, vol. 9, no. 11, pp. 5759-5777, 2018.

[10] A. G. Roy et al. “ReLayNet: Retinal Layer and Fluid Segmentation of Macular Optical Coherence Tomography using Fully Convolutional Networks,” Biomed. Opt. Express, vol. 8, no. 8, pp. 3627–3642, 2017.

[11] F. G. Venhuizen et al. “Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks,” Biomed. Opt. Expres, vol. 8, no. 7, pp. 3292–3316, 2017.

[12] J. Kugelman et al. “Automatic choroidal segmentation in OCT images using supervised deep learning methods,” Sci. Rep., vol. 9, 13298, 2019.

[13] J. Wei et al. “Generative Image Translation for Data Augmentation in Colorectal Histopathology Images,” arXiv:1910.05827, 2019.

[14] V. Sandfort, K. Yan, P. J. Pickhardt and R. M, Summers. “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” Sci. Rep., vol. 9, 16884, 2019.

[15] M. Frid-Adar, E. Klang, M. Amitai, J Goldberger, and H. Greenspan. “Synthetic data augmentation using GAN for improved liver lesion classification,” 15th International Symposium on Biomedical Imaging (ISBI), pp. 289-293, IEEE, 2018.

[16] S. Yamaguchi, S. Kanai, and T. Eda. “Effective Data Augmentation with Multi-Domain Learning GANs,” arXiv:1912.11597, 2019.

[17] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks,” Med. Image Anal., vol. 54, pp. 30-44, 2019.

[18] P. Seeböck et al. “Using cyclegans for effectively reducing image variability across oct devices and improving retinal fluid segmentation,” 16th International Symposium on Biomedical Imaging (ISBI), pp. 605-609, IEEE, 2019.

[19] Cheong et al. “Deshadow GAN: A deep learning approach to remove shadows from optical coherence tomography images,” Transl. Vis. Sci. Technol., vol. 9, no. 2, 23, 2020.

[20] K. J. Halupka et al. “Retinal optical coherence tomography image enhancement via deep learning,” Biomed. Opt. Express, vol. 9, no. 12, pp. 6205-6221, 2018.

[21] Y. Huang et al. “Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network,” Opt. Express, vol. 27, no. 9, pp. 12289-12307, 2019.

[22] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks,” International Conference on Computer Vision, pp. 2223-2232, IEEE, 2017,

[23] D. Romo-Bucheli et al. “Reducing image variability across OCT devices with unsupervised unpaired learning for improved segmentation of the retina,” Biomed. Opt. Express, vol. 11, no. 1, pp. 346-363, 2020.

[24] Ma et al. “Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN,” Biomed. Opt. Express, vol. 9, no. 11, pp. 5129-5146, 2018.

[25] J. Kugelman et al. “Constructing synthetic chorio-retinal patches using generative adversarial networks,” 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8, IEEE, 2019.

[26] C. Zheng et al. “Assessment of Generative Adversarial Networks Model for Synthetic Optical Coherence Tomography Imags of Retinal Disorders,” Transl. Vis. Sci. Technol., vol. 9, no. 2, 29, 2020.

[27] S. A. Read, M. J. Collins, S. J. Vincent, and D. Alonso-Caneiro, “Choroidal thickness in childhood,” Invest. Ophthamol. Vis. Sci., vol. 54, no. 5, pp. 3586-3593, 2013.

[28] S. A. Read, M. J. Collins, S. J. Vincent, and D. Alonso-Caneiro, “Macular retinal layer thickness in childhood,” Retina, vol. 35, pp. 1223-1233, 2015.

[29] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation,” arXiv:1505.04597, 2015.

[30] D. P. Kingma and J. Ba. “Adam: a method for stochastic optimization,” arXiv:1412.6980, 2014.

[31] X. Huang and S. Belongie. “Arbitrary style transfer in real-time with adapative instance normalization,” arXiv:1703.06868, 2017.

[32] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. “Stabilizing training of generative adversarial networks through regularization,” arXiv:1705.09367, 2017.

[33] StyleGAN-Keras. [Online] Available at: https://github.com/manicman1999/StyleGAN-Keras, [Accessed 15 Jul. 2020].