Pathological Retinal Region Segmentation From OCT Images Using Geometric Relation Based Augmentation

Dwarikanath Mahapatra[1]   Behzad Bozorgtabar[2,3,4]

Ling Shao[1]

[1] IIAI, Abu Dhabi   [2] LTS5, EPFL, Lausanne   [3] CIBM, Laussane   [4] CHUV, Lausanne

{dwarikanath.mahapatra, ling.shao}@inceptioniai.org   behzad.bozorgtabar@epfl.ch

Abstract

Medical image segmentation is an important task for computer aided diagnosis. Pixelwise manual annotations of large datasets require high expertise and is time consuming. Conventional data augmentations have limited benefit by not fully representing the underlying distribution of the training set, thus affecting model robustness when tested on images captured from different sources. Prior work leverages synthetic images for data augmentation ignoring the interleaved geometric relationship between different anatomical labels. We propose improvements over previous GAN-based medical image synthesis methods by jointly encoding the intrinsic relationship of geometry and shape. Latent space variable sampling results in diverse generated images from a base image and improves robustness. Given those augmented images generated by our method, we train the segmentation network to enhance the segmentation performance of retinal optical coherence tomography (OCT) images. The proposed method outperforms state-of-the-art segmentation methods on the public RETOUCH dataset having images captured from different acquisition procedures. Ablation studies and visual analysis also demonstrate benefits of integrating geometry and diversity.

1. Introduction

Medical image segmentation is an important task for healthcare applications like disease diagnosis, surgical planning, and disease progression monitoring. While deep learning (DL) methods demonstrate state-of-the-art results for medical image analysis tasks [38], their robustness depends upon the availability of a diverse training dataset to learn different disease attributes such as appearance and shape characteristics. Large scale dataset annotations for segmentation require image pixel labels, which is time consuming and involves high degree of clinical expertise. The problem is particularly acute for pathological images since
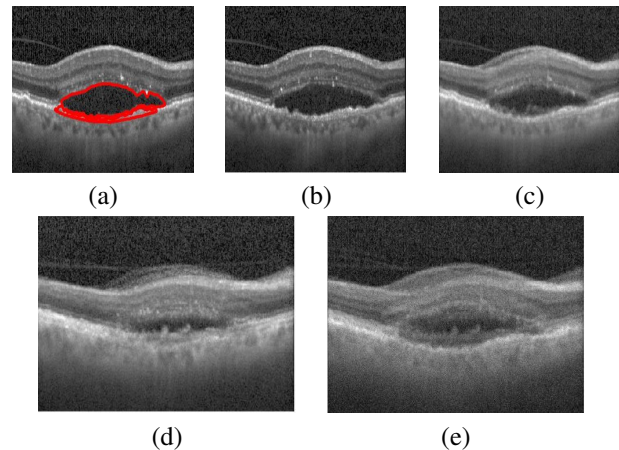

Figure 1. (a) Base image (red contour denotes segmentation mask); Example of generated images using: (b) Our proposed $GeoGAN$ method; (c) Zhao et al. [42]; (d) $DAGAN$ method by [1]; (e) $cGAN$ method by [25].

it is difficult to obtain diverse images for less prevalent disease conditions, necessitating data augmentation. We propose a generative adversarial network (GAN) based approach for pathological images augmentation and demonstrate its efficacy in pathological region segmentation. Figure 1 summarizes the image generation results of our approach and [42], and highlights our superior performance by incorporating geometric information.

Traditional augmentations such as image rotations or deformations have limited benefit as they do not fully represent the underlying data distribution of the training set and are sensitive to parameter choices. Recent works [15, 42, 14, 30] proposed to solve this issue by using synthetic data for augmentation and increase diversity in the training samples. However, certain challenges have not been satisfactorily addressed by these methods.

Zhao et. al. [42] proposed a learning-based registration method to register images to an atlas, use corresponding deformation field to deform a segmentation mask and obtain
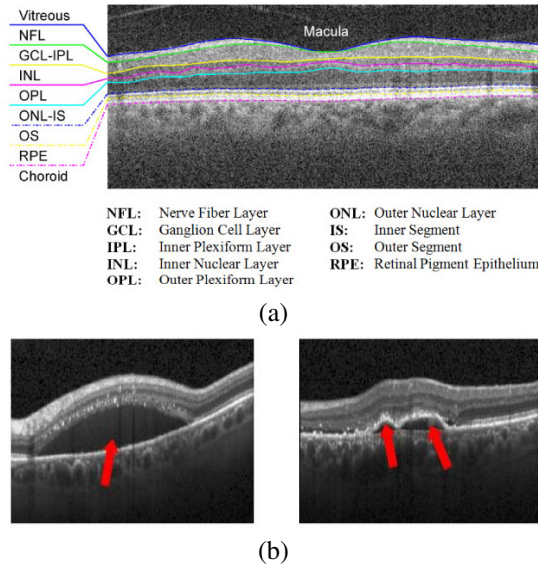
Figure 2. Example of normal and fluid filled OCT images: (a) example control subject image without any abnormalities (taken from [9]); (b) images with accumulated fluid up due to diabetic macular edema and AMD from our dataset. The fluid areas are marked with red arrows.

new data. This approach presents the following challenges: 1) since registration errors propagate to subsequent stages, inaccurate registration can adversely affect the data generation process; 2) with atlas of a normal subject it is challenging to register images from diseased subjects due to appearance or shape changes. This is particularly relevant for layer segmentation in retinal optical coherence tomography (OCT) images, where there is a drastic difference in layer shape between normal and diseased cases. Figure 2 (a) shows the retinal layers of a normal subject, and Figure 2 (b) shows two cases of retinal fluid build up due to diabetic macular edema (DME) and age related macular degeneration (AMD). The retinal layers are severely distorted compared to Figure 2 (a) and registration approaches have limited impact in generating accurate images.

Recent methods for data augmentation [14, 30, 6, 7] using a generative adversarial network (GAN) [13] have shown moderate success for medical image classification. However, they have limited relevance for segmentation since they do not model geometric relation between different organs and most augmentation approaches do not differentiate between normal and diseased samples. Experiments in Section 4.5 show segmentation methods trained on normal subject images (Figure 2 (a)) are not equally effective for diseased cases due to significant shape changes between the two types. Hence there is a need for augmentation methods that consider the geometric relation between different anatomical regions and generate distinct images for diseased and normal cases. Another limitation of current augmentation approaches is that they do not incorporate diversity in a principled manner. In [25] shape mask was incorporated manually for image generation, which is not practical and may lead to unrealistic deformations.

2. Related Work

2.1. Deep Models for Retinal OCT Segmentation

One of the first works to use multi-scale convolutional neural nets (CNNs) on OCT images [35] employed patch-based voxel classification for detecting intraretinal fluid (IRF) and subretinal fluid (SRF) in fully supervised and weakly supervised settings. Fully convolutional neural nets and U-nets were used in [39, 12] to segment IRF, and in [33] to segment both the retinal layers and the fluid. Explicit fluid segmentation methods such as [40] also achieve high classification performance.

2.2. Data Augmentation (DA)

While conventional augmentation approaches are easy to implement and generate a large database, their capabilities are limited in inducing data diversity. They are also sensitive to parameter values [11], variation in image resolution, appearance and quality [22].

Recent DL based methods trained with synthetic images outperform those trained with standard DA over classification and segmentation tasks. Antoniou et al. [1] proposed DAGAN for image generation in few shot learning systems. Bozorgtabar et al. [8] used GAN objective for domain transformation by aligning feature distribution of target data and source domain. Mahapatra et al. [25] used conditional GAN (cGAN) for generating informative synthetic chest Xray images conditioned on a perturbed input mask. GANs have also been used for generating synthetic retinal images [43] and brain magnetic resonance images (MRI) [14, 36], facial expression analysis [5], for super resolution [21, 24], image registration [27, 26, 28] and generating higher strength MRI from their low strength acquisition counterparts [41]. Generated images have implicit variations in intensity distribution but there is no explicit attempt to model attributes such as shape variations that are important to capture different conditions across a population. Milletari et al. [29] augmented medical images with simulated anatomical variations but demonstrate varying performance based on transformation functions and parameter settings.

2.3. Image Generation Using Uncertainty

Kendall et al. [17] used approximate Bayesian inference for parameter uncertainty estimation in scene understanding, but did not capture complex correlations between different labels. Lakshminarayanan et al. [20] proposed a method to generate different samples using an ensemble of M networks while Rupprecht et al. [34] presented a sin-

gle network with M heads for image generation. Sohn et al. [37] proposed a method based on conditional variational autoencoders (cVAE) to model segmentation masks, which improves the quality of generated images. In probabilistic UNet [19], cVAE is combined with UNet [32] to generate multiple segmentation masks, although with limited diversity since randomness is introduced at highest resolution only. Baumgartner et al. [2] introduced a framework to generate images with a greater diversity by injecting randomness at multiple levels.

2.4. Our Contribution

Based on the premise that improved data augmentation yields better segmentation performance in a DL system, we hypothesize that improved generation of synthetic images is possible by considering the intrinsic relationships between shape and geometry of anatomical structures [4]. In this paper we present a Geometry-Aware Shape Generative Adversarial Network (GeoGAN) that learns to generate plausible images of the desired anatomy (e.g., retinal OCT images) while preserving learned relationships between geometry and shape. We make the following contributions:

1. Incorporating geometry information contributes to generation of realistic and qualitatively different medical images and **shapes** compared to standard DA. Other works such as [25, 43] do not incorporate this geometric relationship between anatomical parts.

2. Use of uncertainty sampling and conditional shape generation on class labels to introduce diversity in the mask generation process. Compared to previous methods we introduce diversity at different stages (different from [25, 43, 19]) and introduce an auxiliary classifier (different from [2, 37]) for improving the quality and accuracy of generated images.

3. Method

Our augmentation method: 1) models geometric relationship between multiple segmentation labels; 2) preserves disease class label of original image to learn disease specific appearance and shape characteristics; and 3) introduces diversity in the image generation process through uncertainty sampling. Figure 3 shows the training workflow using a modified UNet based generator network. The set of images and segmentation masks are used to train the generator while the discriminator provides feedback to improve the generator output. Figure 4 depicts generation of synthetic images from the validation image set and their subsequent use in training a UNet for image segmentation at test time.

3.1. Geometry Aware Shape Generation

Let us denote an input image as x, the corresponding manual segmentation masks as s_x and the disease class label of x as l_x. Our method learns to generate a new image and segmentation label map from a base image and its corresponding manual mask. The first stage is a spatial transformer network (STN) [16] that transforms the base mask to a new shape with different attributes of location, scale and orientation. The transformations used to obtain new segmentation mask $s_x^{'}$ are applied to x to get corresponding transformed image $x^{'}$. Since the primary aim of our approach is to learn contours and other shape specific information of anatomical regions, a modified UNet architecture as the generator network effectively captures hierarchical information of shapes. It also makes it easier to introduce diversity at different levels of image abstraction.

The generator $\mathbf{G}_g$ takes input $\mathbf{s}_x$ and a desired label vector of output mask c_g to output an affine transformation matrix $\mathbf{A}$ via a STN, i.e., $\mathbf{G}_g(\mathbf{s}_x, c_g) = \mathbf{A}$. $\mathbf{A}$ is used to generate $s_x^{'}$ and $x^{'}$. The discriminator $\mathbf{D}_{class}$ determines whether output image preserves the desired label c_g or not. The discriminator $\mathbf{D}_g$ is tasked with ensuring that the generated masks and images are realistic. Let the minimax criteria between $\mathbf{G}_g$ and $\mathbf{D}_g$ be $\min_{\mathbf{G}_g} \max_{\mathbf{D}_g} \mathbf{L}_g(\mathbf{G}_g, \mathbf{D}_g)$. The loss function $\mathbf{L}_g$ has three components

$$L_g = L_{adv} + \lambda_1 L_{class} + \lambda_2 L_{shape} \tag{1}$$

where 1) $\mathbf{L}_{adv}$ is an adversarial loss to ensure $\mathbf{G}_g$ outputs realistic deformations; 2) $\mathbf{L}_{class}$ ensures generated image has characteristics of the target output class label (disease or normal); and 3) $\mathbf{L}_{shape}$ ensures new masks have realistic shapes. λ_1, λ_2 balance each term's contribution.

Adversarial loss - $\mathbf{L}_{adv}(\mathbf{G}_g, \mathbf{D}_g)$: The STN outputs $\widetilde{A}$, a prediction for $\mathbf{A}$ conditioned on $\mathbf{s}_x$ and a new semantic map $\mathbf{s}_x \oplus \widetilde{A}(\mathbf{s}_x)$ is generated. $\mathbf{L}_{adv}$ is defined as:

$$L_{adv}(G_g, D_g) = \mathbb{E}_x \left[\log D_g(\mathbf{s}_x \oplus \widetilde{A}(\mathbf{s}_x)) \right]$$
$$+ \mathbb{E}_{\mathbf{s}_x} \left[\log(1 - D_g(\mathbf{s}_x \oplus \widetilde{A}(\mathbf{s}_x))) \right], \tag{2}$$

Classification Loss - $\mathbf{L}_{class}$: The affine transformation $\mathbf{A}$ is applied to the base image $\mathbf{x}$ to obtain the generated image $\mathbf{x}^{'}$. We add an auxiliary classifier when optimizing both $\mathbf{G}_g$ and $\mathbf{D}_g$ and define the classification loss as,

$$L_{class} = \mathbb{E}_{\mathbf{x}', c_g}[-\log D_{class}(c_g | x')], \tag{3}$$

where the term $D_{class}(c_g | x')$ represents a probability distribution over classification labels computed by D.

Shape Loss -$\mathbf{L}_{shape}$: We intend to preserve the relative geometric arrangement between the different labels. The generated mask has regions with different assigned segmentation labels because the base mask (from which the image
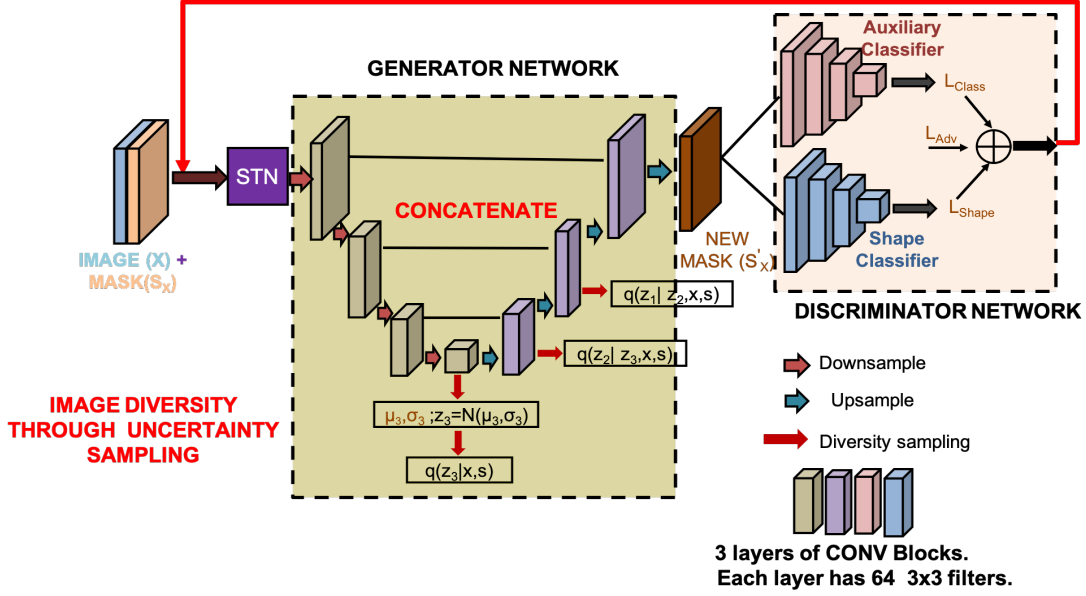
Figure 3. Overview of the steps in the training stage of our method. The images (X) and corresponding segmentation masks (S_X) are input to a STN whose output is fed to the generator network. Generator network is based on UNet architecture, and diversity through uncertainty sampling is injected at different levels. The generated mask S'_X is fed to the discriminator which evaluates its accuracy based on L_{class}, L_{shape} and L_{adv}. The provided feedback is used for weight updates to obtain the final model.
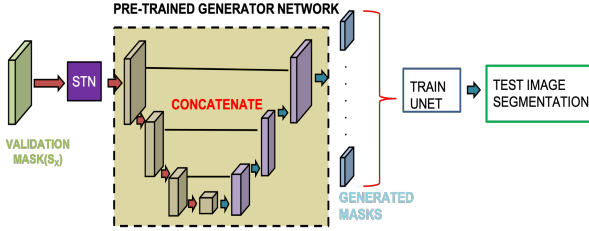

Figure 4. Depiction of mask generation. The trained generator network is used on validation set base images to generate new images that are used to train a segmentation network (UNet or Dense UNet). The model then segments retinal layers from test images.

was generated) already has labeled layers. Let us denote by s_i the image region (or pixels) in $\mathbf{s}_x$ assigned label i. Consider another set of pixels, s_j, assigned label j. We calculate $P_{shape}(l_i|s_j, s_i)$, which is, given regions s_i, s_j, the pairwise probability of s_i being label i. If n denotes the total number of labels, for every label i we calculate the $n-1$ such probability values and repeat it for all n labels. Thus

$$L_{shape} = \frac{1}{n \times (n-1)} \sum_{i,j}^{i \neq j} P_{shape}; \ (i,j) \in \{1, \cdots, n\}$$

(4)

The probability value is determined from a pre-trained modified VGG16 architecture to compute L_{shape} where the input has two separate maps corresponding to the label pair. Each map's foreground has only the region of the corresponding label and other labels considered background.

The conditional probability between the pair of label maps enables the classifier to implicitly capture geometrical relationships and volume information without the need to define explicit features. The geometric relation between different layers will vary for disease and normal cases, which is effectively captured by our approach.

3.2. Sample Diversity From Uncertainty Sampling

The generated mask s'_x is obtained by fusing L levels of the generator G_g (as shown in Figure 3), each of which is associated with a latent variable z_l. We use probabilistic uncertainty sampling to model conditional distribution of segmentation masks and use separate latent variables at multi-resolutions to factor inherent uncertainties. The hierarchical approach introduces diversity at different stages and influences different features (e.g., low level features at the early layers and abstract features in the later layers). Denoting the generated mask as s for simplicity, we obtain conditional distribution $p(s|x)$ for L latent levels as:

$$p(s|x) = \int p(s|z_1, \cdots, z_L) p(z_1|z_2, x) \cdots$$
$$p(z_{L1}|z_L, x) p(z_L|x) dz_1 \cdots dz_L.$$

(5)

Latent variable z_l models diversity at resolution 2^{-l+1} of the original image (e.g. z_1 and z_3 denote the original and $1/4$ image resolution). A variational approximation $q(z|s, x)$ approximates the posterior distribution $p(z|s, x)$ where $z = \{z_1, ..., z_L\}$. $\log p(s|x) = L(s|x) +$

$KL(q(z|s,x)||p(z|s,x))$, where L is the evidence lower bound, and $KL(.,.)$ is the Kullback-Leibler divergence. The prior and posterior distributions are parameterized as normal distributions $\mathcal{N}(z|\mu, \sigma)$.

Figure 3 shows example implementation for $L = 3$. We use 6 resolution levels and $L = 4$ latent levels. Figure 3 shows the latent variables z_l forming skip connections in a UNet architecture such that information between the image and segmentation output goes through a sampling step. The latent variables *are not mapped* to a 1-D vector to preserve the structural relationship between them, and this substantially improves segmentation accuracy. z_l's dimensionality is $r_x 2^{-l+1} \times r_y 2^{-l+1}$, where r_x, r_y are image dimensions.

4. Experimental Results

4.1. Dataset Description

We apply our method to OCT images since retinal disease leads to significant change of retinal layers, while changes due to disease in other modalities, such as Xray or MRI, are not so obvious for mildly severe cases. Moreover, in retinal OCT there is greater interaction between different layers (segmentation labels) which is a good use case to demonstrate the effectiveness of our attempt to model the geometric relation between different anatomical regions. The publicly available RETOUCH challenge dataset [3] is used for our experiments. It has images of the following pathologies: 1) **Intraretinal Fluid (IRF)**: contiguous fluid-filled spaces containing columns of tissue; 2) **Subretinal Fluid (SRF)**: accumulation of a clear or lipid-rich exudate in the subretinal space; 3) **Pigment Epithelial Detachment (PED)**: detachment of the retinal pigment epithelium (RPE) along with the overlying retina from the remaining Bruchs membrane (BM) due to the accumulation of fluid or material in sub-RPE space. It is common for age related macular degeneration (AMD).

OCT volumes were acquired with spectral-domain SD-OCT devices from three different vendors: Cirrus HD-OCT (Zeiss Meditec), Spectralis (Heidelberg Engineering), and T-1000/T-2000 (Topcon). There were 38 pathological OCT volumes from each vendor. Each Cirrus OCT consists of 128 B-scans of 512×1024 pixels. Each Spectralis OCT had 49 B-scans with 512×496 pixels and each Topcon OCT has 128 B-scans of 512×885 (T-2000) or 512×650 (T-1000) pixels. All OCT volumes cover a macular area of 6×6 mm^2 with axial resolutions of: 2μm (Cirrus), 3.9μm (Spectralis), and $2.6/3.5\mu$m (Topcon T-2000/T-1000). We use an additional dataset of 35 normal subjects derived equally $(12, 12, 11)$ from the three device types who had no incidence of retinal disease. The training set consists of 90 OCT volumes, with $24, 24$, and 22 diseased volumes acquired with Cirrus, Spectralis, and Topcon, respectively, with an extra 20 normal subjects $(7, 7, 6$ from each device). The

test set has 57 volumes, 14 diseased volumes from each device vendor and 15 normal subjects (5 from each device type). The distribution of different fluid pathologies (IRF, SRF, PED) and diseases (AMD, RVO) is almost equal in the training and test set.

The total number of images are as follows: 9071 training images (2D scans of the volume) - 7064 diseased and 2007 normal; 5795 test images- 4270 diseased and 1525 normal. Segmentation layers and fluid regions (in pathological images) were manually annotated in each of the $(9071 + 5795 = 14866)$ B-scans. Manual annotations were performed by 4 graders and the final annotation was based on consensus.

4.2. Experimental Setup, Baselines and Metrics

Our method has the following steps: 1) Split the dataset into training (60%), validation (20%), and test (20%) folds such that images of any patient are in one fold only. 2) Use training images to train the image generator. 3) Generate shapes from the validation set and train UNet segmentation network [32] on the generated images. 4) Use trained UNet to segment test images. 5) Repeat the above steps for different data augmentation methods. We trained all models using Adam optimiser [18] with a learning rate of 10^{-3} and batch-size of 12. Batch-normalisation was used. The values of parameters λ_1 and λ_2 in Eqn. 1 were set by a detailed grid search on a separate dataset of 18 volumes (6 from each device) that was not used for training or testing. They were varied between $[0, 1]$ in steps of 0.05 by fixing λ_1 and varying λ_2 for the whole range. This was repeated for all values of λ_1. The best segmentation accuracy was obtained for $\lambda_1 = 0.9$ and $\lambda_2 = 0.95$, which were our final parameter values.

We denote our method as GeoGAN (Geometry Aware GANs), and compare it's performance against other methods such as: 1) rotation, translation and scaling (denoted as DA-Data Augmentation); 2) $DAGAN$ - data augmentation GANs of [1]; 3) $cGAN$ - the conditional GAN based method of [25]; and 4) $Zhao$- the atlas registration method of [42]. Segmentation performance is evaluated in terms of Dice Metric (DM) [10] and Hausdorff Distance (HD) [31]. DSC of 1 indicates perfect overlap and 0 indicates no overlap, while lower values of HD (in mm) indicate better segmentation performance.

Algorithm Baselines. The following variants of our method were used for ablation studies:

1. GeoGAN$_{noL_{class}}$- GeoGAN without classification loss (Eqn.3).

2. GeoGAN$_{noL_{shape}}$- GeoGAN without shape relationship modeling term (Eqn.4).

3. GeoGAN$_{NoSamp}$ - GeoGAN without uncertainty sampling for injecting diversity to determine sampling's relevance to the final network performance.

4. GeoGAN$_{L_{class}}$ - GeoGAN using classification loss (Eqn.3) and adversarial loss (Eqn.2) to determine L_{class}'s relevance to GeoGAN's performance.

5. GeoGAN$_{L_{shape}}$ - GeoGAN using shape loss (Eqn.4) and adversarial loss (Eqn.2) to determine L_{shape}'s contribution to GeoGAN's performance.

6. GeoGAN$_{Samp}$ - GeoGAN using only uncertainty sampling and adversarial loss (Eqn.2). This baseline quantifies the contribution of sampling to the image generation process.

4.3. Segmentation Results And Analysis

We hypothesize that a good image augmentation method should capture the different complex relationships between the anatomies and the generated images leading to the improvement in segmentation accuracy. Average DSC for **pathological images** from all device types are reported in Table 1 for the RETOUCH test dataset. Figure 5 shows the segmentation results using a UNet trained on images from different methods. Figure 5 (a) shows the test image along with the manual mask overlayed and shown as the red contour and Figure 5 (b) shows the manual mask. Figures 5 (c)-(g) show, respectively, the segmentation masks obtained by GeoGAN, $Zhao$ [42], $DAGAN$, $cGAN$ and DA.

Our method outperforms baseline conventional data augmentation and other competing methods by a significant margin. Results of other methods are taken from [3]. GeoGAN's DSC of 0.906 is higher than the DSC value (0.87) of the best performing method (obtained on the Spectralis images of the datasaet). While GeoGAN's average performance is equally good across all three device images, the competing methods rank differently for different devices. GeoGAN's superior segmentation accuracy is attributed to it's capacity to learn geometrical relationship between different layers (through L_{shape}) much better than competing methods. Thus our attempt to model the intrinsic geometrical relationships between different labels could generate superior quality masks.

In a separate experiment we train GeoGAN with images of one device and segment images of the other devices, and repeat for all device types. The average DSC value was 0.893, and HD was 8.6 mm. The decrease in performance compared to GeoGAN in Table 1 is expected since the training and test images are from different devices. However we still do better than $Zhao$ [42] and competing methods on the same dataset.

We repeat the set of experiments in Table 1 using a Dense UNet [23] instead of UNet as the segmentation network. We

	Comparison approaches				Proposed
	DA	DAGAN [1]	cGAN [25]	Zhao [42]	GeoGAN
DM	0.793 (0.14)	0.825 (0.10)	0.851 (0.07)	0.884 (0.09)	**0.906** **(0.04)**
HD	14.3 (4.2)	12.9 (3.8)	10.6 (3.0)	8.8 (3.3)	**7.9** **(2.2)**

Table 1. Segmentation results for **pathological** OCT images from the RETOUCH database. Mean and standard deviation (in brackets) are shown. Best results per metric is shown in bold.

obtain the following average DSC values: GeoGAN -0.917, $Zhao - 0.896$, $cGAN - 0.864$, $DAGAN - 0.834$ and $DA - 0.802$. GeoGAN gives the best results, thus indicating it's better performance irrespective of the backbone segmentation framework.

Ablation Studies. Table 2 shows the segmentation results for different ablation studies. Figure 6 shows the segmentation mask obtained by different baselines for the same image shown in Figure 5 (a). The segmentation outputs are quite different from the ground truth and the one obtained by GeoGAN. In some cases the normal regions in the layers are included as pathological area, while parts of the fluid region are not segmented as part of the pathological region. Either case is undesirable for disease diagnosis and quantification. Thus, different components of our cost functions are integral to the method's performance and excluding one or more of classification loss, geometric loss and sampling loss adversely affects segmentation performance.

	GeoGAN noL_{cls}	GeoGAN noL_{shape}	GeoGAN $noSamp$
DM	0.867(0.07)	0.864(0.09)	0.862(0.09)
HD	9.4(3.0)	9.5(3.3)	9.9(3.2)
	GeoGAN $onlyL_{cls}$	GeoGAN $onlyL_{shape}$	GeoGAN $onlySamp$
DM	0.824(0.08)	0.825(0.07)	0.818(0.06)
HD	11.2(2.9)	11.1(3.0)	12.5(2.8)

Table 2. Mean and standard deviation (in brackets) of segmentation results from ablation studies on **pathological** OCT images from the RETOUCH database. HD is in mm.

4.4. Realism of Synthetic Images

Prior results show GeoGAN could generate more diverse images, which enables the corresponding UNet to show better segmentation accuracy. Figure 1 shows examples of generated synthetic images using $GeoGAN$ and the other image generation methods except DA since it involves rotation and scaling only while Figure 7 shows examples from
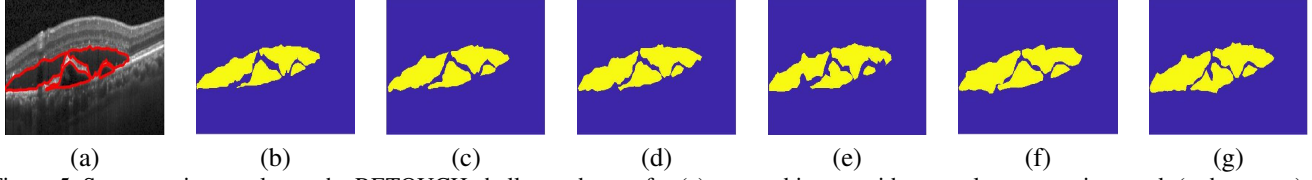
Figure 5. Segmentation results on the RETOUCH challenge dataset for (a) cropped image with manual segmentation mask (red contour); Segmentation masks by (b) ground truth (manual); (c) GeoGAN; (d) Zhao [42]; (e) $DAGAN$; (f) $cGAN$ and (g) conventional DA.
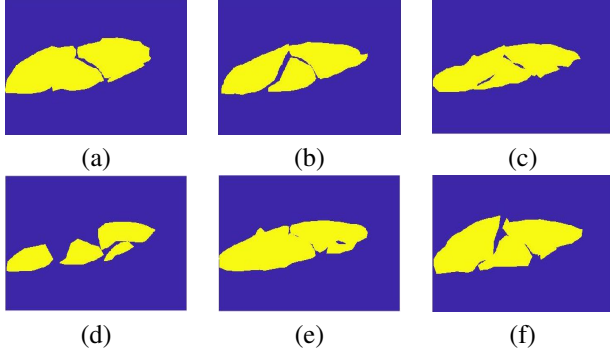

Figure 6. Ablation study results for: (a) $GeoGAN_{noL_{shape}}$; (b) $GeoGAN_{noL_{cls}}$; (c) $GeoGAN_{noSamp}$; (d) $GeoGAN_{onlyL_{shape}}$; (e) $GeoGAN_{onlyL_{cls}}$; (f) $GeoGAN_{onlySamp}$. HD is in mm.

Agreement Statistics	Both Experts	Atleast 1 Expert	No Expert
GeoGAN	**88.0** (440)	**94.6** (473)	**5.4** (27)
Zhao et. al.[42]	**84.8** (424)	**88.2** (441)	**11.8** (59)
cGAN ([25])	**83.2** (416)	**85.4** (427)	**14.6** (73)
DAGAN([1])	**82.2** (411)	**84.2** (421)	**15.8** (79)
DA	**80.4** (402)	**82.4** (412)	**17.6** (88)
$GeoGAN_{noL_{cls}}$	**83.6** (418)	**86.4** (432)	**13.6** (68)
$GeoGAN_{noL_{shape}}$	**83.0** (415)	**85.6** (428)	**14.4** (72)
$GeoGAN_{noSamp}$	**82.8** (414)	**85.0** (425)	**15.0** (75)
$GeoGAN_{L_{cls}}$	**82.2** (411)	**84.0** (420)	**16.0** (80)
$GeoGAN_{L_{shape}}$	**81.2** (406)	**83.4** (417)	**16.6** (83)
$GeoGAN_{Samp}$	**80.4** (402)	**82.8** (414)	**17.2** (86)

Table 3. Agreement statistics for different image generation methods amongst 2 ophthalmologists. Numbers in bold indicate agreement percentage while numbers within brackets indicate actual numbers out of 500 patients.

the ablation models. The base image is the same in both figures. Visual examination shows GeoGAN generated images respect boundaries of adjacent layers in most cases, while other methods tend not to do so.

Only GeoGAN and to some extent $Zhao$ generate images with consistent layer boundaries. Images generated by other methods suffer from the following limitations: 1) tend to be noisy; 2) multiple artifacts exposing unrealistic appearance; 3) smoothed images which distort the layer boundaries; 4) different retinal layers tend to overlap with the fluid area. Segmentation models trained on such images will hamper their ability to produce accurate segmentations.

Two trained ophthalmologists having 4 and 5 years experience in examining retinal OCT images for abnormalities assessed realism of generated images. We present them with a common set of 500 synthetic images from GeoGAN and ask them to classify each as realistic or not. The evaluation sessions were conducted separately with each ophthalmologist blinded to other's answers as well as the image generation model. Results with GeoGAN show one ophthalmologist (OPT 1) identified 461/500 (92.2%) images as realistic while OPT 2 identified 452 (90.4%) generated images as realistic. Both of them had a high agreement with 440 common images (88.0% -"$Both\ Experts$" in Table 3) identified as realistic. Considering both OPT 1 and OPT 2 feedback, a total of 473 (94.6%) unique images were identified as realistic ("$Atleast\ 1\ Expert$" in Table 3). Subsequently, 27/500 (5.4%) of the images were not identified

as realistic by any of the experts ("$No\ Expert$" in Table 3). Agreement statistics for other methods are summarized in Table 3.

The highest agreement between two ophthalmologists is obtained for images generated by our method. For all the other methods their difference from $GeoGAN$ is significant. Zhao et. al. [42] has the best performance amongst them, but has agreement difference of more than 6% (for "$Atleast\ 1\ Expert$") compared to $GeoGAN$ (94.6 vs 88.2). The numbers from Table 3 show a larger difference for the other methods, thus highlighting the importance of modeling geometric relationships in pathological region segmentation.

4.5. Combining Disease And Normal Dataset

Section 4.3 shows results of training the UNet on diseased population shapes to segment diseased shapes. In this section we show the opposite scenario where the training was performed on normal images, the network subsequently used to generate images from the diseased base images and segment test images of a diseased population. Table 4 shows the corresponding results and also for the scenario when the training images were a mix of diseased and normal population, while the test images were from the diseased population. All reported results are for the same set of test images.
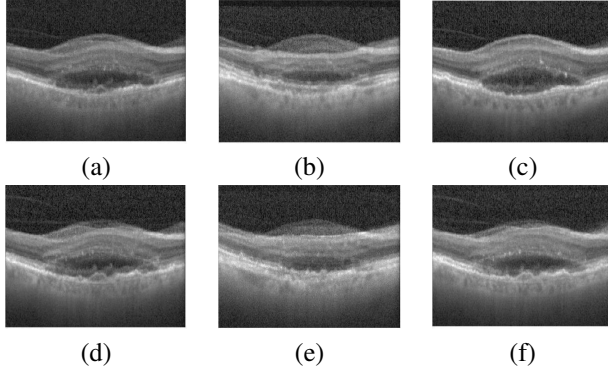
Figure 7. Generated images for ablation study methods: (a) $GeoGAN_{noL_{cls}}$; (b) $GeoGAN_{noL_{shape}}$; (c) $GeoGAN_{noSamp}$; (d) $GeoGAN_{onlyL_{cls}}$; (e) $GeoGAN_{onlyL_{shape}}$; (f) $GeoGAN_{onlySamp}$.

	Train on Normal, Test on Diseased				
	DA	DAGAN [1]	cGAN [25]	[42]	GeoGAN
DM	0.741	0.781	0.802	0.821	**0.856**
HD	15.3	14.5	13.7	11.3	**9.9**
	Train on Mix, Test on Diseased				
	DA	DAGAN [1]	cGAN [25]	[42]	GeoGAN
DM	0.762	0.798	0.820	0.848	**0.873**
HD	14.8	14.0	13.2	10.8	**9.2**

Table 4. Segmentation results for mix of diseased and normal OCT images. Best results per metric is shown in boldface. HD is in mm.

Comparing them with the results in Table 1, the superior performance of training separate networks for diseased and normal population is obvious. Figure 8 (a) shows the segmentation output when training and test image are from the diseased population, while Figure 8 (b) shows the scenario where the training images are from the normal population while the test images are the diseased case. Red contours show the outline of the manual segmentation while the green contours show the output of our method. When training images are from normal population it is more challenging to segment an image from the diseased population. Inaccurate segmentation of the fluid layers can have grave consequences for subsequent diagnosis and treatment plans. Figure 8 (c) shows the results when the training database is a mix of diseased and normal population, which is a more accurate representation of real world scenarios. A mixture of normal and diseased population images in the training set leads to acceptable performance. However, training a network exclusively on disease cases improves segmentation accuracy of pathological regions, which is certainly more critical than segmenting normal anatomical regions. Since it is challenging to obtain large numbers of annotated images, especially for diseased cases, our proposed image augmentation method is a significant improvement over existing methods.
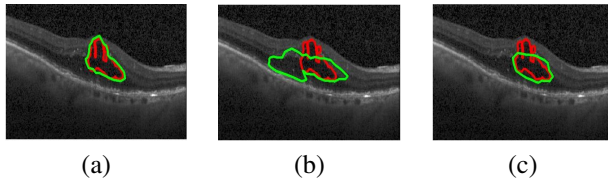

Figure 8. Segmentation results of test images for different training data sources: (a) diseased population only; (b) normal population only; (c) mix of diseased and normal population.

5. Conclusion

We propose a novel approach to generate plausible retinal OCT images by incorporating relationship between segmentation labels to guide the shape generation process. Diversity is introduced in the image generation process through uncertainty sampling. Comparative results show that the augmented dataset from $GeoGAN$ outperforms standard data augmentation and other competing methods, when applied to segmentation of pathological regions (fluid filled areas) in retinal OCT images. We show that synergy between shape, classification and sampling terms lead to improved segmentation and greater visual agreement of experienced ophthalmologists. Each of these terms is equally important in generating realistic shapes. Our approach can be used for other medical imaging modalities without major changes to the workflow.

Despite the good performance of our method we observe failure cases when the base images are noisy due to inherent characteristics of the image acquisition procedure, and when the fluid areas greatly overlap with other layers. Although the second scenario is not very common, it can be critical in the medical context. In future work we aim to evaluate our method's robustness on a wide range of medical imaging modalities such as MRI, Xray, etc. Our method is also useful to generate realistic images for educating clinicians, where targeted synthetic images (e.g. generation of complex cases, or disease mimickers) can be used to speed-up training. Similarly, the proposed approach could be used in quality control of deep learning systems to identify potential weaknesses through targeted high-throughput synthetic image generation and testing.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. In *arXiv preprint arXiv:1711.04340,*, 2017. 1, 2, 5, 6, 7, 8

[2] Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Htker, Urs J. Muehlematter, Khoschy

Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *Proc. MICCAI(2)*, pages 119–127, 2019. 3

[3] Hrvoje Bogunovic and et. al. RETOUCH: The retinal oct fluid detection and segmentation benchmark and challenge. *IEEE Trans. Med. Imag.*, 38(8):1858–1874, 2019. 5, 6

[4] Fred L. Bookstein. Integration, disintegration, and self-similarity: Characterizing the scales of shape variation in landmark data. *Evolutionary Biology*, 42(4):395–426, 2015. 3

[5] Behzad Bozorgtabar, Dwarikanath Mahapatra, and Jean-Philippe. Thiran. Exprada: Adversarial domain adaptation for facial expression analysis. *In Press Pattern Recognition*, 100:15–28, 2020. 2

[6] Behzad Bozorgtabar, Dwarikanath Mahapatra, Hendrik von Teng, Alexander Pollinger, Lucas Ebner, Jean-Philipe Thiran, and Mauricio Reyes. Informative sample generation using class aware generative adversarial networks for classification of chest xrays. *Computer Vision and Image Understanding*, 184:57–65, 2019. 2

[7] Behzad Bozorgtabar, Mohammad Saeed Rad, Hazım Kemal Ekenel, and Jean-Philippe Thiran. Learn to synthesize and synthesize to learn. *Computer Vision and Image Understanding*, 185:1–11, 2019. 2

[8] Behzad Bozorgtabar, Mohammad Saeed Rad, Dwarikanath Mahapatra, and Jean-Philippe Thiran. Syndemo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4210–4219, 2019. 2

[9] Stephanie J. Chiu, Xiao T. Li, Peter Nicholas, Cynthia A. Toth, Joseph A. Izatt, and Sina Farsiu. Automatic segmentation of seven retinal layers in sdoct images congruent with expert manual segmentation. *Opt. Express*, 18(18):19413–19428, 2010. 2

[10] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 5

[11] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Patt. Anal. Mach. Intell.*, 38(9):1734–1747, 2016. 2

[12] G. N. Girish, Bibhash Thakur, Sohini Roy Chowdhurya, Abhishek R. Kothari, and Jeny Rajan. Segmentation of intra-retinal cysts from optical coherence tomography images using a fully convolutional neural network model. *IEEE J. Biomed. Health Inform.*, 23(1):296–304, 2018. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[14] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. Gan-based synthetic brain mr image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738. IEEE, 2018. 1, 2

[15] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–731, 2018. 1

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, pages –, 2015. 3

[17] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *arXiv:1511.02680*, 2015. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980,*, 2014. 5

[19] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Proc. NIPS*, pages 6965–6975, 2018. 3

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NIPS*, pages 6402–6413, 2017. 2

[21] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 2

[22] Kelvin K. Leung, Matthew J. Clarkson, Johnathon W. Bartlett, Shona Clegg, Clifford R. Jack Jr, Michael W. Weiner, Nick C. Fox, Sebastien Ourselin, and A. D. N. Initiative. Robust atrophy rate measurement in alzheimers disease using multi-site serial mri: tissue-specific intensity normalization and parameter selection. *Neuroimage*, 50(2):516–523, 2010. 2

[23] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-DenseUNet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imag.*, 37(12):2663–2674, 2018. 6

[24] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Sajini Hewavitharanage. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In *MICCAI*, pages 382–390, 2017. 2

[25] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Phillipe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *MICCAI*, pages 580–588, 2018. 1, 2, 3, 5, 6, 7, 8

[26] Dwarikanath Mahapatra and Zongyuan Ge. Training data independent image registration with gans using transfer learning and segmentation information. In *In Proc. IEEE ISBI*, pages 709–713, 2019. 2

[27] Dwarikanath Mahapatra and Zongyuan Ge. Training data independent image registration using generative adversarial

networks and domain adaptation. *In press Pattern Recognition*, 100:1–14, 2020. 2

[28] Dwarikanath Mahapatra, Zongyuan Ge, Suman Sedai, and Rajib Chakravorty. Joint registration and segmentation of xray images using generative adversarial networks. In *In Proc. MICCAI-MLMI*, pages 73–80, 2018. 2

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical im- age segmentation. In *Proc. Int. Conf. on 3D vision*, pages 565–571, 2016. 2

[30] Christopher Nielsen and Michal Okoniewski. Gan data augmentation through active learning inspired sample acquisition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 109–112, 2019. 1, 2

[31] Javier Ribera, David Gera, Yuhao Chen, and Edward Delp. Weighted hausdorff distance: A loss function for object localization. In *arXiv preprint arXiv:1806.07564*, 2018. 5

[32] Olaf Ronneberger, Phillip Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *In Proc. MICCAI*, pages 234–241, 2015. 3, 5

[33] Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, and Nassir Navab. Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed. Opt. Express*, 8(8):3627–3642, 2017. 2

[34] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proc. CVPR*, pages 3591–3600, 2017. 2

[35] Thomas Schlegl, Sebastian M. Waldstein, Wolf-Dietrich Vogl, Ursula Schmidt-Erfurth, and George Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *Proc. Int. Conf. Inform. Process. Med. Imag. (IPMI)*, pages 437–438, 2015. 2

[36] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine Andriole, and Mark Michalski. Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks. In *Proc. MICCAI-SASHIMI*, 2018. 2

[37] Kihuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Proc. NIPS*, pages 3483–3491, 2015. 3

[38] Nima Tajbakhsh, Jae. Y. Shin, Suryakant R. Gurudu, R. Todd Hurst, Chrostopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE Trans. Med. Imag.*, 35(5):1299–1312, 2016. 1

[39] Freerk G. Venhuizen, Bram van Ginneken, Bart Liefers, Freekje van Asten, Vivian Schreur, Sascha Fauser, Carel Hoyng, Thomas Theelen, , and Clara I. Sanchez. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomed. Opt. Express*, 9(4):1545–1569, 2018. 2

[40] Xiayu Xu, Kyungmu Lee, Li Zhang, Milan Sonka, and Michael D. Abramoff. Stratified sampling voxel classification for segmentation of intraretinal and sub-retinal fluid in longitudinal clinical oct data. *IEEE Trans. Med. Imag.*, 34(7):1616–1623, 2015. 2

[41] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Med. Imag. Anal.*, 58, 2019. 2

[42] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V. Guttag, and Adrian V. Dalca. Data augmentation using learned transforms for one-shot medical image segmentation. In *In Proc. CVPR*, pages 8543–8552, 2019. 1, 5, 6, 7, 8

[43] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and LiCheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Med. Imag. Anal*, 49:14–26, 2018. 2, 3