Retinal Image Synthesis

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Electronics and Communication Engineering
by Research

by

Anurag Anil Deshmukh
201331117
anurag.deshmukh@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2021

Copyright © Anurag Anil Deshmukh, 2020
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " **Retinal Image Synthesis**" by **Anurag Anil Deshmukh**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Jayanthi Sivaswamy

This thesis is dedicated to my family, advisor and friends for their endless support and encouragement.

Acknowledgements

I would like to express my deepest gratitude to Dr. Jayanthi Sivaswamy for advising and guiding me through every step of this thesis. This work would not have been possible without her constant counsel, support and patience. Her immense dedication towards improving our understanding of the subject, including taking a separate course, Medical Image Processing for just the 3 students made our transition to the field a lot easier. I learned a lot about research from her suggestions and discussions at weekly meetings. I'm eternally grateful for her reassurances and support during the difficult times which kept me moving towards achieving this goal.

I would also like to thank Arunava Chakravarty, A K Pujitha and Karthik Gopinath who helped me a lot during my initial days at the lab. I'm greatful to the my fellow lab mates- Abhinav Dhere, Alphin J Thottupattu, Divya Jyothi, Mythri Venkatesh, Prathyusha Akundi and Sukesh Adiga at MIP group for the insightful discussions at our weekly group meetings.

This acknowledgement wouldn't feel complete without the special mention of my friends Kaveri, Anvesh, Rajat Singla, Kanishk, Apuroop and Aditya Shreekar, who were constantly by my side at every step of the way. I'd also like to thank my wingmates - Rajat Singh, Debayan, Nitish, Abhinav, Kanishk, Pratik and Gautam for making my college life an exciting ride.

Finally, my thesis and journey would not be possible without my loving parents' support, encouragement and guiding hand.

Abstract

Medical imaging has been aiding diagnosis and treatment of diseases by creating visual representations of the interior of the human body. Experts hand-mark these images for abnormalities and diagnosis. Supplementing experts with these rich visualization has enabled detailed clinical analysis and rapid medical intervention. However, deep learning-based methods rely on abundantly large volumes of data for training. Procuring data for medical imaging applications is especially difficult because abnormal cases by definition are rare and the data, in general, requires experts for labelling.

With Deep learning algorithms, data with high class imbalance or of insufficient variability leads to poor classification performance. Thus, alternate approaches like using generative modelling to artificially generate more data have been of interest. Most of these methods are GAN [11] based approaches. While they can be helpful with data imbalance they still require a lot of data to be able to generate realistic images. Additionally, a lot of these methods have been shown to work on natural images where the images are relatively noise-free and smaller artifacts aren't as damaging. Thus, this thesis aims at providing synthesis methods which overcome the limitations of smaller datasets and noisy profile.

We do this for two different modalities, Fundus imaging and Optical Coherence Tomography (OCT). Firstly, we present a fundus image synthesis method aimed at providing paired Optic Cup and Image data for Optic Cup (OC) Segmentation. The synthesis method works well on small datasets by minimising the information to be learnt by leveraging domain-specific knowledge and by providing most of the structural information to the network. We demonstrate this method's advantages over a more direct synthesis method. We show how leveraging domain-specific knowledge can provide higher quality images and annotations. Inclusion of these generated images and their annotations in training of an OC segmentation model showed a significant improvement in performance, thus showing their reliability.

Secondly, we present a novel unpaired image to image translation method which can introduce abnormality (Drusen) to OCT images while avoiding artifacts and preserving the noise profile. Comparison with other state-of-the-art images to image translation methods shows that our method is significantly better at preserving the noise profile and is better at generating morphologically accurate structures.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Generative Modeling . . . 1
1.2 Image Synthesis . . . 2
1.2.1 Conditional Image Synthesis . . . 4
1.3 Image to Image Translation . . . 4
1.4 Image Synthesis for Augmentation / Image Synthesis in Medical Imaging . . . 6
1.4.1 Thesis focus . . . 7
1.4.2 Thesis Contribution . . . 8
1.4.3 Thesis Organisation . . . 8

2 Synthesis of Optical nerve head region of fundus image
A Domain specific vessel bend based Approach . . . 9
2.1 Introduction . . . 9
2.2 Method . . . 11
2.2.1 Training GAN . . . 11
2.2.2 Generator architectures . . . 12
2.2.3 Generating new examples . . . 13
2.3 Experiments and Results . . . 15
2.3.1 Visual Quality . . . 15
2.3.2 Boosting Optic Cup segmentation . . . 16
2.4 Conclusion . . . 18

3 Synthesis of Optical nerve head region of fundus image
A Direct Synthesis Approach . . . 20
3.1 Introduction . . . 20
3.2 Method . . . 20
3.2.1 Training GAN . . . 20
3.3 Experiments and Result . . . 22
3.3.1 Visual Quality of Generated images . . . 22
3.3.2 Quality of synthesised Optic Cup mask . . . 22
3.3.3 Contribution of OD and Vessels in OC synthesis . . . 23
3.3.4 Boosting Cup segmentation . . . 24
3.4 Conclusion . . . 26

4 Drusen Synthesis in Optical coherence tomography 28
4.1 Introduction 28
4.2 Related work 28
4.3 Method 30
4.3.1 Self-Attention 31
4.3.2 ESSIM 32
4.3.3 Network architecture 34
4.3.3.1 Generator 35
4.4 Results and Experiments 36
4.4.1 Dataset 36
4.4.2 Qualitative analysis 37
4.4.3 Quantitative Evaluation 39
4.4.4 Using Generated Images for Augmentation 41
4.5 Conclusion 42

5 Conclusion 43
5.1 Future Scope 44

Bibliography 46

List of Figures

Figure		Page
1.1	Discriminative models vs Generative models (image from [12])	1
1.2	Variational AutoEncoders (image from [33])	2
1.3	Three Major variants of GANs (a) GAN, (b) conditional GAN (CGAN) and (c) auxiliary classifier GAN (ACGAN) architectures. Here x denotes the real image, c is the class label, z is a random noise vector, G is the Generator and D is the Discriminator.	3
1.4	Conditional vs Unconditional Image Synthesis (images from [18])	4
1.5	Different categories of Image-to-image translation tasks	5
1.6	Different applications of generative modeling in medical imaging	6
2.1	Comparison of synthetic image generation using different approaches. Left to right: Results of [6],[47] and the proposed method.	10
2.2	The GAN architecture used to generate images of the ONH region	11
2.3	Two different arrangements for the generators each of the block is a U-net or ResNet.	13
2.4	(a)B-spline registration of OC_A onto OC_B to get the transformation T, (b) Applying the transform T onto V_A to get V'_A	14
2.5	3 sample images generated using each of the four different generators compared	15
2.6	Comparison of vessel detail in the images Generated by $\mathcal{G}_2$ Resnet (left) and $\mathcal{G}_2$ Unet (right)	16
2.7	Generation of synthetic images using R2-Resnet (last column) with vessel mask of real images (first column) and OC masks derived from different images (middle). The green circles indicate vessel bends at the edge of the OC.	17
2.8	DICE scores versus the number of synthetic images used for training	18
3.1	The GAN architecture used to generate Fundus images and OC masks	21
3.2	A sample set of generated images	23
3.3	Images Generated by Proposed architecture(left) and $\mathcal{G}_2$-Resnet from Chapter 2 (right)	24
3.4	OC masks generated by the network marked onto generated images by blue and Ground truth masks marked by green	25
3.5	OC masks generated by the network marked onto generated images by blue and Ground truth masks marked by green	26
3.6	OC masks generated by the network marked onto generated images by blue and Ground truth masks marked by green	27
4.1	(a) Retinal layers visible in OCT, (b) Drusen observed as folds in RPE layer (marked with green arrows) and OPL layer adjusting due to the Drusen (marked with red arrows)	29

4.2 Self-Attention layer (figure from [46]) 31
4.3 Convolutional filters used to calculate histogram 32
4.4 Network Architecture 33
4.5 Generator used by our architecture 35
4.6 Images during the Preprocessing steps (a) is the cropped noisy image (b) is the denoised image and (c) is after retinal flattening 37
4.7 Comparision between different methods to generate Drusen 38
4.8 OPL layer being lifted up in generated images to accommodate for the introduced Drusen. 38
4.9 Attention maps corresponding to the regions marked by the circle in red 39

List of Tables

Table Page

2.1 Mean Q_v values of the images generated by different generators 18
2.2 Cup segmentation Performance comparison on Drishti-GS 19

3.1 DICE score between generated OC with OC corresponding to input vessel and with OC corresponding to input OD 25

4.1 Effect of each addition to ResCycleGAN to the image quality 40
4.2 Comparison with other methods image to image translation methods 40
4.3 Comparison of classification performance using different methods for augmentation . . 41

Chapter 1

Introduction

1.1 Generative Modeling

In recent years Artificial intelligence (AI) specifically Machine Learning (ML) has become a thriving field with a great number of applications and valuable research topics. All of these machine learning algorithms aim to learn the representation of the given data. In the past, all of these representations were highly specific to the task and the method could not be transferred over to new tasks even in the same domain. Hence Deep Learning was introduced to process raw sensory data and learn abstract, high-level representations.

As per the availability of data, Machine Learning algorithms can be divided into two categories - supervised and unsupervised learning. Supervised learning requires each example in the dataset to be labeled. Supervised learning can be further split into classification and regression depending on whether the labels are discrete or not. Whereas in unsupervised learning data labels are absent and the final aim is to explore the underlying structure of the data. These techniques include denoising, density estimation, clustering, and synthesis. Availability of large datasets [7] [24], has made it possible to train supervised learning models that are able to beat even human performance [42] [23] [40].

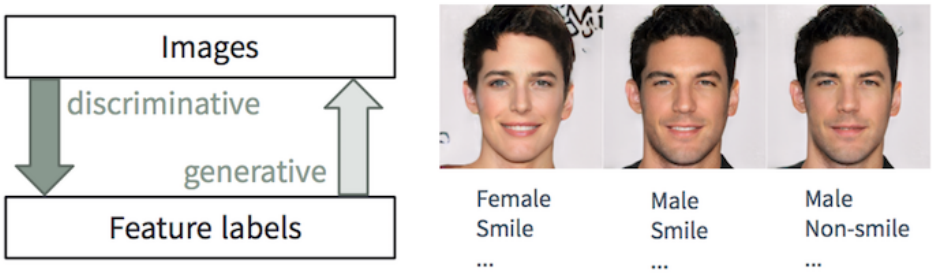

Figure 1.1 Discriminative models vs Generative models (image from [12])

While more data is being made available, it is hard to annotate the ever-increasing number of samples. This is especially true for medical imaging where finding trained medical professionals for an-

notating large scale data is a challenge and multiple experts are required to annotate the same example to make the annotations as reliable as possible. Hence, researchers are more interested in unsupervised learning. In unsupervised learning, generative modelling is one of the most promising avenues. While discriminative models aim at predicting labels given the image/sample, generative models aim at predicting the image/sample given labels/image description (Figure.1.1). Typical generative models were usually based on Markov chains, Maximum Likelihood, and approximate inference. However, these early models [38] [14] are very limited in how well they generalize. In 2014, Goodfellow et al. [11] proposed a novel generative model, named Generative Adversarial Networks (GANs). Based on the game theory, there are two networks in GANs. Due to their flexibility, GANs and its variants have been heavily adopted in Computer Vision.

The main applications of generative models in Computer Vision (CV) can be divided into two categories, Image synthesis and Image to Image translation. Image synthesis focuses on learning the underlying distribution of real images by mapping them to a fixed known distribution (usually a unit Gaussian) or by using Auto Regressive models that predict an image as a sequence of pixels thus making it possible to generate new examples from scratch that were otherwise absent in existing dataset. Image-to-image translation is the task of taking images from one domain and transforming them so they have the style or characteristics of images from another domain.

1.2 Image Synthesis

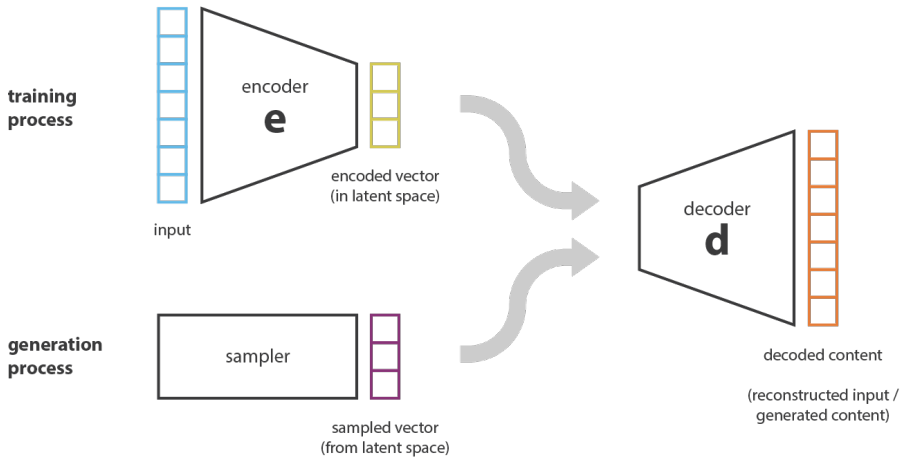

Figure 1.2 Variational AutoEncoders (image from [33])

As mentioned, Image synthesis is the task of generating an image from scratch and thus focuses on learning the underlying distribution. In Computer Vision the three most popular image synthesis methods are Autoregressive models [31], Variational Autoencoders (VAE) [21] and Generative Adversarial Networks (GAN) [11].

Auto Regressive models predict the image sequentially with recurrent layers and residual connections that predict pixels across the vertical and horizontal axes. The architecture models the joint distribution of pixels as a product of conditional distributions of horizontal and diagonal pixels. These models usually require some context based on which the image is generated. This context can be an incomplete image or a damaged image. While the model gives sharp realistic images, they aren't mapped to any latent space thus the model can't capture the underlying data distribution.

A Variational Autoencoders (VAE) (Figure. 1.2) provide a probabilistic manner for describing an observation in regularised latent space. Thus, rather than building an encoder which outputs a single value to describe each latent attribute, VAE formulates an encoder to describe a probability distribution for each latent attribute. A decoder is trained to reconstruct the original image back. This gives the latent space the property of samples being interpolative thus allowing us to generate new samples using the decoder. VAEs are popular as they require much less data than comparable models to train but their generated images lack sharpness.

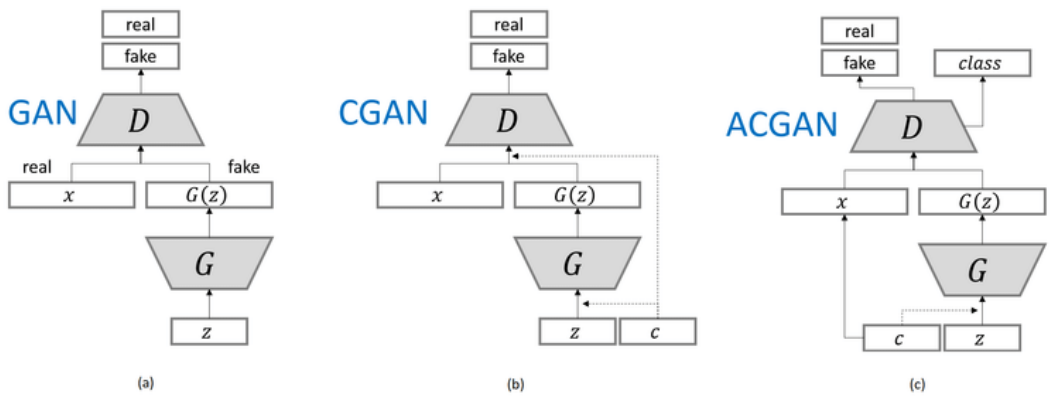

Figure 1.3 Three Major variants of GANs (a) GAN, (b) conditional GAN (CGAN) and (c) auxiliary classifier GAN (ACGAN) architectures. Here x denotes the real image, c is the class label, z is a random noise vector, G is the Generator and D is the Discriminator.

Generative Adversarial Networks (GANs) have gained a lot of attention in the Computer Vision community due to their capability of data generation without explicitly modelling the probability density function. GAN takes a supervised learning approach to do unsupervised learning. The architecture has 2 networks a Generator(G) and a Discriminator (D). The Discriminator(D) is a binary classifier that is trained to tell the real and generated images apart. The Generator (G) samples a random sample from a unit Gaussian and tries to generate a realistic image such that the Discriminator fails in classifying it as a generated image. Since the Generator does not directly observe any of the training samples from the dataset GANs are less prone to over-fitting. The adversarial loss brought by the Discriminator also provides a clever way of incorporating unlabeled samples into training and imposing higher order consistency.

1.2.1 Conditional Image Synthesis

Conditional image generation is the task of generating new images from a dataset conditional on their class labels, attributes or description. Another reason GANs are widely used is that they can also incorporate the class labels as input to the Generator (G) and Discriminator (D). Discriminator is now trained to label images real or synthetic not just based on the realism of the image but also based on the generated image's class. This variant of GAN is referred as conditional GAN (cGAN) [29]. Another variant of GAN, Auxiliary classifier GAN (AC-GAN) [30] incorporates the class label in the Generator (G) and the Discriminator acts as an auxiliary classifier to classify the generated sample along with differentiating the generated and real samples apart.

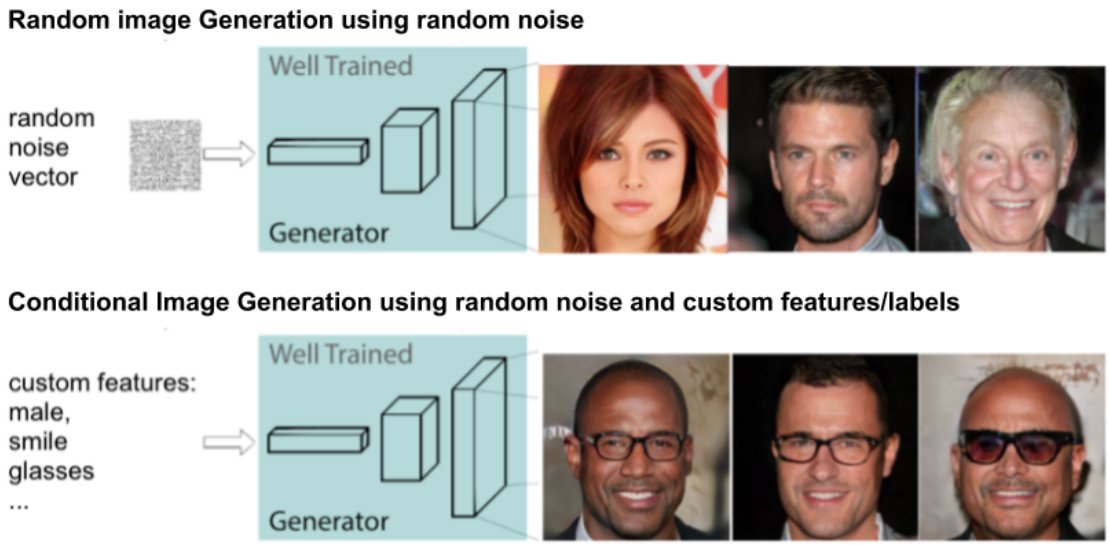

Figure 1.4 Conditional vs Unconditional Image Synthesis (images from [18])

This gives control over the class of the generated image as well as a stronger classifier than otherwise possible with normal training. This flexibility of GANs has proven to be useful in many tasks, such as domain adaptation, data augmentation and also in improving classification classification models. Later proposed variants of AC-GAN can also be used to control exact attributes of the generated images. Architectures like cGAN have also inspired many architectures in Image-to-image translation as explained in the Sub-section 1.3.

1.3 Image to Image Translation

Image-to-image translation is the task of taking images from one domain (source) and transforming them so they have the style or characteristics of images from another domain (target). Numerous varied tasks in Computer Vision can be modelled as Image-to-image translation. Some of these include:

Style transfer: the task of making a photo look like paintings by a particular artist.
Super-resolution: the task of obtaining a High-Resolution image from a Low-Resolution image.
Image denoising: the task of removing noise from an image.

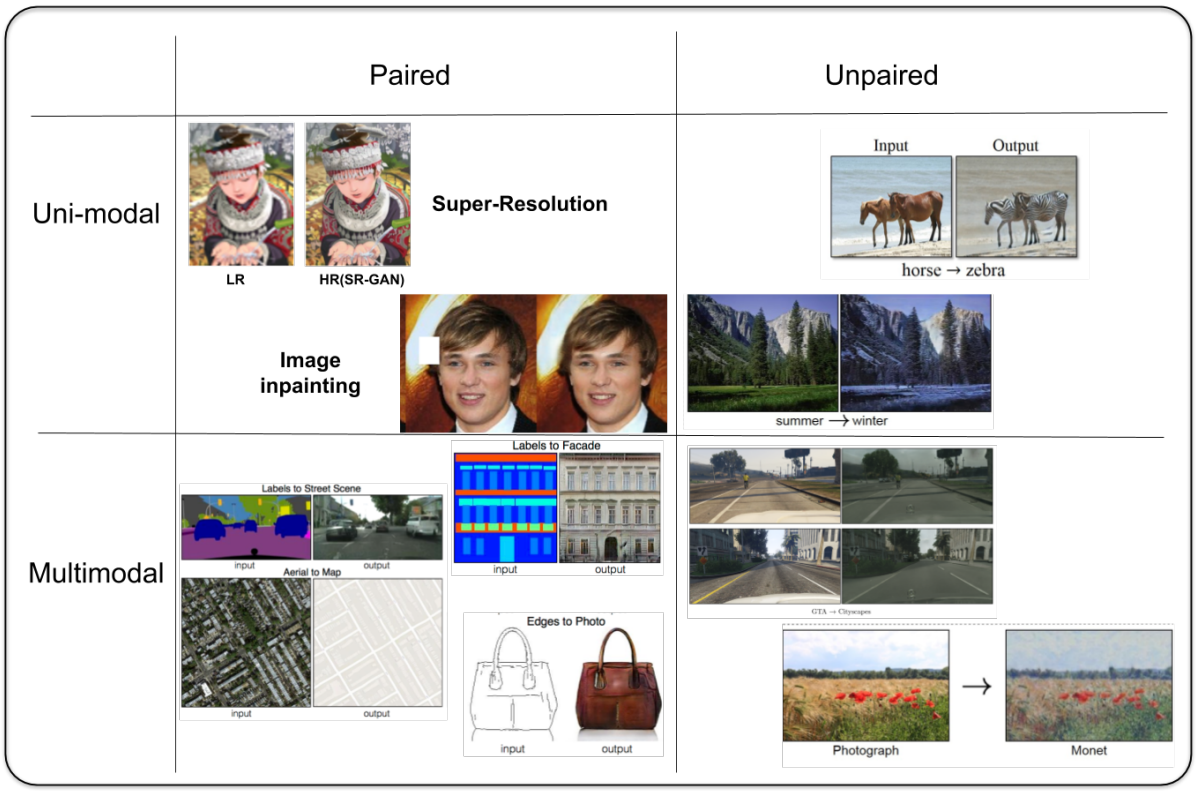

Figure 1.5 Different categories of Image-to-image translation tasks

Most of the Image-to-image translation tasks can be broadly classified into one of the two categories **paired** Image-to-image synthesis or **unpaired** Image-to-image synthesis based on the availability of paired or unpaired data. Tasks like Super-resolution, or image synthesis from available semantic segmentation masks can be modelled as a paired Image-to-image translation. While tasks such as image denoising, style transfer don't have paired data readily available. One of the most popular Image-to-image translation method currently being used is pix2pix [15]. It is inspired from cGAN, the Generator takes in an image from the source domain and the Discriminator takes in the output of the Generator (generated target domain image) as well as the input source image. Image-to-image translation methods can also be categorised into uni-modal and multi-modal based on if the source and target domain belong to the same modality. A few common examples of unimodal Image-to-image translation are image super-resolution and denoising. And a few examples for multi-modal Image-to-image translation are image to painting style transfer. Figure. 1.6 shows the different kinds of paired/unpaired and unimodal/multimodal image to image translation applications.

1.4 Image Synthesis for Augmentation / Image Synthesis in Medical Imaging

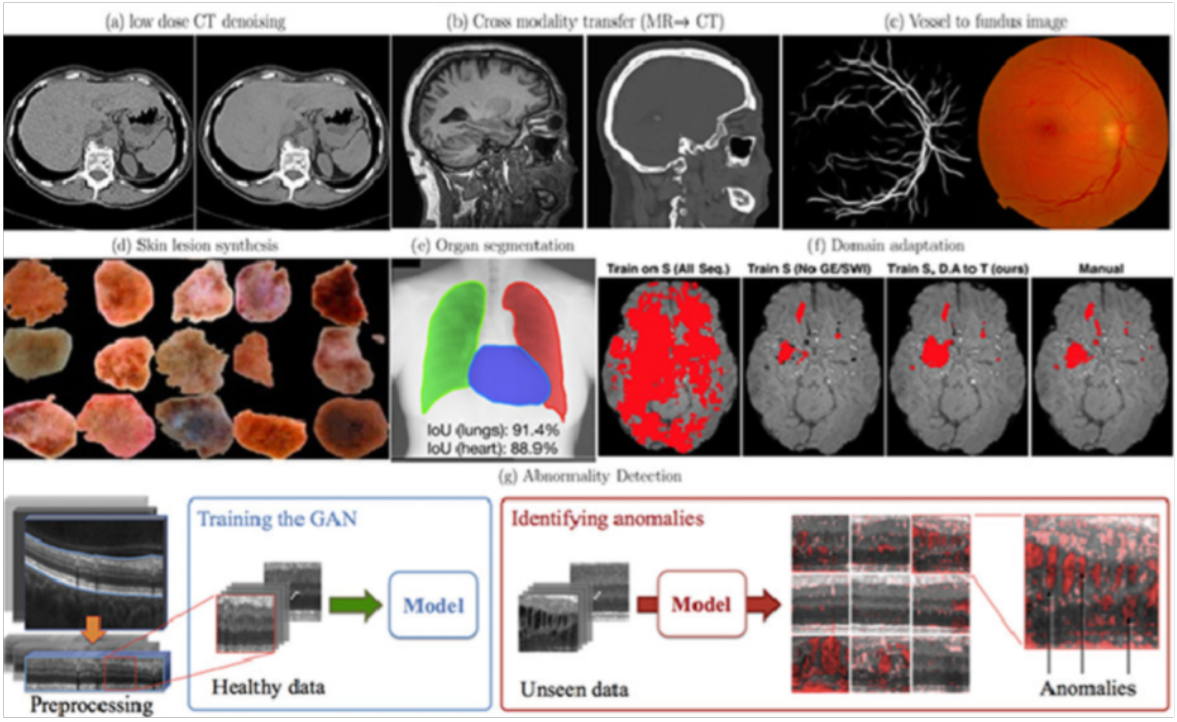

Figure 1.6 Different applications of generative modeling in medical imaging

Medical imaging is the process of creating visual representations of the interior of a body for clinical analysis and medical intervention, as well as a visual representation of the function of some organs or tissues. Medical imaging has been a crucial part of diagnosis and treating diseases. These images are examined by experts for abnormalities and diagnosis. However, since the rapid rise of deep learning many of these tasks are being automated and it is widely known that sufficient data volume is necessary for training successful machine learning algorithms. Thus generative models have been of huge interest. Data with high class imbalance or of insufficient variability leads to poor classification performance. While this is true for every field, three main factors cause this to be a much bigger problem in medical imaging than other domains,

1. In medical imaging abnormal findings are by definition uncommon,
2. The cost of acquiring and annotating these images is much higher as the annotations require multiple experts for a single sample,
3. While most of the natural images are captured using CMOS sensor, different imaging modalities have vastly different acquisition methods, thus making transfer learning more difficult.

Moreover, in the case of image segmentation tasks, the time required to manually annotate data only widens this disparity. Manually segmenting an abnormality can require upwards of five minutes per study making it impractical in a busy radiology practice. The result is a scarcity of annotated data and considerable challenges when attempting to train an accurate model. While traditional data augmentation techniques (e.g., cropping, translation, rotation) can mitigate some of these issues, they fundamentally produce highly correlated image training data.

While data augmentation has been one of the main uses of generative models, they also have been used for many different applications like image reconstruction (low dose CT to high dose CT) [44], cross-modality synthesis (MR to CT), anomaly detection. GANs have also been used for improving classification by using the discriminator of an acGAN as the main classifier [9]. Another way generative models can help in medical imaging is data anonymization.

While there have been some studies aimed at synthesising abnormal cases to better handle class imbalance/data scarcity, most of the current methodologies focus on cross-modality synthesis. Introducing abnormality while keeping an anatomical structure intact is still a challenging task. Moreover, the understanding of the particulars of human anatomy can be leveraged to use more domain-specific knowledge, which hasn't been explored much.

1.4.1 Thesis focus

In this thesis, we focus on retinal image synthesis in two imaging modalities, OCT and fundus imaging. The main aim is to develop learning-based image synthesis method for synthesis of abnormal cases from a normal case and vice-versa. These images can be used for better data augmentation and to reduce the present class imbalance. While traditional augmentation methods can be helpful they create highly correlated data. The generated images should not have any artifacts and should not be blurry. This is especially important in case of medical imaging where the present data is very limited and adding outliers to the distribution might heavily impact the learning methods. Moreover, the generated images should preserve the already present anatomical details and the new abnormalities added should be anatomically correct and representative of an actual case, as more of the learning methods in medical imaging are expected to be as explainable as possible.

For fundus images, we aim to synthesise the optic nerve head region of the retina along with Optic cup mask. Fundus imaging in the Optic nerve head region captures a 3D region projected onto in a 2D plane, thus the effects caused by the transition should be captured in the generated images. For example, the Optic Cup is a depressed region in the retina and thus causes the vessels to bend. Information like this is critical to Glaucoma detection and needs to be captured in the generated images. Capturing the vessels in the generated images is also a challenging task, as they are translucent. Since they are tubular structures they are the most transparent at the center and become more opaque towards the edges. Fundus synthesis is done using a very limited dataset and we employ a method that leverages domain-specific knowledge and compare it to an approach that aims at generating the OC annotation as well.

Optical Coherence Tomography images are corrupted by speckle noise thus their synthesis becomes more challenging. The network has to learn to replicate the noise characteristics as well as avoid creating artifacts. Specifically, we aim at introducing Drusen to normal OCT scans. Drusen are lipid deposits that look like folds in the OPL layer of the retina. The generated images need to be morphologically correct as to not have two overlapping layers. This requires a network with enough complexity to change the structure of the required layers while still affecting and not modifying other parts of the image.

1.4.2 Thesis Contribution

The major contributions of the thesis are:

1. Introducing a novel method to synthesise Optic nerve head region of retina along with its cup masks that works for small datasets.

2. An improved unpaired image-to-image translation method to synthesise Drusen in OCT.

1.4.3 Thesis Organisation

The aim of this thesis is to explore synthesis methods for retinal imaging, specifically for fundus and OCT. Chapter 2 Describes a novel fundus image synthesis method of optic nerve head region that leverages domain-specific knowledge to achieve anatomical correctness. Chapter 3 describes compares a different approach fundus image synthesis that generates the annotations required to the approach described in Chapter 2 and verifies the assumptions made in Chapter 2. Chapter 4 introduces a novel image to image translation method to synthesise Drusen in OCT images.

Chapter 2

Synthesis of Optical nerve head region of fundus image

A Domain specific vessel bend based Approach

The Optic Disc (OD) and Optic Cup (OC) boundaries play a critical role in the detection of glaucoma. However, very few annotated datasets are available for both OD and OC that are required for segmentation. Recently, Convolutional Neural Networks have shown significant improvements in segmentation performance. However, the full potential of CNNs is hindered by the lack of a large amount of annotated training images.

To address this issue, we explore a method to generate synthetic images which can be used to augment the training data. Given the segmentation masks of OD, OC and vessels from arbitrarily different Fundus images, the proposed method employs a combination of B-spline registration and GAN to generate high quality images that ensure that the vessels bend at the edge of the OC in a realistic manner. In contrast, the existing GAN based methods for Fundus image synthesis fail to capture the local details and vasculature in the Optic Nerve Head (ONH) region (Figure. 2.1). The utility of the proposed method in training deep networks for the challenging problem of OC segmentation is explored and an improvement in the dice score from 0.85 to 0.9153 is seen with the inclusion of the synthetic images in the training set.

2.1 Introduction

Realizing the full potential of deep learning requires training over large annotated datasets. This becomes problematic when dealing with medical images where practical issues such as patient privacy, medical regulations, annotator expertise and the cost, make creating and annotating large datasets a challenging task. One such task which suffers from the lack of annotated public datasets is Optic Cup (OC) segmentation, which plays an important role in Glaucoma detection. A lot of times this problem is tackled by using traditional image augmentation techniques like rotation crop and vertical & horizontal flipping. While these methods are somewhat helpful they create highly correlated data.

Another possible solution is to augment synthetically generated images with annotations to the training dataset. GANs [11] have shown that they are highly capable of generating completely new and diverse high resolution images. Conditional GANs have been used to synthesise images for a particular label. This method has seen a lot of success in classification tasks as shown in [10]. This can also be further extended to segmentation tasks as shown by [32]. Instead of generating from scratch, making use of annotations like vessel masks to generate images has also shown to be an effective method [6, 47].

Although the methods in [6, 47] have been successful in retaining most of the global features, they fail to capture the local details of the ONH region (Fig.2.1) like the OD border and the vessel bends. Moreover they still require a considerable amount of data compared to what is available for Optic Cup segmentation. But these methods don't leverage knowledge specific to the domain. This can be leveraged for better performing methods methods.

In case of Optic Cup segmentation this extra domain specific knowledge is that the Optic cup is a bulge in the retina. And thus since Fundus images are 2D projections of these 3D objects the vessels passing through OC boundary have a characteristic bend. Since OC is characterised by 3D depth map, the vessel bends become crucial for indicating the depth change observed at the OC boundary [17]. So instead of generating annotations the proposed method uses different OC, OD and vessel masks from arbitrarily different images as inputs to generate new realistic images. Since the masks are obtained from different images, the vessel bends and the OC edge may not coincide. To ensure that the vessels bend at the edge of the new OC, the vessels are pre-processed(explained in sub-section2.2.3). Since the provided annotations are from real images, they are more reliable and have an anatomically correct structure and thus reduce the information network has to learn in order to synthesise realistic images.

We later show the effect of generator architecture on the image quality and also the usability of those images for optic cup segmentation. In the next chapter we compare this method to a more straight froward image synthesis method where the annotations are generated along with the images.

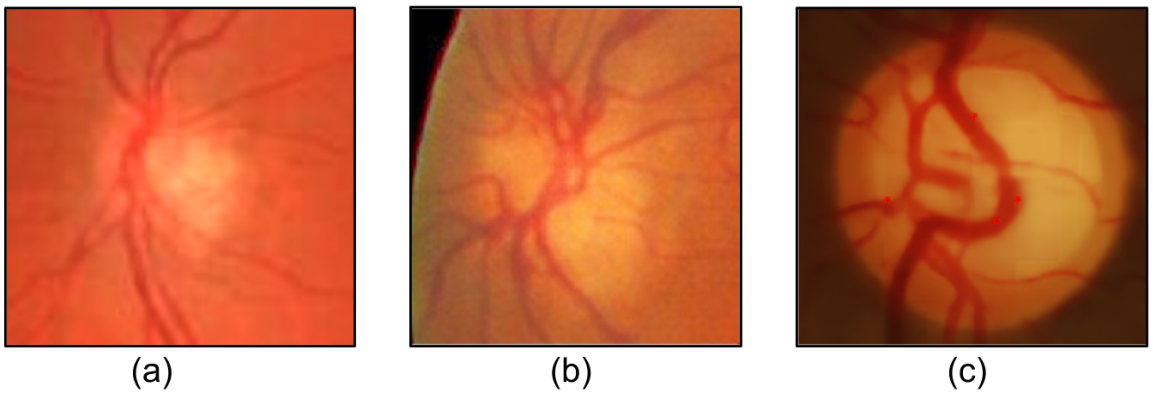

Figure 2.1 Comparison of synthetic image generation using different approaches. Left to right: Results of [6],[47] and the proposed method.

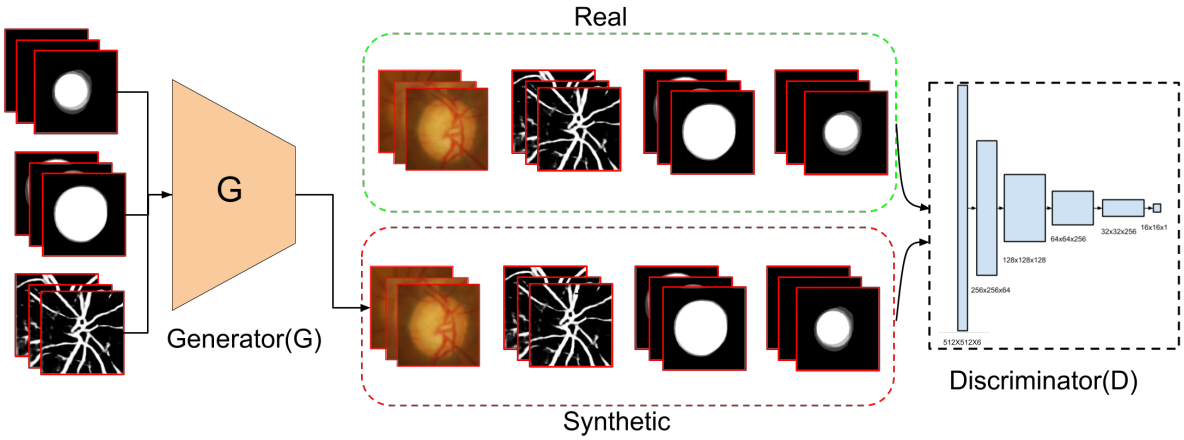

Figure 2.2 The GAN architecture used to generate images of the ONH region

2.2 Method

Given a set of OD, OC and vessel masks as input, the proposed method employs a GAN to generate realistic Fundus image of the ONH region. The network is trained on real images and their corresponding OC, OD and vessel masks. In order to get new examples, various permutations of vessel, OC and OD masks are sampled from different images and preprocessed to ensure that the sampled masks are compatible. The details of the proposed method are explained below.

2.2.1 Training GAN

The GAN architecture used to generate the synthetic images is depicted in Figure.2.2. The generator takes an OC mask, an OD mask and a vessel mask and outputs a 3-channel RGB image. We experimented with four different generators, their details are explained further in sub-section 2.2.2

For all four generator architectures we use the same discriminator. The discriminator, D, is a five layered fully convolutional network (FCN) with each convolution layer using 3x3 kernels and Leaky RelU activation function. The first layer has 32 filters and the number of filters are doubled in each successive layer. The final layer has only 1 filter and a sigmoid activation with output size of 16x16.

The two networks are trained alternately where the discriminator attempts to tell a real image from a synthetic one and the generator tries to generate images that can fool the discriminator. After one iteration (mini batch) of training the discriminator, to train the generator the discriminator layers are frozen and the network optimizes three loss functions : adversarial loss (L_{adv}) , structural loss (L_{SSIM}) and absolute difference (L_{abs}).

Here G denotes the generator and D denotes the discriminator. v , c and d are the vessel mask, OC mask and OD mask and r is their corresponding real image.

$$L_{adv}(G,D) = \mathbb{E}_{(v,c,d),r\sim p_{data}((v,c,d),r)}[log(D((v,c,d),r)] + \\ \mathbb{E}_{(v,c,d)\sim p_{data}(v,c,d)}[1 - log(D((v,c,d),G(v,c,d))]$$

$\mathbb{E}_{(v,c,d),r\sim p_{data}}$ represents the expectation of the log-likelihood of the sample $((v,c,d),r)$ being sampled from the probability distribution of real images p_{data}.

Structural similarity index (SSIM)[41] is used to compare the perceived structural similarity between generated ($G(v,c,d)$) and real (r) images.

$$SSIM(p) = \frac{2\mu_{r_p}\mu_{G(v,c,d)_p} + c1}{{\mu_{r_p}}^2 + {\mu_{G(v,c,d)_p}}^2 + c1} * \frac{2\sigma_{r_p,G(v,c,d)_p} + c2}{{\sigma_{r_p}}^2 + {\sigma_{G(v,c,d)_p}}^2 + c2}$$

where, r_p and $G(v,c,d)_p$ are patches of size 4x4 around the pixel position p; μ represents mean intensity of the patches, σ represents the standard deviation of the intensity of the patches and c1, c2 are constants. L_{SSIM} loss is defined by reducing the average SSIM between all the patches in synthesized and real images.

$$L_{SSIM} = 1 - \frac{1}{N}\sum_{p\in(G(v,c,d),r)} SSIM(p)$$

L_{abs} reduces the absolute difference between the images and captures the finer details.

$$L_{abs} = \mathbb{E}_{(v,c,d),r\sim p_{data}((v,c,d),r)}[|G(v,c,d) - r|]$$

The overall loss function is the weighted sum of above three loss functions[3].

$$Loss = L_{abs} + w1 * L_{SSIM} + w2 * L_{adv},$$

where $w1$, $w2$ determine the contribution of L_{SSIM} and L_{adv} to the total loss function. For our experiments, parameters $w1$ and $w2$ were set to 5 and 1 as we observed that higher $w1$ led to faster convergence.

2.2.2 Generator architectures

The quality of image being generated heavily depends on the choice of generator. In other pix2pix applications a plain U-Net or a Plain Resnet have been used. We arrange these networks in two different arrangements as shown in Figure. 2.3, where each of the block represents either a Unet or a Resnet. Thus, we have four generator architectures: i)$\mathcal{G}_1$-Unet, ii) $\mathcal{G}_1$-Resnet, iii) $\mathcal{G}_2$-Unet and iv) $\mathcal{G}_2$-Resnet. The models in $\mathcal{G}_1$ configuration had 4 times as many filters to their $\mathcal{G}_2$ counterparts thus had approximately the same number of parameters. The $\mathcal{G}_1$ arrangement is a plain Unet/Resnet in which the 3 inputs(OC, OD and vessel) are stacked together along the channel axis and given to the network which

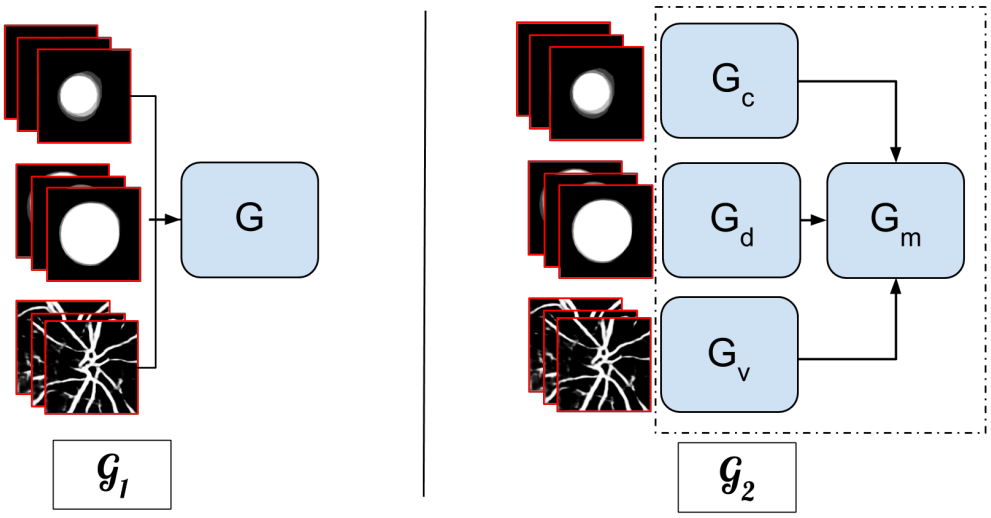

Figure 2.3 Two different arrangements for the generators each of the block is a U-net or ResNet.

outputs an RGB image. Whereas, the $\mathcal{G}_2$ arrangement has three parallel networks, G_c, G_d, G_v each of which takes in an OC mask, an OD mask and a vessel mask respectively. The output of these three networks is then provided as input to another network, G_m which merges them outputs a 3-channel RGB image. For each of the U-Nets each step of the encoder has a convolution layer with 3x3 kernel size and 1x1 stride followed by another convolution layer with 2x2 strides for down-sampling. At each downsampling step the number of filters is doubled. At each step of the decoder we have concatenation of the corresponding feature map from the encoder. This is followed by a convolution layer with 3x3 kernel size and 1x1 stride and a transposed convolution layer with 2x2 kernel size and 2x2 stride, up-sampling the feature maps. At each upsampling step the number of filters is halved. Each layer in both encoder and decoder has Leaky RelU activation and is followed by batch normalization. For $\mathcal{G}_2$The first step of the encoder has 64 filters which are doubled at each step. While, for the decoder the filters are halved at each step. The U-nets finally have a final convolution layer with 3 filters, 1x1 stride and a tanh activation. The Unet in $\mathcal{G}_1$ arrangement follows the same architecture but has four times as many filters at each layer. In the Resnet versions the ResNet has 9 ResNet blocks. Each ResNet block has 2 convolutional filters each of which have 64 filters and are followed by a batch normalization layer. The input the of the renset block is then added to the output of the second convolutional layer on which a Leaky RelU activation is applied.

2.2.3 Generating new examples

To generate new images we randomly pick vessel mask V_A, optic cup mask OC_B and optic disk mask OD_C from three random images A,B and C respectively. It is ensured that the following anatomical constraints are satisfied:

1. OC_B is contained within the boundaries of OD_C

2. The vessels V_A bend at the edge of OC_B.

The first constraint is easily verified by checking intersection of the binary masks. If the sample does not satisfy the first constraint then it is discarded. For the second constraint a transformation T is applied on the vessel mask V_A. Since the vessels in V_A bend at the edges of OC_A, T is obtained by registering OC_A to OC_B using a B-spline registration based on [35]. Then T is applied to V_A to obtain V'_A to ensure that the vessel bends in V'_A are now moved to the OC_A boundary.

The new transformed vessel V'_A along with OC_B and OD_C are given to the trained network to generate new synthetic images.

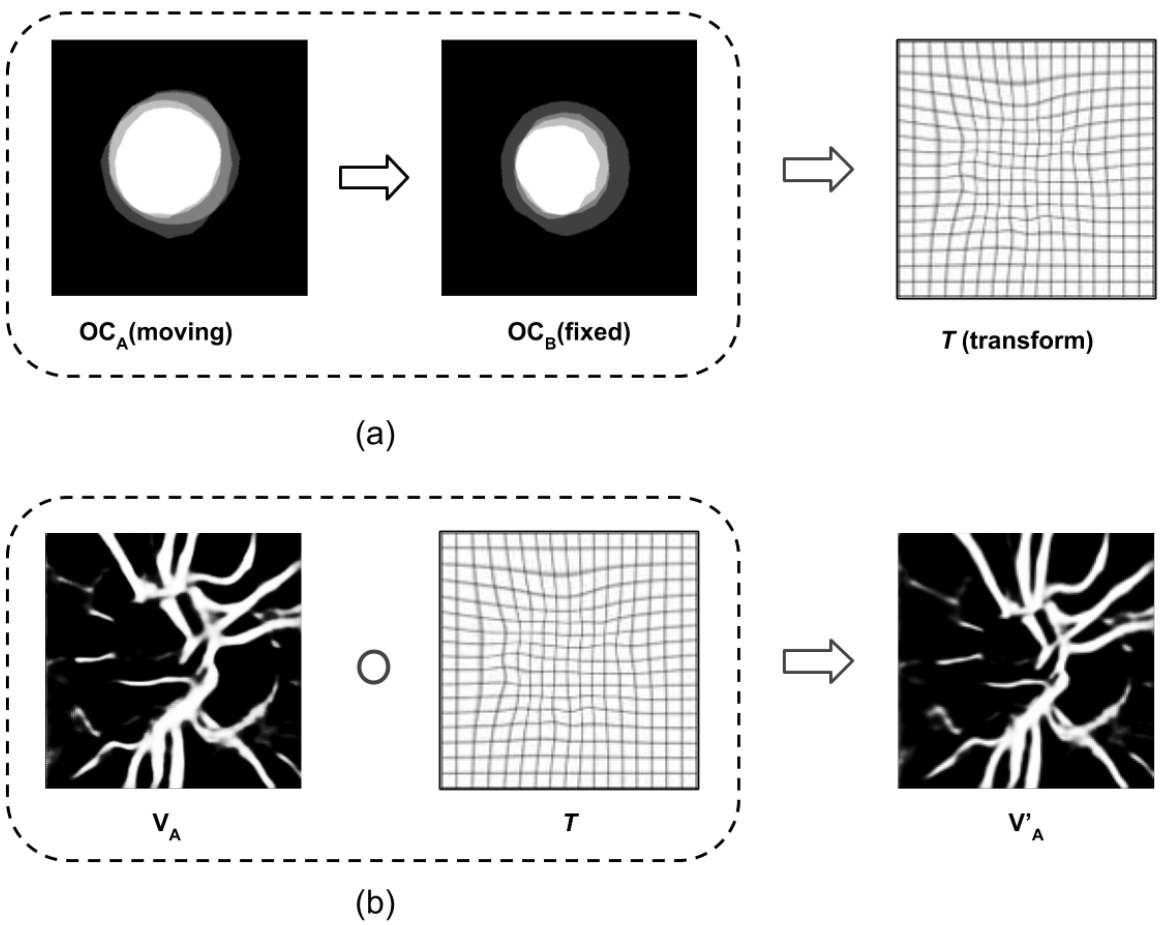

Figure 2.4 (a)B-spline registration of OC_A onto OC_B to get the transformation T, (b) Applying the transform T onto V_A to get V'_A

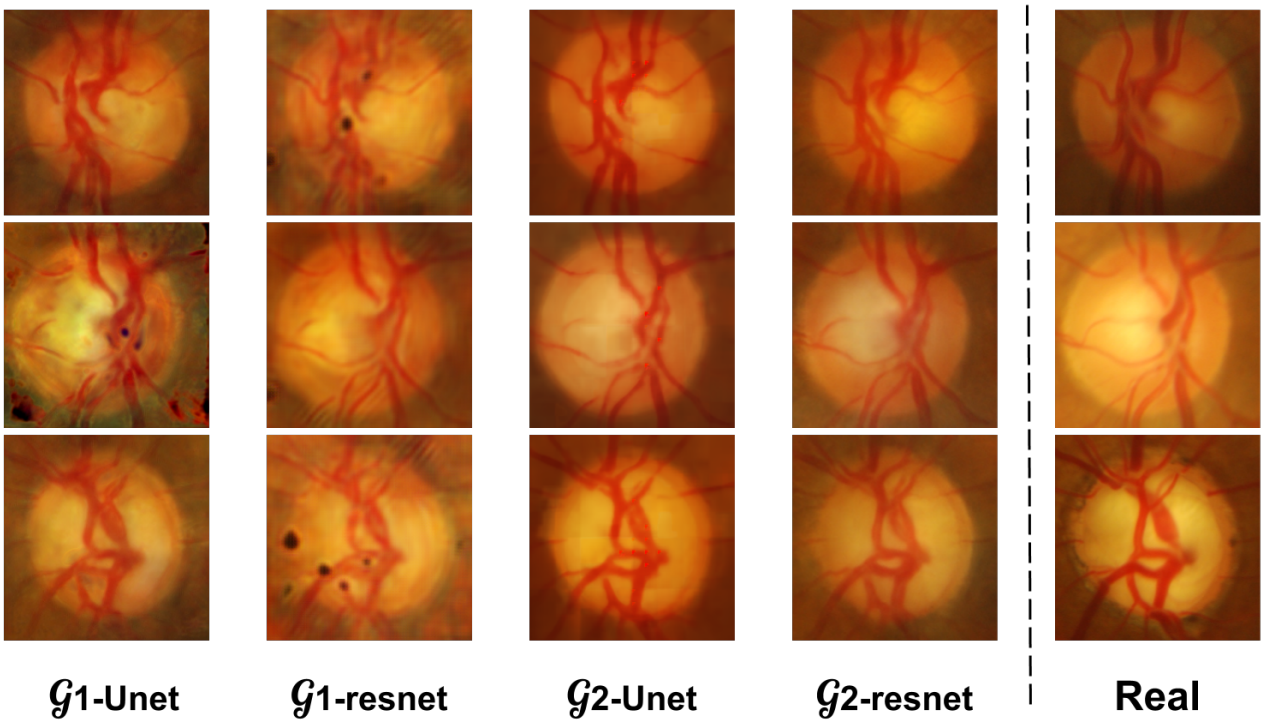

Figure 2.5 3 sample images generated using each of the four different generators compared

2.3 Experiments and Results

To train the network we used images from Drishti-GS [37] dataset. The dataset contains 101 Fundus images with annotated OC and OD masks. The images were cropped around the region of interest i.e. the ONH region with a resolution of 512x512. To get the vessel masks a vessel segmentation U-Net [34] was trained on the DRIVE dataset [39]. Once the vessels were obtained, the dataset was split into a train set containing 66 images and a test set containing 35 images. The GAN was then trained on the train set of 66 images/mask pairs for 2000 epochs. Different sets of OD, OC and vessel masks were sampled from the train set (66) and were pre-processed as explained in the sub-section 2.2.3 and finally given to the trained network to generate new examples.

The GANs were trained on a NVIDIA GTX 1080ti GPU with 11 GB of GPU memory. We show the visual and objective evaluation of the quality of generated images as well as their impact on boosting OC segmentation in the subsequent subsections.

2.3.1 Visual Quality

Figure. 2.5 shows visual comparison between images generated from these four methods and the real image corresponding to the input OC, OD and vessel masks. Both of the $\mathcal{G}_1$ models have heavy artifacts and don't look as realistic even though they have approximately the same number of param-

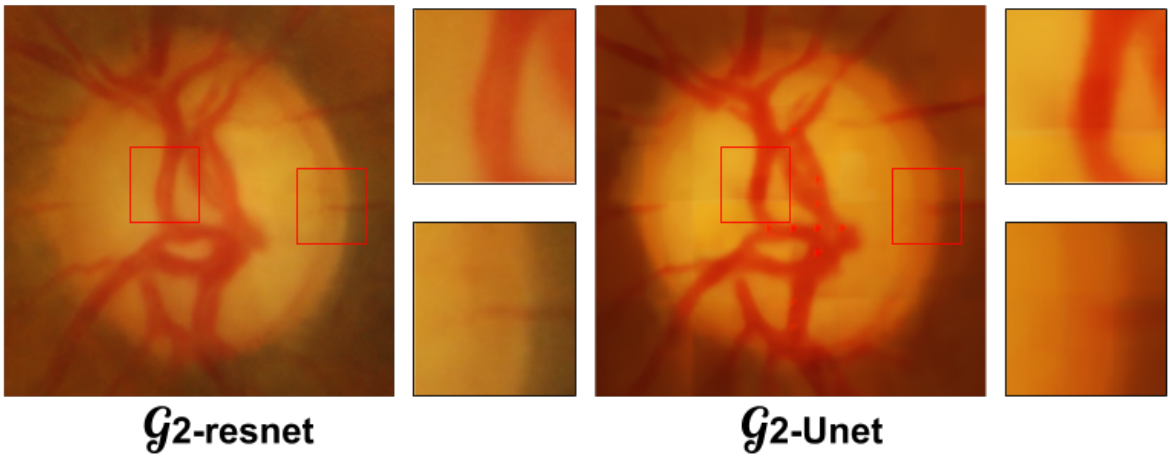

Figure 2.6 Comparison of vessel detail in the images Generated by $\mathcal{G}_2$ Resnet (left) and $\mathcal{G}_2$ Unet (right)

eters compared to their $\mathcal{G}_2$ counterparts. This is because deeper networks have the capacity to learn more complex structures as opposed to shallower with more parameters. $\mathcal{G}_2$-Unet and $\mathcal{G}_2$-ResNet have minimal artifacts and look much closer to the real images however $\mathcal{G}_2$-Unet images tend to have more contrast. This comes at the cost of having less texture detail compared to $\mathcal{G}_2$-ResNet. As seen in Figure. 2.6 the images generated by $\mathcal{G}_2$-Unet have very uniform vessels and the thinner vessels are completely missing. Moreover, the thicker vessels in the image generated by $\mathcal{G}_2$-ResNet are more transparent in the middle. The image also captures the thinner vessels very well. The periphery of the OD has greenish texture similar to real images which is missing in images generated by $\mathcal{G}_2$ Unet.

Figure. 2.7 shows synthetic images generated (last column) using the vessels from real images (first column) and OC mask sampled from a different image (middle column). The vessel bends are highlighted using green circles. The bends can be seen to have moved from the edge of the original OC to the edge of the new OC in the synthetic images.

To evaluate the objective quality of the vessels in the generated images, we use Q_v as proposed in [22]. The metric is a no reference retinal image quality metric that measures the quality of generated images based on the clarity on vessels. It has shown to have high correlation to PSNR and SSIM which are reference based image quality metrics. Table 2.1 shows the comparison between mean Q_v scores of images generated by different generators. Higher Q_v corresponds to higher quality vessels.

2.3.2 Boosting Optic Cup segmentation

For our baseline, we trained a U-Net [34] for OC segmentation for 200 epochs on the training set (66 real images). Choosing random triplets form the train set, as mentioned in the previous section, we generate N synthetic images. We then trained multiple OC segmentation U-NETs [34] with 66 real +

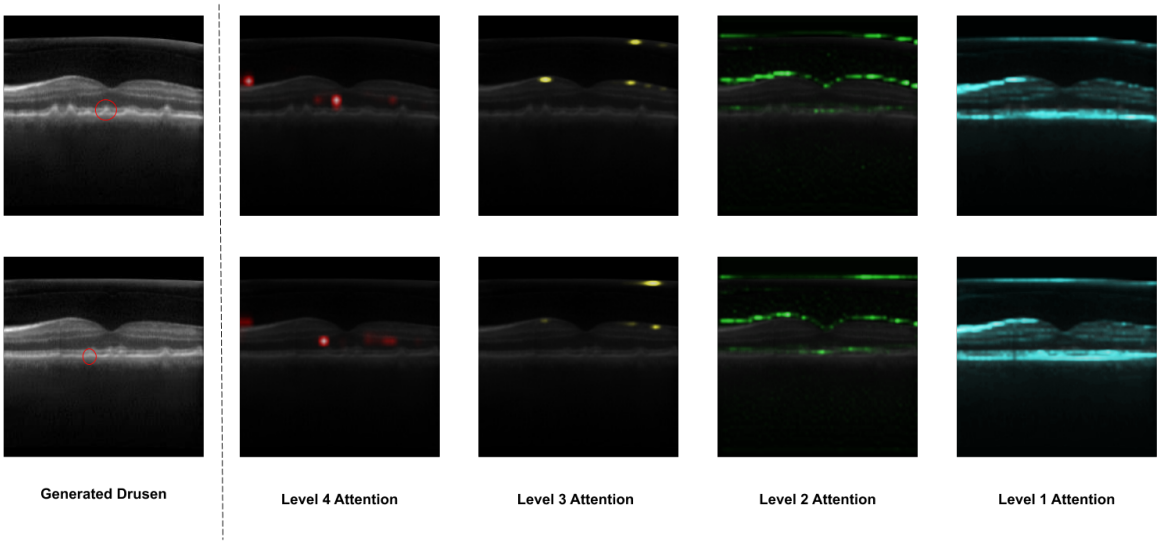

Figure 4.9 Attention maps corresponding to the regions marked by the circle in red

4.4.3 Quantitative Evaluation

We used the Frechet Inception distance(FID) score as our choice of metric for quantitative evaluation. FID is widely used to compare and assess realism on generated images. It has been shown to be consistent with human evaluation in assessing the realism and variation of the generated samples (Heusel et al., 2017). It calculates the Wasserstein-2 distance between the generated images and the real images in the feature space of a pre-trained Inception-v3 network. Lower FID implies closer distance between the synthetic and the real images. The FID between the generated Drusen (D_g) images and i) real Drusen (D_r) as well as ii) real Normal (N_r) images were computed. Ideally the FID score between N_r and D_g should be similar to the distance between N_r and D_r which is **20.93**. This is because except for the RPE layer and regions around Drusen the images are still very similar and we want those regions to be realistic as well. However, the distance between D_g and D_r images should be much closer to zero. Table 4.1 lists the FID scores computed over 1000 images for the base model and its variants. From these results, it is apparent that the proposed network outperforms the variants. The addition of SA is seen to contribute significantly to the performance.

As we can see from the table the most significant difference in performance was due to adding Self attention skip connections even though the skip connection has only one output channel compared to 64+ in vanilla models. In the case of Vanilla RescycleGAN the distance of generated Drusen images is closer to that of normal images. This might be due to the fact that the residual connection and SSIM loss in the generator is making the model lean towards learning identity mapping thus it just outputs a normal image. The improvement gained from adding a reconstruction decoder is significant because the decoder is generating Drusen based on global features learnt by the SA layer rather than applying attention on generated Drusen from the encoder.

Method	**FID: Generated Drusen & Real Drusen**	**FID: Generated Drusen & Real noramal**
Vanilla ResCycleGAN	37	34
ResCyclegan with essim	30	32
Self-Attention based Rescycle-GAN	9	30
Self-Attention RescycleGAN + recon Decoder	**4.98**	**18**
Unsupervised Attention guided GAN	17.36	27.83

Table 4.1 Effect of each addition to ResCycleGAN to the image quality

Method	**FID: Generated Drusen & Real Drusen**	**FID: Generated Drusen & Real normal**
DiscoGAN[20]	87.16	115.78
DualGAN[45]	144.10	166.42
UNIT[26]	29.07	44.54
Hiasa et al.[13]	68.65	40.50
uagGAN[28]	17.36	27.83
Ours	**4.98**	**18**

Table 4.2 Comparison with other methods image to image translation methods

Augmentation data	Area Under ROC Curve
No augmentation	0.909
uagGAN[28] images	0.914
Images generated by Our network	**0.936**

Table 4.3 Comparison of classification performance using different methods for augmentation

Our method was also compared quantitatively with other state of the art methods. The FID scores can be seen in Table 4.2. Since FID score is perceptually correlated, it heavily punishes any artifacts which result in the DualGAN[45] and DiscoGAN[20] having a very low score. The images generated by the CycleGAN trained on gradient correlation loss are much closer to N_r than D_r because the loss function preserves the edges in the synthesised images rather than in the reconstructed image from the inverse mapping network. The UNIT [26] performs comparably to the base model since it does very similar operations except for assuming a common latent space for the images. uagGAN [28] performs better than other methods since it learns attention maps to divide the image into background and foreground and applies the transform only to the foreground, thus preserving the detail in the background. Our method is seen to outperform all other methods by a considerable margin in the table as it benefits from the employment of SA.

4.4.4 Using Generated Images for Augmentation

We tested the usability of our generated images by using them as augmentation sets for Drusen classification. The train set consisted of 7,577 OCT slices containing Drusen and 50,073 OCT slices without Drusen. The test set had 1,000 images with Drusen and 1,000 without Drusen.We used a resnet-18 network pre-trained on imagenet [8]. We replaced the final dense layer with a 1 unit layer with sigmoid activation and fine-tune it to classify Drusen. We used standard augmentation practices like rotation and horizontal flipping. The threshold for classification was varied from 0.0002 to 0.9998 with a difference of 0.0002 to get the Receiver operating characteristic (ROC) curve and Area Under the Curve (AUC) was calculated.

For augmentation we used our trained model to generate 50,073 images with Drusen from available normal images. Thus, including the generated images to our train set the total number of Drusen images become 57,650.

And for comparison, we did the same process using uagGAN [28] as it had the best quality images among other methods. Thus, we have 3 classification models 1st Trained on available training data, 2nd is augmented with Drusen images generated by uagGAN [28] and the 3rd is augmented with Drusen images generated by our generator. The AUC obtained by the models can be seen in the Table 4.3. Images generated by our methods provide a significant increase in AUC.

4.5 Conclusion

A novel method for Image-to-image translation method was proposed for a noisy modality like OCT. This is the first attempt at image synthesis for a noisy case. Our design was based on the hypothesis that incorporating SA and an edge based loss function (which is robust to noise) would achieve the desired quality. The method was demonstrated by synthesising Drusen in OCT slices given Normal slices. The images generated by our method are much more realistic and have FID scores much lower than other state of the art methods. The method also manages to generate images that have upper layers lifted up in order to incorporate for the introduced Drusen, which other methods fail to achieve. We also show that our generated images provide a significant boost to classification performance when included in the training set.

Chapter 5

Conclusion

This thesis explored the practicality of synthesis in medical imaging using very small datasets in two widely used modalities, Fundus and OCT. First, we tackled the challenge of training a Fundus image synthesis network on a small dataset (110 images) which aimed at providing paired data for Optic Cup segmentation. We compared two approaches, one which uses the domain-specific knowledge of aligning vessel bends to OC boundary and another which synthesised OC masks along with the images to provide paired data. The second approach is more flexible but doesn't leverage any domain-specific knowledge. Apart from this, we approached the problem of Drusen synthesis in OCT images by modelling it as an image to image translation task.

ONH region Fundus images for improving OC segmentation exploring two different approaches. Method 1 relies on the domain-specific knowledge of the structure of ONH region i.e. the vessels bend at the OC boundary due to 3D bulging. Method 2 is aimed to synthesise OC masks along with the image given a set of vessel and OD masks. We observed that the first approach yields more reliable OC masks while the image quality remains comparable. The pairs generated using the Method 1 when used for data augmentation in an OC segmentation task, gave a higher performance boost; Method 1 increased the OC segmentation model's performance to 0.9153 DICE score compared to 2nd method's images which increase the DICE score to just 0.88. For method 2, we also tested the dependence of generated OC on the input vessel and input OD. We did this by comparing the OC generated to the OC corresponding to the input vessel and then to the OC corresponding to input OD. The DICE score between the generated OC and the OC corresponding to input vessel was significantly higher than the score between the generated OC and OC corresponding to input OD. Thus reaffirming our belief that the OC has a higher correlation with the vessel structure than with the OD.

We also tested the impact of arranging sub-blocks in the generator and the additional impact of choosing sub-blocks. We chose 2 different sub-block architectures, U-Net and ResNet and two different arrangements of those sub-blocks. We found that arranging the blocks in a hierarchical manner gives higher quality images for the same number of trainable parameters. In our experiments, ResNet gave much more cleaner and detailed images whereas the U-net gave sharper and more colour accurate images but was riddled with numerous artifacts.

With both of these methods, we demonstrated that generative methods can be trained on very small datasets with high quality data which can be used for other applications.

In Chapter 4, we modelled Drusen synthesis as an Image-to-image translation task. Synthesis in OCT hasn't been attempted and the pre-existing Image-to-image translation methods fail to synthesise realistic OCT slices. Synthesis in OCT required a network that could preserve the noise pattern of OCT while also introducing new structural changes to the image. While the added Drusen took a very small portion of the whole image, it depended on and affected the layers around the introduced Drusen. We hypothesised that any vanilla CNN based image to image translation method, due to its small receptive field, won't be able to capture the context required to generate Drusen, so we introduced Self-Attention layers to resolve this issue. Another challenge was to retain the edge detail regardless of noise. ESSIM was used as it was more robust to noise. The FID score based analysis showed that both SA and ESSIM provided a big improvement in generated image quality. Moreover, none of the state of the art Image-to-image translation methods could provide images of the desired quality.

5.1 Future Scope

Although the thesis provides promising results for both Fundus and OCT synthesis, with generated images almost indistinguishable from real ones, this is still a preliminary work and there is scope for improvement in many areas

- The ONH synthesis method was tested on a very small dataset of only 110 images of which only 66 images were used to train the synthesis method. While the vessel bend based approach's novelty comes from the fact it can be used on a small dataset, more data can be used to check if the approach of synthesising cup along with images performs better than the vessel bend based approach.

- The generated images and their corresponding OC mask's quality was tested by using them for augmentation in OC segmentation network. The OC segmentation network could have been incorporated into our network by attaching it to the generator and the gradients of OC segmentation loss could have been used to optimize the generated images for good OC segmentation.

- OCT is a 3D modality, our proposed method works on a slice level. While processing the whole volume can be impractical, including only the slices which contain Drusen can provide a practical way to obtain 3D Drusen synthesis.

- The proposed OCT synthesis method has only been tested for Drusen synthesis. This method can further be extended to synthesise other different lesions like cysts.

Related Publications

- Anurag Deshmukh, Jayanthi Sivaswamy
  Synthesis of Optical Nerve Head Region of Fundus Image accepted at 16th IEEE International Symposium on Biomedical Imaging, April 2019, Venice, Italy.

Bibliography

[1] S. Adiga and J. Sivaswamy. Matching the characteristics of fundus and smartphone camera images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 569–572. IEEE, 2019.

[2] V. Adiga and J. Sivaswamy. Shared encoder based denoising of optical coherence tomography images. In *Proceedings of 11th Indian conference on computer vision, graphics and image processing (ICVGIP'18), ACM*, 2018.

[3] P. Appan and J. Sivaswamy. Retinal image synthesis for cad development. In *International Conference Image Analysis and Recognition*, pages 613–621. Springer, 2018.

[4] G.-H. Chen, C.-L. Yang, L.-M. Po, and S.-L. Xie. Edge-based structural similarity for image quality assessment. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II. IEEE, 2006.

[5] X. Chen, C. Xu, X. Yang, and D. Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018.

[6] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2018.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 38(9):2211–2218, 2019.

[10] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 38(9):2211–2218, 2019.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] S. GUAN. Generating custom photo-realistic faces using ai. *https://medium.com/@joseph.rocca*, 2019.

[13] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato. Cross-modality image synthesis from unpaired data using cyclegan. In *International workshop on simulation and synthesis in medical imaging*, pages 31–41. Springer, 2018.

[14] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[16] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[17] G. D. Joshi, J. Sivaswamy, and S. Krishnadas. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE transactions on medical imaging*, 30(6):1192–1205, 2011.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[20] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilik, G. Michelson, and J. Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 95–100. IEEE, 2013.

[23] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 2019.

[24] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research).

[25] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3668–3677, 2016.

[26] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[27] S. Ma, J. Fu, C. Wen Chen, and T. Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2018.

[28] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018.

[29] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[30] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[31] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[32] S. Pandey, P. R. Singh, and J. Tian. An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation. *Biomedical Signal Processing and Control*, 57:101782, 2020.

[33] J. Rocca. Understanding vae. *https://medium.com/@joseph.rocca*, 2019.

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[35] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[36] S. Sedai, P. K. Roy, D. Mahapatra, and R. Garnavi. Segmentation of optic disc and optic cup in retinal fundus images using shape regression. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 3260–3264. IEEE, 2016.

[37] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 53–56. IEEE, 2014.

[38] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[39] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken. Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[40] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou. Fixing the train-test resolution discrepancy: Fixefficientnet. *arXiv preprint arXiv:2003.08237*, 2020.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[43] J. Yang, A. Kannan, D. Batra, and D. Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*, 2017.

[44] X. Yi and P. Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669, 2018.

[45] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[46] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[47] H. Zhao, H. Li, S. Maurer-Stroh, and L. Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical image analysis*, 49:14–26, 2018.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[49] J. Zilly, J. M. Buhmann, and D. Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics*, 55:28–41, 2017.