Article:
Diaz-Pinto, A, Colomer, A, Naranjo, V et al. (3 more authors) (2019) Retinal Image Synthesis and Semi-supervised Learning for Glaucoma Assessment. IEEE transactions on medical imaging, 38 (9). pp. 2211-2218. ISSN 0278-0062

https://doi.org/10.1109/tmi.2019.2903434

Retinal Image Synthesis and Semi-supervised Learning for Glaucoma Assessment

Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F Frangi

***Abstract*—Recent works show that Generative Adversarial Networks (GANs) can be successfully applied to image synthesis and semi-supervised learning, where, given a small labeled database and a large unlabeled database, the goal is to train a powerful classifier. In this paper, we trained a retinal image synthesizer and a semi-supervised learning method for automatic glaucoma assessment using an adversarial model on a small glaucoma-labelled and large unlabeled database. Various studies have shown that glaucoma can be monitored by analyzing the optic disc and its surroundings, for that reason the images used in this work were automatically cropped around the optic disc. The significance and novelty of this work are a new retinal image synthesizer and a semi-supervised learning method for glaucoma assessment based on the Deep Convolutional Generative Adversarial Network (DCGAN). To the best of the author's knowledge, an unprecedented number of publicly available images (86926 images) are used to train both methods. Synthetic images were qualitatively evaluated using t-SNE plots of features associated with the images and their anatomical consistency were estimated by measuring the proportion of pixels corresponding to the anatomical structures around the optic disc. The resulting image synthesizer is able to generate realistic cropped retinal images and the glaucoma classifier is able to classify them into glaucomatous and normal with high accuracy (AUC=0.9017). The obtained retinal image synthesizer and the glaucoma classifier could be used to generate an unlimited number of cropped retinal images with glaucoma labels.**

***Index Terms*—Glaucoma Assessment, Retinal Image Synthesis, Fundus Images, DCGAN, Medical imaging**

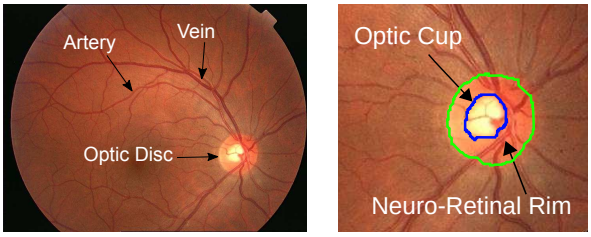
Q#5

I. Introduction

GLAUCOMA is an irreversible eye disease and it is considered the second leading cause of blindness globally [1]. It is mainly characterised by optic nerve fibre loss and that is given by the increased intraocular pressure (IOP) and/or loss of blood flow to the optic nerve. In a fundus image, the optic nerve head or optic disc can be visually separated into two zones, a bright and central zone called optic cup and a peripheral part called neuro-retinal rim. See Fig. 1.

Manuscript received Sept, 2018; revised Month Day, Year. This work was supported by the Ministerio de Economía y Competitividad of Spain, Project ACRIMA [TIN2013-46751-R] and the Project GALAHAD [H2020-ICT-2016-2017, 732613]. In particular, the work of Andres Diaz-Pinto has been supported by the Generalitat Valenciana under the scholarship Santiago Grisolía [GRISOLIA/2015/027]. The work of Adrián Colomer has been supported by the Spanish Government under a FPI Grant [BES-2014-067889]. (Corresponding author: Andres Diaz-Pinto).

Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo and Sandra Morales are with Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain. Yanwu Xu is with Artificial Intelligence Innovation Business, Baidu Inc., China. Alejandro F Frangi is with CISTIB Center for Computational Imaging & Simulation Technologies in Biomedicine, University of Leeds, Leeds LS2 9JT, UK.



(a) (b)

Fig. 1. Digital fundus images. (a) Main structures of an original fundus image and (b) Main structures of the optic disc region.

While the optic disc (OD) and cup are present in all individuals, an abnormal size of the cup with respect to the optic disc is a characteristic of a glaucomatous eye. A deep understanding of the anatomy of the optic disc is crucial for glaucoma diagnosis. For that reason, different approaches have been developed towards optic disc analysis for glaucoma assessment using retinal images. For instance, in a state-of-the-art method developed by Chen et al. [2], they used cropped images to train and evaluate a CNN obtaining an area under the ROC curve of 0.831 on a database of 650 images. However, the amount of available images is a huge problem when trying to generalise. For this reason, one of the main focus of this paper is the development of a retinal image synthesizer algorithm.

II. Background

Retinal image synthesis has been a focus of the scientific community. For instance, Fiorini et al. [3] used a system that generated the retinal background and the fovea and another system to generate the optic disc by using a large dictionary of patches with no vessels that are later registered. After that, the authors developed a complementary work that is mainly focused on vessel generation [4]. Although their method allows the generation of high-quality and large resolution images, the process of concatenating the generation of the main parts of the images is a considerable complex computational algorithm that relies on how well the images are registered.

Another approach to retinal image synthesis is the one developed by Costa et al. [5]. In their work, they trained an adversarial method on vessel networks and their corresponding retinal fundus images. In other words, they learn a transformation between the vessel trees and the retinal fundus. The main limitation of their method is the dependency of an independent algorithm to segment the vessels.

In another paper, Costa et al. presented a method which improves their previous work. Instead of learning a transformation between the vessel trees and the corresponding retinal

image, the authors used the original vessel trees to train an autoencoder. Then, the synthetic vessel trees are used as input to the retinal image synthesizer [6].

Although the latter system proposed by Costa et al. is a substantial improvement in their previous work, both methods are dependent on how well the independent method extracts the vessels. The quality of the segmented vessel tree will affect the synthetic vessel trees and then, the final retinal image. Although Costa's work is focused on synthesizing a bigger field of view, we trained their algorithm on cropped retinal images to compare the images synthesized by their method and our method.

Regarding the glaucoma assessment algorithms available in the literature, there is a great effort in pushing forward the state of the art in this area. For instance, Chen et al. [7] proposed and trained from scratch a CNN architecture to automatically classify glaucomatous fundus images using two databases: ORIGA-(light) (650 images) and SCES (1676 images), obtaining an AUC of 0.831 and 0.887 in the two databases. A study conducted by Alghamdi et al. [8] makes use of eight databases (four public and four private databases) to detect optic disc abnormalities. They developed a method using two CNNs: one CNN was trained to first classify and delimit the optic disc region and the other CNN to classify the optic disc region into normal, suspicious and abnormal classes. Another study worthy to mention was made by Orlando et al. [9], where they showed how two different CNNs, OverFeat and VGG-S, could be used as feature extractors. They also investigate how the performance of these networks behaves when Contrast-Limited Adaptive Histogram Equalization (CLAHE) and vessels deletion are applied to the fundus images. In their work, they used Drishti-GS1 database to test the performance of the fine-tuned CNNs. They observed that OverFeat CNN performed better than VGG-S, obtaining an AUC of 0.7626 and 0.7180, respectively. All these works have obtained great results in detecting glaucoma using glaucoma-labelled images. However, there are no works that take advantage of the huge amount of unlabelled data publicly available.

In this paper, we focused on the development of an image synthesizer and a semi-supervised learning method for glaucoma assessment using cropped retinal fundus images. To reach these goals, we trained two systems on 86926 retinal images cropped around the optic disc using the Deep Convolutional Generative Adversarial Network (DCGAN) [10]: an image synthesizer and a semi-supervised learning method. Synthetic images generated by our method were qualitatively compared with images generated by the Costa's method and the real images by using t-SNE plots. Moreover, quantitative evaluation was carried out by analyzing the structural properties of synthetic and real images. To do this, we measured the proportions of the area occupied by the vessel network and optic disc. The consistency in colour terms between the synthetic and real images is also measured by extracting the 2D-histogram (or chromaticity diagram) and computing the mean-squared error.

Additionally, we compared the performance of the proposed glaucoma classifier obtained from the semi-supervised learning method with the state-of-the-art algorithms. To the best of the author's knowledge, there are no works in the literature that use a semi-supervised learning method and a retinal image synthesizer that are able to generate unlimited number of glaucoma-labelled images.

III. Material and Methods

A. Material

A total of 86926 images from fourteen public databases: ORIGA-light [11], which contains 650 images, Drishti-GS1 [12], which is composed of 101 images (training and test set), RIM-ONE [13], which consists of 455 images, sjchoi86-HRF [14] with 401 images, HRF [15], which contains 45 images, DRIVE [16], which contains 40 images, MESSIDOR [17], which is composed of 1200 images, DR KAGGLE [18] with 82447 images (training and test set), STARE [19] with 195, e-ophtha [20] with 431, ONHSD [21] with 89, CHASEDB1 [22] with 28, DRIONS-DB [23] with 105 and SASTRA [24] with 34 images and a private database, ACRIMA[1], composed of 705 images were used to train the models presented in this work. All these databases are described in detail in Table I.

The reason some of the databases shown in Table I have no images for Glaucoma and Normal categories is because they were used for other tasks such as diabetic retinopathy classification or segmentation. For instance, DR KAGGLE and e-ophtha are databases especially designed for scientific research in Diabetic Retinopathy (DR). On the other hand, MESSIDOR, ONHSD, DRIVE, STARE, CHASEDB1, DRIONS-DB and SASTRA were designed for optic disc, optic cup or vessel segmentation. Q#3

All the fundus images were automatically cropped around the optic disc, except the RIM-ONE database which came originally cropped around the optic disc (See Fig. 5). To do this cropping, we employed the CNN-based method proposed in [25]. In their method, Xu et al. used a basic CNN to find the most probable pixels in the optic disc region. Then, they sort out those candidate pixels via using a threshold. The reasons we used this method are performance and because we want our pipeline to be completely CNN-based. Q#1

In order to fully covered the optic disc, we used a bounding box with ten more pixels around it. After cropping the images, the first author of this work manually discarded cropped images following the next criteria: Q#2

- Images with no presence of optic disc.
- Images with very low resolution in which optic disc is not discernible.
- Images with bright spots that occult a significant part of the optic disc.

For that reason, we used fewer images of the DR KAGGLE database (82447 instead of the 88702 images).

For all the experiments carried out in this work, the open source deep learning library Keras [26] and NVIDIA Titan Xp GPU were used.

[1]The publication of this database is pending of a journal review process

TABLE I
DATABASES USED TO TRAIN THE IMAGE SYNTHESIZER AND SEMI-SUPERVISED LEARNING METHOD

Database	Glaucoma	Normal	Total
ORIGA-light [11]	168	482	650
Drishti-GS1 [12]	70	31	101
RIM-ONE [13]	194	261	455
sjchoi86-HRF [14]	101	300	401
HRF [15]	27	18	45
ACRIMA	396	309	705
DRIVE [16]	-	-	40
MESSIDOR [17]	-	-	1200
DR KAGGLE [18]	-	-	82447
STARE [19]	-	-	195
e-ophtha [20]	-	-	431
ONHSD [21]	-	-	89
CHASEDB1 [22]	-	-	28
DRIONS-DB [23]	-	-	105
SASTRA [24]	-	-	34
	956	1401	86926

B. Generative Adversarial Network

Generative Adversarial Networks, or GAN, are deep neural network architectures comprised of two networks. One is called the generator and the other (the adversary) is called the discriminator. These two networks play a game, where the generator is trained to produce realistic samples, and the discriminator is trained to distinguish generated or synthetic data from real data. They are trained simultaneously, and the competition drives the synthetic samples to be indistinguishable from real data.

For this work, a class of CNN called Deep Convolutional Generative Adversarial Networks (DCGAN) that are based on the adversarial strategy was used. This architecture was a major improvement on the first GAN, generating better quality images and more stability during the training stage. As in the GAN network, synthetic image generation using the DCGAN mainly consists of two phases: a learning phase and generation phase. For the training phase, the generator draws samples from an N-dimensional normal distribution that run through the generator to obtain a synthetic sample and the discriminator attempts to distinguish between images drawn from the generator and images from the training set. A schema of the DCGAN architecture can be seen in Fig. 2.

C. Semi-supervised Classification

Semi-supervised classification is an area in machine learning and a special form of classification in which a large amount of unlabeled data, along with the labeled data, are used to build better classifiers. Other names for this technique are "learning from labeled and unlabeled data" or "learning from partially labeled/classified data" [27].

Semi-supervised learning has been of great interest both in theory and in practice because it requires less human effort and

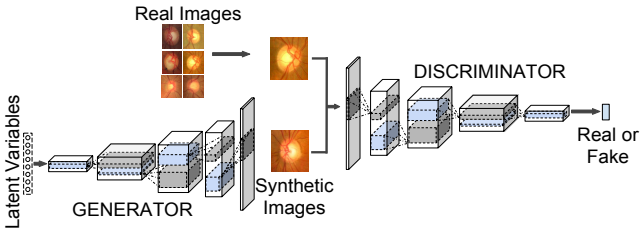


Fig. 2. Schema of the DCGAN architecture. The generator takes as input a vector of latent variables to synthesize retinal images while the discriminator tries to predict whether the input is a real or a generated image.

gives higher accuracy. Given the scarce number of glaucoma-labelled images, this technique can significantly help the development of automatic glaucoma assessment systems using retinal images. For that reason, we decided to use the power of the DCGAN to develop a semi-supervised learning method for training a glaucoma classifier and at the same time an image synthesizer.

IV. PROPOSED METHOD

As it was previously mentioned, we based our work on the DCGAN model. We followed the guidelines to construct the generator and discriminator described in the paper written by Radford et al. [10].

1) Model Architecture and Hyperparameters: The DCGAN architecture has several improvements on the vanilla GAN. Among them are the replacement of all pooling layers with strided convolutions in the discriminator and fractional-strided convolutions in the generator, the use of batch normalization (batchnorm) in both the generator and the discriminator, the replacement of fully connected hidden layers with the average pooling at the end, the use of ReLU activation in the generator for all layers except for the output and the use of LeakyReLU activation for all layers in the discriminator.

TABLE II
THE DISCRIMINATOR AND GENERATOR CNNS USED FOR RETINAL IMAGE SYNTHESIS. CONV STANDS FOR CONVOLUTION, UPCONV STANDS FOR UPCONVOLUTION, FC STANDS FOR FULLY CONNECTED AND BATCHNORM STANDS FOR BATCH NORMALIZATION.

Discriminator D	Generator G
Input 128×128 Color image	Input $\in \mathbb{R}^{100}$
5×5 conv, LeakyReLU (alpha 0.2), stride 2, Dropout 0.4	FC 32×32×256, ReLU, batchnorm
5×5 conv, LeakyReLU (alpha 0.2), stride 2, Dropout 0.4	UpSampling2D size 2 5×5 upconv, ReLU, stride 1, batchnorm
5×5 conv, LeakyReLU (alpha 0.2), stride 1, Dropout 0.4	UpSampling2D size 2 5×5 upconv, ReLU, stride 1, batchnorm
FC-1 output layer, sigmoid activation (Output for **DCGAN**)	5×5 upconv, ReLU, stride 1, batchnorm
FC-3 output layer, softmax activation (Output for **SS-DCGAN**)	5×5 upconv, Tanh, stride 1

The architecture of the image synthesis and semi-supervised learning method differs only on the last output layer (Fully connected layer) of the discriminator: one neuron for image synthesis (Synthetic or Real, **FC-1**) and three neurons for semi-supervised learning method (Normal, Glaucoma and

Synthetic class **FC-3**. See Fig. 3). The architecture details are presented in Table II.

It is worthy to highlight that in Table II are presented two different systems. One system is the DCGAN, that only synthesizes images and a second system that synthesizes and trains a glaucoma classifier (SS-DCGAN).

Regarding image resolution, we modified the architecture to handle 128×128 px, which is closer to the average resolution of the cropped retinal images. No pre-processing was applied to the training images, no data augmentation was used and class weights for the Glaucoma, Normal and Not-labelled images were set to train the semi-supervised learning method.

Although research in adversarial models continues to improve, stability on training these models is still a challenging task. For that reason, we followed the recommendations given in [28] to reach stability on training the DCGAN and the semi-supervised learning method (SS-DCGAN). Recommendations such as normalizing the input images between -1 and 1, using Stochastic Gradient Descent (SGD) optimizer for the discriminator and ADAM optimizer for the generator, using a Gaussian distribution for the latent space and mini-batches containing only all real images or all generated images were used for training the models in this work.

2) Model Losses: As in a regular GAN, the DCGAN model emulates a competition in which the Generator *G* attempts to produce realistic images, while the Discriminator *D* classifies between images from the training set with their corresponding labels and images produced by the generator. The main goal of the DCGAN model is to maximise the miss-classification error of the Discriminator while the generator produces more realistic images trying to fool the discriminator. This competition is also called a two-player minimax game and it can be described as follows:

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

where $\mathbb{E}_{x\sim p_{data}(x)}$ is the expectation over the training data and $\mathbb{E}_{z\sim p_z(z)}$ is the expectation over the data produced by the generator. $D(x)$ represents the probability that x came from the training data rather than the data produced by the generator and $G(z)$ represents the probability of z being produced by the generator. Therefore, the system is trained to minimize $\log(1-D(G(z)))$ and maximise $\log(D(x))$ [29].

However, regarding the semi-supervised learning method using the DCGAN architecture, instead of binary classification, the discriminator is transformed into a K-class classifier [30], [31]. Therefore, the semi-supervised setting loss function is composed of two parts; the supervised and the unsupervised loss function [32]:

$$L = L_{supervised} + L_{unsupervised}, \tag{2}$$

where the supervised loss is defined by the cross-entropy loss function as in a supervised learning setting with K classes:

$$L_{supervised} = -\mathbb{E}_{x,y\sim p_{data}(x,y)} \log(p_{model}(y|x, y < K+1)), \tag{3}$$

and the unsupervised loss function is, in fact, the standard GAN minimax game:

$$L_{unsupervised} = -\left\{\mathbb{E}_{x\sim p_{data}(x)} \log D(x) + \mathbb{E}_{z\sim p_z(z)} \log(1-D(G(z)))\right\}, \tag{4}$$

where $D(x) = 1 - p_{model}(y = K+1|x)$, being $p_{model}(y = K+1|x)$ the model predictive distribution and K the number of real classes.

In other words, the unsupervised loss function is computed to differentiate real training images and fake images and the supervised loss computes the individual real classes probabilities. In this work, these classes are Glaucoma and Normal.

V. Results and Discussion

In this work, we trained an image synthesizer and a semi-supervised learning method on 86926 cropped retinal images from fourteen different databases. In the process of training these models, we tested a range of N-dimensional latent spaces from 32 to 100 latent variables. Each latent space was explored in order to check that the systems do not memorise the training database and, at the same time, it generates plausible retinal images. To accomplish this goal, we used spherical interpolation to evaluate intermediate latent representation points as it was done in [6]. It turns out that using a spherical interpolation, instead of linear interpolation, better results are obtained when finding a path between two samples (z_1 and z_2) [33]. The spherical interpolation (slerp) is defined by the following equation:

$$slerp(z_1, z_2, t) = \frac{sin((1-t)\theta)}{sin(\theta)} z_1 + \frac{sin(t\theta)}{sin(\theta)} z_2 \tag{5}$$

where θ represents the angle between z_1 and z_2 and t is a value ranging from 0 to 1. For $t = 0$, the output of the slerp is equal to z_1, for $t = 1$ the slerp is z_2 and for an intermediate value of t, slerp outputs a spherical interpolated point. Examples of this exploration can be seen in Fig. 4.

It is possible to observe from Fig. 4 that all images resulting from the spherical interpolation are plausible images. This implies the latent space does not contain zones outside the manifold learned during training and the system does not memorize the training set.

Regarding the image size, all the images were rescaled into 128×128 px because this size represents the nearest power of two to the averaged height and width of a retinal image cropped around the optic disc. We utilised a power of two image size to optimise speed and computational performance.

A. Retinal Image Synthesis

Although a great effort to develop objective metrics that correlate with perceived quality measurement has been made in recent years, it is still a challenging task. In the case of quality evaluation of synthetic images, it should be specific for each application [34]. For that reason, we created a database composed of 400 images: 100 synthetic images from the DCGAN, 100 synthetic images from the SS-DCGAN, 100 images from a state-of-the-art method (Costa's method [5]),

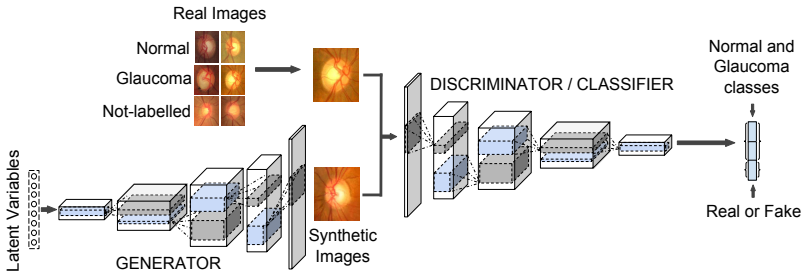


Fig. 3. Schema of the DCGAN architecture used as a Semi-supervised learning method. The DCGAN discriminator is converted into a 3-class classifier (Normal, Glaucoma and Real/Fake class).

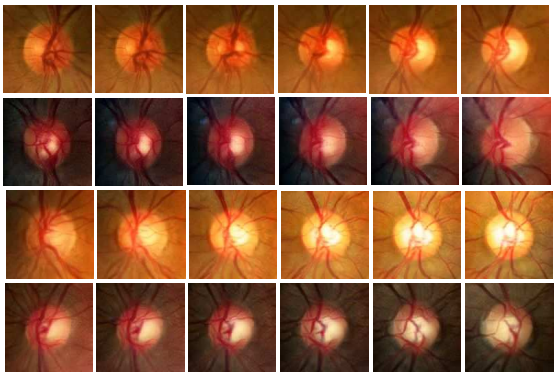


Fig. 4. Examples of the latent space exploration using the spherical interpolation.

and 100 real images (randomly selected from the training set with the exception of ORIGA-light database. It will be used for qualitative evaluation) to perform a qualitative and quantitative evaluation. The synthetic images use for this evaluation were generated after training the DCGAN for 15 Epochs and the semi-supervised learning method for 150 Epochs. To train the SS-DCGAN algorithm, we weighted the classes during the training stage because there is less number of glaucomatous images in the training set than normal and/or images without labels.

Q#7 With regards to the qualitative evaluation, we think that a good way to compare synthetic and real retinal images is by comparing the features extracted by a CNN trained to classify retinal images. Therefore, we fine-tuned the ResNet50 architecture [35] on the ORIGA-light database as a glaucoma classifier. Once this network was fine-tuned, we took 100 features for each image using a fully connected layer with 100 neurons on the top model, in which each neuron's output represents one feature.

Q#7 After obtaining the 100 features for each image, we qualitatively show with t-SNE [36] the feature differences between real images and synthetic images generated by the DCGAN, the semi-supervised learning method (SS-DCGAN) and the Costa's method [5].

Q#7 It is important to highlight that Costa's method [6] was originally presented to synthesize images with a wider field of view and fewer images. For that reason, we retrained their method, following all the recommendations given in their paper, on the 86926 cropped retinal images with a resolution of 128×128 px. Examples of images used for this comparison are shown in Fig. 5.

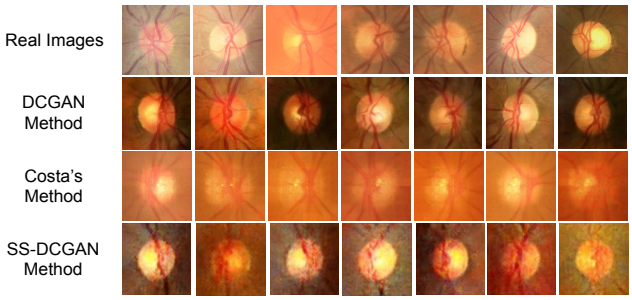
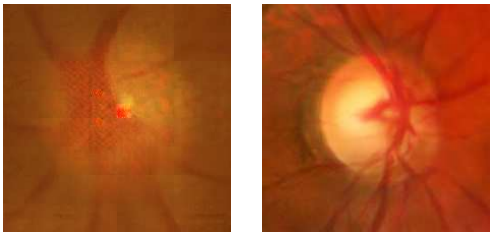


Fig. 5. Examples of real images (first row), synthetic images generated by the DCGAN method (second row), synthetic images generated by Costa's method (third row) and synthetic images generated by the Semi-supervised DCGAN (fourth row).

As it can be seen from the Fig. 5, synthetic images obtained from the DCGAN model are sharper, they present well-defined optic disc shapes, how the blood vessels clearly converge into the optic disc and right/left eye symmetry is evidenced in the resulting images. From this comparison, we found out that synthetic images from the Costa's algorithm have artifacts inside the optic disc as it is shown in Fig. 6



(a) (b)

Fig. 6. Image sample generated by the Costa's (a) and the DCGAN (b) methods. Artifacts inside the optic disc are visible on the image generated by the Costa's method.

These observations can be also qualitatively evaluated making use the t-SNE plots (See Fig. 7). From Fig. 7 it is possible to see that features of the synthetic images generated by the DCGAN architecture are closer to the real images than the other methods and the features of images generated by Costa's

method are closer to the real images than the SS-DCGAN method.

Q#7 The quality of the images generated by the SS-DCGAN was expected to be low due to this method is not only synthesizing images but also training a glaucoma classifier using labelled and not-labelled glaucomatous images. It was empirically demonstrated in [31] that a good semi-supervised learning method and a good generator cannot be obtained at the same time.

Q#1 Regarding the quantitative evaluation, we analyzed two important features: the anatomic characteristics such as vessels and the optic disc and the colour properties of the images. To evaluate the anatomic characteristics, we measured the average proportion of pixels belonging to the vessel and optic disc structures (See Table III). Optic disc masks were manually segmented by clinical experts and the vessel masks were automatically segmented using a method based on morphological operators, curvature evaluation and k-means filtering to detect the vessels [37]. The trade-off between time consuming and performance is the main reason of using this method to segment the vessels.

TABLE III
MEAN AND STANDARD DEVIATION OF PIXEL PROPORTION OCCUPIED BY THE VESSELS, OPTIC DISC, AND BACKGROUND ON THE EVALUATION DATABASE.

	Real Images	DCGAN method	SS-DCGAN method	Costa's method
Vessel proportion	0.1519 ± 0.0306	0.1431 ± 0.0306	0.2224 ± 0.0620	0.1026 ± 0.0195
Optic Disc proportion	0.2456 ± 0.0722	0.1776 ± 0.0339	0.1599 ± 0.0291	0.1851 ± 0.0396
Background	0.6025 ± 0.0795	0.6792 ± 0.0428	0.6177 ± 0.0555	0.7122 ± 0.0437

Q#8 It is possible to observe from Table III that the mean proportions between synthetic images from the DCGAN method and real images are very similar. The small difference between the mean proportion of the DCGAN and real optic discs depends on the normal variation of the optic disc size among real fundus images. Moreover, the vessel proportion obtained from the Costa's images (0.1026) is significantly less than the averaged vessel proportion obtained from the real images (0.1519). It is also possible to see that the mean vessel proportion of the SS-DCGAN images is significantly higher (0.2224) than the mean vessel proportion of the other type of images. These results could also be observed from the Fig. 6, in which vessels of the Costa's images and the optic discs of the SS-DCGAN are not as sharp as in the real or DCGAN images, which may confused the automatic vessel segmentation algorithm.

Q#1 In order to evaluate the colour properties of the synthetic and real images, we also obtained the averaged 2D-histogram [38] of real and synthetic images generated by the DCGAN, SS-DCGAN and Costa's method. These 2D-histograms are a practical way of representing the colour properties of the images, which are constructed by using the red and green channels normalized by the luminance (See Fig. 8).

It can be seen in Fig. 8 that the shape of the histogram obtained from the DCGAN and SS-DCGAN images (Fig. 8(b-c)) are more similar to the shape of the histogram obtained from the real images (Fig. 8(a)) than the shape of the histogram obtained from the images generated by Costa's method (Fig. 8(d)). This means that the colour properties of the images generated by the DCGAN and the SS-DCGAN method are closer to the properties of real retinal images.

Additionally, we calculated the mean-squared error between Q#8 the averaged 2D-histograms and the chromaticity diagram of each of the 400 images of the database (100 Real images, 100 synthetic images using the DCGAN, 100 synthetic images using the SS-DCGAN and 100 images using Costa's method). In other words, we compute for example the mean-squared error between the averaged 2D-histogram of real images and each image synthesized by the DCGAN, the SS-DCGAN and the Costa's method. The obtained results are presented in Table IV.

TABLE IV
AVERAGE AND STANDARD DEVIATION OF THE MEAN-SQUARED ERROR BETWEEN THE AVERAGED 2D-HISTOGRAMS OF THE DCGAN, SS-DCGAN, COSTA'S METHOD, AND ALL IMAGES.

Averaged 2D-histogram	Real Images	DCGAN method	SS-DCGAN method	Costa's method
Real	0.0028 ± 0.000325	0.0036 ± 0.000543	0.0090 ± 0.000540	0.0013 ± 0.000262
DCGAN method	0.0031 ± 0.000461	0.0022 ± 0.000562	0.0078 ± 0.001100	0.0016 ± 0.000439
SS-DCGAN method	0.0026 ± 0.000626	0.0045 ± 0.001100	0.0062 ± 0.001400	0.0015 ± 0.000564
Costa's method	0.0031 ± 0.000126	0.0035 ± 0.000178	0.0091 ± 0.000164	0.0010 ± 0.000163

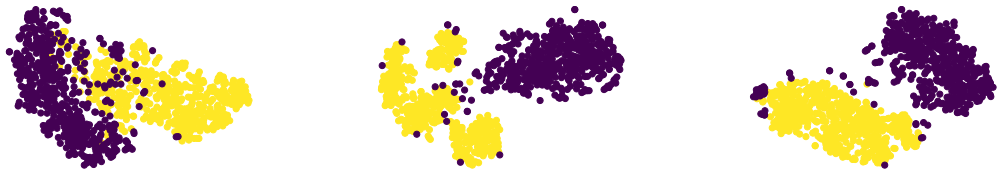
In Table IV it is possible to see that images generated by the DCGAN and SS-DCGAN method are more heterogeneous among them than the images generated by the Costa's method (0.0036 for DCGAN and 0.0090 for the SS-DCGAN images). This is represented by the mean error distance between the averaged 2-D histogram and each image generated by the Costa's method (0.0013).

B. Glaucoma Diagnosis

In the qualitative and quantitative evaluation we showed that although the SS-DCGAN system does not generate synthetic images better than the DCGAN or Costa's method, the resulting discriminator/classifier of the SS-DCGAN could be used as a glaucoma classifier. This classifier is the result of using glaucoma, normal and not-labelled images for training.

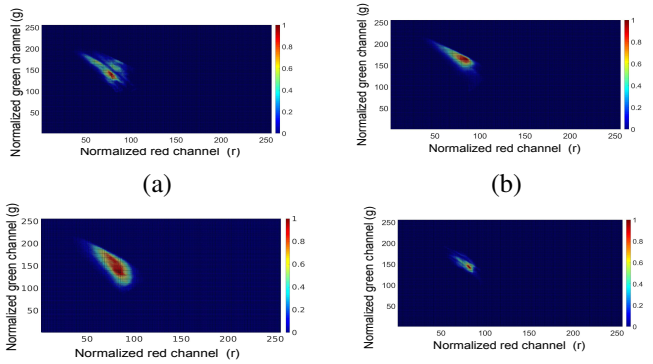
In order to test the performance of the SS-DCGAN as a Q#4 glaucoma classifier, images with glaucoma and normal labels were divided into train and test using a typical division: 70% for training (669 glaucomatous and 981 normal images) and 30% for test (287 glaucomatous and 420 normal images). Using all the unlabelled images (84569) and the 70% of the labelled images, we trained the SS-DCGAN and evaluated the performance of the resulting discriminator/classifier on the test set (30% of labelled data).

We computed the ROC curve, AUC, specificity, sensitivity Q#7 and F-score to evaluate the performance of the proposed glaucoma classifier on the test set. Moreover, the obtained results were compared with other works in the literature such as the CNNs proposed by Chen et al. [7] and Alghamdi et al. [8]. These networks were trained from scratch and tested on the same 70% and 30% of the labelled data. Additionally, we fine-tuned the ResNet50 architecture using the ImageNet weights. The obtained results from those models and our method are presented in Fig. 9 and Table V.



(a) (b) (c)

Fig. 7. t-SNE plots of features associated to the different types of synthetic images. Yellow and blue dots indicate real and synthetic features respectively. Features of synthetic images using (a) DCGAN method, (b) Costa's method and (c) Semi-supervised DCGAN (SS-DCGAN).



(a) (b) (c) (d)

Fig. 8. Averaged 2D-histograms of synthetic and real images. (a) Averaged 2D-histogram of real images, (b) Averaged 2D-histogram of synthetic images generated by the DCGAN model, (c) Averaged 2D-histogram of synthetic images generated by the SS-DCGAN model and (d) Averaged 2D-histogram of synthetic images produced by Costa's method. X-axis represents the normalized red channel and the Y-axis represents the normalized green channel.

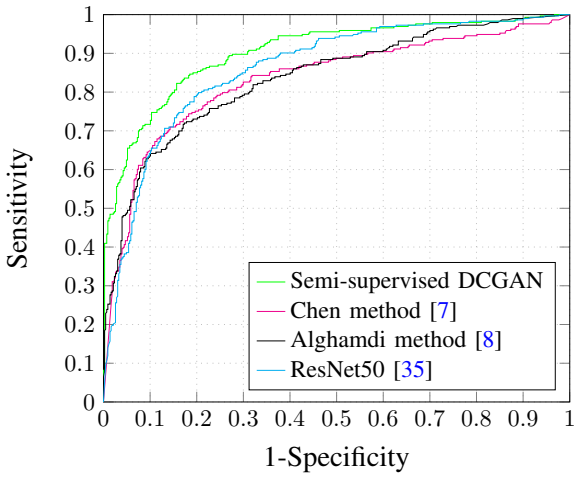


Fig. 9. ROC curve for the glaucoma classifier trained by the Semi-supervised learning method.

It is possible to see, from the Fig. 9 and Table V, that although the obtained results using the ResNet50 model, Chen's and Alghamdi's methods present a high AUC, the proposed glaucoma classifier outperforms them.

It is important to highlight that the architecture of the discriminator/classifier in the SS-DCGAN model is less complex than most of the works in the literature (4 layers). For instance, the CNN proposed by Chen is composed of 6 layers, the CNN proposed by Alghamdi is composed of 10 layers, and the ResNet50 architecture is composed of 50 layers. This improvement is given by the images without label and the synthetic images used to train the semi-supervised DCGAN. It was empirically demonstrated in [31] that generative adversarial networks used as semi-supervised learning method boost the task performance because it uses the synthetic images generated while training the discriminator/classifier.

TABLE V
COMPARISON RESULTS OF THE PROPOSED GLAUCOMA CLASSIFIER.

Model	Specificity	Sensitivity	AUC	F-score
Chen [7]	0.7440	0.8150	0.8330	0.8188
Alghamdi [8]	0.6894	0.8384	0.8365	0.8174
ResNet50 [35]	0.8055	0.7775	0.8607	0.8137
SS-DCGAN	0.7986	0.8290	**0.9017**	0.8429

We made publicly available a dataset of 10.000 images synthesized by the DCGAN and 10.000 samples synthesized by the SS-DCGAN. Labels to the synthetic images were given by the SS-DCGAN classifier and all cropped images used for training our models were also made publicly available at the following link https://figshare.com/s/6e4cbba780b81a59964c

VI. CONCLUSIONS

In this paper, a generative model was trained on cropped retinal images from one private and fourteen public databases (86926 retinal images). In contrast to other approaches to retinal image synthesis, the model presented in this work does not need previous vessel segmentation to generate images and the number of retinal images used during training is significantly greater than any other work in the literature. Qualitative and quantitative evaluation were carried out on the obtained synthetic images, showing an improvement in quality when comparing with the current works in the literature.

Additionally to the image synthesizer, a semi-supervised learning method based on the DCGAN architecture was trained on the 86926 cropped retinal images. An AUC of 0.9017 was obtained from the proposed SS-DCGAN model. After the comparison made with the current works in the literature, the obtained results demonstrate that our method could be used as computer-aided glaucoma diagnosis system.

In summary, a system capable of generating high plausible cropped retinal images and a high discriminative glaucoma

classifier could be used to generate an unlimited number of glaucoma-labelled images. It is possible to utilise the images generated by the DCGAN model and make the SS-DCGAN model to put a label to the generated images.

Future work will focus on using advance generative adversarial networks such as image-to-image translation methods for retinal image synthesis and semi-supervised learning with the aim of improving both the quality of the generated images and the glaucoma classification.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] World Health Organization, "Bulletin of the World Health Organization, Volume 82, Number 11," http://www.who.int/bulletin/volumes/82/11/en/infocus.pdf?ua=1, 2004, accessed 2017-08-01.

[2] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu, "Automatic Feature Learning for Glaucoma Detection Based on Deep Learning," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015, pp. 669–677.

[3] S. Fiorini, M. D. Biasi, L. Ballerini, E. Trucco, and A. Ruggeri, "Automatic Generation of Synthetic Retinal Fundus Images," in *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference*. The Eurographics Association, 2014.

[4] L. Bonaldi, E. Menti, L. Ballerini, A. Ruggeri, and E. Trucco, "Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network," *Procedia Computer Science*, vol. 90, no. Supplement C, pp. 54 – 60, 2016.

[5] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendona, and A. Campilho, "End-to-end Adversarial Retinal Image Synthesis," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2017.

[6] P. Costa, A. Galdran, M. Inês Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, "Towards Adversarial Retinal Image Synthesis," *arXiv: 1701.08974*, Jan. 2017.

[7] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 715–718.

[8] H. S. Alghamdi, H. L. Tang, S. A.Waheeb, and T. Peto, "Automatic optic disc abnormality detection in fundus images: A deep learning approach," in *OMIA3 (MICCAI 2016)*, 2016, pp. 17–24.

[9] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, "Convolutional neural network transfer for automated glaucoma identification," in *SPIE Proceedings*, vol. 10160, 2017, pp. 101 600U–101 600U–10.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv: 1511.06434*, Nov. 2015.

[11] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug 2010, pp. 3065–3068.

[12] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, Ujjwal, and S. T. A., "Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation." in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI).*, 2014, pp. 53–56.

[13] E. Medina-Mesa, M. Gonzalez-Hernandez, J. Sigut, F. Fumero-Batista, C. Pena-Betancor, S. Alayon, and M. G. de la Rosa., "Estimating the amount of hemoglobin in the neuroretinal rim using color images and OCT." *Current Eye Research.*, vol. 41, no. 6, pp. 798–805, 2015.

[14] sjchoi86, "sjchoi86-HRF Database," https://github.com/sjchoi86/retina_dataset/tree/master/dataset, 2017, accessed: 2017-07-02.

[15] 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 *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems.*, 2013, pp. 95–100.

[16] 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*, vol. 23, no. 4, pp. 501–509, 2004.

[17] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed database: the messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014. [Online]. Available: http://www.ias-iss.org/ojs/IAS/article/view/1155

[18] "Kaggle diabetic retinopathy competition," https://www.kaggle.com/c/diabetic-retinopathy-detection/data, 2015, accessed: 2018-02-05.

[19] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, March 2000.

[20] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Lay, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196 – 203, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1959031813000237

[21] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, and L. Kennedy, "Optic nerve head segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 2, pp. 256–264, Feb 2004.

[22] C. G. Owen, A. R. Rudnicka, C. M. Nightingale, R. Mullen, S. A. Barman, N. Sattar, D. G. Cook, and P. H. Whincup, "Retinal Arteriolar Tortuosity and Cardiovascular Risk Factors in a Multi-Ethnic Population Study of 10-Year-Old Children; the Child Heart and Health Study in England (CHASE)," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 31, no. 8, pp. 1933–1938, 2011.

[23] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. M. de-la Casa, "Identification of the optic nerve head with genetic algorithms," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243 – 259, 2008.

[24] K. Narasimhan, K. Vijayarekha, K. JogiNarayana, P. SivaPrasad, and V. SatishKumar, "Glaucoma Detection From Fundus Image Using Opencv," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 24, pp. 5459–5463, 2012.

[25] P. Xu, C. Wan, J. Cheng, D. Niu, and J. Liu, "Optic Disc Detection via Deep Learning in Fundus Images," in *Fetal, Infant and Ophthalmic Medical Image Analysis*. Cham: Springer International Publishing, 2017, pp. 134–141.

[26] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015, accessed: 2017-05-21.

[27] X. Zhu, "Semi-Supervised Learning Literature Survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[28] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, "How to Train a GAN? Tips and tricks to make GANs work," https://github.com/soumith/ganhacks, 2016.

[29] 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 27*. Curran Associates, Inc., 2014, pp. 2672–2680.

[30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," *ArXiv e-prints*, Jun. 2016.

[31] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good Semi-supervised Learning That Requires a Bad GAN," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 6510–6520.

[32] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3581–3589.

[33] T. White, "Sampling Generative Networks," *arXiv: 1609.04468*, Sep. 2016.

[34] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *International Conference on Learning Representations*, Apr 2016. [Online]. Available: http://arxiv.org/abs/1511.01844

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, no. 9, pp. 2579–2605, 2008.

[37] S. Morales, V. Naranjo, A. Navea, and M. Alcañiz, "Computer-Aided Diagnosis Software for Hypertensive Risk Determination Through Fundus Image Processing," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1757–1763, Nov 2014.

[38] A. Colomer, V. Naranjo, and J. Angulo, "Colour normalization of fundus images based on geometric transformations applied to their chromatic histogram," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3135–3139.