Generative Image Inpainting for Retinal Images using Generative Adversarial Networks*

Lucie Charlotte Magister[1] and Ognjen Arandjelović[2]

Abstract— **The diagnosis and treatment of eye diseases is heavily reliant on the availability of retinal imagining equipment. To increase accessibility, lower-cost ophthalmoscopes, such as the Arclight, have been developed. However, a common drawback of these devices is a limited field of view. The narrow-field-of-view images of the eye can be concatenated to replicate a wide field of view. However, it is likely that not all angles of the eye are captured, which creates gaps. This limits the usefulness of the images in teaching, wherefore, artist's impressions of retinal pathologies are used. Recent research in the field of computer vision explores the automatic completion of holes in images by leveraging the structural understanding of similar images gained by neural networks. Specifically, generative adversarial networks are explored, which consist of two neural networks playing a game against each other to facilitate learning. We demonstrate a proof of concept for the generative image inpainting of retinal images using generative adversarial networks. Our work is motivated by the aim of devising more realistic images for medical teaching purposes. We propose the use of a Wasserstein generative adversarial network with a semantic image inpainting algorithm, as it produces the most realistic images.**

***Clinical relevance*— The research shows the use of generative adversarial networks in generating realistic training images.**

I. INTRODUCTION

Retinal images, also known as fundus images, capture the appearance of the inner surface of the eye, where the retina and optical disk are located [1]. The images are crucial for the diagnosis of eye diseases and observing the progression of treatment. High quality, wide field of view images can be captured using expensive, state-of-the-art ophthalmoscopes. In order to make fundus imaging more accessible, especially in countries with emerging economies, lower-cost alternatives such as the Arclight have been developed [2]. However, a limitation of many lower-cost alternatives is that they often only produce a narrow field of view. This means that multiple images of the eye must be taken to capture it fully. These images can be concatenated to obtain a more holistic view, replicating the wide field of view. However, it is likely that gaps exists, as not all angles of the eye are covered. Teaching material for opthalmoscopes, such as the Arclight, are simply artist's impressions of the inner surface of the eye. Our work is motivated by the aim of creating more realistic retinal images for teaching through the automatic, machine learning-driven inpainting of missing areas.

Generative image inpainting is the automatic completion of gaps in images using generative models. Generative models learn to synthesise suitable images based on sample images, which can then be used for inpainting. Recent research has produced a number of inpainting algorithms, which achieve a high level of realism and fewer boundary artefacts by employing generative adversarial networks (GANs). GANs consist of two neural networks, a generator and a discriminator [3]. The generator synthesises artificial image content, while the discriminator differentiates between real and fake images. The two neural networks compete against each other, facilitating learning. We demonstrate a proof of concept for the automatic completion of retinal images using GANs.

We propose a Wasserstein GAN (WGAN) for the generation of synthetic image content, as it has been shown capable of learning and producing diverse and realistic output. The artificial images synthesised by the generator are used to semantically inpaint the retinal images based on contextual and perceptual loss in regards to individual pixels. The success of the inpainting is demonstrated quantitatively and qualitatively.

II. BACKGROUND

A. Generative Models

Generative models are an unsupervised learning technique, where the model learns to synthesise data exhibiting the same properties and structure as the examples in the training dataset. In regards to generative image inpainting, generative models are utilised to synthesise image material for the inpainting of missing areas in an image. Recent research has put forward the use of GANs. GANs are a type of neural network architecture, which can be used to produce synthetic image material for inpainting. The focus on GANs springs from the promising potential of being able to accurately model the probability distribution of data in a dataset [4].

GANs are a deep learning framework proposed by Goodfellow *et al.* [3]. The basic idea behind GANs is that two neural networks, called the generator and discriminator, play a game against each other [3]. The generator is a generative neural network, which manufactures artificial samples. The discriminator is a discriminative neural network, which aims to differentiate between the artificial samples of the generator and samples from the real data distribution. In literature, this is often explained as the generator being a forger trying to

*This work was not supported by any organization.

[1] Lucie Charlotte Magister is with the School of Computer Science, University of St Andrews, St Andrews, KY16 9SX, United Kingdom charlottemagister@googlemail.com

[2] Dr. Ognjen Arandjelović is with the School of Computer Science, University of St Andrews, St Andrews, KY16 9SX, United Kingdom oa7@st-andrews.ac.uk

produce a fake picture, while the discriminator is the police trying to identify counterfeits [3].

Learning in GANs is driven by the generator aiming to fool the discriminator. Let x be the real sample data to be imitated. In order to allow the generator to learn the distribution p_g over the real data x, we define the distribution $p_z(z)$, where z is a vector of random noise [3]. The generator $G(z,\theta_g)$ maps the random input noise z to a data space [3]. $G(z,\theta_g)$ is a differentiable function with respect to its parameter θ_g, which can be implemented via a multilayer neural network [3]. The output of the generator is a fake sample produced from the learned data distribution.

To play against the generator, the discriminator must learn to differentiate between real and fake data samples. The discriminator is defined as $D(x,\theta_d)$, a differentiable function in respect to the parameter θ_d [3]. The output of the discriminator is the probability of the given data sample x being real or fake.

The generator and discriminator contest against each other by respectively trying to minimise and maximise the shared loss function $L(z) = log(1-D(G(z)))$ [3]. The goal state is for the discriminator to converge at an error of 0.5, as this means it cannot differentiate between real and fake data samples [5]. In this case, the generator has reached its optimal state, where $p_g = p_r$ [5]. Gradient descent and backpropagation can be used for training. This is one of the main advantages of GANs, as other architectures require more complex algorithms such as Markov chains [3].

B. Shortcomings of Generative Adversarial Networks

Whilst the theory behind GANs makes a compelling case in regards to their strength and versatility, they are complicated to train. Special attention must be paid to the configuration of the hyperparameters of the model, as GANs are prone to a number of failure modes. An issue faced by GANs is finding the Nash equilibrium [5]. It is hard to reach a Nash equilibrium between the two adversaries, as the discriminator and generator are trained independently causing oscillations in the respective error [5]. Architectures, such as a deep convolutional GAN (DCGAN) [6], have been proposed to address this issue.

Another common problem in training GANs is the vanishing gradient problem. On one side, if the discriminator performs poorly, the generator cannot be trained well as it depends on the accuracy of the discriminator [5]. On the contrary, if the discriminator performs too well the loss drops close to 0, which slows or prohibits the learning of the generator [5]. This failure mode can easily be identified by examining the error in the discriminator and the output produced by the generator.

Another significant failure mode of GANs is mode collapse. Mode collapse is the failure of the generator to produce a variety of samples [5]. The generator learns a particular subset of samples which fool the discriminator and begins producing these over and over again. A number of solutions to this problem have been put forward, such as the WGAN [7] and unrolled GAN [8] architectures.

C. Related Work

The use of GANs in image inpainting for retinal images is novel. Current work on the inpainting of blood vessels in retinal images exists, however, it does not use generative adversarial networks, but a recursive least square dictionary learning algorithm [9]. Current work using GANs for the synthesis of retinal images uses annotated structural drawings of the retina for the synthesis [10]. While this produces highly realistic results, it is ill-suited for the task of inpainting as a high variety of structures must be learned without additional information [10].

III. PROPOSED METHOD

The approach is sectioned into two parts. The first part focuses on deriving a GAN architecture, which produces artificial retinal images of high quality. The second step focuses on completing the original image through an inpainting algorithm using the images generated.

A. Data

The dataset used for training and inpainting is the "Longitudinal diabetic retinopathy screening data" dataset made available by the Rotterdam Ophthalmic Institute [11]. The dataset contains retinal images of 140 eyes, which we split into a training and validation set. Images of the required size are extracted by sliding a fixed window across the large images. Only one sample image per eye is used to avoid duplicate samples.

B. Image Generation

We explored a number of GAN architectures to obtain synthetic images of the highest quality. We found that the original and unrolled GANs were unable to learn to produce sufficiently realistic retinal images, whereas the DCGAN exhibited mode collapse. We deem the WGAN architecture most suitable for the task, as it overcomes these issues and produces images of the highest quality. The code is available at: `https://github.com/CharlotteMagister/GenerativeImageInpainting`.

We chose the WGAN architecture for its robustness. The generator and discriminator follow the architecture of the DCGAN proposed by Radford et al. [6]. However, the loss functions used to train the generator and discriminator are adapted to overcome the issue of mode collapse. WGANs use the Wasserstein distance to quantify the distance between the probability distribution of the training data and the probability distribution modelled by the generator. The discriminator aims to learn the parameter w yielding a suitable K-Lipschitz continuous function f_w from the function family $\{f_w\}_{w \in W}$, allowing to estimate the Wasserstein distance [7]. The generator learns to produce more realistic samples by trying to minimise the Wasserstein distance [7]. However, the WGAN's discriminator cannot be used for inpainting. This is due to the discriminator estimating the parameter w of the K-Lipschitz continuous function rather than the probability of the image being real. Therefore, the design is extended to also train the original discriminator of the

DCGAN. The actual discriminator of the WGAN used for training the generator will therefore be called the critic. Only the critic is involved in the minimax game with the generator, while the discriminator is solely trained for later use in the implementation of the inpainting algorithm. The discriminator is trained in the same manner as in the DCGAN design [6], as its architecture remains unchanged.

C. Image Inpainting

On a broad level we chose to adopt the semantic image inpainting algorithm proposed by Yeh *et al.* [12]. This decision is based on the nature of the specific inpainting challenge at hand. Specifically, it splits the inpainting problem into two tasks, allowing to explore which GAN architecture is best suited for synthesising retinal images independently from inpainting. Moreover, inpainting retinal images requires special attention to image semantics in order to create a high level of realism. For example, the veins should be continuous, wherefore, the context and overall realism of the image must be taken into account. For this task, the summation of the context loss and prior loss is promising [12]. The context loss captures the relation between the inpainted pixels and the surrounding pixels, while the prior loss takes a more holistic approach in punishing the overall realism of the inpainted image [12].

The images inpainted are cropped to 64×64 pixels. Only images of 64×64 pixels can be inpainted, as the inpainting algorithm's loss function requires the selected image and the image synthesised by the GAN to be of the same size. We generate a random mask to remove a section of the image. This artificial creation of gaps in the image allows to perform the proof of concept, showing that retinal images can successfully be inpainted to a certain degree. We then use this image with the hole in the training algorithm to find the best image for later inpainting.

Except for small changes, we closely follow the algorithm described by Yeh *et al.* [12]. The recommended value for λ, the fraction of the prior loss used, is 0.003 [12]. We increase this value to 0.01, as the overall coherence of image features, such as veins, is important. The original algorithm also proposes to stop training after 1500 epochs [12]. As the error still significantly reduces after this, we define a flexible stopping condition. We determine the learning rate experimentally to be 0.1, as this produces the fastest inpainting results without oscillations. Moreover, we explore two blending techniques to remove boundary artefacts. The first blending technique explored is Poisson blending. An alternative blending method applied is the simple weighted averaging of pixels.

D. Experimental Setup

We determine suitable hyperparameters for the models experimentally. The recommended values were used initially. These are then adapted one at a time based on the model's performance. All experiments are run on a machine with 16 GB of RAM using a GeForce GTX 1060 6GB GPU.

IV. RESULTS

GANs cannot be tested and evaluate in the traditional sense of measuring accuracy. This is due to the generated images not constituting a ground truth. The most prevalent method for model evaluation is the examination of the output images and behaviour of the error over time [13]. In order to observe the performance of the model, we perform a qualitative comparison of the visual output. Moreover, we track the progress of training by graphing the loss of the generator and discriminator over a number of epochs.

A. Image Generation

Figure 2 displays the fake images produced by the WGAN generator next to a batch of real images. The artificial images have a high degree of realism and do not appear to repeat. A variety of real images are modelled successfully in regards to the general features portrayed. For example, the optical disk is replicated, as well as different vein structures. Moreover, the hue varies as observed in the real image batch. This output was produced after 80000 epochs.

In reference to Figure 1, it can be observed that training is unstable, as there are oscillation in the error. However, the error almost appears to converge, as the oscillations become smaller when comparing the loss at around 20000 epochs and 80000 epochs. Compared to other architectures, training is more stable. We determine the optimal number of training epochs at 80000, as the images did not appear to gain in quality significantly after this. When examining the images closely, it can be seen that small artefacts remain and the image is slightly more pixilated than the original images.

We put forward the WGAN architecture as the best GAN for synthesising retinal images, as a high variety of realistic images is produced. To quantify the quality of the images produced in comparison to the real images, we calculate the signal-to-noise ratio (SNR) for a batch of images and measure the accuracy of the discriminator. The SNR measures the amount of noise in an image [14]. We calculate the average amount of noise in a batch of real and fake images to be 1.364 and 1.443 decibel, respectively. It can be deduced that the real and fake images are very close in quality, with the fake images having only a slightly higher amount of noise. Lastly, we verify the success of the image generation through the discriminator. An accuracy of 50% is obtained for the discriminator, which is desirable as it shows that the generator successfully fools the discriminator with its images, vouching for the image quality.

B. Image Inpainting

Figure 3 shows an inpainting result produced for an image drawn from the evaluation dataset. The inpainted image depicted in Figure 3c is produced within 10000 epochs of training. The error initially drops quickly, but then begins to plateau, indicating convergence. The final result produced completes the missing gap to some level of detail. The general colour of the image is replicated. Moreover, features such as the optical disk and veins are replicated too. Nevertheless, the inpainted patch appears more pixelated than the

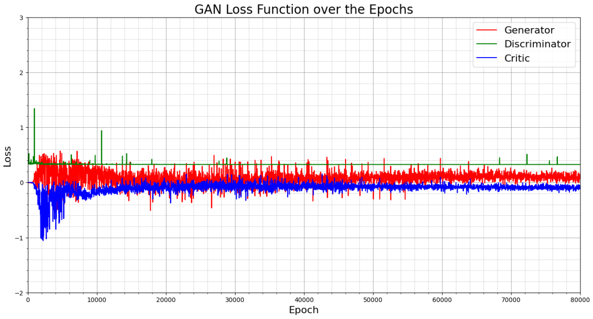


Fig. 1: The loss of the generator and discriminator of the WGAN over 80000 epochs.

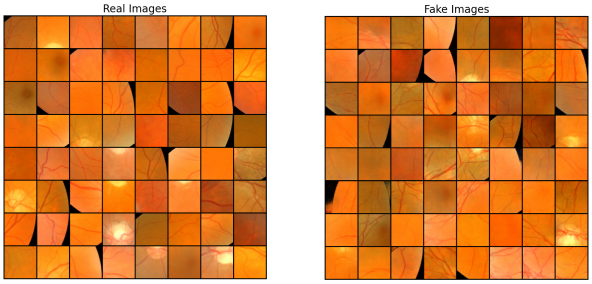
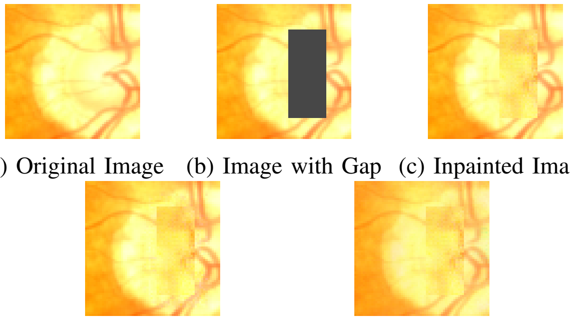


Fig. 2: A batch of real and fake images, sampled from the dataset [11] and trained WGAN generator, respectively.

original image in general. Poisson blending does not appear to improve the output image, while simple blending reduces the boundary artefact slightly, as shown in Figure 3d and 3e respectively. In conclusion, the inpainted section matches the colour of the rest of the image and produces rough features, but fails to replicate fine details.



(a) Original Image (b) Image with Gap (c) Inpainted Image

(d) Poisson Blended Image (e) Blended Image

Fig. 3: The image inpainting process.

Comparing the example output displayed in Figure 3c with artist's impressions, it can be argued that a higher level of realism is achieved, as a real retinal image is inpainted and various structural properties are replicated. However, the inpainted area is more pixelated and has lower resolution. Significant boundary artefacts are produced, even when blending is applied. This draws attention to the inpainted section. Potential improvements on the implementation could improve the resolution and final inpainting produced. Nevertheless, the application of generative image inpainting to retinal images arguably produces more truthful retinal images for teaching.

V. Discussion

We propose a method using a WGAN for inpainting holes in retinal images based on contextual and prior loss. We successfully demonstrate a proof of concept of the synthesis of feigned retinal images and the inpainting of gaps in retinal images. The inpainting results outperform the artist's impression of the retina in some areas. A drawback of the implementation is that it is limited by the image size produced by the GAN, which is an area of future work. Possible future work also includes producing an end-to-end solution for inpainting stitched retinal images. In conclusion, the work has profound implications for the synthesis of high-quality teaching material for medicine and biology.

References

[1] Color Fundus Photography. https://ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/color-fundus-photography/.

[2] K. D. C. Viquez, O. Arandjelovic, A. Blaikie, and I. A. Hwang, "Synthesising Wider Field Images from Narrow-Field Retinal Video Acquired Using a Low-Cost Direct Ophthalmoscope (Arclight) Attached to a Smartphone," in *16th IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, October 2017, pp. 90–98.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, Montreal, Quebec, Canada, December 2014, pp. 2672–2680.

[4] A. Karpathy, P. Abbeel, G. Brockman, P. Chen, V. Cheung, R. Duan, I. Goodfellow, D. Kingma, J. Ho, R. Houthooft, T. Salimans, J. Schulman, I. Sutskever, and W. Zaremba. (2016, March) Generative Models. https://openai.com/blog/generative-models/.

[5] L. Weng. (2019, April) From GAN to WGAN. https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *4th International Conference on Learning Representation (ICLR)*, San Juan, Puerto Rico, November 2016.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, Sydney, Austrailia, December 2017, pp. 214–223.

[8] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled Generative Adversarial Networks," in *5th International Conference on Learning Representation*, Toulon, France, May 2017.

[9] A. Colomer, V. Naranjo, K. Engan, and K. Skretting, "Retinal vessel inpainting using recursive least square dictionary learning algorithm," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Orleans, France, November 2015, pp. 429–433.

[10] H. Zhao, H. Li, S. Maurer-Stroh, and L. Cheng, "Synthesizing retinal and neuronal images with generative adversarial nets," *Medical Image Analysis*, vol. 49, pp. 14–26, October 2018.

[11] K. Adal, P. van Etten, J. Martinez, L. van Vliet, and K. Vermeer, "Accuracy Assessment of Intra and Inter-Visit Fundus Image Registration for Diabetic Retinopathy Screening," *Investigative Ophthalmology and Visual Science*, vol. 56, pp. 1805–1812, March 2015.

[12] R. A. Yeh, C. Chen, T.-Y. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic Image Inpainting with Perceptual and Contextual Losses," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, July 2017, p. 5485–5493.

[13] A. Borji. (2018, June) Pros and Cons of GAN Evaluation Measures. https://arxiv.org/pdf/1802.03446.pdf.

[14] (2020, February) Understanding Frequency Performance Specifications. https://www.ni.com/de-de/support/documentation/supplemental/06/understanding-frequency-performance-specifications.html#section--1434545352.