Digital twins of human corneal endothelium from generative adversarial networks

Eloi Dussy Lachaud, Andrew Caunes, Gilles Thuret, Yann Gavet

Digital twins of human corneal endothelium from generative adversarial networks.

Eloi Dussy Lachaud[a], Andrew Caunes[a], Gilles Thuret[b], and Yann Gavet[a]

[a]Mines Saint-Etienne, Univ. Lyon, CNRS, UMR 5307 LGF, Centre SPIN, F - 42023 Saint-Etienne, France.
[b]Laboratoire Biologie, ingénierie et imagerie de la Greffe de Cornée (BiiGC), Faculté de Médecine, Campus Santé Innovation, Université Jean Monnet, Saint-Etienne, France

ABSTRACT

The human corneal endothelium, the posterior most layer of the cornea, is a monolayer of flat cells that are essential for maintening its transparency over time. Endothelial cells are easily visualized in patients using a specular microscope, a routine device, but accurate cell counting and cell morphometry determination has remained challenging since decades. The first automatic segmentations used mathematical morphology techniques, or the principles of the Fourier transform. In recent years, convolutional neural networks have further improved the results, but they need a large learning database, which takes a long time to collect. Thus, this work proposes a method for simulating digital twins of the images observed in specular microscopy, in order to enrich medical databases.

Keywords: GAN, Corneal Endothelium, Medical imaging, Digital twins

1. INTRODUCTION

Machine learning is getting visibility and popularity in medical imaging at impressive speed. Computational power is still rising exponentially, and people are starting to realize how much better computers are when it comes to processing and identifying patterns in the growing medical image banks. As a matter of fact, the quality of learning is also directly related to the size of the image databases. However, while some fields will naturally and continuously receive data, such as screening operation in radiology, other smaller fields might be lacking in data input. That is where the solution that we are proposing might come into play, allowing the artificial creation of additional realistic data based on what was already collected, and completing classical image processing and analysis methods for medical diagnosis.[1–3]

Our objective is to simulate realistic images of human corneal endothelium visualized by specular microscopy, in order to feed other deep learning networks and improve the quality of image cells segmentation. This technique is a non-invasive way of observing cells and predicting health evolution of the cornea.

1.1 Corneal endothelium

The cornea is the front of the eye. It has two main roles: protection of the eye and refraction of light rays towards the fovea (the cornea is a lens with a power of about 40 diopters). The cornea is made up of several cellular layers, notably the endothelium (the inner part), which is generally responsible for keeping the cornea transparent through the action of its cells. Their number is therefore an essential criterion in verifying the good health of this tissue, and the patient's good vision in general.

1.2 Specular microscopy

The analysis of the endothelium is very important, and analysis techniques have enabled us to evaluate the relevant criteria in the case of observation using specular microscopy on patients. Specular microscopy is a technique that allows us to illuminate the whole cornea, and to take advantage of changes in the indices of the cell layers to observe the reflection corresponding to the endothelium.[4] This technique therefore involves a fairly significant drift in illumination, and is sensitive to surface irregularities in the areas observed.

Send correspondence to Y.G.: gavet@emse.fr.

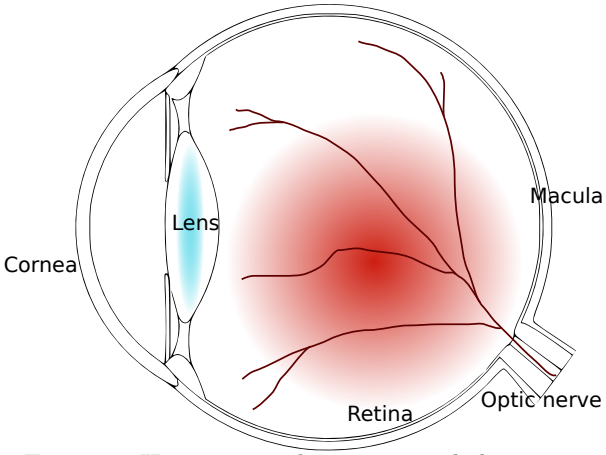


Figure 1: Human eye, the cornea and the retina.

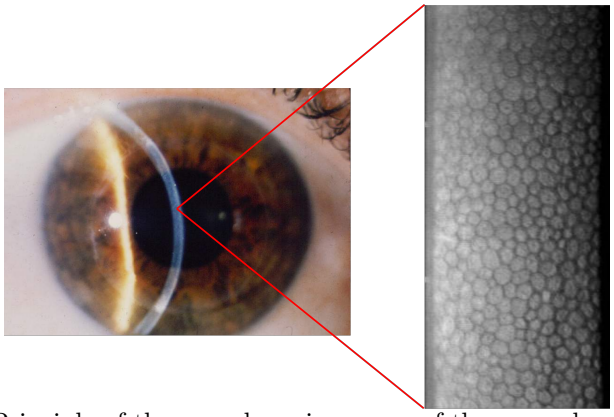


Figure 2: Principle of the specular microscopy of the corneal endothelium.

1.3 Objectives

Some recent studies[5–7] automatically segment the observed images. The results are very promising, but these techniques require access to an important learning base, which does not exist (for the moment) because the manual segmentation time required is very important. We are working in parallel on projects[8] which allow us to replicate the structure of cells and to obtain realistic images of these cells. This work presents the realistic side of the simulations.

2. GENERATION METHOD

2.1 Generative Adversarial Networks

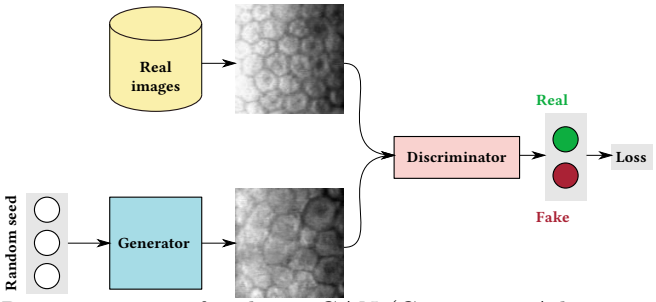


Figure 3: Representation of a classic GAN (Generative Adversarial Network)

GAN (Generative Adversarial Network[9]) are a type of deep learning model that uses unsupervised learning to randomly generate images based on a given image dataset. They are usually used to produce realistic images, hopefully indistinguishable from the real one while also being new and unique.

GANs use two different neural networks. The first one is the Generator, whose purpose is to generate images based on a latent random noise that can fool the discriminator into thinking it is an image from the real dataset. Conversely, the discriminator will try to tell apart the fake from the real image. Their training happens simultaneously, but they are competing against each other, hence the word "Adversarial" in GAN. This peculiarity can make their training troublesome, where one becomes vastly superior to the other or they both chase each other in circles.

2.2 Description of the networks

The Generator and the Discriminator are both convolutional neural networks, composed of a total of 6 layers.

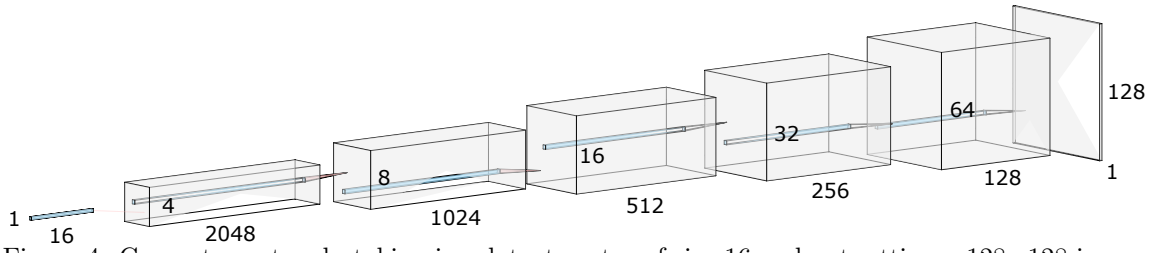


Figure 4: Generator network, taking in a latent vector of size 16 and outputting a 128×128 image.

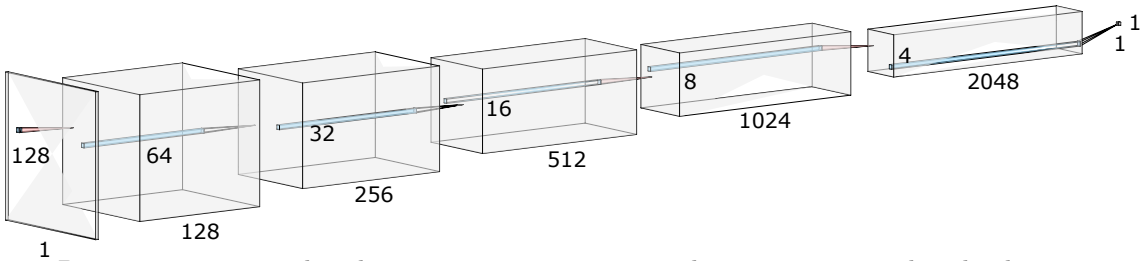


Figure 5: Discriminator network, taking in a 128×128 image and outputting a single value between 0 and 1.

The Generator (Fig.4) takes a latent vector of size 16 as an input and outputs a 128×128 image. All 6 layers are composed by a 2D transposed convolution with a kernel size of 5, followed by a batch normalization as well as a leaky ReLU for the first five layers. The last layer instead uses a hyperbolic tangent to generate the image. The number of channels of the tensor starts at 2048 and goes down to 128 (from which the image is generated) as its size increases. The Discriminator (Fig.5) takes a 128×128 image as its input. This image either comes from the Generator or from the real data set. It outputs a single value comprised between 0 and 1 that corresponds the score of the image, which goes higher the more the Discriminator think the image is real. Its construction is basically the same as the Generator but reversed. All 6 layers are composed of a 2D convolution, as well as a batch normalization and a leaky ReLU that is replaced by a sigmoid function to output the score on the last layer. The tensor starts with 128 channels and goes up to 2048 channels (from which the image score is computed) as its size decreases.

2.3 Construction of the database

We collected about 12000 raw images from the anonymous database of the ophthalmology ward of the University Hospital of Saint-Etienne. Images were taken in patients (aged 5 to 105 years) with either a healty corneas or a corneal disease affecting their endothelial cell density (cells/surface unit) as well as their cell morphometry. The images were randomly selected without any age or pathology criteria. This number was rapidly cut in half by removing all duplicate, black or irrelevant images. We treated and cropped the remaining one using a script

of our making. We then had to carefully go through those thousands of images to remove one by one the ones that were either too blurry, too dark or saturated so that they would not throw off training. This left us with a dataset of 3062 images to train our model (Fig.6). This amount proved to be insufficient. To increase artificially this number, we noticed that the number of cells is quite important in each image, and we don't require the full image to make sense of it. This means we are able to crop the image wherever we want, effectively multiplying our dataset's size (see Fig.7b).

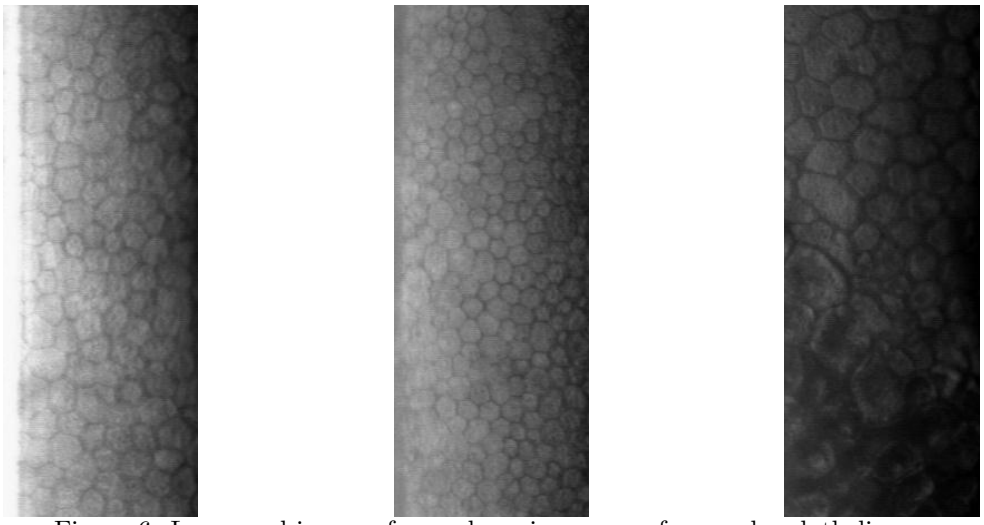
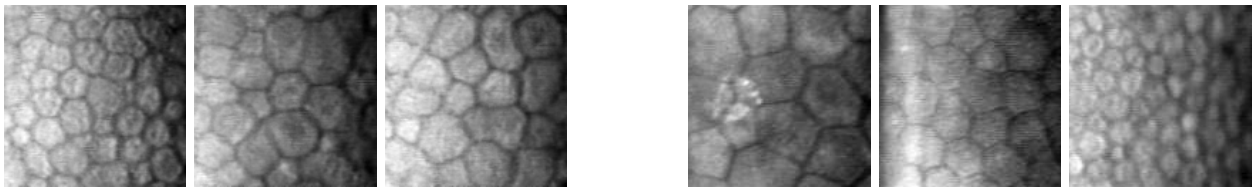


Figure 6: Large real image of specular microscopy of corneal endotheliums.

The model is therefore trained on images of 128×128 pixels, producing similar shaped one. The training appears to be optimal at around 600 to 800 hundreds epochs (iterations) and takes about a dozen hours using a Tesla P100 GPU.

3. RESULTS AND VALIDATION

3.1 Presentation



(a) Images generated from the model. (b) Images sampled from the training dataset.

Figure 7: Dataset of simulated and real images.

The obtained examples (Fig.7a) seem to be indistinguishable from the real images for a non specialist. The generated patterns present diverse features, with varying density and shape, as it is in reality.

The way one should go about evaluating the performance of a GAN is not obvious and is in fact an open field of research. The problem is that it is often hard to produce a metric accurately describing the distance between the real and the fake distributions, respectively followed by real images and artificial examples. It is common nowadays to use both a qualitative and a quantitative method of evaluation, which is what we chose to do here. In the following, we propose two approaches for evaluating the performance: the first one is the quantification of an expert evaluation, the second one is based on the Fréchet Inception Distance, commonly used with GANs.

Table 1: Classification results of the expert ophthalmologist.

	precision	recall	f1-score	support
Simulated image	0.11	0.05	0.07	1000
Real image	0.37	0.56	0.44	1000
Accuracy			0.30	2000

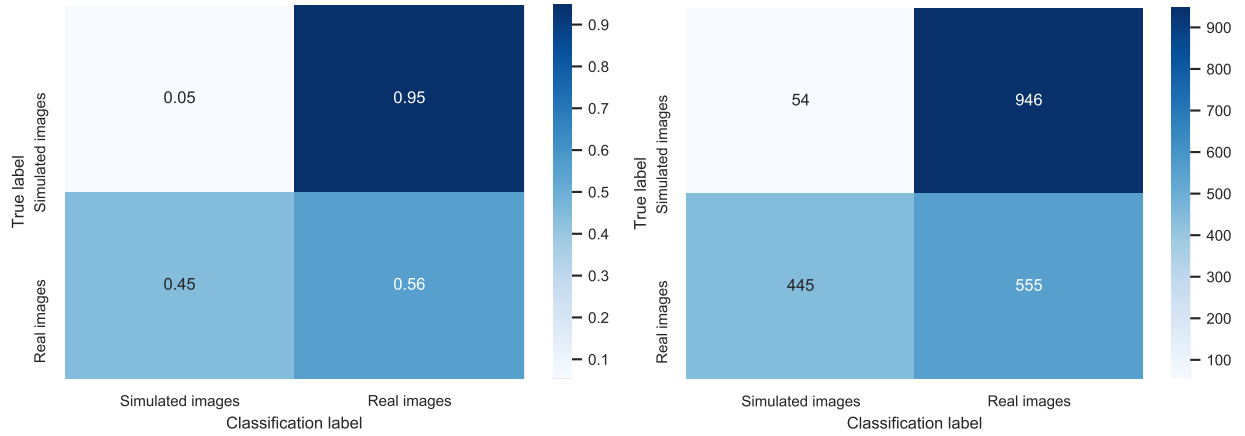
3.2 Validation and results

3.2.1 Qualitative evaluation

A dataset of 2000 images was constructed, with 1000 simulated images, and 1000 real images. The simulated images had the size of 128×128 pixels, the real images were cropped with the same size from a larger one (Fig. 6): the region was randomly selected. This precision is important because real images often present a high gradient of illumination, due to the principle of acquisition.

These images were then classified by an expert ophthalmologist. He has no indication about the number of images of each class.

It is clear that our expert cannot distinguish between real and simulated images. The accuracy (correctly classified images divided by total number of images) is of 0.30, which is really low. It appears that 1501 images have been classified as real. Results are summarized in Table 1 and Fig. 8, and these clearly show that there is no visible difference between real and simulated images.



(a) Normalized representation (on each line). (b) Absolute representation.

Figure 8: Confusion matrix of the classification by an expert ophthalmologist. This shows that even a specialist cannot distinguish simulated images from real images.

3.2.2 Quantitative approach: Frechet Inception Distance

The Fréchet Inception Distance[10] is widely used in GAN performance evaluation nowadays. It uses an Inception v3 classification network trained on the ImageNet database to get a feature representation of the images. By giving as input an image to such a network, we can get a classification as an output (based on ImageNet classes), but we can also stop the computations early and extract the results before the final layer, this way obtaining a vector of lower dimension representing the features of the image. The distributions followed by the representations of real and fake images are assumed to be multivariate normal. Therefore, we can compute their Fréchet Distance as follows :

$$\mathrm{FID}(X_{real}, X_{fake}) = ||\mu_{real} - \mu_{fake}||_2^2 + \mathrm{tr}\left(\sum real + \sum fake - 2\left(\sum real \sum fake\right)^{1/2}\right)$$

This scalar value represents the distance between the distributions of real and fake images, therefore it gives us an idea of how different fake images are from real images.

This quantitative method may seem curious at first since the feature representation uses a classification network trained on natural images, that is, dogs and cars for example, which differ from medical images quite a lot. However it is common to use it with all sorts of GANs and it generally grants satisfying results. This is because the layer of the Inception v3 network which we use to get the representations is somewhat early, and, at this stage in the network, only gives insights about simple shapes and structure in the image, and not yet about complex natural objects.

We did simple control tests to verify that the method wasn't giving irrelevant results when comparing noise to real images or real images to themselves. When comparing real images together, we found FID values of around 10, while when comparing noise to real images, we found FID values of more than 350. This gives us an idea of what kind of values to expect when actually measuring FID for our generated images.

We used a Pytorch adaptation[11] of the original FID code.

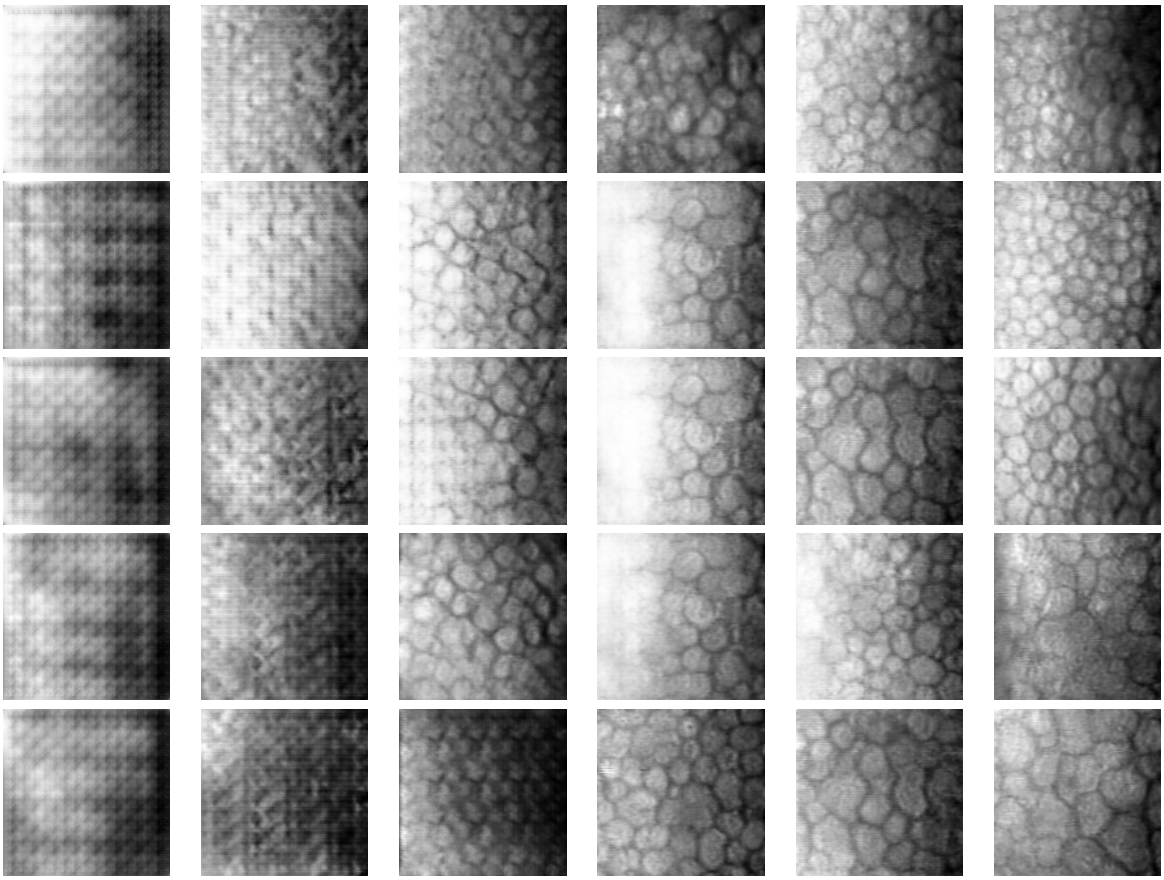


Figure 9: Generated images with their associated FID values.

Fig. 9 shows samples of generated images and their FID values. It is interesting to see that FID correlates well with human observation. With higher FID values, we can see a clear drop in quality of the generated images. FID also takes diversity into account (since the Fréchet Distance involves the standard deviations of the distributions) as can be seen when the FID gets below 100. When FID = 90, the images are realistic looking but we see reoccurring features across the samples, indicative of a mode collapse effect. However, when FID = 35, the diversity has improved while the quality remained high. Fig.10 illustrates the FID value according to the number of epochs. After 100 epochs, there is no more improvements.

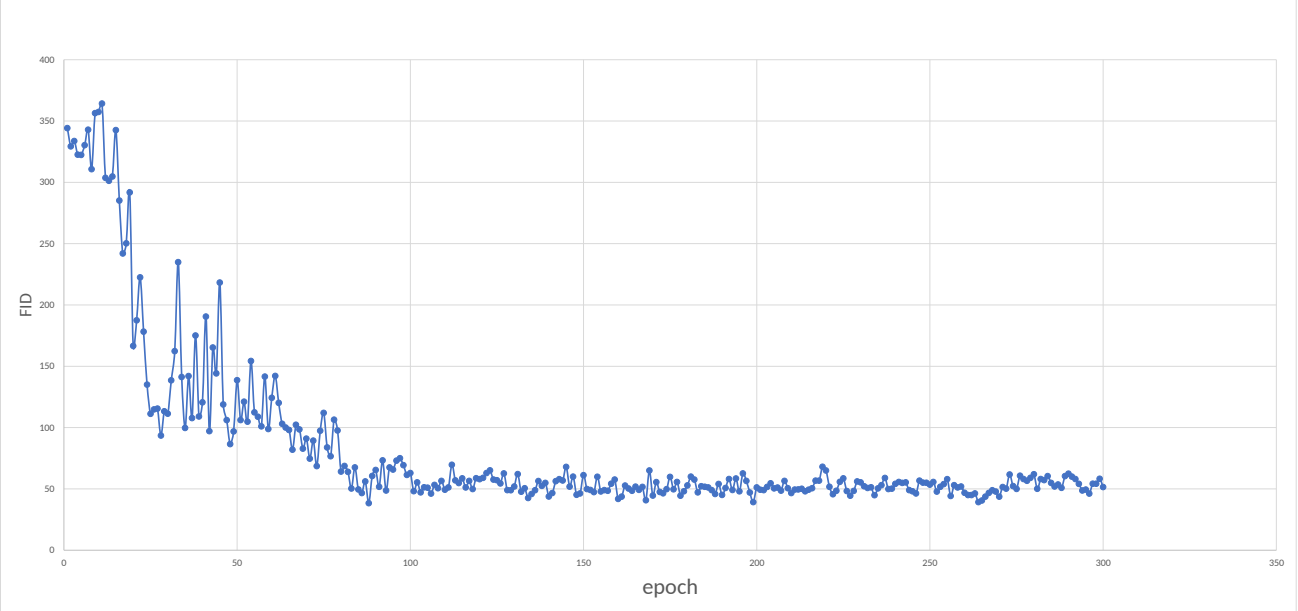


Figure 10: FID scores of a model from epoch 1 to 300 of training. As expected, the score is decreasing at a slower and slower pace with time, with oscillations.

On our best models, we could achieve FID scores of roughly 35 for 3000 images in each set. These values seem relevant, considering those obtained by other papers using GANs in a similar manner.[12] More importantly, we were able to observe a correlation between the FID scores and the quality and diversity of the images as evaluated by human observation.

4. CONCLUSION AND PERSPECTIVES

We proposed a way of generating images with the same look as real corneal endothelium images observed by specular microscopy. GANs are an actual way of performing this task, but this method still lacks control on biological of physiological properties that could be reproduced. To be more precise, our objectif is to be able to simulate images representing a specific pathology or a patient of a given age.

Another aspect that could be improved upon is the generation of labeled images. A possibility would be the numerical generation of a pattern similar to the ones of corneal endothelium based on given parameters, which would be transformed in a realistic looking image usable for training using Cycle-GAN. This would allow us to know the parameters and features (e.g. cells borders) of the generated images, allowing supervised learning.

Such generated images could greatly help future machine learning algorithm applied to corneal endothelium, as well as other similar imaging field.

REFERENCES

[1] Gain, P., Thuret, G., Kodjikian, L., Gavet, Y., Turc, P. H., Theillere, C., Acquart, S., Le Petit, J. C., Maugery, J., and Campos, L., "Automated tri-image analysis of stored corneal endothelium," *British Journal of Ophthalmology* **86**(7), 801–808 (2002).

[2] Gavet, Y. and Pinoli, J.-C., "Visual perception based automatic recognition of cell mosaics in human corneal endotheliummicroscopy images," *Image Analysis & Stereology* **27**(1), 53–61 (2008).

[3] Ruggeri, A., Grisan, E., and Schroeter, J., "Evaluation of repeatability for the automatic estimation of endothelial cell density in donor corneas," *Br J Ophthalmol* **0** (2007).

[4] Bourne, W. M. and Kaufman, H. E., "Specular microscopy of human corneal endothelium in vivo.," *American journal of ophthalmology* **81**(3), 319–323 (1976).

[5] Vigueras-Guillén, J. P., Sari, B., Goes, S. F., Lemij, H. G., van Rooij, J., Vermeer, K. A., and van Vliet, L. J., "Fully convolutional architecture vs sliding-window cnn for corneal endothelium cell segmentation," *BMC Biomedical Engineering* **1**(1), 1–16 (2019).

[6] Nurzynska, K., "Deep learning as a tool for automatic segmentation of corneal endothelium images," *Symmetry* **10**(3), 60 (2018).

[7] Fabijańska, A., "Segmentation of corneal endothelium images using a u-net-based convolutional neural network," *Artificial intelligence in medicine* **88**, 1–13 (2018).

[8] Penaud-Polge, V. and Gavet, Y., "Robustness study of edge segmentation andcompletion by cnn based method for tessellationimages," in [*10th International Symposium on Signal, Image, Video and Communications*], (2021). to be published.

[9] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial networks," (2014).

[10] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," (2018).

[11] Seitzer, M., "pytorch-fid: FID Score for PyTorch." https://github.com/mseitzer/pytorch-fid (August 2020). Version 0.1.1.

[12] Middel, L., Palm, C., and Erdt, M., "Synthesis of medical images using gans," in [*Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*], 125–134, Springer International Publishing, Cham (2019).