CHAPTER 15

Multimodal reconstruction of retinal images over unpaired datasets using cyclical generative adversarial networks

Álvaro S. Hervella[a,b], José Rouco[a,b], Jorge Novo[a,b], and Marcos Ortega[a,b]
[a]CITIC Research Center, University of A Coruña, A Coruña, Spain
[b]VARPA Research Group, Biomedical Research Institute of A Coruña (INIBIC), University of A Coruña, A Coruña, Spain

15.1 Introduction

The recent rise of deep learning has revolutionized medical imaging, making a significant impact on modern medicine [1]. Nowadays, in clinical practice, medical imaging technologies are key tools for the prevention, diagnosis, and follow-up of numerous diseases [2]. There exist a large variety of imaging modalities that allow to visualize the different organs and tissues in the human body [3]. Thus, clinicians can select the most adequate imaging modality to study the different anatomical or pathological structures in detail. Nevertheless, the detailed analysis of the images can be a tedious and difficult task for a clinical specialist. For instance, many diseases in their early stages are only evidenced by very small lesions or subtle anomalies. In these scenarios, factors such as the clinicians' expertise and workload can affect the reliability of the final analysis. Thus, the use of deep learning algorithms allows to accelerate the process and helps to produce a more reliable analysis of the images. Ultimately, this will result in a better diagnosis and treatment for the patients.

Deep neural networks (DNNs) have been demonstrated to provide a superior performance for numerous image analysis problems in comparison to more classical methods [4]. For instance, nowadays, deep learning represents the state-of-the-art approach for typical tasks, such as image segmentation [5] or image classification [6]. Besides the remarkable improvements in these canonical image analysis problems, deep learning also makes possible the emergence of novel applications. For instance, these algorithms can be used for the transformation of images among different modalities [7], or the training of future clinical professionals using realistic generated images [8]. These novel applications, among others, certainly benefit from the particular advantages of generative adversarial networks (GANs) [9]. This creative setting, consisting of different networks with opposite objectives, have been demonstrated to be able to further exploit the capacity of the DNNs.

Generative Adversarial Networks for Image-to-Image Translation
https://doi.org/10.1016/B978-0-12-823519-5.00014-2

Multimodal reconstruction is a novel application driven by DNNs that consists in the translation of medical images among complementary modalities [7]. Nowadays, complementary imaging modalities, representing the same organs or tissues, are commonly available in most medical specialties [3]. The differences among modalities can be due to the use of different capture devices, and also due to the use of contrasts that enhance certain tissues. The clinicians choose the most adequate imaging modality according to different factors, such as the target organs or tissues, the evidence of disease, or the risk factors of the patient. In this sense, it is particularly important to consider the properties of the different anatomical and pathological structures, given that some structures can be enhanced in one modality and be completely missing in other. This significant change in the appearance, dependent on the properties of the tissues and organs, can make the translation among modalities very challenging. However, this challenge that complicates the training of the multimodal reconstruction is beneficial if we are interested in using the task for representation learning purposes. This is due to the fact that a harder task will enforce the network to learn more complex representations during the training. In this regard, the multimodal reconstruction has already demonstrated a successful performance as pretraining task for transfer learning in medical imaging [10].

In this chapter, we study the use of GANs for the multimodal reconstruction between complementary imaging modalities. In particular, the multimodal reconstruction is addressed by using a cyclical GAN methodology, which allows training the adversarial setting with independent sets of two different image modalities [11]. Nowadays, GANs represent the quintessential approach for image-to-image translation tasks [12]. However, these kinds of applications are typically focused on producing realistic and aesthetically pleasing images. In contrast, in the multimodal reconstruction of medical images, the realism and aesthetics of the generated images are not as important as producing medically accurate reconstructions. In particular, this means that the generated color patterns and textures must be coherent with the expected visualization of the real organs or tissues in the target modality. Additionally, this may involve the omission of certain structures, or even the enhancement of those that are only vaguely appreciated in the original modality. We evaluate all these aspects in order to assess the validity of the studied cyclical GAN method for the multimodal reconstruction.

The study presented in this chapter is focused on ophthalmic imaging. In particular, we use the retinography and the fluorescein angiography as the original and target imaging modalities in the multimodal reconstruction. These imaging modalities, which represent the eye fundus, are useful for the study of important ocular and systemic diseases, such as glaucoma or diabetes [2]. A representative example of retinography and fluorescein angiography for the same eye is depicted in Fig. 15.1. The main difference between them is that the fluorescein angiography uses a contrast dye, which is injected to the patient, to produce the fluorescence of the blood. Thus, the fluorescein angiography depicts an enhanced representation of the retinal vasculature and related lesions.

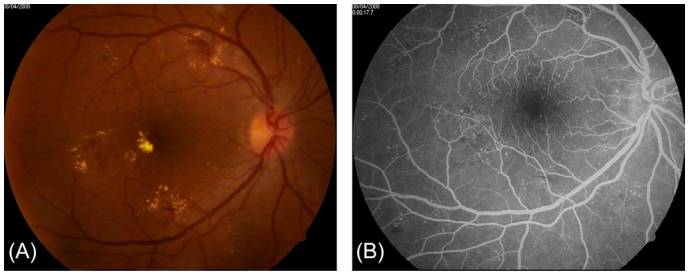

Fig. 15.1 Example of retinography and fluorescein angiography for the same eye: (A) retinography and (B) angiography.

In this context, the successful training of a deep neural network in the multimodal reconstruction of the angiography from the retinography will provide a model able to produce a contrast-free estimation of the enhanced retinal vasculature. Additionally, due to the challenges of the transformation, which is mainly mediated by the presence of blood flow in the different tissues, the neural networks will need to learn rich high level representations of the data. This represents a remarkable potential for transfer learning purposes [13, 14].

The presented study includes an extensive evaluation of the cyclical GAN methodology for the multimodal reconstruction between complementary imaging modalities. For this purpose, two different multimodal datasets containing both retinography and fluorescein angiography images are used. Additionally, in order to further analyze the advantages and limitations of the methodology, we present an extensive comparison with a state-of-the-art approach for the multimodal reconstruction of these ophthalmic images [15]. In contrast with the cyclical GAN methodology, this other approach requires the use of multimodal paired data for training, i.e., retinography and angiography of the same eye. Therefore, the cyclical GAN presents an important advantage, avoiding not only the necessity of paired data but also the unnecessary preprocessing for the alignment of the different image pairs.

15.2 Related research

Generative adversarial networks (GANs) represent a relatively new deep learning framework for the estimation of generative models [16]. The original GAN setting consists of two different networks with opposite objectives. In particular, a discriminator that learns to distinguish between real and fake samples and a generator that learns to produce fake samples that the discriminator misclassifies as real. Based on this original idea, several variations were developed in posterior works, aiming at applying the novel paradigm in different scenarios [17].

In recent years, GANs have been extensively used for addressing different vision problems and graphics tasks. The use of GANs has been especially groundbreaking for computer graphics applications due to the visually appealing results that are obtained. Similarly, a kind of vision problem that has been revolutionized by the use of GANs is image-to-image translation, which consists of performing a mapping between different image domains or imaging modalities [12]. An early work addressing this problem with GANs, known as Pix2Pix [18], relied on the availability of paired data for learning the generative model. In particular, Isola et al. [18] show that their best results are achieved by combining a traditional pixel-wise loss and a conditional GAN framework. Given the difficulty of gathering the paired data in many application domains, posterior works have proposed alternatives to learn the task by using unpaired training data. Among the different proposals, the work of Zhu et al. [19], known as CycleGAN, has been especially influential. CycleGAN compensates for the lack of paired data by learning not only the desired mapping function but also the inverse mapping. This allows introducing a cycle-consistency loss whereby the subsequent application of both mapping functions must return the original input image. Concurrently, this same idea with different naming was also proposed in DualGAN [20] and DiscoGAN [21]. Additionally, besides the cycle-consistency alternative, other different proposals have been presented in different works [12] although the use of these other alternatives is not as extended in posterior applications.

In medical imaging, GANs have also been used for different applications, including the mapping between complementary imaging modalities. In particular, GANs have been successfully applied in tasks such as image denoising [22], multimodal reconstruction [11], segmentation [23], image synthesis [24], or anomaly detection [25]. Among these different tasks, several of them can be directly addressed as an image-to-image translation [8]. In these cases, the adaption of those state-of-the-art approaches that already demonstrated a good performance in natural images has been common. In particular, numerous works in medical imaging are based on the use of Pix2Pix or CycleGAN methodologies [8]. Similarly to other application domains, the choice between one or other approach is conditioned by the availability of paired data for training. However, in medical imaging, the paired data is typically easy to obtain, which is evidenced by the prevalence of paired approaches in the literature [8]. With regard to the multimodal reconstruction, the difficulty in these cases is to perform an accurate registration of the available image pairs.

An important concern regarding the use of GANs in medical imaging is the hallucination of nonexistent structures by the networks [8]. This is a concomitant risk with the use of GANs due to the high capacity of these frameworks to model the given training data. Cohen et al. [26] demonstrated that this risk is especially elevated when the training data is heavily unbalanced. For instance, a GAN framework that is trained for multimodal reconstruction with a large majority of pathological images will tend to hallucinate pathological structures when processing healthy images. This behavior can be in part mitigated by the addition of pixel-wise losses if paired data is available. Nevertheless, regarding the multimodal reconstruction, even when the paired data is available, most

of the works still use the GAN framework together with the pixel-wise loss [8]. In this regard, the work of Hervella et al. [15] is an example of multimodal reconstruction without GANs and using instead the Structural Similarity (SSIM) for the loss function. The motivation for this is, for many applications in medical imaging, it is not necessary to generate realistic or aesthetically pleasing images. In this context, the results obtained in Ref. [15] show that, without the use of GANs, the generated images lack realism and can be easily identified as synthetic samples.

15.3 Multimodal reconstruction of retinal images

Multimodal reconstruction is an image translation task between complementary medical imaging modalities [7]. The objective of this task is, given a certain medical image, to reconstruct the underlying tissues and organs according to the characteristics of a different complementary imaging modality. Particularly, this chapter is focused on the multimodal reconstruction of the fluorescein angiography from the retinography. These two complementary retinal imaging modalities represent the eye fundus, including the main anatomical structures and possible lesions in the eye. The main difference between retinography and angiography is that the latter requires the injection of a contrast dye before capturing the images. The injection of this contrast dye results in an enhancement of the retinal vasculature as well as those pathological structures with blood flow. Simultaneously, those other retinal structures and tissues where there is a lack of blood flow may be attenuated in the resulting images. Thus, there is an intricate relation between retinography and angiography, given that the visual transformation between the modalities depends on physical properties such as the presence of blood flow in the different tissues. As a reference, the transformation between retinography and angiography for the main anatomical and pathological structures in the retina can be visualized in Fig. 15.2.

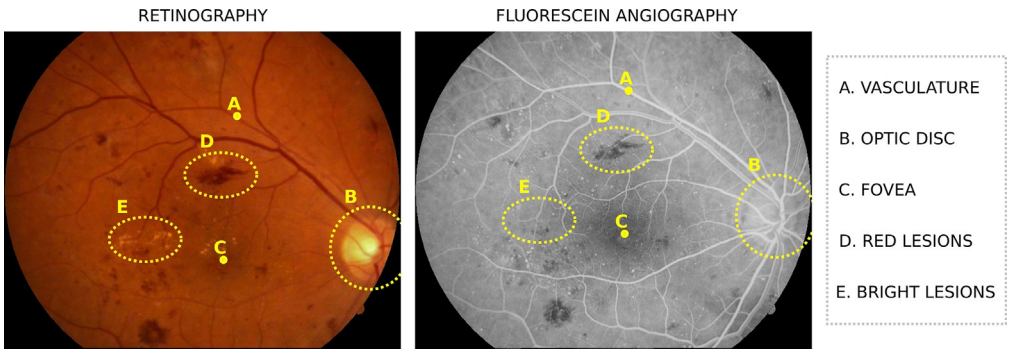

Fig. 15.2 Example of retinography and fluorescein angiography for the same eye. The included images depict the main anatomical structures as well as the two main types of lesions in the retina.

Recently, the difficulty of performing the multimodal reconstruction between retinography and angiography has been overcome by using DNNs [7]. In this regard, the required multimodal transformation can be modeled as a mapping function that $G_{R2A}: R \rightarrow A$ given a certain retinography $r \in R$ returns the corresponding angiography $a = G_{R2A}(r) \in A$ for the same eye. In this scenario, the mapping function G_{R2A} can be parameterized by a DNN. Thus, the function parameters can be learned by applying an adequate training strategy. In this regard, we present two different deep learning-based approaches for learning the mapping function G_{R2A}, the cyclical GAN methodology [11] and the paired SSIM methodology [15].

15.3.1 Cyclical GAN methodology

The cyclical GAN methodology is based on the use of generative adversarial networks (GANs) for learning the mapping function from retinography to angiography [11]. In this regard, GANs have demonstrated to be useful tools for learning the data distribution of a certain training set, allowing the generation of new images that resemble those contained in the training data [16]. This means that, by using GANs and a sufficiently large training set of unlabeled angiographies, it is possible to generate new fake angiographies that are theoretically indistinguishable from the real ones. However, in the presented multimodal reconstruction, the generated images do not only need to resemble real angiographies but, also, need to represent the physical attributes given by a particular retinography. Thus, in contrast with the original GAN approach [16], the presented methodology does not generate new images from a random noise vector, but rather from another image with the same spatial dimensions as the one that is being generated. In practice, this image-to-image transformation is achieved by using an encoder-decoder network as the generator, whereas the discriminator is still a decoder network as in the original GAN approach. Applying this setting, the multimodal reconstructions could theoretically be trained by using two independent unlabeled sets of images, one of the retinographies, and the other of angiographies.

An inherent difficulty of training an image-to-image GAN is that, typically, the generator network has enough capacity to generate a variety of plausible images while ignoring the characteristics of the network input. In the case of the multimodal reconstruction, this would mean that the physical attributes of the retinographies are not successfully transferred to the generated angiographies. In this regard, early image-to-image GAN approaches addressed the issue by explicitly conditioning the generated images on the network input [18]. In particular, this is achieved by using a paired dataset instead of two independent datasets for training. For instance, the use of retinography-angiography pairs, instead of independent retinography and angiography samples, allows training a discriminator to distinguish between fake and real angiographies conditioned on a given real retinography. The use of such a discriminator will force the generator to analyze and take

into account the attributes of the input retinography. Additionally, in Ref. [18], the use of paired datasets is even further exploited by complementing the adversarial feedback to the generator with a pixel-wise similarity metric between the generator output and the available ground truth. However, in this case, it is not only necessary to have paired data, but also the available image pairs must be aligned.

In contrast with previous alternatives, the presented cyclical GAN methodology addresses the issue of the generator potentially ignoring the characteristics of its input in a different manner that does not require the use of paired datasets. In particular, the cyclical GAN solution is based on the use of a double transformation [19]. The idea is to simultaneously learn G_{R2A} and its inverse mapping function $G_{A2R}: A \rightarrow R$ that given a certain angiography $a \in A$ produces a retinography $r = G_{A2R}(a) \in R$ of the same eye. Then, the subsequent application of both transformations should be equivalent to the identity function. For instance, if a retinography is transformed into angiography and, then, it is transformed back into retinography, the resulting image should be identical to the original retinography that is used as input. However, if any of the two transformations ignores the characteristics of their input, the resulting retinography will differ from the original. Therefore, it is possible to ensure that the input image characteristics are not being ignored by enforcing the identity between the original retinography and the one that is transformed back from angiography. This is referred to as cycle-consistency, and it can be applied by using any similarity metric between both original and reconstructed input image. An important advantage of this solution is that it does not require the use of paired datasets, only being necessary two independent sets of unlabeled retinographies and angiographies.

In order to obtain the best performance for the multimodal reconstruction, the presented cyclical GAN methodology involves the use of two complementary training cycles: (1) from retinography to angiography to retinography ($R2A2R$) and (2) from angiography to retinography to angiography ($A2R2A$). A flowchart showing the complete training procedure is depicted in Fig. 15.3. It is observed that two different generators, G_{R2A} and G_{A2R}, and two different discriminators, D_A and D_R, are used during the training. The discriminators D_A and D_R are trained to distinguish between generated and real images. Simultaneously, the generators G_{R2A} and G_{A2R} are trained to generate images that the discriminators misclassify as real. This adversarial training is performed using a least square loss, which has demonstrated to produce a more stable learning process in comparison to the original loss in regular GANs [27]. Regarding the discriminator training, the target values are 1 for the real images and 0 for the generated images. Thus, the adversarial training losses for the discriminators are defined as

$$L_{D_A}^{adv} = E_{r\sim R}\left[D_A(G_{R2A}(r))^2\right] + E_{a\sim A}\left[(D_A(a)-1)^2\right] \tag{15.1}$$

$$L_{D_R}^{adv} = E_{a\sim A}\left[D_R(G_{A2R}(a))^2\right] + E_{r\sim R}\left[(D_R(r)-1)^2\right] \tag{15.2}$$

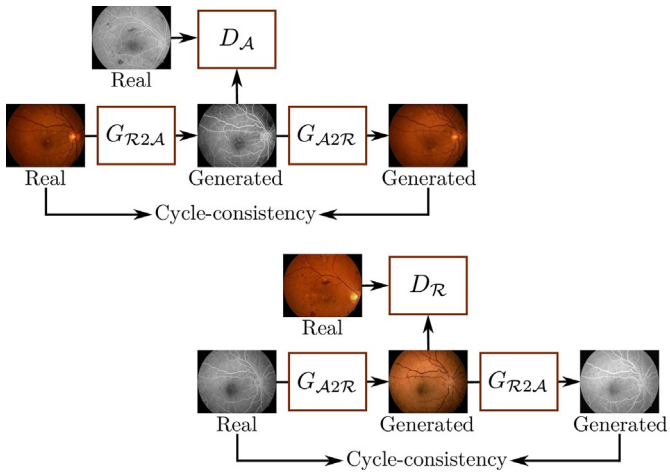

Fig. 15.3 Flowchart for the complete training procedure in the cyclical GAN methodology. This approach involves the use of two complementary training cycles that only differ in which imaging modality is being used as input and which one is the target. For each training cycle, the appearance of the target modality in the generated images is enforced by the feedback of the discriminator. Simultaneously, the cycle-consistency is used to ensure that the input image characteristics, such as the anatomical and pathological structures, are not being ignored by the networks.

In the case of the generator training, the objective is that the discriminator assigns a value 1 to the generated images. Thus, the adversarial training losses for the generators are defined as

$$L_{G_{R2A}}^{adv} = E_{r\sim R}\left[\left(D_A\left(G_{R2A}(r)\right)-1\right)^2\right] \tag{15.3}$$

$$L_{G_{A2R}}^{adv} = E_{a\sim A}\left[\left(D_R\left(G_{A2R}(a)\right)-1\right)^2\right] \tag{15.4}$$

Regarding the cycle consistency in the presented approach, the L1-norm between the original image and its reconstructed version is used as a loss function. In particular, the complete cycle-consistency loss, including both training cycles, is defined as

$$L^{cyc} = E_{r\sim R}\left[\left\|G_{A2R}\left(G_{R2A}(r)\right)-r\right\|_1\right] + E_{a\sim A}\left[\left\|G_{R2A}\left(G_{A2R}(a)\right)-a\right\|_1\right] \tag{15.5}$$

As it can be observed in previous equations as well as in Fig. 15.3, there is a strong parallelism between both training cycles, *R2A2R* and *A2R2A*. In particular, the only difference is the imaging modality that each training cycle starts with, what sets which imaging modality is being used as input and which one is the target.

Finally, the complete loss function that is used for simultaneously training all the networks is defined as

$$L = L_{G_{R2A}}^{adv} + L_{D_A}^{adv} + L_{G_{A2R}}^{adv} + L_{D_R}^{adv} + \lambda L^{cyc} \tag{15.6}$$

where λ is a parameter that controls the relative importance of the cycle-consistency loss and the adversarial losses. For the experiments presented in this chapter, this parameter is set to a value of $\lambda = 10$, which was also previously adopted in Ref. [19].

The optimization of the loss function during the training is performed with the Adam algorithm [28]. Regarding the hyperparameters of Adam, the weight decays are $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In comparison to the original values recommended by Kingma et al. [28], this set of values has demonstrated to provide a more stable learning process when training GANs [29]. The optimization is performed with a batch size of 1 image. The learning rate is set to an initial value of $\alpha = 2e-4$ and it is kept constant for 200,000 iterations. Then, following the approach previously adopted in Ref. [19], the learning rate is linearly reduced to zero for the same number of iterations. The number of iterations before starting to reduce the learning rate is established empirically through the analysis of both the learning curves and the generated images in a training subset that is reserved for validation.

Finally, a data augmentation strategy is applied to avoid possible overfitting to the training set. In particular, random spatial and color augmentations are applied to the images. The spatial augmentations consist of affine transformations and the color augmentations are linear transformations of the image channels in HSV (Hue-Saturation-Value) color space. In the case of the angiographies, which have one single channel, a linear transformation is directly applied over the raw intensity values. This augmentation strategy has been previously applied for the analysis of retinal images, demonstrating a good performance avoiding overfitting with limited training data [10, 30]. The particular range for the transformations was validated before training in order to ensure that the augmented images still resemble valid retinas.

15.3.2 Paired SSIM methodology

An alternative methodology for the multimodal reconstruction between retinography and angiography was proposed in Ref. [7]. In this case, the authors avoid the use of GANs by taking advantage of existing multimodal paired data. In particular, a set of retinography-angiography pairs where both images correspond to the same eye. The motivation for this lies in the fact that, in contrast to other application domains, in medical imaging the paired data is easy to obtain. Nowadays, in modern clinical practice, the use of different imaging modalities is broadly extended across most medical services. In this sense, although for many patients the use of a single imaging modality can be enough for diagnostic purposes, there is still a large number of cases where the use of several imaging modalities is required. In this latter scenario, it is also common to use more complex or invasive techniques, such as those requiring the injection of contrasts. This is the case of retinography and angiography in retinal imaging. While retinography is a broadly extended modality, typically used in screening programs, angiography is only used when

it is clearly required. However, each time the angiography is taken for a patient, retinography is typically also available. This facilitates the gathering of these paired multimodal datasets.

Technically, the advantage of using paired training data is that it allows directly comparing the network output with a ground truth image. In particular, during the training, for each retinography that is fed to the network, there is also available an angiography of the same eye. Thus, the training feedback can be obtained by computing any similarity metric between generated and real angiography. In order to facilitate this measurement of similarity, the retinography and angiography within each multimodal pair are registered. The registration produces an alignment of the different retinal structures between the retinography and the angiography. Consequently, there will also be an alignment between the network output and the real angiography that is used as ground truth. This allows the use of common pixel-wise metrics for the measurement of the similarity between the network output and the target image.

In the presented methodology [15], the registration is performed following a domain-specific method that relies on the vascular structures of the retina [31]. In particular, this registration method presents two different steps. The first step is a landmark-based registration where the landmarks are the crossings and the bifurcations of the retinal vasculature. This first registration produces a coarse alignment of the images that is later refined by performing a subsequent intensity-based registration. This second registration is based on the optimization of a similarity metric of the vessels between both images. The complete registration procedure allows generating a paired and registered multimodal dataset, which is used for directly training the generator network G_{R2A}. The complete methodology for training the multimodal reconstruction is depicted in Fig. 15.4. As it is observed, an advantage of this methodology is that only a single neural network is required.

Regarding the training of the generator, the similarity between the network output and target angiography is evaluated by using the structural similarity (SSIM) [32]. This metric, which was initially proposed for image quality assessment, measures the similarity between images by independently considering the intensity, contrast, and structural

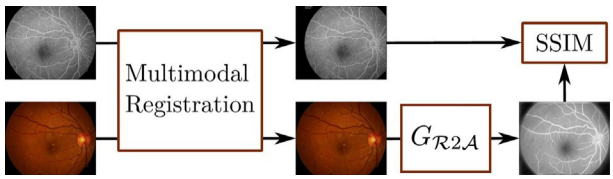

Fig. 15.4 Flowchart for the complete training procedure of the paired SSIM methodology. The first step is the multimodal registration of the paired retinal images, which can be performed off-line before the actual network training. Then, the training feedback is provided by the structural similarity (SSIM), which is a pixel-wise similarity metric.

information. The measurement is performed at a local level considering a small neighborhood for each pixel. In particular, an SSIM map between two images (x, y) is computed with a set of local statistics as

$$\mathrm{SSIM}(x, y) = \sum \frac{\left(2\mu_x\mu_y + C_1\right) + \left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{15.7}$$

where μ_x and μ_y are the local averages for x and y, respectively, σ_x and σ_y are the local standard deviations for x and y, respectively, and σ_{xy} is the local covariance between x and y. These local statistics are computed for each pixel by weighting its neighborhood with an isotropic two-dimensional Gaussian with $\sigma = 1.5$ pixels [32].

Then, given that SSIM is a similarity metric, the loss function for training G_{R2A} is defined by using the negative SSIM:

$$L^{\mathrm{SSIM}} = E_{r,a\sim(R,A)}[-\mathrm{SSIM}(G_{R2A}(r), a)] \tag{15.8}$$

The optimization of the loss function during the training is performed with the Adam algorithm [28]. Regarding the hyperparameters of Adam, the weight decays are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are the default values recommended by Kingma et al. [28]. The optimization is performed with a batch size of 1 image. The learning rate is set to an initial value of $\alpha = 2e-4$ and then it is reduced by a factor of 10 when the validation loss ceases to improve for 1250 iterations. Finally, the training is early stopped after 5000 iterations without improvement in the validation loss. These hyperparameters are established empirically according to the evolution of the learning curves during the training.

Finally, a data augmentation strategy is also applied to avoid possible overfitting to the training set. In particular, random spatial and color augmentations are applied to the images. The spatial augmentations consist of affine transformations and the color augmentations are linear transformations of the image channels in HSV (Hue-Saturation-Value) color space. In this case, the color augmentations are only applied to the retinography, which is the only imaging modality being used as input to a neural network. In contrast, the same affine transformation is applied to the retinography and the angiography in each multimodal image pair. This is necessary to keep the alignment between the images and make possible the measurement of the pixel-wise similarity, namely SSIM, between the network output and the target angiography. As in the cyclical GAN methodology, the particular range for the transformations is validated before training in order to ensure that the augmented images still resemble valid retinas.

15.3.3 Network architectures

Regarding the neural networks, the same network architectures are used for the two presented methodologies, cyclical GAN and paired SSIM. This eases the comparison

between the methodologies, excluding the network architecture as a factor in the possible performance differences. In particular, the experiments that are presented in this chapter are performed with the same network architectures that were previously used in Ref. [19]. The generator, which is used in both cyclical GAN and paired SSIM, is a fully convolutional neural network consisting of an encoder, a decoder, and several residual blocks in the middle of them. A diagram of the network and the details of the different blocks are depicted in Fig. 15.5 and Table 15.1, respectively. In contrast with other common encoder-decoder architectures, this network presents a small encoder and decoder, which is compensated by the large number of layers that are present in the middle residual blocks. As a consequence, there is also a small spatial reduction of the input data through the network. In particular, the height and width of the internal representations within the network are reduced up to a factor of 4. This relatively low spatial reduction allows keeping an adequate level of spatial accuracy without the

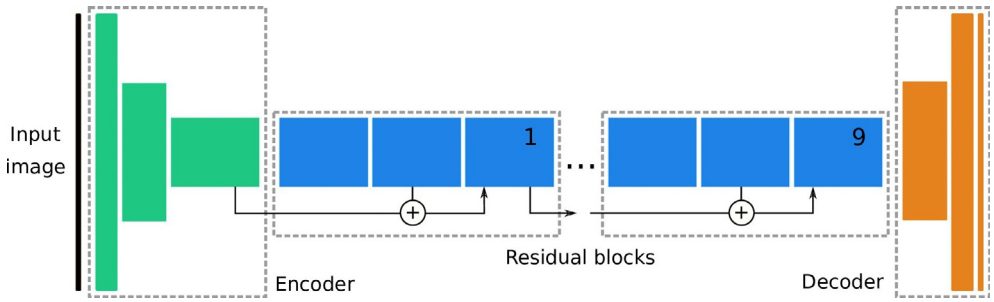

Fig. 15.5 Diagram of the network architecture for the generator. Each colored block represents the output of a layer in the neural network. The width of the blocks represents the number of channels whereas the height represents the spatial dimensions. The details of the different layers are in Table 15.1.

Table 15.1 Building blocks of the generator architecture.

Block	Layers	Kernel	Stride	Out features
Encoder	Conv/IN/ReLU	7×7	1	64
	Conv/IN/ReLU	3×3	2	128
	Conv/IN/ReLU	3×3	2	256
Residual	Conv/IN/ReLU	3×3	1	256
	Conv/IN	3×3	1	256
	Residual addition	–	–	256
Decoder	ConvT/IN/ReLU	3×3	2	128
	ConvT/IN/ReLU	3×3	2	64
	Conv/IN/ReLU	7×7	1	Image channels

Conv, convolution; IN, instance normalization [33]; ConvT, convolution transpose.

necessity of additional features such as skip connections [34]. Another particularity of the network is the use of instance normalization [33] layers after each convolution, in contrast to the more extended use of batch normalization. In this regard, instance normalization was initially proposed for improving the performance of style-transfer applications and has demonstrated to be also effective for cyclical GANs. Additionally, these normalization layers could be seen as an effective way of dealing with the problems of using batch normalization with small batch sizes. In this sense, it should be noticed that both the experiments presented in this chapter as well as the experiments in Ref. [19] are performed with a batch size of 1 image.

In contrast with the generator, the discriminator network is only used in the cyclical GAN methodology. The selected architecture is the one that was also used in Ref. [19]. In particular, the discriminator is a fully convolutional neural network, which allows working on arbitrarily sized images. This kind of discriminator architecture is typically known as PatchGAN [18], given that the decision of the discriminator is produced at the level of overlapping image patches. A diagram of the network and the details of the different layers are depicted in Fig. 15.6 and Table 15.2, respectively. The characteristics of the different layers are similar to those in the generator network. The main difference is the use of Leaky ReLU instead of ReLU as an activation function, which has

Fig. 15.6 Diagram of the network architecture for the discriminator. Each colored block represents the output of a layer in the network. The width of the blocks represents the number of channels whereas the height represents the spatial dimensions. The details of the different layers are in Table 15.2.

Table 15.2 Layers of the discriminator architecture.

Layers	Kernel	Stride	Out features
Conv/Leaky ReLU	4×4	2	64
Conv/IN/Leaky ReLU	4×4	2	128
Conv/IN/Leaky ReLU	4×4	2	256
Conv/IN/Leaky ReLU	4×4	1	512
Conv	4×4	1	1

Conv, convolution; IN, instance normalization.

demonstrated to be a useful modification for the adequate training of GANs [29]. With regard to the discriminator output, this architecture provides a decision for overlapping image patches of size 70×70.

15.4 Experiments and results

15.4.1 Datasets

The experiments presented in this chapter are performed on a multimodal dataset consisting of 118 retinography-angiography pairs. This multimodal dataset is created from two different collections of images. In particular, half of the images are taken from a public multimodal dataset provided by Isfahan MISP [35] whereas the other half have been gathered from a local hospital [15].

The Isfahan MISP collection consists of 59 retinography-angiography pairs including both pathological and healthy cases. In particular, 30 image pairs correspond to patients that were diagnosed with diabetic retinopathy whereas the other 29 images pairs correspond to healthy retinas. All the images in the collection present a size of 720×576 pixels.

The private collection consists of 59 additional retinography-angiography pairs. Most of the images correspond to pathological cases, including representative samples of several common ophthalmic diseases. Additionally, the original images presented different sizes and, therefore, they were resized to a fixed size of 720×576. This collection of images has been gathered from the ophthalmic services of Complexo Hospitalario Universitario de Santiago de Compostela (CHUS) in Spain.

To perform the different experiments, the complete multimodal dataset is randomly split into two subsets of equal size, i.e., 59 image pairs each. One of these subsets is held out as a test set and the other is used for training the multimodal reconstruction. Additionally, the training image pairs are randomly split into a validation subset of nine image pairs and a training subset of 50 image pairs. The purpose of this split is to control the training progress through the validation subset, as described in Section 15.3.

Finally, it should be noticed that, although the same subset of image pairs is used for the training of both methodologies, the images are considered as unpaired for the cyclical GAN approach.

15.4.2 Qualitative evaluation of the reconstruction

Firstly, the quality and coherence of the generated angiographies are evaluated through visual analysis. To that end, Figs. 15.7 and 15.8 depict some representative examples of generated images together with the original retinographies and angiographies. The examples are taken from the holdout test set. In general, both methodologies were able to learn an adequate transformation for the main anatomical structures in the retina, namely, the vasculature, fovea, and optic disc. In particular, it is observed that the retinal

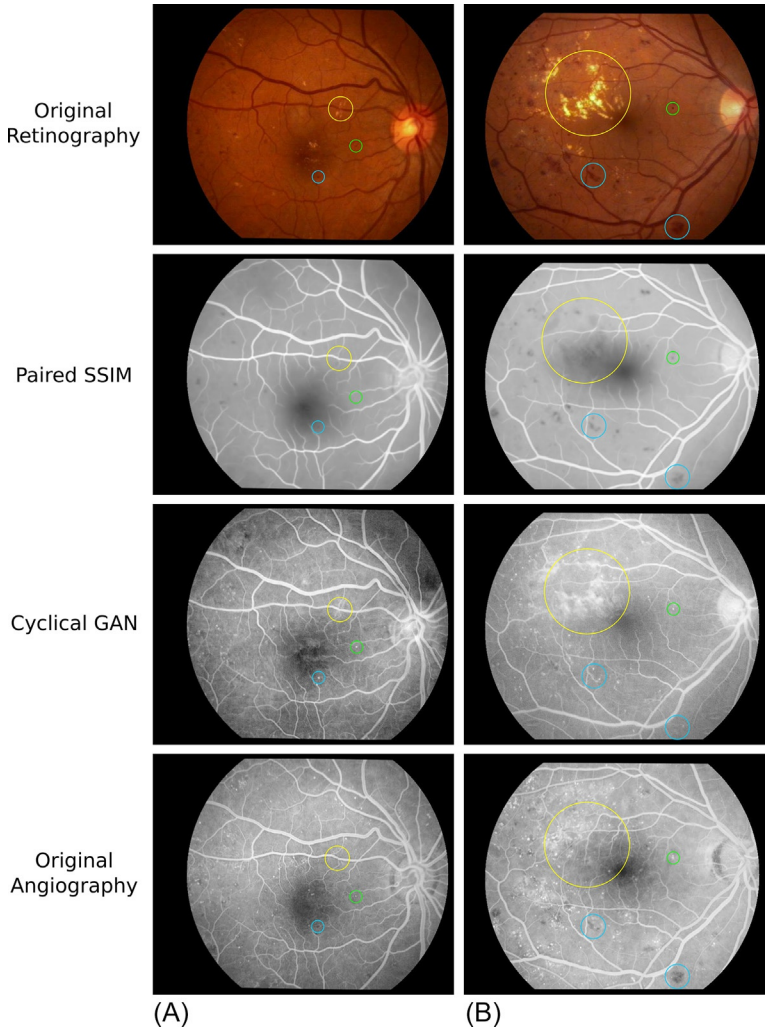

Fig. 15.7 Examples of generated angiographies together with the corresponding original retinographies and angiographies. Some representative examples of microaneurysms *(green)*, microhemorrhages *(blue)*, and bright lesions *(yellow)* are marked with circles.

vasculature is successfully enhanced in all the cases, which is one of the main characteristics of real angiographies. This vascular enhancement evidences a high-level understanding of the different structures in the retina, given that other dark-colored structures in retinography, such as the fovea, are mainly kept with a dark tone in the reconstructed angiographies. This means that the applied transformation is structure-specific and guided by the semantic information in the images instead of low-level

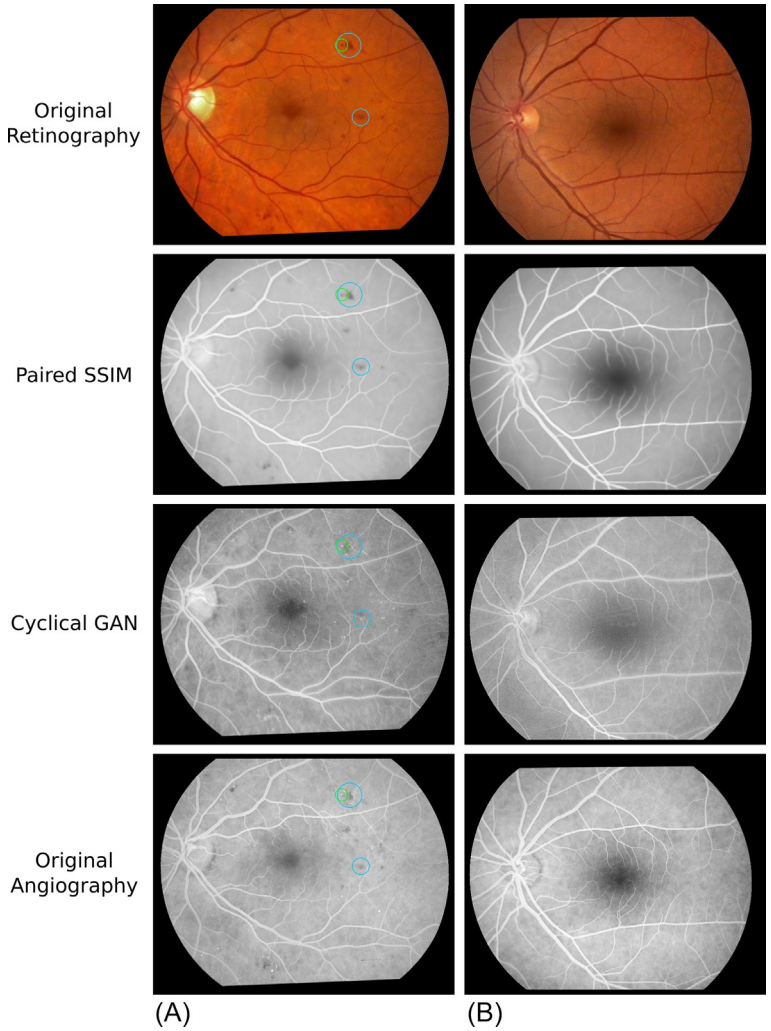

Fig. 15.8 Examples of generated angiographies together with the corresponding original retinographies and angiographies. Some representative examples of microaneurysms *(green)* and microhemorrhages *(blue)* are marked with circles.

information such as the color. In contrast with the vasculature, the reconstructed optic discs are not as similar as those in the real angiographies. However, this can be explained by the fact that the appearance of the optic disc is not as consistent among angiographies. In this sense, both methodologies learn to reconstruct the optic disc with a slight higher intensity, which may indicate that this is the predominant appearance of this anatomical structure in the training set.

With regard to the pathological structures, there are greater differences between the presented methodologies. For instance, microaneurysms are only generated or enhanced by the cyclical GAN methodology. Microaneurysms are tiny vascular lesions that, in contrast to other pathological structures, remain connected to the bloodstream. Therefore, they are directly affected by the injected contrast dye in the angiography. As it is observed in Fig. 15.7, the cyclical GAN methodology is able to enhance these small lesions. However, neither all the microaneurysms in the ground truth angiography are reconstructed nor all the reconstructed microaneurysms are present in the ground truth. This may indicate that part of these microaneurysms are artificially created by the network or that small microhemorrhages are being misidentified as microaneurysms. Nevertheless, it must be considered that the detection of microaneurysms is a very challenging task in the field. Thus, despite the possible errors, the fact that these small structures were identified by the cyclical GAN methodology is a significative outcome.

In contrast to the previous analysis about microaneurysms, the examples of Fig. 15.7 evidence that the paired SSIM methodology provides a better reconstruction for other pathological structures. In particular, bright lesions that are present in the retinography should not be visible in the angiography. However, the cyclical GAN approach fails to completely remove these lesions, especially if they are large such as those in the top-left quarter or the retina shown in Fig. 15.7B. The paired SSIM approach provides a more accurate reconstruction regarding these kinds of lesions although in the previous case there still remains a show in the area of the lesion. Finally, regarding the microhemorrhages, these kinds of lesions are also more accurately reconstructed by the paired SSIM approach. In particular, these lesions present a dark appearance in both retinography and angiography. In the depicted examples, it is observed that paired SSIM reconstructs the microhemorrhages, as expected. However, the cyclical GAN approach tends to remove these lesions. Additionally, in some cases, the small microhemorrhages are reconstructed with a bright tone like the microaneurysms.

Besides the anatomical and pathological structures in the retina, the main difference that is observed between both methodologies is the general appearance of the generated angiographies. In this regard, the images generated by the cyclical GAN present a more realistic look and they could be easier misidentified as real angiographies. The main reason for this is the texture in the images. In particular, cyclical GAN produces a textured retinal background that mimics the appearance of a real angiography. In contrast, the retinal background in the angiographies generated by paired SSIM is very homogeneous, which gives away the synthetic nature of the images. The explanation for this difference between both approaches is the use of GANs in the cyclical GAN methodology. In this sense, the discriminator network has the capacity to learn and distinguish the main characteristics of the angiography, including the textured background. Thus, a synthetic angiography with a smooth background would be easily identified as fake by the discriminator. Consequently, during the training, the generator will learn to generate

the textured background in order to trick the discriminator. In the case of the paired SSIM, the presented results show that SSIM does not provide the feedback that is required to learn this characteristic. Additionally, according to the results presented in Ref. [15], the use of L1-norm or L2-norm in the loss function does not provide that feedback either. In this regard, it should be noticed that these are full-reference pixel-wise metrics that directly compare the network output against a specific ground truth image. Thus, even if an angiography-like texture is generated, this will not necessarily minimize the loss function if the generated texture does not exactly match the one in the provided ground truth. It could be the case that the specific texture of each angiography was impossible to infer from the corresponding retinography. In that scenario, the generator could never completely reduce the loss portion corresponding to the textured background. The resulting outcome could be the generation of a homogeneous background that minimizes the loss throughout the training set. This explanation fits with what is observed in Figs. 15.7 and 15.8.

15.4.3 Quantitative evaluation of the reconstruction

The multimodal reconstruction is quantitatively evaluated by measuring the reconstruction error between the generated and the ground truth angiographies. In particular, the reconstruction is evaluated by means of SSIM, mean average error (MAE), and mean squared error (MSE), which are common evaluation metrics for image reconstruction and image quality assessment. The presented evaluation is performed on the paired data of the holdout test set.

When comparing the two presented methodologies, it must be considered that the paired SSIM relies on the availability of paired data for training. The paired data represent a richer source of information in comparison to the unpaired counterpart and, therefore, it is expected that the paired SSIM provided better performance than cyclical GAN for the same number of training samples. Additionally, it should be also considered that the paired data, despite being commonly available in medical imaging, is inherently harder to collect than the unpaired counterpart. For these reasons, the presented evaluation not only compares the performance of both methodologies when using the complete training set but, also, it compares the performance when there are more unpaired than paired images available for training. This is an expected scenario in practical applications.

The results of the quantitative evaluation are depicted in Fig. 15.9. In the case of paired SSIM, the presented results correspond to several experiments with a varying number of training samples, ranging from 10 to 50 image pairs. In the case of cyclical GAN, the presented results are obtained after training with the complete training subset, i.e., 50 image pairs. Firstly, it is observed that the paired SSIM always provides better results than the cyclical GAN considering SSIM although that is not the case for MAE and MSE. Considering these two metrics, the paired SSIM obtains similar

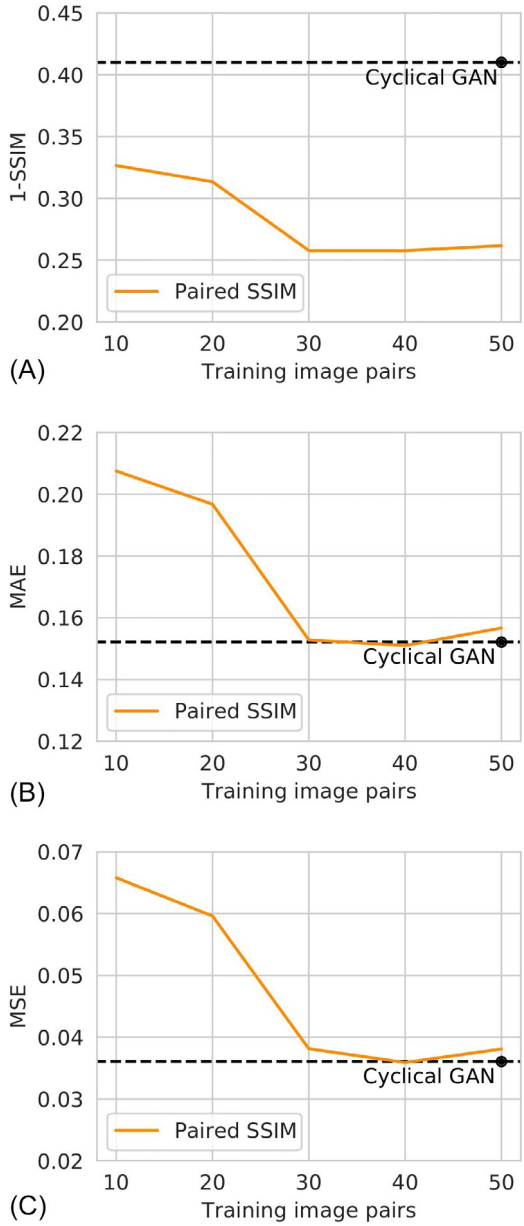

Fig. 15.9 Comparison of cyclical GAN and paired SSIM with a varying number of training samples for paired SSIM. The evaluation is performed by means of (A) SSIM, (B) MAE, and (C) MSE.

or worse results depending on the number of training samples. In general, it is clear that, up to 30 image pairs, the paired SSIM experiments a positive evolution with the addition of more training data. Then, between 30 and 50 image pairs, the evolution stagnates and there is no improvement with the addition of more images. In the case of MAE and MSE, the final results to which the paired SSIM converges are approximately the same as those obtained by the cyclical GAN. This may indicate an existent upper bound in the performance of the multimodal reconstruction with this experimental setting. Regarding the comparison by means of SSIM, there is an important difference between both methodologies independently of the number of training images for paired SSIM. On the one hand, this may be explained by the fact that the generator of the paired SSIM has been explicitly trained to maximize SSIM. Thus, this network excels when it is evaluated by means of this metric. On the other hand, however, it must be considered that SSIM is a more complex metric in comparison to MAE or MSE. In particular, SSIM does not directly measure the difference between pixels but, instead, it measures local similarities that include high-level information such as structural coherence. Thus, it could be possible that subtle structural errors, which are not evidenced by MAE or MSE, contribute to the worse performance of cyclical GAN considering SSIM.

15.4.4 Ablation analysis of the generated images

In order to better understand the obtained results, we present a more detailed quantitative analysis in this section. In particular, the presented analysis considers the possible differences in error distribution among different retinal regions. As it was shown in Section 15.4.2, both methodologies seem to provide a similar enhancement of the retinal vasculature. However, there are important differences in the reconstructed retinal background and certain pathological structures. Therefore, it is interesting to study how the reconstruction error is distributed between the vasculature and the background, and whether this distribution is different between both methodologies. To that end, the reconstruction errors are recalculated using a binary vascular mask to separate between vasculature and background regions. Given that only a broad approximation of the vasculature is necessary, the vascular mask is computed by applying some common image processing techniques. First, the multiscale Laplacian operator proposed in Ref. [31] is applied to the original angiography. This operation further enhances the retinal vasculature, resulting in an image with much greater contrast between vasculature and background [36]. Then, the vascular region is dilated to ensure that the resulting mask not only includes the vessels but also their surrounding pixels. This way, the reconstruction error in the vasculature will also include the error due to inappropriate vessel edges. Finally, the vascular mask is binarized by applying Otsu's thresholding method [37]. An example of the produced binary vascular mask together with the original angiography is depicted in Fig. 15.10.

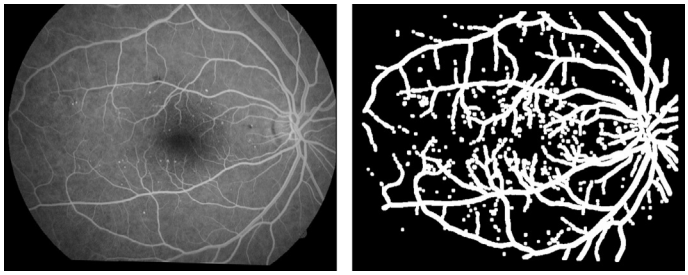

Fig. 15.10 Example of vascular mask used for evaluation: (A) angiography and (B) resulting vessel mask for (A).

The results of the quantitative evaluation using the computed vascular masks are depicted in Fig. 15.11. Firstly, it is observed that, in all the cases, the reconstruction error is greater in the vessels than in the background. This may indicate that the reconstruction of the retinal background is an easier task in comparison to the retinal vasculature. In this regard, it must be noticed that the retinal vasculature is an intricate network with numerous intersection and bifurcations, which increases the difficulty of the reconstruction. The background also includes some pathological structures, which can be a source of errors as seen in Section 15.4.2. However, these pathological structures neither are present in all the images nor occupy a significantly large area of the background. Moreover, the bright lesions in the angiography, i.e., the microaneurysms, are included within the vascular mask, as can be seen in Fig. 15.10. This balances the contribution of the pathological structures between both regions. Regarding the comparison between cyclical GAN and paired SSIM, the analysis is the same as in the previous evaluation. This happens independently of the retinal region that is analyzed, vasculature or background. In particular, the performance of paired SSIM experiments the same evolution with the increase in the number of training images. Considering MAE and MSE, paired SSIM converges again to the same results that are achieved by cyclical GAN, resulting in a similar performance. In contrast, there is still an important difference between the methodologies when considering SSIM.

Finally, it is interesting to observe that the error distribution between regions is the same for paired SSIM and cyclical GAN, even when there is a clear visual difference in the reconstructed background between both methodologies (see Fig. 15.7). This shows that the more realistic look provided by the textured background does not necessarily lead to a better reconstruction in terms of full-reference pixel-wise metrics. In particular, the same reconstruction error can be achieved by producing a homogeneous background with an adequate tone, as paired SSIM does. This explains why the use of these metrics as a loss function does not encourage the generator to produce a textured background. Moreover, in the case of SSIM, which is the metric used by paired SSIM during training, the reconstruction error for the textured background is even greater than that of the homogeneous version.

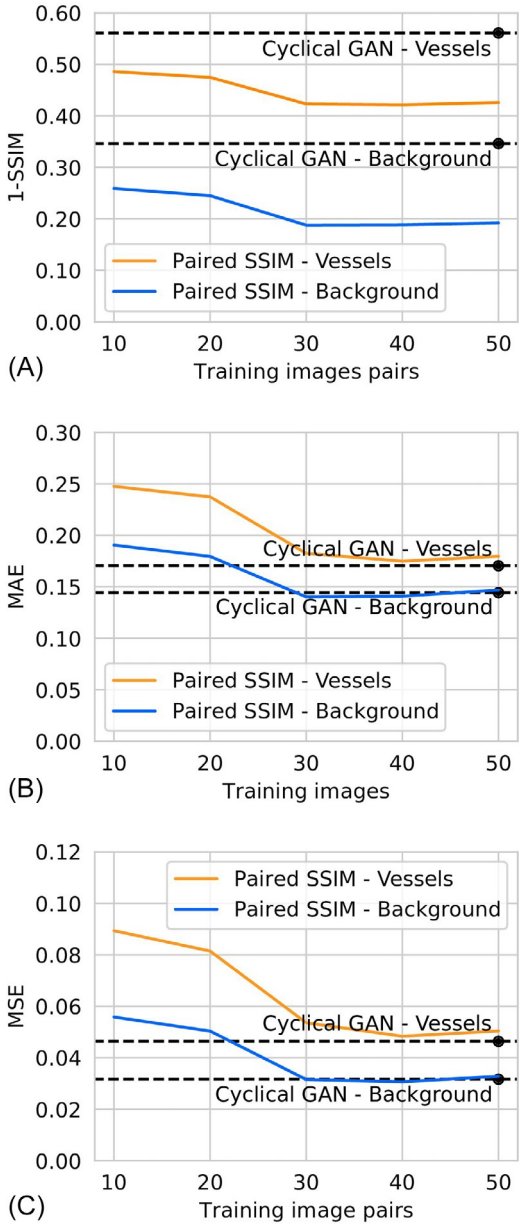

Fig. 15.11 Comparison of cyclical GAN and paired SSIM with a varying number of training samples for paired SSIM. The evaluation is conducted independently for vessels and the background of the images. The evaluation is performed by means of (A) SSIM, (B) MAE, and (C) MSE.

15.4.5 Structural coherence of the generated images

An observation that remains to be explained after the previous analyses is the different results obtained whether the evaluation is performed by means of SSIM or MAE/MSE. In particular, both methodologies achieve similar results in MAE and MSE, although paired SSIM always performs better in terms of SSIM. Given that SSIM is characterized by including higher level information such as the structural coherence between images, the generated images are visually inspected to find possible structural differences. Fig. 15.12 depicts some composite images using a checkerboard pattern that is used to perform the visual analysis. In particular, the depicted images show the generated angiography together with the original retinography (Fig. 15.12A and C) as well as the generated angiography together with the ground truth angiography (Fig. 15.12B and D). At a glance, it seems that both angiographies, from paired SSIM and cyclical GAN, are perfectly reconstructed. However, on closer examination, it is observed that in the angiographies generated by cyclical GAN there are small displacements with respect to the originals. Examples of these displacements are shown in detail in Fig. 15.12. As it is observed, the displacement occurs, at least, in the retinal vasculature. Moreover, it can be observed that the displacement is consistent among the zoomed patches even when they are distant in the images. This indicates that the observed displacement could be the result of an affine transformation.

With regard to the cause of the displacement, an initial hypothesis is based on the fact that cyclical GAN does not put any hard constraint on the structure of the generated angiography. The only requirements are that the image must look like a real angiography and that it must be possible to reconstruct the original retinography from it. Thus, although the more straightforward way to reconstruct the original retinography seems to be to keep the original structure as it is, nothing enforces the networks to do so. Nevertheless, it must be considered that if G_{R2A} applies any spatial transformation to the generated angiographies, and then G_{A2R} must learn to apply the inverse transformation when reconstructing the original retinography. This synergy between the networks is necessary to still minimize the cycle-consistency loss in the cyclical GAN methodology. Although not straightforward, this situation seems plausible given that the observed displacement is very subtle. The presented situation may initiate if the first network, G_{R2A}, starts to reconstruct the vessels of the angiography over the vessel edges of the input retinography. This is likely to happen given the facility of a neural network to detect edges in an image. Moreover, the vessel edges are easier to detect than the vessel centerlines. To verify this hypothesis, the angiographies generated during the first stages of the training have been revised. A representative example of these images is depicted in Fig. 15.13. As it can be observed, there are some bright lines that seems to be drawn over the edges of the subtle dark vessels. This evidences the origin of the issue, although the ultimate cause is the underconstrained training setting of cyclical GAN.

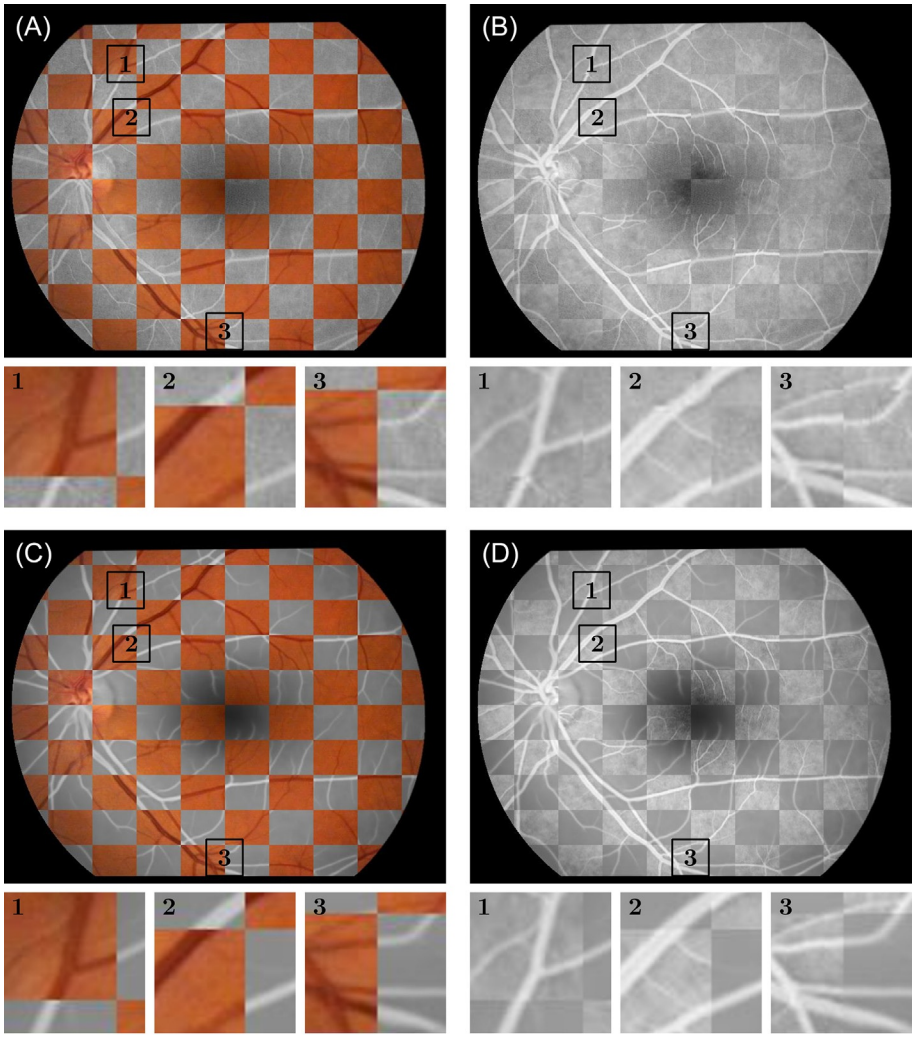

Fig. 15.12 Comparison of generated angiographies against (A, C) the corresponding original retinographies and (B, D) the corresponding ground truth angiographies. (A, B) Angiography generated using paired SSIM. (C, D) Angiography generated using cyclical GAN. Additionally, cropped regions are depicted in detail for each case.

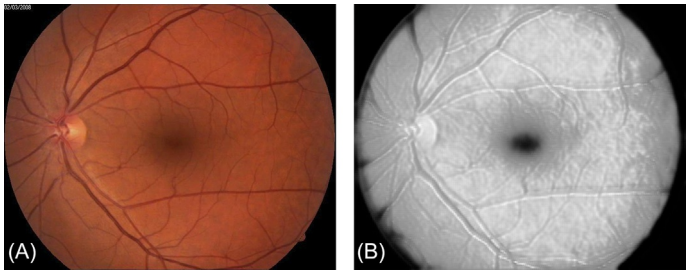

Fig. 15.13 Representative example of generated angiography on the first stages of training for cyclical GAN: (A) original retinography and (B) generated angiography.

15.5 Discussion and conclusions

In this chapter, we have presented a cyclical GAN methodology for the multimodal reconstruction of retinal images [11]. This multimodal reconstruction is a novel task that consists of the translation of medical images between complementary modalities [7]. This allows the estimation of either more invasive or less affordable imaging modalities from a readily available alternative. For instance, this chapter addresses the estimation of fluorescein angiography from retinography, where the former requires the injection of a contrast dye to the patients. Despite the recent technical advances in the field, the direct use of generated images in clinical practice is still only a future potential application. However, there are several other possible applications where this multimodal reconstruction can be taken advantage of. For instance, multimodal reconstruction has already demonstrated to be a successful pretraining task for transfer learning in medical image analysis [13, 14]. This is an important application that reduces the necessity of large collections of expert-annotated data in medical imaging [10].

In order to provide a comprehensive analysis of the cyclical GAN methodology, we have also presented an exhaustive comparison against a state-of-the-art approach where no GANs were used [15]. This way, it is possible to study the particular advantages and disadvantages of using GANs for multimodal reconstruction. The provided comparison is performed under the fairest conditions, by using the same dataset, network architectures, and training strategies. In this regard, the only differences are those intrinsically due to the methodologies themselves. Regarding the presented results, it is seen that both approaches are able to produce an adequate estimation of the angiography from retinography. However, there are important differences in several aspects of the generated angiographies. Moreover, the requirements for training each one of both approaches must also be considered in the comparison.

Regarding the requirements for the training of both approaches, the main difference is the use of unpaired data in cyclical GAN and paired data in paired SSIM. In broad domain applications, i.e., performed in natural images, this would represent an insurmountable obstacle for the paired SSIM methodology. However, in medical imaging, the paired data can be relatively easy to obtain due to the common use of complementary imaging modalities in clinical practice. In this case, however, the disadvantage of paired SSIM is the necessity of registered image pairs where the different anatomical and pathological structures must be aligned. The multimodal registration method that is applied in paired SSIM has demonstrated to be reliable for the alignment of retinography-angiography pairs [31]. Moreover, it has been successfully applied for the registration of the multimodal dataset that is used in the experiments herein described. However, the results presented in Ref. [31] also show that, quantitatively, the registration performance is lower for the most complex cases, which can be due to, e.g., low-quality images or severe pathologies. This could potentially limit the variety of images in an extended

version of the dataset including more challenging scenarios. Additionally, the registration method in paired SSIM is domain-specific and, therefore, cannot be directly applied to other types of multimodal image pairs. This means that the use of paired SSIM in other medical specialties would require the availability of adequate registration methods. Although image registration is a common task in medical imaging, the availability of such multimodal registration algorithms cannot be taken for granted. In contrast, cyclical GAN can be directly applied to any kind of multimodal setting without the need for registered or paired data.

Another important difference between the presented approaches is the complexity of the training procedure. In this sense, cyclical GAN represents a more complex approach including four different neural networks and two training cycles, as described in Section 15.3.1. In comparison, once the multimodal image registration is performed, paired SSIM only requires the training of a single neural network. The use of four different networks in cyclical GAN means that, computationally, more memory is required for training. In a situation of limited resources, which is the common practical scenario, this will negatively affect the size and number of images that is possible for each batch during the training. Moreover, in practice, cyclical GAN also requires longer training times than paired SSIM, which further increases the computational costs. This is in part due to the use of a single network in paired SSIM but also to the use of a full-reference pixel-wise metric for the loss functions. The feedback provided by this more classical alternative results in a faster convergence in comparison to the adversarial training.

Regarding the performance of the multimodal reconstruction, the examples depicted in Figs. 15.7 and 15.8 show that both methodologies are able to successfully recognize the main anatomical structures in the retina. In that sense, despite the evident aesthetic differences, the transformations applied to the anatomical structures are adequate in both cases. Thus, both approaches show a similar potential for transfer learning regarding the analysis of the retinal anatomy. However, when considering the pathological structures, there are important differences between both methodologies. In this case, none of the methodologies perfectly reconstruct all the lesions. In particular, the examples depicted in Fig. 15.7 indicate that each methodology gives preference to different types of lesions in the generated images. Thus, it is not clear which alternative would be a better option toward the pathological analysis of the retinal images. In this regard, given the mixed results that are obtained, future works could explore the development of hybrid methods for the multimodal reconstruction of retinal images. The objective, in this case, would be to combine the good properties of cyclical GAN and paired SSIM.

One of the main differences between cyclical GAN and paired SSIM is the appearance of the generated angiographies. Due to the use of a GAN framework in cyclical GAN, the generated angiographies look realistic and aesthetically pleasing. In contrast, the angiographies generated by paired SSIM present a more synthetic appearance. The importance

of this difference in the appearance of the generated angiographies depends on the specific application. On the one hand, for representation learning purposes, the priority is the proper recognition of the different retinal structures. Additionally, even for the potential clinical interpretation of the images, realism is not as important as the accurate reconstruction of the different structures. On the other hand, there exist potential applications such as data augmentation or clinical simulations where the realism of the images is of great importance.

Finally, a relevant observation presented in this chapter is the fact that cyclical GAN does not necessarily keep the exact same structure of the input image. This is a known possible issue, given the underconstrained training setting in cyclical GANs. Nevertheless, in this chapter, we have presented empirical evidence of this issue in the form of small displacements for the reconstructed blood vessels. According to the evidence presented in Section 15.4.5, it is not possible to predict whether these displacements will happen or how they will exactly be. In this sense, the particular structural displacements produced by the networks is affected by the stochasticity of the training procedure. Moreover, although we have only noticed these structural incoherence in the blood vessels, it would be possible to note the existence of similar subtle structural transformations for other elements in the images. In line with prior observations in the presented comparison, the importance of these structural errors depends on the specific application for which the multimodal reconstruction is applied. For instance, this kind of small structural variations should not significantly affect the quality of the internal representations learned by the network. However, they would impede the use of cyclical GAN as a tool for accurate multimodal image registration. The development of hybrid methodologies, as previously discussed, could also be a solution to this structural issue while keeping the good properties of GANs. For instance, according to the results presented in Section 15.4.3, the addition of a small number of paired training samples could be sufficient for improving the structural coherence of the cyclical GAN approach. Additionally, a hybrid approach of this kind could still incorporate those more challenging paired images that may not be successfully registered.

To conclude, the presented cyclical GAN approach has been demonstrated to be a valid alternative for the multimodal reconstruction of retinal images. In particular, the provided comparison shows that cyclical GAN has both advantages and disadvantages with respect to the state-of-the-art approach paired SSIM. In this regard, these two approaches are complementary to each other when considering their strengths and weaknesses. This motivates the future development of hybrid methods aiming at taking advantage of the strengths of both alternatives.

Acknowledgments

This work was supported by Instituto de Salud Carlos III, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project, and