UNIVERSIDADE DA CORUÑA

Deep Learning techniques for automated analysis and processing of high resolution medical imaging

Doctoral Thesis

Álvaro Suárez Hervella
2021

PhD advisors:
Jorge Novo Buján
José Rouco Maseda

Deep Learning techniques for automated analysis and processing of high resolution medical imaging

Doctoral Thesis

Álvaro Suárez Hervella

2021

PhD advisors:
Jorge Novo Buján
José Rouco Maseda

PhD degree in Computer Science

UNIVERSIDADE DA CORUÑA

Facultade de Informática
Campus de Elviña s/n
15071, A Coruña (Spain)

José Rouco Maseda, Assistant Professor at the Department of Computer Science and Information Technologies of University of A Coruña,

Jorge Novo Buján, Associate Professor at the Department of Computer Science and Information Technologies of University of A Coruña,

and

Marcos Ortega Hortas, Associate Professor at the Department of Computer Science and Information Technologies of University of A Coruña,

HEREBY CERTIFY

that the present PhD Thesis entitled ***Deep Learning techniques for automated analysis and processing of high resolution medical imaging***, submitted to the University of A Coruña by *Álvaro Suárez Hervella*, has been carried out under the supervision of José Rouco Maseda and Jorge Novo Buján and the tutoring of Marcos Ortega Hortas, and fulfills all the requirements for the award of the degree of *PhD in Computer Science with International Mention.*

José Rouco Maseda	Jorge Novo Buján	Marcos Ortega Hortas
Advisor	Advisor	Tutor

Acknowledgments

First of all, I would like to thank my PhD advisors, José Rouco and Jorge Novo, for their guidance during this journey. Ultimately, their invaluable mentorship is what makes it possible for me to write these words today. I am especially thankful for their wise advice and patience, which have allowed me to learn and grow during this important stage of my life. It was an honor for me to share their scientific knowledge and learn from their experience.

I am also very grateful to all my present and past colleagues of the VARPA research group. I feel really fortunate to have met all of them, not only for their technical or administrative support during these years, but also for their human qualities and the good times we have had together. I would especially like to thank my colleagues at CITIC for their warm welcome and their priceless company throughout these years.

I would also like to thank Jaime Cardoso and the people of the VCMI research group at INESC TEC for the warm welcome. In spite of the circumstances, being able to share other perspectives has been an enriching experience for me.

This PhD represents a long journey full of enriching experiences. A journey that began even before I considered it, long before I arrived here. I want to thank all those people who crossed my path and served me as support and guide along the way. Among others, I would especially like to thank my past colleagues at CEIT.

Finally, I would like to thank my family and friends who have shown me their support over the years.

Great journeys require great travel companions. Thank you all.

The most exciting phrase to hear in science,
the one that heralds the most discoveries,
is not "Eureka!" but "That's funny..."
Isaac Asimov

Abstract

Medical imaging plays a prominent role in modern clinical practice for numerous medical specialties. For instance, in ophthalmology, different imaging techniques are commonly used to visualize and study the eye fundus. In this context, automated image analysis methods are key towards facilitating the early diagnosis and adequate treatment of several diseases. Nowadays, deep learning algorithms have already demonstrated a remarkable performance for different image analysis tasks. However, these approaches typically require large amounts of annotated data for the training of deep neural networks. This complicates the adoption of deep learning approaches, especially in areas where large scale annotated datasets are harder to obtain, such as in medical imaging.

This thesis aims to explore novel approaches for the automated analysis of medical images, particularly in ophthalmology. In this regard, the main focus is on the development of novel deep learning-based approaches that do not require large amounts of annotated training data and can be applied to high resolution images. For that purpose, we have presented a novel paradigm that allows to take advantage of unlabeled complementary image modalities for the training of deep neural networks. Additionally, we have also developed novel approaches for the detailed analysis of eye fundus images. In that regard, this thesis explores the analysis of relevant retinal structures as well as the diagnosis of different retinal diseases. In general, the developed algorithms provide satisfactory results for the analysis of the eye fundus, even when limited annotated training data is available.

Resumen

Las técnicas de imagen tienen un papel destacado en la práctica clínica moderna de numerosas especialidades médicas. Por ejemplo, en oftalmología es común el uso de diferentes técnicas de imagen para visualizar y estudiar el fondo de ojo. En este contexto, los métodos automáticos de análisis de imagen son clave para facilitar el diagnóstico precoz y el tratamiento adecuado de diversas enfermedades. En la actualidad, los algoritmos de aprendizaje profundo ya han demostrado un notable rendimiento en diferentes tareas de análisis de imagen. Sin embargo, estos métodos suelen necesitar grandes cantidades de datos etiquetados para el entrenamiento de las redes neuronales profundas. Esto complica la adopción de los métodos de aprendizaje profundo, especialmente en áreas donde los conjuntos masivos de datos etiquetados son más difíciles de obtener, como es el caso de la imagen médica.

Esta tesis tiene como objetivo explorar nuevos métodos para el análisis automático de imagen médica, concretamente en oftalmología. En este sentido, el foco principal es el desarrollo de nuevos métodos basados en aprendizaje profundo que no requieran grandes cantidades de datos etiquetados para el entrenamiento y puedan aplicarse a imágenes de alta resolución. Para ello, hemos presentado un nuevo paradigma que permite aprovechar modalidades de imagen complementarias no etiquetadas para el entrenamiento de redes neuronales profundas. Además, también hemos desarrollado nuevos métodos para el análisis en detalle de las imágenes del fondo de ojo. En este sentido, esta tesis explora el análisis de estructuras retinianas relevantes, así como el diagnóstico de diferentes enfermedades de la retina. En general, los algoritmos desarrollados proporcionan resultados satisfactorios para el análisis de las imágenes de fondo de ojo, incluso cuando la disponibilidad de datos de entrenamiento etiquetados es limitada.

Resumo

As técnicas de imaxe teñen un papel destacado na práctica clínica moderna de numerosas especialidades médicas. Por exemplo, en oftalmoloxía é común o uso de diferentes técnicas de imaxe para visualizar e estudar o fondo de ollo. Neste contexto, os métodos automáticos de análises de imaxe son clave para facilitar o diagnóstico precoz e o tratamento adecuado de diversas enfermidades. Na actualidade, os algoritmos de aprendizaxe profunda xa demostraron un notable rendemento en diferentes tarefas de análises de imaxe. Con todo, estes métodos adoitan necesitar grandes cantidades de datos etiquetos para o adestramento das redes neuronais profundas. Isto complica a adopción dos métodos de aprendizaxe profunda, especialmente en áreas onde os conxuntos masivos de datos etiquetados son máis difíciles de obter, como é o caso da imaxe médica.

Esta tese ten como obxectivo explorar novos métodos para a análise automática de imaxe médica, concretamente en oftalmoloxía. Neste sentido, o foco principal é o desenvolvemento de novos métodos baseados en aprendizaxe profunda que non requiran grandes cantidades de datos etiquetados para o adestramento e poidan aplicarse a imaxes de alta resolución. Para iso, presentamos un novo paradigma que permite aproveitar modalidades de imaxe complementarias non etiquetadas para o adestramento de redes neuronais profundas. Ademais, tamén desenvolvemos novos métodos para a análise en detalle das imaxes do fondo de ollo. Neste sentido, esta tese explora a análise de estruturas retinianas relevantes, así como o diagnóstico de diferentes enfermidades da retina. En xeral, os algoritmos desenvolvidos proporcionan resultados satisfactorios para a análise das imaxes de fondo de ollo, mesmo cando a dispoñibilidade de datos de adestramento etiquetados é limitada.

Contents

Acronyms

3D Three-Dimensional

AMD Age-related Macular Degeneration

CAD Computer-Aided Diagnosis

CNN Convolutional Neural Network

DNN Deep Neural Network

FA Fluorescein Angiography

GAN Generative Adversarial Network

MR Multimodal Reconstruction

MSL Multi-Scale Laplacian

NCC Normalized Cross-Correlation

OCT Optical Coherence Tomography

ROI Region Of Interest

SLO Scanning Laser Ophthalmoscopy

List of Figures

Chapter 1

Introduction

In accordance with the regulations of the University of A Coruña for doctoral studies, this PhD dissertation is structured as a compilation thesis consisting of several published research articles. In that regard, this PhD thesis presents a first chapter that briefly summarizes and discusses the research work included in the thesis. First, this introductory chapter provides the motivation and context for the research work that is included in the thesis as well as the intended objectives. Then, in order to provide coherence and consistency to the compilation thesis, a brief discussion about the different research articles is included. Finally, general conclusions are drawn and potential future works derived from this PhD thesis are discussed.

1.1 Background and Motivation

Medical imaging plays a prominent role in modern clinical research and practice [1]. Nowadays, multiple imaging modalities are commonly used in medicine to facilitate the diagnosis, treatment, and follow-up of the patients [1, 2]. These techniques allow the visualization and study of the different organs and tissues in the human body [1]. Thus, they can be used by the clinicians to analyze the different anatomical structures that may be affected by a disease or to find potential pathological lesions. However, in many cases, the analysis of these images is a very challenging and tedious task [3, 4]. For instance, it is common for many diseases to only show subtle abnormalities or very small lesions at their earliest stages. In order to detect and adequately analyze these subtle evidences of disease, the analysis of the images must be carefully done by experienced clinicians. In this regard, automated medical image analysis tools arise as a crucial aid to the clinicians, helping to alleviate their workload and potentially improve the reliability of the diagnosis [5, 6, 7].

The use of multiple imaging modalities is broadly extended in the study of the

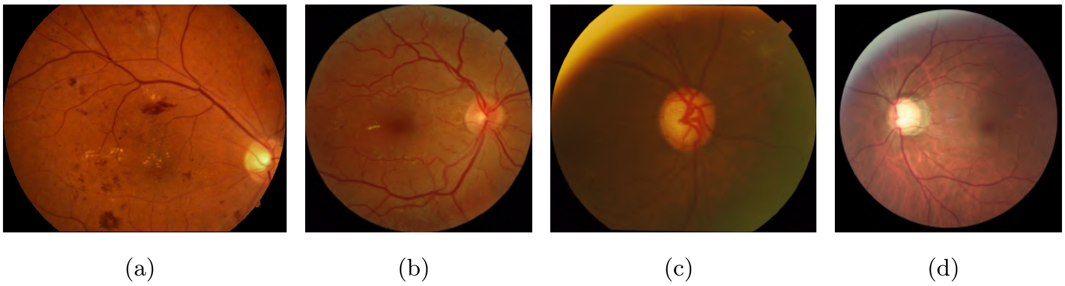


Figure 1.1: Representative examples of retinal images for different eyes.

human eye [2, 1]. The analysis of retinal (or eye fundus) images is crucial for the diagnosis of numerous pathological conditions [8], including ophthalmic disorders such as glaucoma [9] or Age-related Macular Degeneration (AMD) [10] as well as systemic diseases that affect the eye such as diabetes [11] or hypertension [12]. Nowadays, the most affordable and widely available retinal imaging modality is color retinography (or color fundus photography) [13, 6]. These retinal images are color photographs of the eye fundus, i.e. the back surface of the eye, that depict relevant anatomical structures such as the retinal microvasculature, the fovea, or the optic disc. Additionally, pathological structures that are relevant for the diagnosis of numerous diseases, such as hemorrhages, exudates, or drusen are also easily observable in these images. Figure 1.1 depicts representative examples of eye fundus images for different eyes.

Besides color retinography, there are also other retinal imaging modalities such as Fluorescein Angiography (FA), Scanning Laser Ophthalmoscopy (SLO), or Optical Coherence Tomography (OCT) [2, 13]. These techniques typically provide some advantages regarding the visualization of the retinal structures and tissues. However, they are also much less commonly available due to the necessity of more complex equipment or invasive procedures for the patients. In contrast, color retinography is a non invasive technique that can be performed with relatively cheap equipment. In this regard, nowadays, it is even possible to take fundus photographs with specialized portable devices [14]. For these reasons, color retinography represents a valuable tool in the context of preventive health programs and the screening of large populations [5, 6].

In recent years, there has been a great interest in the development of automated methods for the analysis of eye fundus images [15, 16, 17]. In this regard, there are several examples of Computer-Aided Diagnosis (CAD) systems being used in different health services or screening programs worldwide [5]. Currently, the most successful approaches are those that are based on deep learning algorithms [18, 17]. Similarly to other computer vision areas, the use of Deep Neural Networks (DNNs) has resulted in a significantly improvement in terms of performance for many medical applications [18]. Additionally, these algorithms typically provide more straightforward and adaptable methodologies, avoiding ad-hoc processing steps such as the hand-engineering of features that is required for classical machine learning algorithms [19, 20].

The recent rise and spread of deep learning has been motivated by different factors, including technical developments that facilitated the training of DNNs, the availability of large-scale datasets, or the increased computational power that is usually available [21, 22]. In this regard, for instance, the field has been pushed forward by the celebration of several challenges that resulted in large annotated datasets being publicly available for researchers worldwide [23, 24, 25]. However, simultaneously, the availability of annotated data is still a key limiting factor for the successful application of deep learning algorithms in numerous areas. This is a particular prevalent issue in medical imaging, given that the manual annotation of the images requires profound medical knowledge and a high level of expertise [26, 18, 27]. In this regard, the image labeling should ideally be performed by clinical specialists with years of practical experience in the kind of analysis that is required. Additionally, given the high inter-expert variability that can be expected for some specially challenging analyses, it is commonly required to have a consensus taken into account the annotations of several experts [3, 4]. These factors typically limit the size of the annotated datasets that are available in medical imaging.

The scarcity of annotated data in medical imaging can be alleviated following different approaches [18, 27]. Firstly, in the case of image-level annotations, additional labels may be distilled from clinical reports of existing patients [18]. However, this approach cannot be applied for pixel-level annotations, which are required for tasks such as image segmentation. Additionally, the manual annotation of pixel-level labels is specially tedious and challenging, which is reflected in the significantly lower number of annotated images for these kinds of tasks [28, 29]. Secondly, data augmentation strategies are commonly used in the field and represent a key tool to achieve a successful performance with limited training data [18]. These approaches aim at simulating new plausible samples by applying color and spatial transformations to the available annotated images [30]. In this context, there is also an increasing interest in the development of automated methods for the generation of synthetic data samples using DNNs. These kinds of approaches aim at increasing the variability of samples that can be obtained without the necessity of any ad-hoc image processing [31]. However, they present the inherent risk of producing non-plausible contents in the images [32]. Finally, a broadly extended approach for the training of DNNs is transfer learning [18, 27, 33]. In general terms, transfer learning consist in leveraging the knowledge acquired from the training of one task to solve another related problem. This approach is typically applied in a sequential fashion by first pre-training a DNN in an auxiliary task with a large annotated dataset and then fine-tuning the network in a target task with limited annotations. However, it is also possible to take advantage of this approach by simultaneously training both tasks. In this multi-task setting, both tasks could benefit from the training data of each other [34, 35]

For years, the usual approach for transfer learning in medical imaging has been the use of a fully-supervised pre-training performed on a large-scale dataset of natural images [36, 37]. In this regard, the use of classification networks pre-trained on the ImageNet dataset [24] is broadly extended and considered a standard procedure. Despite the different nature of the pre-training images, this approach has demonstrated to provide a rich set of learned features that facilitates the training of numerous target tasks regardless of the final application domain [33]. Still, it could be argued that an in-domain pre-training of similar characteristics should provide more relevant high level representations, resulting in an improved transfer learning performance.

Recently, self-supervised learning has arisen as a promising alternative to traditional fully-supervised approaches for transfer learning [38, 39]. Following the self-supervised paradigm, the training targets (or labels) are automatically derived from the raw unlabeled training data. Thus, a standard supervisory signal can be provided to the network without involving any manual annotation. This facilitates the pre-training of a DNN using images of the final application domain, resulting in learned features that are potentially more useful for the desired target task. Existing self-supervised approaches could be broadly split into either generative or contrastive tasks [40]. The generative self-supervised family is based on the prediction of hidden samples of the data or the prediction of hidden relations among different data samples [40]. For instance, this type of self-supervised learning can be performed by predicting masked or shuffled regions in an input image [41, 42], predicting geometric relations between different object candidates [43], or predicting the temporal relation among different frames in a video [44, 45]. Additionally, during this PhD thesis, we have also proposed a novel self-supervised alternative consisting on the prediction of a complementary medical imaging modality [46, 47]. With regards to the contrastive self-supervised family, the training objective is to obtain a high level representation that maximizes the similarity among related data samples [40, 39].

These related samples are typically obtained by applying standard data augmentation techniques to the raw unlabeled data. This type of self-supervised learning has been recently explored in several works, which proposed different network architectures and training procedures aiming at better taking advantage of the contrastive learning paradigm [48, 49].

1.2 Objectives

The herein presented PhD thesis is focused on the development of novel deep learning approaches for the automated analysis of medical images. Particularly, the aim is to apply the developed methodologies to the automated analysis of retinal images. The main objectives of the PhD thesis can be summarized as follows:

- Development of novel deep learning-based approaches for medical image analysis that reduce the necessity of large-scale manually annotated datasets and can be applied to high resolution images.

- Development of novel methodologies for medical image analysis to improve the prevention and diagnosis of ophthalmic and vascular diseases.

Also, the following specific objectives of the PhD thesis are defined:

- Improve the detection and analysis of anatomical and pathological structures in color retinography.

- Obtain an automated enhancement of the retinal micro-vasculature in color retinography.

- Explore the use of multiple imaging modalities for the developed algorithms.

- Development of methodologies that do not require large-scale manually annotated datasets.

- Development of methodologies that can be applied to high resolution images.

1.3 Research and General Discussion

This section provides the reader with an overview of the research work included in the PhD thesis. Particularly, the section provides a brief summary and a general discussion of all the appended publications that constitute this compilation thesis. The research work included in the compilation thesis comprises 4 JCR-indexed Journal

Papers, 1 Book Chapter, and 4 Proceedings Papers of International Conferences. Attending to their particular research topics, these publications are organized into 3 different blocks: Parte I - Reconstrutión Multimodal de Imágenes de la Retina, Parte II - Análisis de Estructuras de la Retina, Parte III - Diagnóstico Asistido por Ordenador.

1.3.1 Part I - Multimodal Reconstruction of Retinal Images

The scarcity of annotated data is a relevant issue for the development of deep learning algorithms in medical imaging [27, 18, 37]. During several years, fully-supervised pre-training on the ImageNet dataset has been the go-to approach for addressing this issue [27, 36, 37]. However, this approach still relies on the availability of large amounts of annotated data, though from a different application domain. For this reason, in recent years there is a increasing interest in the development of self-supervised approaches that do not require manual labeling of the images [40, 38]. In this regard, currently, there are numerous different self-supervised auxiliary tasks that can be used for transfer learning in medical imaging. However, besides the typical generative and contrastive approaches that are also used in natural images [40], additional potential free sources of supervision can be found in medical imaging.

In modern clinical practice, it is common the use of complementary imaging modalities that provide alternative visualizations of the same organs or tissues [2, 50, 13]. The different visual characteristics between these modalities are mainly due to the use of different capture devices o injected contrasts that enhance the visualization of certain tissues in the images. In this regard, the clinicians must choose the most adequate image modality for each case, attending to different factors such as the patient risk level or the particular anatomical structure to be analyzed. Additionally, in the most complex cases, it is common the use of multiple complementary imaging modalities. This eases the gathering of multimodal image collections for research purposes. Nevertheless, these multimodal datasets are typically only used when labels for the images are also available. In this regard, there are several examples in the literature of automated approaches that make a prediction based on a multimodal input consisting of several complementary imaging modalities [51]. In this case, the multimodal data must be available both at training and inference time and should also be annotated for the training phase. However, the differences between complementary imaging modalities represent a rich source of supervision in itself, without any manually annotated label being involved. For instance, in this PhD thesis, we have propose a novel approach for self-supervised learning purposes that consists in the prediction of one image modality from another [46]. In order

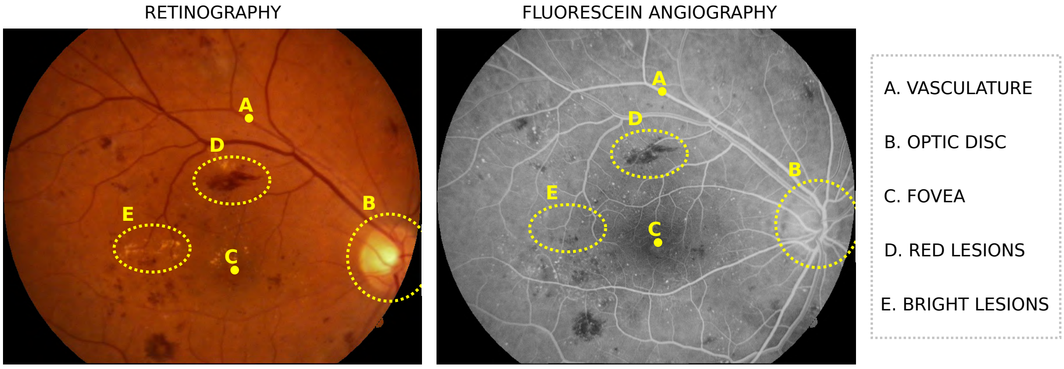


Figure 1.2: Representative example of retinography and FA for the same eye. The main anatomical structures in the retina and relevant lesions are highlighted.

to solve this task, a DNN must first recognize the different elements that compose the input image, including different anatomical and pathological structures. Then, the neural network must apply the most adequate transformation for each of the identified elements and generate as output the compositions of all the transformed elements. This complex process requires the ability to recognize numerous domain-specific patterns in the images as well as a high level knowledge of the image contents. Thus, this multimodal reconstruction of complementary image modalities can be used as a self-supervised auxiliary task for transfer learning purposes. Additionally, the multimodal reconstruction itself provides a potentially valuable estimation of an additional imaging modality.

During this PhD thesis, we explored the idea of the multimodal reconstruction in the context of retinal image analysis. In particular, our aim was the development of novel methods for the analysis of color retinography, hence this image modality is used as input in the proposed multimodal reconstruction approach. As target image modality, we used FA, a complementary modality that provides an enhanced visualization of the retinal microvasculature. In this case, the injection of a contrast dye produces a drastic change in the appearance of the different anatomical and pathological structures in the images. In this regard, besides the evident changes in intensity and color, several structures that are almost imperceptible in one modality may be clearly visible in the other. This can be seen in Figure 1.2, which depicts a representative example of retinography and FA for the same eye.

The proposed MR methodology is based on the use of paired multimodal data, particularly retinography-FA pairs where both images correspond to the same eye. These paired data can be easily gathered due to the fact that retinography is also typically available when a FA is taken for a certain patient [2]. In order to completely take advantage of the paired images, the multimodal image pairs are aligned to establish a pixel-wise correspondence between modalities. This facilitates the training of a DNN in the MR by allowing the use of common full-reference metrics as loss function.

For the purpose of alignment of the multimodal image pairs, we proposed a novel multimodal image registration methodology in [52]. The proposed methodology is an hybrid approach that combines landmark-based and intensity-based registration methods. Additionally, particular characteristics of the retinal images, such as the complexity of the retinal vasculature tree, are exploited for both registration steps. In the first step, vessel crossings and bifurcations in the retinal vasculature tree are used as landmarks. The detection of these landmarks is performed by adapting the method of Ortega et al. [53] to the multimodal scenario. This method has previously demonstrated to be successful in the field of retinal image verification using retinography [53]. Figure 1.3 depicts some examples of detected landmarks for a retinography-FA image pair. Then, the landmarks-based registration is performed by matching the corresponding points between images and estimating a rigid transformation between them. In order to discard outlier points a RANSAC algorithm is used for the point-matching and transformation estimation. For the second registration step, a Multi-Scale Laplacian (MSL) transformation is applied to the images. This transformation converts both image modalities to a common image space where the retinal microvasculature is enhanced in the images. Figure 1.4 depicts some examples of MSL representations for both retinography and FA. This common representation for both modalities allows the use of standard similarity metrics between the images. In this particular case, we use Normalized Cross-Correlation (NCC). The intensity-based registration is performed by estimating the spatial transformation that maximizes the similarity between both images. In this case, both rigid and deformable transformations are used. Figure 1.5 depicts some representative examples of registered multimodal image pairs.

The methodology for the MR using paired and registered multimodal images was presented in [46]. This methodology is based on the use of a standard Convolutional Neural Network (CNN). In particular, we adopted the U-Net [54] architecture that is commonly used in medical imaging and represents a well-known baseline. For the training of the network, we explored different loss functions. In particular, we considered both L1 and L2 metrics, which have been previously used in several problems of similar characteristics. Besides these metrics, we also explored the use of the negative Structure Similarity (SSIM) as loss function. SSIM [55] is a similarity metric originally proposed for image quality assessment. This metrics presents the

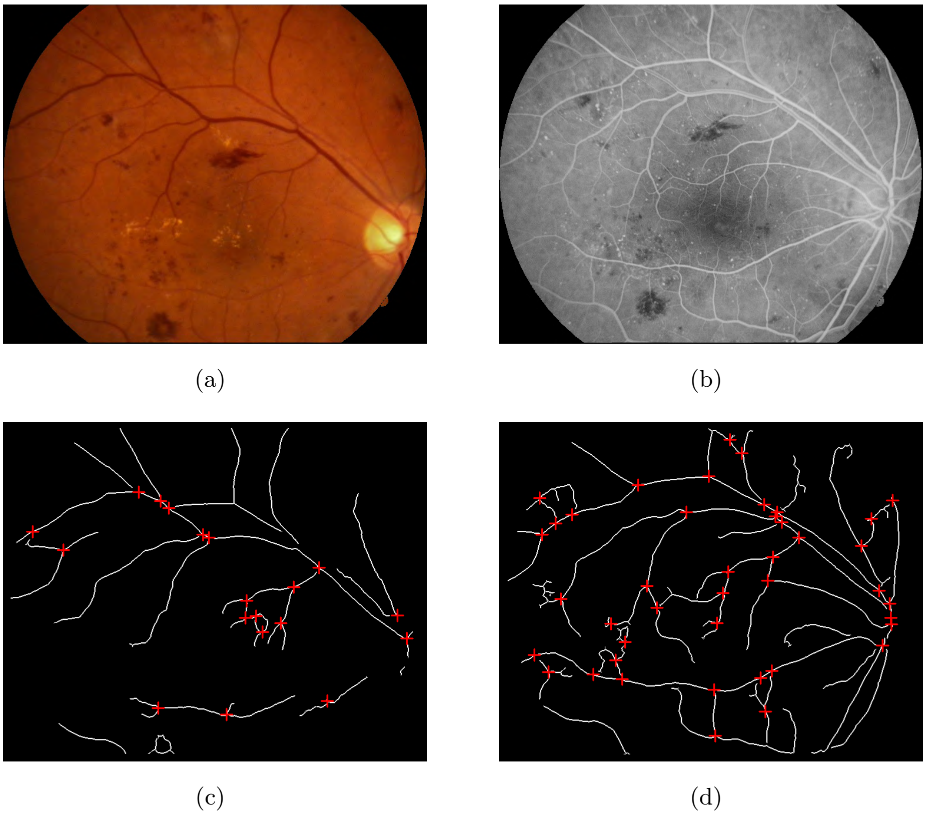


Figure 1.3: Example of vessel tree and detected landmarks for a retinography-FA pair from a diabetic retinopathy patient. (a) Retinography. (b) FA. (c) Vessel tree and landmarks from (a). (d) Vessel tree and landmarks from (b).

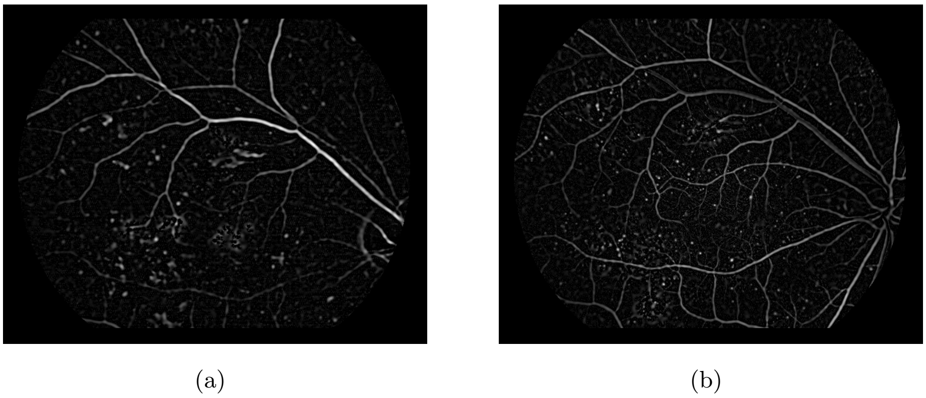


Figure 1.4: Example of MSL maps for the retinography-FA pair depicted in Figure 1.3. (a) MSL map for the retinography. (b) MSL map for the FA.

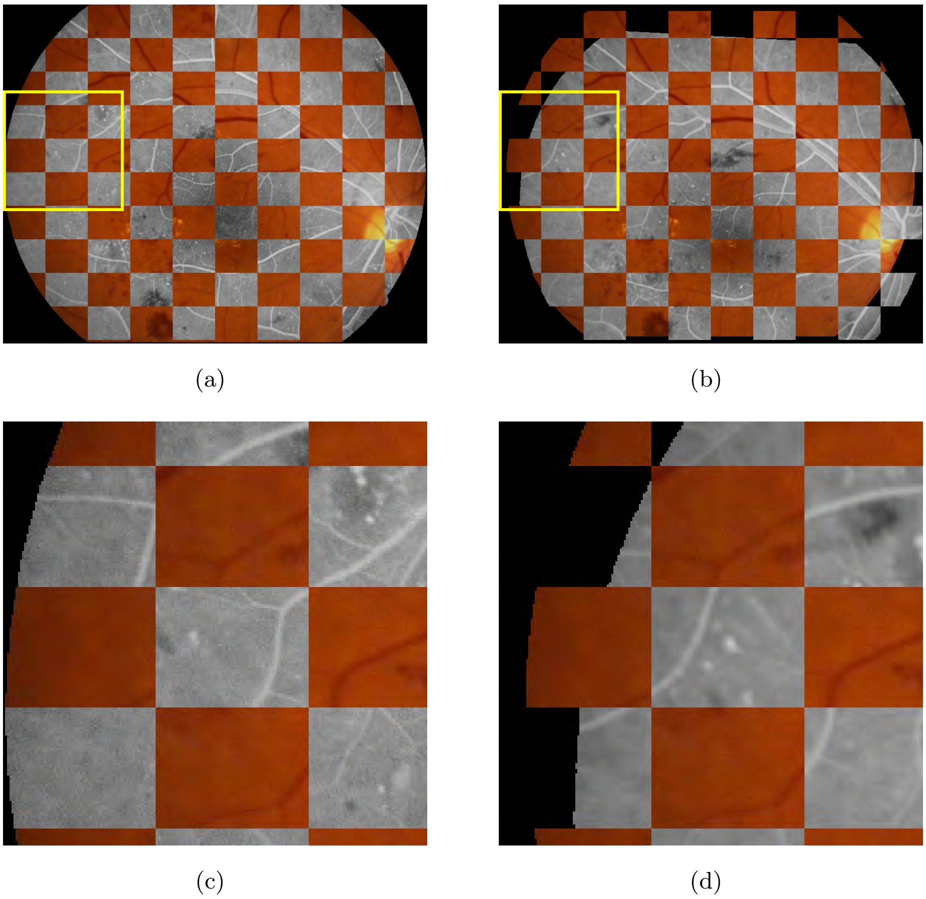
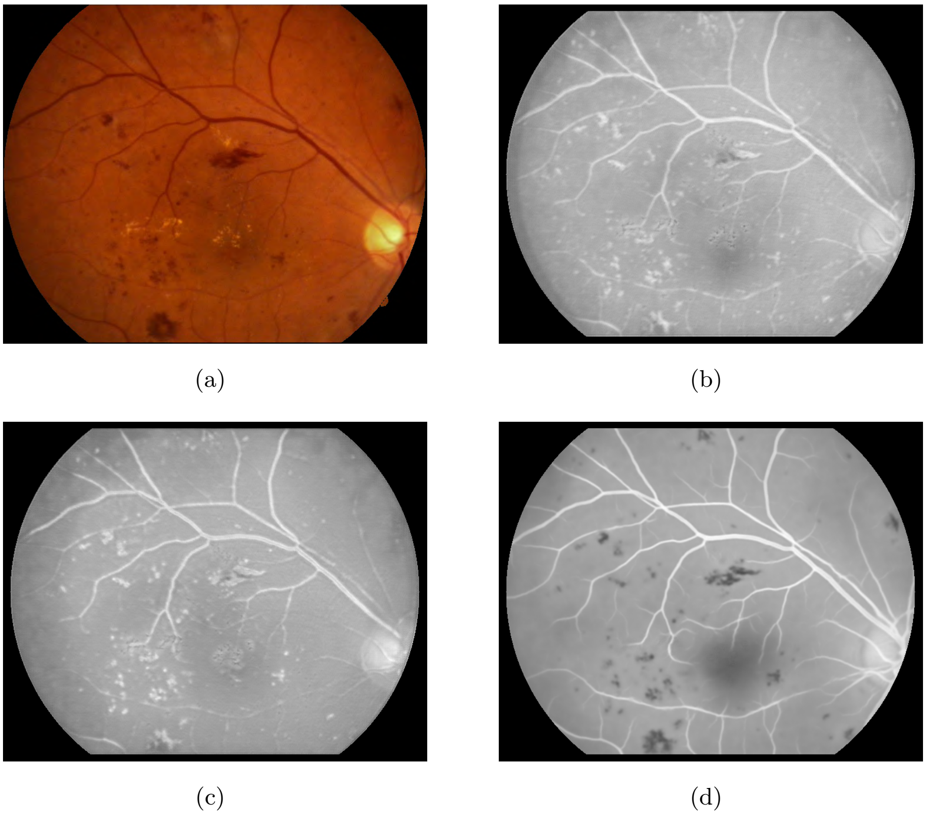


Figure 1.5: Example of the multimodal registration for a retinography-FA pair. (a) Before the registration. (b) After the registration. (c) Detail from (a). (d) Detail from (b).

particularity of considering the intensity, contrast, and structural differences between images. To accomplish this, SSIM is computed using a set of local statistics for each pixel position. For instance, the mean is included for measuring the intensity, the variance for the contrast, and the covariance for the image structure. These measures provided a more complete picture of the differences between images, which can potentially overcome some of the limitations of L1 and L2. For instance, it is known that L2 typically leads to generation of blurry images [21]. Examples of generated FA images using the proposed MR methodology are depicted in Figure 1.6. It is observed that L1 and L2 produce a similar outcome, whereas SSIM produces a more detailed output with sharper structures. In this regard, the best results are clearly achieved by SSIM, which can be explained by the previous discussed factors. Additionally, it should also be considered that SSIM may be more robust to slight misalignments that may remain between the images after the multimodal registration.



(a)

(b)

(c)

(d)

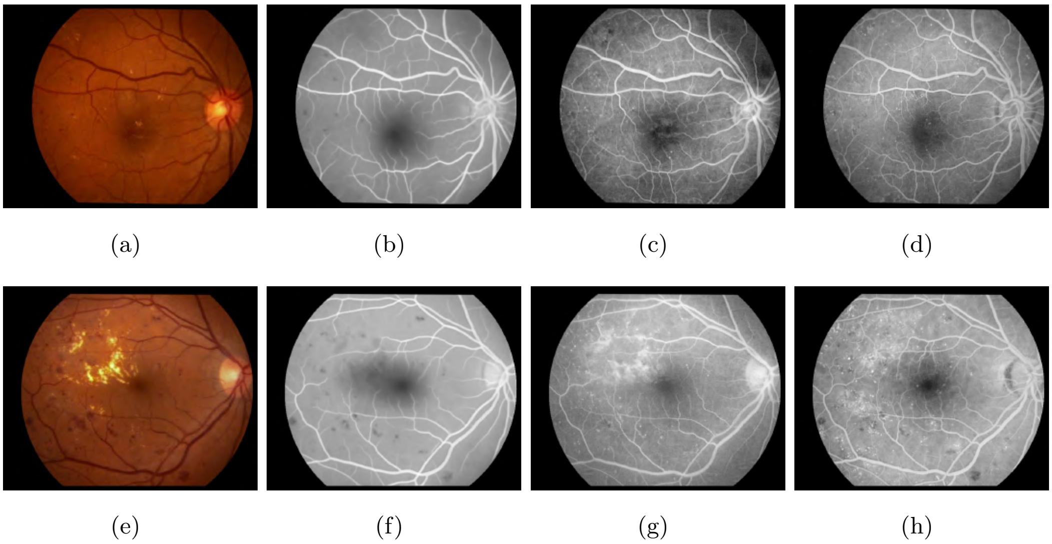
Figure 1.6: Representative examples of generated FAs. (a) Original retinography. (b) Using the L2 training loss. (c) Using the L1 training loss. (d) Using the SSIM training loss.

The proposed MR was initially explored and tested using a publicly available dataset of 59 retinography-FA pairs. Later, a more comprehensive analysis of the methodology was performed using an extended dataset including 59 additional image pairs provided by a local hospital. The additional dataset includes several examples of severe pathological lesions and lower quality images, which allows to test the robustness of the methodology. This comprehensive analysis, for both the multimodal registration and the MR using DNNs, was presented in [56]. Additionally, in this work, the recognition of the retinal microvasculature directly using the predicted FA was evaluated. This is motivated by the fact that FA is, in itself, an enhanced representation of the blood vessels and related lesions. Thus, a satisfactory estimation of FA should provide a significantly improvement in the visualization of the vasculature. This was evaluated using different datasets with pixel-wise retinal vasculature annotations.

During the development of this PhD thesis, several approaches for image-to-image translation have been proposed by different authors. These kinds of approaches have been typically focused on the realism of the generated images, leaving in second place the structural and semantic accuracy of the results. For this reason, it is common the use of Generative Adversarial Networks (GANs) [57, 58], which nowadays represent the go-to approach for realistic image generation [59]. However, GANs also present the risk of hallucinating non-existent structures, which is more likely to happen when the different image patterns in the training dataset are heavily unbalanced [60]. Nevertheless, an important advantage of some GAN-based approaches is that they allow the learning of an image-to-image mapping without the necessity of paired training data [61]. This is key in many application domains with natural images because the paired samples are difficult to obtain. In contrast, in medical imaging, paired image collections are easier to gather due to the common use of complementary modalities in clinical practice, specially for the most complex cases. However, taking complete advantage of the paired data also requires to successfully perform a multimodal registration of the images, which may fail in the most complex scenarios either due to the presence of severe pathologies or low quality images. For these reasons, we also explored the used of unpaired GAN-based approaches for the MR of retinal images.

Regarding the use of unpaired GAN-based approaches for the MR of retinal images, we presented a complete study comparing both paired and unpaired methodologies in [62, 63]. In this case, for the unpaired methodology, we adopted the well-known CycleGAN [61] approach. In brief, CycleGAN compensates the lack of paired data by simultaneously learning two different transformations. The one from input to target image modality as well as the inverse mapping. This allows to introduce a cycle-consistency loss that aims at ensuring the structural and semantic coherence between input and generated images. At the same time, the modality-specific aspect of the images is enforced by the adversarial networks of the GAN framework. The comparison of the paired and unpaired approaches is performed by evaluating the reconstruction fidelity and quality of the generated images. Figure 1.7 depicts some representative examples of FA images generated with each approach.

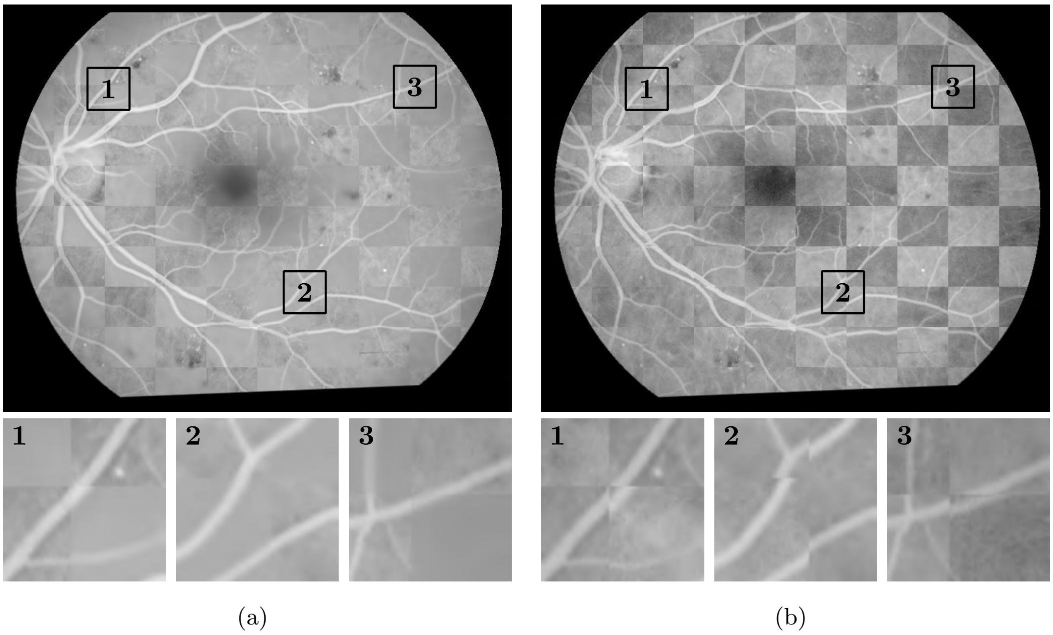
The unpaired alternative produces more realistic generated samples, mainly due to the background texture in the images. However, it also improves the recognition of some small lesions in the images. These outcomes are consequence of the GAN framework. However, there are some structural inaccuracies between the input and generated images. These can be easily seen in the displacement of the blood vessels (Figure 1.8)). Thus, although the use of a GAN framework may provide some advantages, the cycle-consistency loss that enables the use of unpaired data is not enough to ensure the structural and semantic consistency between input and output.



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

(e) (f) (g) (h)

Figure 1.7: Examples of generated FAs together with the corresponding original retinography and the real FA from the test set. ((a),(e)) Original retinography. ((b),(f)) Generated FA using paired training data. ((c),(g)) Generated FA using unpaired training data. ((d),(h)) Real FA.



(a) (b)

Figure 1.8: Examples of generated FAs against the real FA from the test set. Additionally, cropped regions are depicted in detail for each case. (a) Generated FA using paired data against the real FA. (b) Generated FA using unpaired data against the real FA.

1.3.2 Part II - Retinal Image Understanding

The analysis of the different anatomical structures in the retina plays a prominent role in the diagnosis and follow-up of numerous diseases [64]. For instance, pathological lesions may appear around certain anatomical regions that must be adequately identified to provide a diagnosis. Additionally, some eye disorders directly produce morphological changes in the retinal anatomy. In these cases, it is convenient to detect and characterize the affected retinal structures in order to assess the effects of the disease [10, 9, 11].

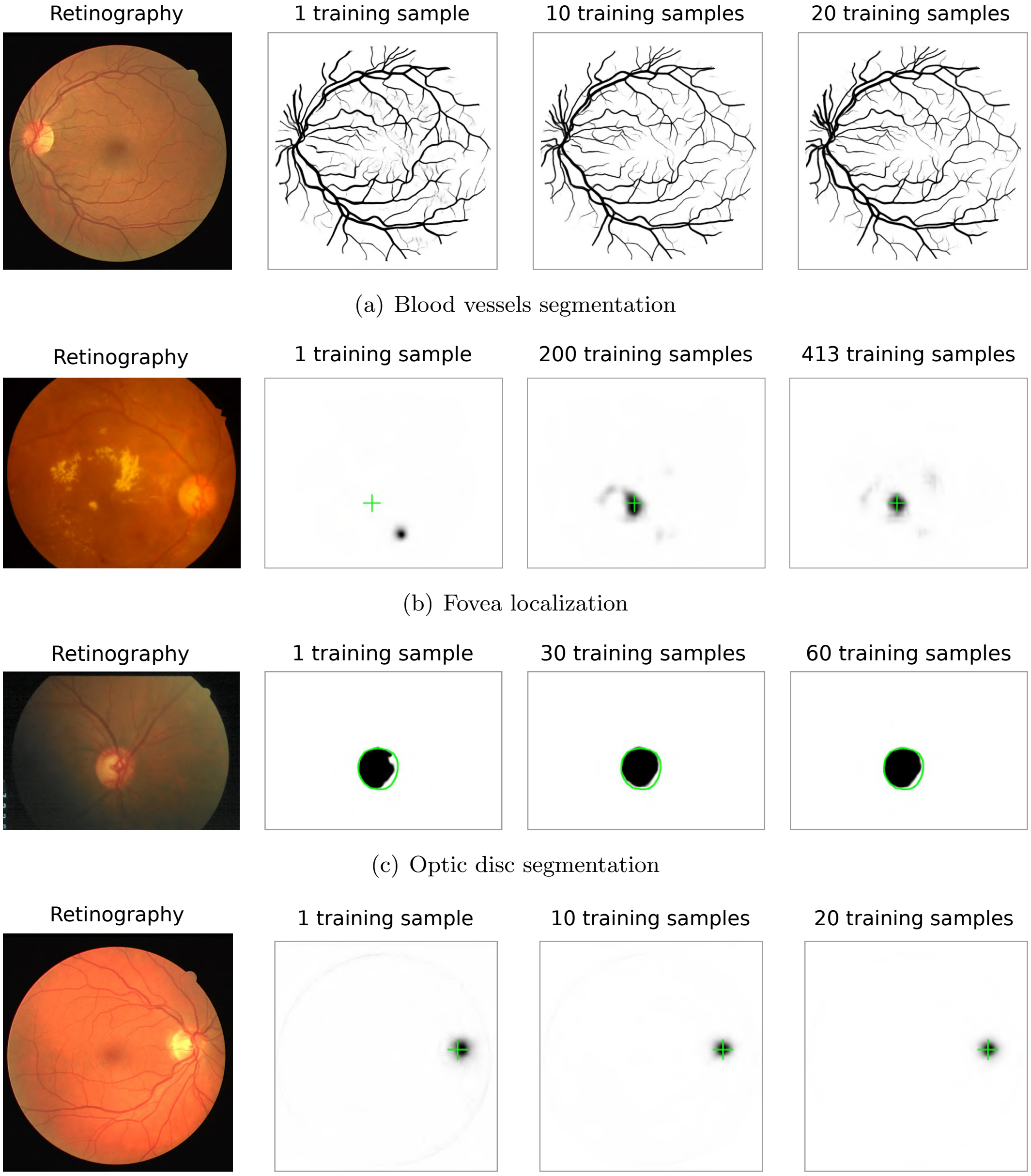
In broad terms, the major anatomical structures in the retina are the microvasculature, the optic disc, and the fovea [8]. The retinal microvasculature is involved in the study of several ophthalmic and systemic diseases. The particular clinical relevance of this anatomical structure is due to the fact that the retina is the only organ of the human body that allows the study of the vascular system in vivo and without invasive procedures [65]. Thus, the analysis of the retinal microvasculature has received a lot of interest in the literature. In this regard, the main task regarding the vasculature is the segmentation of the blood vessels. Nowadays, this task can be easily solve by using modern DNNs. However, the segmentation of the smallest vessels in the images still remains a challenge. Additionally, the manual labeling of the retinal microvasculature is a particularly tedious task due to the high number of small vessels and the low contrast in some regions of the images. Besides the importance of the microvasculature for diagnostic purposes, the retinal vascular tree is also commonly exploited for other applications. For instance, the crossings and bifurcations of the blood vessels in the retina are commonly used as landmarks for image registration algorithms or for retinal verification approaches [52, 53].

Regarding the optic disc, this retinal structure is particularly important for the diagnosis of glaucoma. In fact, a broadly extended biomarker for the assessment of glaucoma, the cup-to-disc ratio, can be solely obtained from the morphological analysis of the optic disc and its inner components [66]. Particularly, the optic disc can be split into two different subregions, the optic cup and the neuroretinal rim. In the literature, numerous works have addressesed the automated segmentation of these two regions, aiming at facilitating the diagnosis of glaucoma via the use of morphological biomarkers [66]. Additionally, the localization or segmentation of the optic disc is also commonly used as an intermediate procedure within more complex pipelines for diagnosis purposes or for the analysis of other retinal structures [67]. Similarly, the localization of the fovea is also commonly used as part of more complex pipelines. In particular, the identification of the foveal (or macular) region is of great interest for the diagnosis of several diseases that lead to the development of different lesions in that area, such as e.g. AMD or diabetic macular edema [10].

In the context of this PhD thesis, the localization and segmentation of the anatomical structures in the retina was used for demonstrating the advantages of the proposed MR for transfer learning purposes [47]. In this regard, to successfully perform the MR, a DNN must learn different low and high level retinal patterns. Thus, using the MR as pre-training task, this domain-specific knowledge can be taken advantage of for different downstream tasks focusing on the analysis of the retinal anatomy. We explored this idea in [47], where the MR was used as self-supervised pre-training task for the segmentation of the blood vessels, detection of the fovea, and segmentation and detection of the the optic disc.

The methodology presented in [47] is based on a U-Net [54] architecture, which is known to provide state-of-the-art performance for segmentation of the blood vessels or the localization of the fovea [68, 69]. All the tasks are trained following the same training procedure, including network architecture, data augmentation, and optimization hyperparameters. The only difference between tasks is the formulation of the training objective and the loss function. The segmentation of the blood vessels and the optic disc are performed following the most standard approach, i.e. a binary pixel-wise classification using cross-entropy as loss function [70]. Regarding the localization of the optic disc and the fovea, there is a greater variety of approaches in the existing literature. In this case, we approach the localization as a distance map regression, where the value of each pixel depends on the distance to the target location [69]. In particular, we compute the corresponding euclidean distances for every pixel and then we apply a hyperbolic tangent function to the obtained values. This approach results in a distance map that is steeper near the target location and flattens out in the farther regions. In order to evaluate the proposed transfer learning approach, we performed experiments using different amounts of labeled training data, ranging from a single training sample to the whole training set. The obtained results demonstrate that the MR pre-training contributes to the different tasks, significantly improving the performance when the annotated data available for training is scarce. Figure 1.9 depicts some representative examples of predictions made using different levels of annotated training data.

Besides the previously mentioned experiments, in [71], we also tested the use of the MR pre-training for the segmentation of the optic disc and optic cup. This is a particular challenging problem due to the ill-defined boundary of the optic cup in the images and the morphological differences between healthy and glaucomatous eyes [66]. In order to address this task, we followed a similar methodology to the one used for the segmentation of blood vessels and optic disc in [47]. The only



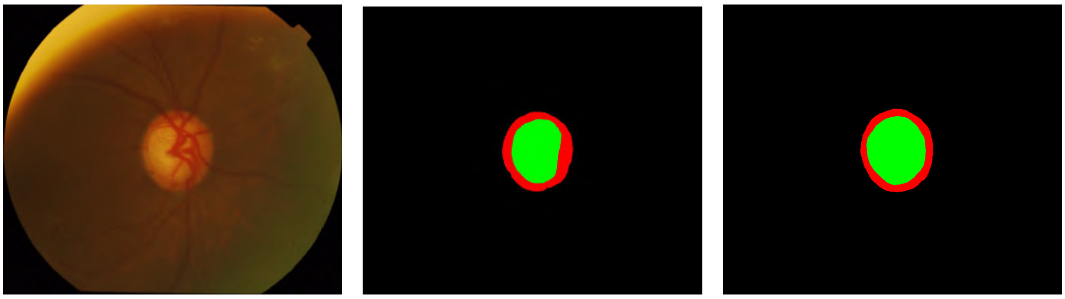
(a) Blood vessels segmentation

(b) Fovea localization

(c) Optic disc segmentation

(d) Optic disc localization

Figure 1.9: Examples of predicted segmentation and location maps for different number of training samples using the proposed MR pre-training. The green crosses and circles depict the ground truth annotations.



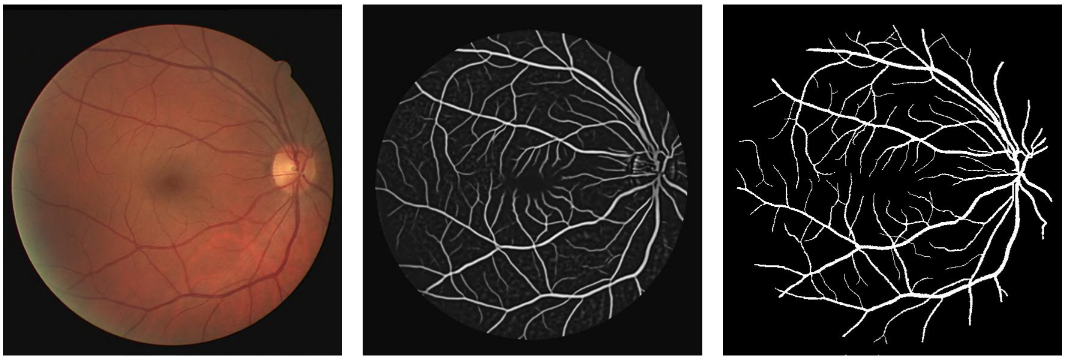
(a) Input image (b) Prediction (c) Ground truth

Figure 1.10: Examples of prediction and corresponding ground truth for optic disc and cup segmentation.

difference is that, in this case, the segmentation is approached as a pixel-wise multi-class classification. In particular, three classes are considered, namely the optic cup, the neuroretinal rim, and the background. Then, the optic disc is defined as the sum of the cup and the rim. In this case, the experimental results also show that the MR pre-training improves the performance of the segmentation task for both optic disc and optic cup. Figure 1.10 depicts some representative examples of predicted segmentations.

Regarding the retinal microvasculature, we also explored novel alternatives for segmenting the blood vessels using DNNs without any annotated data. In this regard, we proposed in [72] a novel approach for self-supervised retinal vessel segmentation that is motivated by two previous developments. First, in [52], we proposed a MSL transformation that significantly enhances the retinal microvasculature for both retinography and FA. In this case, a better vascular map is obtained for the FA because, in this modality, the blood vessels are already highlighted due to the fluorescence of the injected contrast dye. Second, in [46] , we propose the MR approach that generates and estimated FA for any given retinography, hence highlighting the blood vessels in the images. Finnaly, in [72], we combine these two approaches in order to further improve the enhancement of the retinal microvasculature in the images. In particular, the methodology consists in training a DNN in the prediction of the MSL of FA using retinography as input. This way, the network learns to produce a highly enhanced representation of the blood vessels directly from retinography and without using any manually annotated label. In this case, the labels for training are automatically derived from the unlabeled multimodal data. Figure 1.11 depicts some representative examples of blood vessels maps obtained with the approach proposed in [72].

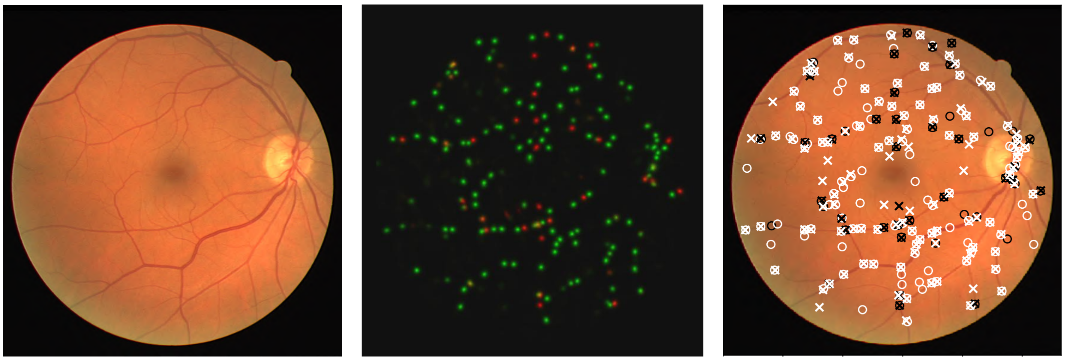
Finally, regarding the analysis of the retinal anatomy, we also explored the de-



(a) Input image (b) Prediction (c) Ground truth

Figure 1.11: Examples of prediction and corresponding ground truth for optic disc and cup segmentation.

tection of the vessel crossings and bifurcations in the eye fundus [73]. In this case, previous approaches in the literature typically relied on extensive ad-hoc processing even when using DNNs. Additionally, it was common to separate the problem into two different tasks, the detection of the vessel junctions and their subsequent classification between crossings and bifurcations [74]. In this context, in [73], we proposed an approach to simultaneously detect and identify the crossings and bifurcations in a single step using DNNs. In particular, the detection task was formulated as a multi-instance heatmap regression where each junction is represented by an individual blob in the predicted heatmap. The precise location of each junctions is given by the point of maximum value within each blob. To provide an adequate heuristic for learning the heatmap regression, the heatmap values are progressively lowered in the pixels surrounding the junction location. Additionally, we explored two different alternatives for generating the target heatmaps, using either a Gaussian or a Radial Hyperbolic Tangent (Radial Tanh) convolutional kernel. The differentiation between crossings and bifurcations is made by simultaneously predicting two independent heatmaps, one for each type of junction. The experimental results show that both Gaussian and Radial Tanh provide similar results when the scale of the kernel is adequately adjusted. However, the Radial Tanh alternative is more robust to these changes, providing a more stable performance. Additionally, the proposed approach significantly outperforms previous methods for both the detection and identification of the vessel crossings and bifurcations. Figure 1.12 depicts some representative examples of generated multi-instance heatmaps and detected crossings and bifurcations.



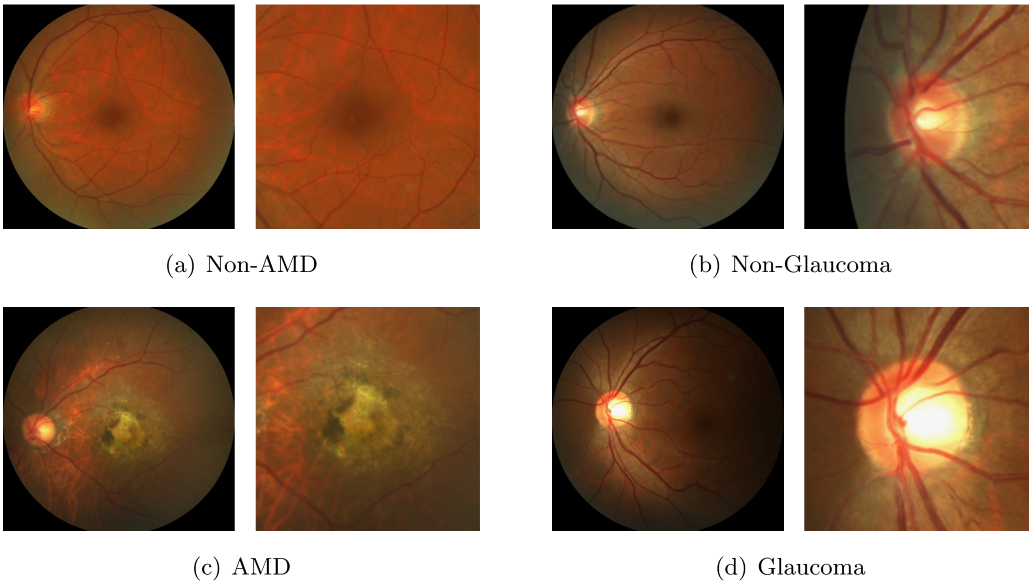
(a) Input image (b) Generated heatmap (c) Detected landmarks

Figure 1.12: Examples of generated heatmap and detected vascular landmarks. (a) Input image. (b) Generated multi-instance heatmap where green denotes bifurcation and red denotes crossing. (c) Detected (white) bifurcations and (black) crossings. Circles denote the ground truth.

1.3.3 Part III - Retinal Computer-Aided Diagnosis

Deep learning represents a fundamental tool for modern CAD systems. In this regard, DNNs have significantly improved the results that could be achieved with traditional approaches for the diagnosis of numerous diseases [75]. For instance, in ophthalmology, deep learning-based approaches have been successfully applied for the diagnosis of AMD, glaucoma, or diabetic retinopathy among other diseases [75]. However, the success of these approaches is strongly linked to the availability or large annotated datasets for training DNNs [18]. In this context, during this PhD thesis, we presented a novel transfer learning approach for retinal CAD systems using the previously proposed MR [76]. Similarly to the previous use of the MR pre-training for the analysis of the retinal anatomy, the idea is to take advantage of the domain-specific knowledge that a DNN acquires from the unlabeled multimodal data during the training of the MR. However, in this case, the final application is the diagnosis of several retinal diseases, i.e. different image classification tasks. This kind of application presents different technical requirements, such as e.g. the network architecture, that make necessary a different transfer learning methodology.

In this PhD thesis, the proposed transfer learning approach for retinal CAD is applied to the diagnosis of AMD and glaucoma. These are two important eye disorders that affect different areas of the retina and lead to significant vision loss if they remain untreated. In particular, AMD is a degenerative eye disorder affecting the macula, which represents the area surrounding the fovea in the retina. This disease is characterized by the presence of different pathological structures or lesions


(a) Non-AMD

(b) Non-Glaucoma

(c) AMD

(d) Glaucoma

Figure 1.13: Examples of retinographies and ROIs used for the diagnosis of ((a),(c)) AMD and ((b),(d)) glaucoma. For each image pair the retinography is in the left and the cropped ROI in the right.

in this area, such as drusen, exudates, or epithelial abnormalities among others. Thus, the diagnosis is typically performed by analyzing the eye fundus looking for these pathological structures [10]. In contrast, glaucoma is characterized by an increased intra-ocular pressure that produces damage to different tissues and retinal structures, such as the optic nerve head. In this regard, glaucoma can be diagnosed by analyzing the eye fundus images looking for morphological changes in the optic disc, such as the reduction of the neuroretinal rim and the increase of the optic cup [9].

The transfer learning methodology for retinal CAD was presented in [76]. The proposed methodology is adapted to each disease by focusing the analysis on the Region Of Interest (ROI) that is required according to the clinical criteria. In particular, a squared ROI is cropped around the fovea and the optic disc for the diagnosis of AMD and glaucoma, respectively. The detection of the fovea and the optic disc is automatically performed following the approach that we previously proposed in [47]. The extracted ROIs are used for the target image classification task as well as the MR pre-training using unlabeled multimodal data. Thus, the pre-training phase is tailored for the study of each disease. Figure 1.13 depicts some representative examples of retinal images and cropped ROIs for the diagnosis of AMD or glaucoma. Similarly, Figure 1.14 depicts some representative examples of

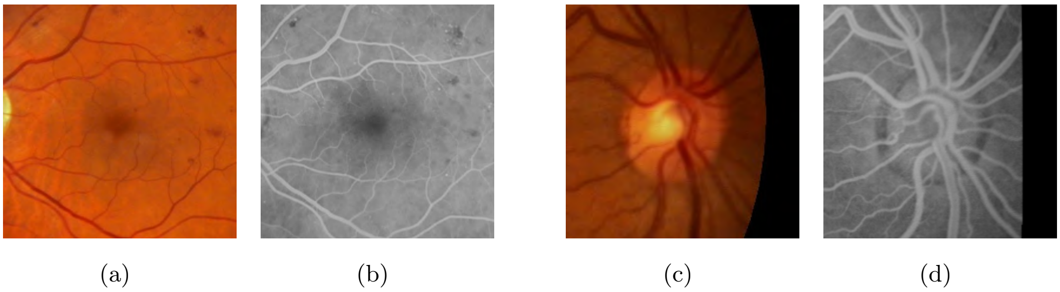


Figure 1.14: Representative examples of multimodal image pairs cropped around the ROI required for each disease. ((a),(b)) AMD: foveal region. ((c),(d)) Glaucoma: optic disc region.

cropped ROIs for a multimodal image pair retinography-FA.

As in [47], the MR pre-training is performed using a U-Net network architecture. However, for image classification, the typical network design mainly consists of a convolutional encoder followed by some fully connected layers to make the final prediction. Thus, in this case, only the encoder of the pre-trained network is reused for the final target tasks. An additional issue that must be considered, regarding the network architecture, is the effect of the skip connections of U-Net in the proposed transfer learning approach. In this regard, although the skip connections facilitate the network training, they also make possible that some relevant information never reach the last layers of the encoder. In the proposed setting, where only the pre-trained network encoder is reused for the target tasks, this could have a detrimental effect in the transfer learning performance. We have studied this issue in [76]. The obtained results show that, in some cases, the use of all the skip connections may compromise the performance of the final target task. However, removing all the skip connections also presents a detrimental effect due to the difficulty for successfully performing the MR pre-training. Thus, the most robust results are achieved following an intermediate approach. Finally, the proposed methodology was validated by comparing its performance against training the network from scratch and using an ImageNet pre-training. The results show that the proposal has a positive impact in the performance of the different tasks in the context of retinal CAD.

1.4 General Conclusions

The analysis of eye fundus images, such as color retinography, is a key step in the prevention, diagnosis, and follow-up of numerous eye disorders. In recent years, there is an increasing interest in the development of automated tools for the analysis of these images. These automated tools help the clinicians in providing more reliable diagnoses and facilitate the conduction of preventive healthcare programs.

In this PhD thesis, we have presented several methodological developments for improving the automated analysis of eye fundus images using deep learning techniques. DNNs have demonstrated to provide remarkable performance in numerous vision problems and represent the current go-to approach for the automated analysis of medical images. In this context, the lack of annotated training data represents one of the main limitations for the successful application of deep learning-based approaches in medical imaging. Considering this, we have proposed a novel paradigm for training DNNs in a self-supervised fashion using unlabeled multimodal visual data. This proposal takes advantage of multimodal image pairs that are commonly available in ophthalmology. The presented approach allows the prediction of FA images from color retinography and can be used as pre-training for any downstream target task performed on retinography.

In order to take advantage of the available multimodal paired data, first we developed a novel methodology for the multimodal registration of retinal images. In particular, we presented a hybrid approach consisting of both landmark-based and intensity-based registration steps. This methodology allows the construction of multimodal datasets with paired and aligned image pairs, which are later used for the training of DNNs. Then, the methodology proposed for the MR allows the prediction of FA images from color retinography and can be used as pre-training for any downstream target task performed on retinography. Additionally, we also explored the use of unpaired multimodal data for performing the MR. Our experiments demonstrated that the use of paired and aligned data remains advantageous.

Taking into consideration the previous results, we explored the use of the MR as pre-training for different pixel-wise and image-wise prediction tasks performed on color retinography. First, we addressed the segmentation and localization of different anatomical structures in the retina, which are a common initial step in numerous retinal image analysis procedures. In particular, our experiments were focused on the retinal microvasculature, the fovea, and the optic disc, which represent the main anatomical structures or regions in the eye fundus images. This experimentation shows that the proposed transfer learning approach reduces the amount of annotated data that is required to achieve satisfactory results in all the tasks. This a strong result that indicates that the proposed approach can facilitate the application of deep learning algorithms for new problems with limited annotated data. Additionally, the same transfer learning approach has also demonstrated to be advantageous for the segmentation of the optic disc and optic cup. This represents a particularly challenging task that is useful for the diagnosis of glaucoma.

With regards to the use of the MR as pre-training for image-wise prediction tasks, such as e.g. image classification, we proposed a transfer learning methodology for retinal CAD systems. In particular, we addressed the diagnosis of two important eye disorders such as AMD and glaucoma. The diagnosis of these two diseases requires very different types of analyses, hence providing complementary scenarios for a robust evaluation of our proposal. The results show that the proposed transfer learning approach using unlabeled multimodal image pairs is advantageous for the diagnosis of these diseases. Additionally, overall, it provides a more robust performance than other alternatives such as fully-supervised pre-training on the ImageNet dataset.

In order to provide a more complete understanding of the eye fundus, we also addressed the detection and identification of the vessel crossings and bifurcations. In this case, we proposed a novel methodology that allows to further take advantage of DNNs for the detection and identification of the vessel landmarks. In this regard, besides significantly outperforming previous approaches, our proposal provides a more straightforward procedure that avoids any ad-hoc processing of the data.

Finally, regarding the retinal anatomy and, particularly, the retinal microvasculature, we also proposed a novel approach for the segmentation of the blood vessels using automatically generated labels. This approach takes advantage of other developments made during this PhD thesis as well as the availability of unlabeled retinography-FA image pairs.

In summary, in this PhD thesis, we have proposed different methodologies to perform a complete analysis of the eye fundus and reduce the necessity of large-scale annotated datasets for training DNNs. In this regard, given the success of the proposed transfer learning approaches using unlabeled multimodal data, in future works we consider to extend this idea to additional applications. For instance, it would be interesting to explore these kind of multimodal self-supervised techniques for the detection and characterization of different lesions or the diagnosis of other particularly challenging diseases such as diabetic retinopathy. These works may be accompanied by additional technical developments to further improve the proposed paradigm. Additionally, we also consider to extend the proposed paradigm to other medical areas where multimodal imaging is common. In this case, it could also be possible to take advantage of the 3D visual data that are common in other medical areas. Another future research direction that we consider is to explore different transfer learning paradigms, e.g. applying multi-task learning. In contrast to the pre-training and fine-tuning approach, multi-task learning allows both tasks to exploit the annotated data available for each other. Thus, in this case, the necessity of large-scale annotated datasets could be reduced by combining different supervised tasks with complementary objectives.

1.5 Structure of the Thesis

This dissertation is structured by chapters and according to the following sequence. Chapter B presents a brief introduction to the PhD thesis. First, this chapter provides the motivation and context for the research work herein described. Then, the main objectives of the PhD thesis are clearly described. Finally, a brief discussion about the research work in this PhD thesis is provided. This discussion aims at providing consistency and coherence among the different publications that compose this dissertation. Chapter 2 includes the detailed description of the methodologies and experimentation for the MR of retinal images using unlabeled multimodal data. Chapter 3 includes the detailed description of the methodologies and experimentation for the analysis of the anatomical structures in the retina. Chapter 4 presents the proposed transfer learning methodology for retinal CAD systems, including the experimentation and analysis of the obtained results.

Chapter 2

Multimodal Reconstruction of Retinal Images - Published Papers

2.1 Conference Paper: Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement

Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, Campus de Elviña s/n, A Coruña 15071, Spain
[2] Department of Computer Science, University of A Coruña, Campus de Elviña s/n, A Coruña 15071, Spain

Conference: 22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES 2018
Location: Belgrade - Serbia
Date: September 03-05, 2018
Pages: 97–104
Editorial: Elsevier
ISSN: 1877-0509

This is a post-peer-review, pre-copyedit version of an article published in Procedia Computer Science, volume 126 (2018), 97–104. The final authenticated version is available online at: https://doi.org/10.1016/j.procs.2018.07.213.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

ELSEVIER

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia

Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Marcos Ortega[a,b]

[a] *CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña 15071, Spain*
[b] *Department of Computer Science, University of A Coruña, A Coruña 15071, Spain*

Abstract

The analysis of different image modalities is frequently performed in ophthalmology as they provide complementary information for the diagnosis and follow-up of relevant diseases, like hypertension or diabetes. This work presents an hybrid method for the multimodal registration of color fundus retinography and fluorescein angiography. The proposed method combines a feature-based approach, using domain-specific landmarks, with an intensity-based approach that employs a domain-adapted similarity metric. The methodology was tested on a dataset of 59 image pairs containing both healthy and pathological cases. The results show a satisfactory performance of the proposed combined approach in the multimodal scenario, improving the registration accuracy achieved by the feature-based and the intensity-based methods.

Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

Multimodal medical image registration is important in the context of diagnosis and follow-up of many relevant diseases. An accurate multimodal registration allows the integration of information obtained from different image modalities, providing complementary information to the clinicians, and improving the diagnostic capabilities. Ophthalmology benefits from this fact given the significant number of existing retinal image modalities: color fundus retinography, fluorescein angiography, autofluorescence fundus retinography or red-free fundus retinography, among others. These modalities offer different visualizations of the eye fundus anatomical structures, lesions and pathologies, without the possibility of achieving the combined multimodal information using only one of the modalities.

In general, registration algorithms can be classified in two groups: feature-based registration (FBR) and intensity-based registration (IBR)[1]. FBR methods use interest points, such as landmarks, along with the intensity profiles at their neighbourhoods to find point correspondences and estimate the spatial transformation between the images. For the detection of interest points, common algorithms as Harris corner detector[2], SIFT[3 4], SURF[5] as well as variations

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: a.suarezh@udc.es

of them[6 7] have been used in different proposals for retinal images. These algorithms detect a large number of interest points in the images. However, as the detected points are not necessarily representative characteristics of the retinal images contents, many of them may not be present across the different modalities. An excessive number of non-representative detected points increases the computational cost of the posterior point matching and increases the likelihood of matching wrong correspondences. The application of generic descriptors is also limited by the differences among retinal image modalities, requiring a preprocessing for its use in multimodal scenarios[7]. Some proposals solve this issue with the design of domain-specific descriptors[8 2] but they still rely on non-specific methods for the detection of interest points. The use of algorithms aimed to the detection of common retinal structures can provide more representative and repeatable characteristics. The detection of line structures, mainly from vessels and disease boundaries, may be seen as a first approach to detect representative characteristics of the retinal images[9]. However, those boundaries do not show clear correspondence among all the modalities. More representative characteristics can be obtained with the extraction of natural landmarks, such as vessel bifurcations. This idea was tested by Laliberté et al.[10] for the registration of retinal images, although their method, that also requires the detection of the optic disk, was not robust enough and failed in several images. The use of these natural landmarks was not explored in posterior works to our knowledge, even though its successful application can greatly reduce the number of detected candidate points for the matching process.

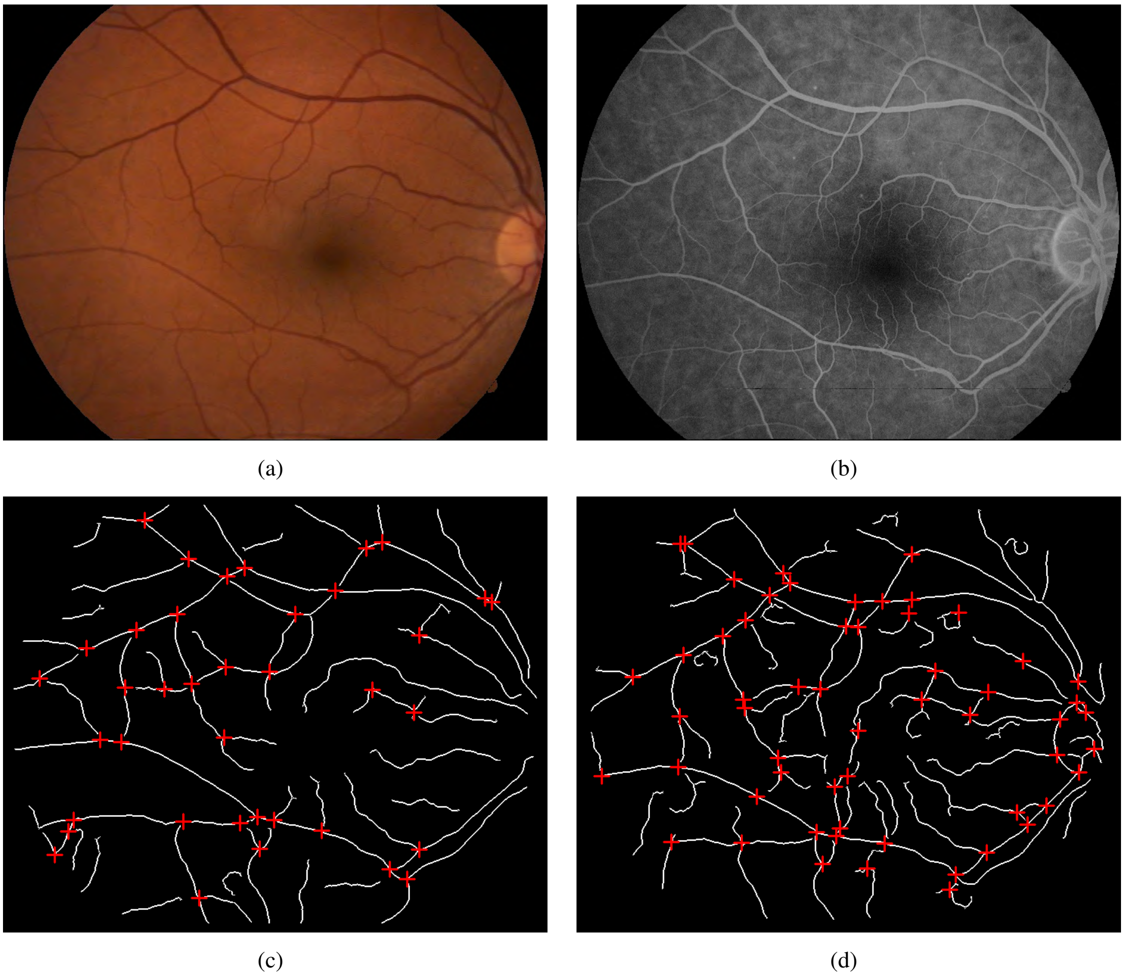
IBR methods use similarity metrics that take into account the intensity values of the whole images instead of sparse local neighbourhoods. This allows to perform the registration with high order transformations[1], as it prevents the risk of overfitting to a small number of points. The registration is performed by optimizing a similarity measure, as intensity differences or cross-correlation for monomodal cases, or mutual information (MI) for multimodal cases. Nevertheless, the application in multimodal scenarios depends on the complexity of the image modalities and the relation between their intensity distributions. Specifically for retinal images, Legg et al.[11] found that in some cases there is an inconsistency between the MI value and the accuracy of the registration, existing transformations with better MI scores that the ground truth registration. These difficulties may explain the reduced number of IBR proposals for multimodal retinal image registration. Other use of the IBR approach may be in combination with FBR methods, being combined in hybrid methodologies that try to exploit the capabilities of both strategies[12 13].

In this work, we propose an hybrid methodology for the multimodal registration of color fundus retinographies and fluorescein angiographies. The method combines a initial FBR approach driven by domain-specific landmarks, with a IBR refinement that uses a domain-adapted similarity metric. Both approaches exploit the presence of the retinal vascular tree in the retinographies and the angiographies. The proposed FBR is based on the detection of landmarks present in both the retinal image modalities, i.e. vessel bifurcations and crossovers. These landmarks can be detected with high specificity, which greatly reduces the number of detected points and facilitates the subsequent point matching. We completely avoid the descriptor computational step, as the point matching is performed with only the geometric information already obtained from the detection algorithm. The latter IBR aims to refine the registration through the estimation of a high order transformation. To perform the IBR over the multimodal images, a domain adapted similarity metric is used. This adaptation consist in the enhancement of vessel regions that transform retinography and angiography to a common image space where the similarity metrics from the monomodal scenarios can be employed. Experiments are conducted to evaluate the performance of the hybrid approach and the improvement over the independent application of the FBR and IBR methods.

2. Methodology

2.1. Landmarks-based Registration

The retinal vascular tree is a complex network of arteries and veins that branch and intersect frequently. The intersection points of the blood vessel segments are natural characteristic points of in the retina and have proven to be a reliable biometric pattern[14]. These intersection points, consisting of vessel bifurcations and crossovers, are used as landmarks. The detection and matching of these domain-specific landmarks is performed following an approach proposed for retinal biometric authentication[14]. The original method was applied in a monomodal scenario with optic disc centered images to compute the similarity between different retinographies. The multimodal registration shows the reverse problem, as it is known that both images are from the same individual and the similarity between them



(a)

(b)

(c)

(d)

Fig. 1. Example of multimodal image pair and result of the landmarks detection method: a) retinography; b) angiography; c,d) binary vessel tree and detected landmarks for the retinography(c) and the angiography(d).

must be maximized. This implies that a higher accuracy in the localization of the landmarks is needed. The mentioned method is adapted to detect landmarks in both retinography and angiography with specific modality modifications.

Retinal images can be seen as landscapes where vessels appear as creases (ridges and valleys). In retinographies, the vessels are valleys in the landscape while in angiographies they are ridges. Defining the images as level functions, valleys (or ridges) are formed in the points of negative minima (or positive maxima) curvature. The local curvature minima and maxima are detected using the MLSEC-ST operator[15]. The vessel tree is given by the set of valleys (or ridges) for retinography (or angiography). The result is a binary image for each modality, consisting of 1 pixel width vessel segments,

The obtained vessel tree is fragmented at some points. Discontinuities appear at crossovers and bifurcations where vessels with different directions meet, and in the middle of a single vessel due to illumination and contrast variations of the image. Bifurcation and crossover detections are approached by joining the segmented vessels as described in[16]. Bifurcations are established where an extended segment under a given maximum distance intersects another segment. Crossovers, instead, are considered as double bifurcations. They are detected at positions where two bifurcations are closer than a given distance and the relative angle between their directions is below a certain threshold. Fig. 4 shows an example of retinography/angiography pair and the result of the vessel tree and landmarks obtained.

This detection method results in a low number of suitable detected points and it allows to immediately perform the transformation estimation without an additional computation of descriptors. Bifurcations and crossovers are used to estimate the transformation between image pairs with a RANSAC point matching algorithm. The applied transformation is a restricted form of affine transformation that only considers translation, rotation and isotropic scaling. Therefore, the transformation has 4 degrees of freedom and can be computed with only two pairs of matched points.

For each previously detected bifurcation or crossover, the position and vessel orientation are known as they are directly obtained from the detection method. These two characteristics are enough to perform the registration, without the need of an specific descriptor computation stage. The high specificity of the detection method leads to a low number of detected landmarks per image. Thus, it is practical to consider all possible matching pairs. The number of possibilities is additionally reduced by taking into account a maximum and minimum scaling factor, which can be is computed in advance as the ratio between the distance of two points in an image and the distance of any other two points in the other image. Relative angles between points, derived from the vessel orientation, are also used for additional restrictions[14].

2.2. Intensity-based Registration

The registration accuracy of the proposed FBR method is limited by the complexity of the transformation considered and the landmark localization precision. A refinement stage that considers high order transformations is proposed to improve the registration accuracy. In order to estimate higher order transformations is convenient to use an IBR approach considering all the pixels of the image pairs. A new domain-adapted similarity metric is constructed combining a vessel enhancement preprocessing with the normalized cross-correlation (NCC). The vessel enhancement transforms images from both modalities to a common image space where the NCC can be successfully employed. This whole operation is named as VE-NCC.

The vessel enhancement is motivated by the fact that the vessels are present in both the retinography and the angiography in form of tubular regions of low or high intensity values, respectively. These vessels vary in thickness throughout the image and can appear in any direction. This motivates the use of a multiscale analysis. A scale-space is defined as $I(x, y; t) = I(x, y) * G(x, y; t)$ where t is the scale parameter and G is a gaussian kernel[17]. The enhancement of the vessel regions is performed using the Laplace operator ∇^2. The Laplacian image, $\nabla^2 I$, will have a high response at nearby positions of the image edges, like those at the vessel boundaries. The distance from the Laplacian peaks to the edges depends on the scale used to compute $\nabla^2 I$. The vessel centerlines achieve maximum response at the scales where the peaks from both vessel boundaries concur. The scale parameter t, therefore, allows to control the scale of the vessels to enhance. The normalized Laplacian scale-space is defined as:

$$L(x, y; t) = t^2 \nabla^2 I(x, y; t) \tag{1}$$

Where t^2 is a normalization factor. A property of scale-space representation is that the amplitude of spatial derivatives decreases with the scale[17]. The normalization factor allows the comparison and combination of the magnitude at different scales. Finally, the maximum value across scales for every point is computed as:

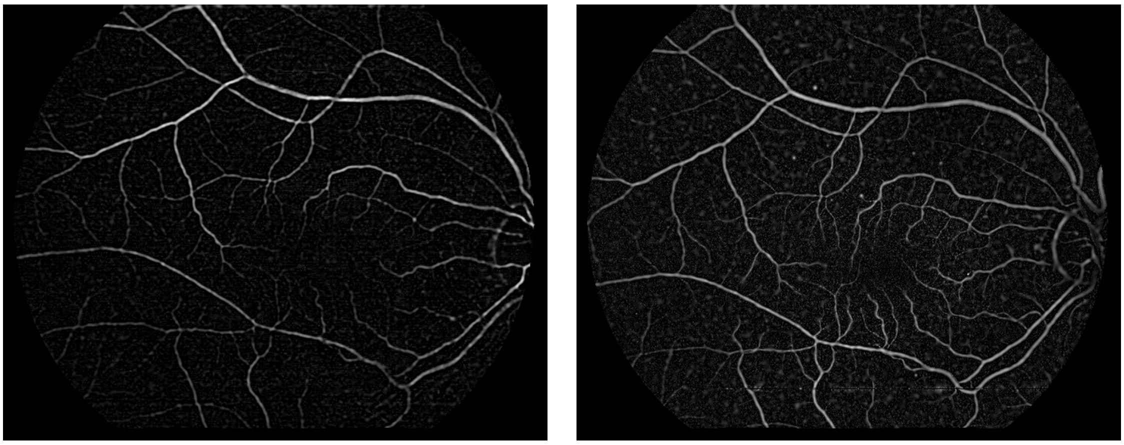
$$L(x, y) = max_{t \in S} \lceil mL(x, y; t) \rceil_{\emptyset} \tag{2}$$

where $m = 1$ for angiography and $m = -1$ for retinography, and $\lceil \cdot \rceil_{\emptyset}$ denotes halfwave rectification. The rectification is used to avoid the negative Laplacian peaks outside the vessel regions, so that only the vessel interiors are represented in the enhanced images. This results in a common representation for retinography and angiography, with enhanced vessel regions and the same intensity level pattern. Fig. 2 shows the result of the vessel enhancement operation applied to the retinography/angiography pair from Fig. 4.

The transformation between images is obtained through minimization of the negative VE-NCC. It is important to initialize the algorithm with a proper initial transformation. The estimated transformation from the FBR serves as initialization for the IBR. Two different transformation models are considered to perform the IBR: Affine Transformation (AT) and Free Form Deformation(FFD). AT allows translation, rotation, anisotropic scaling and shearing, having 6 degrees of freedom. Differently, FFD uses a grid of control points that are moved individually along the image to define a high order transformation.

3. Experiments and Results

For the evaluation of the proposed methodology, we used the publicly available Isfahan MISP dataset of retinography and angiography images of diabetic patients[18]. This dataset consists of 59 image pairs divided in two collections



(a)

(b)

Fig. 2. Exampled of the vessel enhancement operation applied to a multimodal image pair: a) retinography; b) angiography.

Table 1. Average and standard deviation VE-NCC for the different configurations tested.

FBR	IBR-AT	IBR-FFD	VE-NCC	
			Healthy cases	Phatological cases
•	•	•	0.6123 ± 0.0815	0.4758 ± 0.1419
•	•		0.5980 ± 0.0865	0.4661 ± 0.1406
•		•	0.5668 ± 0.0828	0.4401 ± 0.1381
•			0.5266 ± 0.0928	0.3961 ± 0.1416
	•		0.0673 ± 0.0500	0.0930 ± 0.1065
	•	•	0.0733 ± 0.0627	0.1005 ± 0.1250
		•	0.0581 ± 0.0323	0.0656 ± 0.0497
			0.0481 ± 0.0159	0.0518 ± 0.0220

of healthy and pathological cases. The pathological cases correspond to patients with mild and moderate retinal diseases due to diabetic retinophaty. The images have a resolution of 720 × 576 pixels. The division of the dataset in healthy and pathological cases allows to analyze the effect of the pathologies in the registration performance.

Several experiments were conducted to evaluate the hybrid methodology as well as the performance of the FBR and IBR methods. Regarding the IBR method, both the affine transformation (IBR-AT) and the free form deformation (IBR-FFD) variations were applied. We propose the hybrid method formed by the sequential application of FBR, IBR-AT and IBR-FFD, and alternative variations over this by removing one or two steps at a time. This results in 7 different methods as reported in Table 1. The table shows the average and standard deviation VE-NCC for each method in healthy and pathological cases. Figure. 3 depicts the cumulative distribution of the VE-NCC values. The best result are achieved by the proposed hybrid method. There is a large difference between the experiments that perform the initial FBR and the ones that directly apply IBR. For the latter experiments the registration failed in most cases. Most of the image pairs do not significantly change their VE-NCC values by applying IBR alone, and only a few of them obtained values over the minimum that was achieved by the initial FBR. These results indicate that, with the use of IBR and high order transformations, more accurate registrations can be achieved. However, they also evidence the importance of a proper initialization for the convergence of the optimization algorithm, which is provided by the initial FBR. Moreover, the IBR-FFD also benefits from the previous IBF-AT, as the order of the applied transformation directly fixes the search space dimensionality, increasing the complexity of the optimization. Figure 4 exposes some representative examples of the images registered with the proposed hybrid method. Both the raw images and the vessel enhanced images provide qualitative evidence of a satisfactory multimodal registration with the hybrid approach in healthy and pathological scenarios.

Additionally, we performed a more in-depth analysis of the effect of the different steps in the proposed hybrid configuration. Fig. 5 shows scatter plots of the VE-NCC values before and after each step of the hybrid method for both healthy and pathological cases. It is observed that the biggest contribution comes from the initial FBR. The improvement decreases with each step as minor adjustments in the estimated transformation are required. The

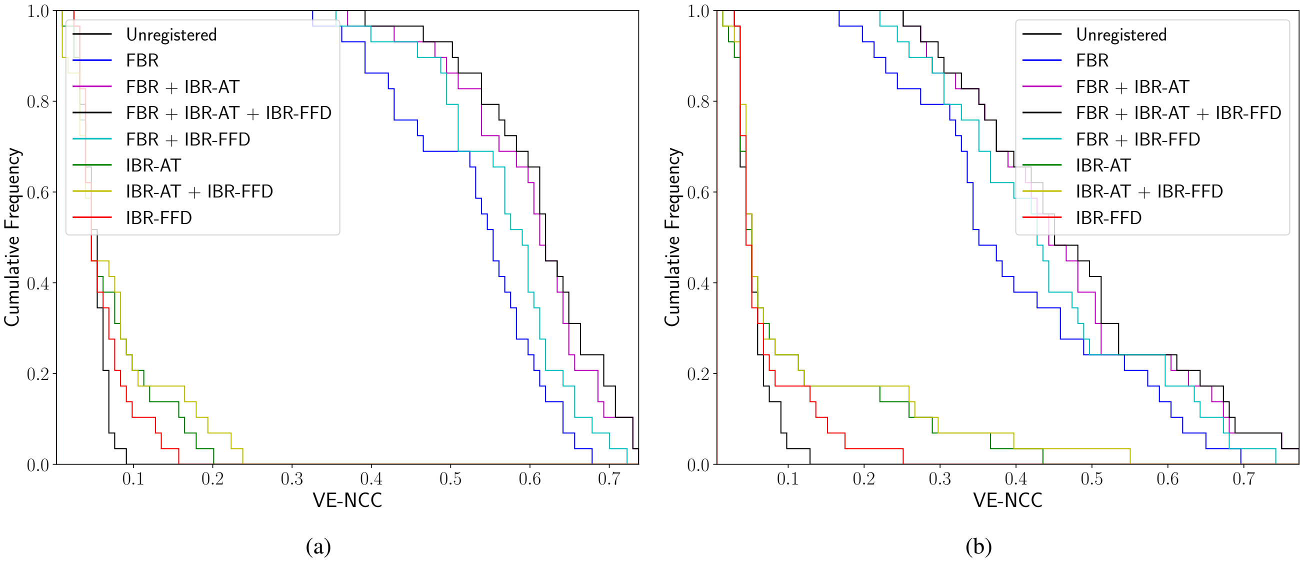


Fig. 3. Cummulative distribution of the VE-NCC: a) healthy cases: b) pathological cases.

presence of pathologies in the images does not affect the general behaviour of the proposed hybrid method, as similar conclusions can be drawn from sets of both scatter plots. However, the average VE-NCC values are slightly lower for the pathological cases, at the same time that the variance is slightly higher. This may be an indication of the slightly influence of the pathological structures in the VE-NCC. The maximum value is not necessarily the same for every image pair, although this does not affect the optimal transformation and a successful registration or the retinography/angiography pairs is achieved.

4. Conclusions

The joint analysis of color fundus retinography and fluorescein angiography usually requires the registration of the images. In this work, an hybrid method for the multimodal registration of pairs of retinographies an angiographies is presented. Domain-specific solutions, exploiting the presence of the retinal vasculature in both image modalities, were proposed for both the feature and intensity-based registration steps that constitute the hybrid proposal. The use of a domain-adapted similarity metric allows the estimation of high order transformations that increase the accuracy of the registration. Simultaneously, accurate registration is only feasible departing from the initial registration with domain-specific landmarks. Different experiments were conducted to validate the suitability of the proposed method and to evaluate the contribution of each registration steps. The results demonstrated that the hybrid method outperforms the individual application of each of its constituting approach.

Acknowledgments

This work is supported by I.S. Carlos III, Government of Spain, and the ERDF of the EU through the DTS15/00153 research project, and by the MINECO, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the ERDF and ESF of the EU, and the Xunta de Galicia through Centro Singular de Investigacin de Galicia, accreditation 2016-2019, ref. ED431G/01 and Grupo de Referencia Competitiva, ref. ED431C 2016-047 research projects, and the predoctoral grant contract ref. ED481A-2017/328.

References

1. Oliveira, F., Tavares, J.. Medical image registration: A review. *Computer methods in biomechanics and biomedical engineering* 2014; **17**:73–93.

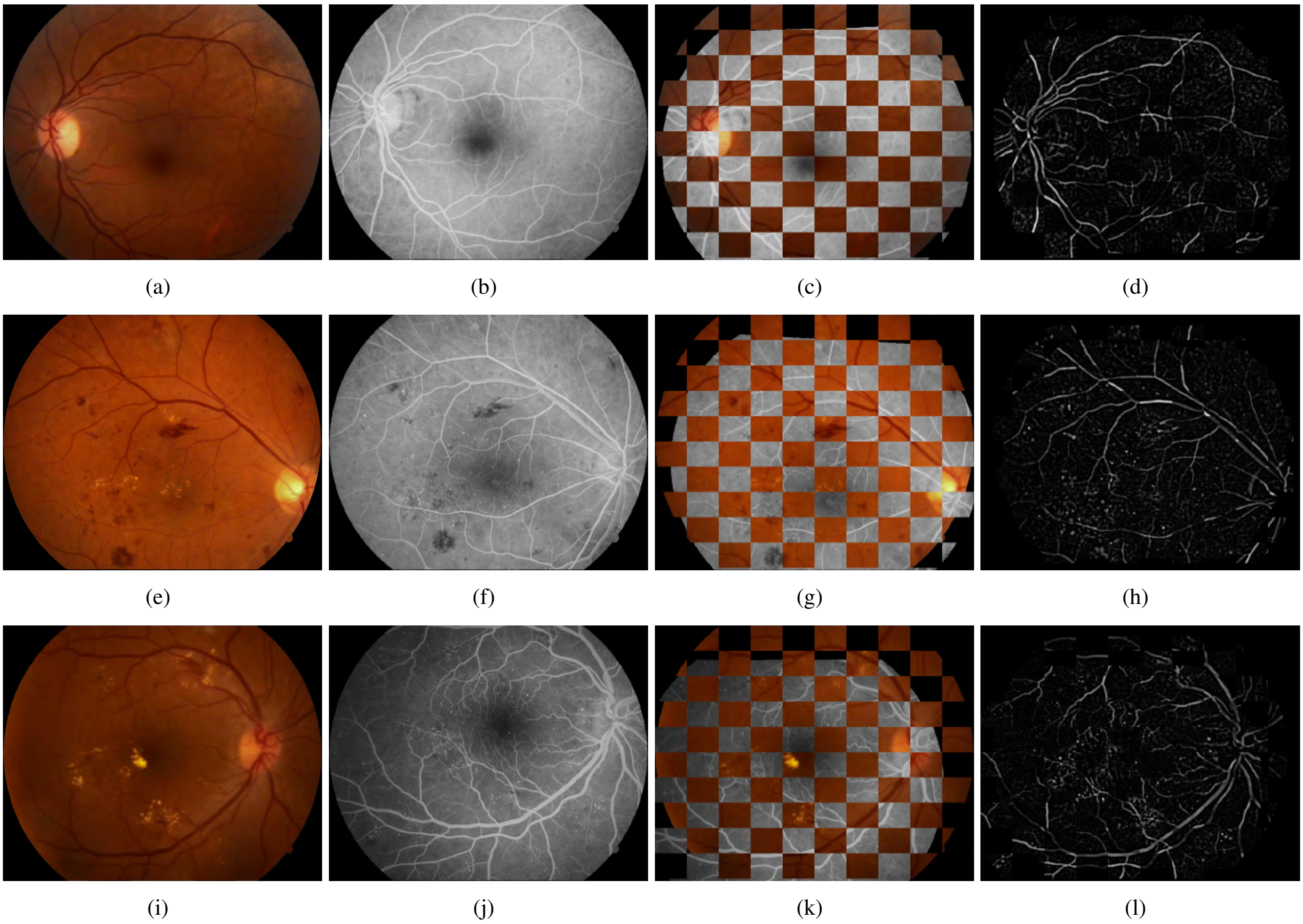


Fig. 4. Examples of the multimodal registration with the hybrid method: a,c,i) retinographies; b,f,j) angiographies; c,g,k) registered image pairs; d,h,l) results of the vessel enhancement operation applied to the registered image pairs.

2. Chen, J., Tian, J., Lee, N., Zheng, J., Smith, R.T., Laine, A.F.. A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engineering* 2010;**57**(7):1707–1718. doi:10.1109/TBME.2010.2042169.
3. Yang, G., Stewart, C.V., Sofka, M., Tsai, C.L.. Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2007;**29**(11):1973–1989. doi:10.1109/TPAMI.2007.1116.
4. Tsai, C.L., Li, C.Y., Yang, G., Lin, K.S.. The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence. *IEEE Transactions on Medical Imaging* 2010;**29**(3):636–649. doi:10.1109/TMI.2009.2030324.
5. Wang, G., Wang, Z., Chen, Y., Zhao, W.. Robust point matching method for multimodal retinal image registration. *Biomedical Signal Processing and Control* 2015;**19**:68–76.
6. Ghassabi, Z., Sedaghat, A., Shanbehzadeh, J., Fatemizadeh, E.. An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors. *EURASIP J Image and Video Processing* 2013;**2013**:25. URL: `https://doi.org/10.1186/1687-5281-2013-25`. doi:10.1186/1687-5281-2013-25.
7. Ma, J., Jiang, J., Chen, J., Liu, C., Li, C.. Multimodal retinal image registration using edge map and feature guided gaussian mixture model. In: *2016 Visual Communications and Image Processing (VCIP)*. 2016, p. 1–4. doi:10.1109/VCIP.2016.7805491.
8. Lee, J.A., Cheng, J., Lee, B.H., Ong, E.P., Xu, G., Wong, D.W.K., et al. A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, p. 1046–1053. doi:10.1109/CVPR.2015.7298707.
9. Hernandez, M., Medioni, G., Hu, Z., Sadda, S.. Multimodal registration of multiple retinal images based on line structures. In: *2015 IEEE Winter Conference on Applications of Computer Vision*. 2015, p. 907–914. doi:10.1109/WACV.2015.125.
10. Laliberte, F., Gagnon, L., Sheng, Y.. Registration and fusion of retinal images-an evaluation study. *IEEE Transactions on Medical Imaging* 2003;**22**(5):661–673. doi:10.1109/TMI.2003.812263.
11. P.A. Legg P.L. Rosin, D.M., Morgan, J.. Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation. *Computerized Medical Imaging and Graphics* 2013;**37**:597–606.
12. Chanwimaluang, T., Fan, G., Fransen, S.R.. Hybrid retinal image registration. *IEEE Transactions on Information Technology in Biomedicine* 2006;**10**(1):129–142. doi:10.1109/TITB.2005.856859.
13. Kolar, R., Harabis, V., Odstrcilik, J.. Hybrid retinal image registration using phase correlation. *The Imaging Science Journal* 2013; **61**:369–384.

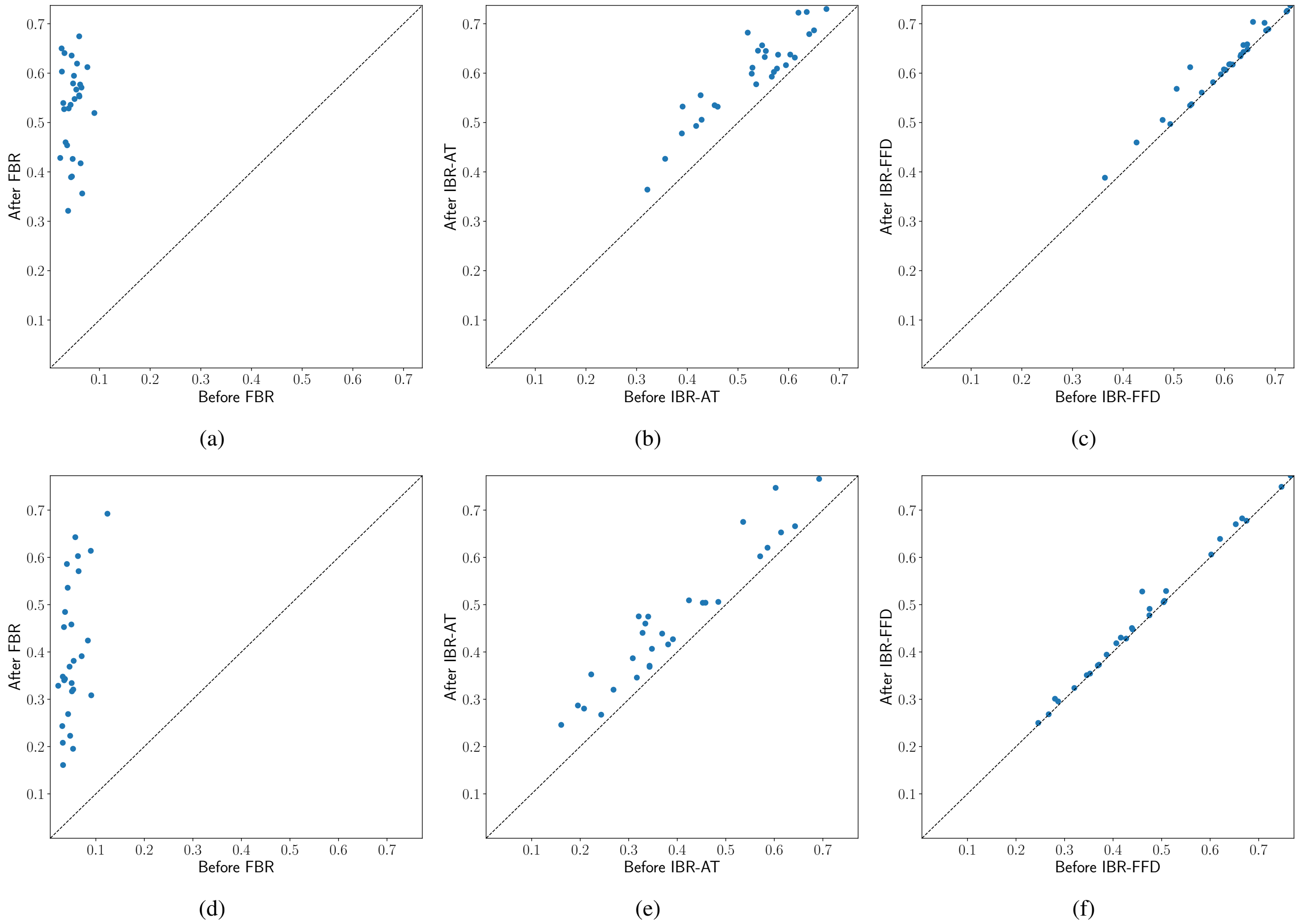


Fig. 5. Step-by-step analysis of the proposed hybrid method: a,b,c) healthy cases: d,e,f) pathological cases.

14. Ortega, M., Penedo, M.G., Rouco, J., Barreira, N., Carreira, M.J.. Retinal verification using a feature points-based biometric pattern. *EURASIP Advances in Signal Processing* 2009;**2009**(1):235746. URL: `https://doi.org/10.1155/2009/235746`. doi:10.1155/2009/235746.
15. López, A.M., Lloret, D., Serrat, J., Villanueva, J.J.. Multilocal creaseness based on the level-set extrinsic curvature. *Computer Vision and Image Understanding* 2000;**77**(2):111–144. URL: `https://doi.org/10.1006/cviu.1999.0812`. doi:10.1006/cviu.1999.0812.
16. Ortega, M., Rouco, J., Novo, J., Penedo, M.G.. Vascular landmark detection in retinal images. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A., editors. *Computer Aided Systems Theory - EUROCAST 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04772-5; 2009, p. 211–217.
17. Lindeberg, T.. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision* 1998;**30**(2):117–156. URL: `https://doi.org/10.1023/A:1008097225773`. doi:10.1023/A:1008097225773.
18. Golmohammadi, H., Kashefpur, M., Kafieh, R., Jorjandi, S., Khodabande, Z., Abbasi, M., et al. Isfahan misp dataset. *Journal of Medical Signals and Sensors* 2017;**7**(1):43–48.

2.2 Conference Paper: Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction

Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Conference: 21st International Conference on Medical Image Computing and Computer Assisted Intervention - MICCAI 2018
Location: Granada - Spain
Date: September 16-20, 2018
Pages: 321–328
Editorial: Springer, Cham
ISBN: 978-3-030-00928-1

This is a post-peer-review, pre-copyedit version of an article published in Lecture Notes in Computer Science, volume 11070 (2018), 321–328. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-00928-1_37.

Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction

Álvaro S. Hervella[1,2]✉, José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)
{a.suarezh, jrouco, jnovo, mortega}@udc.es

Abstract. The successful application of deep learning-based methodologies is conditioned by the availability of sufficient annotated data, which is usually critical in medical applications. This has motivated the proposal of several approaches aiming to complement the training with reconstruction tasks over unlabeled input data, complementary broad labels, augmented datasets or data from other domains. In this work, we explore the use of reconstruction tasks over multiple medical imaging modalities as a more informative self-supervised approach. Experiments are conducted on multimodal reconstruction of retinal angiography from retinography. The results demonstrate that the detection of relevant domain-specific patterns emerges from this self-supervised setting.

Keywords: self-supervised, multimodal, retinography, angiography.

1 Introduction

Nowadays, deep learning-based solutions are commonly used for a significant variety of computer vision applications. Deep Neural Networks (DNN) are able to recognize complex patterns from raw input images and signals, and to hierarchically learn suitable representations of the underlying information at different levels [1]. In order to do so, DNNs train a large number of parameters using also large datasets that include representative annotated data. While large datasets exist for many computer vision applications, they are a scarce resource in clinical environments. The size of annotated medical imaging datasets is usually limited given the significant cost of hand labeling the data. Annotations must be performed by expert clinicians, whose time and expertise is not efficiently used if it is invested in tedious and time-consuming tasks like manual labeling large datasets. Besides, expert-annotated images are better used for the clinical validation of the resulting medical image analysis methods. In contrast, a large amount of unlabeled medical images is readily available from the daily clinical practice, along with the patient clinical condition, which can be used as a broad label for the image. However, detailed marking of the images is still needed to provide relevant information for the detection and classification of lesions and anatomical structures in the images.

Several approaches have been proposed to alleviate the scarcity of annotated data, some of which have been applied to medical imaging. A common approach is the application of transfer learning, which consists in the reuse of trained models from different domains of application. For example, pretrained networks for ImageNet classification are usually employed for this purpose, either using the first layers as feature extractors or using the whole network as initialization [2]. However, this approach is limited by the differences between natural and medical images. Self-supervised reconstruction of unlabeled input images, e.g., with stacked autoencoders, is used for the same initialization and feature extraction purposes [3]. The advantage is that the image domain remains the same, but it is not guaranteed that the reconstruction relies on relevant features for the target application. A possible solution to this is the simultaneous training of the auxiliary task, i.e. the self-supervised reconstruction, along with the target task [4]. Although other multitask learning settings are also possible [5]. For example, the simultaneous learning of several supervised tasks over the same input, some of which may be based on global labels, augments the labeled data improving the performance on all the tasks [6]. This latter approach allows a more efficient use of the labeled data, but may further benefit from auxiliary self-supervised tasks that are relevant for the target application. A different approach consists in artificially increasing the dataset with synthetic images and labels. This data augmentation is usually performed through basic spatial and intensity transformations. However, the use of generative deep learning models has also been explored to create new plausible sample images [7].

In this work, we explore an innovative source of additional self-supervised learning information for medical environments, which has not been previously used to complement scarce datasets. In many medical environments, it is common that the diagnosis and follow-up of a disease involves the use of multiple image modalities. This eases the gathering of multimodal image datasets. Multimodal image reconstruction, from one image modality to another, using aligned images of the same patient, is a self-supervised task that can provide information about the relevant image objects. On the one hand, each modality provides a complementary view of the same real world object, without a trivial reconstruction path between them. Training the reconstruction may give rise to rich representations involving the joint properties of the imaged objects. On the other hand, some modalities may be more informative about some specific anatomical contents through, e.g., the use of injected contrasts. Our idea is to use these invasive modalities as the target output for the reconstruction from a non-invasive alternative. Thus, the contrast can be seen as a pseudo-label, and the trained network can be used as a non-invasive estimator of the invasive modality.

The proposed experiments in the work herein described are focused on two ophthalmological image modalities: retinography and fluorescein angiography. These two modalities offer complementary information about the structures and pathological lesions of the retina. The angiography is an invasive technique as it requires the injection of fluorescein to the patient, limiting its use to cases with clear symptoms or patients that are already diagnosed. The contrast enhances the visualization of the retinal vasculature and makes the angiography a more suitable modality for the diagnosis and follow-up of cardiovascular diseases. In contrast, the retinography is an affordable and non-invasive modality, suitable for screening programs and regular check-ups. The self-supervised multimodal reconstruction of angiography from retinography can be used to extract relevant retinal patterns and produce a non-invasive estimation of the angiography. Experiments performed in this work focus on this self-supervised reconstruction without adding additional tasks or training data. A rough segmentation of the retinal vasculature shows, nonetheless, an important improvement due to the self-supervised training. Both qualitative and quantitative evaluations demonstrate that retinal image understanding emerges from the multimodal reconstruction.

2 Materials and Methods

2.1 Dataset Preparation

The publicly available Isfahan MISP dataset [8] is used. It contains 59 retinography/angiography pairs divided in healthy and pathological cases. The pathological images correspond to patients with diabetic retinopathy. The images have a resolution of 720×576 pixels.

Multimodal Registration. Each of the eye fundus images displays the retina in a different pose. The registration of the retinography-angiography pairs is needed for building a pixel-wise correspondence. The multimodal registration is performed following the methodology proposed in [9]. An initial registration is estimated using domain-specific landmarks that consist of bifurcations and crossovers of the vessel tree, followed by the application of a RANSAC matching algorithm. Afterwards, a refined transformation is computed using an image-domain similarity metric based on a multiscale enhancement of the vessel regions.

Multimodal ROI. Eye fundus photographies display the retina in a circular region of interest (ROI). The multimodal registration aligns the ROI contents of both images but not the ROI shapes that may no completely overlap. Thus, a multimodal ROI is computed as the intersection of the individual ROIs.

2.2 Network Architecture

For the proposed multimodal reconstruction setting, we adapted the U-Net architecture proposed in [10]. The U-Net model is a fully convolutional network that heavily relies on downsampling and upsampling operations to obtain dense predictions. The core of the model is a convolutional autoencoder with a contractive and a expansive part. In the contractive part, spatial max pooling operations are interleaved between convolutional blocks, leading to an internal space with high depth and reduced width. This contraction of the space forces the model

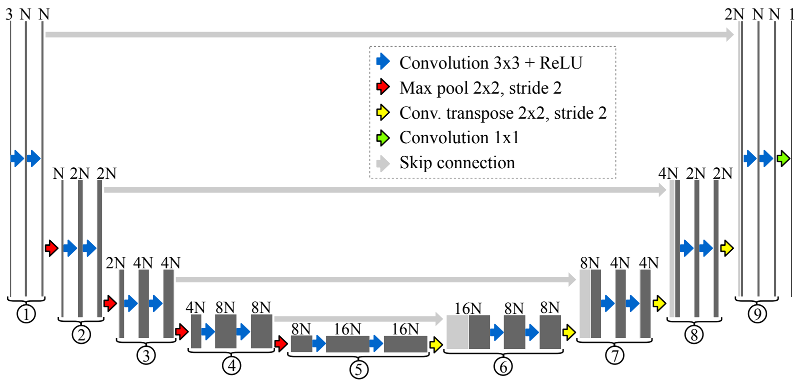
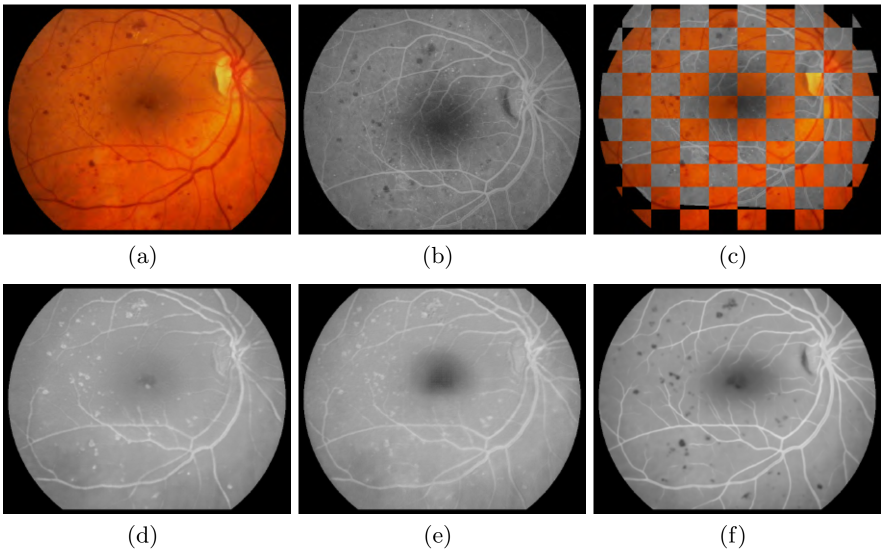


Fig. 1. U-Net architecture, where N is the number of base channels.

to learn high level representations. Conversely, the expansive part has upsampling operations in between convolutional blocks. The expansive decoder allows to generate the output image from the internal space representation. The convolutions are followed by ReLU activations except the last layer that is linear. As the contractive part performs spatial pooling, the precise location of the patterns in the input image is compromised. The U-Net solves this by creating skip connections between layers of the same resolution in the contractive and expansive parts. This allows to bypass the spatial information, improving structural correspondence between the input and output image spaces. Figure 1 shows the U-Net architecture that we used with the default value of $N = 64$ base channels.

Multimodal Reconstruction Loss. Three different metrics are considered for the network loss: L1 norm , L2 norm and Structural Similarity (SSIM) index [11]. L1 and L2 norms are commonly used in deep learning image generation and reconstruction applications. On the contrary, SSIM is commonly used for image quality assesment. It evaluates the structural differences between images comparing local statistics instead of measuring pointwise distances, which leads to a better correlation with the human visual perception [11]. As SSIM measures similarity, the negative SSIM is used as loss. The value of the three losses is computed over the multimodal ROI. The remaining pixels are not considered given they do not contain multimodal information.

Network Training. Network parameters are initialized using the He et al. [12] method, and the optimization is performed using the Adam algorithm [13]. The multimodal dataset was randomly divided into training and validation sets using a 4 to 1 ratio. Early stopping is performed based on the validation loss. The high number of free parameters in the model in relation to the number of samples in the dataset makes it easy for the model to overfit. To resolve this situation we use dropout, with a rate of 0.5, after the convolutional blocks 4 and 5 (see Fig. 1), as well as data augmentation. The applied data augmentation consists of random small elastic and affine transformations over the images. No other preprocessing is applied.


(a) (b) (c)

(d) (e) (f)

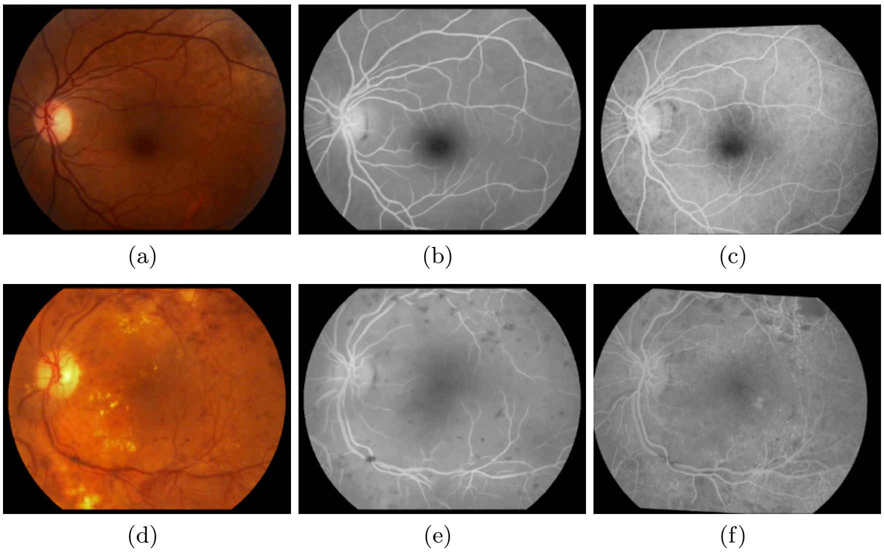
Fig. 2. Examples of generated pseudo-angiography after training with the different losses: a) input retinography; b) corresponding angiography; c) registered images; d,e,f) pseudo-angiography results with L1(d), L2(e) and SSIM(f).

3 Results and Discussion

3.1 Qualitative Evaluation

Figure 2 shows an example of a registered image pair and the generated images with the networks using L1, L2 and SSIM losses. These images are part of the validation set. It is observed that SSIM generates sharper images with a greater presence of thin vessels. L1 and L2 losses offer similar visual appearances and tend to overenhance the vessel borders.

Figure 3 shows more examples from the validation set using the SSIM loss. Each image is accompanied by the original retinography and angiography. The network learned different transformations for the vasculature, fovea, optic disc, pathological structures and retinal background. This provides evidence of an underlying understanding of important retinal patterns. The vasculature is enhanced with respect to the retinographies and more small vessels are present. This visual improvement is also present for vessels with poor visibility as, e.g., Fig. 3(b). Bright color pathologies are absent in the reconstruction, as in the actual angiography (e.g., Fig. 3(e)). Red pathologies are reconstructed with low intensity values, despite that they may have different appearance in classical angiographies (Figs. 2(f) and 3(e)).



(a) (b) (c)

(d) (e) (f)

Fig. 3. Two examples of pseudo-angiography generation after training with SSIM: a,d) retinography; b,e) pseudo-angiography; c,f) registered angiography.

Table 1. Training loss comparison in terms of the validation loss.

Training	Validation loss		
loss	L1	L2	1−SSIM
L1	0.0914	0.0125	0.3411
L2	0.0895	0.0121	0.3310
1−SSIM	0.0856	0.0110	0.2768

3.2 Quantitative Evaluation

Table 1 shows the validation losses after training the network with L1, L2 and SSIM losses. The model trained with SSIM outperforms the others even when the comparison is based on the L1 or L2 losses, indicating that SSIM helps training the network.

In order to quantify the complexity of the transformation achieved by the self-supervised multimodal reconstruction we proposed an additional experiment. As the angiography enhances the vasculature, a rough vessel segmentation can be obtained through plain thresholding with an appropiate threshold value. This is not the case with retinography. Thus, it would be expected that a Receiver Operating Characteristic (ROC) analysis of this segmentation provides a higher Area Under Curve (AUC) for the angiography than for the retinography. We apply this ROC analysis comparison to both the retinography and the estimated pseudo-angiography. This evaluation is performed using the DRIVE image database [14], which consists of 40 retinographies of size 565×584 including

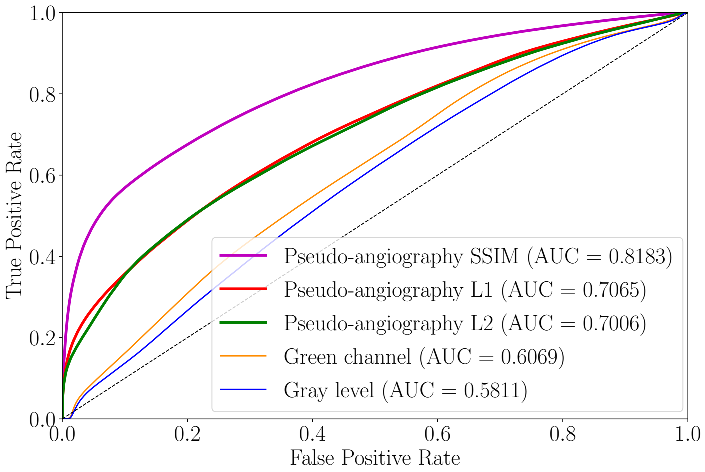


Fig. 4. ROC curves of the quantitative evaluation.

ground truth vasculature segmentations. As the evaluated models are trained in a self-supervised way in the Isfahan dataset, the whole DRIVE dataset is used as test set for this analysis. Figure 4 shows the ROC curves for retinography and pseudo-angiography. Both green channel and grayscale image are compared for retinography, as they are common choices for vessel segmentation. Pseudo-angiography curves correspond to the models trained with L1, L2 and SSIM losses. These results show that the pseudo-angiography provides additional information about the vessel structures and the network is not providing a trivial solution. Thus, the model has learned to recognize relevant patterns in the retina.

4 Conclusions

Motivated by the scarcity of annotated medical imaging datasets and the common availability of multiple imaging modalities, in this work we proposed the use of self-supervised multimodal reconstruction as a more informative alternative to self-supervised reconstruction of the input images. Experiments were performed on retinal angiography reconstruction from aligned retinographies, giving rise to a pseudo-angiography estimator that enhances the vascular structures of the retina. Quantitative and qualitative results indicate that the obtained transformation provides additional understanding of the relevant retinal patterns, noting that it is not a mere intensity mapping. Apart from the potential aplications of the self-supervised task on multitask and transfer learning, the generated pseudo-angiography may have important clinical applications as it simulates the angiography enhancement without the need of the invasive contrast injection.

Acknowledgments This work is supported by I.S. Carlos III, Government of Spain, and the ERDF of the EU through the DTS15/00153 research project, and by the MINECO, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the ERDF and ESF of the EU, and the Xunta de Galicia through Centro Singular de Investigación de Galicia, accreditation 2016-2019, ref. ED431G/01 and Grupo de Referencia Competitiva, ref. ED431C 2016-047 research projects, and the predoctoral grant contract ref. ED481A-2017/328.

References

1. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering (19) (2017) 221–248
2. Karri, S.P.K., Chakraborty, D., Chatterjee, J.: Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomedical Optics Express **8**(2) (2017) 579–592
3. Shin, H.., Orton, M.R., Collins, D.J., Doran, S.J., Leach, M.O.: Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. IEEE Trans. PAMI **35**(8) (2013) 1930–1943
4. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: Advances in Neural Information Processing Systems 28. (December 2015) 3546–3554
5. Ruder, S.: An overview of multi-task learning in deep neural networks. CoRR **abs/1706.05098** (2017)
6. Tan, J.H., Acharya, U.R., Bhandary, S.V., Chua, K.C., Sivaprasad, S.: Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. Journal of Computational Science **20** (2017) 70 – 79
7. Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendonça, A.M., Campilho, A.: End-to-end adversarial retinal image synthesis. IEEE Transactions on Medical Imaging **37**(3) (2018) 781–791
8. Alipour, S.H.M., Rabbani, H., Akhlaghi, M.R.: Diabetic retinopathy grading by digital curvelet transform. Comp. and Math. Methods in Medicine **2012** (2012)
9. Hervella, A.S., Rouco, J., Novo, J., Ortega, M.: Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES). (2018)
10. Ronneberger, O., Fischer, P., Brox, T.: U net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). (October 2015) 234–241
11. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Proc. **13**(4) (2004) 600–612
12. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision (ICCV). (December 2015) 1026–1034
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (May 2015)
14. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge based vessel segmentation in color images of the retina. IEEE Trans. on Medical Imaging **23**(4) (2004) 501–509

2.3 Journal Paper: Self-supervised multimodal reconstruction of retinal images over paired datasets

Self-supervised multimodal reconstruction of retinal images over paired datasets

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña (Spain)

[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Journal: Expert Systems with Applications
Volume: 161
Pages: 113674
Editorial: Elsevier
ISSN: 0957-4174
Year: 2020

Self-supervised multimodal reconstruction of retinal images over paired datasets

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Marcos Ortega[a,b]

[a] *CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain*
[b] *Department of Computer Science, Universidade da Coruña, A Coruña, Spain*

Abstract

Data scarcity represents an important constraint for the training of deep neural networks in medical imaging. Medical image labeling, especially if pixel-level annotations are required, is an expensive task that needs expert intervention and usually results in a reduced number of annotated samples. In contrast, extensive amounts of unlabeled data are produced in the daily clinical practice, including paired multimodal images from patients that were subjected to multiple imaging tests. This work proposes a novel self-supervised multimodal reconstruction task that takes advantage of this unlabeled multimodal data for learning about the domain without human supervision. Paired multimodal data is a rich source of clinical information that can be naturally exploited by trying to estimate one image modality from others. This multimodal reconstruction requires the recognition of domain-specific patterns that can be used to complement the training of image analysis tasks in the same domain for which annotated data is scarce.

In this work, a set of experiments is performed using a multimodal setting of retinography and fluorescein angiography pairs that offer complementary information about the eye fundus. The evaluations performed on different public datasets, which include pathological and healthy data samples, demonstrate that a network trained for self-supervised multimodal reconstruction of angiography from retinography achieves unsupervised recognition of important retinal structures. These results indicate that the proposed self-supervised task provides relevant cues for image analysis tasks in the same domain.

Keywords: self-supervised learning, eye fundus, deep learning, multimodal, retinography, angiography.

1. Introduction

The increment in data availability has a prominent role in the recent rise and spread of deep learning algorithms, allowing the end-to-end training of solutions that achieve unprecedented results in a substantial number of vision problems (Guo et al., 2016). However, data scarcity is still a common limiting factor for the successful training of modern Deep Neural Networks (DNNs) (Litjens et al., 2017). Although there are some large-scale annotated datasets for vision problems in which deep learning was successfully applied (Deng et al., 2009; Patterson & Hays, 2016; Everingham et al., 2010), it is usually challenging to gather an equivalent amount of data for several tasks and application domains. This leads to an increasing interest in the development of techniques that allow an effective use of the virtually unlimited amount of unlabeled images and videos (Litjens et al., 2017).

Annotated data is an especially scarce resource in medical imaging domains (Tajbakhsh et al., 2016; Litjens et al., 2017), where the common size of annotated datasets is orders of magnitude lower than that of the broad domain datasets. The main reason is that the appropriate labeling of medical images requires

[*] Corresponding author

Email addresses: a.suarezh@udc.es (Álvaro S. Hervella), jrouco@udc.es (José Rouco), jnovo@udc.es (Jorge Novo), mortega@udc.es (Marcos Ortega)

knowledge and expertise. Manual image labeling is a tedious and time consuming task that usually requires the intervention of experienced specialists, and the professionals with the required knowledge usually cannot invest large periods of time in the manual labeling of large image collections. Additionally, a significant amount of the annotated data must be held out for the clinical validation of the proposed methods, which further reduces the amount of data that is available for training and calibration.

In contrast, medical imaging is commonly used for the diagnosis and follow-up of patients in the daily clinical practice, which produces extensive amounts of unlabeled data. Also, increasingly large weakly-labeled datasets start to be available due to the use of clinical diagnoses as broad labels for the images. Nevertheless, detailed expert annotations are usually required for the precise localization of relevant anatomical structures and lesions. Additionally, routine clinical tests usually involve different image modalities, which results in the availability of paired multimodal medical image datasets. The different modalities offer complementary representations of anatomical structures and lesions, providing additional sources of relevant information for the clinicians. These paired datasets have been previously used as input for image analysis methods requiring the multimodal information (Liu et al., 2015). However, the unlabeled multimodal data can be additionally used to gain insight about relevant image contents, even for applications that do not need the multimodal information as input. This possibility has not been previously explored, being the focus of the work herein described.

The described situation of data scarcity in medical imaging motivates the application of methods for improving the training of DNNs with reduced datasets (Litjens et al., 2017; Shin et al., 2016). Data augmentation strategies are frequently used in the field, being often a key contribution to the good performance of the trained systems (Litjens et al., 2017). The common approach implies performing color and spatial transformations that produce alternative appearances of the images for which labels are available (Jamaludin et al., 2017). These transformations can simulate new acquisition conditions, but they do not increase the variability of the anatomical structures and lesions in the images. Some recent works also explored the augmentation of datasets using synthetic data samples (Costa et al., 2018), which may increase the variability of the image contents but may also produce non-plausible anatomical structures.

Network pretraining is another extensively applied strategy when annotated data is scarce. This technique consists in the initialization of the network with parameters that result from the training of an additional task for which a large amount of data is available. This strategy has been shown to improve the performance in comparison to random initialization (Tajbakhsh et al., 2016). Despite the differences between natural and medical images, ImageNet (Deng et al., 2009) classification is a commonly used pretraining task in medical imaging, as it produces good feature extractors in the first layers of the networks (Shin et al., 2016; Tajbakhsh et al., 2016). A different pretraining approach consists in using autoencoders for the self-supervised reconstruction of the input data (Shin et al., 2013; Xu et al., 2016). This unsupervised pretraining benefits from additional unlabeled data samples and it has the potential to learn useful representations of domain-specific patterns from the implicit structure of the data.

Multi-task learning is another commonly applied strategy to extend the available training data. It consists in the simultaneous training of complementary tasks over the same application domain (Twinanda et al., 2017; Jamaludin et al., 2017). This setting allows increasing the number of labels that are available for learning a shared representation among the tasks (Ruder, 2017). Moreover, the targets of some of the auxiliary tasks may provide relevant information for the main task. This strategy has demonstrated to improve the performance with respect to the individual training of single tasks (Twinanda et al., 2017). Similarly, common pretraining tasks, such as self-supervised input reconstruction, demonstrate further contribution if they are simultaneously trained with the target task (Rasmus et al., 2015).

Weakly-supervised approaches have been recently explored as an alternative when detailed annotations are not available (Jamaludin et al., 2017). In these approaches, broad image labels are used to identify the image regions that contribute the most to the target global classification. Hence, the localization of some image contents can be roughly estimated in the absence of more detailed annotations.

Despite of the existing alternatives, the training of DNNs for medical image applications would further benefit from new approaches taking advantage of the available unlabeled data. In that sense, pretraining and multi-tasking strategies have demonstrated their ability to transfer the knowledge acquired in additional tasks. However, they are limited by the degree of domain-related information that an auxiliary task is able to extract in the absence of human supervision. Thus, it is desired the development of new complementary tasks able to learn relevant domain-specific patterns from the unlabeled data. In this work, we propose a novel approach based on self-supervised multimodal reconstruction. This reconstruction task may be used to complement the training of DNNs using both pretraining and multi-tasking strategies.

1.1. Related work

An effective way to learn representations from unlabeled data using neural networks is the use of self-supervised tasks. The idea is to design complex supervised machine learning tasks in which the supervisory signal can be automatically derived from the input data. Classical approaches like autoencoders with equal input and output fall into this paradigm. In autoencoders, an information bottleneck is enforced at the hidden layers to perform data compression and, more importantly, to avoid learning a trivial identity solution between the input and the output (Bengio et al., 2013). Adding corruption to the input data or regularization penalties to the network loss may also improve the bottleneck effect (Bengio et al., 2013). However, these additions do not usually make the reconstruction task complex enough to enforce the learning of domain-specific patterns and semantics from the input data. The current trend to address this issue is to use more complex tasks that exploit additional sources of self-supervisory signals (Fernando et al., 2017; Noroozi & Favaro, 2016).

Spatio-temporal arrangement of the input data is a common source of self-supervision. Time series prediction tasks are classical examples of this. Some recent works approach this paradigm in the form of video frame prediction (Lotter et al., 2017). Although simpler classification tasks, detecting video sequences with shuffled frames (Misra et al., 2016), or with odd events (Fernando et al., 2017) have been also proposed. Similarly, in some approaches the image contents are directly reconstructed from the surrounding spatial context (Pathak et al., 2016), while in others, simpler tasks consisting in the prediction of relative patch positions (Doersch et al., 2015), or solving random jigsaw puzzles (Noroozi & Favaro, 2016), are proposed.

Other self-supervised approaches use complementary sources of information in the input data. For example, color information is used to define a colorization pretext task in (Zhang et al., 2016), which was later used to complement learning approaches in medical imaging applications (Ross et al., 2018). Complementary view information was used in (Sermanet et al., 2018) to learn pose-invariant features. Information from different modalities has been also used to provide self-supervisory signals, in approaches relating the image information with sound (Owens et al., 2016), depth (Wang et al., 2017), or motion information (Agrawal et al., 2015). In this work, we propose a self-supervised task of this kind that aims to reconstruct one image modality from another of the same patient.

The idea under the multimodal image reconstruction is that both image modalities provide complementary visual representations of the same anatomical structures and lesions of interest. In general, given two or more complementary visual representations of the same real world object, the estimation of one of these representations from the others involves the extraction of relevant object features if no trivial path between the representations exists. This means that the color and structural transformations that ideally map one modality to the other would depend on the semantic content of the images. Thus, learning this multimodal transformation involves the recognition of high level patterns related to the image contents. Furthermore, the estimation of other image modalities has value besides the induced representation learning, as a good enough estimation will provide extended information without the need of additional equipment or acquisition procedures.

In this sense, while many of the previously proposed tasks are only used for representation learning, the proposed multimodal reconstruction has the additional contribution of providing an estimate of the output modality.

1.2. Proposed work

The proposed self-supervised multimodal reconstruction paradigm naturally fits medical image applications, given the extensive use of multimodal visual data in many clinical specialties. This implies that the same patients are subjected to multiple imaging tests, allowing the gathering of paired multimodal data. These datasets only require a multimodal registration procedure to allow the training of the multimodal reconstruction.

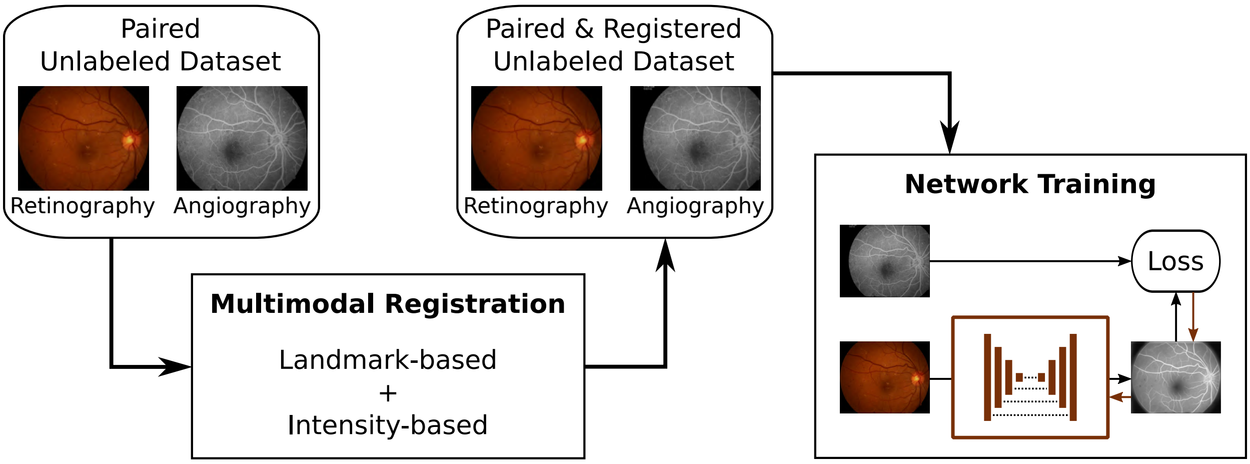


Figure 1: Proposed self-supervised approach using unlabeled multimodal data. First, the paired multimodal dataset is registered. The resulting registered dataset is used to train a DNN in the multimodal reconstruction of angiography from retinography.

In the work herein described, the proposed paradigm is applied to ophthalmology, where the use of several image modalities is the standard in clinical practice routine. In particular, we use the multimodal setting formed by color retinography and fluorescein angiography. These image modalities provide complementary visual representations of the eye fundus. The retinography is a color photography of the eye fundus that provides information of the retinal anatomical structures and lesions as seen in an ophthalmoscope. The angiography, instead, is a fluorescence image captured after that a fluorescein contrast dye is injected into the patient. Fluorescein increases the visibility of the blood vessels of the eye, giving additional information that is used to diagnose diseases affecting the circulatory system. Both modalities are used by the clinicians for the diagnosis and follow-up of many relevant diseases specific to the eye or systemic, such as age-related macular degeneration or diabetic retinopathy, for reference. However, despite its suitability for vascular analyses, the invasive nature of the angiography limits its use to patients with clear symptoms or already diagnosed. On the contrary, the retinography is affordable and non-invasive. Thus, it is suitable for periodic check-ups and screening programs, representing the most widely used ophthalmological image modality.

In this multimodal setting, we propose the self-supervised reconstruction of the angiography from a retinography of the same patient. These image modalities show important differences in the appearance of anatomical structures and lesions. The injected contrast has a different effect for each retinal structure and, therefore, the retinography-angiography appearance relation is structure-specific. This implies that the estimation of the transformation between retinography and angiography requires the recognition of the retinal structures, i.e., a trivial solution to the reconstruction does not exist.

The proposed approach for the the self-supervised reconstruction of angiography from retinography is summarized in the diagram of Figure 1. The multimodal reconstruction is performed using a U-Net fully convolutional neural network (Ronneberger et al., 2015). The network is trained using paired and aligned retinographies and angiographies of the same patient. The paired images are obtained from the publicly available Isfahan MISP dataset (Alipour et al., 2012) and from an additional private dataset. The alignment of the images is performed using the multimodal retinography-angiography registration algorithm proposed by Hervella et al. (2018a). The evaluation of the proposed setting is based on the unsupervised detection of the retinal vasculature. This evaluation is performed on two reference public datasets with vasculature annotations, DRIVE (Staal et al., 2004) and STARE (Hoover et al., 2000). Preliminary results of this work have been presented in (Hervella et al., 2018b). However, this paper presents important differences and additional contributions. Firstly, we provide a comprehensive contextualization of the proposal and a significantly more detailed description of the applied methodology. With respect to Hervella et al. (2018b), we have improved the data augmentation strategy for the network training by increasing the variety through additional color transformations. Also, in order to further evaluate the potential of the proposal, we provide a novel method in the evaluation that significantly improves the unsupervised recognition of the retinal vasculature. Finally, regarding the provided experiments, we have also studied important factors that may affect the performance, including the network size, number of training samples, and complexity of the images. In particular, the latter is possible due to the addition of two new datasets with more severe pathological cases.

The rest of the work is structured as follows. In Section 2, the algorithm for the multimodal registration of retinography-angiography pairs is described. In section 3, the proposed self-supervised multimodal reconstruction is detailed, including the description of the network architecture, the reconstruction loss, and the network training. Section 4 comprises the results and discussion for the different performed experiments. Finally, conclusions are drawn in Section 5.

2. Multimodal retinal image registration

The alignment of the multimodal image pairs is automatically performed following a recently proposed multimodal methodology for retinal images (Hervella et al., 2018a). The difference in intensity profiles for retinographies and angiographies prevents the direct comparison of pixel intensities between paired images. The intensity comparison is typically used for image registration in monomodal scenarios. Multimodal registration, instead, requires the transformation of the images to a common representation space. To that end, the applied methodology takes advantage of the presence of retinal vascular structures in both modalities. The methodology is divided into two steps, combining landmark-based and intensity-based registration approaches (Hervella et al., 2018a). The first step provides an initial low-order transformation that corrects the bulk of the misalignment between images. The second step computes a high-order transformation employing the initial transformation as initialization for the optimization of a similarity metric. This combination allows a robust and accurate registration of the images in this multimodal scenario.

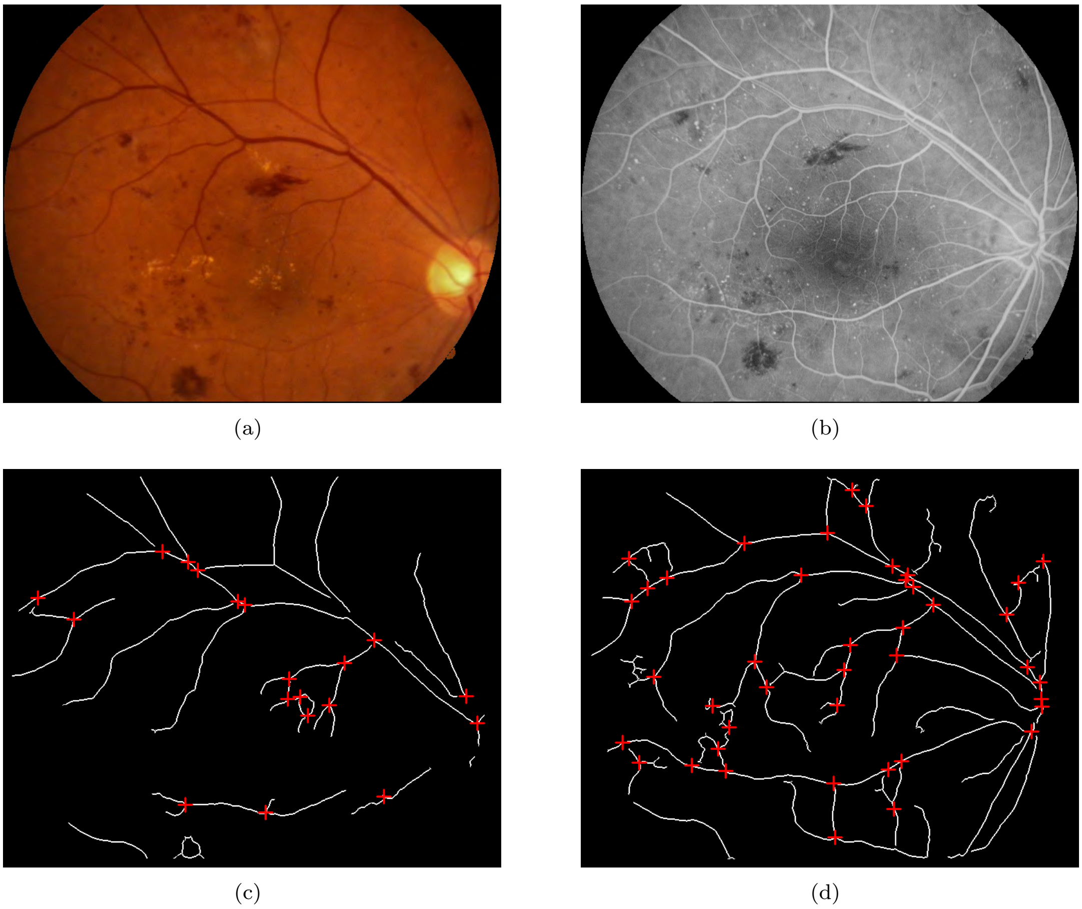
2.1. Initial registration

First, an initial landmark-based registration is performed using the bifurcations and crossovers of the vasculature. The automatic detection and matching of these domain-specific landmarks is based on a well-proven algorithm that was initially proposed for biometric authentication (Ortega et al., 2009). This algorithm treats the retinal image as a topological relief whose level curves are given by the intensity values in the image. The vessel centerlines are detected as the points of minima (in retinography) or maxima (in angiography) level curve curvature. After removing spurious points, an approximated vessel tree is formed. Then, the vessel intersection points, corresponding to bifurcations and crossovers, are identified in these trees. Examples of the detected vessel tree and landmarks for a retinography-angiography pair are depicted in Figure 2. Finally, the estimation of the spatial transformation between the images is computed by matching the bifurcation and crossover landmarks from both images. The considered transformation consists of translation, rotation, and isotropic scaling, only requiring the correct matching of two landmark pairs. This produces an initial estimation of the geometric transformation between the images that, although globally accurate, lacks some precision in the details.

2.2. Refined registration

The second step consists in an intensity-based registration that maximizes a pixel-wise similarity measure between the images. Due to the different intensity profiles of retinographies and angiographies, a transformation that maps both modalities to a common representation is applied. This transformation is performed with a Laplacian-based operation that enhances the vascular regions. This makes possible the direct comparison of pixel intensities between modalities.

The Laplacian is a second-order filter that produces high responses for tubular regions, such as the vessels in the retinal fundus. A vascular region is properly enhanced when the peak Laplacian response is obtained for the vessel centerline, which only happens if the scale of analysis fits the vessel width. Given that vessels



(a)

(b)

(c)

(d)

Figure 2: Example of vessel tree and detected landmarks for a retinography-angiography pair from a diabetic retinopathy patient. (a) Retinography. (b) Angiography. (c) Vessel tree and landmarks from (a). (d) Vessel tree and landmarks from (b).

with different widths are present in retinal images, multiple Laplacian scales are used for the analysis. Given an image $\mathbf{x}$, the Laplacian response at a scale t is defined as:

$$L(\mathbf{x}; t) = t^2 \Delta G(t) * \mathbf{x} \tag{1}$$

where $G(t)$ is a Gaussian kernel with scale parameter t, Δ denotes Laplacian, and $*$ denotes the convolution. The Gaussian kernel is defined as:

$$G(a, b; t) = \frac{1}{2\pi t} e^{-\frac{a^2+b^2}{2t}} \tag{2}$$

where (a, b) are the pixel coordinates with respect to the kernel center. The use of multiple scales requires the normalization of individual responses with a t^2 factor so their magnitudes are comparable (Lindeberg, 1998). Then, the maximum response across scales for each pixel is gathered in a multiscale Laplacian map computed as:

$$MSL(\mathbf{x}, m) = max_{t \in S} \lceil mL(\mathbf{x}; t) \rceil_{\emptyset} \tag{3}$$

where $\lceil \cdot \rceil_{\emptyset}$ denotes halfwave rectification, and m is a sign factor with values of $m = 1$ for retinographies and $m = -1$ for angiographies. The rectification is used to avoid the negative Laplacian peaks outside the vessel regions. The sign factor m is used to take into account that vessels appear as dark regions over light background in retinographies, whereas they present the inverse relation in angiographies. Figure 3 depicts examples of multiscale Laplacian maps for the retinography and angiography in Figure 2.

Once the multiscale Laplacian maps are computed for both modalities, the Normalized Cross-Correlation (NCC) is used as similarity metric for their comparison. The NCC is defined as:

$$NCC(\mathbf{x}, \mathbf{y}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{(x_{i,j} - \mu_{\mathbf{x}})(y_{i,j} - \mu_{\mathbf{y}})}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two single channel images, $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the averages of $\mathbf{x}$ and $\mathbf{y}$ respectively, $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ are the standard deviations of $\mathbf{x}$ and $\mathbf{y}$ respectively, and H and W are the height and width image dimensions. The refined spatial transformation, consisting in an affine transform followed by a free-form deformation, is obtained through the optimization of this metric with a gradient descent algorithm. The final transformation is obtained as:

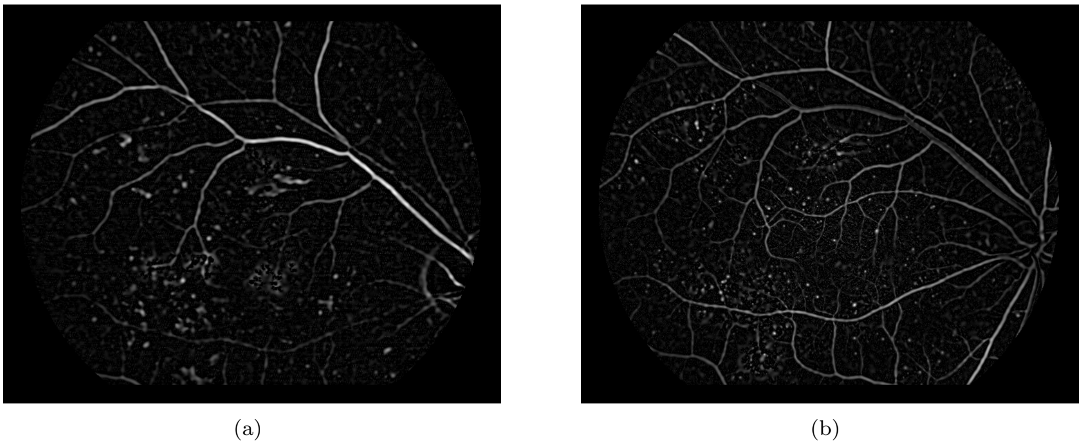
$$T^* = \arg\max_{T} NCC(MSL(\mathbf{r}, 1), MSL(T(\mathbf{a}, -1))) \tag{5}$$

where $(\mathbf{r}, \mathbf{a})$ is an unregistered retinography-angiography pair, and T is the transformation that produces the aligned pair $(\mathbf{r}, T(\mathbf{a}))$. Although the multiscale Laplacian also produces response for other structures different from the vessels, it has proven to be accurate enough for a NCC-driven registration when a proper initialization is given (Hervella et al., 2018a). This initialization is provided by the previously described landmark-based registration.

3. Self-supervised multimodal reconstruction of retinal images

The proposed multimodal reconstruction task consists in the estimation of an angiography from a retinography of the same eye. This task can be formulated as learning an image-to-image transformation $G : \mathcal{R} \rightarrow \mathcal{A}$ that maps a retinography $\mathbf{r} \in \mathcal{R}$ to its corresponding angiography $\mathbf{a} \in \mathcal{A}$.

Figure 4 depicts the main retinal structures in representative examples of the two considered modalities. It can be observed that the appearance of these retinal structures differs from one image modality to the other. As an illustration, the vasculature, red lesions, and fovea share similar color and intensity profiles in the retinography, whereas their intensity features are different in the angiography. The presence of the contrast dye in the bloodstream also produces some structural changes between both image modalities. The vasculature appears slightly thickened in the angiography and the small vessels, which can be hardly perceived in the retinography, are clearly visible. Simultaneously, the bright lesions observed in


(a)

(b)

Figure 3: Example of multiscale Laplacian maps for the retinography-angiography pair depicted in Figure 2. (a) Multiscale Laplacian map for the retinography. (b) Multiscale Laplacian map for the angiography.

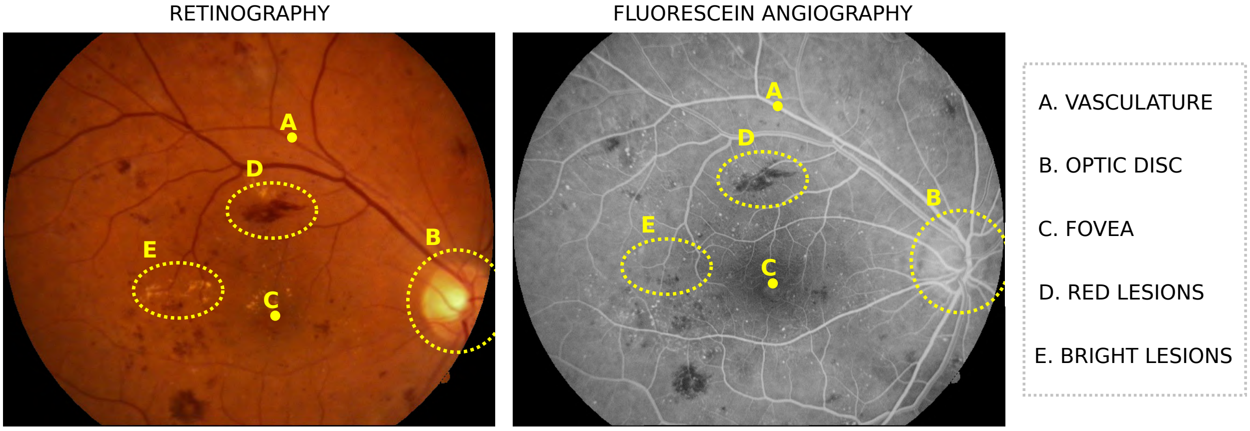


Figure 4: Example of color retinography and fluorescein angiography from the same eye of a diabetic retinopathy patient. The appearance of the retinal structures, such as vasculature, optic disc, fovea, red lesions and bright lesions is different from one image modality to the other. The transformation between retinography and angiography requires the identification of these structures in the image.

the retinography are not visible in the angiography. These differences indicate that both image modalities provide complementary information about the same retinal structures. Additionally, they evidence that the multimodal reconstruction between retinography and angiography is not trivial and requires the recognition of relevant patterns for this application domain.

A neural network trained for multimodal reconstruction should, therefore, be able to recognize this relevant patterns. This recognition ability may be exploited in other applications of the same domain through transfer of multi-task learning approaches. Furthermore, the estimated transformation G can be directly used to produce a pseudo-angiography representation $\hat{\mathbf{a}} = G(\mathbf{r})$ that shares the visual properties of an actual angiography, but with the advantage of being obtained without additional equipment or invasive procedures.

3.1. Network architecture

The proposed multimodal reconstruction is performed using an U-Net fully convolutional neural network (Ronneberger et al., 2015). This network architecture is characterized by using a contractive convolutional

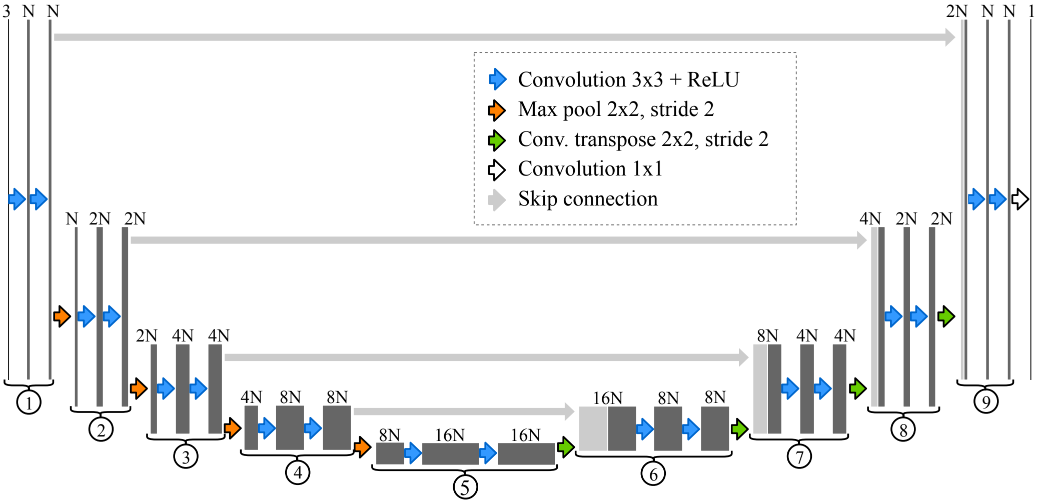


Figure 5: U-Net architecture as implemented for the experiments of this work. The number of channels is indicated for each feature map. The numbers below identify the convolutional blocks.

encoder followed by an expansive convolutional decoder, with additional skip connections that preserve the spatial localization of the learned patterns.

In the initial contractive path, the width and height image dimensions are sequentially reduced, creating a spatial bottleneck that helps with extracting relevant data patterns and learning high level representations. In the expansive path, the input space dimensionality is recovered with a progressive upsampling, producing a network output in the same scale of the input image. This yields a symmetric architecture where both parts of the network, encoder and decoder, have similar complexity. The downsampling operations are performed with spatial max pooling whereas the upsampling with transpose convolutions.

The downside of the created spatial bottleneck is that the precise localization of extracted data patterns is compromised. U-Net solves this issue transferring some additional information between the encoder and the decoder. Particularly, the feature maps extracted just before each max pooling are transferred to the corresponding layer in the decoder, through the use of skip connections. This creates an alternative path in the network that effectively skips part of the innermost layers and max pooling operations, ensuring that fine details are not lost.

A scheme of the used network is depicted in Figure 5. The network comprises nine convolutional blocks. Each block is composed of two convolutional layers followed by a downsampling or upsampling operation, for the encoder or decoder parts, respectively. All the convolutional layers have 3×3 kernels, following the same strategy proposed in VGG-Net (Simonyan & Zisserman, 2015). The hidden layers have ReLU activation functions. The output layer activation is linear to allow the whole range of values for the regression. The first convolutional block of the decoder has N output channels. The number of channels increases for subsequent blocks as the spatial dimensions of the feature maps decrease. The symmetric relation is held for the decoder blocks. For the experiments in this work, $N = 64$ unless stated otherwise.

3.2. Multimodal reconstruction loss

The multimodal reconstruction task is trained with a paired multimodal set of aligned retinography-angiography pairs $\{(\mathbf{r}, \mathbf{a})_1, ..., (\mathbf{r}, \mathbf{a})_n\}$. For each retinography $\mathbf{r}$, its corresponding angiography $\mathbf{a}$ acts as a pseudolabel. A pixel-wise loss between the network output and the pseudolabel is used as supervisory signal.

This self-supervised setting is enabled by the registration of the training data, aligning both image modalities using the algorithm described in Section 2. Retinal images are characterized for displaying the eye fundus in a circular region of interest (ROI) usually centered respect to the image frame. After the multimodal registration, the same eye pose is observed in both images, but the ROIs are likely to not

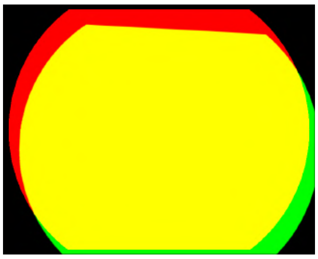


Figure 6: Example of Multimodal ROI, in yellow, where multimodal data is available. The retinography comprises the red and yellow areas, whereas the angiography comprises the green and yellow.

completely overlap. Then, a multimodal ROI Ω_M is defined as the intersection between the retinography and the angiography ROIs, Ω_R and Ω_A respectively, so that the set of pixels that contain information from both modalities is identified. An example of this is depicted in Figure 6. Thus, the loss is only computed for the pixels contained in Ω_M. However, whole retinographies are fed to the network, as every pixel in Ω_R provides valuable contextual information for the estimation of individual pixels in Ω_A.

For any pair $(\mathbf{r}, \mathbf{a})$ of the training set, the multimodal reconstruction loss is given by:

$$\mathcal{L}^{\mathcal{E}} = \sum_{\Omega_M} \mathcal{E}(G(\mathbf{r}), \mathbf{a}) \tag{6}$$

where $\mathcal{E}(G(\mathbf{r}), \mathbf{a})$ is an error map computed with the error function $\mathcal{E}$. The sum over all pixels in Ω_M is used instead of the average because $|\Omega_M|$ varies between training samples, and the average error would give more weight to the pixels of less overlapped image pairs.

For the error function $\mathcal{E}$, three different alternatives are considered. As the proposed reconstruction is a regression problem, it is natural to consider the L2-norm, which is defined as:

$$L2(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2^2 \tag{7}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two single channel images. The L1-norm is another common choice for regression, which approximates the output to a median representation instead of the mean approximated by L2-norm. It is defined as:

$$L1(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| \tag{8}$$

The third alternative is the optimization of the Structural Similarity (SSIM) index (Wang et al., 2004). SSIM is a similarity metric initially proposed for image quality assessment that is commonly used as test metric for the evaluation of image reconstruction, super-resolution or image synthesis tasks. However, SSIM is rarely chosen as optimization objective. Zhao et al. (2017) proposed the optimization of SSIM for image restoration, reporting improved results with respect to other common loss functions. Given that SSIM is a measure of similarity, the negative SSIM is used as reconstruction loss. The SSIM is defined as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1) + (2\sigma_{\mathbf{xy}} + C_2)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C_2)} \tag{9}$$

where $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the local averages of $\mathbf{x}$ and $\mathbf{y}$ respectively, $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ are the local standard deviations of $\mathbf{x}$ and $\mathbf{y}$ respectively, and $\sigma_{\mathbf{xy}}$ is the local covariance between $\mathbf{x}$ and $\mathbf{y}$. These statistics are computed locally for each image point using a Gaussian window with $\sigma = 1.5$ (Wang et al., 2004). The main difference of SSIM with respect to the other considered functions is that the error value for each pixel is conditioned by the intensity distribution in a small neighborhood. Therefore, the used SSIM loss could be seen as a local metric, opposite to L1 and L2 losses that are strictly point-wise.

3.3. Network training

For training, network parameters are randomly initialized following the method proposed by He et al. (2015). The Adam algorithm is used for the optimization with decay rates for the first and second order moments of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively, as proposed by Kingma & Ba (2015). The training data is randomly split in training and validation subsets with a 4 to 1 ratio. The starting learning rate is set to $\alpha = 1e\text{-}4$, being reduced by a factor of 10 each time the validation loss ceases to improve for 50 epochs. Finally, the training is stopped when the validation loss has not reached at least its best value for 100 epochs. These values were tuned by the analysis of learning curves in the training dataset.

Dropout and data augmentation techniques are used to avoid overfitting. Dropout layers are included after the convolutional blocks 3, 4 and 5 (depicted in Figure 5). In these layers, the activations are randomly set to zero following a Bernoulli distribution with probability $p = 0.2$. Random spatial and color data augmentations, similar to the ones used in other proposals (Jamaludin et al., 2017; Urban et al., 2017), are performed during training. The spatial augmentation consists in random affine transformations with rotation, scaling and shearing components. Color data augmentation consists in random linear transformations of the image components in HSV space as applied by Urban et al. (2017). The range for the transformations has been chosen beforehand to increase the variability of the image appearances while ensuring that they still resemble valid retinal visualizations.

4. Results and discussion

4.1. Training datasets

Two different datasets are used for training the multimodal reconstruction. One of the datasets is from the Isfahan MISP database (Alipour et al., 2012), which is publicly available. It is composed of 59 retinography and angiography pairs, including both healthy and pathological cases. The latter are from patients diagnosed with diabetic retinopathy. The size of the images is 720×576 pixels. The other dataset is a private collection of 59 retinography and angiography pairs provided by the Complexo Hospitalario Universitario de Santiago de Compostela (CHUS), Galicia, Spain. These images present mild and severe pathological cases of different diseases. The size of the images is 768×576 pixels. Both datasets provide unaligned image pairs that must be registered to enable the self-supervised multimodal reconstruction.

All the experiments performed in this work, except for the ones in Section 4.8, use the public Isfahan MISP dataset for training the multimodal reconstruction. For the experiments in Section 4.8 both datasets are used.

4.2. Quantitative evaluation

In order to quantitatively evaluate whether the trained multimodal reconstruction networks have learned about the domain, an analysis of their capability for retinal vasculature detection is performed.

In particular, one important characteristic of angiographies is the improved visibility of the retinal vessels with respect to retinographies. It is expected, therefore, that the multimodal reconstruction networks will be able to generate a pseudo-angiography with this same property from any given retinography. In such case, a rough vessel segmentation could be performed on the pseudo-angiography using a global threshold with appropriate value. The same thresholding procedure over the retinography should produce much worse results.

The evaluation of this segmentation is used as a measurement of the saliency of the retinal vessels in the images. The segmentation performance is evaluated with respect to the ground truth using Receiver Operator Characteristic (ROC) and Precision-Recall (PR) analyses. Both analyses employ a variable threshold to produce multiple binary maps where the segmentation is evaluated. The results obtained for all the individual thresholds are aggregated in ROC and PR curves.

ROC curves plot False Positive Rate (FPR) against True Positive Rate (TPR). In this scenario, the FPR is the ratio of non-vessel pixels incorrectly classified as vessels. The values can be obtained for each threshold as:

$$FPR = \frac{FalsePositives}{FalsePositives + TrueNegatives} \tag{10}$$

The TPR is the ratio of true vessel pixels that are correctly classified. The values can be obtained for each threshold as:

$$TPR = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{11}$$

PR curves, instead, plot Recall against Precision. Recall is the same measurement as TPR in Equation 11. Precision is the ratio of output vessel pixels correctly classified. It can be computed for each threshold as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{12}$$

Finally, ROC and PR curves can be summarized with their Area Under Curve (AUC). Both curves are typically used to evaluate the performance of algorithms in binary decision problems. The main difference between the results presented by these curves takes place when positive and negative examples are unbalanced. If the number of negative examples exceeds the number of positives examples, as happens with vessels and non-vessels in retinal images, PR curves are more sensitive to changes in the number of false positives, i.e. background pixels incorrectly classified as vessels.

4.3. Test datasets

The quantitative evaluation is performed using two different publicly available datasets, DRIVE (Staal et al., 2004) and STARE (Hoover et al., 2000), for which ground truth vessel segmentations are available. The DRIVE dataset is a collection of 40 retinographies with their corresponding ground truth vessel segmentations. This dataset is divided between training and test subsets. The training samples include a single ground truth annotation whereas the test samples present two annotations from two different human observers. The size of the images is 565 × 584 pixels.

The STARE dataset has 20 retinographies with associated ground truth vessel segmentations from two different human observers. The images in STARE correspond to mild and severe pathological cases. The size of the images is 700×605 pixels. Given that there is a significant variability between both annotations, we decided to use them as two independent datasets. By default, they are named STARE AH and STARE VK, being "AH" and "VK" referenced to the names of the human annotators.

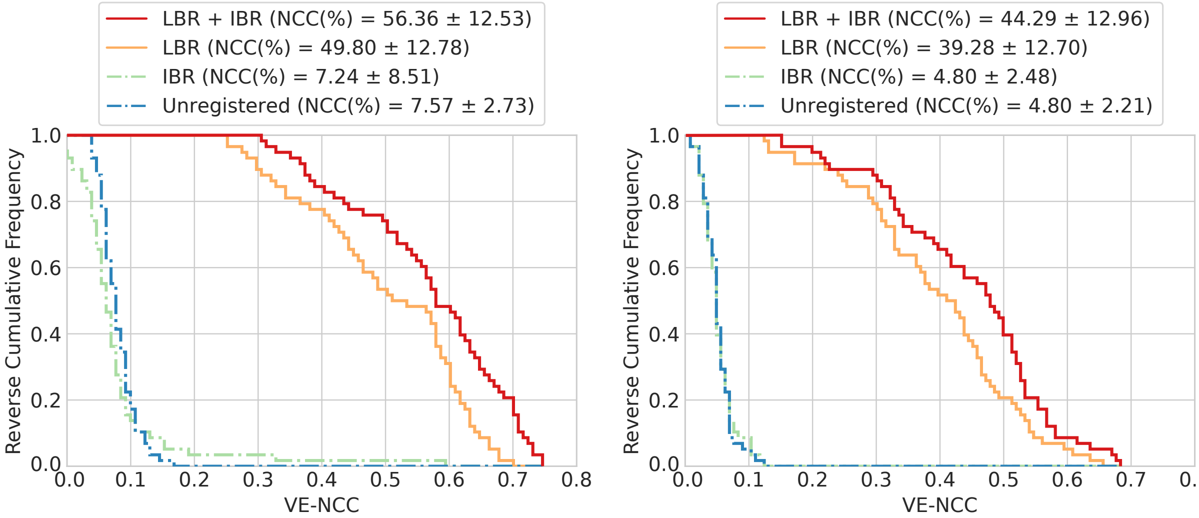
These datasets are usually split into training and test subsets. In this work, however, as the network training is performed using the unlabeled multimodal datasets described in Section 4.1, the whole datasets are used for testing purposes in the quantitative evaluation. The use of different datasets for training and test also allows evaluating the generalization ability of the proposed setting.

4.4. Multimodal registration results

The multimodal registration is evaluated using the NCC between paired retinographies and angiographies after applying the vessel enhancement described in Section 2. This operation is defined as VE-NCC. A better alignment is reflected by a higher VE-NCC value due to the matching of the retinal vascular structures between paired images.

Figure 7 depicts the reversed cumulative histograms for the VE-NCC before and after the multimodal registration in the training datasets. The plots also include the results of performing a registration with only the individual steps described in Section 2: the landmark-based registration (LBR) and the intensity-based registration (IBR). It is observed that the applied methodology, with two steps, achieves the best results. The sole application of the LBR greatly increases the VE-NCC with respect to the unregistered images. However, it produces worse results than the combined approach. This indicates that the LBR alone is able to produce a rough registration that is latter successfully refined. On the other hand, the independent application of the IBR only improves the VE-NCC for a few images, failing to register the images when a large transformation is required. These results evidence that the IBR can reach a more accurate registration than the LBR but it is highly dependent on the initialization. In this case, the initial transformation is provided by the LBR. This demonstrates the suitability of the combined approach.

The results also show a high variability among image pairs for the measured VE-NCC. This is due to the fact that the vessel enhancement produces some response for other retinal structures besides the vessels, and



(a) (b)

Figure 7: Results of the multimodal registration for the training datasets in terms of the VE-NCC . (a) Isfahan MISP. (b) CHUS.

this additional response depends on the individual characteristics of the images. It is observed, for example, that the achieved VE-NCC values after the registration are worse for the CHUS dataset, whose images comprise more pathological manifestations. Despite these differences, the maximum VE-NCC achieved for each image pair produces an adequate registration when visually evaluated.

Figure 8 shows an example of the multimodal registration including intermediate and final results. It is observed that the images are globally registered after the LBR. However, they are not completely aligned, which is evidenced in the vessels when they are observed in detail. The IBR after the LBR corrects these misalignments. This agrees with the previous analysis of the VE-NCC values for the whole datasets.

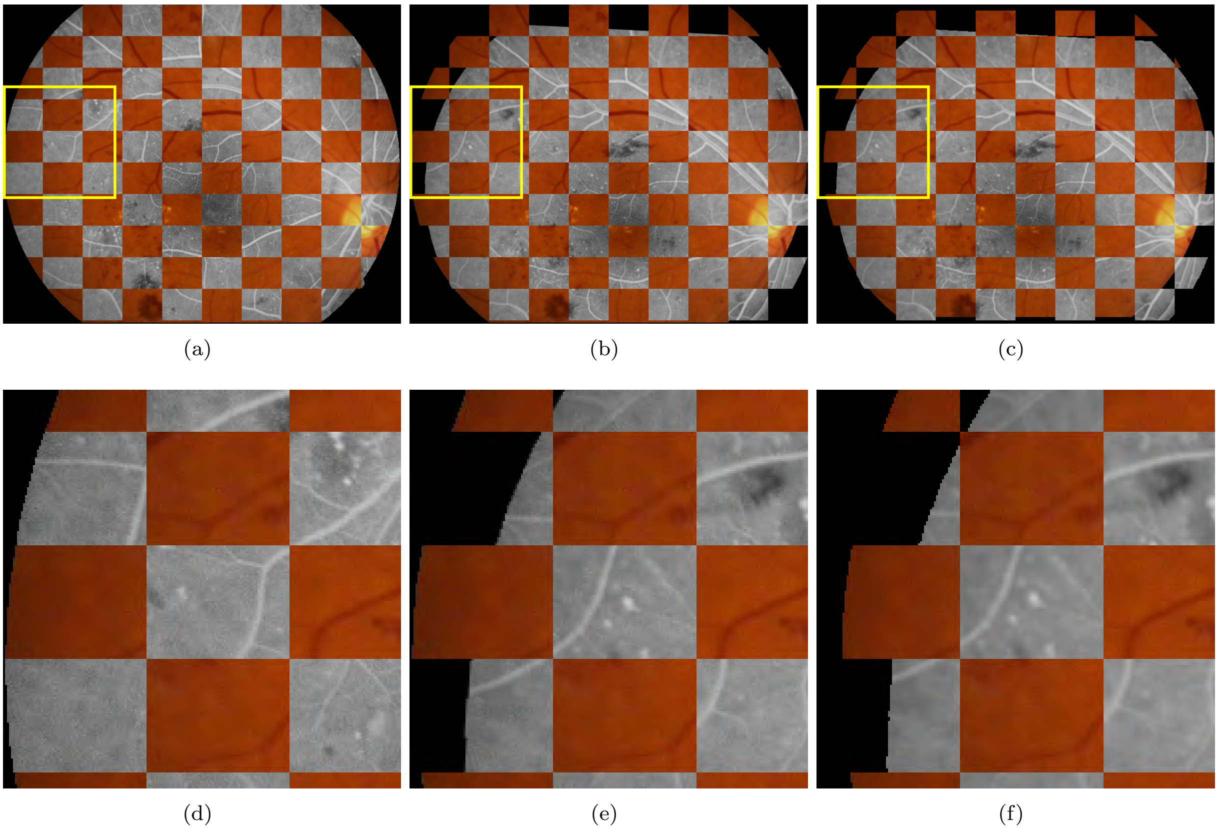
4.5. Comparison of loss functions

Figure 9 shows an example of the generated pseudo-angiographies using the models trained with the three losses described in Section 3.2. The input image corresponds to the retinography depicted in Figure 4, which is part of the validation set. It can be observed that the models trained with L2 and L1 generate blurred images with less small vessels visible. Also, these models reconstruct the vasculature and red lesions in a similar manner, while the appearance of these structures differs in the target angiography (Figure 4). On the contrary, the model trained with SSIM generates sharper images, with a higher rate of small vessels visible. The red lesions, in this case, can be distinguished from the vasculature by their intensity level.

The validation errors obtained after training with the different loss functions are shown in Table 1. The model trained with SSIM obtains better results even when the comparison is performed in terms of L2 and L1 loss values. This indicates that SSIM provides better properties for the self-supervised multimodal reconstruction training.

The results for the quantitative evaluation described in Section 4.2 are depicted in Figure 10. These curves show a comparison of the three considered training losses when evaluated in the test datasets. It is observed that SSIM outperforms the other losses in all the experiments. Training with SSIM leads to a greater vasculature saliency, which eases the threshold based segmentation of the vessels. Despite the lower performance, L1 and L2 obtain similar results in all the experiments.

The comparison of the results for the different test datasets reveals that the gap between SSIM and the other losses is greater in STARE than in DRIVE. These results are explained by the fact that the models



(a)

(b)

(c)

(d)

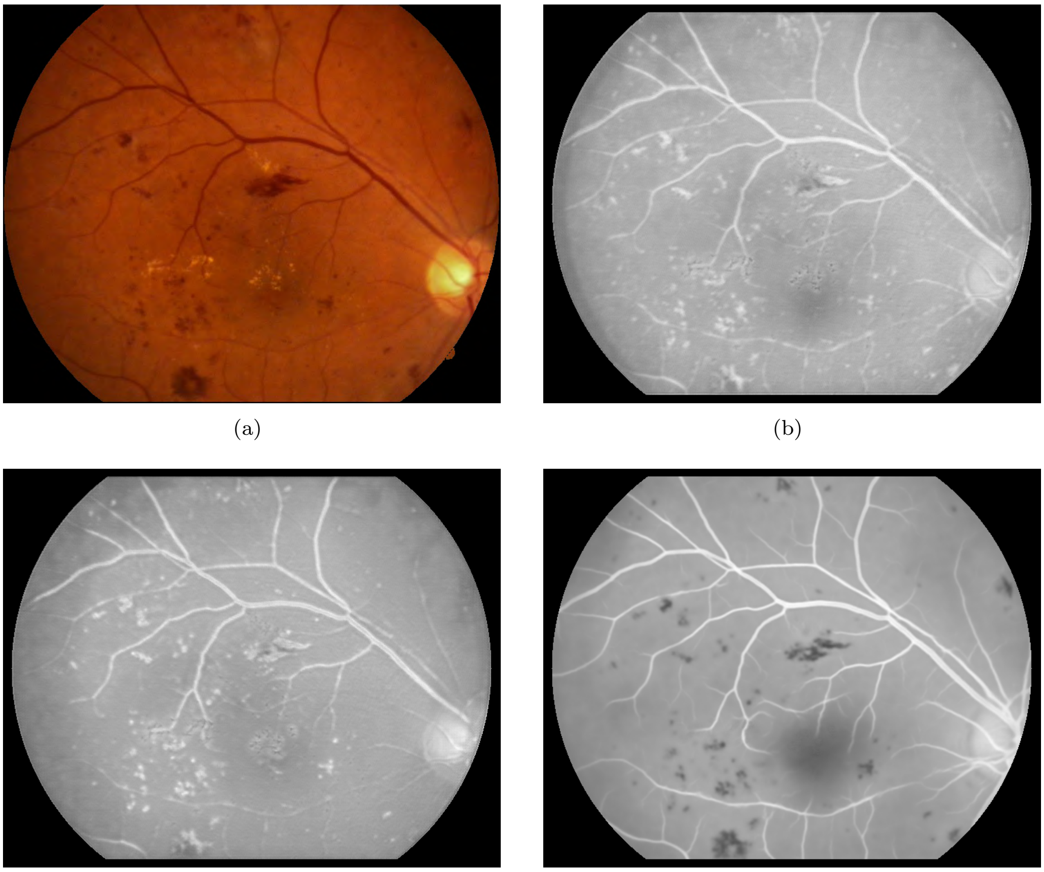
(e)

(f)

Figure 8: Example of the multimodal registration for a retinography-angiography pair. (a) Before the registration. (b) After the inital registration (LBR). (c) After the refined registration (LBR+IBR). (d) Detail from (a). (e) Detail from (b). (f) Detail from (c).

Table 1: Cross-comparison of error functions. The values in the table are computed as the average pixel loss in the validation set after training.

Training	Validation loss		
loss	L2	L1	SSIM
L2	0.0378	0.1646	−0.6805
L1	0.0375	0.1628	−0.6859
SSIM	**0.0217**	**0.1161**	**−0.7642**



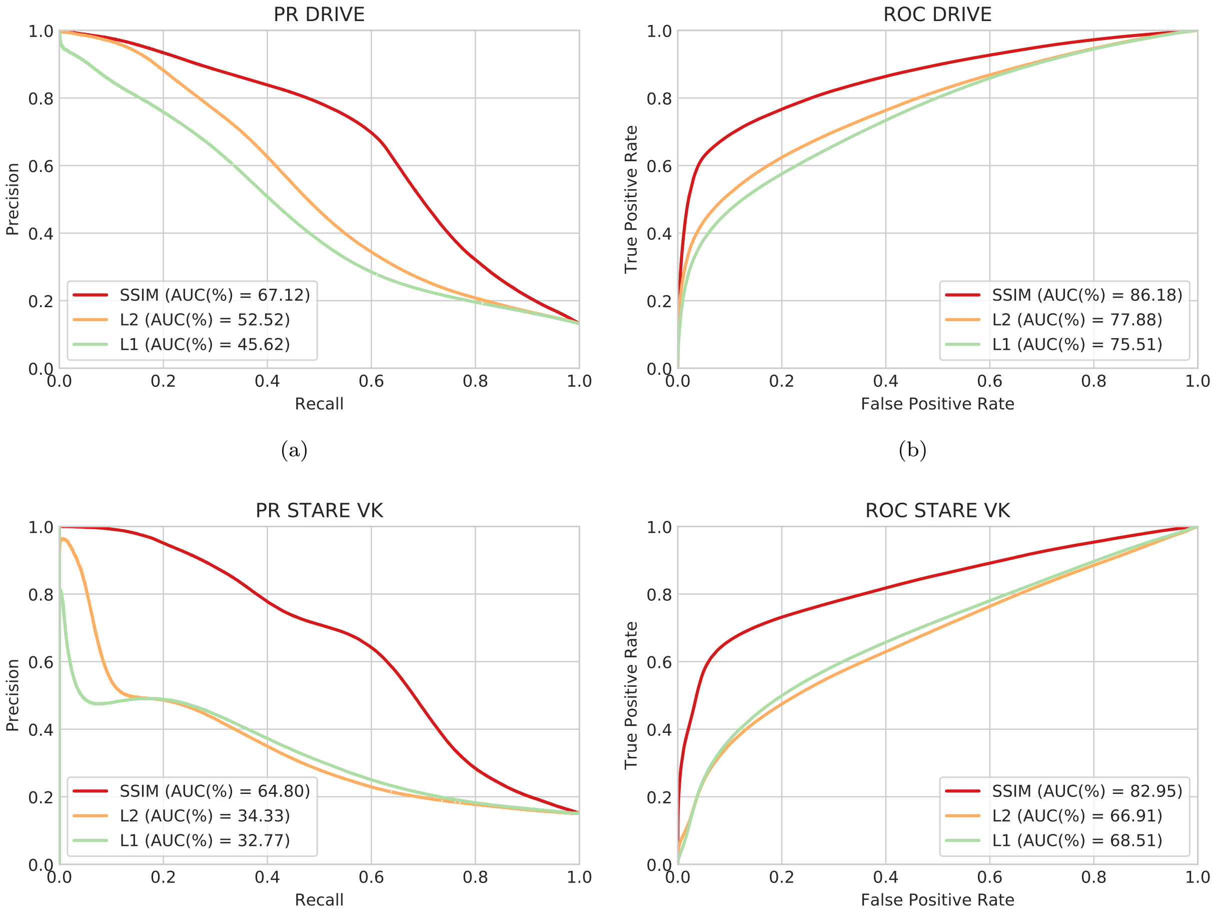
(a)

(b)

(c)

(d)

Figure 9: Example of generated pseudo-angiographies. (a) Original retinography. (b) Using the L2 training loss. (c) Using the L1 training loss. (d) Using the SSIM training loss. L2 and L1 produce blurred images with similar appearance, whereas SSIM produces sharper images where the different retinal structures are easily identified.



(a)

(b)

(c)

(d)

Figure 10: Comparison of the different training losses. The graphics depict PR ((a), (c)) and ROC ((b), (d)) curves. (a)-(b) Using the DRIVE images as test set. (c)-(d) Using the STARE images as test set with the VK ground truth. The curves obtained for STARE AH are similar to those of figures (c) and (d).

Table 2: Experiments performed to study the effect of the network size varying the parameter N. AUC-PR and AUC-ROC values are measured in the DRIVE, STARE AH and STARE VK datasets. The results indicate that the performance is improved with the increased size, also reducing the variability.

N	Parameters	DRIVE		STARE AH		STARE VK	
		PR (%)	ROC (%)	PR (%)	ROC (%)	PR (%)	ROC (%)
2	30k	60.52±4.86	63.64±1.67	46.77±20.95	61.91±14.51	47.31±19.96	60.09±12.36
4	122k	63.78±4.57	77.03±1.37	58.97±4.75	75.20±12.79	58.32±3.96	71.89±11.11
8	489k	65.96±1.88	84.04±0.78	61.60±4.15	82.19±1.68	59.88±3.58	77.65±1.38
16	2M	65.23±1.16	84.67±0.49	61.00±2.65	83.38±1.33	59.24±2.17	78.73±1.02
32	8M	65.78±0.52	85.18±0.29	63.27±1.50	85.28±0.98	61.55±1.40	80.51±0.82
64	32M	65.85±1.29	**85.59±0.50**	**66.43±1.06**	87.35±0.52	**64.40±1.07**	82.37±0.57
128	128M	**66.03±0.94**	85.46±0.35	65.46±1.81	**87.56±0.47**	63.38±1.47	**82.38±0.46**

trained with L1 or L2 fail to differentiate between the vasculature and the red lesions. The images from DRIVE include less pathological structures, thus the performance in this dataset is less penalized.

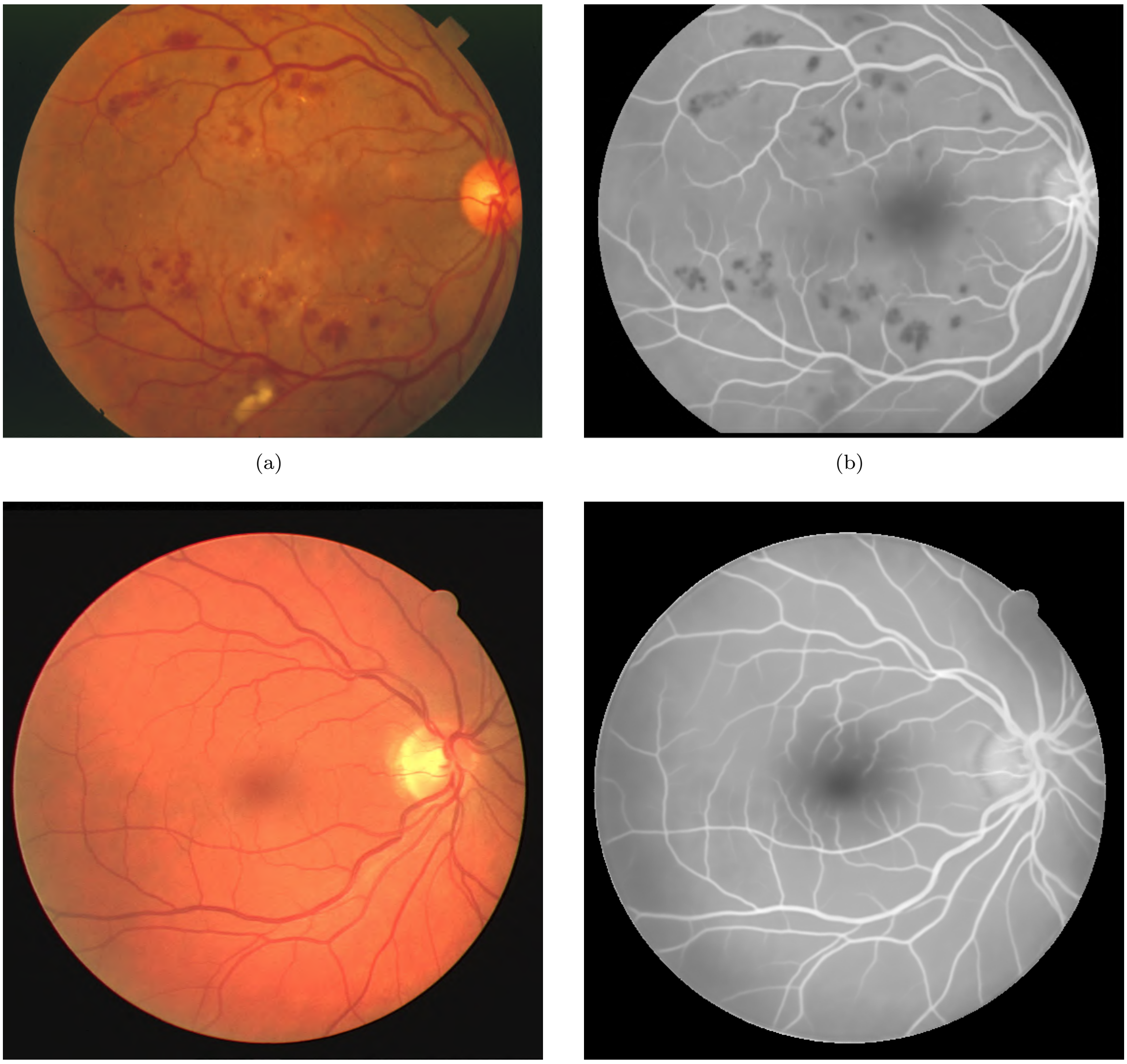
4.6. Unsupervised recognition of retinal patterns

The example shown in Figure 9(d) reveals that the network trained for multimodal reconstruction using SSIM has learned to identify and transform significant retinal structures. Additional examples using the SSIM model on DRIVE and STARE test images are shown in Figure 11. The vasculature is reconstructed with increased saliency, even for the small vessels. The reconstructed fovea and optic disc resemble the original colors of the angiography. Note, as reference, that the foveal region is clearly marked even if it is not easily perceived in the original retinography. The pathological structures are also reconstructed in a non-trivial manner. The red lesions are reconstructed with low intensity value and can be easily distinguished from the vessels and the background. Bright lesions, on the other hand, are reconstructed resembling the background, as happens in the angiographies. These retinal structures experiment an independent transformation from their retinography to the pseudo-angiography. This demonstrates that the multimodal reconstruction involves an understanding of the retinal structures. The recognition of the retinography patterns allows the generation of an image that resembles the target angiography, simulating the effect of the injected contrast.

The increment in the vasculature saliency, from retinography to pseudo-angiography, can be measured using the proposed quantitative evaluation method. Figure 12 depicts the quantitative results obtained with the SSIM pseudo-angiography in comparison with alternative methods. The pseudo-angiography curves represent the mean and standard deviation over 5 training repetitions with different random initializations. It is observed that thresholding over the pseudo-angiography provides better vessel extraction than using thresholding over the inverse retinography. This is the expected behavior if we compared the retinography with an actual angiography. However, simple vessel enhancement (VE) algorithms, like the multiscale Laplacian explained in Section 2 can also provide a fair vessel extraction from retinographies. For this reason, the comparison also includes an evaluation of the VE when applied to the retinography and the pseudo-angiography. It is observed that the VE retinography performs better than the raw pseudo-angiography. However, applying the VE over the pseudo-angiography provides the best results. This indicates that the trained network applies a complex processing that is able to remove the VE artifacts related to the presence of pathologies or other anatomical structures. Thus, these results evidence that the self-supervised multimodal reconstruction provides an unsupervised way to extract relevant retinal patterns, providing more information about the vasculature than the original retinography.

4.7. Effects of the network size

Experiments varying the network size are performed to evaluate how it affects to the learning of the required patterns. The parameter N in the U-Net architecture (Figure 5) is used to control the size of the



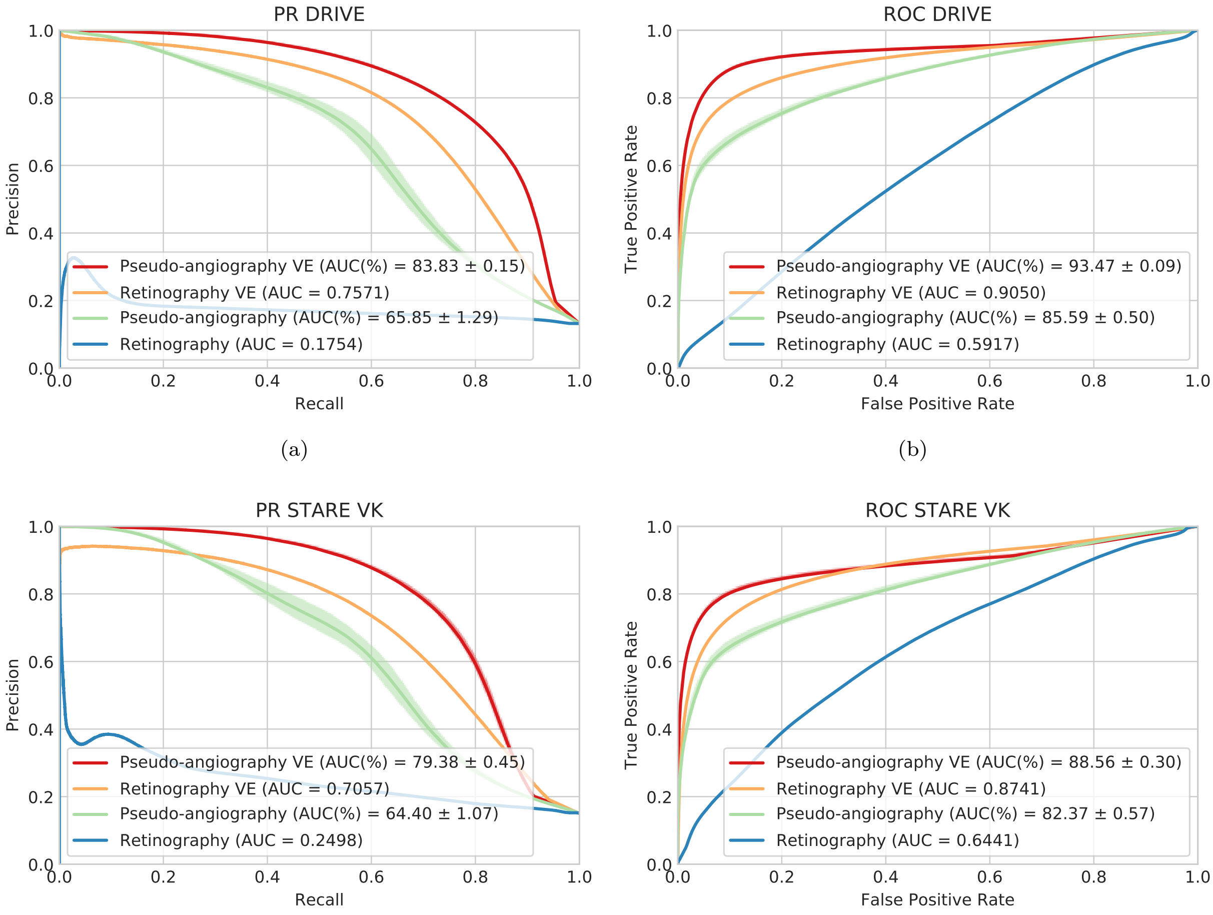
(a)

(b)

(c)

(d)

Figure 11: Examples of generated pseudo-angiographies on images from the test datasets, using the SSIM model. (a) Retinography from the STARE dataset. (b) Generated pseudo-angiography from (a). (c) Retinography from the DRIVE dataset. (d) Generated pseudo-angiography from (c).



(a)

(b)

(c)

(d)

Figure 12: Evaluation of the generated pseudo-angiography for the unsupervised recognition of vessel structures. (a)-(b) Using the DRIVE images as test set. (c)-(f) Using the STARE images as test set with the AH ground truth. The pseudo-angiography curves represent the mean and standard deviation over five training repetitions. The pseudo-angiography performs better than the original retinography but worse than using the vessel enhancement (VE) over the retinography. However, applying the VE over the pseudo-angiography provides the best results.

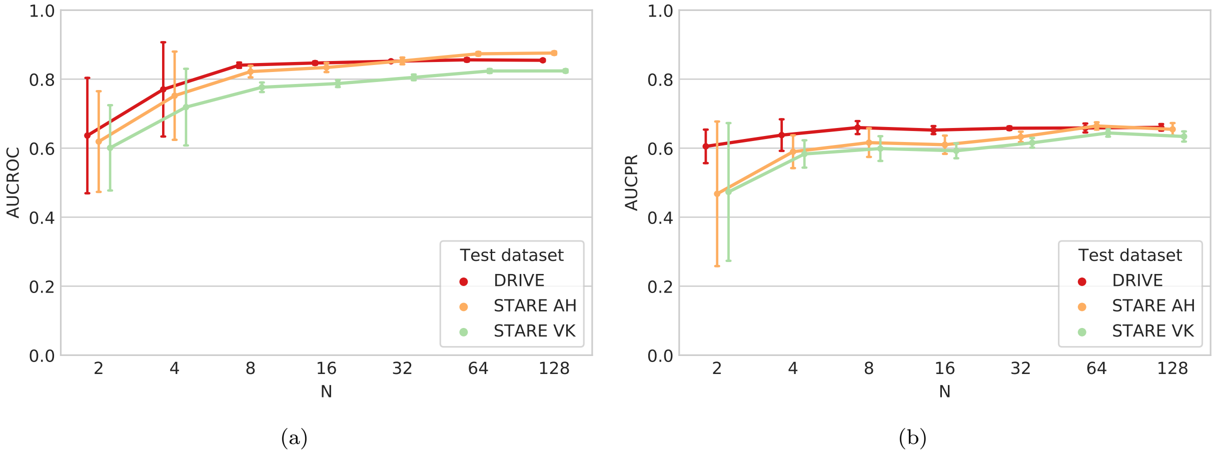


Figure 13: Evaluation of the network size. (a) AUC-ROC with varying N. (b) AUC-PR with varying N. The plots represent the mean and standard deviation over five training repetitions. The increased network size improves the average results and reduces the variance. The improvement is higher for the more complex datasets.

network. This parameter controls the network width while keeping the network depth and receptive field size constant. Networks with N values varying from $N = 2$ to $N = 128$ were trained on the Isfahan MISP dataset and evaluated using the quantitative procedure of Section 4.2 over the DRIVE and the STARE datasets. This training was repeated five times with different random initializations. Table 2 summarizes the obtained results, along with the number of parameters in each network configuration. These results are also presented in the plots in Figure 13. The best results are obtained for the largest networks, with very similar values for N=64 and N=128. It is observed that the variance is higher for low N values, decreasing at the time N increases. Also, the increased performance presents a higher impact in the STARE dataset, which is considerably more heterogeneous and complex than the DRIVE dataset. Thus, larger networks seem to extrapolate better to more complex cases and be more independent on the initialization.

4.8. Effects of additional training data

Additional experiments varying the number of training samples are conducted to study how this parameter affects the proposed multimodal reconstruction. Both training datasets described in Section 4.1 are used with that purpose, creating 3 different training configurations: Isfahan MISP (59 image pairs), CHUS (59 image pairs) and both (118 image pairs). This also allows to study how the use of different data sources may affect the performance.

The main results of these experiments are depicted in Figure 14. Each configuration is trained with 5 repetitions using different random initializations. It is observed that the highest AUC-PR and AUC-ROC are obtained with the largest training data. This indicates that the proposed setting benefits from larger datasets. This is an interesting result as the main advantage of the proposed setting is the ease of gathering additional data. The relative improvement is larger for the STARE dataset, which is a more complex scenario and benefits more from the increased diversity of the training data.

The comparison between Isfahan MISP and CHUS datasets shows that the source of data slightly affects the performance. From the six analyses summarized in Figure 14, only in one of the models trained with the CHUS dataset achieved better performance than those trained with Isfahan MISP dataset. As both datasets contain the same number of images, the different results must be explained by the different distribution of retinal characteristics and quality of the images. The CHUS dataset presents a higher rate of pathological structures, with a higher variation in the angiographies appearance. The Isfahan MISP dataset, instead, is more homogeneous, producing a more consistent enhancement of the vasculature. Nevertheless, the use of additional training samples improves the performance of both independent datasets.

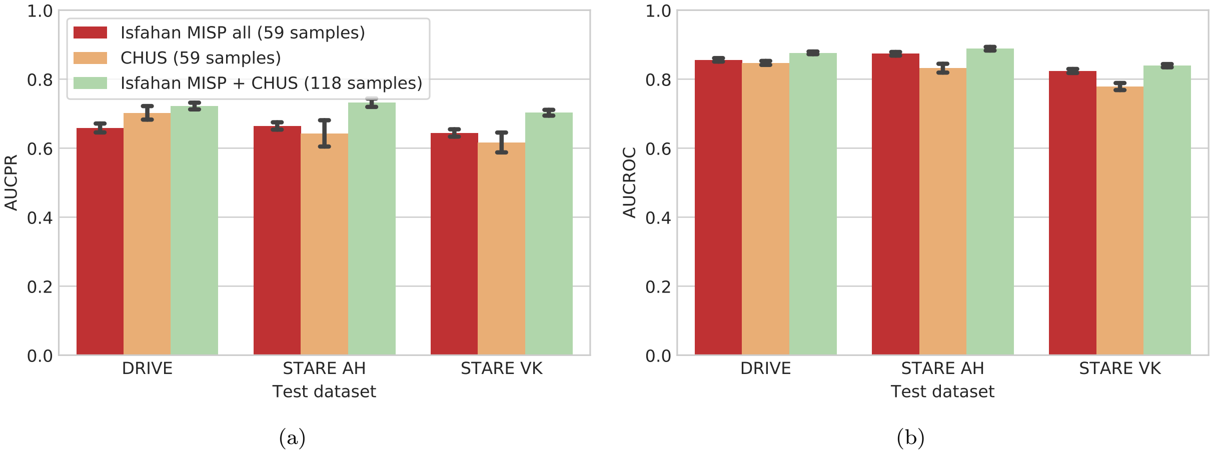


Figure 14: Evaluation of additional training data. (a) AUC-PR. (b) AUC-ROC.

5. Conclusions

The scarcity of annotated data in medical imaging motivates the development of solutions that target the successful training of DNNs with minimum human labeling. In this work, we proposed the multimodal reconstruction as a self-supervised task that can be automatically constructed given a set of paired images of different modalities. This approach naturally suits to medical imaging given that the multimodal scenario is frequent in the daily clinical practice of many specialities, which eases the data gathering. In our particular case, we performed experiments with the multimodal image setting formed by retinography and fluorescein angiography. Networks trained in the reconstruction of angiographies from retinographies of the same patient learn to identify important retinal structures and to simulate the effect of an injected contrast dye. The paired multimodal data for training the networks was obtained from public and private datasets that include healthy and pathological samples. For the evaluation of the trained networks additional public datasets were employed. The complexity of the learned transformations is evidenced by the qualitative analysis of the generated pseudo-angiographies. Exhaustive quantitative evaluation, based on the ability to detect the retinal vasculature, confirms that the multimodal reconstruction serves as a pretext task to learn important domain-specific patterns.

The obtained results show that, besides the new generated representation, the proposed multimodal reconstruction presents significant potential as a complementary task for training DNNs in situations of data scarcity. In this regard, a future research direction involves the application of the proposed approach in transfer learning or multitask settings. The aim would be to facilitate the use of DNNs with scarce annotated data and to improve the automated diagnosis of important retinal diseases. Additionally, given the availability of multimodal data in medical imaging, another future research direction is the application of the proposed paradigm in other medical domains. In this regard, it should be considered that, while the multimodal reconstruction is learned end-to-end with a DNN, the previous multimodal registration follows a domain-specific approach. Thus, this registration step could be seen as a limitation for the application of the paradigm in other medical domains. The solution, in this case, would be the adoption of adequate registration algorithms, which are potentially available due to the common use of registration techniques in medical imaging. Finally, we expect that the multimodal reconstruction will be helpful for the training of numerous image analysis tasks in the field.

Acknowledgments

This work is 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 by

Ministerio de Economía, Industria y Competitividad, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the ERDF and Xunta de Galicia through Grupo de Referencia Competitiva, ref. ED431C 2016-047, and from the European Social Fund (ESF) of the EU and Xunta de Galicia through the predoctoral grant contract ref. ED481A-2017/328. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

Conflict of interest

The authors declare no conflicts of interest.

References

Agrawal, P., Carreira, J., & Malik, J. (2015). Learning to see by moving. In *International Conference on Computer Vision (ICCV)*.

Alipour, S. H. M., Rabbani, H., & Akhlaghi, M. R. (2012). Diabetic retinopathy grading by digital curvelet transform. *Computational and Mathematical Methods in Medicine*, *2012*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1798–1828.

Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., & Campilho, A. (2018). End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, *37*, 781–791.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision (ICCV)*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Fernando, B., Bilen, H., Gavves, E., & Gould, S. (2017). Self-supervised video representation learning with odd-one-out networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27 – 48.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2018a). Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)*.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2018b). Retinal image understanding emerges from self-supervised multimodal reconstruction. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Hoover, A. D., Kouznetsova, V., & Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, *19*, 203–210.

Jamaludin, A., Kadir, T., & Zisserman, A. (2017). Spinenet: Automated classification and evidence visualization in spinal mris. *Medical Image Analysis*, *41*, 63 – 73.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Lindeberg, T. (1998). Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, *30*, 117–156.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60 – 88.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M. J., & ADNI (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*, 1132–1140.

Lotter, W., Kreiman, G., & Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations (ICLR)*.

Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision (ECCV)*.

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*.

Ortega, M., Penedo, M. G., Rouco, J., Barreira, N., & Carreira, M. J. (2009). Retinal verification using a feature points-based biometric pattern. *EURASIP Advances in Signal Processing*, *2009*.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., & Torralba, A. (2016). Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision (ECCV)*.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Patterson, G., & Hays, J. (2016). COCO attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision (ECCV)*.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., & Raiko, T. (2015). Semi-supervised learning with ladder networks. In *International Conference on Neural Information Processing Systems (NIPS)*.

Ronneberger, O., P.Fischer, & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., & Maier-Hein, L. (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, *13*, 925–933.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, *abs/1706.05098*.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., & Levine, S. (2018). Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, .

Shin, H. C., Orton, M. R., Collins, D. J., Doran, S. J., & Leach, M. O. (2013). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1930–1943.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*, 1285–1298.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

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

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*, 1299–1312.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., & Padoy, N. (2017). Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, *36*, 86–97.

Urban, G., Geras, K. J., Kahou, S. E., Aslan, O., Wang, S., Caruana, R., Mohamed, A., Philipose, M., & Richardson, M. (2017). Do deep convolutional nets really need to be deep and convolutional? In *International Conference on Learning Representations (ICLR)*.

Wang, W., Wang, N., Wu, X., You, S., & Neumann, U. (2017). Self-paced cross-modality transfer learning for efficient road segmentation. In *International Conference on Robotics and Automation (ICRA)*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., & Madabhushi, A. (2016). Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, *35*, 119–130.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European Conference on Computer Vision (ECCV)*.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, *3*, 47–57.

2.4 Book Chapter: Multimodal reconstruction of retinal images over unpaired datasets using cyclical generative adversarial networks

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

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies,
University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Book: Generative Adversarial Networks for Image-to-Image Translation
Pages: 347–376
Editorial: Elsevier
ISBN: 9780128235195
Year: 2021

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], Marcos Ortega[a,b]

[a] *Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain*
[b] *VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña, Spain*

Abstract

Nowadays, the most successful approaches for image-to-image translation are those based on the use of generative adversarial networks (GANs). These novel deep learning frameworks represent a reference technique for learning generative models. In particular, GANs allow the training of image-to-image translation tasks using unpaired data, which enables the use of these approaches in numerous application domains where the paired data is difficult to obtain. Nevertheless, in medical imaging the paired data can be easily gathered due to the common use of complementary imaging techniques in the modern clinical practice. For instance, the availability of paired data has been successfully exploited for the multimodal reconstruction of retinal images, which consists in an image-to-image translation between complementary retinal imaging modalities. In this context, the multimodal reconstruction does not only provide an estimate of an additional modality, but it also allows to learn relevant retinal patterns that are useful for transfer learning purposes.

In this chapter, the use of GANs for the multimodal reconstruction of retinal images is studied. In particular, we present a cyclical GAN methodology that allows the training of the multimodal reconstruction using unpaired data. In this regard, despite the ease of gathering the paired retinal images, taking advantage of them still requires the alignment of the different image pairs. This alignment represents a challenging task in itself, which can compromise the actual availability of paired training data. The presented methodology avoids this issue by leveraging the high modeling capacity of GANs. In order to provide a comprehensive analysis of the presented approach, an exhaustive comparison against the state-of-the-art methodology for the multimodal reconstruction of retinal images is presented. This latter approach, which relies on the availability of paired training data, does not use GANs. Therefore, the provided comparison will directly highlight the advantages and disadvantages of using GANs for the multimodal reconstruction. Additionally, the presented experiments allow to analyze the extent to which the use of GANs compensates for the lack of paired training data.

Keywords: medical imaging, cyclical GANs, deep learning, retinal imaging

1. Introduction

The recent rise of deep learning has revolutionized medical imaging, making a significant impact in 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.

[*] Corresponding author
Email address: a.suarezh@udc.es (Álvaro S. Hervella)

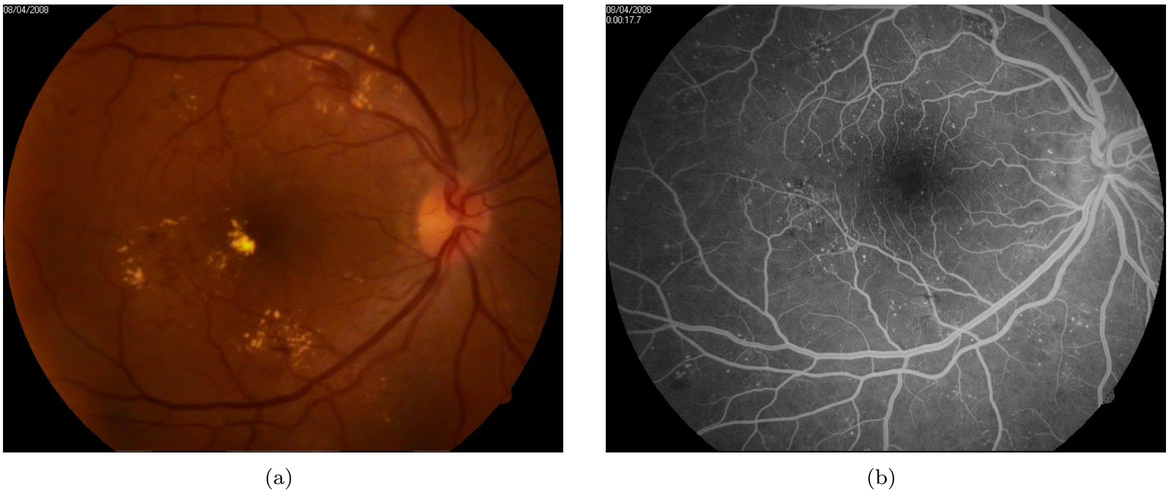
For instance, many diseases in their early stages are only evidenced by very small lesions or subtle anomalies. In these scenarios, factors such as the clinicians' expertise and workload can affect the reliability of the final analysis. Thus, the use of deep learning algorithms allows to accelerate the 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) has 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 demonstrated to be able to further exploit the capacity of the DNNs.

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, but also 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 pre-training 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 to train 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 Figure 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. 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



(a) (b)

Figure 1: Example of retinography and fluorescein angiography for the same eye. (a) Retinography. (b) Angiography.

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 pre-processing for the alignment of the different image pairs.

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 ground-breaking 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 in 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 to introduce a cycle-consistency loss whereby the subsequent application of

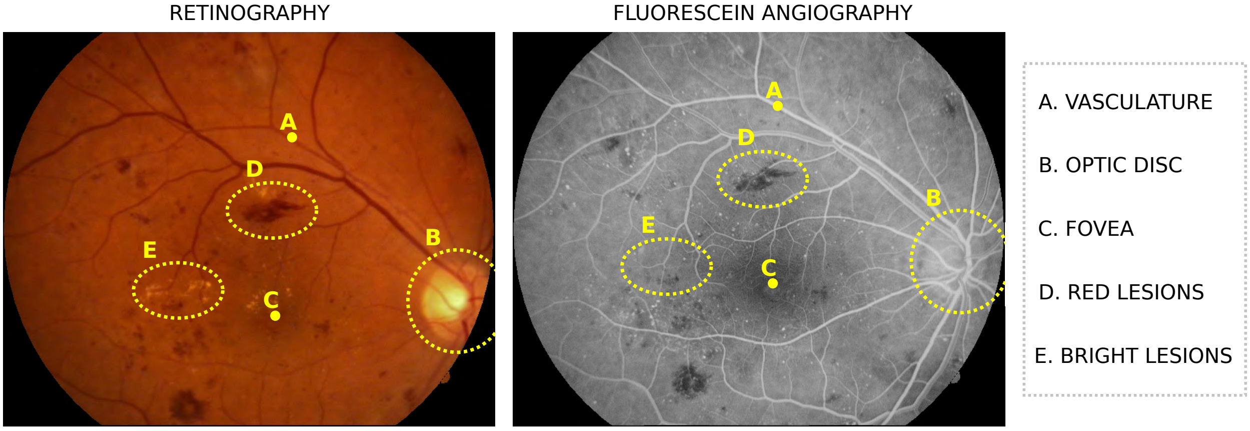


Figure 2: Example of retinography and fluorescein angiography for the same eye. The included images depict the main anatomical structures as well as the two main types of lesions in the retina.

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, it has been common the adaption of those state-of-the-art approaches that already demonstrated a good performance in natural images. 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 or 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 that, for many applications in medical imaging, it is not necessary to generate realistic or aesthetically pleasing images. In this context, the results obtained in [15] show that, without the use of GANs, the generated images lack realism and can be easily identified as synthetic samples.

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 reference, the transformation between retinography and angiography for the main anatomical and pathological structures in the retina can be visualized in Figure 2.

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 $G_{\mathcal{R}2\mathcal{A}} : \mathcal{R} \rightarrow \mathcal{A}$ that given a certain retinography $r \in \mathcal{R}$ returns the corresponding angiography $a = G_{\mathcal{R}2\mathcal{A}}(r) \in \mathcal{A}$ for the same eye. In this scenario, the mapping function $G_{\mathcal{R}2\mathcal{A}}$ can be parameterized by a DNN. Thus, the function parameters can be learned by applying the adequate training strategy. In this regard, we present two different deep learning-based approaches for learning the mapping function $G_{\mathcal{R}2\mathcal{A}}$, the Cyclical GAN methodology [11] and the Paired SSIM methodology [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, they 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 retinographies and 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 to train 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 [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 $G_{\mathcal{R}2\mathcal{A}}$ and its inverse mapping function $G_{\mathcal{A}2\mathcal{R}} : \mathcal{A} \rightarrow \mathcal{R}$ that given a certain angiography $a \in \mathcal{A}$ produces a retinography $r = G_{\mathcal{A}2\mathcal{R}}(a) \in \mathcal{R}$ of the same eye. Then, the

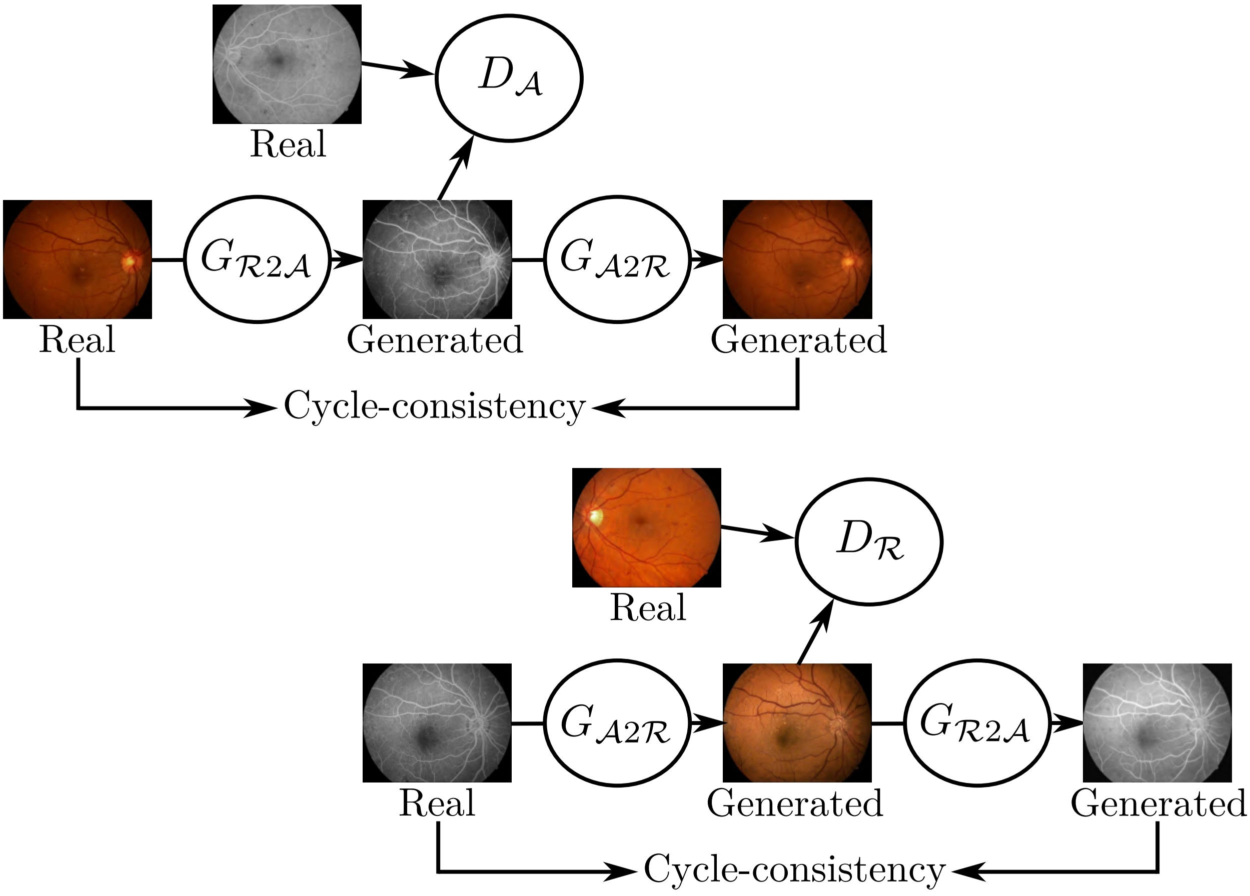


Figure 3: Flowchart for the complete training procedure in the Cyclical GAN methodology. This approach involves the use of two complementary training cycles that only differ in which imaging modality is being used as input and which one as target. For each training cycle, the appearance of the target 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.

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 ($\mathcal{R}2\mathcal{A}2\mathcal{R}$) and (2) from angiography to retinography to angiography ($\mathcal{A}2\mathcal{R}2\mathcal{A}$). A flowchart showing the complete training procedure is depicted in Figure 3. It is observed that two different generators, $G_{\mathcal{R}2\mathcal{A}}$ and $G_{\mathcal{R}2\mathcal{A}}$, and two different discriminators, $D_{\mathcal{A}}$ and $D_{\mathcal{R}}$, are used during the training. The discriminators $D_{\mathcal{A}}$ and $D_{\mathcal{R}}$ are trained to distinguish between generated and real images. Simultaneously, the generators $G_{\mathcal{R}2\mathcal{A}}$ and $G_{\mathcal{R}2\mathcal{A}}$ 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:

$$\mathcal{L}_{D_{\mathcal{A}}}^{adv} = \mathbb{E}_{r\sim\mathcal{R}}[D_{\mathcal{A}}(G_{\mathcal{R}2\mathcal{A}}(r))^2] + \mathbb{E}_{a\sim\mathcal{A}}[(D_{\mathcal{A}}(a) - 1)^2] \tag{1}$$

$$\mathcal{L}_{D_{\mathcal{R}}}^{adv} = \mathbb{E}_{a\sim\mathcal{A}}[D_{\mathcal{R}}(G_{\mathcal{A}2\mathcal{R}}(a))^2] + \mathbb{E}_{r\sim\mathcal{R}}[(D_{\mathcal{R}}(r) - 1)^2] \tag{2}$$

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

$$\mathcal{L}_{G_{\mathcal{R}2\mathcal{A}}}^{adv} = \mathbb{E}_{r\sim\mathcal{R}}[(D_{\mathcal{A}}(G_{\mathcal{R}2\mathcal{A}}(r)) - 1)^2] \tag{3}$$

$$\mathcal{L}_{G_{\mathcal{A}2\mathcal{R}}}^{adv} = \mathbb{E}_{a\sim\mathcal{A}}[(D_{\mathcal{R}}(G_{\mathcal{A}2\mathcal{R}}(a)) - 1)^2] \tag{4}$$

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

$$\mathcal{L}^{cyc} = \mathbb{E}_{r\sim\mathcal{R}}[||G_{\mathcal{A}2\mathcal{R}}(G_{\mathcal{R}2\mathcal{A}}(r)) - r||_1] + \mathbb{E}_{a\sim\mathcal{A}}[||G_{\mathcal{R}2\mathcal{A}}(G_{\mathcal{A}2\mathcal{R}}(a)) - a||_1] \tag{5}$$

As it can be observed in previous equations as well as in Figure 3, there is a strong parallelism between both training cycles, $\mathcal{R}2\mathcal{A}2\mathcal{R}$ and $\mathcal{A}2\mathcal{R}2\mathcal{A}$. 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 as target.

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

$$\mathcal{L} = \mathcal{L}_{G_{\mathcal{R}2\mathcal{A}}}^{adv} + \mathcal{L}_{D_{\mathcal{A}}}^{adv} + \mathcal{L}_{G_{\mathcal{A}2\mathcal{R}}}^{adv} + \mathcal{L}_{D_{\mathcal{A}}}^{adv} + \lambda\mathcal{L}^{cyc} \tag{6}$$

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 [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 [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 in 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.

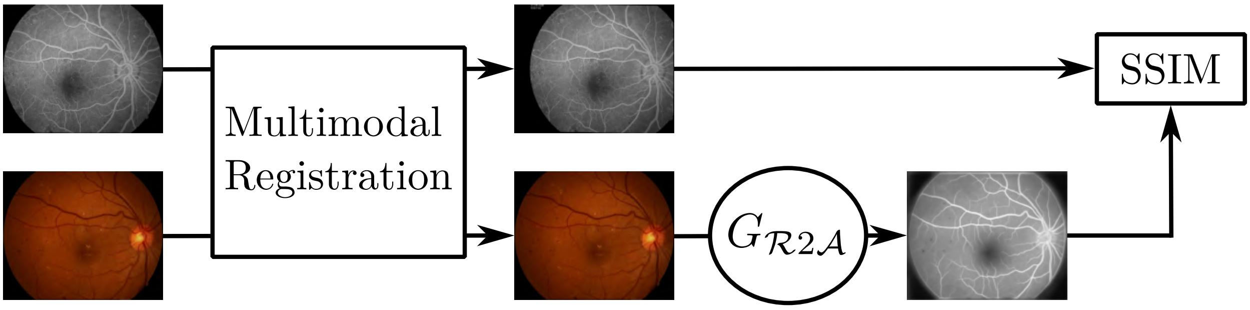


Figure 4: Flowchart for the complete training procedure of the Paired SSIM methodology. The first step is the multimodal registration of the paired retinal images, which can be performed off-line before the actual network training. Then, the training feedback is provided by the Structural Similarity (SSIM), which is a pixel-wise similarity metric.

3.2. Paired SSIM methodology

An alternative methodology for the multimodal reconstruction between retinography and angiography was proposed in [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 of the 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 the use of more complex or invasive techniques, such as, e.g., those requiring the injection of contrasts. This is the case of the retinography and the angiography in retinal imaging. While the retinography is a broadly extended modality, typically used in screening programs, the angiography is only used when it is clearly required. However, each time the angiography is taken for a patient, the 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 to directly compare 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 to generate a paired and registered multimodal dataset, which is used for directly training the generator network $G_{\mathcal{R}2\mathcal{A}}$. The complete methodology for training the multimodal reconstruction is depicted in Figure 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 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 information. The measurement is performed at a local level considering a

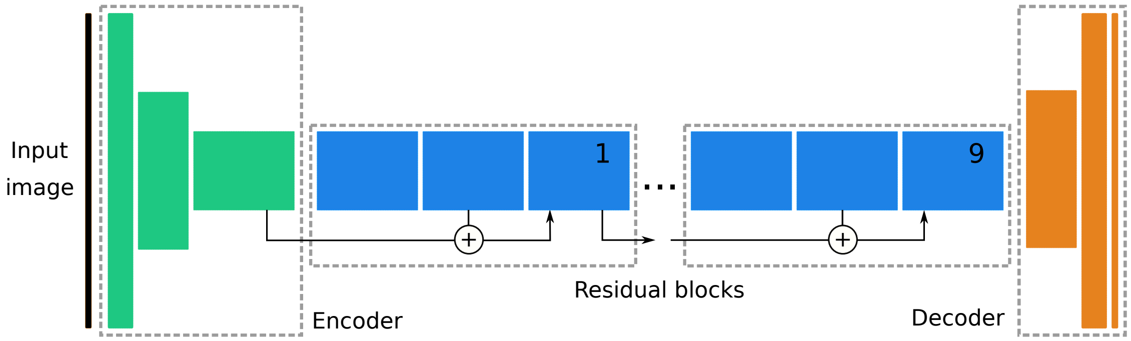


Figure 5: Diagram of the network architecture for the generator. Each colored block represents the output of a layer in the neural network. The width of the blocks represents the number of channels whereas the height represents the spatial dimensions. The details of the different layers are in Table 1

small neighborhood for each pixel. In particular, a SSIM map between two images (x, y) is computed with a set of local statistics as:

$$SSIM(x, y) = \sum \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{7}$$

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 $G_{\mathcal{R}2\mathcal{A}}$ is defined by using the negative SSIM:

$$\mathcal{L}^{SSIM} = \mathbb{E}_{r,a\sim(\mathcal{R},\mathcal{A})}[-SSIM(G_{\mathcal{R}2\mathcal{A}}(r), a)] \tag{8}$$

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 Ba [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 $1,250$ iterations. Finally, the training is early stopped after $5,000$ 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 in 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.

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

Table 1: Building blocks of the generator architecture. Conv is convolution, IN is instance normalization [33], and ConvT is Convolution Transpose.

Block	Layers	Kernel	Stride	Out features
Encoder	Conv / IN / ReLU	7x7	1	64
	Conv / IN / ReLU	3x3	2	128
	Conv / IN / ReLU	3x3	2	256
Residual	Conv / IN / ReLU	3x3	1	256
	Conv / IN	3x3	1	256
	Residual Addition	-	-	256
Decoder	ConvT / IN / ReLU	3x3	2	128
	ConvT / IN / ReLU	3x3	2	64
	Conv / IN / ReLU	7x7	1	Image channels

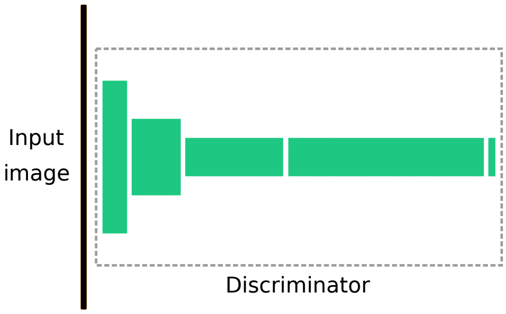


Figure 6: Diagram of the network architecture for the discriminator. Each colored block represents the output of a layer in the network. The width of the blocks represents the number of channels whereas the height represents the spatial dimensions. The details of the different layers are in Table 2.

that are presented in this chapter are performed with the same network architectures that were previously used in [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 Figure 5 and Table 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 to keep an adequate level of spatial accuracy without the 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 [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 [19]. In particular, the discriminator is a fully convolutional neural network, which allows to work 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 Figure 6 and Table 2, respectively. The characteristics of the different layers are similar to

Table 2: Layers of the discriminator architecture. Conv is convolution and IN is instance normalization.

Layers	Kernel	Stride	Out features
Conv / Leaky ReLU	4x4	2	64
Conv / IN / Leaky ReLU	4x4	2	128
Conv / IN / Leaky ReLU	4x4	2	256
Conv / IN / Leaky ReLU	4x4	1	512
Conv	4x4	1	1

those in the generator network. The main difference is the use of Leaky ReLU instead of ReLU as activation function, which has 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.

4. Experiments and results

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 pixels. 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 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 9 images 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 3.

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

4.2. Qualitative evaluation of the reconstruction

Firstly, the quality and coherence of the generated angiographies is evaluated through visual analysis. To that end, Figures 7 and 8 depict some representative examples of generated images together with the original retinographies and angiographies. The examples are taken from the hold out 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 vasculature is successfully enhanced in all the cases, which is one of the main characteristics of the 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 information such as, e.g., the color. In contrast with the vasculature, the reconstructed optic discs are not as similar as those in the real

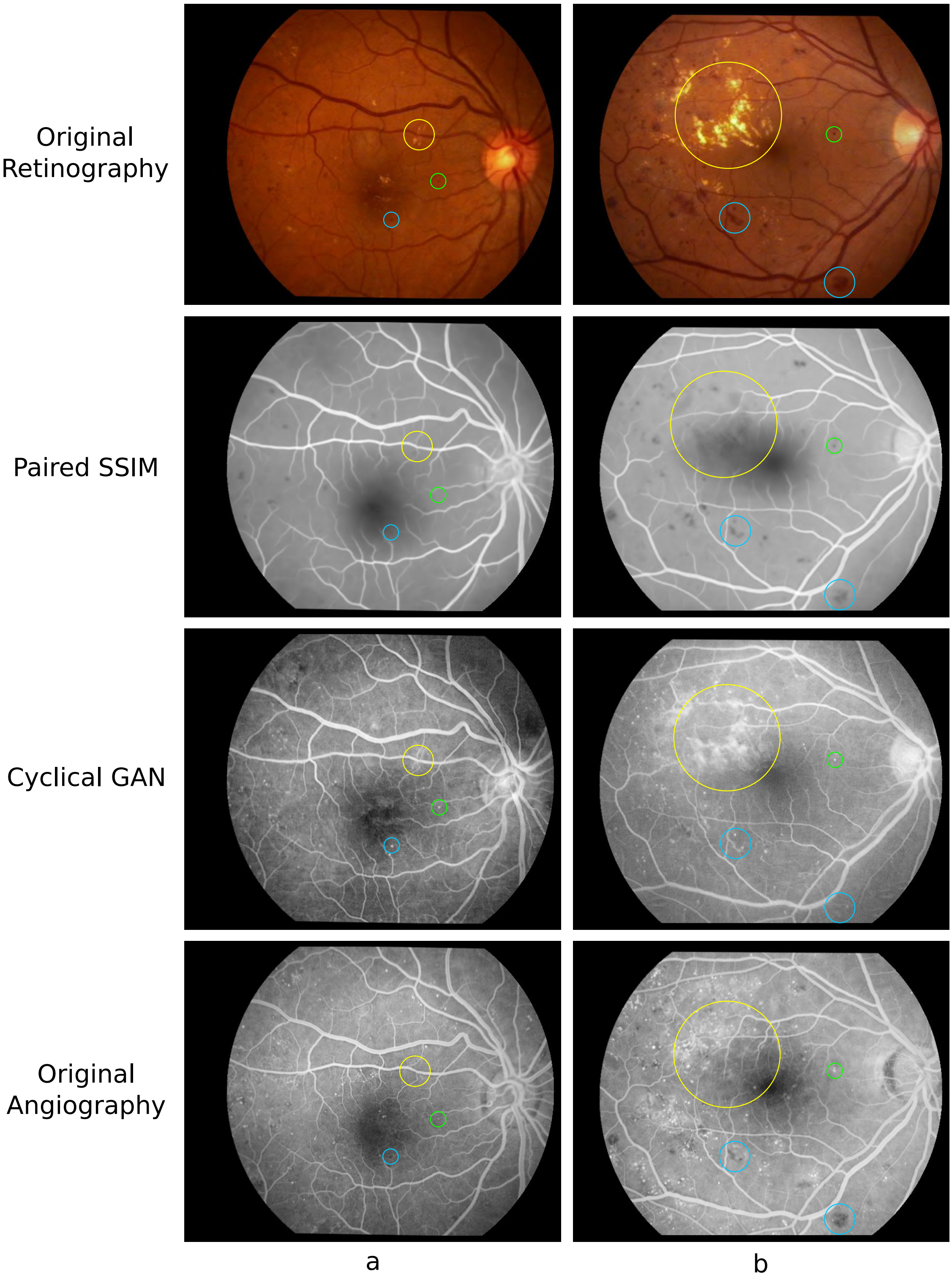


Figure 7: Examples of generated angiographies together with the corresponding original retinographies and angiographies. Some representative examples of microaneurysms (green), microhemorrhages (blue), and bright lesions (yellow) are marked with circles.

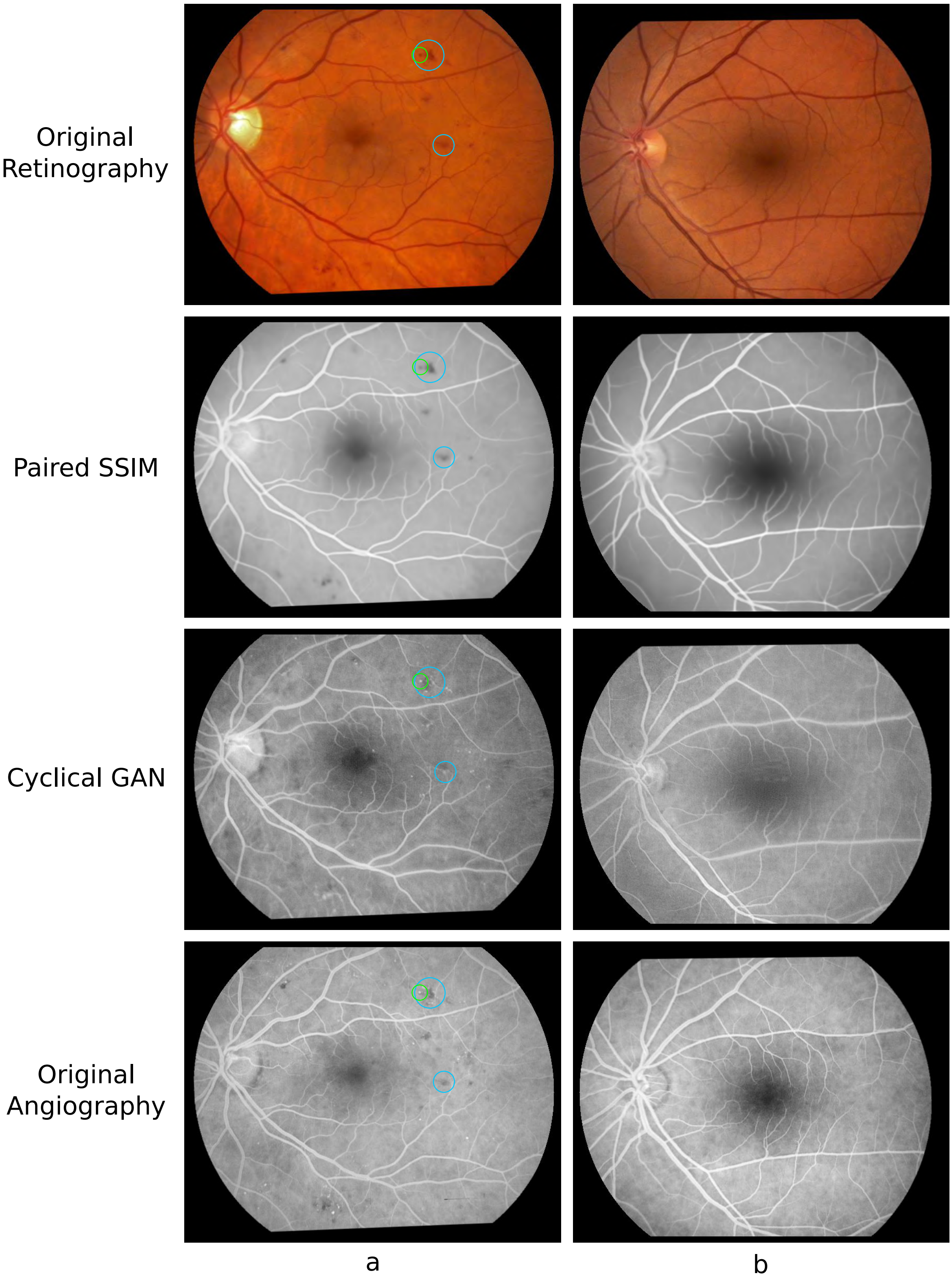


Figure 8: Examples of generated angiographies together with the corresponding original retinographies and angiographies. Some representative examples of microaneurysms (green) and microhemorrhages (blue) are marked with circles.

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 slighter 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 Figure 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 this 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 Figure 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 in Figure 7(b). The Paired SSIM approach provides a more accurate reconstruction regarding these kind of lesions, although in the previous case there is still a remaining shadow 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 [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 Figures 7 and 8.

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 hold out 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 a 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 Figure 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 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 higher level information such as the 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.

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 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 applying some common image processing techniques. First, the Multi-Scale Laplacian operator proposed in [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

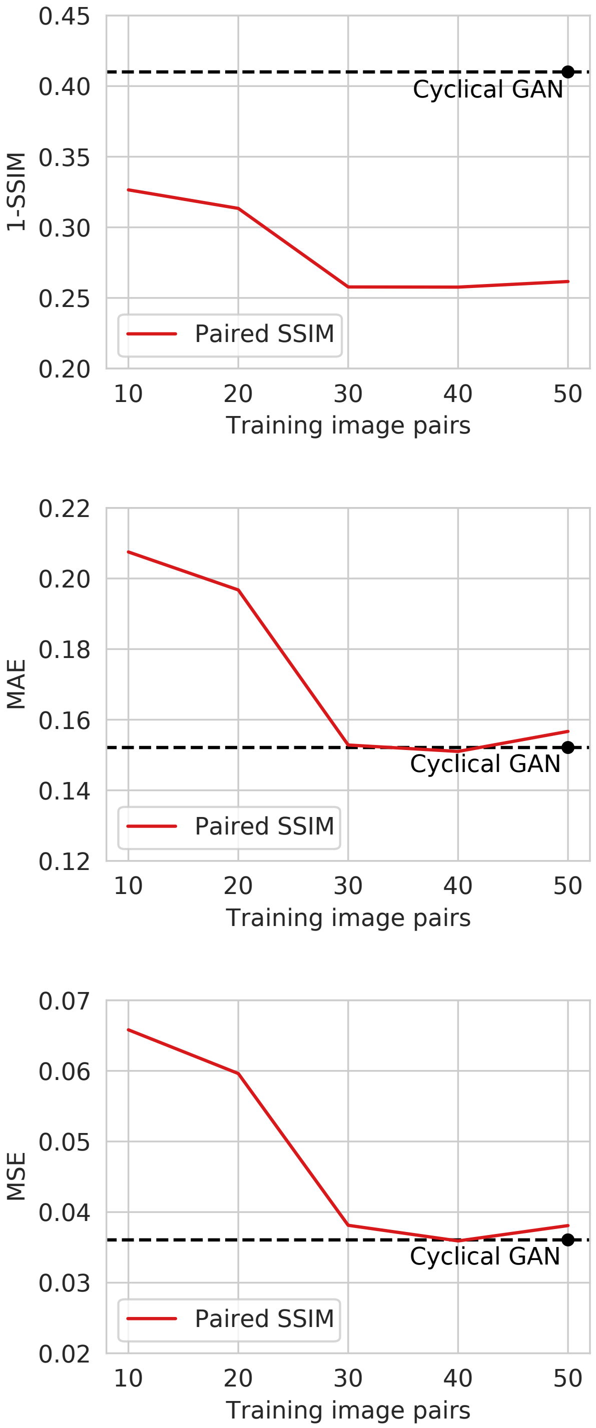


Figure 9: Comparison of Cyclical GAN and Paired SSIM with a varying number of training samples for Paired SSIM. The evaluation is performed by means of (a) SSIM, (b) MAE, and (c) MSE.

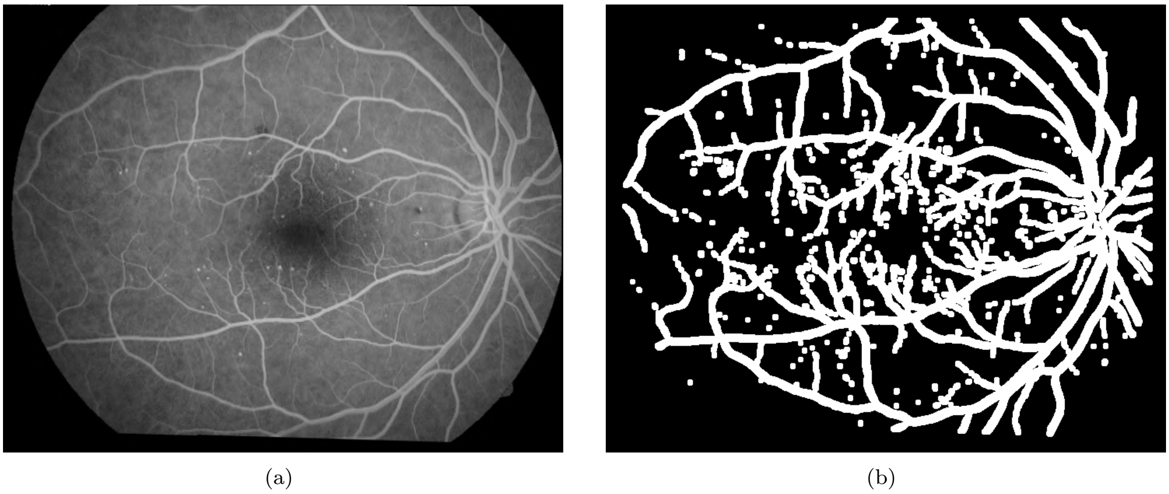


Figure 10: Example of vascular mask used for evaluation. (a) Angiography. (b) Resulting vessel mask for (a).

binarized by applying the Otsu's thresholding method [37]. An example of produced binary vascular mask together with the original angiography is depicted in Figure 10.

The results of the quantitative evaluation using the computed vascular masks are depicted in Figure 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 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 it can be seen in Figure 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 Figure 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 the adequate tone, as Paired SSIM does. This explains why the use of these metrics as loss function does not incentive 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.

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.

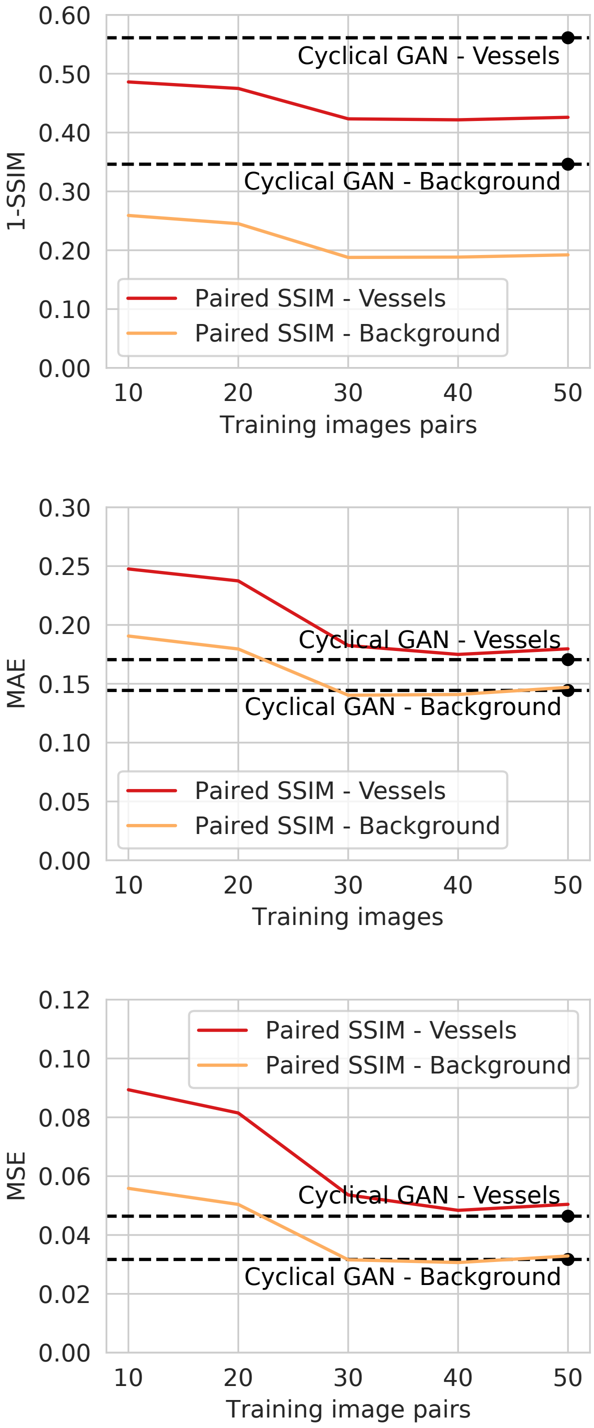
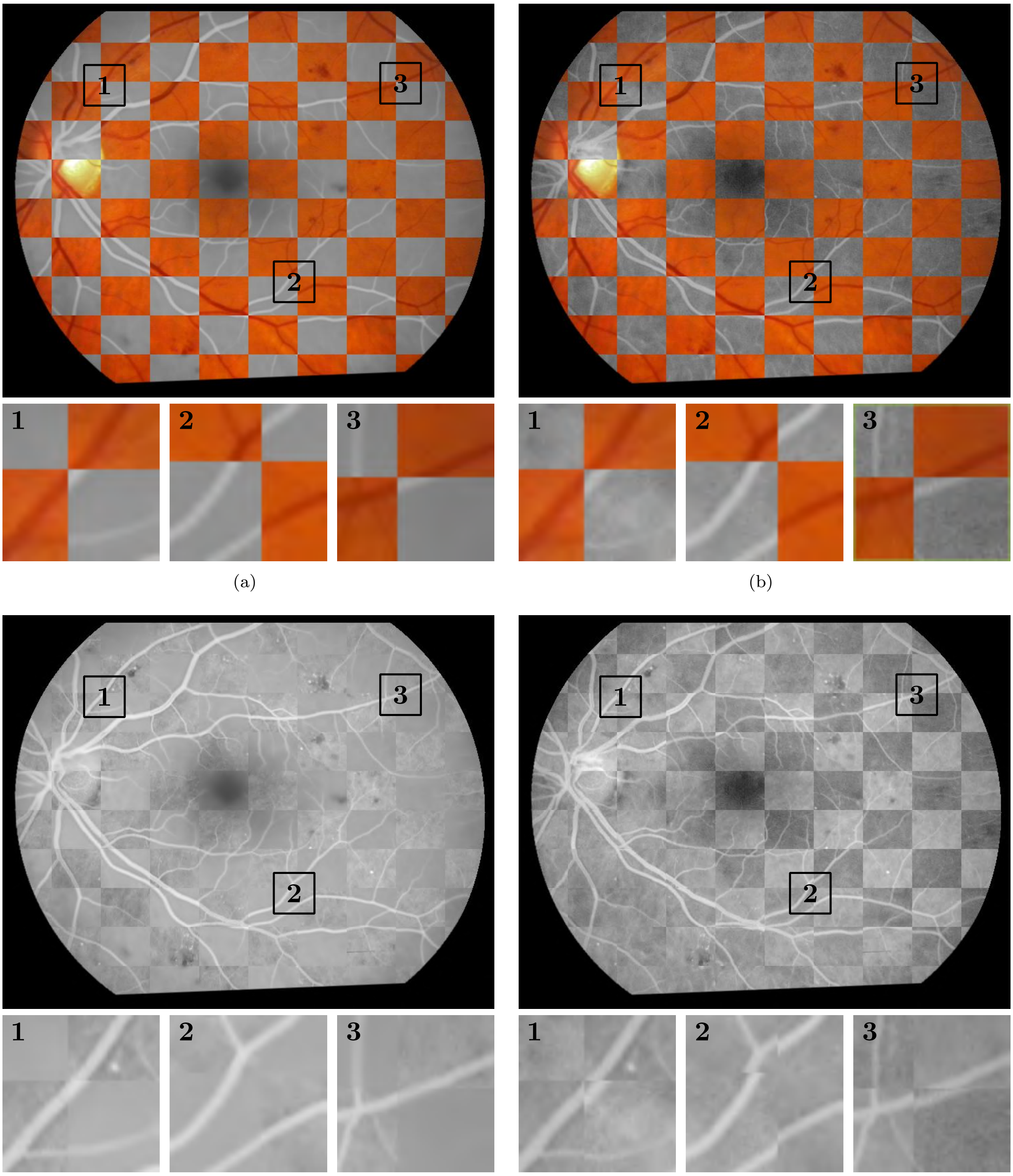


Figure 11: Comparison of Cyclical GAN and Paired SSIM with a varying number of training samples for Paired SSIM. The evaluation is conducted independently for vessels and background of the images. The evaluation is performed by means of (a) SSIM, (b) MAE, and (c) MSE.



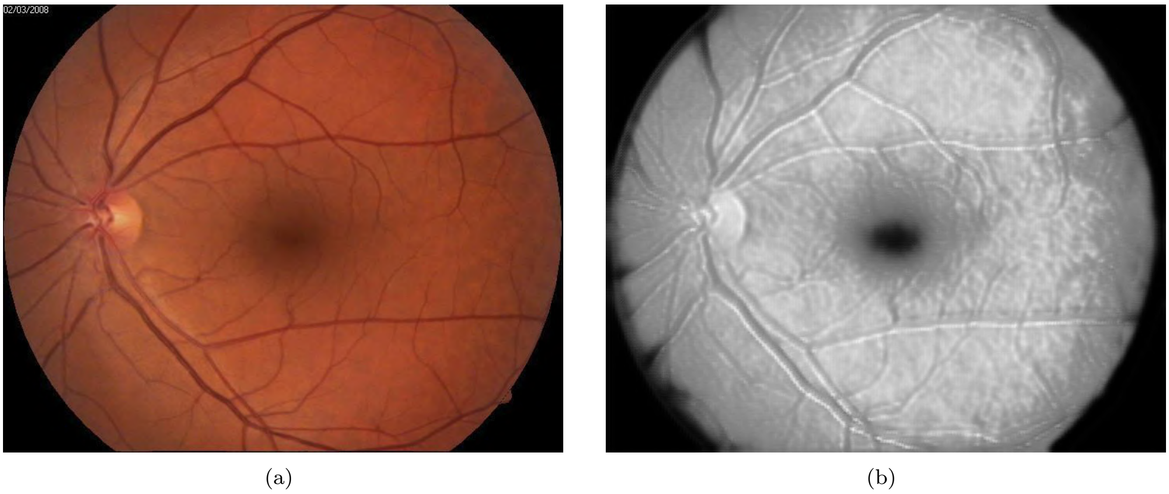
(a)

(b)

(c)

(d)

Figure 12: Comparison of generated angiographies against ((a),(b)) the corresponding original retinographies and ((c),(d)) the corresponding ground truth angiographies. ((a),(c)) Angiography generated using Paired SSIM. ((b),(d)) Angiography generated using Cyclical GAN. Additionally, cropped regions are depicted in detail for each case.



(a)

(b)

Figure 13: Representative example of generated angiography on the first stages of training for Cyclical GAN. (a) Original retinography. (b) Generated angiography.

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. Figure 12 depicts some composite images using a checkerboard pattern that are used to perform the visual analysis. In particular, the depicted images show the generated angiography together with the original retinography (Figures 12(a) and 12(b)) as well as the generated angiography together with the ground truth angiography (Figures 12(c) and 12(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 Figure 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 $G_{\mathcal{R}2\mathcal{A}}$ applies any spatial transformation to the generated angiographies, then $G_{\mathcal{A}2\mathcal{R}}$ 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, $G_{\mathcal{R}2\mathcal{A}}$, 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 Figure 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 under-constrained training setting of Cyclical GAN.

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 in 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 the 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, the multimodal reconstruction has already demonstrated to be a successful pre-training 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 the 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 the 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 [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, it 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 can not 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 4 different neural networks and 2 training cycles, as described in Section 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 4 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 Figures 7 and

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 Figure 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 towards 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 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, the 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 under-constrained training setting in cyclical GANs. Nevertheless, in this chapter, we have presented an empirical evidence of this issue in the form of small displacements for the reconstructed blood vessels. According to the evidence presented in Section 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 incoherences in the blood vessels, it would be possible 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 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 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 of 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 is supported by Instituto de Salud Carlos II, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project, and by Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the RTI2018-095894-B-I00 research project. The authors of this work also receive financial support from the ERDF and European Social Fund (ESF) of the EU and Xunta de Galicia through Centro de Investigación de Galicia, ref. ED431G 2019/01, and the predoctoral grant contract ref. ED481A-2017/328.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60 – 88, doi: 10.1016/j.media.2017.07.005.

[2] E. D. Cole, E. A. Novais, R. N. Louzada, N. K. Waheed, Contemporary retinal imaging techniques in diabetic retinopathy: a review, Clinical & Experimental Ophthalmology 44 (4) (2016) 289–299, doi:10.1111/ceo.12711.

[3] T. Farncombe, K. Iniewski, Medical Imaging: Technology and Applications, CRC Press, 2017.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27 – 48, doi:10.1016/j.neucom.2015.09.116.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Applied Soft Computing 70 (2018) 41–65, ISSN 15684946, doi:10.1016/j.asoc.2018.05.018.

[6] W. Rawat, Z. Wang, Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, Neural Computation 29 (9) (2017) 2352–2449, doi:10.1162/neco_a_00990.

[7] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-030-00928-1_37, 2018.

[8] S. Engelhardt, L. Sharan, M. Karck, R. D. Simone, I. Wolf, Cross-Domain Conditional Generative Adversarial Networks for Stereoscopic Hyperrealism in Surgical Training, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, doi:10.1007/978-3-030-32254-0_18, 2019.

[9] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Medical Image Analysis 58 (2019) 101552, ISSN 1361-8415, doi:10.1016/j.media.2019.101552.

[10] Á. S. Hervella, J. Rouco, J. Novo, M. Ortega, Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction, Applied Soft Computing 91 (2020) 106210, doi:10.1016/j.asoc.2020.106210.

[11] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Deep Multimodal Reconstruction of Retinal Images Using Paired or Unpaired Data, in: International Joint Conference on Neural Networks (IJCNN), doi:10.1109/IJCNN.2019.8852082, 2019.

[12] L. Wang, W. Chen, W. Yang, F. Bi, F. R. Yu, A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks, IEEE Access 8 (2020) 63514–63537, doi:10.1109/ACCESS.2020.2982224.

[13] A. S. Hervella, L. Ramos, J. Rouco, J. Novo, M. Ortega, Multi-Modal Self-Supervised Pre-Training for Joint Optic Disc and Cup Segmentation in Eye Fundus Images, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), doi:10.1109/ICASSP40776.2020.9053551, 2020.

[14] J. Morano, A. S. Hervella, N. Barreira, J. Novo, J. Rouco, Multimodal Transfer Learning-based Approaches for Retinal Vascular Segmentation, in: 24th European Conference on Artificial Ingelligence (ECAI), 2020.

[15] Á. S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Systems with Applications (2020) 113674doi:https://doi.org/10.1016/j.eswa.2020.113674.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in Neural Information Processing Systems (NIPS) 27, 2672–2680, 2014.

[17] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, Y. Zheng, Recent Progress on Generative Adversarial Networks (GANs): A Survey, IEEE Access 7 (2019) 36322–36333, doi:10.1109/ACCESS.2019.2905015.

[18] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), doi:10.1109/CVPR.2017.632, 2017.

[19] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, in: 2017 IEEE International Conference onComputer Vision (ICCV), doi:10.1109/ICCV.2017.244, 2017.

[20] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: Unsupervised Dual Learning for Image-To-Image Translation, in: The IEEE International Conference on Computer Vision (ICCV), doi:10.1109/ICCV.2017.310, 2017.

[21] T. Kim, M. Cha, H. Kim, J. K. Lee, J. Kim, Learning to Discover Cross-Domain Relations with Generative Adversarial Networks, in: Proceedings of the 34th International Conference on Machine Learning, vol. 70, 1857–1865, 2017.

[22] J. M. Wolterink, T. Leiner, M. A. Viergever, I. Išgum, Generative Adversarial Networks for Noise Reduction in Low-Dose CT, IEEE Transactions on Medical Imaging 36 (12) (2017) 2536–2545, doi:10.1109/TMI.2017.2708987.

[23] Y. Xue, T. Xu, H. Zhang, L. Long, X. Huang, Neuroinformatics 16 (3-4) (2018) 383–392, ISSN 1539-2791, doi:10.1007/s12021-018-9377-x.

[24] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321 – 331, ISSN 0925-2312, doi:10.1016/j.neucom.2018.09.013.

[25] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks, Medical Image Analysis 54 (2019) 30 – 44, ISSN 1361-8415, doi:10.1016/j.media.2019.01.010.

[26] J. Cohen, M. Luck, S. Honari, Distribution Matching Losses Can Hallucinate Features in Medical Image Translation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-030-00928-1_60, 2018.

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least Squares Generative Adversarial Networks, in: The IEEE International Conference on Computer Vision (ICCV), doi:10.1109/ICCV.2017.304, 2017.

[28] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: International Conference on Learning Representations (ICLR), 2015.

[29] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, in: 4th International Conference on Learning Representations, ICLR 2016, 2016.

[30] Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Computer Methods and Programs in Biomedicine 186 (2020) 105201, ISSN 0169-2607, doi:https://doi.org/10.1016/j.cmpb.2019.105201.

[31] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal Registration of Retinal Images Using Domain-Specific Landmarks and Vessel Enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), doi:10.1016/j.procs.2018.07.213, 2018.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612, doi:10.1109/TIP.2003.819861.

[33] D. Ulyanov, A. Vedaldi, V. S. Lempitsky, Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 4105–4113, doi:10.1109/CVPR.2017.437, 2017.

[34] O. Ronneberger, P.Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-319-24574-4_28, 2015.

[35] S. H. M. Alipour, H. Rabbani, M. R. Akhlaghi, Diabetic Retinopathy Grading by Digital Curvelet Transform, Computational and Mathematical Methods in Medicine 2012, doi:10.1155/2012/761901.

[36] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data, in: International Joint Conference on Neural Networks (IJCNN), doi:10.1109/IJCNN.2019.8851844, 2019.

[37] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man and Cybernetics 9 (1) (1979) 62–66, doi:10.1109/TSMC.1979.4310076.

Chapter 3

Retinal Image Understanding - Published Papers

3.1 Conference Paper: Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data

Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Conference: 2019 International Joint Conference on Neural Networks - IJCNN 2019
Location: Budapest - Hungary
Date: July 14-19, 2019
Pages: 1–8
Editorial: IEEE
ISSN: 2161-4407

This is a post-peer-review, pre-copyedit version of an article published in Proceedings of the 2019 International Joint Conference on Neural Networks. The final authenticated version is available online at: https://doi.org/10.1109/IJCNN.2019.8851844.

Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data

Álvaro S. Hervella*†, José Rouco*†, Jorge Novo*†, Marcos Ortega*†

*CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña
†Department of Computer Science, Universidade da Coruña
A Coruña, Spain
Email: {a.suarezh, jrouco, jnovo, mortega}@udc.es

***Abstract*—This paper presents a novel approach that allows training convolutional neural networks for retinal vessel segmentation without manually annotated labels. In order to learn how to segment the retinal vessels, convolutional neural networks are typically trained with a set of pixel-level labels annotated by a clinical expert. This annotation is a tedious and error-prone task that limits the number of available training samples.**

To alleviate this problem, we propose the use of unlabeled multimodal data for learning about the retinal vasculature. Instead of using manually annotated labels, the networks learn to segment the retinal vessels from a complementary image modality where the vasculature is already highlighted. In this complementary modality, a vessel map can be easily constructed with simple image processing techniques. Then, a convolutional neural network is trained to learn the cross-modal mapping from the original modality to the automatically derived vessel maps. Using this strategy, the supervisory signal for training is automatically obtained from the unlabeled multimodal data. Thus, the number of training samples can be increased without any human annotation effort.

Several experiments were conducted to evaluate the performance of the networks that were trained with the automatically derived labels, obtaining competitive results for retinal vessel segmentation in relevant public datasets. Furthermore, the results are promising towards including the presented approach in semi-supervised methods.

I. Introduction

Retinal vessel segmentation plays a fundamental role in the automatic analysis of eye fundus images. In particular, the vasculature is used as reference for the detection of other anatomical or pathological structures [1]. Furthermore, as the retinal vessels are part of the vascular system, their analysis is not only useful for the analysis of opthalmic diseases but also for the analysis of relevant systemic diseases such as hypertension or diabetes, among others [1].

Due to its importance, the automatic segmentation of the retinal vessels has been studied in several works, applying a variety of methodologies [2]. Lately, there is a trend with many works using Convolutional Neural Networks (CNNs), that are trained on a set of manually annotated vessel labels [3], [4]. In contrast with traditional approaches, which rely on the use of hand-crafted features, CNNs are capable of learning the required features from the raw input images. The end-to-end training of these networks led to an increase in performance over previously studied approaches [5]. The application of deep learning methodologies implies, nevertheless, the limitation of requiring a large amount of annotated data.

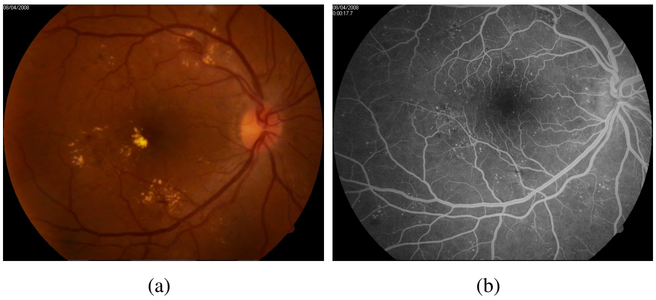

(a) (b)

Fig. 1. Example of retinography and fluorescein angiography from the same eye. (a) Retinography. (b) Angiography.

The concern for the scarcity of annotated data is not limited to medical imaging, which is reflected in the numerous recent works proposing strategies for training CNNs without manual annotations [6]. It is true, nevertheless, that medical annotations are harder to obtain due to the commonly required expertise in the field [7]. Also, some applications require the annotation of fine details in low contrast image regions, e.g., small vessels in eye fundus images as the scope of this work, which makes the manual annotation a difficult, tedious and time-consuming task.

In contrast with the scarcity of annotated data, large amounts of unlabeled images results from the routine clinical practice [8]. Among these unlabeled data, there are complementary image modalities representing the same organs and tissues. In ophthalmology, the classical retinography is the most commonly used image modality. Although the study of the retinal vasculature may require the use of the fluorescein angiography image modality, which is an invasive alternative. The angiography requires the injection of a contrast dye into the bloodstream, which increases the visibility of the blood vessels as well as vascular lesions [9]. An example of retinography and angiography for the same eye of a patient is depicted in Fig. 1. It is observed that, in the angiography, the vasculature is highlighted and more small vessels are visible. This facilitates the detection of the retinal vessels with common image processing techniques [10].

The multimodal data have been typically used as complementary sources of information for the image analysis algorithms. Nevertheless, in the context of CNNs, the multimodal data could be exploited in more creative ways to alleviate the scarcity of annotations. As example, the reconstruction between retinography and angiography has been recently presented as a potential tool for gaining domain-specific knowledge [11]. Besides this multimodal reconstruction task, we argue that the multimodal data can be used to directly learn target representations derived from a complementary modality.

In this work, multimodal sets are used for learning to generate retinal vessel representations from retinography without any manually annotated label. This is achieved using angiography-derived vessel maps as targets for training CNNs. These angiography-derived labels are automatically obtained with basic edge detection filters. Therefore, the proposed task automatically obtains the supervisory signals from the originally unlabeled multimodal data, resulting in a self-supervised training. The trained networks are tested on public datasets of reference for retinal vessel segmentation, where only a single image modality is available. The conducted experiments demonstrate that training with the multimodal data helps with learning the required patterns for retinal vessel segmentation.

II. Methodology

The proposed self-supervised training consists in learning the mapping from retinography to angiography-derived vessel maps. The vessel maps are automatically obtained for each angiography with a Multiscale Laplacian (MSL) operation [10]. Then, these generated maps are aligned with the retinography to allow the supervised training of a CNN with common pixelwise metrics. After training, the CNN is able to estimate a vessel segmentation from retinography. The methodology is summarized in Fig. 2.

A. Vessel enhancement with Multiscale Laplacian

The vessel maps are obtained with a vessel enhancement operation applied over the angiography. In this image modality, blood vessels are represented as high intensity tubular regions surrounded by a low intensity background. The vessel regions are enclosed, therefore, between two edges with opposite intensity transitions (low-to-high and high-to-low). Convolving these images with a second-order derivative filter, e.g. a Laplacian, results in two peak responses for each of the vessel edges, one inside the vessel and other outside. The two peaks inside the vessels, of the same sign, may overlap depending on how well the scale of the Laplacian filter fits the vessel width. For the adequate combination of vessel width and Laplacian scale, both peaks will completely overlap and thus the whole vessel is enhanced.

Given that the vessel widths vary throughout the image, we apply the Laplacian at multiple scales, obtaining different responses that are later aggregated into a single vessel map. The scale of the Laplacian is controlled convolving with a Gaussian filter with variable sigma, given rise to a Laplacian of Gaussian filter (LoG). Examples of eye fundus images convolved with LoG filters at different scales are depicted in Fig. 3. The single-scale LoG responses are combined by taken the maximum across scales for each pixel of the angiography, resulting in a Multiscale Laplacian (MSL) map in which the vessels of all widths are enhanced. This is computed as:

$$MSL(\mathbf{a}) = max_{t \in S} \lceil t^2 LoG(\mathbf{a}; t) \rceil_{\emptyset} \qquad (1)$$

where $LoG(\mathbf{a}; t)$ denotes the result of applying a LoG filter of scale t to the angiography image $\mathbf{a}$, S is the set or scales considered for the Multiscale Laplacian and $\lceil \cdot \rceil_{\emptyset}$ is a halfwave rectification to avoid the negative peaks outside the vessels. The factor t^2 is applied to make the results of the LoG at different scales comparable [12].

B. Label alignment

A generated vessel map and its corresponding retinography are aligned using the methodology proposed in [10]. This multimodal registration methodology consists of two different steps. The first step aims at computing a coarse transformation that globally corrects most of the misalignment between retinography and angiography. For such purpose, vessel crossings and bifurcations in the eye fundus are used as landmarks. These domain-specific landmarks are extracted from both images [13], being the corresponding landmark pairs matched to compute an initial transformation between the images. The second step aims at refining the initial transformation, in order to achieve an adequate pixelwise correspondence between the retinography and the generated label. In this step, we apply the MSL to both the retinography and the angiography. Then, the Normalized Cross-Correlation (NCC) between both MSL representations is maximized using a gradient ascent algorithm over the transformation parameter space. The MSL of the retinography is obtained using the intensity inverted grayscale retinography as input. Although both representations differ in the level of noise and detail for the vasculature, the maximization of a similarity measure between them has demonstrated to produce an adequate registration [10]. Fig. 4 depicts an example of retinography and generated vessel map before and after the alignment.

C. Self-supervised training for retinal vessel segmentation

An automatically annotated set $\{(\mathbf{r}, \mathbf{v_a})_1, ..., (\mathbf{r}, \mathbf{v_a})_N\}$, where $\mathbf{r}$ denotes retinography and $\mathbf{v_a} = MSL(\mathbf{a})$, is used for training a CNN. The network training is approached in two different ways: as a regression and as a classification problem.

For the classification approach, $\mathbf{v_a}$ is interpreted as a probability map that specifies the likelihood of belonging to a vessel region for each pixel. In contrast with a manually annotated binary map, the generated labels are noisy and neither the vessel pixels have probability 1 nor the background pixels

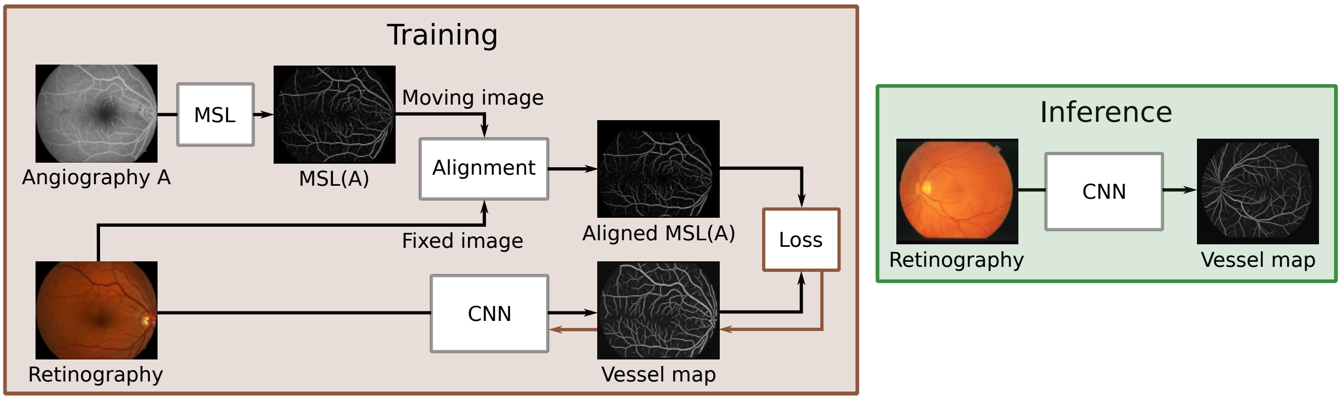


Fig. 2. Summary of the proposed approach for learning to segment the blood vessels in retinography. During the training, the supervisory signal is automatically obtained from the unlabeled multimodal data. After the training, the network is able to estimate a vessel segmentation from retinography.

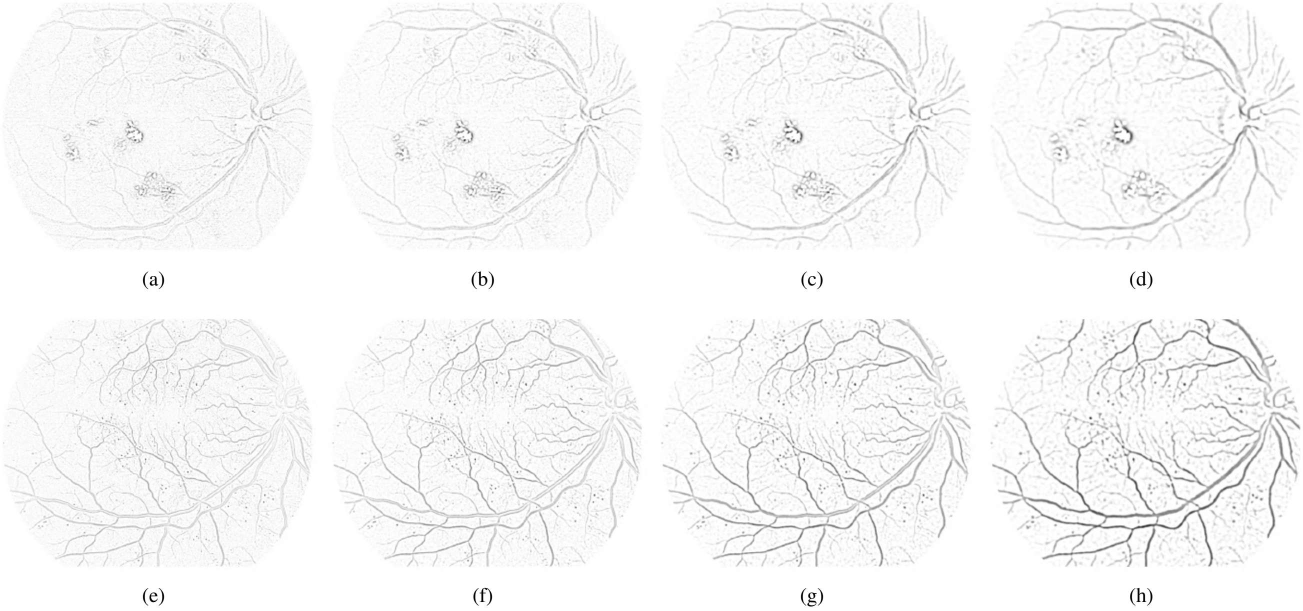


Fig. 3. Results of applying Laplacian of Gaussian filters with different scales over the retinography (1^{st} row) and the angiography (2^{nd} row) in Fig. 1. (a)-(e) $t = 1$, (b)-(f) $t = \sqrt{2}$, (c)-(g) $t = 2$, and (d)-(h) $t = 2\sqrt{2}$ where t is the scale of the filter. The intensity of the images has been inverted for a better visualization.

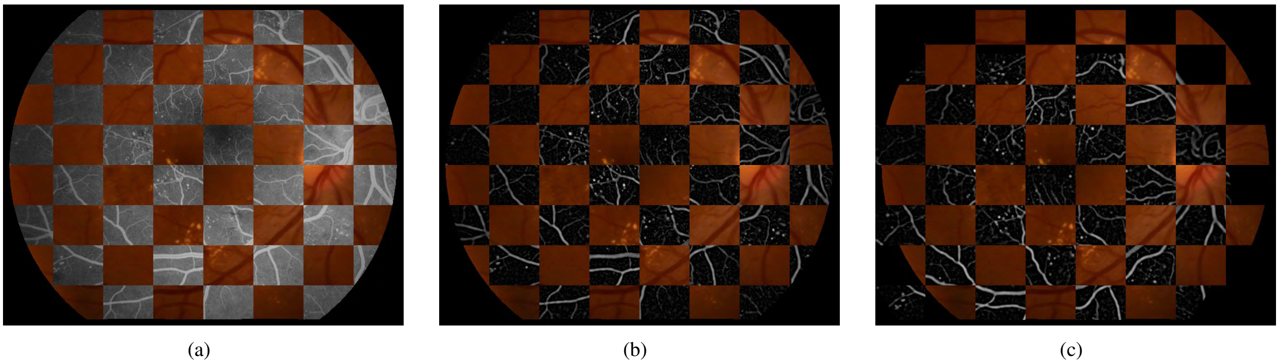


Fig. 4. Example of automatic generation of vessel map for a retinography. (a) Original retinography-angiography pair that is misaligned. (b) Retinography and generated vessel map before the alignment. (c) Retinography and generated vessel map after the alignment.

probability 0. The network must, therefore, learn to distinguish the vessel structures within these noisy annotations. In this approach, the network is trained using the binary cross-entropy (BCE) loss between the network output and the labels. For each pair $(\mathbf{r}, \mathbf{v_a})$, the loss to minimize is obtained as:

$$\mathcal{L}(f(\mathbf{r}), \mathbf{v_a}) = -\sum \mathbf{v_a} log(f(\mathbf{r})) + (1 - \mathbf{v_a})(log(1 - f(\mathbf{r}))) \tag{2}$$

where $f(\mathbf{r})$ is the predicted probability map.

For the regression approach, the training objective is to predict the $\mathbf{v_a}$ values directly, using distance measures between the network output and the labels. In this work, we experiment with the L1, L2, and Structural Similarity (SSIM) metrics. L1 and L2 are common metrics for the training and evaluation of regression problems. Their corresponding losses are computed as:

$$\mathcal{L}^{L1} = \sum |f(\mathbf{r}) - \mathbf{v_a}| \tag{3}$$

$$\mathcal{L}^{L2} = \sum ||f(\mathbf{r}) - \mathbf{v_a}||_2^2 \tag{4}$$

where $f(\mathbf{r})$ is the network output. SSIM is commonly used as image quality assessment metric but it has recently shown good performance for training a similar task with multimodal data [11]. The SSIM loss, for each pair $(\mathbf{r}, \mathbf{v_a})$, is obtained as:

$$\mathcal{L}^{SSIM}(f(\mathbf{r}), \mathbf{v_a}) = -\sum SSIM(f(\mathbf{r}), \mathbf{v_a}) \tag{5}$$

In order to compute a local SSIM value for each pixel, we take into account a small neighborhood to compute the required statistics. The SSIM map between two images can be computed as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_\mathbf{x}\mu_\mathbf{y} + C_1) + (2\sigma_\mathbf{xy} + C_2)}{(\mu_\mathbf{x}^2 + \mu_\mathbf{y}^2 + C_1)(\sigma_\mathbf{x}^2 + \sigma_\mathbf{y}^2 + C_2)} \tag{6}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two single channel images, $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ are the local averages of $\mathbf{x}$ and $\mathbf{y}$, respectively. $\sigma_\mathbf{x}$ and $\sigma_\mathbf{y}$ are the local standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, whereas $\sigma_\mathbf{xy}$ is the local covariance between $\mathbf{x}$ and $\mathbf{y}$. These values are computed for each pixel weighting its neighborhood with a Gaussian window of $\sigma = 1.5$ [14].

D. Network architecture and training

As neural network architecture, we use U-Net [15]. This network has the ability to integrate multi-scale patterns at full resolution, and it is typically taken as reference for segmentation tasks in medical imaging. U-Net is a fully convolutional neural network characterized by having a symmetric encoder-decoder structure and skip connections. We implemented U-Net, as in [15], with nine convolutional blocks. Each convolutional block has two convolutions of 3×3 kernels followed by ReLU activation functions. In the encoder, the downsampling is produced with spatial max pooling, whereas the upsampling in the decoder uses transpose convolutions. The activation of the output layer is sigmoid for the classification problems and linear for the regression counterparts.

The Adam algorithm [16] is used for the optimization, with parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to $\alpha = 1e-4$. A learning rate schedule that reduced the learning rate by a factor of 10 when the validation loss ceases to improve for 25 epochs is applied. Early stopping is used when the validation loss does not improve for 100 epochs. These parameters were empirically established as those that provided the desired performance.

Data augmentation is also applied to reduce the overfitting of the network to the training set. This data augmentation consists of spatial and color transformations applied online during training. The spatial augmentations are random affine transformations whereas the color augmentations are random global variations of image components in HSV color space.

III. Experiments and Results

A. Datasets

The public multimodal dataset of eye fundus images provided by Isfahan MISP [17] is used for the network training. This dataset comprises 59 retinography-angiography pairs of a significative healthy and pathological representation. In particular, 29 of the image pairs are from healthy individuals whereas the remaining 30 image pairs are from individuals that were diagnosed with diabetic retinopathy. The images present a resolution of 720×576 pixels. For the evaluation of the trained networks, the DRIVE [18] and STARE [19] datasets, containing manually labeled vasculature, are used. These are usual datasets of reference for vessel segmentation in retinography [2].

The DRIVE public dataset consists of 40 retinographies with manually annotated vessel segmentations. The images present a resolution of 565×584 pixels. This dataset is, by default, divided into training and test subsets of 20 images each. Although the presented approach does not requires the training subset, for an adequate comparison with other works we only evaluate the networks using the 20 images of the test subset.

The STARE public dataset consists of 20 retinographies with manually annotated vessel segmentations. The images present a resolution of 700×605 pixels. Different strategies are used in the literature for the training and evaluation using this dataset, being common to follow a leave-one-out approach. Given that no annotated training data is required in the presented approach, we use the whole dataset for the evaluation of the trained networks.

B. Experiments

The proposed approach consists in training the prediction of angiography-derived vessel maps, i.e., the automatically generated labels, from retinography (R2MSL(A)). In order to better understand the obtained results, we perform an analysis including alternatives to our main proposal. Table I summarizes the considered alternatives. One alternative is to train the generation of angiography from retinography, using regression, and to apply the vessel enhancement as a post-processing step (R2A+MSL). This post-processing can also be applied over the original retinography without requiring the training of any network (R+MSL). Finally, it is also possible

TABLE I
ALTERNATIVE METHODS CONSIDERED IN THE PERFORMED EXPERIMENTS.

Name	Training		Post-processing
	Input	Target	
R2MSL(A)	**r**	MSL(**a**)	-
R2A+MSL	**r**	**a**	MSL
R+MSL	-	-	MSL
R2MSL(R)	**r**	MSL(**r**)	-

to train a neural network in the prediction of a vessel enhanced retinography without using the multimodal data (R2MSL(R)).

Additionally, an analysis of the considered classification and regression training losses is performed independently for R2MSL(A) and R2MSL(R). In the case of R2A+MSL, the most adequate training loss is already known to be SSIM, which results from the experiments in [11].

In order to evaluate the vessel segmentation task, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) analysis are performed. ROC curves are typically used for the evaluation of retinal vessels segmentation [2], given that they allow the evaluation of raw probability maps without choosing the decision threshold. PR curves are included because they provide a complementary evaluation criteria that is more sensible to the variation of false positives in unbalanced classification problems. This is the case with vessel segmentation in retinal images, where the background class is significantly more probable than the vessel class.

C. Results and discussion

Table II shows the results of the evaluation using both DRIVE and STARE datasets for all the studied combinations. First of all, these results allow us to select the best training losses for R2MSL(A) and R2MSL(R). For R2MSL(A), similar results are obtained with the different losses. In particular, BCE classification produces the best results in DRIVE, whereas SSIM regression produces the best results in STARE. Regarding R2MSL(R), there is a slightly larger difference among the results that were obtained with the different losses. In this case, the best performance is clearly achieved using L2 regression. These experiments show that the most adequate training loss depends on the specific followed approach, without the possibility of establishing a best overall solution. Furthermore, there are also differences between the evaluation in DRIVE and STARE. This is explained by the fact that STARE is a more heterogeneous dataset than DRIVE, including higher rates of severe pathological cases.

The PR and ROC curves of the studied alternatives are depicted in Fig. 5. Only the best training losses are selected for each case: BCE and SSIM for R2MSL(A), L2 for R2MSL(R) and SSIM for R2A+MSL. It is observed that the best results are produced with R2MSL(A). Furthermore, R2MSL(A) produces better results than the other alternatives independently of the training loss, as it can be seen in Table II. These means that the improvement in performance is due to the proposed self-supervised multimodal training with the automatically generated labels. The second best approach is R2A+MSL, which is the only other alternative including multimodal data for training. This demonstrates that the angiography provides valuable information for learning about the retinal vasculature, regardless of whether the vessel enhancement is learned or applied as post-processing. Simultaneously, this also evidences that CNNs can successfully learn relevant patterns from multimodal data with the strategy of using one of the modalities as target, implicitly (R2MSL(A)) or explicitly (R2A+MSL).

With regards to the alternatives that do not include the multimodal data for training, R+MSL clearly achieves better results than R2MSL(R).

Examples of the predicted vessel maps using networks trained for R2MSL(A) with BCE classification and SSIM regression are depicted in Fig. 6. These examples are obtained from the test subsets of DRIVE and STARE and include the manually annotated ground truths for comparison. It is observed that the networks learned to represent the retinal vasculature with great detail, even including significantly small blood vessels that are frequently missed by many approaches. The visual analysis shows that training with SSIM leads to obtain a more precise detection in the cases of small vessels than training with BCE. This performance is stable, being observed both in DRIVE and STARE. However, in return, there is also slightly more noise in the background when training with SSIM, which penalizes the results. In general, the most important visual differences between the predicted vessels and the ground truths are the background noise and the non-binary nature of the generated predictions.

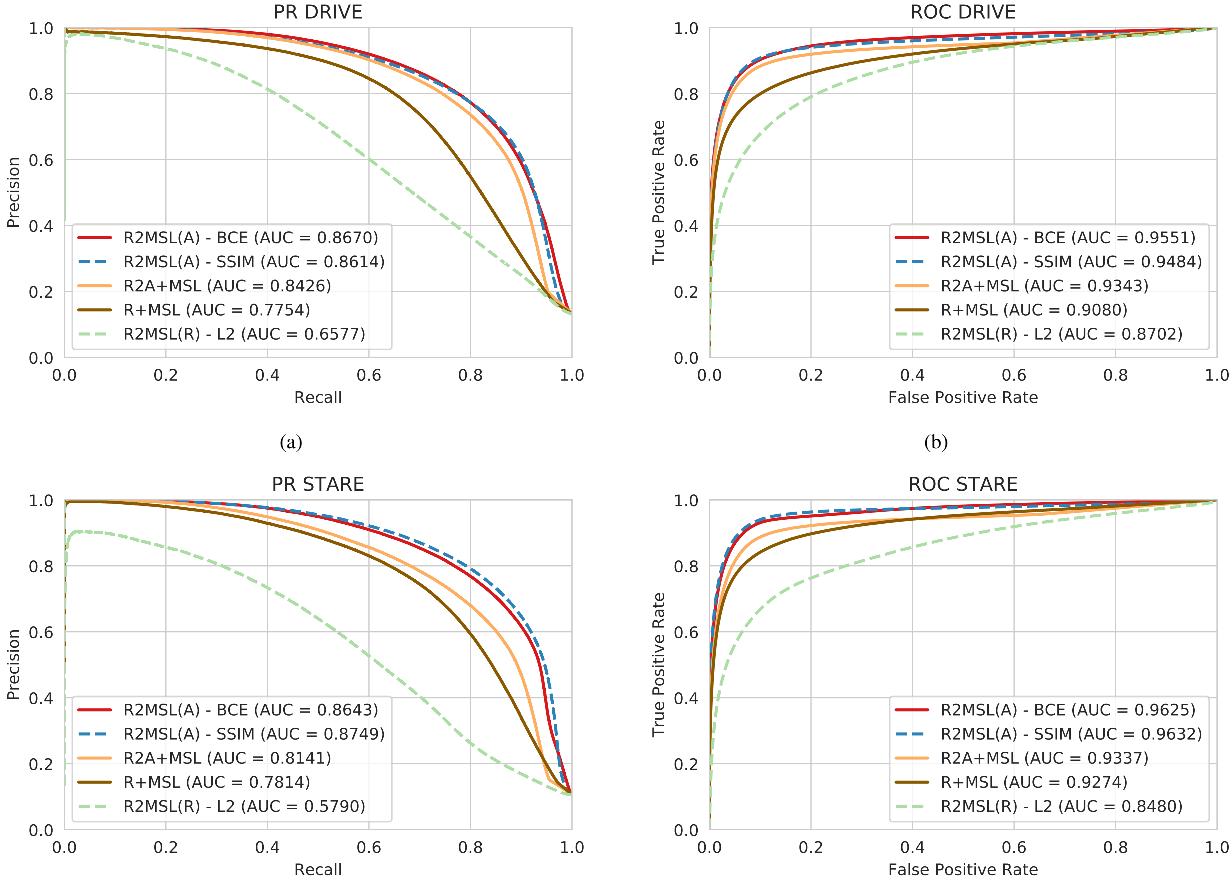
Table III shows a comparison with fully-supervised methods where manually annotated labels were used for training. We would like to remark that deep learning-based methods that do not require manual annotations should not be directly compared with standard supervised approaches. Nevertheless, the comparison with these state-of-the-art works allows for a better analysis of the obtained results. It is observed that the works in Table III produce better results that out proposal but when they are trained and tested on the same dataset. Although the AUC-ROC values are within a close margin. We have to consider the worse scenario of our proposal, which is always trained and tested on different datasets. In fact, if the fully-supervised methods are trained and tested on different datasets, their performance decreases and the gap with our self-supervised approach is reduced. In particular, for the STARE dataset, the performance of the proposed approach is within the ranges that are reported in the literature. It must be say that this scenario is also the most interesting towards the practical use of the methods.

With regards to AUC-PR values, only one of the works in the literature reported results for this metric (only for training and test on the same dataset). In comparison with AUC-ROC, the relative performance of our approach in AUC-PR is lower, which may be due to a high number of false positives. This is consistent with the slight background noise that is observed in the examples of Fig. 6.

TABLE II

Quantitative results for the evaluation of the retinal vessel segmentation task in both DRIVE and STARE public datasets.

		DRIVE		STARE	
Approach	Training loss	AUC-PR(%)	AUC-ROC(%)	AUC-PR(%)	AUC-ROC(%)
R2MSL(A)	BCE	**86.70**	**95.51**	86.43	96.25
	SSIM	86.14	94.84	**87.49**	**96.32**
	L2	86.38	95.17	85.65	96.15
	L1	85.90	93.80	85.81	95.17
R2MSL(R)	BCE	64.90	**87.03**	56.56	83.50
	SSIM	62.46	85.56	53.40	83.75
	L2	**65.77**	87.02	**57.90**	**84.80**
	L1	62.27	84.69	54.29	83.29
R2A+MSL	SSIM	84.26	93.43	81.41	93.37
R+MSL	-	77.54	90.80	78.14	92.74



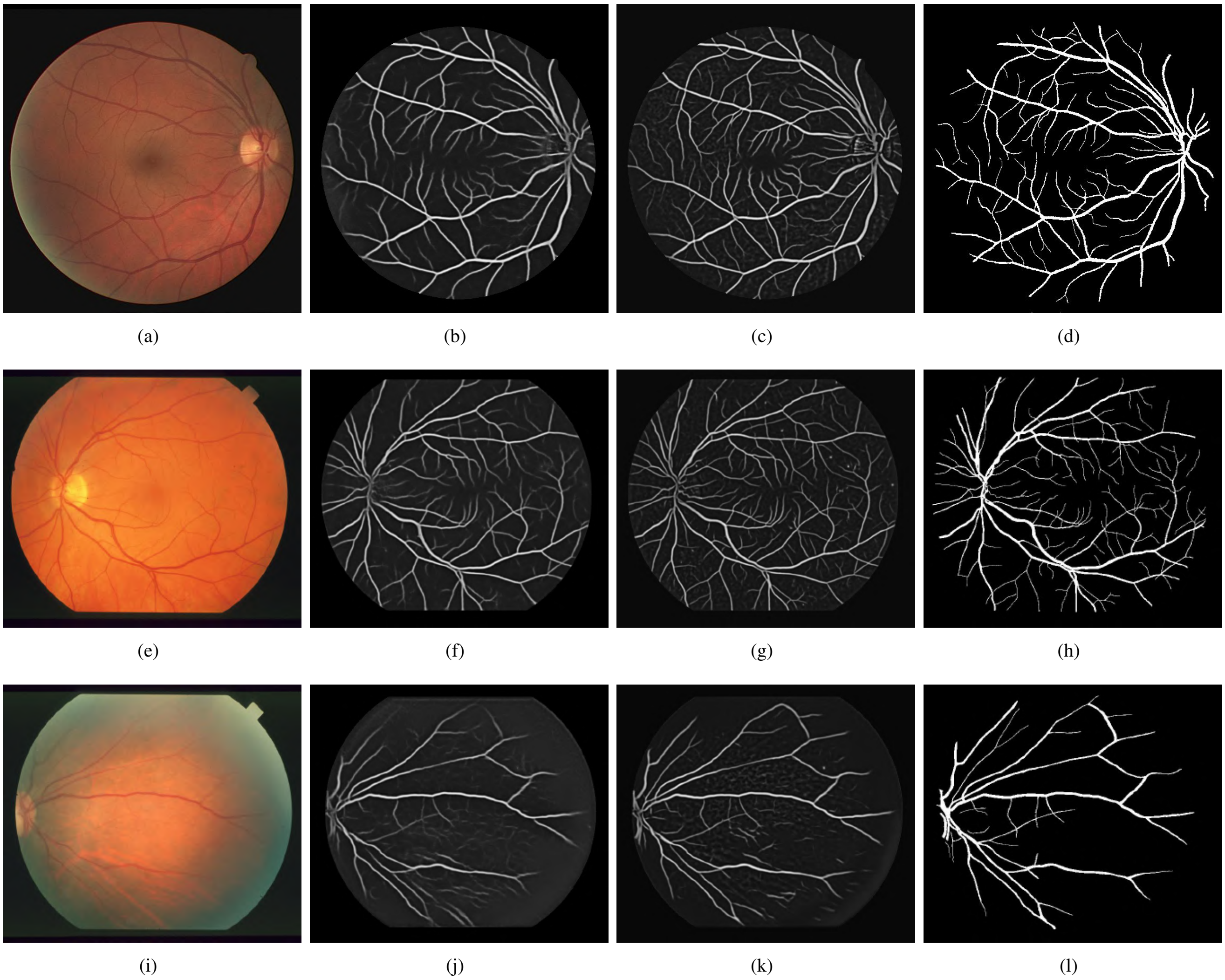
(a)

(b)

(c)

(d)

Fig. 5. PR ((a),(c)) and ROC ((b),(d)) curves from the evaluation of the blood vessel segmentation task in both DRIVE ((a),(b)) and STARE ((c),(d)) datasets.


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

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 6. Examples of predicted vessel maps after the self-supervised training of R2MSL(A). (1^{st} column) Original retinographies. (2^{nd} column) Predicted vessel maps after training with BCE. (3^{rd} column) Predicted vessel maps after training with SSIM. (4^{th} column) Manually annotated ground truths. The images are taken from the test subsets of (1^{st} row) DRIVE and (2^{nd} and 3^{rd} rows) STARE.

TABLE III
COMPARISON OF METHODS FOR VESSEL SEGMENTATION IN RETINOGRAPHY.

		DRIVE		STARE	
Method	Training data	AUC-PR(%)	AUC-ROC(%)	AUC-PR(%)	AUC-ROC(%)
Fully-supervised (Human annotated labels)					
Fraz et al. [20]	DRIVE	-	97.47	-	96.60
Likowski et al. [5]		-	97.90	-	95.95
Li et al. [21]		-	97.38	-	96.71
Maninis et al. [3]		90.64	97.93	-	-
Juan Mo et Lei Zhang [4]		-	97.82	-	97.51
Fraz et al. [20]	STARE	-	96.97	-	97.68
Likowski et al. [5]		-	96.05	-	99.28
Li et al. [21]		-	96.77	-	98.79
Maninis et al. [3]		-	-	92.46	98.72
Juan Mo et Lei Zhang [4]		-	96.53	-	98.85
Self-supervised (Automatically generated labels)					
R2MSL(A)-BCE (ours)	Multimodal (Isfahan MISP)	86.70	95.51	86.43	96.25
R2MSL(A)-SSIM (ours)		86.14	94.84	87.49	96.32

The presented results have shown that CNNs trained with the automatically generated labels from multimodal data are able to extract and represent the retinal vasculature with great detail. Simultaneously, it is also evident that not seeing any binary vessel map during the training penalizes the performance of these networks. This could be overcome with the addition of a refinement step or including a complementary loss that penalizes the distance of the network output to a desirable target distribution. Nevertheless, the results are promising for including the proposed approach as a complementary task in semi-supervised settings.

IV. Conclusions

In this paper, we proposed a novel approach that allows training neural networks for retinal vessel segmentation without using manual annotations. Instead, the proposed approach takes advantage of the existent multimodal image pairs in medical imaging. Specifically, the networks are trained to produce angiography-derived vessel maps from retinography. The training samples for this self-supervised task are automatically generated from the originally unlabeled data. Thus, the size of the training dataset can scale up free of cost.

The CNNs trained on the proposed self-supervised task learn to extract and represent with great detail the retinal vasculature. This is demonstrated with the experiments in different public datasets, including healthy and pathological samples. Additionally, the obtained results show a great potential of the proposed self-supervised multimodal training towards improving the segmentation of the retinal vasculature. In that sense, as future work, we plan to combine the proposed approach with some annotated samples in a semi-supervised setting.

Acknowledgment

This work is 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 by the Ministerio de Economía, Industria y Competitividad, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the European Regional Development Fund (ERDF) and European Social Fund (ESF) of the EU, and the Xunta de Galicia through Centro Singular de Investigación de Galicia, accreditation 2016-2019, ref. ED431G/01 and Grupo de Referencia Competitiva, ref. ED431C 2016-047 research projects, and the predoctoral grant contract ref. ED481A-2017/328.

References

[1] R. Besenczi, J. Tóth, and A. Hajdu, "A review on automatic analysis techniques for color fundus photographs," *Computational and Structural Biotechnology Journal*, vol. 14, pp. 371–384, 2016.

[2] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, "Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 71–91, 2018.

[3] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Deep Retinal Image Understanding," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.

[4] J. Mo and L. Zhang, "Multi-level deep supervised networks for retinal vessel segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 12, pp. 2181–2193, dec 2017.

[5] P. Liskowski and K. Krawiec, "Segmenting Retinal Blood Vessels with Deep Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.

[6] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2070–2079.

[7] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[9] E. D. Cole, E. A. Novais, R. N. Louzada, and N. K. Waheed, "Contemporary retinal imaging techniques in diabetic retinopathy: a review," *Clinical & Experimental Ophthalmology*, vol. 44, no. 4, pp. 289–299, may 2016.

[10] A. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)*, vol. 126, 2018, pp. 97–104.

[11] A. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Retinal image understanding emerges from self-supervised multimodal reconstruction," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 11070, 2018, pp. 321–328.

[12] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 117–156, Nov 1998.

[13] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, and M. J. Carreira, "Retinal verification using a feature points-based biometric pattern," *EURASIP Advances in Signal Processing*, vol. 2009, no. 1, p. 235746, Mar 2009.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, apr 2004.

[15] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, May 2015.

[17] S. H. M. Alipour, H. Rabbani, and M. R. Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Comp. and Math. Methods in Medicine*, vol. 2012, 2012.

[18] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge based vessel segmentation in color images of the retina," *IEEE Trans. on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

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

[20] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, sep 2012.

[21] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A Cross-Modality Learning Approach for Vessel Segmentation in Retinal Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 109–118, jan 2016.

3.2 Journal Paper: Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction

Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies,
University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Journal: Applied Soft Computing
Volume: 91
Pages: 106210
Editorial: Elsevier
ISSN: 1568-4946
Year: 2020

Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Marcos Ortega[a,b]

[a] *CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain*
[b] *Department of Computer Science, Universidade da Coruña, A Coruña, Spain*

Abstract

Deep learning is becoming the reference paradigm for approaching many computer vision problems. Nevertheless, the training of deep neural networks typically requires a significantly large amount of annotated data, which is not always available. A proven approach to alleviate the scarcity of annotated data is transfer learning. However, in practice, the use of this technique typically relies on the availability of additional annotations, either from the same or natural domain. We propose a novel alternative that allows to apply transfer learning from unlabeled data of the same domain, which consists in the use of a multimodal reconstruction task. A neural network trained to generate one image modality from another must learn relevant patterns from the images to successfully solve the task. These learned patterns can then be used to solve additional tasks in the same domain, reducing the necessity of a large amount of annotated data.

In this work, we apply the described idea to the localization and segmentation of the most important anatomical structures of the eye fundus in retinography. The objective is to reduce the amount of annotated data that is required to solve the different tasks using deep neural networks. For that purpose, a neural network is pre-trained using the self-supervised multimodal reconstruction of fluorescein angiography from retinography. Then, the network is fine-tuned on the different target tasks performed on the retinography. The obtained results demonstrate that the proposed self-supervised transfer learning strategy leads to state-of-the-art performance in all the studied tasks with a significant reduction of the required annotations.

Keywords: deep learning; eye fundus; self-supervised learning; optic disc; blood vessels; fovea; medical imaging; transfer learning

1. Introduction

The analysis of the anatomical structures in the retina represents an essential step for the diagnosis and screening of important ocular and systemic diseases. The morphology of the anatomical structures, such as blood vessels, fovea, or optic disc, can in itself provide evidence of the presence of certain diseases. Additionally, they can be used as reference for the localization of lesions as well as for the assessment of their severity [1].

The retinal anatomy can be studied using eye fundus photography, or retinography, which is a non-invasive and affordable imaging technique. These reasons motivate its widespread use in many clinical services, and make it an interesting target for the development of image analysis algorithms [2]. In this regard, several works have approached the automatic analysis of eye fundus images, including the localization or segmentation of the different anatomical structures [1]. Similarly to other medical fields, the number of methods based on neural networks has grown significantly in the last few years, which carried an improvement of the obtained results [3, 4, 5]. Currently, the use of deep neural networks (DNNs) is the standard approach

[*] Corresponding author
Email address: a.suarezh@udc.es (Álvaro S. Hervella)

in many computer vision applications when the required annotated data is available. DNNs have not only improved the results obtained with traditional methods, but have also brought a new simplified paradigm where no feature design is needed [6]. Instead, the focus has shifted to the design or selection of the most suitable network architectures, training losses and training strategies [7].

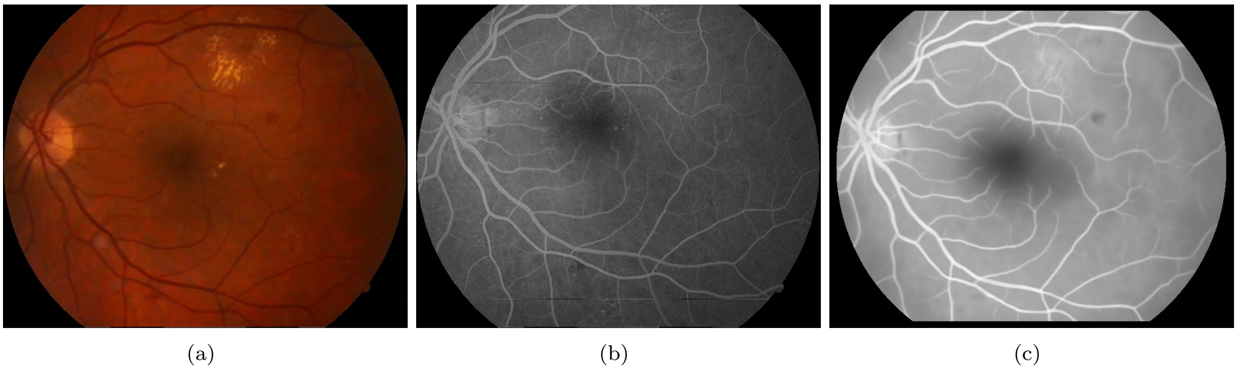
Regarding the automatic analysis of representative anatomical structures in retinography, the main limitation for the early use of DNNs was the scarcity of annotated data [3]. In that sense, the available datasets typically present a small number of annotated samples due to the difficulty of hand-labeling the retinal images in detail. Moreover, despite that some large datasets have been gathered, in practice, the annotated data usually present a meager representation of pathological cases [8], given that those images are typically of higher variability and complexity.

The scarcity of annotated data is not specific to retinal imaging. Instead, this is a broadly relevant issue in medical imaging, where a high level of expertise is required for the reliable labeling of the medical data [9]. Conversely, a large amount of medical images is produced everyday in the different medical services due to the widespread use of imaging techniques in modern clinical practice [7]. This directly produces the availability of large unlabeled datasets, which may be used for the training of neural networks in unsupervised or semisupervised settings. Moreover, the medical images are typically accompanied by clinical reports describing the patient's conditions, which may be used for distilling image-level labels [7]. In contrast, pixel-level labels require to be annotated on purpose by, at least, one clinical expert. Moreover, the manual annotation of pixel-level labels represents a difficult task, being more tedious and time-consuming than the manual annotation of image-level labels. This is reflected in the number of annotated samples that are provided in common medical imaging datasets [1], being significantly smaller when the required annotations are more detailed [10].

The limited annotated data in medical imaging is typically alleviated using extensive data augmentation and transfer learning [7]. The use of data augmentation techniques including, e.g., rigid transformations, elastic deformations or color transformations, has become a key component of successful deep learning methods [7]. Transfer learning, on the other hand, has been applied since the earliest deep learning approaches on medical imaging. Early works used the first layers of pre-trained classification networks as feature extractors [11]. These networks are trained in a broad domain application with extensive available data, such as ImageNet classification [12]. Posterior works additionally performed fine-tuning of the pre-trained layers together with additional layers that are specialized for the target task [3]. Multi-task learning techniques have also been recently explored for their ability to combine complementary tasks over data of the same domain [13]. Multi-task settings can be seen as a special case of transfer learning where the transference of knowledge is bidirectional and simultaneous between the involved tasks. In this case, the amount of labeled data is increased by using heterogeneous labels (for each task) over data of the same domain. However, these additional heterogeneous labels from the same domain can be also exploited using a regular pre-training and fine-tuning approach, to achieve improved results on the later tasks [14]. In the case of training multiple tasks of this kind, with varying difficulty, the training order may have an impact in the final performance. In this sense, some works have also proposed to optimize the sampling order or the different tasks to improve the final outcome [15].

Self-supervised methods are a recent alternative that allows the use of unlabeled data for transfer learning [16]. These approaches rely on the use of innovative complementary tasks which labels can be automatically computed from the unlabeled datasets and, thus, can be trained without the need of manual annotations. The purpose of training these self-supervised tasks is to learn relevant patterns of the domain from the data, and then use the learned patterns to improve the desired tasks through transfer of multitask learning. Existent proposals in medical imaging have exploited, as reference, the color information in images using a colorization task [17] or the relation among longitudinal data by learning patient embeddings [18].

A rich source of information that has still not been exploited for self-supervised transfer learning is the unlabeled multimodal data in medical imaging. In modern clinical practice, it is common to analyze and diagnose the patients using multiple imaging techniques. This results in the availability of multimodal sets in which samples from complementary image modalities are available for the same patient. The availability of these multimodal data can be exploited using a self-supervised multimodal reconstruction task where a neural network is trained to generate one image modality from other. If the two involved modalities are



(a) (b) (c)

Figure 1: Example of (a) retinography, (b) fluorescein angiography, and (c) pseudo-angiography for the the same eye. The pseudo-angiography (c) is generated from (a) using the method proposed in Hervella et al. [19].

different enough, the network has necessarily to learn the recognition of relevant domain-related patterns to successfully solve the task. Then, the learned models can be further adjusted to solve additional target tasks over the same input modality.

In particular, in this work, we experiment with these ideas in the context of the localization and segmentation of anatomical structures of the eye fundus in retinography. The objective is to reduce the amount of annotated data that is required to solve these tasks with a DNN, and to that end we propose to use the self-supervised multimodal reconstruction for transfer learning. Specifically, we pre-train the networks to generate fluorescein angiography from retinography. The retinography and angiography are complementary image modalities, both providing visualizations of the eye fundus. However, the angiography is an invasive modality that requires the injection of a contrast dye to the patients, providing additional information about the retinal vasculature and related lesions. In the proposed paradigm, both unlabeled image modalities are used to pre-train the networks. However, the target tasks are performed using a single image modality, which in this case is the retinography. Moreover, the unlabeled multimodal data for pre-training and the task-specific data for fine-tuning do not need to belong to the same patients. This allows the use of any multimodal dataset available in the same domain, independently of the target tasks.

With regards to the multimodal reconstruction, Hervella et al. [19] demonstrated that a pseudo-angiography representation can be generated from a given retinography using a DNN. Moreover, the vascular enhancement in the angiography can also be directly exploited to produce an approximate representation of the vascular tree in retinography [20], requiring an additional pre-processing of the target angiographies. None of the previous works, however, have taken advantage of the domain-specific patterns that a neural network must learn in order to perform the multimodal reconstruction. The idea proposed in this work exploits those patterns learned from the unlabeled multimodal data for transfer learning purposes. This represents a novel alternative to complement the training of a DNN and reduce the amount of annotated data that is required. As reference, an illustrative example of retinography, fluorescein angiography, and generated pseudo-angiography for the same eye is depicted in Figure 1.

In order to demonstrate the advantages of the proposed self-supervised transfer learning strategy, we use the multimodal reconstruction as a common self-supervised pre-training for: (1) the localization of the fovea, (2) the localization and (3) segmentation of the optic disc, and (4) the segmentation of the retinal vasculature. Additionally, we aim at solving all these target tasks with the same standard methodology, including the network architecture and training strategy. In order to study the efficient use of annotated data with our proposal, we conducted an extensive experimentation with progressive amounts of annotated training data. The objective is to demonstrate that the self-supervised multimodal reconstruction successfully reduces the amount of annotations required to solve the considered target tasks.

1.1. State-of-the-art

In the literature, several works have approached the automatic analysis of the most important anatomical structures in retinography [1]. Previous works typically focus on the localization or segmentation of a single anatomical structure. However, the localization of the fovea has been traditionally approached together with the localization of the optic disc. This is motivated by the use of the optic disc location as reference to detect the fovea [21, 22]. Additionally, the retinal vascular tree has also been used as reference for the localization of both the optic disc and the fovea [23, 22].

Regarding the optic disc, some proposals exploit its characteristic circular shape. For instance, edge detection filters can be used to obtain optic disc boundary candidates [23, 24]. Then, these boundaries allow to derive both the segmented area and the center coordinates after a refinement step, e.g., using a hough transformation [24, 23] or measuring the distance to some pre-computed templates [25]. In this context, Dashtbozorg et al. [26] proposes specific filters in order to better match the optic disc shape. Alternatively, the characteristic color patterns of the optic disc are also exploited by applying histogram matching [27]. Additionally, Qureshi et al. [25] explores the use of an ensemble of previously proposed algorithms to improve the results. In contrast, the most recent proposals use deep learning for both the localization [4, 5] and the segmentation [3] of the optic disc. In the localization task, a convolutional network with fully-connected output layers can be used to predict the fovea coordinates [4]. However, instead, Meyer et al. [5] reformulates the problem as a heatmap regression task, which can be performed using fully-convolutional networks. The latter approach is the one that we have adopted in this work for the localization of both optic disc and fovea.

With regards to the fovea localization, traditional approaches typically rely on the previous detection of the optic disc to reduce the search area [22, 28, 21]. Additionally, Gegundez-Arias et al. [22] also makes use of the extracted retinal vascular tree to perform a better initial estimate of the foveal region. The final localization is usually performed exploiting the characteristic shape and color of the foveal region. For instance, Niemeijer et al. [21] uses a k-NN regressor and features extracted from both the retinal image and the segmented blood vessels, whereas Gegundez-Arias et al. [22] uses thresholding techniques and features from the original image. In addition, the fovea and the optic disc can be detected using template matching with the same template filter but of opposite responses [28]. Similarly to the optic disc, the most recent proposals use DNNs for the regression of the fovea coordinates [4] or the prediction of a full-image size distance map [5].

In the case of the retinal vasculature segmentation, traditional approaches have typically relied on the characteristic tubular shape of blood vessels. This characteristic can be exploited using the gradients of the image or Gabor filter responses, among other techniques [29]. However, recent works have successfully solve this task using DNNs, either fully convolutional [3], fully connected [30] or convolutional with fully-connected output layers [31]. In this regard, the novelty of recent works is related to the use of specific network designs or training objectives, including, as reference, the use of class-balanced losses [3] or the supervision to intermediate layers [32].

The rest of the manuscript is organized as follows: A general overview of the proposed approach, along with a description of the pre-training and target tasks is depicted in Section 2. The network architecture and the training strategy are also detailed in this section. The description of the conducted experiments and the obtained results are presented in Section 3. Section 4 is focused on the discussion of results and the final conclusions are drawn in Section 5.

2. Methodology

A general scheme that summarizes the proposed methodology is depicted in Figure 2. Particularly, the self-supervised reconstruction of fluorescein angiography from retinography is used as pre-training. Then, the pre-trained neural network is fine-tuned on the different target tasks. Given that the multimodal reconstruction covers the whole anatomy of the retina, it is expected that the internal neural network representations that are used for the reconstruction are also useful for the detection and segmentation of the different anatomical structures.

The exact same network architecture and training strategy are employed for all the considered tasks, with the only difference of the loss function. In particular, for the pre-training task a reconstruction loss is

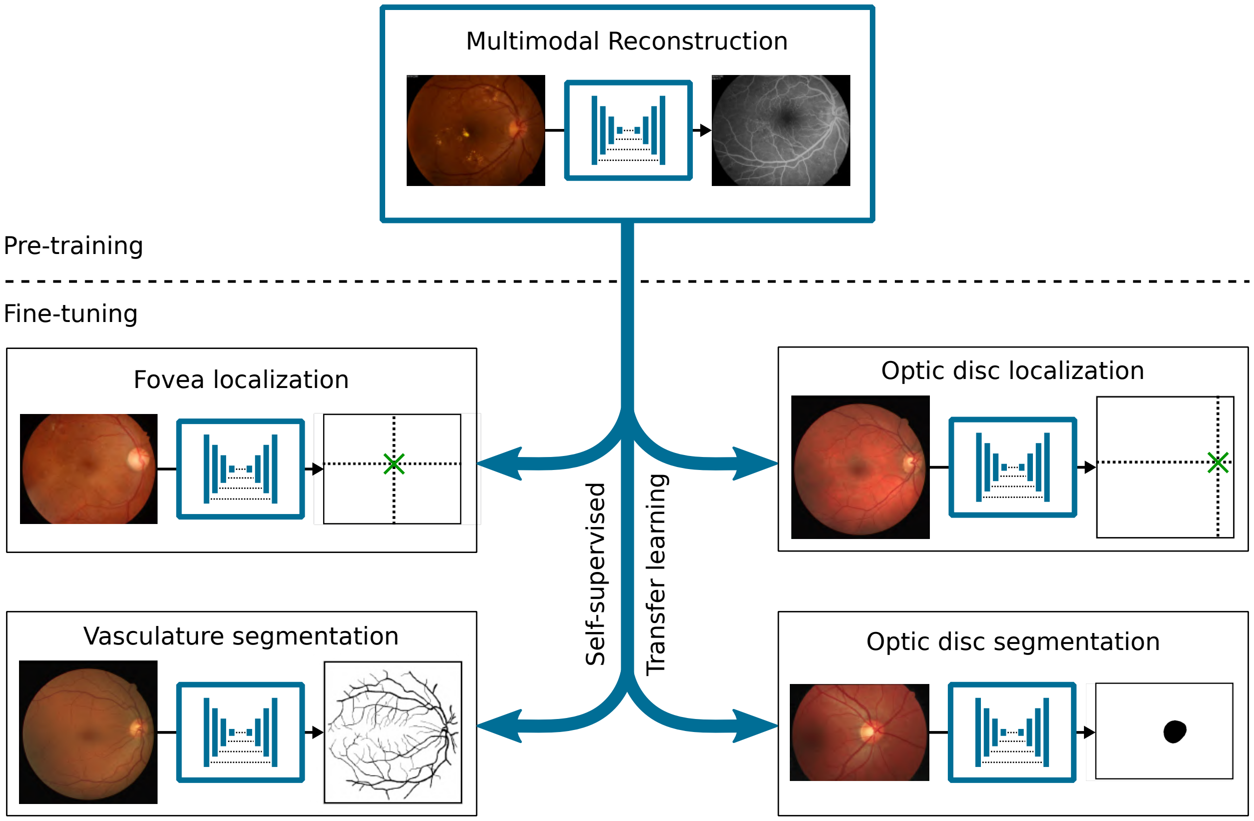


Figure 2: Scheme of the proposed methodology. The self-supervised multimodal reconstruction of angiography from retinography is used as pre-training task. The pre-trained network is fine-tuned on different target tasks aiming at the analysis of the main anatomical structures in retinography.

used, whereas for the target tasks two different losses are used depending on the objective: a localization loss and a segmentation loss.

2.1. Self-supervised multimodal reconstruction

The multimodal reconstruction of fluorescein angiography from retinography is conceived as a self-supervised task due to the use of aligned retinography-angiography pairs from the same eye [19]. In this scenario, there is a pixel-wise correspondence between the input retinography and the target angiography. This enables the use of full-reference metrics for the reconstruction loss, which provides a supervisory training signal that involves fine image details and does not need any human labeling effort.

The aligned multimodal data for training the network is obtained after the registration of retinographies and angiographies of the same eye. This registration is performed following a domain-specific methodology that relies on the presence of retinal vessels in both image modalities [33]. This registration methodology is divided into two main steps: an initial landmark-based registration that globally aligns the images followed by a refined pixel-wise registration that corrects the remaining small misalignments between the images.

Both retinography and angiography display the eye fundus in a circular Field of View (FOV). After the image alignment, the area containing information from both modalities, denoted as the multimodal FOV, Ω_M, will be typically smaller than the individual FOVs of the original images. This area is defined as:

$$\Omega_M = \Omega_R \cap \Omega_A \tag{1}$$

where Ω_R and Ω_A denote the circular FOVs of the retinography and the angiography respectively. Consequently, Ω_M represents the region where the reconstruction loss is computed during the training. The reconstruction loss $\mathcal{L}_R(\mathbf{g}(\mathbf{r}), \mathbf{a})$ is given by:

$$\mathcal{L}_R(\mathbf{g}(\mathbf{r}), \mathbf{a}) = -\sum_{\Omega_M} \mathbf{SSIM}(\mathbf{g}(\mathbf{r}), \mathbf{a}) \tag{2}$$

where $\mathbf{r}$ is the input retinography, $\mathbf{a}$ the target angiography, $\mathbf{g}(\mathbf{r})$ the output of the network, and **SSIM** the Structural Similarity (SSIM) index map between the target angiography and the network output [19]. SSIM is frequently used as test metric for the evaluation of deep learning models that were trained with other losses. However, in our context, the direct optimization of the SSIM has demonstrated an improved performance with respect to other common metrics in the presented task [19]. The **SSIM** map is obtained as:

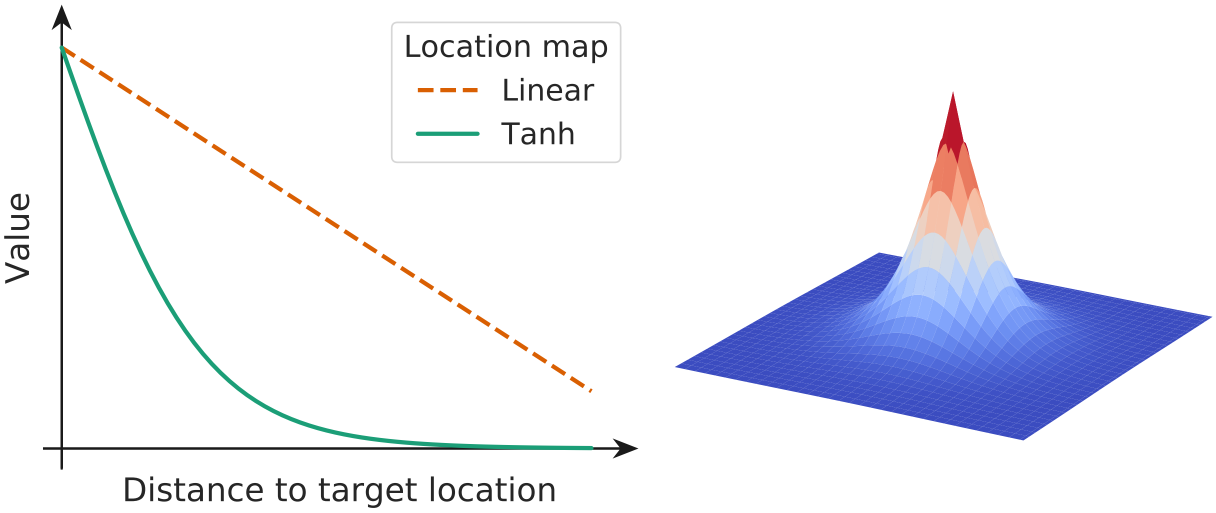
$$\mathbf{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1) + (2\sigma_{\mathbf{xy}} + C_2)}{({\mu_{\mathbf{x}}}^2 + {\mu_{\mathbf{y}}}^2 + C_1)({\sigma_{\mathbf{x}}}^2 + {\sigma_{\mathbf{y}}}^2 + C_2)} \tag{3}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote two single channel images, $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ the local averages of $\mathbf{x}$ and $\mathbf{y}$ respectively, $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ the local standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, $\sigma_{\mathbf{xy}}$ the local covariance between $\mathbf{x}$ and $\mathbf{y}$, and C_1 and C_2 are constant values used to avoid instability when the denominator terms are close to zero [34]. The local statistics for each pixel are computed using a Gaussian window with $\sigma = 1.5$ [34].

2.2. Localization of anatomical structures of the retina

The localization of the fovea and optic disc centers is obtained following the same task formulation. In this regard, the localization tasks consist in the regression of pixel coordinates, which can be directly approached using a DNN with fully connected layers that produce the coordinate values. However, this kind of regression settings can be difficult to train, and does not take full advantage of the shared weights and local connectivity of convolutional networks. An straightforward alternative is to predict a target map with two classes: the pixel of the target location and the rest of the image. In this case, the difficulty is that the target maps are heavily unbalanced. An alternative to improve this is to augment the ground truth annotations by the means of a distance map to the target pixel [5]. Using the Euclidean norm, this distance map is given by:

$$d_T(x_i, y_i) = \sqrt{(x_i - x_T)^2 + (y_i - y_T)^2} \tag{4}$$



(a)

(b)

Figure 3: (a) Value of the location map as a function of the distance to the target location. The hyperbolic tangent (tanh) version is the one used in this work, whereas the linear version is provided for comparison. (b) The location map represented as a three-dimensional surface.

where (x_T, y_T) are the coordinates of the target pixel and (x_i, y_i) the coordinates of each pixel in the image. The distance map $\mathbf{d}_T$ provides additional information for training the localization task. Nevertheless, the accurate prediction of the norm values for the most distant pixels is difficult given that less visual cues are present. This has a negative effect on the global accuracy of the prediction due to the excessive importance given to the less relevant distant pixels. Thus, we use a location map with higher variability near the target location, which is obtained by applying an exponential decay that saturates at the distant pixels. The proposed location map $\mathbf{y}_L$ is defined as:

$$\mathbf{y}_L = 1 + tanh\left(-\mathbf{d}_T \frac{\pi}{\beta}\right) \tag{5}$$

where $tanh$ is a hyperbolic tangent function, β the saturation distance, and $\mathbf{d}_T$ the original Euclidean distance map. For the experiments in this work, we set the saturation distance β to the value of the approximate optic disc radius. An illustration of the proposed location map for a given target location is shown in Figure 3. The localization tasks are then trained using a mean squared error (MSE) loss between the target location map $\mathbf{y}_L$ and the network output.

A straightforward approach can be used to recover the resulting location coordinates from the predicted location map by detecting the pixel of maximum response.

2.3. Segmentation of anatomical structures of the retina

The segmentation of the retinal vasculature and the optic disc is approached following the same formulation. Both tasks consist in the prediction of pixel-level labels within two categories: the anatomical structure of interest and the background. The training of these tasks is performed with a set $\{(\mathbf{r}, \mathbf{y}_s)_1, ..., (\mathbf{r}, \mathbf{y}_s)_N\}$ where $\mathbf{r}$ denotes the fundus image and $\mathbf{y_s}$ denotes its corresponding ground truth segmentation map. The objective is to obtain the transformation $\mathbf{f}_s$ that assigns the likelihood of belonging to the anatomical structure of interest to each pixel of the fundus image.

These binary classifications are trained optimizing the cross-entropy loss between ground truth and

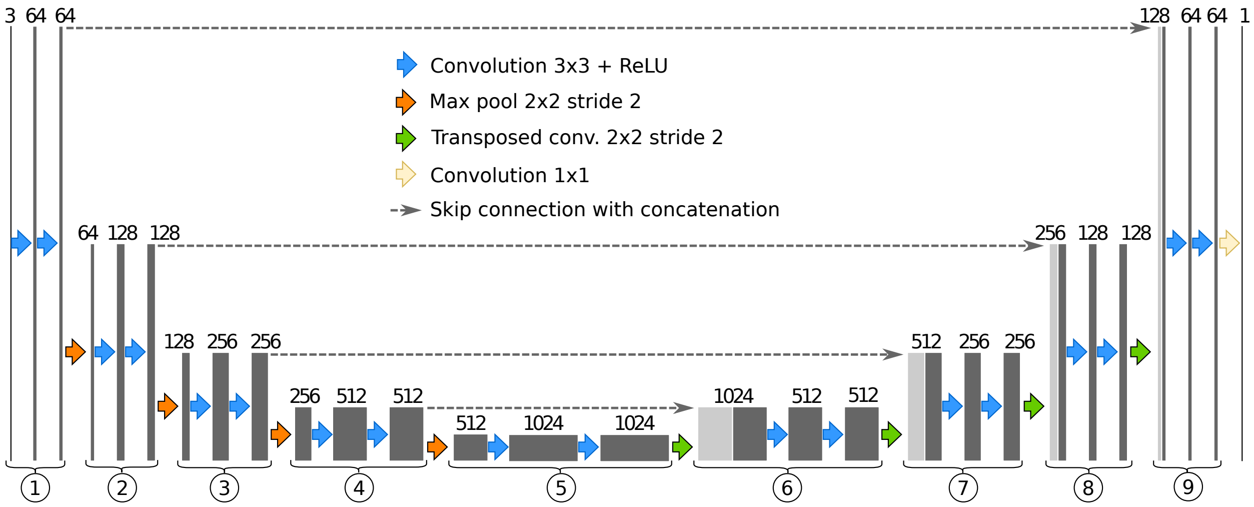


Figure 4: Description of the U-Net architecture.

network output, defined as:

$$\mathcal{L}_S(\mathbf{f}_s(\mathbf{r}),\mathbf{y}_s) = -\sum_{\Omega_R} \mathbf{y}_s log(\mathbf{f}_s(\mathbf{r})) + (1-\mathbf{y}_s)(log(1-\mathbf{f}_s(\mathbf{r}))) \tag{6}$$

where $\mathbf{r}$ is the input retinography, $\mathbf{y}_s$ the corresponding ground truth binary map, $\mathbf{f}_s(\mathbf{r})$ the output of the network, and Ω_R the retinography FOV where the loss is computed.

2.4. Network architecture

In this work, we use the U-Net architecture [35] for all the reconstruction, localization, and segmentation tasks. U-Net is a commonly used network in many medical imaging applications, and a well-known and proven baseline. In that sense, in order to ensure an strongly validated baseline, we use the same exact network that was proposed by Ronneberger et al. [35], including the same number of layers and channels, without any additional adjustments. The only exception is the number of output channels, which inevitably depends on the output that is required for each specific problem. A general scheme of the network, including details of the different layers, is depicted in Figure 4. Specifically, U-Net is a fully convolutional neural network with output and input of the same size. This allows the estimation of a full size target image map, which represents an useful property for segmentation or reconstruction tasks, as well as for the prediction of location maps.

This architecture presents a multiscale encoder-decoder structure, featuring skip connections between their respective inner blocks. In the encoder part, the width and height image dimensions are progressively reduced by half at subsequent blocks, using max pooling operations. Following the idea of the VGG networks [36], these blocks are composed of two convolutional layers with kernel size 3×3 followed by the spatial max pooling operation. The objective of the progressive reduction in space is to enforce the learning of broad and abstract patterns from the data. This helps to produce a hierarchical representation from low to high level features in which the input data is transformed. The decoder part progressively recovers the width and height of the input images, by building the output from the high level abstractions to the low level details. The progressive upsampling is produced with strided transposed convolutions that increase the spatial dimensions by a factor of 2 at each block. These transposed convolutions are interleaved between convolution layers like those in the encoder.

The width and height variations across the network create a bottleneck effect that enforces the learning of high level patterns. However, the spatial contraction penalizes the tracking of the precise localization of the extracted features. U-Net successfully improves the localization and generation of small details with the inclusion of skip connections between encoder and decoder. These connections transfer features from the encoder to the decoder at different resolutions, providing alternative paths to propagate precise spatial localizations.

All the convolutional layers of the network are followed by ReLU activation functions except for the last layer. In the case of the segmentation tasks, a sigmoid activation function is used at the output layer of the network, whereas for the localization and the multimodal reconstruction tasks a linear activation function is used instead.

2.5. Network training

When the network is trained from scratch, the parameters are randomly initialized following the method proposed by He et al. [37]. The Adam [38] algorithm is used for the optimization of the loss functions. The decay rates for the first and second order moments of Adam are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively, as originally proposed by Kingma and Ba [38]. The initial learning rate is set to $\alpha = 1e\text{-}4$ for the multimodal reconstruction and $\alpha = 1e\text{-}5$ for all the localization and segmentation tasks. The learning rate schedule is the same for all the experiments. It consists in the reduction of the learning rate by a factor of 10 when the validation loss does not improve for 2500 iterations. Each iteration consists in a network parameters update due to the presentation of a training minibatch, which is fixed to consist of one image in all the experiments. The training stops when the validation loss stalls after reaching a learning rate of $\alpha = 1e\text{-}7$. These parameters were empirically established as those that were observed to provide enough training for all the tasks.

For the target tasks the datasets are initially divided into training and hold-out test sets, whereas for the pre-training task the whole dataset is used during training. In order to control the learning rate schedule and the stopping criteria, the training sets are additionally divided into training and validation subsets. In this work, several experiments are performed varying the number of training samples used. Therefore, for each experiment, the samples that are not selected for training are included into the validation subset. In the experiments where the whole training data is used, there is no validation subset, and the schedule resulting from the previous experiment with more data samples is applied.

To avoid excessive overfitting, data augmentation techniques and dropout are also used in both the target and pre-training tasks. In that sense, we apply the same data augmentation techniques as Hervella et al. [20], including color and spatial augmentations. The color augmentations consists in random linear transformations of the image channels using the HSV color representation. The spatial augmentation consists in random affine transformations with scaling, rotation, and shearing components. Dropout layers with probability $p = 0.2$ are added to the network after the convolutional blocks 2,3,4,5, and 6, which are depicted in the Figure 4.

3. Experiments and results

In order to quantify and demonstrate the advantages of the proposed approach, the self-supervised multimodal pretraining is compared against training the networks from scratch, which is the standard alternative without requiring additional annotated data. In this way, the same experiments were conducted for two different frameworks:

- **Multimodal reconstruction**: The neural network is pre-trained on the unlabeled multimodal data using the self-supervised multimodal reconstruction. Then, the network is fine-tuned using the annotated data of the target task.

- **Random initialization**: The neural network is randomly initialized and trained from scratch using the annotated data of the target task.

In order to guarantee an adequate and fair comparison, the same network architecture and training strategy was used for both frameworks, as described in Section 2. Additionally, the same settings were also used in all the studied tasks, except for the training loss and the output layer of the network, which require to be specific for each task objective (segmentation or localization).

In general terms, several experiments were mainly performed to study whether the use of the multimodal reconstruction as self-supervised pre-training may alleviate the impact of having a very small number of annotated samples. To that end, we performed experiments with a varying progressive number of training samples, ranging from a single image to the whole training set, while keeping the same hold-out test set for the evaluation. In most of these experiments, only a subset of the available training data is actually used for training. Thus, for each experiment, different combinations of the available training samples are possible. The variability regarding the selection of these training samples may have an effect in the performance of the networks. In order to take this variability into account, we performed 5 repetitions for each experiments using 5 different training subsets. These subsets are randomly selected from all the possible combinations of the available training data. The only exception to this procedure was the experiment with the whole training set, where all the training data was used for a single repetition. Additionally, in order to ensure a fair comparison, the same randomly selected training subsets were used for both frameworks.

Finally, the performance of both frameworks is compared against that of state-of-the-art approaches for fovea localization, optic disc localization, vessel segmentation, and optic disc segmentation. The objective of this comparison is to ensure that the proposed methods, despite being general and of straightforward use, can reach state-of-the-art performance in the tested tasks.

3.1. Datasets

The experiments presented in this paper were all conducted using five of the most representative publicly available datasets, which are described below:

- Isfahan MISP [39]: This dataset was used for the self-supervised pre-training consisting in the multimodal reconstruction between retinography and angiography. The dataset comprises 59 retinography-angiography pairs with image sizes of 720×576 pixels. Half of the samples correspond to pathological cases that were obtained from patients diagnosed with diabetic retinopathy. The other half correspond to healthy cases. All the images in this dataset are used for training.

- DRIVE [40]: This dataset was used for the training and evaluation of the blood vessel segmentation and optic disc localization. DRIVE is a collection of 40 retinographies with their corresponding ground truth vessel segmentations. The ground truth optic disc locations, instead, are not publicly available and were manually annotated by a clinical expert in our case. These annotations consist of the pixel coordinates for the optic disc center. The images present a size of 565×584 pixels and the approximate optic disc radius is 40 pixels. This value is used as the saturation distance β in Equation 5 to compute the optic disc location maps. We use the standard split for this dataset, which results in 20 images used as training set and the remaining 20 images hold out for the evaluation.

- DRIONS [41]: This dataset was used for the training and evaluation of the optic disc segmentation. DRIONS includes a collection of 110 retinographies with their corresponding ground truth optic disc segmentations. The images have a size of 700×605 pixels. We use the same data split that Maninis et al. [3], consisting of 60 images for training and the remaining 50 images hold out for the evaluation.

- IDRiD [42]: This dataset was used for the training and evaluation of the fovea localization. IDRiD contains 516 retinographies including different grades of diabetic retinopathy. The provided ground truth annotations for the fovea localization consist in the pixel coordinates of the fovea center. The images have a size of 4288×2848 pixels, being, therefore, significantly larger than the images from the Isfahan MISP dataset used for pre-training. The size of the retinal structures in the images also differs. For this reason, the images are rescaled to a fixed size of 858×570, for which the approximate optic disc radius is 50 pixels. This value is used as the saturation distance β in Equation 5 to compute the fovea location maps. We use the standard split for this dataset, consisting of 413 images for training and the remaining 103 images hold out for the evaluation.

- MESSIDOR [8]: This dataset was used for the evaluation of the fovea localization. MESSIDOR is a collection of 1200 retinographies including different grades of diabetic retinopathy. From them, we use

1136 images, for which the ground truth fovea localizations were provided by Gegundez-Arias et al. [22]. The dataset includes images of three different sizes. As happens with IDRiD, the scale of the retinal structures is significantly different to that of the pre-training dataset. Therefore, the original image sizes of 2240 × 1488, 1440 × 960, and 2304 × 1536 are rescaled to 1120 × 744, 1080 × 720, and 1152 × 768, respectively, to match the scale of the other datasets. The approximate optic disc radii are also provided by Gegundez-Arias et al. [22] and are rescaled in the same proportion than the images. In this case, all the images are used as test set for comparison with the state-of-the-art.

3.2. Evaluation metrics

For the localization of the optic disc and the fovea, the performance was evaluated following the strategy that is typically used in the literature [23, 4]. First, the euclidean distance between the predicted location and the ground truth location is computed. If this distance is lower than a certain threshold, the prediction is considered successful. The accuracy, defined as the ratio between the successful predictions and the total number of images, is used for the assessment of the performance. In order to obtain a more complete analysis, this accuracy is computed using different progressive thresholds. Particularly, we use R, R/2, and R/4, where R denotes the approximate optic disc radius, which is indicated for each dataset in Section 3.1. Additionally, the average distance in pixels is also used as evaluation metric.

Regarding the segmentation tasks, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were used to assess the performance. Both curves are commonly used in binary decision problems, allowing the evaluation of the generated probability maps without selecting the decision threshold. Note that the difference between ROC and PR curves is significative when the target classes are unbalanced. In our case, for the vessels and the optic disc segmentation, the number of samples from the positive class, i.e., vessels or optic disc, is significantly lower than the number of samples from the negative class, i.e., background. In this scenario, PR curves are more sensitive to variations in the false positive number, which leads to a greater performance discrimination ability. Despite this, ROC curves are widely used in the literature as a default metric in retinal imaging, specially for vessel segmentation [31, 32]. For such reason, we include both complementary curves in our evaluation. Additionally, the area under the ROC curve (AUC-ROC) and the area under the Precision-Recall curve (AUC-PR) were used.

Finally, for all the target tasks, mean values and standard deviations of the evaluation metrics are computed from the 5 repetitions with 5 different trainings subsets that are performed for each experiment. Additionally, in the case of the segmentation tasks, mean ROC and PR curves are also computed. The only exception to this procedure happens for the experiments with the whole training set, given that all the training samples are used for a single repetition in that case.

3.3. Results

The results for the fovea localization and the optic disc localization are depicted in Figure 5 and Figure 6, respectively. It is observed that the use of the self-supervised multimodal pre-training improves the performance of the localization process of both anatomical structures. In particular, this improvement happens in terms of both average value and standard deviation. In the case of the fovea (Figure 5), the improvement is significant for any number of training samples, whereas in the case of the optic disc (Figure 6), the random initialization approach reaches the performance of the proposed method only when all the training data is used. The latter is due to the fact that the multimodal reconstruction framework has already almost converged to the maximum performance with a smaller number of annotated samples.

The results for blood vessel segmentation and optic disc segmentation are depicted in Figure 7 and Figure 8, respectively. It is observed that the use of the self-supervised multimodal pre-training also improves the performance for the segmentation of both anatomical structures. In the case of the optic disc segmentation (Figure 8), the random initialization approach reaches the performance of the proposed method when half the training data is used. As with the optic disc localization, this is due to the fact that the multimodal reconstruction framework has already converged. Regarding the blood vessel segmentation (Figure 7), the improvement is obtained using any number of training samples. In fact, as illustrated in the plots of Figures 7(b) and 7(c), the trend may have continued if more training samples were also used.

A notorious difference between the localization and the segmentation results is that the latter show a smaller difference between the two frameworks in the comparison. This is a consequence of the high performance that is already achieved by training the networks from scratch, which leaves only a little gap for improvement. However, even in this highly competitive scenario, the self-supervised multimodal pre-training gets to improve the performance.

In addition, the comparison with state-of-the-art methods is respectively shown in Table 1 for the fovea localization, Table 2 for the optic disc localization, Table 3 for the blood vessel segmentation, and Table 4 for the optic disc segmentation. It is observed that both the multimodal reconstruction and the random initialization frameworks reached competitive performance in all the studied tasks. However, we would like to remark that the proposed self-supervised multimodal pre-training approach leads to state-of-the-art performance with much less annotated data.

Regarding the fovea localization, our experiments were performed using the recently published IDRiD dataset. In order to perform a comparison with state-of-the-art approaches we include additional results of our proposal evaluated on the MESSIDOR dataset. This additional evaluation is performed using the networks that were previously trained using the IDRiD dataset. Table 1 shows that the results obtained for MESSIDOR are better than those obtained for IDRiD. We have to consider, in this case, the higher percentage of pathological cases and advanced severity stages that is present in the IDRID dataset.

In the case of the optic disc localization, existent approaches evaluated on the DRIVE dataset only report accuracy for a distance threshold of value R, i.e., the approximate optic disc radius. Thus, as additional reference, we include two representative works that were evaluated using the MESSIDOR dataset and manually labeled ground truths. In this case, the labels were not publicly available.

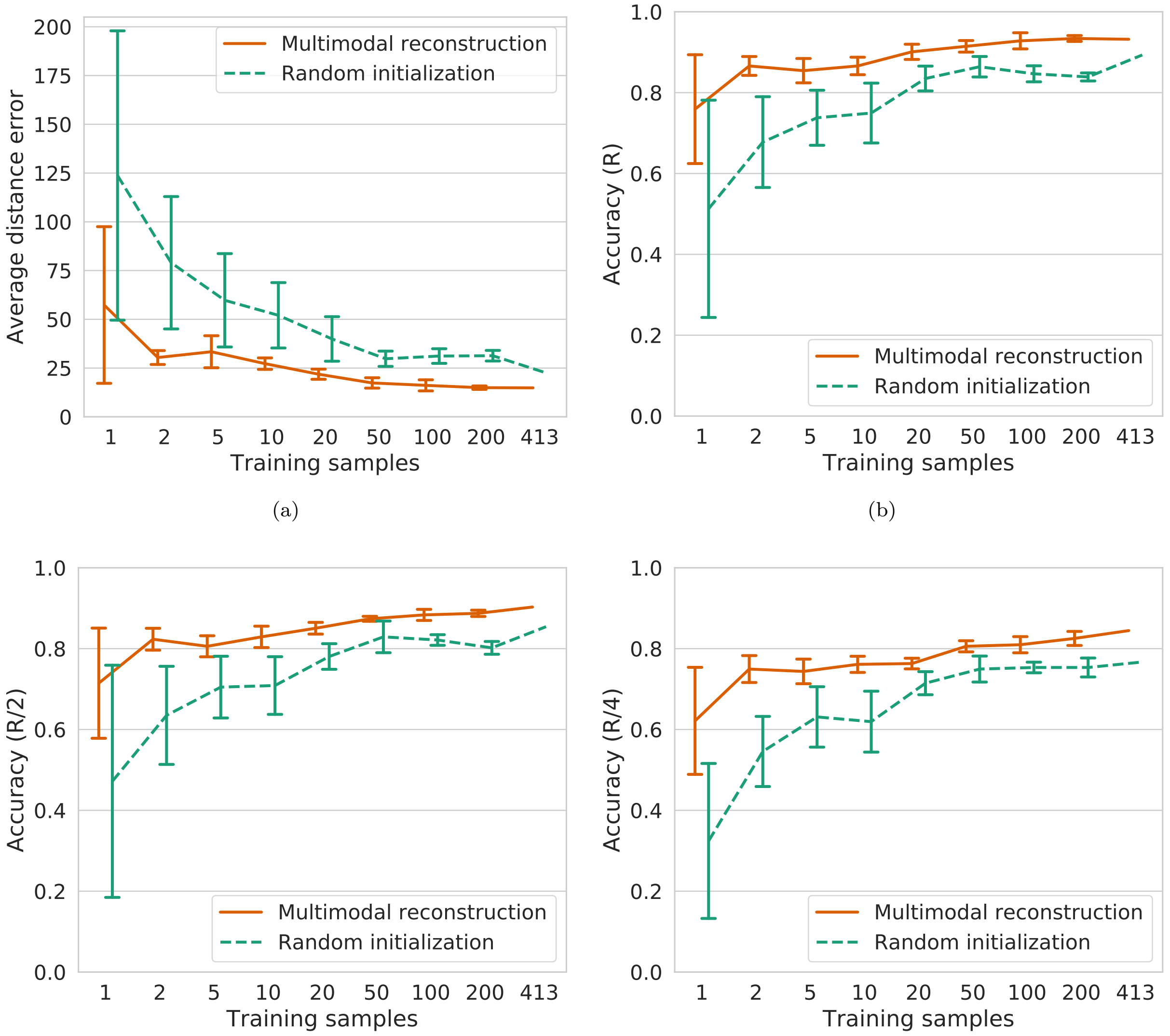
Finally, as illustration for qualitative comparison, examples of results obtained with the multimodal reconstruction and the random initialization frameworks are provided in Figures 9, 10, 11, and 12. In particular, Figures 9 and 10 depict representative examples of predicted location maps for the fovea and the optic disc, respectively. In addition, Figures 10 and 11 depict representative examples of predicted segmentation maps for the vasculature and the optic disc, respectively. All the examples correspond to images from the evaluation sets, and the ground truth annotations are provided as reference.

In general, it is observed that the multimodal reconstruction approach produces similar or even better results than the random initialization approach when all the training data is used. Nevertheless, due to the competitive performance of both frameworks, the visual comparison of the results can be difficult, requiring a more detailed analysis that is out of the scope of this paper. In contrast, when the training data is reduced, the contribution of the self-supervised multimodal pre-training is easier to appreciate with a rough visual analysis. In that sense, the improvement is especially significant when a single training sample is used, which represents the most challenging scenario in this scope.

Additionally, even greater improvement is that of the example in Figure 9 (b). In this case, the multimodal reconstruction leads to an important improvement when all the training data is used with respect to the random initialization counterpart. This is caused by the presence of lesions in the retina, which evidences that the proposed self-supervised multimodal pre-training presents the potential of being especially helpful in the more complex pathological cases.

4. Discussion

In this work, we address the problem of training DNNs for the localization and segmentation of the main anatomical structures of the eye fundus in retinography using scarce annotated data. To that end, we propose the use of the multimodal reconstruction between retinography and fluorescein angiography as a common self-supervised pre-training task; and the later fine-tuning of the pre-trained DNN for fovea localization, optic disc localization, blood vessel segmentation, and optic disc segmentation using a limited amount of task-specific annotated data. Given that obtaining the best possible results is not our main objective, we use the same network and training methodology for all the considered tasks. The only difference is the training loss, which requires to be specific for each kind of task: reconstruction, localization, or segmentation. Additionally, as neural network architecture, we employ the original U-Net [35], which is a reliable baseline



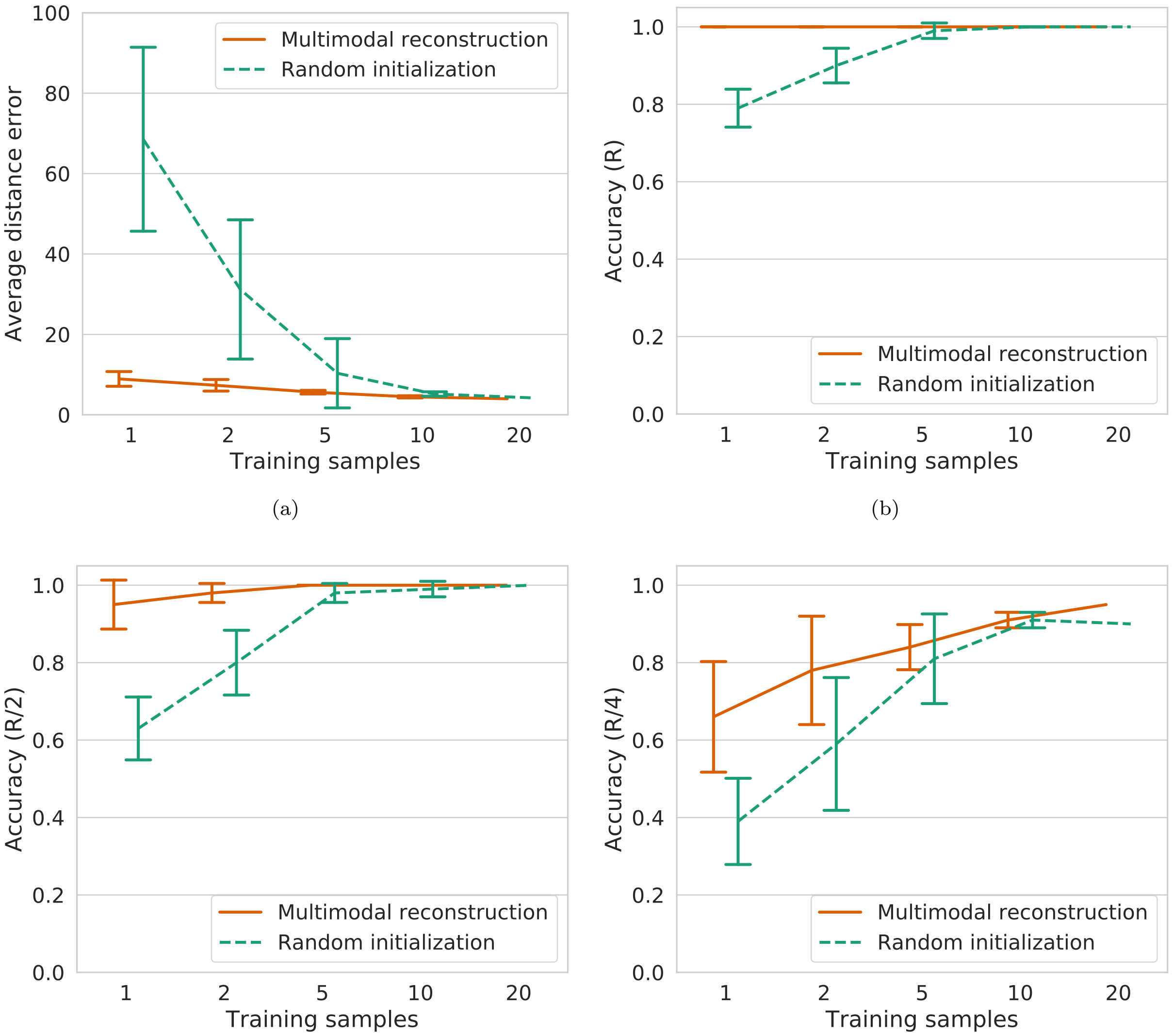
(a)

(b)

(c)

(d)

Figure 5: Results of the fovea localization for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Average distance error in pixels and ((b),(c),(d)) accuracy for (b) R, (c) R/2, and (d) R/4 criteria. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.



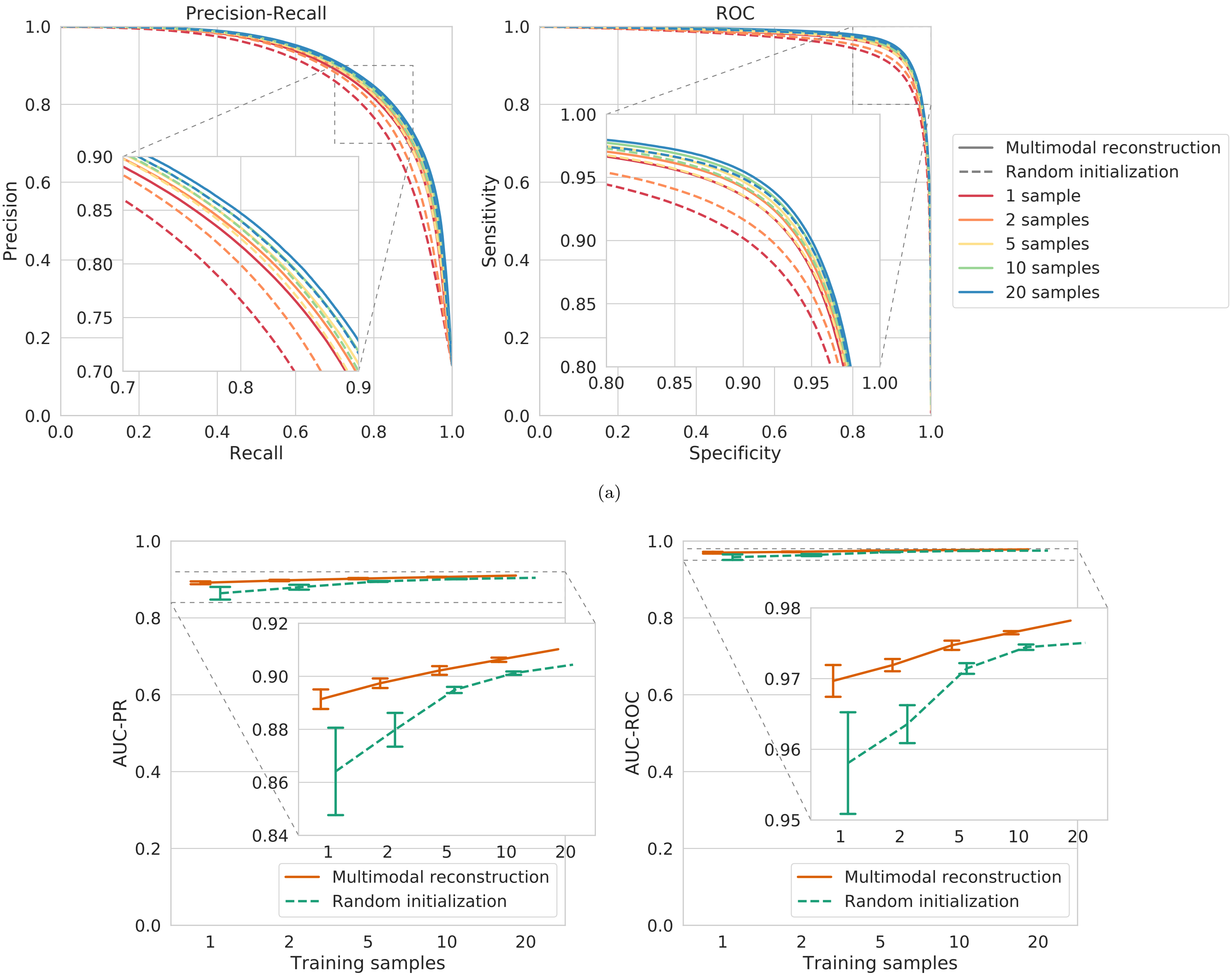
(a)

(b)

(c)

(d)

Figure 6: Results of the optic disc localization for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Average distance error in pixels and ((b),(c),(d)) accuracy for (b) R, (c) R/2, and (d) R/4 criteria. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

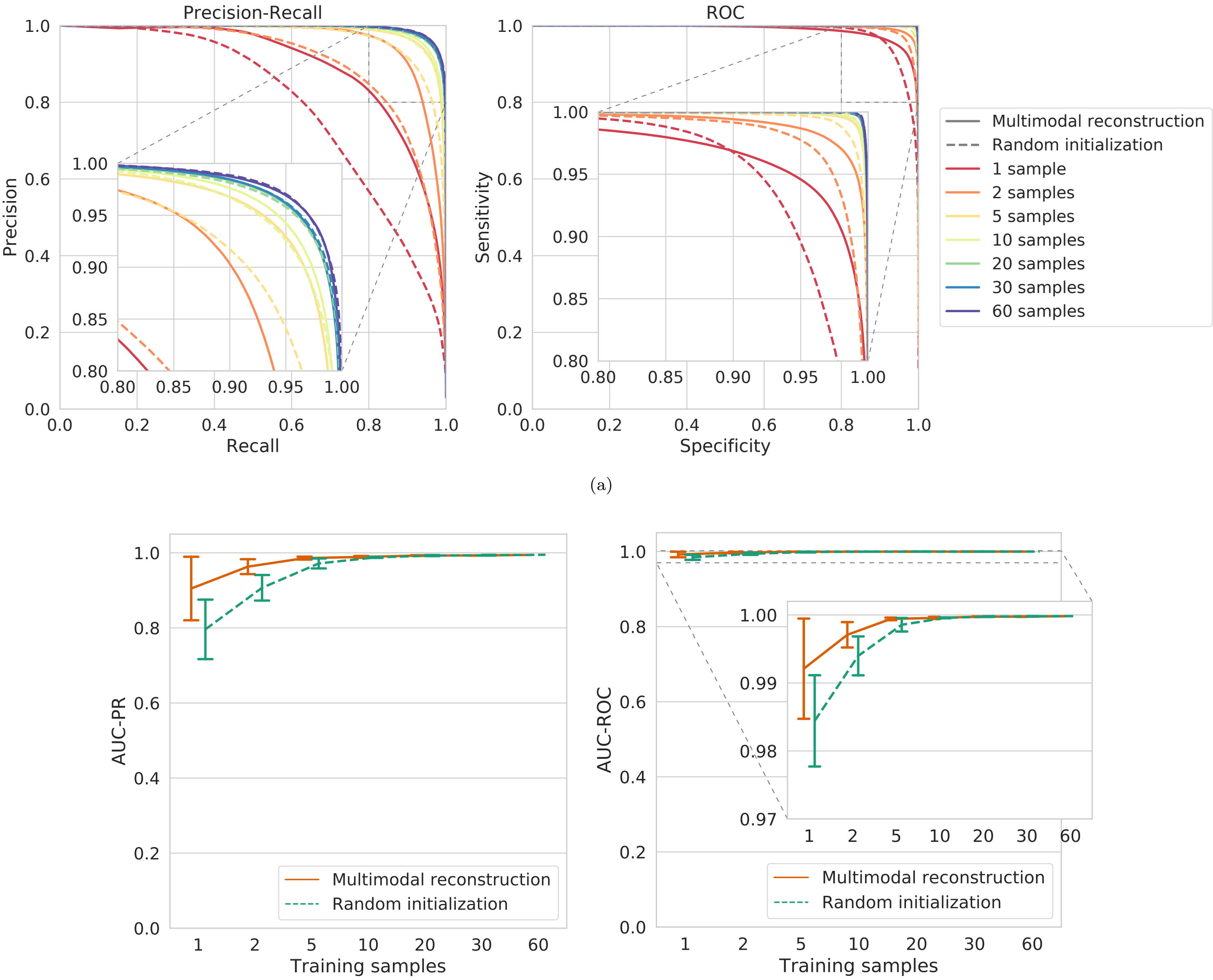


(a)

(b)

(c)

Figure 7: Results of the blood vessels segmentation for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Mean PR and ROC curves, (b) AUC-PR, and (c) AUC-ROC for a varying number of training samples. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.



(a)

(b)

(c)

Figure 8: Results of the optic disc segmentation for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Mean PR and ROC curves, (b) AUC-PR, and (c) AUC-ROC for a varying number of training samples. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

Table 1: Comparison with state-of-the-art methods for the fovea localization. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

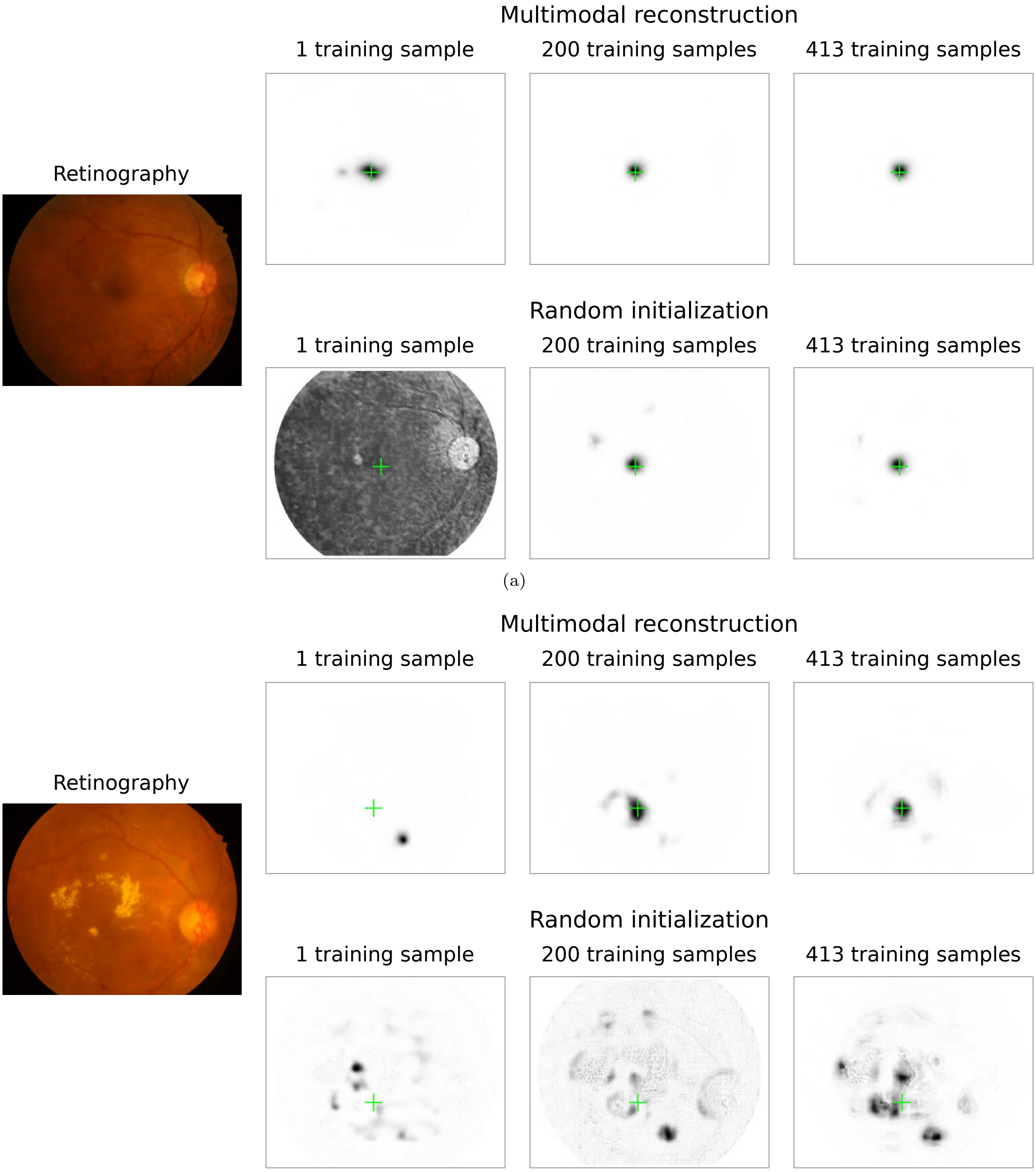
		Accuracy (%)		
		R	R/2	R/4
Evaluation on MESSIDOR				
Gegundez-Arias et al. [22]		96.50	95.88	94.25
Yu et al. [28]		98.00	94.00	64.88
Niemeijer et al. [21]		97.38	96.00	93.25
Dashtbozorg et al. [26]		98.87	93.75	66.50
Al-Bander et al. [4]		96.60	91.40	66.80
Meyer et al. [5]		99.74	97.71	94.01
Ours (1 image)	Random init.	54.52 ± 36.23	53.86 ± 36.01	48.98 ± 32.08
	Multimodal	86.09 ± 19.02	85.49 ± 19.18	80.07 ± 21.20
Ours (2 images)	Random init.	83.64 ± 9.41	83.17 ± 9.36	78.93 ± 9.51
	Multimodal	98.33 ± 0.59	97.94 ± 0.52	94.35 ± 1.21
Ours (200 images)	Random init.	99.47 ± 0.06	99.26 ± 0.09	97.02 ± 0.38
	Multimodal	99.84 ± 0.07	99.54 ± 0.13	97.80 ± 0.15
Ours (413 images)	Random init.	99.91	99.56	97.54
	Multimodal	100.00	99.65	97.98
Evaluation on IDRiD				
Ours (1 image)	Random init.	51.26 ± 26.87	47.18 ± 28.72	32.43 ± 19.17
	Multimodal	75.92 ± 13.46	71.46 ± 13.63	62.14 ± 13.24
Ours (2 images)	Random init.	67.77 ± 11.23	63.50 ± 12.14	54.56 ± 8.67
	Multimodal	86.60 ± 2.33	82.33 ± 2.70	74.95 ± 3.33
Ours (200 images)	Random init.	83.88 ± 0.99	80.19 ± 1.58	75.34 ± 2.35
	Multimodal	93.40 ± 0.73	88.74 ± 0.78	82.52 ± 1.74
Ours (413 images)	Random init.	89.32	85.44	76.70
	Multimodal	93.20	90.29	84.47

Table 2: Comparison with state-of-the-art methods for the optic disc localization. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

		Accuracy (%)		
		R	R/2	R/4
Al-Bander et al. [4]	(MESSIDOR)	97.00	95.00	83.60
Marin et al. [23]	(MESSIDOR)	99.75	99.50	97.75
Zhu et al. [24]		90.00	-	-
Qureshi et al. [25]		100.00	-	-
Dehghani et al. [27]		100.00	-	-
Ours (1 image)	Random init.	79.00 ± 4.90	63.00 ± 8.12	39.00 ± 11.14
	Multimodal	100.00 ± 0.00	95.00 ± 6.32	66.00 ± 14.28
Ours (2 images)	Random init.	90.00 ± 4.47	80.00 ± 8.37	59.00 ± 17.15
	Multimodal	100.00 ± 0.00	98.00 ± 2.45	78.00 ± 14.00
Ours (10 images)	Random init.	100.00 ± 0.00	99.00 ± 2.00	91.00 ± 2.00
	Multimodal	100.00 ± 0.00	100.00 ± 0.00	91.00 ± 2.00
Ours (20 images)	Random init.	100.00	100.00	90.00
	Multimodal	100.00	100.00	95.00

Table 3: Comparison with state-of-the-art methods for the blood vessels segmentation. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

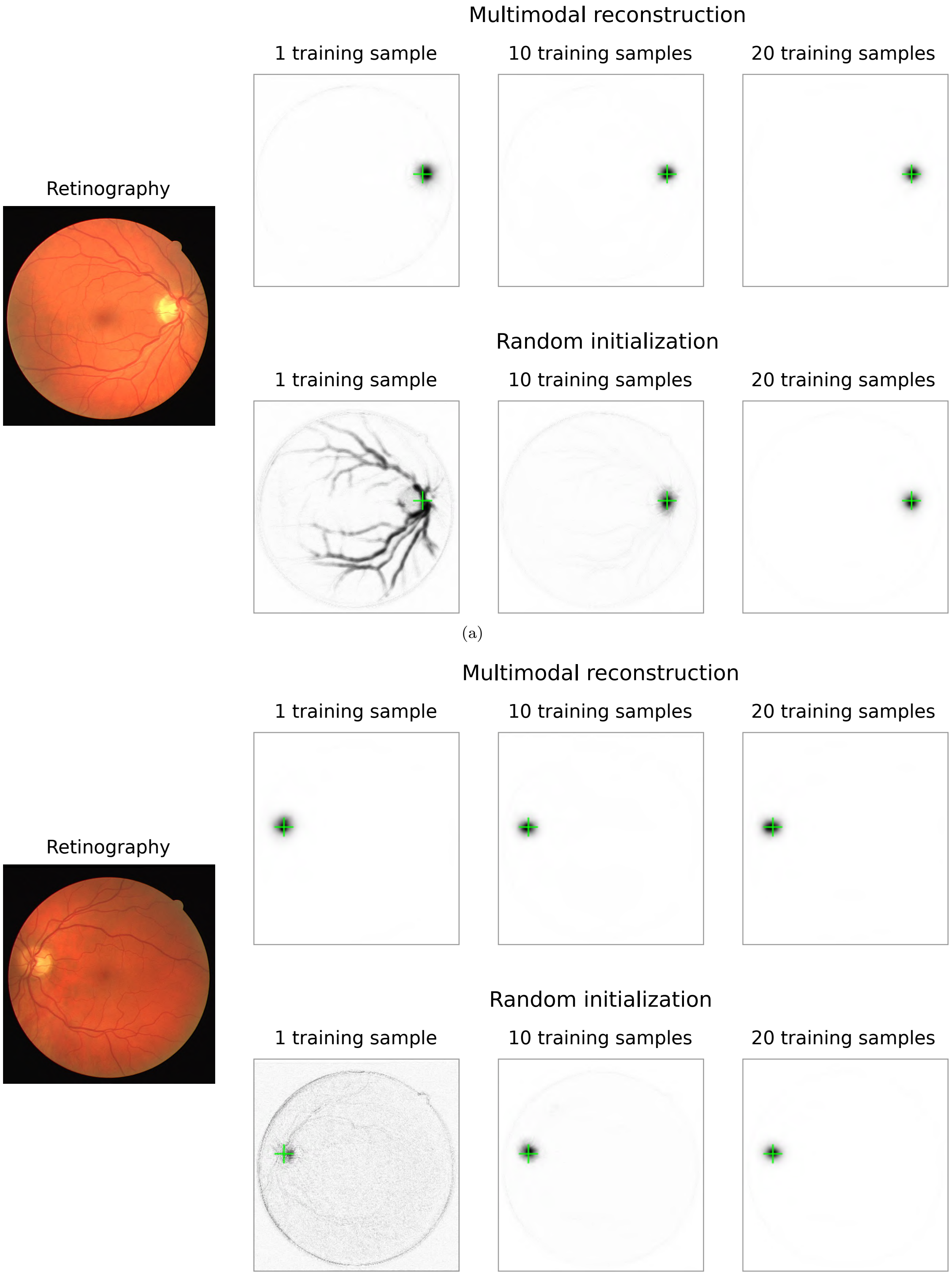
		AUC-PR (%)	AUC-ROC (%)
Fraz et al. [29]		-	97.47
Liskowski and Krawiec [31]		-	97.90
Li et al. [30]		-	97.38
Maninis et al. [3]		90.64	97.93
Mo and Zhang [32]		-	97.82
Ours (1 image)	Random init.	86.41 ± 1.65	95.81 ± 0.72
	Multimodal	89.14 ± 0.37	96.97 ± 0.22
Ours (2 images)	Random init.	87.98 ± 0.64	96.36 ± 0.27
	Multimodal	89.74 ± 0.18	97.19 ± 0.08
Ours (10 images)	Random init.	90.12 ± 0.06	97.44 ± 0.04
	Multimodal	90.62 ± 0.08	97.65 ± 0.02
Ours (20 images)	Random init.	90.44	97.51
	Multimodal	91.02	97.82



(a)

(b)

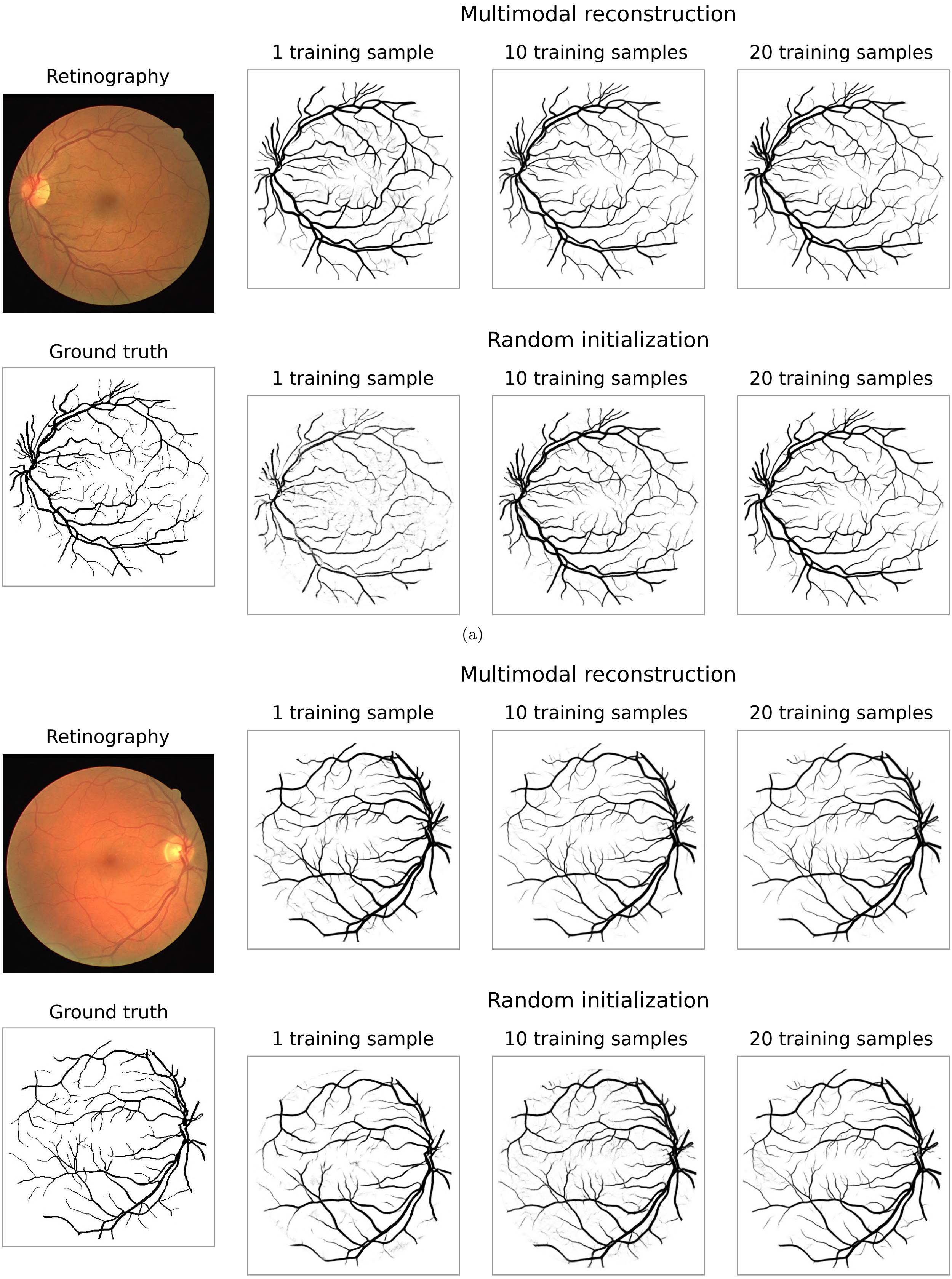
Figure 9: Examples of predicted location maps for the fovea using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The green cross depicts the ground truth location.



(a)

(b)

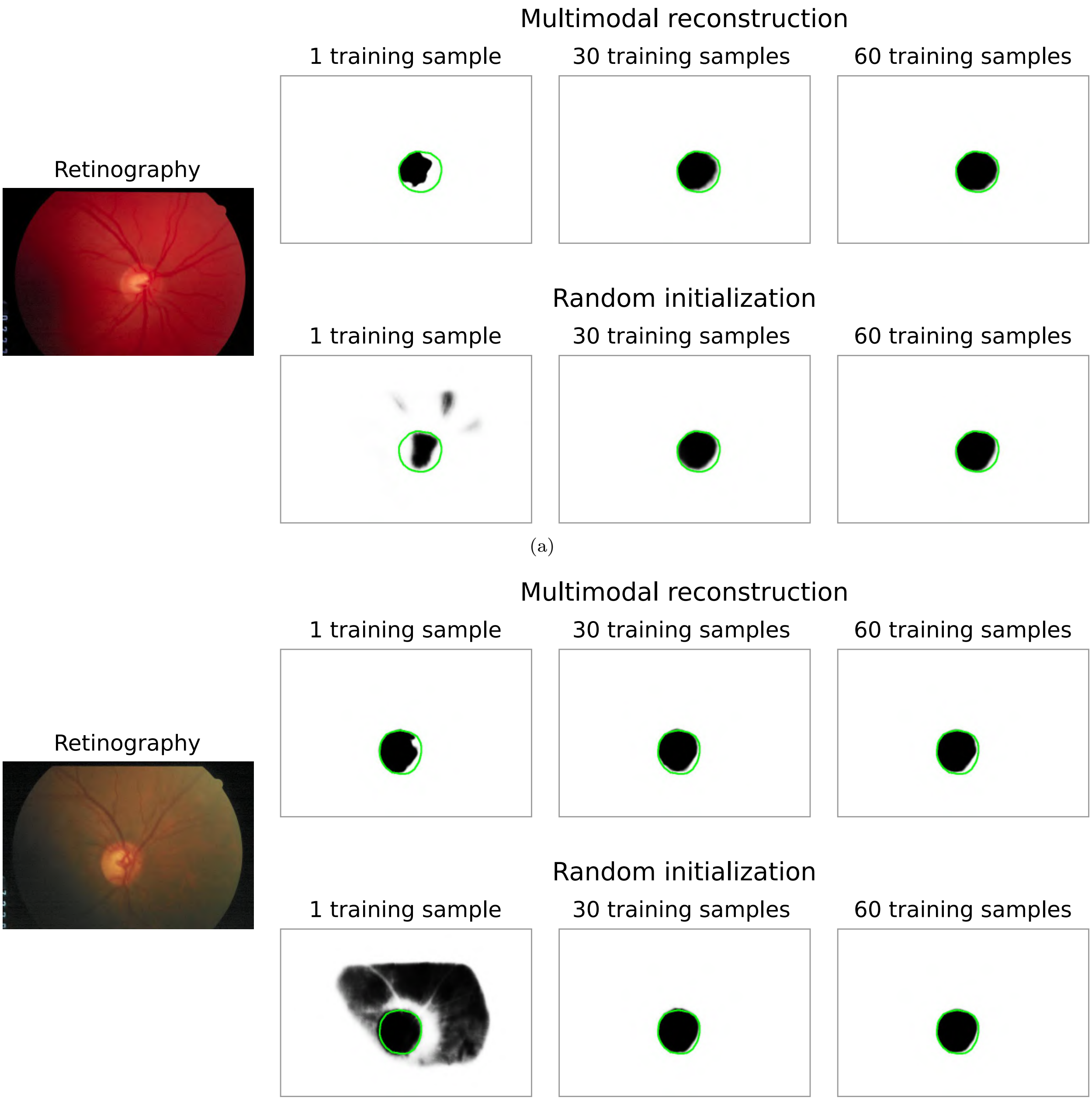
Figure 10: Examples of predicted location maps for the optic disc using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The green cross depicts the ground truth location.



(a)

(b)

Figure 11: Examples of predicted segmentation maps for the retinal vasculature using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization).



(a)

(b)

Figure 12: Examples of predicted segmentation maps for the optic disc using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The boundary of the ground truth segmentation is depicted in green.

Table 4: Comparison with state-of-the-art methods for the optic disc segmentation. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

		AUC-PR (%)	AUC-ROC (%)
Maninis et al. [3]		99.57	99.98
Ours (1 image)	Random init.	79.62 ± 7.93	98.44 ± 0.67
	Multimodal	90.50 ± 8.46	99.21 ± 0.74
Ours (2 images)	Random init.	90.69 ± 3.41	99.40 ± 0.29
	Multimodal	96.33 ± 1.98	99.71 ± 0.19
Ours (30 images)	Random init.	99.37 ± 0.10	99.98 ± 0.00
	Multimodal	99.31 ± 0.03	99.98 ± 0.00
Ours (60 images)	Random init.	99.49	99.98
	Multimodal	99.45	99.98

that was previously applied in this retinal context with a satisfactory performance [19]. Incidentally, the experimental results demonstrate that state-of-the-art performance can be achieved in all the studied tasks with the same network architecture and training strategy without further specific tuning.

From the comparison between the multimodal reconstruction and the random initialization frameworks, it is observed that the proposed self-supervised multimodal pre-training improves the obtained performance in all the studied tasks. Nevertheless, the extent of this improvement is not the same for all the tasks or training data sizes. The most remarkable improvement is observed in all the tasks when only few annotated images are used for training. In fact, the results that are obtained training from scratch with all the annotated data can be achieved using a fraction of the annotations if the networks are, instead, pre-trained with the proposed multimodal reconstruction. A negligible improvement of the proposed approach happens only for the cases where highly competitive performance is already obtained by the random initialization counterpart. Naturally, the beneficial effect of using the multimodal reconstruction as pre-training is limited by the room left for improvement by the baseline approach. For example, this is the case of some experiments involving the optic disc. However, in any case, the multimodal reconstruction approach converges to the maximum performance with less annotated data. In this regard, the results indicate that the optic disc localization and segmentation tasks are easier in comparison to the others in our experiments.

The provided comparison with state-of-the-art works shows that both the multimodal reconstruction and the random initialization frameworks produce competitive results when using all the training data. In that sense, the strong baseline ensures the practical relevance of the conclusions drawn from our analysis. Additionally, for some experiments, the random initialization framework behaves reasonably well with moderate reductions in the training data. This shows that modern data augmentation practices, adequate training schedules, and well designed loss functions are key to the successful application of DNNs to standard medical image analysis applications, without even needing any bells and whistles to fine tune the network architecture.

Regarding the self-supervised multimodal pre-training, the provided comparisons demonstrate that competitive results can also be achieved using a fraction of the total annotated training data. This is a strong result, indicating that clinical applications based on deep learning methods can be produced without requiring large amounts of manually annotated images. Additionally, we have demonstrated the advantages of the multimodal reconstruction as stand-alone transfer learning strategy. However, the proposed pre-training could also be applied together with other complementary self-supervised tasks in settings similar to those already explored in other domains [16]. In that sense, future works could explore the complementary application of the multimodal reconstruction and other self-supervised approaches in the medical domain.

Finally, other benefit of the proposed self-supervised pre-training is that, in general, the variability due to the use of different training samples is significantly reduced. However, this variability is still high when fewer annotations are used. Incidentally, this indicates that some images are considerably more adequate for training than others in order to achieve a better generalization. Thus, despite that a competitive performance can be achieved with very scarce annotations, for some applications this labeled data efficiency could be limited by the appropriate selection of particular training samples. In those situations, it would be interesting to explore the use of techniques aiming at the selection of the most informative images for being annotated.

5. Conclusions

Despite the great success of deep neural networks, the scarcity of annotated data is still a significant limiting factor to apply deep learning solutions to new clinical applications. In this regard, we propose to use the multimodal reconstruction as a self-supervised pre-training for different target tasks in the same application domain. We demonstrate the advantages of this proposal in the context of retinal image analysis. In particular, this work focuses on the localization and the segmentation of the main anatomical structures of the eye fundus, namely the fovea, the retinal vasculature, and the optic disc. For that purpose, we use the self-supervised multimodal reconstruction between retinography and fluorescein angiography to pre-train the networks.

The performed experiments demonstrate that using the multimodal reconstruction as self-supervised pre-training improves the performance of the considered target tasks. In particular, the proposed self-supervised transfer learning strategy allows to produce state-of-the-art results with a significant reduction of the annotated training data. This outcome has remarkable implications for future applications of neural networks in many fields of medical imaging where multimodal data can be easily gathered.

Acknowledgments

This work is 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 by Ministerio de Economía, Industria y Competitividad, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the ERDF and Xunta de Galicia through Grupo de Referencia Competitiva, ref. ED431C 2016-047, and from the European Social Fund (ESF) of the EU and Xunta de Galicia through the predoctoral grant contract ref. ED481A-2017/328. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

Conflict of interest

The authors declare no conflicts of interest.

References

[1] R. Besenczi, J. Tóth, A. Hajdu, A review on automatic analysis techniques for color fundus photographs, Computational and Structural Biotechnology Journal 14 (2016) 371–384, doi:10.1016/j.csbj.2016.10.001.

[2] E. D. Cole, E. A. Novais, R. N. Louzada, N. K. Waheed, Contemporary retinal imaging techniques in diabetic retinopathy: a review, Clinical & Experimental Ophthalmology 44 (4) (2016) 289–299, doi:10.1111/ceo.12711.

[3] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, L. V. Gool, Deep Retinal Image Understanding, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-319-46723-8_17, 2016.

[4] B. Al-Bander, W. Al-Nuaimy, B. M. Williams, Y. Zheng, Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc, Biomedical Signal Processing and Control 40 (2018) 91–101, doi: 10.1016/j.bspc.2017.09.008.

[5] M. I. Meyer, A. Galdran, A. M. Mendon, A Pixel-Wise Distance Regression Approach for Joint Retinal Optical Disc and Fovea Detection, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), doi: 10.1007/978-3-030-00934-2, 2018.

[6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Applied Soft Computing 70 (2018) 41–65, ISSN 15684946, doi:10.1016/j.asoc.2018.05.018.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60 – 88, doi:10.1016/j.media.2017.07.005.

[8] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, J.-C. Klein, Feedback on a punlicly distributed image database: the MESSIDOR database, Image Analysis & Stereology 33 (3) (2014) 231, doi:10.5566/ias.1155.

[9] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?, IEEE Transactions on Medical Imaging 35 (5) (2016) 1299–1312, doi:10.1109/TMI.2016.2535302.

[10] A. D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Transactions on Medical Imaging 19 (3) (2000) 203–210, doi:10.1109/42.845178.

[11] B. van Ginneken, A. A. A. Setio, C. Jacobs, F. Ciompi, Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: International Symposium on Biomedical Imaging (ISBI), doi:10.1109/ISBI.2015.7163869, 2015.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), doi:10.1109/CVPR.2009.5206848, 2009.

[13] A. I. Namburete, W. Xie, M. Yaqub, A. Zisserman, J. A. Noble, Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning, Medical Image Analysis 46 (2018) 1–14, doi:10.1016/j.media.2018.02.006.

[14] K. C. Wong, T. Syeda-Mahmood, M. Moradi, Building medical image classifiers with very limited data using segmentation networks, Medical Image Analysis 49 (2018) 105–116, doi:10.1016/J.MEDIA.2018.07.010.

[15] G. Maicas, A. P. Bradley, J. C. Nascimento, I. D. Reid, G. Carneiro, Training Medical Image Analysis Systems like Radiologists, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), doi:10.1007/978-3-030-00928-1_62, 2018.

[16] C. Doersch, A. Zisserman, Multi-task Self-Supervised Visual Learning, in: International Conference on Computer Vision (ICCV), doi:10.1109/ICCV.2017.226, 2017.

[17] T. Ross, D. Zimmerer, A. Vemuri, F. Isensee, M. Wiesenfarth, S. Bodenstedt, F. Both, P. Kessler, M. Wagner, B. Müller, H. Kenngott, S. Speidel, A. Kopp-Schneider, K. Maier-Hein, L. Maier-Hein, Exploiting the potential of unlabeled endoscopic video data with self-supervised learning, International Journal of Computer Assisted Radiology and Surgery 13 (6) (2018) 925–933, doi:10.1007/s11548-018-1772-0.

[18] A. Jamaludin, T. Kadir, A. Zisserman, Self-supervised Learning for Spinal MRIs, 294–302, doi:10.1007/978-3-319-67558-9_34, 2017.

[19] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-030-00928-1_37, 2018.

[20] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data, in: International Joint Conference on Neural Networks (IJCNN), 2019.

[21] M. Niemeijer, M. D. Abràmoff, B. van Ginneken, Fast detection of the optic disc and fovea in color fundus photographs, Medical Image Analysis 13 (6) (2009) 859–870, doi:10.1016/J.MEDIA.2009.08.003.

[22] M. E. Gegundez-Arias, D. Marin, J. M. Bravo, A. Suero, Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques, Computerized Medical Imaging and Graphics 37 (5-6) (2013) 386–393, doi:10.1016/j.compmedimag.2013.06.002.

[23] D. Marin, M. E. Gegundez-arias, A. Suero, J. M. Bravo, Obtaining optic disc center and pixel region by automatic thresholding methods on morphologically processed fundus images, Computer Methods and Programs in Biomedicine 118 (2) (2014) 173–185, doi:10.1016/j.cmpb.2014.11.003.

[24] X. Zhu, R. M. Rangayyan, A. L. Ells, Detection of the Optic Nerve Head in Fundus Images of the Retina Using the Hough Transform for Circles, Journal of digital imaging 23 (3) (2010) 332–341, doi:10.1007/s10278-009-9189-5.

[25] R. J. Qureshi, L. Kovacs, B. Harangi, B. Nagy, T. Peto, A. Hajdu, Combining algorithms for automatic detection of optic disc and macula in fundus images, Computer Vision and Image Understanding 116 (1) (2012) 138–145, doi:10.1016/j.cviu.2011.09.001.

[26] B. Dashtbozorg, J. Zhang, F. Huang, B. M. ter Haar Romeny, Automatic Optic Disc and Fovea Detection in Retinal Images Using Super-Elliptical Convergence Index Filters, doi:10.1007/978-3-319-41501-7_78, 2016.

[27] A. Dehghani, H. A. Moghaddam, M. S. Moin, Optic disc localization in retinal images using histogram matching, Eurasip Journal on Image and Video Processing 2012 (2012) 1–11, doi:10.1186/1687-5281-2012-19.

[28] H. Yu, S. Barriga, C. Agurto, S. Echegaray, M. Pattichis, G. Zamora, W. Bauman, P. Soliz, Fast localization of optic disc and fovea in retinal images for eye disease screening, in: Proceedings of SPIE, doi:10.1117/12.878145, 2011.

[29] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, S. A. Barman, An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation, IEEE Transactions on Biomedical Engineering 59 (9) (2012) 2538–2548, doi:10.1109/TBME.2012.2205687.

[30] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A Cross-Modality Learning Approach for Vessel Segmentation in Retinal Images, IEEE Transactions on Medical Imaging 35 (1) (2016) 109–118, doi:10.1109/TMI.2015.2457891.

[31] P. Liskowski, K. Krawiec, Segmenting Retinal Blood Vessels with Deep Neural Networks, IEEE Transactions on Medical Imaging 35 (11) (2016) 2369–2380, doi:10.1109/TMI.2016.2546227.

[32] J. Mo, L. Zhang, Multi-level deep supervised networks for retinal vessel segmentation, International Journal of Computer Assisted Radiology and Surgery 12 (12) (2017) 2181–2193, doi:10.1007/s11548-017-1619-0.

[33] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal Registration of Retinal Images Using Domain-Specific Landmarks and Vessel Enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), doi:10.1016/j.procs.2018.07.213, 2018.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612, doi:10.1109/TIP.2003.819861.

[35] O. Ronneberger, P.Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), doi:10.1007/978-3-319-24574-4_28, 2015.

[36] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: International Conference on Learning Representations, 2015.

[37] K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in: International Conference on Computer Vision (ICCV), doi:10.1109/ICCV.2015.123, 2015.

[38] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: International Conference on Learning Representations (ICLR), 2015.

[39] S. H. M. Alipour, H. Rabbani, M. R. Akhlaghi, Diabetic Retinopathy Grading by Digital Curvelet Transform, Computational and Mathematical Methods in Medicine 2012, doi:10.1155/2012/761901.

[40] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, B. van Ginneken, Ridge based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging 23 (4) (2004) 501–509, doi:/10.1109/TMI.2004.825627.

[41] E. J. Carmona, M. Rincón, J. García-Feijoó, J. M. Martínez-de-la Casa, Identification of the optic nerve head with genetic algorithms, Artificial Intelligence in Medicine 43 (3) (2008) 243–259, doi:10.1016/j.artmed.2008.04.005.

[42] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau, Indian Diabetic Retinopathy Image Dataset (IDRiD), doi:10.21227/H25W98, 2018.

3.3 Conference Paper: Multi-Modal Self-Supervised Pre-Training for Joint Optic Disc and Cup Segmentation in Eye Fundus Images

Multi-Modal Self-Supervised Pre-Training for Joint Optic Disc and Cup Segmentation in Eye Fundus Images

Álvaro S. Hervella[1,2], Lucía Ramos[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, l.ramos, jnovo, jrouco, mortega}@udc.es

[1] Centro de Investigación CITIC, Universidade da Coruña, Coruña, A Coruña (Spain)

[2] VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña (Spain)

Conference: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2020
Location: Budapest - Hungary
Date: May 4-8, 2020
Pages: 961–965
Editorial: IEEE
ISBN: 978-1-5090-6631-5

This is a post-peer-review, pre-copyedit version of an article published in Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. The final authenticated version is available online at: https://doi.org/10.1109/ICASSP40776.2020.9053551.

MULTI-MODAL SELF-SUPERVISED PRE-TRAINING FOR JOINT OPTIC DISC AND CUP SEGMENTATION IN EYE FUNDUS IMAGES

Álvaro S. Hervella*† *Lucía Ramos**† *José Rouco**† *Jorge Novo**† *Marcos Ortega**†

* Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain
† VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC),
Universidade da Coruña, A Coruña, Spain

ABSTRACT

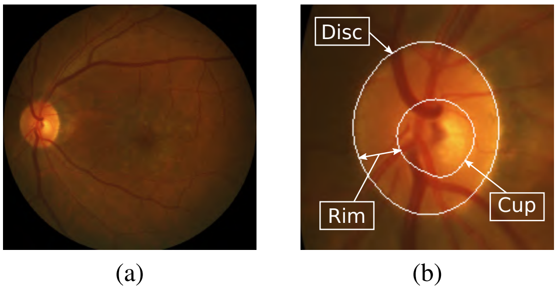
This paper presents a novel approach for the segmentation of the optic disc and cup in eye fundus images using deep learning. The accurate segmentation of these anatomical structures in the eye is important towards the early detection of glaucoma and, therefore, potentially avoiding severe vision loss. In order to improve the segmentation of the optic disc and cup, we propose a novel self-supervised pre-training consisting in the multi-modal reconstruction of eye fundus images. This novel approach aims at facilitating the segmentation task and avoiding the necessity of excessively large annotated datasets.

To validate the proposal, we perform several experiments on different public datasets. The results show that the proposed multi-modal self-supervised pre-training leads to a significant improvement in the performance of the segmentation task. Consequently, the presented approach shows remarkable potential towards further improving the interpretable and early diagnosis of a relevant disease as is glaucoma.

Index Terms— Deep learning, self-supervised learning, segmentation, eye fundus, glaucoma

1. INTRODUCTION

The analysis of the optic disc and cup is crucial in the diagnosis of glaucoma. This widely-spread eye disorder is one of the leading causes of irreversible vision loss in the world and it is characterized by an increased intraocular pressure [1]. One of the main consequences of this high intraocular pressure is the deformation of the optic disc, including the enlargement of the optic cup and the reduction of the neuroretinal rim [1]. For reference, Fig. 1 depicts a representative example of the optic disc region. This physical evidence of the disease has motivated the proposal of different biomarkers derived from the segmentation of the optic disc and cup [1].



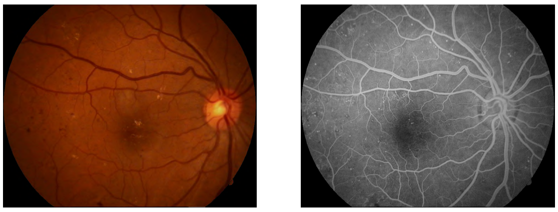
(a) (b)

Fig. 1. (a) Eye fundus. (b) Optic disc region from (a).

Given the prevalence and severity of the disease, there is an increasing interest in the development of automated methods for the screening and diagnosis of glaucoma. In that sense, the direct diagnosis from retinographies has been explored in some works [2]. However, the segmentation of relevant structures and the extraction of biomarkers is highly important towards producing an interpretable diagnosis and assist the clinicians in their decisions.

Recently, the use of Deep Neural Networks (DNNs) has been explored for the segmentation of the optic disc and cup, surpassing the performance of other more traditional methods [3]. In contrast with previous alternatives [4], the deep learning-based approaches do not require the ad-hoc design of complex algorithms. However, it may still be necessary to design adequate network architectures and training strategies. In this regard, previous works have explored the use of different network architectures, most of them consisting in modifications of the original U-Net [5], e.g., adding dense blocks [6], residual blocks [7], or multi-scale input-outputs [3]. Additionally, the lack of annotated data has motivated the development of novel strategies aimed at increasing the information for training DNNs. In this sense, the estimation of retinal depth maps has been proposed as a means to provide addi-

This work is 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 by Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the DPI2015-69948-R and RTI2018-095894-B-I00 research projects. The authors of this work also receive financial support from the ERDF and European Social Fund (ESF) of the EU, and Xunta de Galicia through Centro Singular de Investigación de Galicia, accreditation 2016-2019, ref. ED431G/01, Grupo de Referencia Competitiva, ref. ED431C 2016-047, and the predoctoral grant contract ref. ED481A-2017/328.



(a) (b)

Fig. 2. (a) Retinography and (b) FA from the same eye.

tional information to the networks [8]. Alternatively, Wang et al. [9] proposed a domain adaption technique that improves the generalization across datasets. This approach facilitates the use of additional annotated data when there is a lack of annotations in the target dataset, which is a common scenario in the literature.

In this context, we propose a novel approach for improving the segmentation of the optic disc and cup using DNNs without increasing the annotations. Instead, we use unlabeled multi-modal image pairs consisting of retinography and fluorescein angiography (FA) images. The latter represents an imaging modality complementary to retinography that requires the injection of a contrast dye. This invasive procedure facilitates the analysis of the vascular system while drastically changing the appearance of the retinal structures and lesions in the images. This effect can be observed in the examples of Fig. 2. Previous works have demonstrated that a DNN is able to learn the estimation of FA from retinography [10]. Given that this estimation is non-trivial, the DNN must have learned some knowledge about the different retinal structures. It should be possible, therefore, to exploit that learned knowledge for transfer learning purposes. In that sense, we propose to use the multi-modal reconstruction as pre-training task for the segmentation of the optic disc and cup. The idea is that if a DNN learns first to recognize the different retinal structures using the unlabeled multi-modal data, it should then be easier to learn how to segment the optic disc and optic cup.

In summary, in this work, we propose a novel and robust approach for the precise segmentation of the optic disc and cup in eye fundus images using DNNs. In particular, we use a self-supervised pre-training consisting in the multi-modal reconstruction of FA from retinography. This novel approach aims at improving the segmentation of the optic disc and cup using unlabeled multi-modal image pairs. In order to validate our proposal, we perform several experiments on different public datasets including both macula-centered and optic disc-centered retinographies.

2. METHODOLOGY

The proposed methodology is summarized in the diagram of Fig. 3. The pre-training phase consists in the generation of the invasive FA from retinography. This multi-modal reconstruction of the eye fundus is a self-supervised task that does not require manually annotated data for training. Instead, it takes advantage of the unlabeled multi-modal image pairs. The objective of the pre-training phase is to learn domain-specific patterns that are useful for the segmentation of the optic disc and cup in retinographies. Then, in the fine-tuning phase, the network training continues in the segmentation task using the annotated data and common supervised approaches. Finally, the trained network is able to precisely predict the optic disc and cup segmentation from the retinographies.

2.1. Pre-training: multi-modal reconstruction

The multi-modal reconstruction of FA from retinography is trained following the approach proposed by [10]. In that sense, we use a set of registered image pairs, which allows the use of full-reference metrics between the target and the network output as loss function. Specifically, in this case, we use the negative Structural Similarity (SSIM) as loss function:

$$\mathcal{L}^{Rec} = -\frac{1}{N}\sum_{n}^{N} SSIM(p_n, y_n) \tag{1}$$

where $\mathbf{p}$ denotes the predicted FA, $\mathbf{y}$ the target FA, and N the number of pixels. The SSIM between both images is computed as described in [10]. Despite that SSIM was initially proposed for quality assessment purposes [11], it has demonstrated superior performance in comparison to other alternatives regarding the multi-modal reconstruction [10].

2.2. Fine-tuning: optic disc and cup segmentation

A particularity of the optic disc and cup segmentation is that the optic cup is contained within the optic disc (see Fig. 1). This means that some pixels in the images belong to both optic disc and cup regions, which complicates to directly approach the problem as a multi-class classification. Thus, instead of directly predicting the optic disc region, we approach the prediction of the optic cup and the neuroretinal rim regions, which together cover the whole optic disc. In particular, the neural network is trained to predict the likelihood of each pixel belonging to: the background, the optic cup, and the neuroretinal rim. The training is performed as in [8] using the multi-class cross-entropy as loss function:

$$\mathcal{L}^{Seg} = -\frac{1}{N}\sum_{n}^{N}\sum_{c}^{C} y_{n,c} log(p_{n,c}) \tag{2}$$

where $\mathbf{p}$ denotes the network output, $\mathbf{y}$ the corresponding ground truth, N the number of pixels, and C the number of classes. Then, the likelihood of a pixel belonging to the optic disc is computed as the summation of the individual likelihoods for the optic cup and the neuroretinal rim. Additionally, the normalized likelihoods in the network output are obtained using a Softmax activation function.

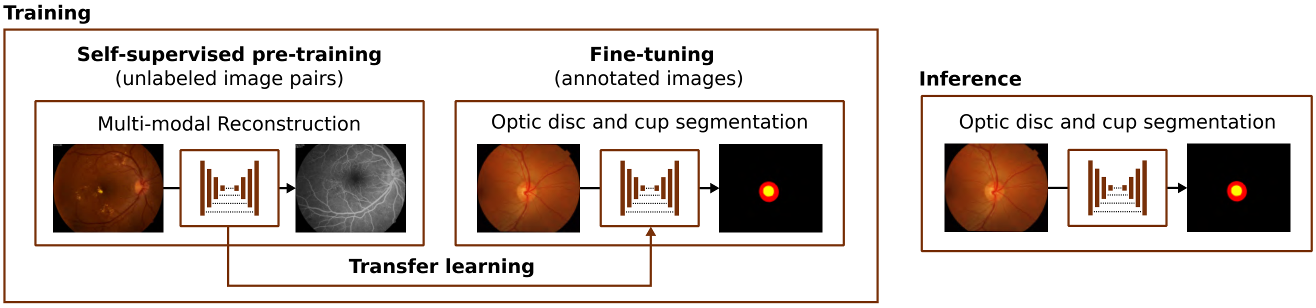

Training
Self-supervised pre-training
(unlabeled image pairs)
Multi-modal Reconstruction
Fine-tuning
(annotated images)
Optic disc and cup segmentation
Transfer learning
Inference
Optic disc and cup segmentation

Fig. 3. Methodology for the joint optic disc and cup segmentation using the proposed self-supervised pre-training.

2.3. Network architecture and training details

In this work, we adapt the original U-Net [5]. This neural network represents a reliable baseline that is commonly used to solve segmentation problems in different domains, including the analysis of the eye fundus [12]. In brief, U-Net is a fully convolutional network characterized by a symmetric encoder-decoder structure and skip connections between the layers of the encoder and the decoder. The network in our experiments presents the same number of layers and channels as the original one [5]. The only difference is the output layer, which must be adapted for each specific task. In the case of the multi-modal reconstruction, we use a 1-channel output with linear activation function. Instead, for the segmentation task, we use a 3-channel output with Softmax activation function.

The network initialization is performed following the approach proposed by He et al. [13]. Then, for the network training, we use the Adam optimization algorithm [14] with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the multi-modal reconstruction, we use the learning rate settings proposed in [10]. For the segmentation task, the learning rate is initialized with a value of $1e-5$ and reduced by a factor of 10 when the validation loss does not improve for 10 epochs. Then, the training is stopped after 20 epochs without improvement. Additionally, we apply data augmentation consisting of random scaling, rotation, and color-intensity transformations, in both pre-training and fine-tuning.

3. EXPERIMENTS, RESULTS AND DISCUSSION

The unlabeled multi-modal data consists of 59 image pairs from the public Isfahan MISP [15] database. Of these images, half of them belong to diabetic retinopathy patients whereas the other half are from healthy individuals. The image pairs are registered following the approach proposed in [16].

For the optic disc and cup segmentation, we use two different public datasets: DRISHTI-GS [17] and REFUGE [18]. DRISHTI-GS contains 101 annotated images centered at the optic disc, of which 70 correspond to glaucomatous eyes. REFUGE contains annotated 800 images centered at the macula, of which 81 correspond to glaucomatous eyes. Additionally, REFUGE includes 400 *Test* images for which the ground truth is not publicly available. These latter images are not included in the experiments.

In order to analyze all the images at the same scale, the images are rescaled to a Field of View (FOV) diameter of 720 pixels. This is the original scale for the Isfahan MISP images.

Regarding the quantitative evaluation, we use the most common metrics in the literature [7], namely the Dice score (D) and the Jaccard index (J).

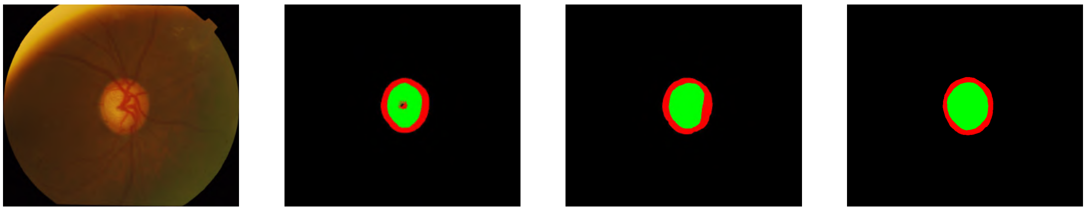
3.1. Evaluation of the proposed approach

To evaluate the proposed approach, we compare the performance of the networks pre-trained on the multi-modal reconstruction against the networks trained from scratch on the segmentation task.

Figure 4 depicts representative examples of predicted segmentations. These examples show that the proposed approach leads to successfully detect and precisely segment the optic disc and cup in the images. In contrast, the network trained from scratch does not always produce an adequate prediction, resulting in some cases in inconsistent segmentations. This evidences that the proposed multi-modal self-supervised pre-training provides the network with a better understanding of the retinal anatomy.

In order to produce a robust and reliable quantitative evaluation, we follow a 5×2-fold cross-validation approach, where 10 experiments with different training-test splits are performed. Table 1 shows the results of these experiments. It is observed that the proposed approach significantly improves the segmentation of the optic disc and cup in both datasets. Nevertheless, the improvement is greater for the optic cup and also for the experiments in the DRISHTI-GS dataset.

With regards to the difference between optic disc and optic cup, it must be noticed that the segmentation of the optic cup represents a harder problem. This is mainly due to the less defined boundary of the optic cup (see Fig. 1). Also, the optic cup is the region affected in the cases with glaucoma, resulting in higher variability. In that sense, the results show that it is precisely in the more challenging optic cup segmentation where the proposed approach is more beneficial. Regarding



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

Fig. 4. (a) Retinography and corresponding ((b),(c)) predictions and (d) ground truth. (b) From scratch. (c) Proposed.

Table 1. Quantitative results by means of Jaccard index (%).

Dataset	Method	Optic cup	Optic disc
DRISHTI-GS	From scratch	74.81 ± 1.91	88.15 ± 1.67
	Proposed	81.23 ± 0.72	91.01 ± 1.33
REFUGE	From scratch	76.51 ± 1.83	91.71 ± 0.25
	Proposed	79.02 ± 0.43	92.25 ± 0.22

the differences between both datasets, it must be considered that the REFUGE dataset is significantly larger. This results in more annotated data for training the network. In this sense, the results show that the proposed approach is more beneficial when the annotated data for training is more limited. Consequently, this means that the multi-modal self-supervised pre-training successfully compensates the lack of annotations.

3.2. Comparison with the state-of-the-art

Table 2 shows the comparison of the proposed approach against state-of-the-art methods in the DRISHTI-GS dataset. For this comparison, we use the original training-test split of DRISHTI-GS [17]. It is observed that the proposed approach offers competitive performance, even obtaining the best results for the segmentation of the optic cup. Moreover, it is important to notice that our approach achieves these competitive results using considerably fewer annotations than any of the other works. In that sense, previous works typically compensate for the small size of the DRISHTI-GS dataset by gathering additional annotated data. In contrast, the proposed self-supervised pre-training successfully compensates for the lack of annotations by taking advantage of the unlabeled multi-modal images. Additionally, in our experiments, the networks are directly applied over the whole images, whereas previous works typically operate by first cropping a patch containing the optic disc region. Although this step facilitates the segmentation, it adds complexity to the final methodology. In that sense, our proposal allows for a more straightforward solution, completely based on end-to-end learning.

4. CONCLUSIONS

In this work, we propose a novel approach for the segmentation of the optic disc and cup in eye fundus images using deep learning. In particular, we use a novel self-supervised pre-training consisting in the multimodal reconstruction of complementary eye fundus images. This multi-modal self-supervised pre-training provides the network with a better understanding of the retinal anatomy, which facilitates the segmentation task and reduces the necessity of excessively large annotated datasets.

In order the validate the proposal, we perform several experiments on different public datasets, including macula-centered and optic disc-centered images. The results show that the multi-modal self-supervised pre-training significantly improves the segmentation performance. Hence, the proposed approach presents an important potential towards improving the interpretable and early diagnosis of glaucoma. In that sense, future works will explore the automated extraction of relevant biomarkers for this eye disorder.

Table 2. State-of-the-art comparison for the DRISHTI-GS dataset by means of Jaccard index (J) and Dice score (D).

Method	Training data	Optic cup		Optic disc	
		$J(\%)$	$D(\%)$	$J(\%)$	$D(\%)$
Al-Bander et al. (2018) [19]	455 annotated images	71.13	82.82	90.42	94.90
Yu et al. (2019) [7]	655 + 50 annotated images	80.42	88.77	94.92	97.38
Shankaranarayana et al. (2019) [8]	325 + 50 annotated images	–	84.8	–	96.3
Wang et al. (2019) [9]	400 annotated images + 50 unlabeled images	–	90.1	–	97.4
Ours (Proposed)	50 annotated images + 59 unlabeled image pairs	82.29	90.29	92.43	96.07
Ours (From scratch)	50 annotated images	75.36	85.95	88.19	93.72

5. REFERENCES

[1] M. C. V. Stella Mary, E. B. Rajsingh, and G. R. Naik, "Retinal fundus image analysis for diagnosis of glaucoma: A comprehensive survey," *IEEE Access*, vol. 4, pp. 4327–4354, 2016.

[2] Yuki Hagiwara, Joel En Wei Koh, Jen Hong Tan, Sulatha V. Bhandary, Augustinus Laude, Edward J. Ciaccio, Louis Tong, and U. Rajendra Acharya, "Computer-aided diagnosis of glaucoma using fundus images: A review," *Computer Methods and Programs in Biomedicine*, vol. 165, pp. 1 – 12, 2018.

[3] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

[4] Niharika Thakur and Mamta Juneja, "Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma," *Biomedical Signal Processing and Control*, vol. 42, pp. 162 – 189, 2018.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[6] Baidaa Al-Bander, Bryan M. Williams, Waleed Al-Nuaimy, Majid A. Al-Taee, Harry Pratt, and Yalin Zheng, "Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis," *Symmetry*, vol. 10, no. 4, pp. 87, 2018.

[7] Shuang Yu, Di Xiao, Shaun Frost, and Yogesan Kanagasingam, "Robust optic disc and cup segmentation with deep learning for glaucoma detection," *Computerized Medical Imaging and Graphics*, vol. 74, pp. 61 – 71, 2019.

[8] S. M. Shankaranarayana, K. Ram, K. Mitra, and M. Sivaprakasam, "Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1417–1426, July 2019.

[9] S. Wang, L. Yu, X. Yang, C. Fu, and P. Heng, "Patch-based output space adversarial learning for joint optic disc and cup segmentation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[10] Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega, "Retinal image understanding emerges from self-supervised multimodal reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018.

[11] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Self-supervised deep learning for retinal vessel segmentation using automatically generated labels from multimodal data," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision (ICCV)*, 2015.

[14] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012.

[16] Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega, "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," *Procedia Computer Science*, vol. 126, pp. 97 – 104, 2018.

[17] Jayanthi Sivaswamy, SR Krishnadas, and Arunava Chakravarty, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, 2015.

[18] José Ignacio Orlando et al., "REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs," *Medical Image Analysis*, p. 101570, 2019.

[19] Baidaa Al-Bander, Waleed Al-Nuaimy, Bryan M. Williams, and Yalin Zheng, "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc," *Biomedical Signal Processing and Control*, vol. 40, pp. 91–101, 2018.

3.4 Journal Paper: Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images

Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], Manuel G. Penedo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, jrouco, mgpenedo, mortega}@udc.es

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña (Spain)
[2] Department of Computer Science, University of A Coruña, A Coruña (Spain)

Journal: Computer Methods and Programs in Biomedicine
Volume: 181
Pages: 105201
Editorial: Elsevier
ISSN: 0169-2607
Year: 2020

This is a post-peer-review, pre-copyedit version of an article published in the journal Computer Methods and Programs in Biomedicine. The final authenticated version is available online at: https://doi.org/10.1016/j.cmpb.2019.105201.

Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Manuel G. Penedo[a,b], Marcos Ortega[a,b]

[a] *CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain*
[b] *Department of Computer Science, Universidade da Coruña, A Coruña, Spain*

Abstract

Background and objectives The analysis of the retinal vasculature plays an important role in the diagnosis of many ocular and systemic diseases. In this context, the accurate detection of the vessel crossings and bifurcations is an important requirement for the automated extraction of relevant biomarkers. In that regard, we propose a novel approach that addresses the simultaneous detection of vessel crossings and bifurcations in eye fundus images.

Method We propose to formulate the detection of vessel crossings and bifurcations in eye fundus images as a multi-instance heatmap regression. In particular, a deep neural network is trained in the prediction of multi-instance heatmaps that model the likelihood of a pixel being a landmark location. This novel approach allows to make predictions using full images and integrates into a single step the detection and distinction of the vascular landmarks.

Results The proposed method is validated on two public datasets of reference that include detailed annotations for vessel crossings and bifurcations in eye fundus images. The conducted experiments evidence that the proposed method offers a satisfactory performance. In particular, the proposed method achieves 74.23% and 70.90% F-score for the detection of crossings and bifurcations, respectively, in color fundus images. Furthermore, the proposed method outperforms previous works by a significant margin.

Conclusions The proposed multi-instance heatmap regression allows to successfully exploit the potential of modern deep learning algorithms for the simultaneous detection of retinal vessel crossings and bifurcations. Consequently, this results in a significant improvement over previous methods, which will further facilitate the automated analysis of the retinal vasculature in many pathological conditions.

Keywords: deep learning, eye fundus, blood vessels, crossings, bifurcations, landmark detection

1. Introduction

The retinal vascular tree is a complex structure formed by arteries and veins that intersect and bifurcate frequently over all the eye fundus. The analysis of this structure plays an important role in the diagnosis and follow-up of numerous diseases. In particular, the retina is the only organ of the human body where the vascular system can be studied in vivo and without invasive procedures [1]. This makes the analysis of the retinal vasculature relevant for the clinical assessment of both ocular and systemic diseases, such as age-related macular degeneration, diabetes, hypertension, or atherosclerosis, among others [2].

An exhaustive analysis of the retinal vasculature requires the recognition of the vessel crossings and bifurcations, representing the landmarks where blood vessels intersect or bifurcate, respectively. As reference, Figure 1 depicts representative examples of these characteristic points in the eye fundus. The localization and identification of these landmarks has important clinical applications. For instance, the analysis of

[*] Corresponding author
Email address: a.suarezh@udc.es (Álvaro S. Hervella)

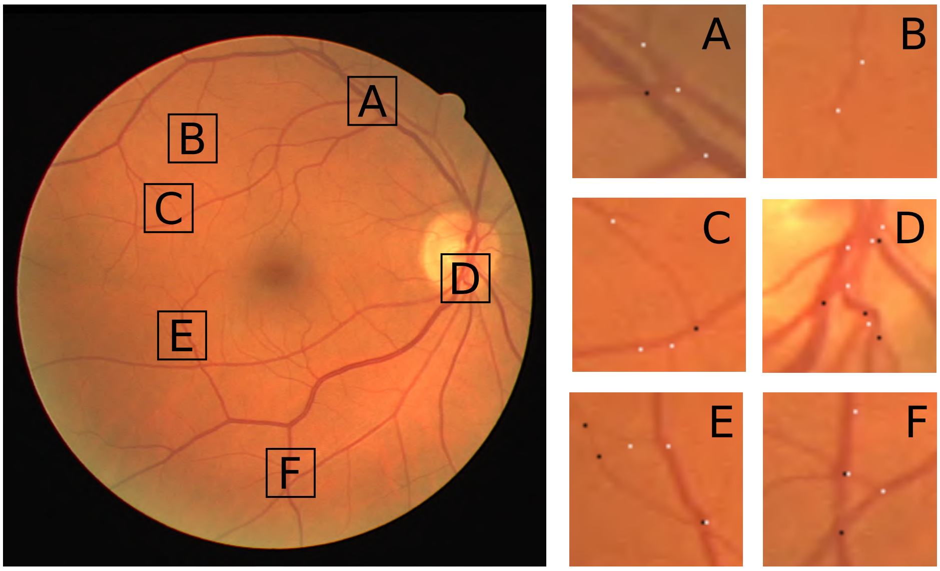


Figure 1: Example of eye fundus image including cropped regions that depict vessel crossings and bifurcations in detail. The black dots represent crossings whereas the white dots represent bifurcations.

the bifurcations provides measurements like the bifurcation angles which have been studied as biomarkers for hypertension and other cardiovascular diseases [1]. The identification of the crossings, instead, allows studying the presence of arteriovenous nicking, which happens when an artery compresses a vein. This pathological condition is associated with the development of retinal vein occlusion and it is also indicative of hypertension, among other relevant diseases [3].

Besides the direct analysis of the vessel crossings and bifurcations, these characteristic points are commonly used as reference in many heterogeneous procedures related to the automated analysis of the retinal vasculature [4, 5]. Moreover, vessel-tracking techniques that are commonly used for the measurement of vessel widths and tortuosity estimation may be affected by an inadequate identification of the constituent crossings and bifurcations [6, 1]. Additionally, these characteristic points can be used as landmarks for the registration of eye fundus images using point matching algorithms [7]. The complexity of the retinal vascular tree, which is unique for each eye, also allows the use of these landmarks as a reliable biometric pattern [8].

The importance of the vessel crossings and bifurcations means that the improvements in their identification present a potential carryover to numerous applications. In that sense, related significative problems such as vasculature segmentation [9] or microaneurysm detection [10] have benefited from the use of Deep Neural Networks (DNNs). The deep learning-based approaches do not require the ad-hoc design of complex algorithms and typically provide an improved performance in comparison with traditional methods [11]. However, the novel use of DNNs may not always be straightforward.

In the case of tasks such as segmentation or classification, a DNN can be directly trained by optimizing a similarity metric between the network outputs and the target binary labels. However, the ground truth labels for the detection of crossings and bifurcations consist of two independent sets of pixel coordinates, one for each type of landmark. In that case, the selection of the most adequate training objective is not straightforward. Additionally, both the number of vascular landmarks in the images and their approximate spatial distribution are unknown, given that the patterns described by the retinal vascular tree are unique for each eye. Thus, the challenge of this task is to adequately formulate the problem to take full advantage of the capacity of a DNN.

In this work, we propose to formulate the detection of retinal vessel crossings and bifurcations as a multi-instance heatmap regression. In that sense, we convert the prediction of pixel coordinates into the regression of heatmaps representing the location of multiple landmarks. The prediction of these multi-instance heatmaps can be easily learned by a DNN using common regression metrics as loss function. Then, the precise location of the crossings and bifurcations is obtained by extracting the local maxima in the predicted multi-instance heatmaps. This novel approach allows to address the prediction of an unknown number of landmarks while using a DNN applied over full images of arbitrary sizes. In this setting, the simultaneous detection of crossings and bifurcations is directly enabled by training the network to predict multiple heatmaps, one for each type of landmark. Therefore, the proposed approach allows to perform the detection and distinction of vessel crossings and bifurcations integrated into a single step. In order to validate our proposal, several representative experiments are performed using two public datasets of reference that include ground truth manual annotations for both vessel crossings and bifurcations.

1.1. Related work

In the literature, several works have approached the detection of vessel crossings and bifurcations in eye fundus images. The most commonly followed strategy is to split the problem into two different tasks: the general detection of vessel junctions, and the later classification of the detected junctions as crossings or bifurcations [12, 13]. Additionally, there are several works that only tackle the first task, without facing the complex and difficult distinction between both types of landmarks [14, 15].

Regarding the first task, a recurrent approach for the detection of vessel junctions is to start by segmenting the blood vessels. Then, a thinning algorithm is used to obtain the skeleton of the vascular tree, being the vessel junctions extracted after a topological analysis of this skeleton [16, 17]. In this regard, Fahti et al. [18] propose to perform a joint analysis of both the skeleton and the segmented vessels. In these skeleton-based approaches, the most challenging part corresponds to the identification of the vessel crossings, given that, in the obtained skeletons, many crossings are represented as two close bifurcations [16]. In that sense, the classification between crossings and bifurcations is typically performed using geometrical features such as the connectivity [16], the vessel angles [19], and the vessel widths [19]. Alternatively, the vessel landmarks can be directly extracted from the segmented vascular tree by using the adequate combination of shifted Gabor filter responses [15]. Nevertheless, this approach does not allow to distinguish between crossings and bifurcations.

A common drawback of the methods applied over the segmented vessels is that their performance critically depends on the accuracy of the previous vessel tree segmentation. In that sense, several works directly assume that an accurate vascular segmentation is available and evaluate the proposed landmark detection algorithms over manually labeled blood vessels [18, 15]. However, in practice, these ground truth segmentations are not commonly available, being the manual labeling unfeasible in clinical practice routine. An alternative that does not require an explicit segmentation of the vasculature is to use a vessel tracking algorithm guided by the intensity patterns of the retinal vessels [20]. Additionally, junction likelihood maps can be produced from the eye fundus images by using wavelets to compute orientation scores [12]. Abbasi et al. [12] combine this approach with a skeleton-based method to detect the vessel landmarks. These landmarks are later classified as crossings or bifurcations using the previously obtained dominant orientations.

The generation of junction likelihood maps has also been attempted by using DNNs [14]. However, the successful training of this task with common deep learning approaches is challenging. As reference, Uslu et al. [14] trained a multi-task network that predicts a rough estimation of the junction patterns. However, the extraction of the vessel landmarks from the network output still requires significant post-processing, similar to that applied in skeleton-based methods.

A different approach to solve the landmark detection with DNNs consists in training a patch-wise classifier [13]. Then, the predictions of overlapping patches are aggregated to obtain the final landmark estimations. Pratt et al. [13] combine this approach with a subsequent network to predict whether the patches that are identified as containing landmarks correspond to crossings or bifurcations. In this case, the vessel landmarks are both detected and classified. However, the method does not take advantage of the DNNs capacity to simultaneously perform both tasks, neither of their ability to integrate more representative information from larger contexts in comparison to the reduced analysis in small local patches.

In contrast with previous approaches, our proposal allows to successfully generate both the crossings and bifurcations likelihood maps from the raw eye fundus images. In that sense, the detection and distinction of the vascular landmarks is integrated into a single step. Besides the computational benefits, the use of a

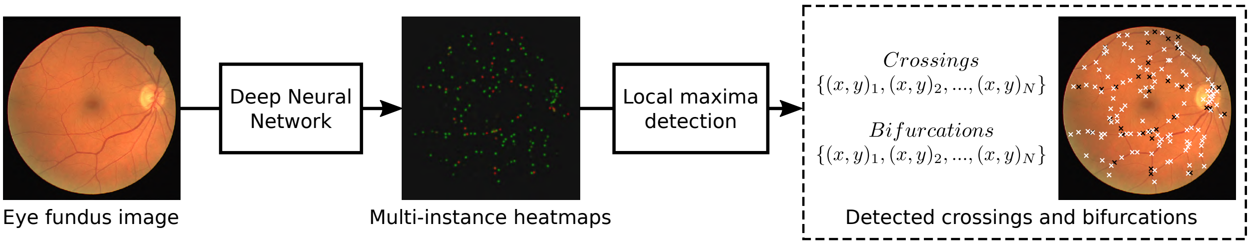


Figure 2: Methodology for the detection of vessel crossings and bifurcations in eye fundus images using the proposed multi-instance heatmap regression.

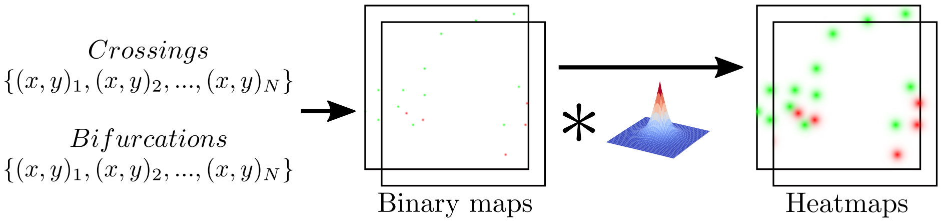

Crossings
$\{(x,y)_1,(x,y)_2,...,(x,y)_N\}$
Bifurcations
$\{(x,y)_1,(x,y)_2,...,(x,y)_N\}$
Binary maps
Heatmaps

Figure 3: Generation of the target heatmaps from the annotated pixel coordinates.

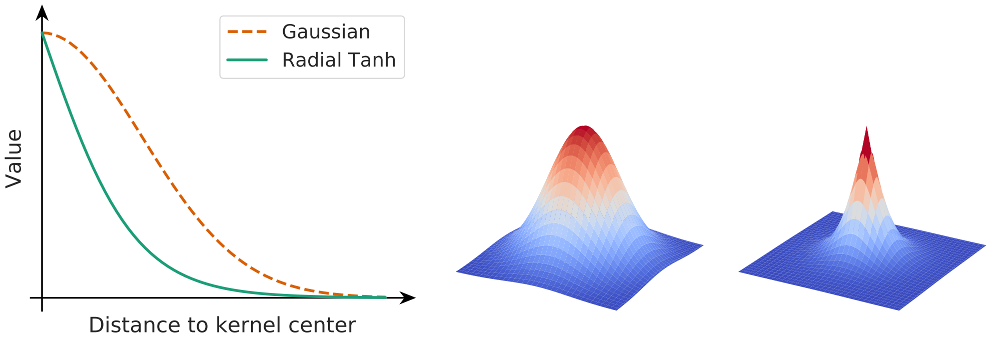
single network applied over large contexts significantly increases the feedback for learning the recognition of the vessel landmarks, which benefits the final performance. This is achieved by training a DNN in the prediction of multi-instance heatmaps that are automatically derived from the annotated pixel coordinates.

The use of a heatmap regression as surrogate task for the localization of landmarks has been previously explored in other domains. In particular, human pose estimation [21] and facial landmark detection [22] have been successfully approached by predicting landmark-derived heatmaps. Nevertheless, these tasks are typically performed over previously detected bounding boxes, which allows to only target the estimation of a known number of landmarks at a fixed scale. In contrast, the size of the blood vessels varies throughout the eye fundus whereas the number of vessel landmarks significantly varies among images. Thus, in our proposal, the networks learn to detect the required patterns at multiple scales and to generate output heatmaps containing multiple instances of the same target landmark type.

2. Materials and methods

2.1. Multi-instance heatmap regression

The detection of vessel crossings and bifurcations in eye fundus images requires the prediction of each landmark location as well as the distinction between the two possible types of landmarks: crossings and bifurcations. Moreover, the number of vascular landmarks present in the images is unknown. The straightforward alternative to tackle the detection of these landmarks using fully convolutional networks would imply the prediction of binary maps where only the pixels corresponding to the ground truth location of each landmark are labeled as positive class. However, those target binary maps are heavily unbalanced given that the number of landmark coordinates is much lower than the total number of pixels in the images. As a consequence, the labels provide limited feedback for training a DNN and over-penalize wrong but close predictions to the ground truth landmarks. An improved alternative is to transform the binary ground truth maps into heatmaps where the maximum values correspond to the labeled locations and progressively lower values are assigned to the surrounding pixels. The resulting heatmaps are defined as multi-instance heatmaps because they represent the location of multiple landmarks. The improved heuristic strategy provided by these heatmaps increases the information from the labels that is available to the network, improving the feedback for learning the detection task. Additionally, the heatmap approach takes into account the potential noise in the labels, transforming the hard binary labels into soft labels that better model the



(a) (b) (c)

Figure 4: (a) Comparison of the different kernel profiles. ((b),(c)) Kernels represented as a three-dimensional surface. (b) Gaussian. (c) Radial Tanh.

likelihood of a pixel being a target landmark location. For instance, in the considered task, the patterns that represent each crossing or bifurcation comprise several pixels and, therefore, the precise labeling of its center is error-prone, especially for thick vessels that cover a wide region (e.g., Figure 1(A)). In addition, many of the thin vessels present low contrast, which also makes difficult the labeling (e.g., Figure 1(B)). Hence, the use of soft labels may be beneficial in these frequent scenarios.

Figure 2 depicts a general overview of our methodology using the proposed multi-instance heatmap regression. Additionally, the generation of the ground truth heatmaps is summarized in the diagram of Figure 3. In particular, the annotated pixel coordinates are used to create the binary maps with the target locations labeled as the positive class. Then, the ground truth heatmaps are generated convolving the original binary maps with an isotropic kernel of convex and monotonic decreasing kernel profile. Given that there is no prior evidence of the most adequate specific kernel profile for the considered task, we explore the use of two different alternatives: a Gaussian kernel and a Radial Hyperbolic Tangent (Radial Tanh) kernel. The Gaussian kernel has been previously explored for the localization of landmarks in other application domains [21] whereas the Radial Tanh kernel is an alternative depicting a sharper profile, which may facilitate the detection task. Figure 4 depicts a visual comparison between both kernel types. The Gaussian (K_G) and Radial Tanh (K_{RT}) kernel are defined as:

$$K_G(x, y; \sigma) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

$$K_{RT}(x, y; \alpha) = 1 + tanh\left(-\frac{\pi\sqrt{x^2+y^2}}{\alpha}\right) \tag{2}$$

where (x, y) are the pixel coordinates with respect to the kernel center, σ is the standard deviation for the Gaussian kernel, and α is the saturation distance for the Radial Tanh kernel. Both the standard deviation of the Gaussian kernel and the saturation distance of the Radial Tanh kernel allow to control the region of influence for each landmark. In order to facilitate the comparison between both alternatives, we define an equivalent saturation distance for the Gaussian kernel. In particular, we empirically set this parameter to a value of 2.5 standard deviations, i.e., $\sigma = 0.4\alpha$.

Regarding the distinction between crossings and bifurcations, it is approached by the prediction of two independent heatmaps, one for each type of landmark. In this case, the neural network has to generate a two-channel output. Nevertheless, this setting strongly penalizes the misidentification of a crossing as bifurcation, or vice versa. For instance, using common regression metrics, the error when predicting a crossing in the bifurcation channel would be higher than the error when not predicting any landmark at all. Although this seems to be adequate for the final trained network, it complicates the learning process in the

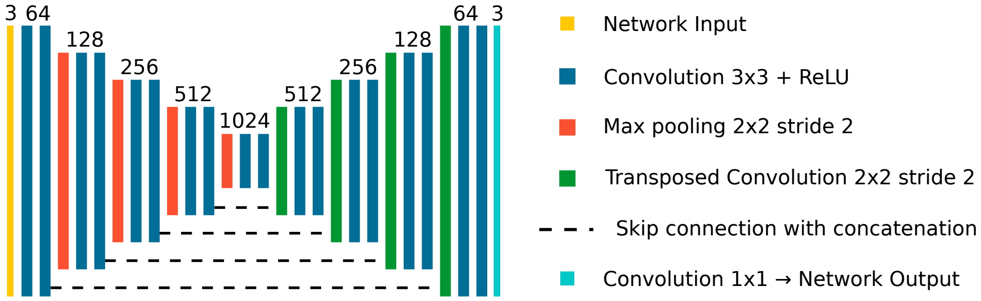

3
64
128
256
512
1024
512
256
128
64
3
Network Input
Convolution 3x3 + ReLU
Max pooling 2x2 stride 2
Transposed Convolution 2x2 stride 2
Skip connection with concatenation
Convolution 1x1 → Network Output

Figure 5: Diagram of the U-Net neural network depicting the number of output channels for each convolutional block.

early stages of the training. Thus, the neural network is trained to predict a third channel that includes both landmarks, which further encourages the detection of vessel landmarks regardless of their type.

The simultaneous regression of the three multi-instance heatmaps is trained using the mean squared error (MSE) between the predicted and the target heatmaps as loss. Thus, the training loss is defined as:

$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathbf{y}; \alpha) = ||\mathbf{f}(\mathbf{x}) - \mathbf{y} * \mathbf{K}(\alpha)||_2^2 \tag{3}$$

where $\mathbf{x}$ is an eye fundus image, $\mathbf{y}$ the corresponding target binary map, $\mathbf{f}$ the transformation given by a DNN that generates the predicted heatmaps, and $\mathbf{K} \in \{\mathbf{K}_G, \mathbf{K}_{RT}\}$ the convolutional kernel used to generate the target heatmaps.

The pixel coordinates of the target landmarks are recovered from the heatmaps by directly detecting the local maxima. In particular, we use a maximum filter and an intensity threshold to only retrieved the most salient local maxima. The threshold is required for the predicted heatmaps given the likely slight background noise that is produced by the network, preserving only the significative landmark detections. Additionally, this threshold allows to calibrate the proposed method to different operating points according to the requirements of each specific application. The half-size of the maximum filter must be, at most, lower than the minimum expected distance between landmarks of the same type. The minimum distance between different types of landmarks, i.e., between crossings and bifurcations in this case, does not affect because they are predicted in different output channels of the network.

2.2. Network architecture and training

In order to validate the proposed multi-instance heatmap regression for the identification of crossings and bifurcations, we use a standard network architecture and training procedure. In that sense, the experiments in this work are conducted using an U-Net network architecture [23]. This network represents a reliable baseline, being commonly used in many medical image analysis procedures. Particularly, U-Net has demonstrated to produce satisfactory results for related tasks performed on eye fundus images [24, 25]. Hence, it is expected to be also adequate for the detection of crossings and bifurcations in the same domain. A diagram of the network that is used in our experiments is depicted in Figure 5. In brief, U-Net is characterized by an encoder-decoder structure, including skip connections between the inner layers of the encoder and the decoder. These skip connections concatenate feature maps taken from the encoder with those of the same spatial resolution in the decoder. The main building blocks of the network consists of convolutional layers with 3 × 3 kernels and ReLU activation functions, following the idea of the VGG networks. We use a network of the same size as the original one proposed in [23]. However, our network presents a 3-channel input, required for the eye fundus images, and a 3-channel output, required for the three multi-instance heatmaps described in Section 2.1. Additionally, the output layer presents a linear activation function.

The network parameters are initialized with a zero-centered normal distribution following the method proposed by He et al. [26]. Then, the network is trained with full resolution images and batch size of one image. Additionally, we use a validation set composed of the 25% of the available training data. For the optimization, we use the Adam algorithm [27] with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which represent the default setting proposed by the authors. We also apply a learning rate schedule that reduces the learning rate by a factor of 10 when the validation loss plateaus. The initial learning rate is set to $\alpha = 1e-4$ and the patience for the learning rate schedule is 2500 batches. These hyperparameters are empirically established as those that, relying on the evolution of the validation loss, provide stable learning. The training is stopped after reaching a final learning rate of $\alpha = 1e - 7$, given that no significant changes are produced in the validation loss after that point. To avoid overfitting during training, we use spatial data augmentation consisting of random affine transformations applied to the input eye fundus images and the ground truth pixel coordinates of the target landmarks. Additionally, we also use color data augmentation consisting of random transformations of the image components in HSV color space, similar to the satisfactory application in the same domain of [28].

2.3. Datasets

The experiments in this work are performed using the publicly available DRIVE and IOSTAR datasets. In particular, the ground truth annotations for the identification of crossings and bifurcations in both datasets are provided by [12][1]. The DRIVE dataset [29] comprises 40 color fundus images that are divided by default into balanced training and test sets of 20 images each. The images present a field of view of 45° and a resolution of 565 × 584 pixels. In contrast, the IOSTAR dataset [12] is a collection of 24 scanning laser ophthalmoscope (SLO) images with a field of view of 45°. The images present varying resolutions but keep the same scale as the DRIVE dataset. SLO is a variant of eye fundus imaging that provides increased contrast with respect to traditional color fundus. In particular, the images of IOSTAR have been captured using green and infrared lasers.

The locations of the crossings and bifurcations have been annotated and reviewed by three different experts for both datasets [12]. In particular, the DRIVE dataset presents an average of 100 bifurcations and 30 crossings per image, whereas the IOSTAR dataset presents an average of 55 bifurcations and 23 crossings per image.

Following the common practices in previous works [12, 14], the DRIVE training set is used for training the networks, whereas the DRIVE test set and the IOSTAR dataset are held out for evaluation purposes.

2.4. Evaluation

The evaluation of the proposed approach is performed by comparing the detected crossings and bifurcations against the ground truth annotations. In that regard, an independent analysis is performed for each type of landmark (crossings or bifurcations). As gold standard, a detected landmark is considered a True Positive (TP) when it is located within a specified distance d of a ground truth landmark and a False Positive (FP) otherwise. Each ground truth landmark can only be detected once, i.e., we establish a one-to-one correspondence between the set of predictions and the set of ground truth landmarks. In case of several landmarks within the range d of a prediction, the closest one is considered as its corresponding. The ground truth landmarks that remain undetected are considered False Negatives (FN). Then, TP, FP, and FN measures are used to compute Precision and Recall, which are defined as:

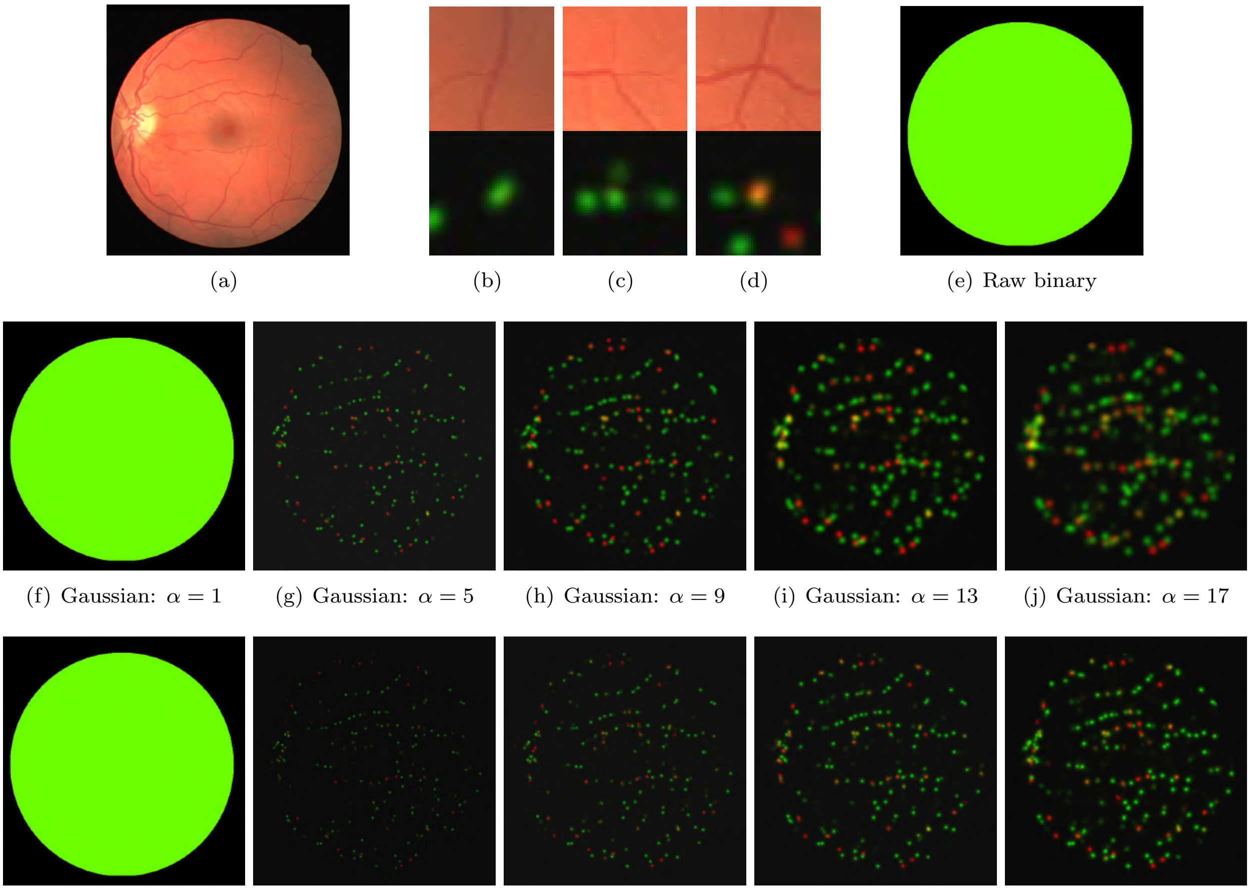
$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Additionally, we compute the F-score (F_1), which is the harmonic mean of Precision and Recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

[1] www.retinacheck.org/datasets



(a) (b) (c) (d) (e) Raw binary

(f) Gaussian: $\alpha = 1$ (g) Gaussian: $\alpha = 5$ (h) Gaussian: $\alpha = 9$ (i) Gaussian: $\alpha = 13$ (j) Gaussian: $\alpha = 17$

(k) Radial Tanh: $\alpha = 1$ (l) Radial Tanh: $\alpha = 5$ (m) Radial Tanh: $\alpha = 9$ (n) Radial Tanh: $\alpha = 13$ (o) Radial Tanh: $\alpha = 17$

Figure 6: Examples of predicted heatmaps where crossings are represented in the red channel and bifurcations in the green channels. (a) Eye fundus image from the DRIVE test set. (b-d) Regions cropped from (a) that depict both the original image and the predicted heatmaps in detail. (e) Prediction for (a) using the raw binary targets for training. (f-j) Predicted heatmaps for (a) using the Gaussian kernel at varying scales. (k-o) Predicted heatmaps for (a) using the Radial Tanh kernel at varying scales.

The described analysis is performed using a distance of 5 pixels ($d = 5$) as criteria to consider the detected landmarks as valid, as defined in other works [14]. This represents an approximate real distance of $125\mu m$ and $140\mu m$ for the DRIVE and IOSTAR datasets, respectively [12].

Additionally, we also measure the localization error for the detected landmarks, which is especially relevant for applications such as registration, vascular change detection, or authentication. The localization error is computed as the average Euclidean distance between the detected ground truth landmarks and their corresponding predictions. The higher bound for this localization error is given by the maximum distance required to consider a detection as valid, which in this case is 5 pixels.

3. Results and discussion

Figure 6 depicts representative examples of predicted heatmaps for networks that were trained using Gaussian or Radial Tanh kernels. In particular, predicted heatmaps corresponding to varying kernel sizes are depicted for each kernel type. The different kernel sizes are specified by the saturation distance parameter defined in Section 2.1. In the examples, the crossings are represented in the red channel whereas the bifurcations are represented in the green channel. Each one of the blobs depicted in the images corresponds to an identified crossing or bifurcation, whose most likely location is given by the local maximum in the center of the blob region. It is observed that for some experiments the output of the network is nearly constant, which is due to the network failing to converge during training (see Figure 6 (e),(j)). This only happens for very small kernels, which make the task very similar to the prediction of the raw binary targets. As reference, Figure 6 (b) depicts a representative example of network output when the raw binary targets are used for training. In that case, the network also failed to converge.

Additionally, in contrast with the use of binary targets, the prediction of heatmaps offers more useful output feedback. Regarding the general appearance of the predicted heatmaps, most of the blobs present similar shape and intensity values, although some exceptions are observed. In this regard, there are elongated or low-intensity blobs that differ from the model that the network learns during training. Given that the network learns to generate a specific pattern only when a crossing or bifurcation is detected, the generation of an altered output may evidence a less confident prediction. Thus, an elongated blob may indicate uncertainty in the precise location of the detected landmark (e.g., Figure 6(b)), whereas the low-intensity blobs may indicate uncertainty regarding the presence of that landmark (e.g., Figure 6(c)). Additionally, the example of Figure 6(d) shows how the network successfully deals with overlapping crossings and bifurcations. In this case, the predicted crossing and bifurcation blobs partially overlap, which results in a yellowish tone in the output of this depicted example.

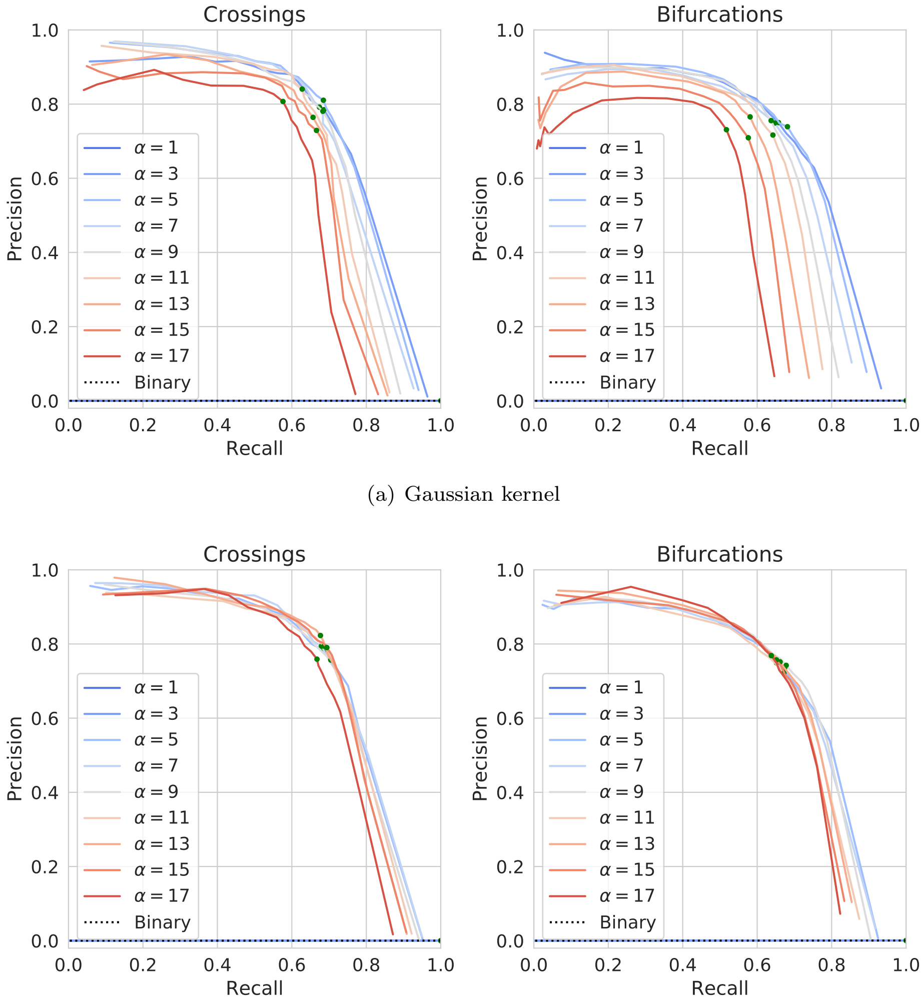
Regarding the comparison between Gaussian and Radial Tanh kernels, it is observed that, for the same kernel scale, the Gaussian kernel results in the generation of apparently larger and blurrier blobs. This effect is due to the more disperse distribution produced by the Gaussian kernel in comparison to the sharper one produced by the Radial Tanh variant, being the latter more concentrated around the specific identified landmark location.

In order to quantitatively evaluate the final objective of the proposed methodology, we perform the analysis described in Section 2.4. The local maxima are extracted from the predicted heatmaps as indicated in Section 2.1. In this case, we use a variable threshold, which allows to plot the Precision-Recall (PR) curves represented in Figure 7. The curves are depicted for both Gaussian and Radial Tanh kernels, as well as for different kernel scales. Additionally, the maximum F-score is computed for every experiment, which provides a representative operating point for subsequent comparisons. Simultaneously, we measure the average localization error for each experiment and threshold value in the PR curves. Figure 8 depicts these results by plotting the localization error against the recall measures. To facilitate the comparison with other results, the points of maximum F-score are also indicated. As reference, for both Figure 7 and Figure 8, we also include the results obtained when the raw binary targets are used for training.

As previously seen in the examples of Figure 6, the experiments with the smallest kernels do not converge and result in almost zero precision for any applied threshold. This matches with the constant output depicted in Figures 6(e),(j). Also, for those experiments, the localization error is 5 pixels, which is the maximum for the performed evaluation. Additionally, the same results are obtained when using the raw binary maps. However, once the kernel size is increased to the minimum required for convergence, the performance of the multi-instance heatmaps improves drastically. In this regard, it should be noticed that the smallest kernels will produce little change in the original binary maps, providing limited feedback for training the networks. However, slightly increasing the kernel size, the region of influence for each landmark is also increased. This results in an improved heuristic for learning the detection task.

In the case of the Gaussian variant, the best performance is obtained for the smallest kernels (after removing the non-convergence case) and it is gradually reduced with the increase of the kernel size. This happens in terms of both PR analysis and localization error. In contrast, for the Radial Tanh variant, a similar performance is obtained for the different kernel sizes. The only exception is the largest kernel when evaluating the detection of crossings. Nevertheless, if the analysis is reduced to the high recall region, the smaller kernels are able to produce higher recall values. This trend is similar to that of the Gaussian kernels, albeit on a smaller scale.

Figure 9 depicts representative examples of detected crossings and bifurcations over an analyzed eye fundus image from the DRIVE test set. The detected landmarks are represented with crosses whereas the ground truth landmarks are represented with circles. At the same time, the black color denotes crossings and the white denotes bifurcations. The provided examples correspond to the operating points with the highest



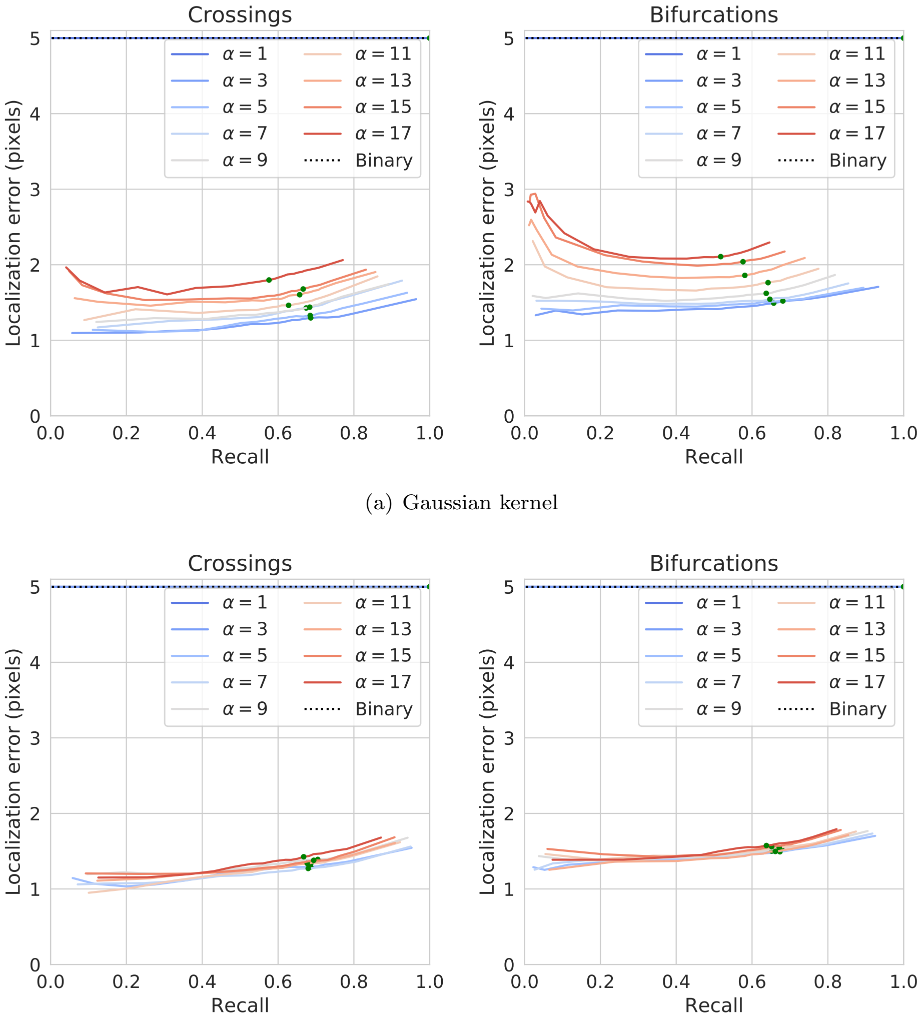
(a) Gaussian kernel

(b) Radial Tanh kernel

Figure 7: Precision-Recall curves for the detection of crossings and bifurcations in the DRIVE test set at varying kernel scales. The green dots represent the operating points of maximum F-score.

F-score, which are marked in the plots of Figures 7 and 8. These examples show that the method detects the majority of the landmarks, while simultaneously it distinguishes between crossings and bifurcations. Regarding the missing landmarks and false detections, most of them correspond to secondary tiny vessels (as reference, see Figure 9(c)). In these cases, the crossings and bifurcations are very difficult to appreciate and, therefore, their analysis is typically not considered in the clinical practice. Moreover, the small size and low contrast of these tiny vessels also makes the labeling more error-prone, which complicates both the training and evaluation. Discarding these extreme scenarios, in general, the method offers an adequate performance for both main and secondary branches of the vascular tree. Additionally, the examples show that the results obtained with the two different kernels are similar, at least when an adequate kernel scale is selected. In particular, many of the missing landmarks and false detections are the same for both variants.

In summary, the obtained results demonstrate that the multi-instance heatmap regression approach is adequate for the detection of crossings and bifurcations in eye fundus images. In the performed experiments, the use of very small kernels led to the networks failing to converge during training. However, as said before, the smallest kernels in our experiments are almost equivalent to not using any kernel at all and, instead,



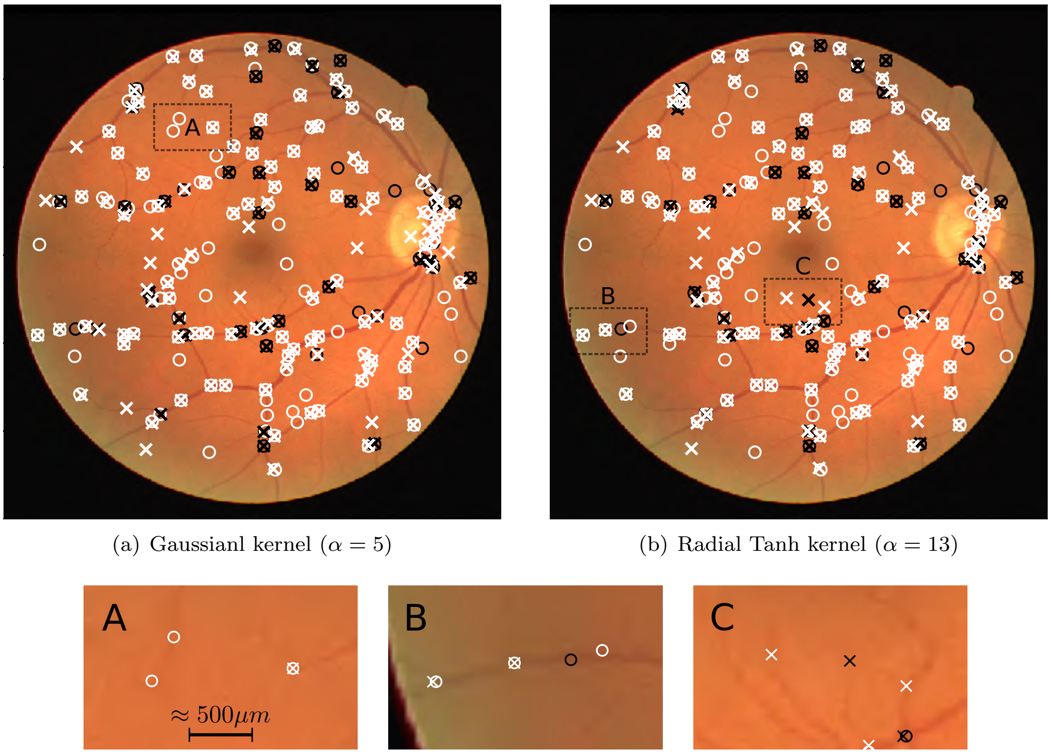
(a) Gaussian kernel

(b) Radial Tanh kernel

Figure 8: Localization error (in pixels) against Recall for the detection of crossings and bifurcations in the DRIVE test set at varying kernel scales. The green dots represent the operating points of maximum F-score.

directly training the prediction of the binary target maps. In fact, the same outcome was obtained when directly using the raw binary maps for training. This means that it is precisely the proposed approach which makes possible the detection of vessel crossings and bifurcations using fully convolutional networks.

Regarding the comparison between both types of kernels, the main difference is the higher dependency of the Gaussian variant with respect to the kernel size. In that sense, even though the proposed approach requires the selection of an adequate kernel scale, the Radial Tanh variant demonstrated a robust and stable performance for a significant range of kernel sizes. In contrast, even if the same or superior performance can be achieved using the Gaussian kernel, in practice its use requires more tuning of the kernel scale. In that regard, the advantage of the Radial Tanh kernel is due to the sharper profile. This kernel produces well-defined maxima even when the kernel size is significantly increased. At the same time, it still facilitates the training of the detection task. Finally, a trend that is observed for both kernels in the high recall region of the PR curves (Figure 7) is the reduction in recall with the increase of the kernel size. This may be explained by a less defined maxima when the generated blobs get larger as well as the possible overlap of very close landmarks of the same type, which makes it extremely complicated to differentiate each one of



(a) Gaussianl kernel ($\alpha = 5$)

(b) Radial Tanh kernel ($\alpha = 13$)

(c) Cropped regions in detail

Figure 9: Examples of detected crossings (in black) and bifurcations (in white) over an eye fundus image from the DRIVE test set. The circles denote ground truth annotations whereas the crosses denote detected landmarks. (a-b) Complete eye fundus images. (c) Cropped regions from (a) and (b) depicting representative examples of missing landmarks and false detections.

them. Nevertheless, this happens to a lesser extent for the Radial Tanh kernel, given the mentioned genuine sharper profile.

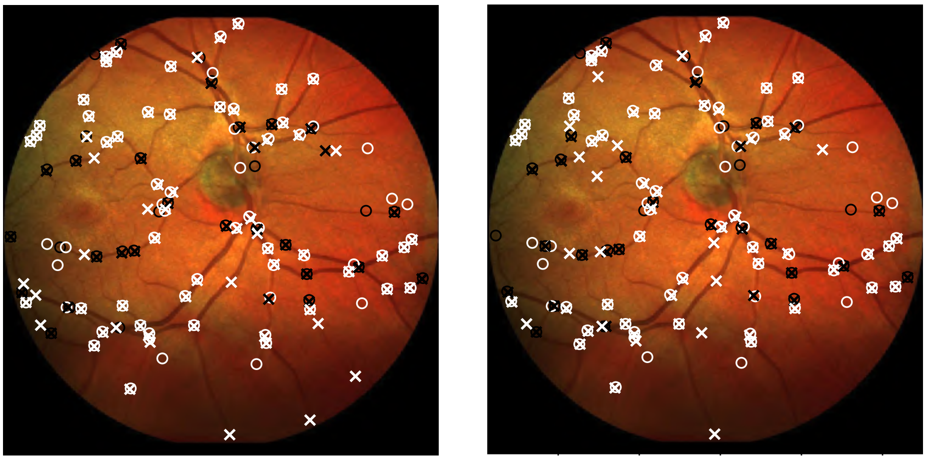
3.1. Comparison with the state-of-the-art

In this section, we compare the performance of the proposed approach against those state-of-the-art works that were evaluated on the same public datasets. To that end, we select the kernel sizes that provide the best performance by means of maximum F-score on the DRIVE training set. Then, the comparison is performed for both the DRIVE test set and the IOSTAR dataset. As reference, Figure 10 depicts examples of detected crossings and bifurcations for the IOSTAR dataset.

In contrast with the proposed approach, previous works typically address the detection of junctions followed by their classification between crossings and bifurcations. This is reflected in their evaluation, which is independently performed for these two steps (detection and classification). To provide an adequate comparison, we reevaluate the trained networks as junctions detectors by merging the predicted sets of crossings and bifurcations. Additionally, the performance as binary classifiers is evaluated over the set of correctly detected junctions. In this case, the crossings are considered as positive samples and the bifurcations as negative ones [12, 13].

Figure 11 depicts the comparison for the detection of junctions. It is observed that the proposed method significantly outperforms previous approaches in both the DRIVE and IOSTAR dataset. Furthermore, the improvement is independent of the selected operating point, given that the performance of the other approaches is always under the PR curves of the proposed method.

In the literature, there are some additional works that reported competitive performance regarding the detection of junctions. However, it should be considered that, in some cases, the evaluation datasets



(a) Gaussian kernel ($\alpha = 5$)

(b) Radial Tanh kernel ($\alpha = 13$)

Figure 10: Examples of detected crossings (in black) and bifurcations (in white) over an eye fundus image from the IOSTAR dataset. The circles denote ground truth annotations whereas the crosses denote detected landmarks.

present significantly less detailed annotations [16], whereas, in others, the methods are applied over manually segmented vessels [18]. In that regard, Uslu et al. [14] evaluate their method on both eye fundus images and manually labeled vessels. However, in order to produce an even comparison among all the methods, we do not include the results corresponding to the manual segmentations. Additionally, regarding the provided comparisons, Pratt et al. [13] report individual results for the annotations of three different experts. In this case, we only include the results with the highest accuracy, which is provided by the first expert in their work.

Table 1 depicts the results and comparison for the binary classification between crossings and bifurcations. Given that the classification is evaluated over the correctly detected junctions, the results can vary depending on the operating point for the detection of junctions. Thus, we report classification results for several recall levels in the detection of junctions. These results show that the proposed method outperforms previous approaches at the same level of detection sensitivity. Additionally, our approach also keeps an adequate performance when the detection sensitivity is increased, i.e., when more landmarks are detected.

In summary, the proposed approach leads to a remarkable improvement over previous existing methods. In that sense, although the use of DNNs had been previously explored, existent works did not achieve a significant improvement over other methodologies. This evidences that the advantage of the presented method is not merely due to the use of DNNs but, instead, to the proposed multi-instance heatmap regression. In particular, this novel approach allows to detect vessel crossings and bifurcations using the whole eye fundus images as input to the network, which increases the contextual information that is available for each landmark. Additionally, the use of heatmaps, instead of binary labels, provides more feedback for training the network, as well as an improved heuristic strategy. Furthermore, the proposed approach is more direct and efficient than other alternatives, given that a single neural network is able to detect and distinguish the vessel crossings and bifurcations.

Finally, the results provided in this section show that the performance for the detection of junctions on the IOSTAR dataset is not as good as that on the DRIVE test set. In this case, it should be considered that these datasets correspond to two slightly different image modalities, namely color fundus and SLO. Moreover, following the approach of previous works [14, 13], we reserve the whole IOSTAR dataset for evaluation due to its small size. Hence, there is a certain domain shift between training and test in the case of the evaluation on IOSTAR.

Additionally, it should be noticed that the validation of existent methods for the detection of vessel crossings and bifurcations is conditioned by the currently available datasets. In that sense, both DRIVE

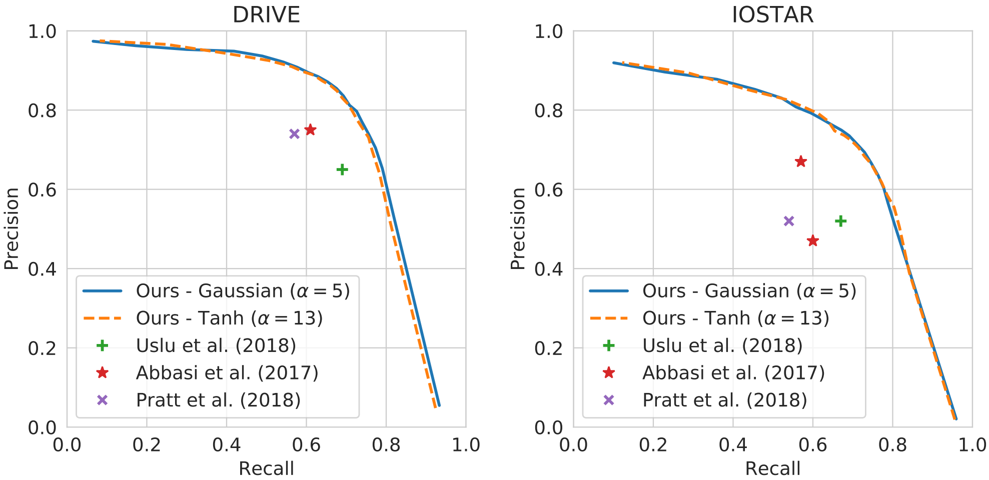


Figure 11: Precision-Recall curves for the detection of vessel junctions without considering their distinction between crossings and bifurcations. Comparison of state-of-the-art works and the proposed approach.

and IOSTAR lack annotations regarding the presence of pathological lesions. Therefore, as future work, we consider the elaboration of a more complete database, including expert annotations of vascular landmarks and lesions for the same images. This would allow to study the performance of the algorithms in the presence of different pathological conditions.

4. Conclusions

The automated detection of vessel crossings and bifurcations in eye fundus images represents an important task with numerous practical applications. In that sense, despite the direct analysis for clinical purposes, the detection of these representative landmarks is commonly required as an intermediate step for several automated procedures. In this work, we propose a novel methodology that addresses the simultaneous detection of crossings and bifurcations in eye fundus images. In particular, we reformulate the detection task as a multi-instance heatmap regression, which is performed using a deep neural network. This novel approach allows to make predictions using full images and integrates into a single step the detection and distinction of the vascular landmarks.

Several experiments are conducted to analyze the proposed approach, including the study of different alternatives to construct the multi-instance heatmaps for training the neural networks. In order to validate the proposal, we use two public datasets of reference with detailed annotations of vessel crossings and bifurcations. The proposed method achieves 74.23% and 70.90% F-score for the detection of crossings and bifurcations, respectively, in color fundus images. These results represent a significant improvement over previous existent methods. Moreover, in the presented approach, the detection and distinction of the vessel crossings and bifurcations is integrated into a single step, being not only more effective but also more efficient than other alternatives.

Acknowledgments

This work is 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 by Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the DPI2015-69948-R and RTI2018-095894-B-I00 research projects. The authors of this work also receive financial support from the ERDF and European Social Fund (ESF) of the EU, and Xunta de Galicia through Centro Singular de

Table 1: Performance for the binary classification between crossings (positive samples) and bifurcations (negative samples). Comparison of state-of-the-art works and the proposed approach. Acc, Sp, and Sn denote accuracy, specificity, and sensitivity, respectively.

Method	Acc (%)	Sp (%)	Sn (%)	Support set
Evaluation on DRIVE				
Abbasi et al. (2016) [12]	83.00	91.00	59.00	Detected with 61.00% recall
Pratt et al. (2018) [13]	80.27	84.82	69.89	All*
Ours – Gaussian	93.56	96.91	85.88	Detected with 60.90% recall
	93.83	97.09	86.17	Detected with 71.01% recall
	95.93	97.42	92.27	Detected with 82.61% recall
Ours – Radial Tanh	94.39	97.22	87.53	Detected with 59.24% recall
	93.82	96.62	87.05	Detected with 70.94% recall
	94.93	96.60	90.76	Detected with 81.55% recall
Evaluation on IOSTAR				
Abbasi et al. (2016) [12]	83.00	93.00	67.00	Detected with 57.00% recall
Pratt et al. (2018) [13]	64.79	61.27	74.35	All*
Ours – Gaussian	94.22	95.87	90.37	Detected with 59.14% recall
	92.83	95.22	87.36	Detected with 70.76% recall
	95.59	97.70	90.57	Detected with 80.19% recall
Ours – Radial Tanh	93.93	96.17	88.60	Detected with 61.20% recall
	92.72	94.60	88.24	Detected with 71.23% recall
	95.29	96.53	92.29	Detected with 81.43% recall

* The classifier was evaluated on the whole set of ground truth annotations.

Investigación de Galicia, accreditation 2016-2019, ref. ED431G/01, Grupo de Referencia Competitiva, ref. ED431C 2016-047, and the predoctoral grant contract ref. ED481A-2017/328.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] N. Patton, T. M. Aslam, T. MacGillivray, I. J. Deary, B. Dhillon, R. H. Eikelboom, K. Yogesan, I. J. Constable, Retinal image analysis: Concepts, applications and potential, Progress in Retinal and Eye Research 25 (2006) 99 – 127.

[2] M. D. Abramoff, M. K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Reviews in Biomedical Engineering 3 (2010) 169–208.

[3] T. Y. Wong, R. Klein, B. E. Klein, J. M. Tielsch, L. Hubbard, F. Nieto, Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality, Survey of Ophthalmology 46 (2001) 59 – 80.

[4] S. Akbar, M. U. Akram, M. Sharif, A. Tariq, U. ullah Yasin, Arteriovenous ratio and papilledema based hybrid decision support system for detection and grading of hypertensive retinopathy, Computer Methods and Programs in Biomedicine 154 (2018) 123 – 141.

[5] E. Grisan, M. Foracchia, A. Ruggeri, A novel method for the automatic grading of retinal vessel tortuosity, IEEE Transactions on Medical Imaging 27 (2008) 310–319.

[6] S. Kalaie, A. Gooya, Vascular tree tracking and bifurcation points detection in retinal images using a hierarchical probabilistic model, Computer Methods and Programs in Biomedicine 151 (2017) 139 – 149.

[7] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), 2018.

[8] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, M. J. Carreira, Retinal verification using a feature points-based biometric pattern, EURASIP Journal on Advances in Signal Processing 2009 (2009) 235746.

[9] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, L. V. Gool, Deep Retinal Image Understanding, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016.

[10] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, A. Hunter, Microaneurysm detection using fully convolutional neural networks, Computer Methods and Programs in Biomedicine 158 (2018) 185 – 192.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60 – 88.

[12] S. Abbasi-Sureshjani, I. Smit-Ockeloen, E. Bekkers, B. Dashtbozorg, B. t. H. Romeny, Automatic detection of vascular bifurcations and crossings in retinal images using orientation scores, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 189–192.

[13] H. Pratt, B. M. Williams, J. Y. Ku, C. Vas, E. McCann, B. Al-Bander, Y. Zhao, F. Coenen, Y. Zheng, Automatic detection and distinction of retinal vessel bifurcations and crossings in colour fundus photography, Journal of Imaging 4 (2018).

[14] F. Uslu, A. A. Bharath, A Multi-task Network to Detect Junctions in Retinal Vasculature, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 92–100.

[15] G. Azzopardi, N. Petkov, Automatic detection of vascular bifurcations in segmented retinal images using trainable COSFIRE filters, Pattern Recognition Letters 34 (2013) 922–933.

[16] D. Calvo, M. Ortega, M. G. Penedo, J. Rouco, Automatic detection and characterisation of retinal vessel tree bifurcations and crossovers in eye fundus images, Computer Methods and Programs in Biomedicine 103 (2011) 28 – 38.

[17] A. M. Aibinu, M. I. Iqbal, A. A. Shafie, M. J. E. Salami, M. Nilsson, Vascular intersection detection in retina fundus images using a new hybrid approach, Computers in Biology and Medicine 40 (2010) 81–89.

[18] A. Fathi, A. R. Naghsh-Nilchi, F. A. Mohammadi, Automatic vessel network features quantification using local vessel pattern operator, Computers in Biology and Medicine 43 (2013) 587–593.

[19] H. Hamad, D. Tegolo, C. Valenti, Automatic detection and classification of retinal vascular landmarks, Image Analysis & Stereology 33 (2014) 189–200.

[20] Chia-Ling Tsai, C. V. Stewart, H. L. Tanenbaum, B. Roysam, Model-based method for improving the accuracy and repeatability of estimating vascular bifurcations and crossovers from retinal fundus images, IEEE Transactions on Information Technology in Biomedicine 8 (2004) 122–130.

[21] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: International Conference on Computer Vision, 2015.

[22] D. Merget, M. Rock, G. Rigoll, Robust facial landmark detection via a fully-convolutional local-global context network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[23] O. Ronneberger, P.Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[24] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal image understanding emerges from self-supervised multimodal reconstruction, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018.

[25] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised deep learning for retinal vessel segmentation using automatically generated labels from multimodal data, in: International Joint Conference on Neural Networks (IJCNN), 2019.

[26] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: International Conference on Computer Vision (ICCV), 2015.

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.

[28] P. Liskowski, K. Krawiec, Segmenting Retinal Blood Vessels with Deep Neural Networks, IEEE Transactions on Medical Imaging 35 (2016) 2369–2380.

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

Chapter 4

Retinal Computer-Aided Diagnosis - Published Papers

4.1 Journal Paper: Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis

Journal Paper: Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis

Álvaro S. Hervella[1,2], José Rouco[1,2], Jorge Novo[1,2], and Marcos Ortega[1,2]
{a.suarezh, jrouco, jnovo, mortega}@udc.es

[1] Centro de Investigación CITIC, Universidade da Coruña, Coruña, A Coruña (Spain)

[2] VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña (Spain)

Journal: Expert Systems with Applications
Volume: 185
Pages: 115598
Editorial: Elsevier
ISSN: 0957-4174
Year: 2021

This is a post-peer-review, pre-copyedit version of an article published in the journal Expert Systems with Applications. The final authenticated version is available online at: https://doi.org/10.1016/j.eswa.2021.115598.

Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Marcos Ortega[a,b]

[a] *Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain*
[b] *VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña, Spain*

Abstract

Computer-aided diagnosis using retinal fundus images is crucial for the early detection of many ocular and systemic diseases. Nowadays, deep learning-based approaches are commonly used for this purpose. However, training deep neural networks usually requires a large amount of annotated data, which is not always available. In practice, this issue is commonly mitigated with different techniques, such as data augmentation or transfer learning. Nevertheless, the latter is typically faced using networks that were pre-trained on additional annotated data.

An emerging alternative to the traditional transfer learning source tasks is the use of self-supervised tasks that do not require manually annotated data for training. In that regard, we propose a novel self-supervised visual learning strategy for improving the retinal computer-aided diagnosis systems using unlabeled multimodal data. In particular, we explore the use of a multimodal reconstruction task between complementary retinal imaging modalities. This allows to take advantage of existent unlabeled multimodal data in the medical domain, improving the diagnosis of different ocular diseases with additional domain-specific knowledge that does not rely on manual annotation.

To validate and analyze the proposed approach, we performed several experiments aiming at the diagnosis of different diseases, including two of the most prevalent impairing ocular disorders: glaucoma and age-related macular degeneration. Additionally, the advantages of the proposed approach are clearly demonstrated in the comparisons that we perform against both the common fully-supervised approaches in the literature as well as current self-supervised alternatives for retinal computer-aided diagnosis. In general, the results show a satisfactory performance of our proposal, which improves existing alternatives by leveraging the unlabeled multimodal visual data that is commonly available in the medical field.

Keywords: deep learning, medical imaging, self-supervised learning, eye fundus, transfer learning, computer-aided diagnosis

1. Introduction

Deep learning has become a fundamental part of modern computer-aided diagnosis (CAD) systems. The use of deep neural networks (DNNs) has improved the performance over traditional methods without requiring the ad-hoc design of complex processing algorithms. However, in return, DNNs need to be fed with expert knowledge in the form of large annotated datasets (Litjens et al., 2017; Jing & Tian, 2020).

Gathering enough annotated data for training a DNN can be challenging. In fact, in medical imaging, the manual labeling of the images should be performed by expert clinicians. This requirement commonly

[*] Corresponding author. Email: a.suarezh@udc.es; Tel.: +34-981-16-70-00 (ext. 5522)

Email addresses: a.suarezh@udc.es (Álvaro S. Hervella), jrouco@udc.es (José Rouco), jnovo@udc.es (Jorge Novo), mortega@udc.es (Marcos Ortega)

leads to a limited number of available annotated samples, given the time that takes to produce high-quality annotations (Tajbakhsh et al., 2016). This issue has motivated the adoption of numerous techniques aiming at the improvement of the deep learning methods without the necessity of additional annotated data (Cheplygina et al., 2019). For instance, data augmentation techniques, which consist in creating new data samples through a set of plausible transformations for the application domain, are applied by default in order to successfully train DNNs (Litjens et al., 2017; Bloice et al., 2019). Additionally, transfer learning techniques, which consist in taking advantage of already trained models for other applications, are also commonly employed to further improve the performance of the networks (Litjens et al., 2017; Cheplygina et al., 2019).

A common approach to transfer learning in image analysis is the use of DNNs that were pre-trained on extensive annotated datasets (Cheplygina et al., 2019; Houssein et al., 2020). However, the available datasets of this kind are typically focused on broad domain applications, such as the natural image classification challenge in the ImageNet dataset (Deng et al., 2009). It can be argued that the different nature of these images with respect to, for example, medical images, can represent a limiting factor for transfer learning purposes. In fact, when large scale annotated datasets of medical images are available, the performance benefit due to ImageNet pre-training is very limited (Raghu et al., 2019). However, in practice, the medical image datasets typically present a reduced number of annotations. In these scenarios, ImageNet classification pre-training has demonstrated to provide a general knowledge that improve the training of very deep convolutional networks in the medical imaging field (Cheplygina et al., 2019; Houssein et al., 2020).

Another alternative is to exploit the availability of heterogeneous or complementary labels within the same application domain, e.g. segmentation and classification labels. This allows to produce supervised auxiliary tasks that are restricted to the target application domain (Cheplygina et al., 2019). In this case, applying transfer or multi-task learning, the target task benefits from the increased amount of domain-specific knowledge. This auxiliary task alternative within the same domain has demonstrated to be superior than the use of additional data from broad domain natural images, even when the total number of involved annotations is lower (Wong et al., 2018). The inconvenience in this case is that the additional labels in the target application domain carry an additional annotation effort.

Recently, self-supervised learning has arisen as a promising alternative to the traditional supervised approaches for transfer learning (Jing & Tian, 2020). Self-supervised learning is based on the use of pretext tasks that are trained with conventional supervised methods but do not require manual annotations. Instead, the training labels for theses tasks are automatically generated from the unlabeled data. This allows the learning of useful representations for a target task using unlabeled data from the same application domain. Nowadays, the most common approaches to self-supervised learning are focused on either the prediction of hidden portions of the data or the prediction of hidden relations in the data (Jing & Tian, 2020). For instance, a representative self-supervised pretext task is colorization (Zhang et al., 2016), which consists in the prediction of the different color components from the grayscale input image. Similarly, it is also possible to create an image inpainting task that requires the prediction of the original content of masked regions in the input image (Pathak et al., 2016). Alternatively, solving jigsaw puzzles of the input image (Noroozi & Favaro, 2016) or predicting the geometric relationship between automatically extracted object proposals (Oh et al., 2019) are representative examples of predicting the relations in the data. In this line, an emerging trend is the use of instance discrimination tasks (Ye et al., 2019; Chen et al., 2020) that are performed via contrastive learning (Hadsell et al., 2006). In these cases, the network learns to discriminate the individual images after being substantially altered via common data augmentations pipelines.

In medical imaging, given the difficulty for gathering large annotated datasets, there is an increasing interest for exploring these approaches. In particular, several works have adapted or extended existing paradigms previously proposed for natural images. For instance, Ross et al. (2018) propose colorization as auxiliary task for improving the segmentation of endoscopic video data. More recently, Taleb et al. (2020) extended several state-of-the-art self-supervised approaches to 3D medical data, including, e.g., jigsaw puzzles or instance discrimination with contrastive learning. Similarly, the instance discrimination paradigm was extended by including the synthesis of a complementary image modality as an additional transformation of the input image (Li et al., 2020). Additionally, other novel self-supervised paradigms have been directly proposed in the medical imaging field. For instance, Chen et al. (2019) propose a context restoration task

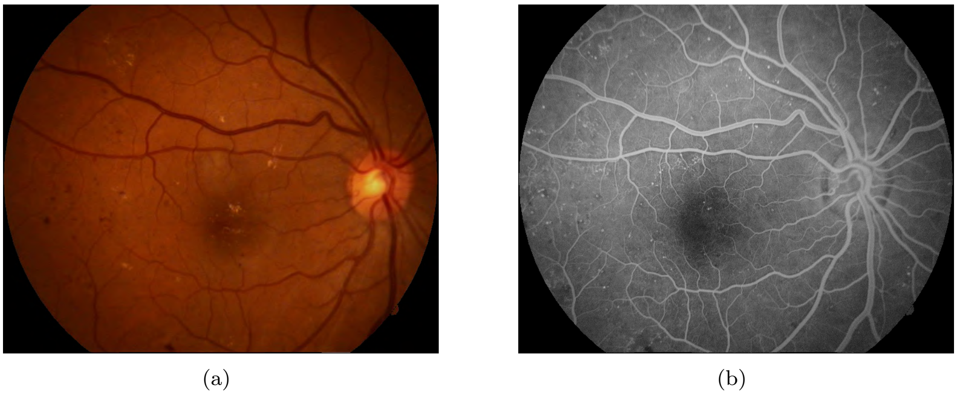


Figure 1: Representative example of (a) retinography and (b) fluorescein angiography for the same eye.

that requires to predict the original content of an image where different random patches are swapped. In contrast, Chaitanya et al. (2020) extended the contrastive learning paradigm to local features, producing a more adequate auxiliary task for segmentation.

Alternatively, instead of building the pretext task by manipulating the input image, in medical imaging it is also possible to directly use multimodal visual data for self-supervised learning purposes. In particular, Hervella et al. (2020a) propose the multimodal reconstruction between complementary image modalities as auxiliary task for segmentation or localization. The use of different imaging techniques is common in the modern clinical practice, including the use of complementary image modalities that represent the same organs or tissues. These complementary image modalities can be used to create a self-supervised multimodal reconstruction task consisting in the prediction of one image modality from other Hervella et al. (2020b). In order to solve this complex task, a neural network will have to learn relevant domain-specific patterns from the unlabeled data. Hence, the internal representations learned during this self-supervised task should be useful to improve the training of other target tasks in the same application domain.

However, self-supervision based on multimodal reconstruction of medical images has not yet been explored for improving deep learning CAD systems. In this regard, although Li et al. (2020) aim at using complementary modalities to aid self-supervision, their setting is not based on direct prediction. Instead, they explore the use of synthetic complementary image modalities as an additional augmentation strategy in a contrastive learning instance discrimination setting. In that case, it is expected that the learned representations are invariant to the synthetic multimodal transformation. However, that approach does not provide any incentive for the network to detect all the important patterns involved in the complex causal relations between modalities. In this sense, the network could just represent the patterns that are evident and similar in both modalities, disregarding the particular image contents that evidence the different complementary visualizations of the same reality. Additionally, due to the use of a synthetic image modality, the network only has access to a rough estimate of the true multimodal data. In contrast, the multimodal reconstruction task directly provides the network with the true multimodal data and the network must precisely learn the complex relationship between modalities to solve the task. Thus, the self-supervised multimodal reconstruction provides the network with a deep understanding of the image contents, which is expected to further facilitate the training of the desired deep learning CAD systems.

In this work, we propose to use the multimodal reconstruction between complementary retinal image modalities as self-supervised pre-training for deep learning-based retinal CAD systems. Specifically, we use the multimodal reconstruction between retinography and fluorescein angiography (Hervella et al., 2020b). Figure 1 depicts a representative example of retinography and fluorescein angiography for the same eye. These two imaging modalities are obtained with different capture processes and represent complementary information about the different anatomical structures and pathological lesions in the retina. In particular, the retinography is directly obtained as a color photograph of the retina, whereas the angiography is captured

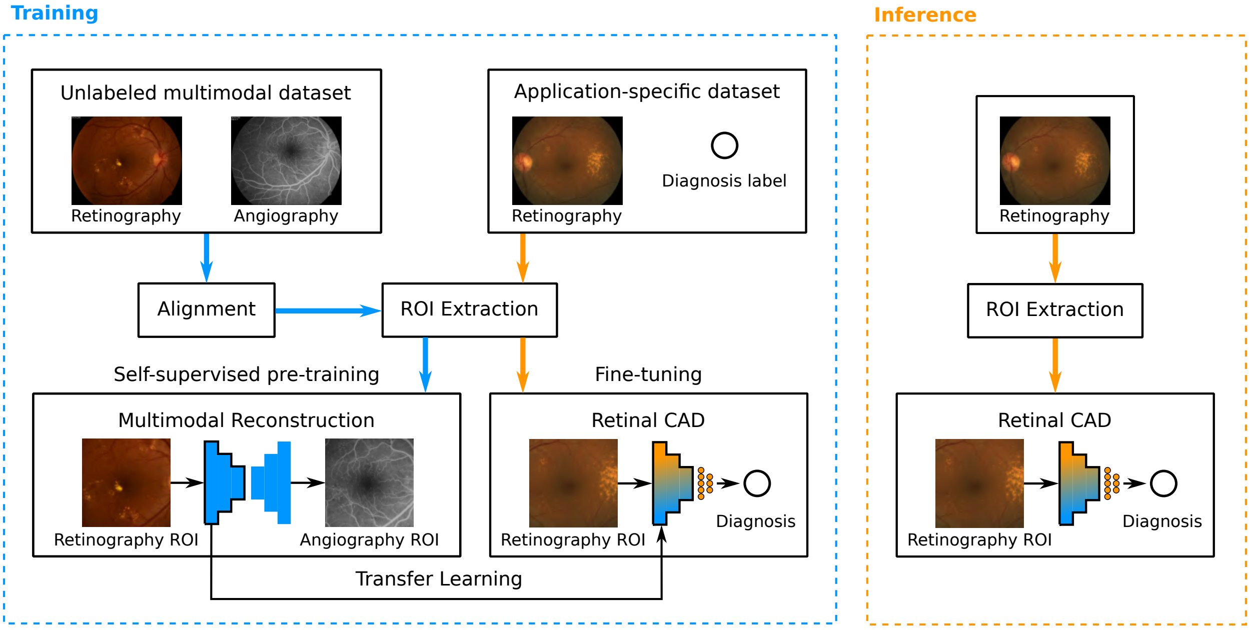


Figure 2: Scheme of deep learning-based retinal CAD system using the proposed multimodal self-supervised pre-training.

after injecting a contrast dye into the patient's bloodstream. The proposed approach exploits this kind of existent unlabeled multimodal image pairs for learning useful representations of the data. This idea has been previously explored for improving pixel-wise prediction tasks, such as segmentation and localization, where the same neural network can be used for pre-training and target tasks (Hervella et al., 2020a). However, CAD systems require a completely different network architecture, which prevents the direct adoption of existing methodologies. In this regard, we provide a complete methodology for taking advantage of the multimodal reconstruction and improve the training of deep learning-based retinal CAD systems. In particular, this work focuses on the diagnosis of different retinal diseases, including two of the most prevalent impairing ocular disorders: glaucoma and age-related macular degeneration (AMD). In this context, we perform several experiments that allow to better understand the proposed approach and we perform a comparison against two common fully-supervised approaches: training the network from scratch in the target task and pre-training in the annotated ImageNet dataset. Additionally, we also provide a comparison against previous self-supervised approaches in retinal image analysis.

2. Methodology

The proposed multimodal self-supervised transfer learning paradigm for the training of retinal CAD systems is summarized in the scheme of Figure 2. The objective of the retinal CAD system is to predict the clinical diagnosis for a certain disease using the retinography of the patient as single input data. This target classification task is trained using an application-specific dataset containing annotated retinographies. In order to improve the performance of the target task and reduce the necessity of a large annotated dataset, we propose a domain-specific pre-training using unlabeled images. In particular, the pre-training task consists in the generation of fluorescein angiography from retinography. This multimodal reconstruction of the eye fundus is a self-supervised task that does not require manually annotated data for the training. Instead, it takes advantage of existent unlabeled multimodal image pairs. In order to take advantage of these image pairs, the retinography and the angiography of the same eye are aligned together. This establishes a pixel-wise correspondence between both images, resulting in a richer source of information in comparison with the unaligned counterparts.

Regarding the retinal CAD system, the study of different diseases typically requires the analysis of different regions in the retinal images. Thus, the Region Of Interest (ROI) for each disease is automatically extracted from the input retinography before feeding the image to the neural network. Simultaneously, the ROI for each disease is also extracted in the images of the unlabeled multimodal dataset. Thus, during the pre-training phase, the neural network will have to learn retinal patterns similar to those required for the target application. The proposed transfer learning paradigm is applied by fine-tuning, in the target classification task, the previously trained multimodal reconstruction network. In particular, a fully convolutional encoder-decoder network is used for the multimodal reconstruction. Then, the encoder part of the network is reused for the target classification task. In this regard, given the different network architecture requirements of both tasks, we explore different alternatives for performing an effective transfer learning between multimodal reconstruction and classification.

2.1. Deep learning-based retinal CAD

The automated diagnosis of AMD and glaucoma from the retinography is approached as a binary classification task. Therefore, for each disease, a neural network is trained to predict whether an input retinography is healthy or pathological. In this regard, in clinical practice, AMD is typically diagnosed by the presence of certain pathological structures or lesions around the macula, such as drusen, exudates, or epithelial abnormalities, among others (AREDS Research Group, 2001). In contrast, glaucoma is typically diagnosed after a detailed analysis of the optic disc morphology, including the optic cup and rim (Weinreb et al., 2014). These clinical criteria are adopted by cropping squared ROIs centered at the macula and the optic disc for the cases of AMD and glaucoma, respectively. Following the clinical standards, the cropped regions present a size of four times the average optic disc diameter for AMD (AREDS Research Group, 2001) and two times the average optic disc diameter for glaucoma (Weinreb et al., 2014). The automated detection of the macula center and the optic disc is performed following the method proposed in Hervella et al. (2020a).

Representative examples of retinographies and the corresponding ROIs for AMD and glaucoma are depicted in Figure 3. In the case of AMD, these examples show the great variety of pathological structures that may be present in this disease, ranging from very tiny lesions to larger structures that cover a substantial area in the macula. With regards to glaucoma, the examples show the common subtle differences between glaucomatous and non-glaucomatous eyes. In this case, the differences are typically focused on the internal optic disc morphology.

For each disease, the network training is performed using the binary cross-entropy (BCE) as loss function. Thus, the training loss for diagnosis is computed as:

$$\mathcal{L}_D = BCE(\mathbf{f}(\mathbf{r}), \mathbf{y}) \tag{1}$$

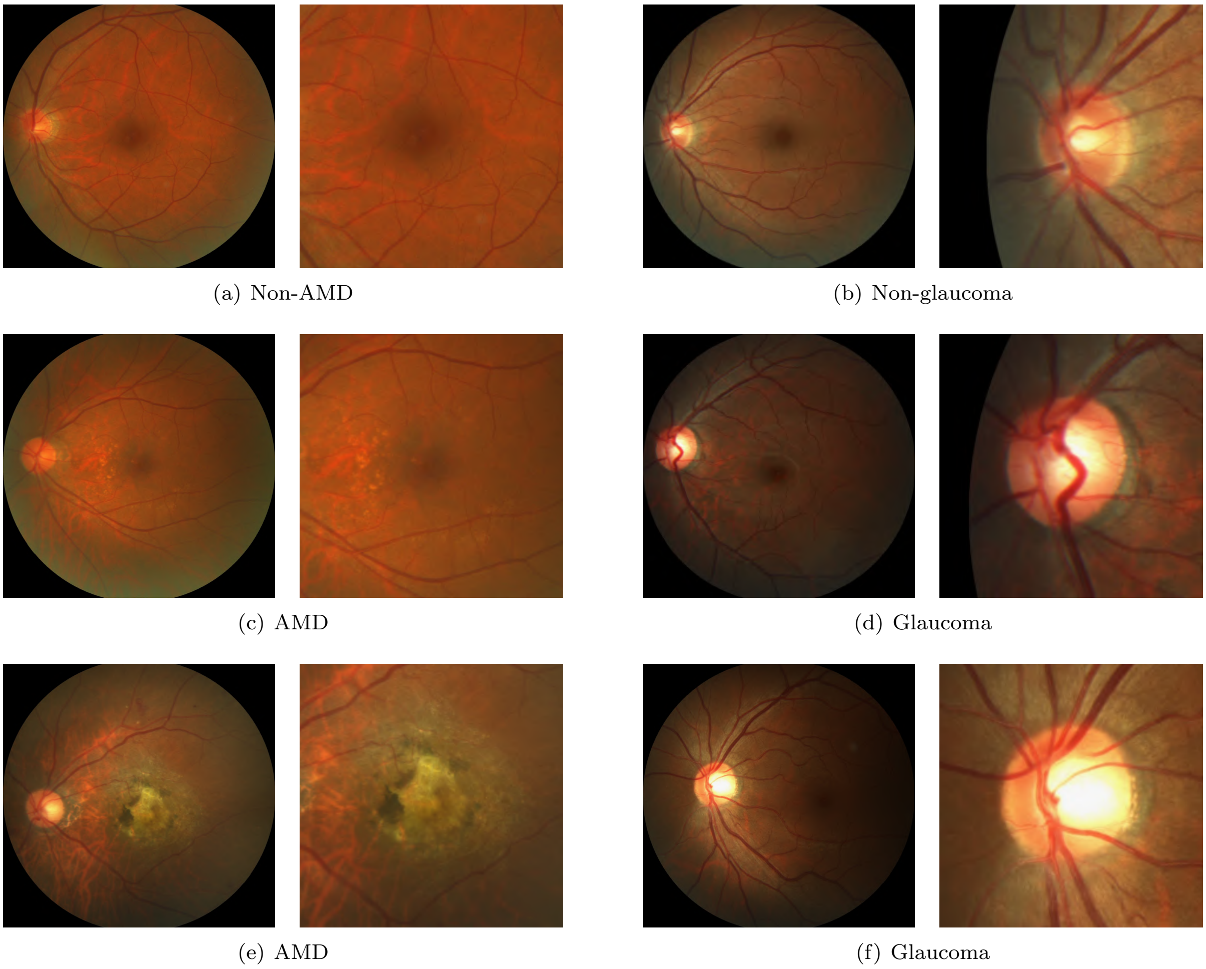
where $\mathbf{r}$ denotes the cropped retinography ROI, $\mathbf{y}$ its corresponding ground truth label, and $\mathbf{f}$ the transformation that assigns to each retinography $\mathbf{r}$ the likelihood of being a pathological sample.

2.2. Self-supervised multimodal pre-training

The multimodal reconstruction of fluorescein angiography from retinography is approached by using aligned retinography-angiography pairs as training data. The use of aligned image pairs results in a strong pixel-level supervision for learning the multimodal reconstruction task, as it allows the use of full-reference metrics between the network output and the aligned target image as loss function (Hervella et al., 2018b). The alignment of the multimodal image pairs is automatically performed following the domain-specific methodology proposed in Hervella et al. (2018a).

The multimodal reconstruction pre-training is applied to the specific ROI of each target task. The ROI required for the analysis of each disease is extracted from both the retinography and the angiography following the same criteria and methods indicated for the target classification task (Section 2.1). In this way, during the pre-training, the neural network will learn to recognize those retinal structures that are relevant for the specific target application.

For each disease, the multimodal reconstruction pre-training is performed using the negative Structural Similarity (SSIM) as loss function. The use of SSIM has demonstrated to provide a superior performance



(a) Non-AMD

(b) Non-glaucoma

(c) AMD

(d) Glaucoma

(e) AMD

(f) Glaucoma

Figure 3: Examples of retinographies and ROIs used for the diagnosis of ((a),(c),(e)) AMD and ((b),(d),(f)) glaucoma. For each image pair the retinography is in the left and the cropped ROI in the right.

for the multimodal reconstruction in comparison to other common metrics Hervella et al. (2018b). SSIM is a similarity metric that takes into account intensity, contrast, and structural differences between the images. For that purpose, SSIM requires the computation of a series of local statistics at each pixel position, such as the mean and the variance in each individual image, and the covariance between the images. These statistics are computed locally considering a small neighborhood for each pixel. Then, given a pair of pixels (x, y), the SSIM value between x and y is computed as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1) + (2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

where μ_x and μ_y denote the local means of x and y respectively, σ_x^2 and σ_y^2 the local variances of x and y respectively, σ_{xy} the local covariance between x and y, and c_1 and c_2 are constant values used to avoid instability when the denominator terms are close to zero (Wang et al., 2004). To avoid artifacts in the output, the local statistics are computed weighting the neighborhood of each pixel with a Gaussian window of $\sigma = 1.5$ (Wang et al., 2004).

Finally, the training loss for the multimodal reconstruction is computed as the negative mean SSIM

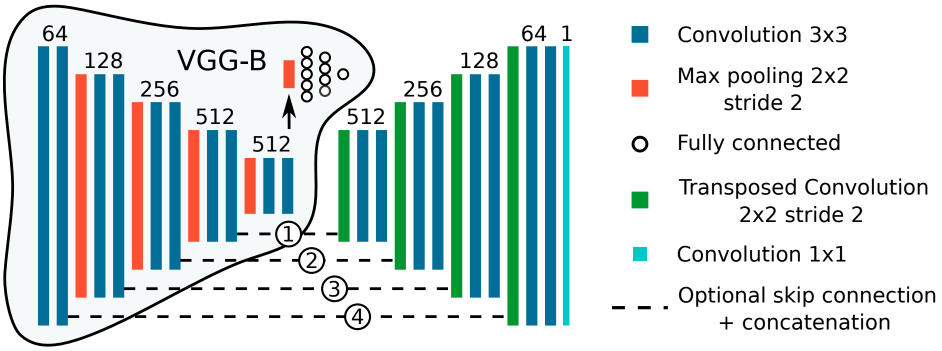


Figure 4: Network architectures that are used in this work. The multimodal self-supervised pre-training and the target classification task are performed using U-Net and VGG-Net, respectively. Both network share the layers of the convolutional encoder.

between the network prediction and the target:

$$\mathcal{L}_{MR} = -\frac{1}{N}\sum_{n=1}^{N} SSIM(\mathbf{g}(\mathbf{r})_n, \mathbf{a}_n) \tag{3}$$

where $\mathbf{r}$ denotes the cropped retinography ROI, $\mathbf{a}$ the corresponding angiography ROI, $\mathbf{g}$ the transformation that maps each retinography to its angiography counterpart, and N the number of pixels in the ROI.

2.3. Network architecture

In order to demonstrate the advantages of the proposed approach and provide a reference well-proven baseline, we adopt standard network architectures for both target and pre-training tasks. In particular, we use VGG-Net (Simonyan & Zisserman, 2015) for the target classification tasks and U-Net (Ronneberger et al., 2015) for the multimodal reconstruction pre-training. Both VGG-Net and U-Net represent well-proven network architectures for image-level and pixel-level prediction tasks, respectively (Houssein et al., 2020; Tariq et al., 2020). Additionally, both networks share numerous characteristics due to the fact that the U-Net layers are precisely based on the design of VGG-Net. As consequence, these networks allow for a straightforward transfer learning strategy by directly reusing the pre-trained U-Net encoder as the encoder of the VGG-Net in the target classification task. Figure 4 depicts a joint diagram of these networks that shows the close relationship between them.

Particularly, in this work, we use a VGG-Net with 13 layers (VGG-B) (Simonyan & Zisserman, 2015). This network consists of 10 convolutional layers followed by 3 fully connected layers. All the convolutions present kernels of size 3×3 and after every two convolutions there is a max pooling operation. In comparison with the 1000 classes of the ImageNet challenge (Deng et al., 2009), for which the network was originally designed, the classification tasks in this work only require the prediction of two classes: healthy or pathological. Thus, we adapt the network architecture by reducing the number of units in the 3 fully connected layers to 512, 128 and 1. A sigmoid activation function is used in the last layer to generate the binary prediction whereas the other layers have ReLU activation functions.

Regarding U-Net, this network architecture has already extensively demonstrated to be adequate for both the multimodal reconstruction (Hervella et al., 2020b) and transfer learning in this same application domain (Hervella et al., 2020a). In particular, U-Net is a fully convolutional network with a symmetric encoder-decoder structure and skip connections between encoder and decoder. These skip connections concatenate feature maps from the encoder with those of the same spatial resolution in the decoder. This particular design provides two main benefits. Firstly, precise spatial locations of the different extracted patterns are available in the decoder through the skip connections. This allows the precise generation of subtle details in the network output. Secondly, the skip connections ease the gradients back-propagation towards the early layers, which improves the network training.

The different convolutional blocks in U-Net are like those in VGG-Net. In particular, all the convolutions present kernels of size 3×3 and, in the encoder, there is a max pooling operation after every two convolutions. Similarly, in the decoder, there is a transpose convolution for upsampling every two convolutions. Then, the last layer consists of a convolution with kernel of size 1×1 and a linear activation function, whereas the other layers have ReLU activation functions.

Regarding the previously described transfer learning strategy, which consists in reusing the pre-trained U-Net encoder, we argue that it may be negatively affected by the skip connections in U-Net. In this sense, we should consider the effect of the skip connections in the high level representations learned by the encoder. Despite the positive effects in the network training, some relevant information related to the patterns that are learned in the early layers may never reach the network bottleneck (i.e., the encoder output), as they are directly forwarded to the decoder through the skip connections. In this case, the high level representations in the encoder will lack some information that may be relevant for the target classification task.

In this work, besides the standard U-Net architecture, we consider some variations of this network with a reduced number of skip connections for pre-training. The aim of this is to enforce that most of the relevant information reaches the network bottleneck during the multimodal reconstruction training. Additionally, this alternative keeps unchanged the classification network, which facilitates the comparison with other standard approaches for the target task. Initially, focusing on the requirements of the target classification task, it could be argued that avoiding all the skip connections would be the best alternative. However, this may excessively complicate the multimodal reconstruction training due to the difficulty of propagating precise spatial locations through the low resolution network bottleneck. In this regard, an inadequate pre-training could compromise the learning of representations that are useful for the target task. Thus, in order to study the most adequate configuration for transfer learning, we perform experiments with a varying numbers of skip connections, ranging from 0 to 4. In particular, in the performed experiments, the skip connections are added one at a time from the innermost to the outermost, as indicated in the numbering of Figure 4.

2.4. Training details

Regarding the neural networks, the initial parameters are drawn from a zero-centered normal distribution following the approach proposed by He et al. (2015). The optimization is performed using the Adam optimization algorithm (Kingma & Ba, 2015), with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and batch size of one image. The initial learning rate is set to $\alpha = 1e\text{-}4$ for the multimodal reconstruction and $\alpha = 1e\text{-}5$ for the classification tasks. Additionally, we apply a learning rate schedule that reduces the learning rate by a factor of 10 when the validation loss does not improve for 25 epochs. Finally, the training is stopped after 100 epochs without improvement in the validation loss. These particular values are empirically set by taking as reference previous works in the literature and analyzing the learning curves during the training. In order to apply the described settings, 25% of the training data is used as validation subset. Additionally, in order to adapt the images to the input requirements of the classification network, the cropped ROI for each disease is rescaled to a size of 224 × 224 pixels.

To avoid overfitting in both pre-training and target tasks, we apply online data augmentation consisting of random spatial and color transformations. The spatial transformations are comprised of rotation, scaling, and shearing for the multimodal reconstruction and rotation and shearing for the classification tasks. The color transformations were applied in HSV color space as proposed in Hervella et al. (2020b). Additionally, the range of the transformation parameters was selected so that transformed images are still considered as valid in appearance. Finally, in order to take into account the stochasticity of the networks training, we perform 5 repetitions with different random seeds for each experiment in the target classification tasks.

3. Experiments and results

In order to validate the proposed approach we perform a set of experiments focused on three main aspects. First, we evaluate the effect of the U-Net skip connections in the proposed multimodal reconstruction pre-training. We use AMD and glaucoma diagnosis from retinographies as case study for this evaluation. Second, under this same AMD and glaucoma use cases, we compare the proposed self-supervised pre-training method with commonly used fully-supervised baseline approaches, based on initializing the diagnosis network with random weights, and using ImageNet classification pre-training. Finally, our work is compared with Li et al. (2020) which, to the best of our knowledge, is the only related work in the literature using self-supervised approaches for retinal image analysis. Specifically. In order to provide comparable results, we follow the exact same experimental setting used in Li et al. (2020), to provide AMD and pathological myopia (PM) diagnosis from retinographies. In this case, the proposed multimodal self-supervised pre-training framework is directly applied without bells and whistles. The following sections provide the specific details and the obtained results for these three experimental settings.

3.1. Datasets

3.1.1. Multimodal reconstruction pre-training

For the proposed multimodal self-supervised pre-training, we use 59 retinography-angiography pairs from the public Isfahan MISP database (Alipour et al., 2012). In this dataset, half of the images correspond to patients diagnosed with diabetic retinopathy, an eye condition that arises as a complication of diabetes (Rahim et al., 2015, 2019). The other half of the images correspond to healthy individuals. All the images in this dataset are used for training/validation in the multimodal reconstruction.

3.1.2. Retinal CAD

For the diagnosis of glaucoma, we use 800 annotated retinographies from the public REFUGE dataset (Orlando et al., 2019). The prevalence of glaucoma in this dataset is 10%. The dataset includes a default split into two sets of 400 images each, named *Training* and *Validation*. In our experiments, we use the 400 images of *Training* as training data and the 400 images of *Validation* as hold-out test data.

For the diagnosis of AMD, we use 400 annotated retinographies from the public ADAM dataset (Fu et al., 2020). These images correspond to the *Training* split of this dataset, which is also used for the experiments in Li et al. (2020). The prevalence of AMD in this dataset is 23%. Similarly to glaucoma, we randomly split the dataset into two sets of 200 images with the same prevalence of AMD, one for training and the other as hold-out test data.

For the comparison with the state-of-the-art, we include an additional collection of images from the public PALM dataset (Fu et al., 2019). This dataset contains representative samples of retinas with pathological myopia (PM). In particular we use the 400 annotated retinographies from the *Training* split, which were also used in Li et al. (2020). The prevalence of PM in this dataset is 50%.

3.2. Evaluation

In order to quantitatively evaluate the proposed approach, the performance of the neural networks in the target classification tasks is evaluated using Receiver Operator Characteristic (ROC) analysis. This allows to directly evaluate the network predictions, which can be seen as the likelihood of the input samples being pathological, without the necessity of applying any specific decision threshold. In this way, we generate the ROC curves, which plot sensitivity and specificity for different decision thresholds. Additionally, we also compute the Area Under Curve (AUC) for ROC, which is commonly used to summarize the performance of the method into a single value.

3.3. Results

Figure 5 depicts the results obtained for the diagnosis of AMD and glaucoma using the proposed multimodal self-supervised pre-training. The performance is evaluated by means of AUC-ROC for a varying number of skip connections in the multimodal reconstruction network. In these experiments, the skip connections are added one at a time, starting with the innermost layers and following the order described in Figure 4. Additionally, for each number of skip connections, Figure 5 also depicts the performance of the multimodal reconstruction, i.e., the pre-training task, by means of SSIM in the validation set. In order to better appreciate the differences between the considered alternatives, Figure 6 depicts the complete ROC curves for each experiment in the target classification tasks.

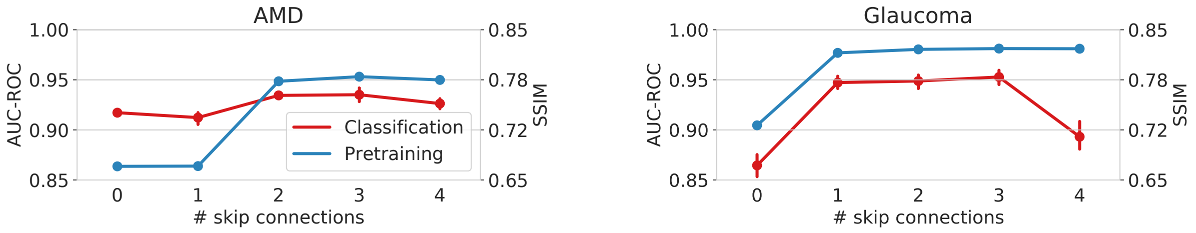


Figure 5: Performance of the target classification tasks and their corresponding multimodal self-supervised pre-training for a varying number of skip connections. The depicted results for the classification task represent the mean value and standard deviation for 5 repetitions of the experiments.

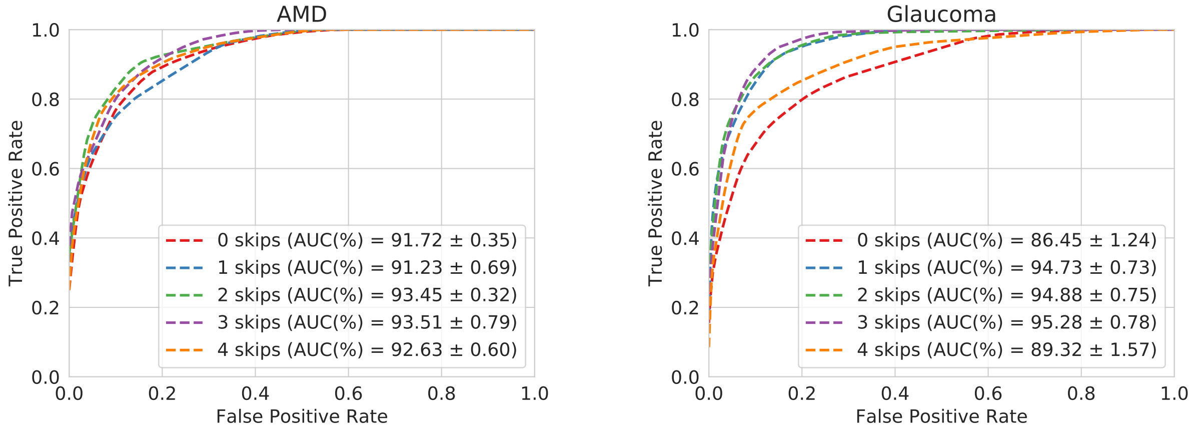
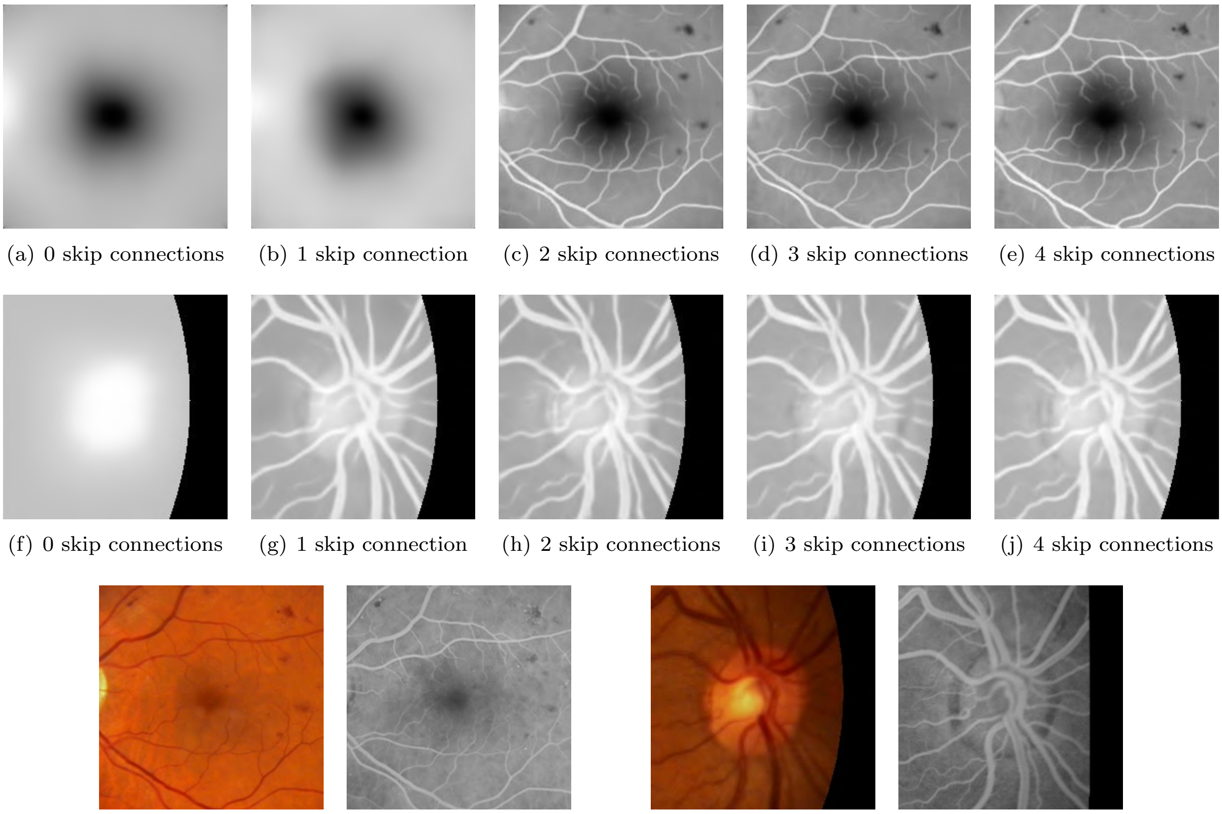


Figure 6: ROC analysis of the target classification tasks for a varying number of skip connections. The results are obtained from 5 repetitions of the experiments and the depicted ROC curves represent the average performance for each case.

In general, the obtained results show that the proposed approach produces a satisfactory performance for both AMD and glaucoma classification. However, the performance is not equally satisfactory in all the experiments. The best results are achieved with an intermediate number of skip connections. Particularly, $\{2, 3\}$ in the case of AMD and $\{1, 2, 3\}$ in the case of glaucoma. A lower number of skip connections results in a reduced performance for the target classification tasks and, also, for the multimodal reconstruction. The latter is expected due to the importance of the skip connections in the generation of detailed outputs at full resolution. In that sense, the lack of skip connections adversely affects the multimodal reconstruction, which, in turn, seems to compromise the learning of useful representations for the target classification task. Using all the skip connections also results in a reduction in the performance of the classification task, despite that the performance of the multimodal reconstruction is not significantly altered. These results fit with the idea that an extensive use of skip connections in the pre-training network may be detrimental for transfer learning purposes if only the pre-trained encoder is going to be reused.

To better understand the quantitative results, Figure 7 depicts representative examples of generated angiographies for a varying number of skip connections. It can be observed that the quality of the generated angiographies fits perfectly well with the quantitative multimodal reconstruction results depicted in Figure 5. In that sense, a minimum number of skip connections seems to be necessary to facilitate the adequate convergence of the multimodal reconstruction task. Nevertheless, it can be observed that, even in the worst cases, the network learns to recognize some important retinal structures and to generate a coarse representation of them. In particular, Figures 7(e) and 7(f) show that the network recognizes the center of the macula, whereas Figure 7(j) shows that the network broadly recognizes the optic disc. However, in these cases, many details such as the vasculature or lesions are missing.



(a) 0 skip connections (b) 1 skip connection (c) 2 skip connections (d) 3 skip connections (e) 4 skip connections

(f) 0 skip connections (g) 1 skip connection (h) 2 skip connections (i) 3 skip connections (j) 4 skip connections

(k) (l) (m) (n)

Figure 7: Representative examples of generated angiographies for a varying number of skip connections. (a-e) Examples corresponding to the macula region (AMD pre-training). (f-j) Examples corresponding to the optic disc region (glaucoma pre-training). (k) Input retinography and (l) target angiography for (a-e). (m) Input retinography and (n) target angiography for (f-j).

Regarding the comparison between AMD and glaucoma, it can be observed in Figures 5 and 6 that glaucoma classification is more sensitive to changes in the number of skip connections. It that sense, the classification of AMD keeps an adequate performance for all the experiments, even when no skip connections are used in the pre-training network. However, this is not the case for the classification of glaucoma, which experiments an important boost when the adequate pre-training settings are being used.

3.4. Comparison with fully-supervised approaches

To further analyze the advantages of the proposed approach, we perform a comparison against the most common alternatives in the literature, namely training the classification tasks from scratch and pre-training the networks on the annotated ImageNet dataset (Deng et al., 2009). Regarding the training from scratch, we use the initialization method proposed by He et al. (2015). In the case of the ImageNet pre-training, we use the pre-trained VGG-B network that is provided in the computer vision library of the PyTorch project (Paszke et al., 2017). It should be noticed that this network has been pre-trained in a fully-supervised fashion using more than a million annotated images. In contrast, the proposed multimodal self-supervised pre-training represents a novel alternative that only requires additional unlabeled data, and uses a dataset that is several orders of magnitude lower, counting with only 59 multimodal image pairs. In this regard, an important advantage of the proposed approach is that the size of the pre-training dataset could be increased without any human labeling effort.

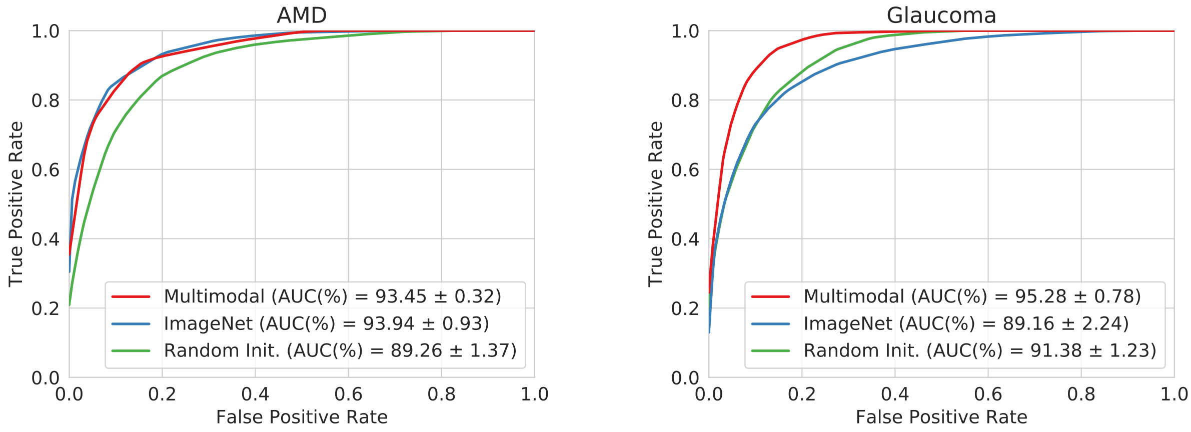


Figure 8: ROC curves for the target classification tasks using the proposed multimodal self-supervised pre-training (Multimodal), ImageNet classification pre-training (ImageNet) and training from scratch (Random Init.). The results are obtained from 5 repetitions of the experiments and the depicted ROC curves represent the average performance for each alternative.

Figure 8 depicts the comparison of the proposed approach against training from scratch and ImageNet pre-training for both AMD and glaucoma diagnosis. This comparison is performed using the best empirical configuration for each of the methods. In particular, the results for the multimodal self-supervised pre-training correspond to the number of skip connections that provides the best performance for each disease (i.e., 2 skip connections for AMD and 3 skip connections for glaucoma). In the case of the ImageNet pre-training, it is common to apply a normalization scheme to the input images based on the statistics of the ImageNet dataset. In this work, we explored fine-tuning on the application-specific datasets both with and without this normalization. The results presented in Figure 8 correspond to the configuration that provides the best performance for each disease, which is the default ImageNet normalization for AMD and no normalization for glaucoma. Regarding the obtained results, it is observed that the proposed approach outperforms the training from scratch by a significant margin in both AMD and glaucoma diseases. This evidences that the patterns learned for the multimodal reconstruction are also useful for the detailed analysis of important retinal areas such as the macula or the optic disc. Thus, the obtained results demonstrate that the proposed approach is able to successfully take advantage of the unlabeled multimodal data for transfer learning purposes. Regarding the comparison with the ImageNet pre-training, the proposed self-supervised pre-training achieves a similar performance for the diagnosis of AMD despite not requiring any additional annotated data. Moreover, in the case of glaucoma, the proposed approach even outperforms the ImageNet pre-training by a significant margin. In this regard, it should be noticed that most of the self-supervised alternatives in the state-of-the-art are not able to equal the performance of the ImageNet pre-training Jing & Tian (2020). Considering this, the proposed approach offers a remarkable performance.

The results presented in Figure 8 show that the ImageNet pre-training provides a satisfactory performance improvement for the diagnosis of AMD, while in the case of glaucoma the performance is slightly lower than that of the random initialization. These two ocular diseases require the analysis of different areas of the retina, namely the macula and the optic disc, involving features of different nature. Additionally, AMD is typically diagnosed by detecting the presence of certain local pathological lesions, whereas the diagnosis of glaucoma is typically performed by analyzing the morphology of the optic disc. Thus, a plausible explanation could be the recently demonstrated bias of the ImageNet pre-trained networks towards recognizing textures rather than shapes (Geirhos et al., 2019). In that sense, in our experiments, the ImageNet pre-trained network excels when the detection of subtle abnormalities is required (AMD) but falls behind when the morphological properties become more important (glaucoma).

In practice, we have also observed that the networks pre-trained on ImageNet present a significantly larger generalization gap in comparison to the other alternatives. This indicates that the networks' predictions

Table 1: Comparison with state-of-the-art self-supervised approaches for retinal computer-aided diagnosis.

Method	AMD	PM
	AUCROC(%)	AUCROC(%)
Li et al. (2020)	83.17	98.41
Proposed	89.57 ± 3.22	99.48 ± 0.58

tend to rely more on patterns that are specific to particular images, instead of those that are common to all the images of the same class (healthy or pathological). Thus, although ImageNet pre-training provides a very rich set of patterns that improves the network's training, this does not necessarily translate to a better performance in unseen images. This issue seems to be aggravated in the case of glaucoma due to the atypical morphological analysis that is required. Additionally, it should be noticed that the performance for the diagnosis of glaucoma also decreases when the proposed multimodal self-supervised pre-training does not properly converge due to the lack of skip connections in the network architecture (compare AUC-ROC values of Figures 6 and 8). Therefore, in our experiments, it is clear that providing an adequate pre-training for the diagnosis of glaucoma is more challenging than doing the same for the diagnosis of AMD.

3.5. Comparison with state-of-the-art self-supervised approaches

In this section, we perform a comparison of the proposed approach against existing self-supervised approaches in the literature. In particular, only one prior work has recently proposed an alternative self-supervised approach for retinal image analysis (Li et al., 2020). In order to adequately produce a fair comparison with this approach, we perform additional experiments using the same configuration that is adopted in Li et al. (2020). In particular, for these experiments, the whole images are used as input to the network. For this, the images are rescaled to a size of 224 × 224 pixels. The experiments are performed for the diagnosis of AMD and pathological myopia (PM) following a 5-fold cross-validation approach. In this case, we avoid to specifically tailor the methodology for each particular diseases and use the U-Net variant with 2 skip connections for all the experiments.

The comparison with the state-of-the-art is depicted in Table 1. In our case, we provide both the mean and standard deviation of the obtained results. It is observed that the proposed approach clearly outperforms the state-of-the-art alternative in both datasets. The difference in performance is greater for AMD, however, it is significant in both cases, especially considering the reduced standard deviation of our approach in PM. Moreover, besides the remarkable improvement in performance, our approach offers additional important advantages. Particularly, our proposal directly exploits the available paired multimodal data in a single pre-training step, whereas the method proposed in Li et al. (2020) requires two separate training stages and several neural networks in the pre-training phase. Thus, our proposal represents a more efficient and straightforward alternative. Moreover, to achieve the results reported in Li et al. (2020), the authors required to combine the multimodal synthesis augmentation with regular augmentation approaches used in broad domain self-supervised approaches (Ye et al., 2019), so the contribution of the multimodal-based self-supervision is not as clearly exploited in that approach. Instead the contribution of the multimodal information is clearly exploited in our method, without the need of complementary self-supervision tasks. Finally, the proposed approach, which is based on a pixel-level prediction (the multimodal reconstruction) allows the simultaneous pre-training of both image-level and pixel-level target tasks, such as, e.g., classification and segmentation. Thus, in contrast to the previous alternative, the proposed approach is also adequate for all kind of multi-task settings.

4. Conclusions

Nowadays, deep learning algorithms are commonly used in CAD systems. However, the performance of these methods is limited by the availability of sufficient annotated data. In order to mitigate this issue, we propose a self-supervised pre-training for deep learning-based retinal CAD systems consisting in the multimodal reconstruction between complementary imaging modalities. This approach exploits common existent unlabeled multimodal data in the medical domain for learning useful domain-specific representations.

The advantages of the proposed approach are mainly demonstrated in the context of two of the most prevalent impairing ocular diseases: AMD and glaucoma. We performed several experiments to analyze this novel transfer learning paradigm, including the study of important factors regarding the network architectures. In order to demonstrate the relevance of the proposed approach, we performed a comparison against two common fully-supervised approaches, namely training the network from scratch and pre-training on the annotated ImageNet dataset. Additionally, we also provide a comparison against existing self-supervised alternatives in retinal image analysis, including experiments in additional scenarios such as pathological myopia. The obtained results demonstrate that the proposed approach offers a satisfactory performance in all the pathological scenarios. Moreover, the multimodal reconstruction pre-training significantly outperforms both the training from scratch and the state-of-the-art alternatives, while it also demonstrates to be an overall superior approach to ImageNet pre-training.

Finally, given the excellent results that were obtained in all the pathological scenarios, in future work we plan to study the application of the proposed approach in other medical domains where multimodal visual data is also commonly available.

Acknowledgments

This work is 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; Ministerio de Ciencia e Innovación, Government of Spain, through the RTI2018-095894-B-I00 and PID2019-108435RB-I00 research projects; Xunta de Galicia and the European Social Fund (ESF) of the EU through the predoctoral grant contract ref. ED481A-2017/328; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, through Grupos de Referencia Competitiva, grant ref. ED431C 2020/24. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

Conflict of interest

The authors declare no conflicts of interest.

References

Alipour, S. H. M., Rabbani, H., & Akhlaghi, M. R. (2012). Diabetic retinopathy grading by digital curvelet transform. *Computational and mathematical methods in medicine*, *2012*. doi:doi:10.1155/2012/761901.

AREDS Research Group (2001). The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the age-related eye disease study report number 6. *American Journal of Ophthalmology*, *132*, 668–681. doi:doi:10.1016/S0002-9394(01)01218-1.

Bloice, M. D., Roth, P. M., & Holzinger, A. (2019). Biomedical image augmentation using Augmentor. *Bioinformatics*, *35*, 4522–4524. doi:doi:10.1093/bioinformatics/btz259.

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems (NIPS)*. volume 33.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, *58*, 101539. doi:doi:10.1016/j.media.2019.101539.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Cheplygina, V., de Bruijne, M., & Pluim, J. P. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, *54*, 280 – 296. doi:doi:10.1016/j.media.2019.03.009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fu, H., Li, F., Orlando, J. I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., & Zhang, X. (2019). Palm: Pathologic myopia challenge. doi:doi:10.21227/55pk-8z03.

Fu, H., Li, F., Orlando, J. I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., & Zhang, X. (2020). Adam: Automatic detection challenge on age-related macular degeneration. doi:doi:10.21227/dt4f-rt59.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2018a). Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Computer Science*, *126*, 97 – 104. doi:doi:10.1016/j.procs.2018.07.213.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2018b). Retinal image understanding emerges from self-supervised multimodal reconstruction. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. doi:doi: 10.1007/978-3-030-00928-1_37.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2020a). Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. *Applied Soft Computing*, (p. 106210). doi:doi:10.1016/j.asoc.2020.106210.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2020b). Self-supervised multimodal reconstruction of retinal images over paired datasets. *Expert Systems with Applications*, (p. 113674). doi:doi:10.1016/j.eswa.2020.113674.

Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2020). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, (p. 114161). doi:doi:10.1016/j.eswa.2020.114161.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–1). doi:doi:10.1109/TPAMI.2020.2992393.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Li, X., Jia, M., Islam, M. T., Yu, L., & Xing, L. (2020). Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, (pp. 1–1). doi:doi:10.1109/TMI.2020.3008871.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60 – 88. doi:doi:10.1016/j.media.2017.07.005.

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*.

Oh, C., Ham, B., Kim, H., Hilton, A., & Sohn, K. (2019). Ocean: Object-centric arranging network for self-supervised visual representations learning. *Expert Systems with Applications*, *125*, 281 – 292. doi:doi:10.1016/j.eswa.2019.01.073.

Orlando, J. I. et al. (2019). REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. *Medical Image Analysis*, (p. 101570). doi:doi:10.1016/j.media.2019.101570.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. (2016). Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 3347–3357). volume 32.

Rahim, S. S., Palade, V., Almakky, I., & Holzinger, A. (2019). Detection of diabetic retinopathy and maculopathy in eye fundus images using deep learning and image augmentation. In *Machine Learning and Knowledge Extraction* (pp. 114–127). Cham: Springer International Publishing.

Rahim, S. S., Palade, V., Jayne, C., Holzinger, A., & Shuttleworth, J. (2015). Detection of diabetic retinopathy and maculopathy in eye fundus images using fuzzy image processing. In Y. Guo, K. Friston, F. Aldo, S. Hill, & H. Peng (Eds.), *Brain Informatics and Health* (pp. 379–388). Cham: Springer International Publishing.

Ronneberger, O., P.Fischer, & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., & Maier-Hein, L. (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, *13*, 925–933. doi:doi:10.1007/s11548-018-1772-0.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*, 1299–1312. doi:doi:10.1109/TMI.2016.2535302.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., & Lippert, C. (2020). 3d self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems (NIPS)*. volume 33.

Tariq, M., Iqbal, S., Ayesha, H., Abbas, I., Ahmad, K. T., & Niazi, M. F. K. (2020). Medical image based breast cancer diagnosis: State of the art and future directions. *Expert Systems with Applications*, (p. 114095). doi:doi:10.1016/j.eswa.2020.114095.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.

Weinreb, R. N., Aung, T., & Medeiros, F. A. (2014). The pathophysiology and treatment of glaucoma: a review. *JAMA*, *311*, 1901–11.

Wong, K. C., Syeda-Mahmood, T., & Moradi, M. (2018). Building medical image classifiers with very limited data using segmentation networks. *Medical Image Analysis*, *49*, 105–116. doi:doi:10.1016/j.media.2018.07.010.

Ye, M., Zhang, X., Yuen, P. C., & Chang, S.-F. (2019). Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European Conference on Computer Vision (ECCV)*.

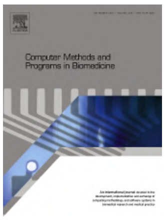
Appendix A

Publications and other mentions

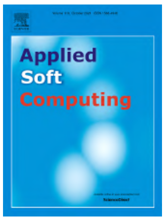
In this appendix, we list all the articles published during the PhD period. For the conferences, we provide their rank according to CORE Conference Ranking [77]. For each journal, we detail its impact factor according to Journal Citation Reports (JCR) index [78] and its corresponding quartile.

A.1 JCR-indexed Journals

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Systems with Applications, Volume 161, 2020. IF 2020: 6.954 (Q1).



Álvaro S. Hervella, José Rouco, Jorge Novo, Manuel G. Penedo, Marcos Ortega, Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Computer Methods and Programs in Biomedicine, Volume 186, 2020. IF 2020: 5.428 (Q1).



Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega, Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction, Applied Soft Computing, Volume 91, 2020. IF 2020: 6.725 (Q1).

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega, Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis, Expert Systems with Applications, Volume 185, 2021. IF 2020: 6.954 (Q1).

José Morano, Álvaro S. Hervella, José Rouco, Jorge Novo, Simultaneous segmentation and classification of the retinal arteries and veins from color fundus images, Artificial Intelligence in Medicine, Volume 118, 2021. IF 2020: 5.326 (Q1).

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega, Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images, Information Fusion, 2021. IF 2020: 12.975 (Q1).

A.2 Book Chapters

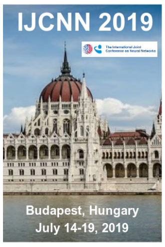
Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega, Chapter 15 - Multimodal reconstruction of retinal images over unpaired datasets using cyclical generative adversarial networks, Generative Adversarial Networks for Image-to-Image Translation, Academic Press, 2021, 347-376, ISBN 9780128235195

A.3 International Conferences

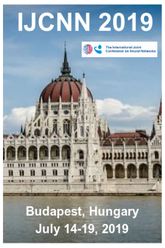
Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. Procedia Computer Science: Knowledge-Based and Intelligent Information & Engineering Systems (KES), 126, 97-104, 2018. CORE 2018: B.

Alejandro F. Frangi · Julia A. Schnabel
Christos Davatzikos · Carlos Alberola-López
Gabor Fichtinger (Eds.)
LNCS 11070
Medical Image Computing and Computer Assisted Intervention – MICCAI 2018
21st International Conference
Granada, Spain, September 16–20, 2018
Proceedings, Part I
1 Part I
MICCAI
Springer

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Retinal Image Understanding Emerges from Self-Supervised Multimodal Reconstruction. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science 11070, 2018. CORE 2018: A.



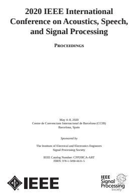
Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Self-Supervised Deep Learning for Retinal Vessel Segmentation Using Automatically Generated Labels from Multimodal Data. 2019 International Joint Conference on Neural Networks (IJCNN), CORE 2019: A.



Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Deep Multimodal Reconstruction of Retinal Images Using Paired or Unpaired Data. 2019 International Joint Conference on Neural Networks (IJCNN), CORE 2019: A.

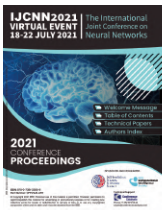
Roberto Moreno-Díaz
Franz Pichler
Alexis Quesada-Arencibia (Eds.)
LNCS 12014
Computer Aided Systems Theory – EUROCAST 2019
17th International Conference
Las Palmas de Gran Canaria, Spain, February 17–22, 2019
Revised Selected Papers, Part II
2 Part II
Springer

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Impact of the Circular Region of Interest on the Performance of Multimodal Reconstruction of Retinal Images. Computer Aided Systems Theory – EUROCAST 2019, Lecture Notes in Computer Science 12014, 2019.


2020 IEEE International Conference on Acoustics, Speech, and Signal Processing
Proceedings
May 4–8, 2020
Centre de Convencions Internacional de Barcelona (CCIB)
Barcelona, Spain
Sponsored by
The Institute of Electrical and Electronics Engineers
Signal Processing Society
IEEE Catalog Number: CFP20ICA-ART
ISBN: 978-1-5090-6631-5
IEEE
IEEE Signal Processing Society

Álvaro S. Hervella, Lucía Ramos, José Rouco, Jorge Novo, Marcos Ortega. Multi-Modal Self-Supervised Pre-Training for Joint Optic Disc and Cup Segmentation in Eye Fundus Images. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 961-965, 2020

José Morano, Álvaro S. Hervella, Noelia Barreira, Jorge Novo, José Rouco. Multimodal Transfer Learning-Based Approaches for Retinal Vascular Segmentation. European Conference on Artificial Ingelligence – ECAI 2020, Frontiers in Artificial Ingelligence and Applications, 325, 1866-1873, CORE 2020: A.



Daniel I. Morís, Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Context encoder self-supervised approaches for eye fundus analysis. 2021 International Joint Conference on Neural Networks (IJCNN), CORE 2020: A.

A.4 National Conferences

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Learning Retinal Patterns from Multimodal Image. I Congreso XoveTIC, MDPI Proceedings, 2, 2018.

Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Paired and unpaired deep generative models on multimodal retinal image reconstruction. II Congreso XoveTIC, MDPI Proceedings, 21, 2019.

Álvaro S. Hervella, Lucía Ramos, José Rouco, Jorge Novo, Marcos Ortega. Joint Optic Disc and Cup Segmentation Using Self-Supervised Multimodal Reconstruction Pre-Training. III Congreso XoveTIC, MDPI Proceedings, 54, 2020.

José Morano, Álvaro S. Hervella, Noelia Barreira, José Rouco, Jorge Novo. Enhancing Retinal Blood Vessel Segmentation through Self-Supervised Pre-Training. III Congreso XoveTIC, MDPI Proceedings, 54, 2020.

engineering proceedings
The 4th XoveTIC Conference
A Coruña, Spain | 7-8 October 2021
https://xovetic.citic.udc.es/en/home/
MDPI

José Morano, Álvaro S. Hervella, José Rouco, Jorge Novo. Deep Multi-Segmentation Approach for the Joint Classification and Segmentation of the Retinal Arterial and Venous Trees in Color Fundus Images. IV Congreso XoveTIC, Engineering Proceedings, 7, 2021.

A.5 Under review process

- Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images, 2021.
- Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Balanced Multi-Task Learning Using a Multi-Adaptive Optimization Strategy, 2021.
- Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega. Multimodal image encoding pre-training for diabetic retinopathy grading, 2021

Appendix B

Extended Summary in Spanish

De acuerdo con la normativa de la Universidad de A Coruña para los estudios de doctorado, en este anexo se presenta un resumen extendido de la tesis doctoral en castellano. Esta tesis doctoral se estructura como una tesis por compendio de publicaciones y este anexo resume brevemente el trabajo de investigación incluido en la tesis. En primer lugar, se proporciona la motivación y el contexto del trabajo de investigación que se incluye en la tesis, así como los objetivos previstos. A continuación, para dar coherencia y consistencia a la tesis por compendio de publicaciones, se incluye una breve discusión sobre los diferentes artículos de investigación. Por último, se extraen conclusiones generales y se discuten los posibles trabajos futuros derivados de esta tesis doctoral.

B.1 Introducción y motivación

Las técnicas de imagen médica tienen un papel destacado en la investigación y práctica clínica moderna [1]. Hoy en día, es habitual el uso de múltiples modalidades de imagen para facilitar el diagnóstico, tratamiento y seguimiento de los pacientes [1, 2]. Estas técnicas permiten visualizar y estudiar los diferentes órganos y tejidos del cuerpo humano [1]. Así, estas técnicas pueden ser utilizadas por los clínicos para analizar las diferentes estructuras anatómicas que pueden estar afectadas por una enfermedad o para encontrar posibles lesiones. Sin embargo, en muchos casos, el análisis de las imágenes es difícil y laborioso [3, 4]. Por ejemplo, muchas enfermedades sólo muestren anonalías sutiles o lesiones muy pequeñas en sus etapas más tempranas. Para detectar y analizar adecuadamente estas sutiles evidencias de la enfermedad, el análisis de las imágenes debe ser realizado cuidadosamente por clínicos con amplia experiencia. En este sentido, las herramientas automáticas para el análisis de imagen médica representan una ayuda crucial para los clínicos, ayudando a aliviar su carga de trabajo y mejorando potencialmente la fiabilidad del diagnóstico [5, 6, 7].

El uso de múltiples modalidades de imagen está ampliamente extendido en el estudio del ojo humano [2, 1]. El análisis de las imágenes de la retina (o del fondo de ojo) es crucial para el diagnóstico de numerosas patologías [8], incluyendo trastornos oftálmicos como glaucoma [9] o la degeneración macular asociada a la edad (DMAE) [10], así como enfermedades sistémicas que afectan al ojo como la diabetes [11] o la hipertensión [12]. Hoy en día, la modalidad de imagen de la retina más asequible y ampliamente disponible es la retinografía [13, 6]. Estas imágenes de la retina son fotografías a color del fondo del ojo que muestran estructuras anatómicas relevantes como la microvasculatura de la retina, la fóvea o el disco óptico. Además, estructuras patológicas que son relevantes para el diagnóstico de numerosas enfermedades, como hemorragias, exudados o drusas, también se pueden observar en estas imágenes.

Además de la retinografía, existen otras modalidades de imagen de la retina como la angiografía con fluoresceína (AF), la oftalmoscopia láser de barrido o la tomografía de coherencia óptica [2, 13]. Estas técnicas suelen ofrecer algunas ventajas en cuanto a la visualización de las estructuras y los tejidos de la retina. Sin embargo, no suelen ser tan comunes debido a la necesidad de equipos más complejos o procedimientos invasivos para los pacientes. Por el contrario, la retinografía es una técnica no invasiva que se puede realizar con equipos relativamente asequibles. En este sentido, hoy en día, incluso es posible obtener retinografías con dispositivos portátiles especializados [14]. Por estas razones, la retinografía representa una herramienta valiosa en el contexto de los programas preventivos de salud y el cribado de grandes poblaciones [5, 6].

En los últimos años, ha habido un gran interés en el desarrollo de métodos automáticos para el análisis de imágenes de fondo de ojo [15, 16, 17]. En este sentido, hay varios ejemplos de sistemas de diagnóstico asistido por ordenador (DAO) que se utilizan en diferentes servicios de salud o programas de cribado en todo el mundo [5]. Actualmente, los métodos más exitosos son los que se basan en algoritmos de aprendizaje profundo [18, 17]. Al igual que en otras áreas de la visión por ordenador, el uso de redes neuronales profundas (RNPs) ha supuesto una mejora significativa para muchas aplicaciones médicas [18]. Además, estos algoritmos suelen dar lugar a metodologías más sencillas y adaptables, evitando la ingeniería manual de características que se requiere para los algoritmos clásicos de aprendizaje automático [19, 20].

El reciente auge y difusión del aprendizaje profundo ha estado motivado por diferentes factores, entre los que se encuentran los desarrollos técnicos que facilitaron el entrenamiento de las RNPs, la disponibilidad de conjuntos masivos de datos o el aumento de la potencia computacional comúnmente disponible [21, 22]. Sin embargo, en muchas áreas, la disponibilidad de datos etiquetados sigue siendo un factor limitante para la aplicación exitosa de algoritmos de aprendizaje profundo. Este problema es especialmente común en el análisis de imagen médica, dado que el etiquetado manual de las imágenes requiere un profundo conocimiento médico y alto nivel de experiencia [26, 18, 27]. En este sentido, lo ideal es que el etiquetado de las imágenes sea realizado por especialistas clínicos con años de experiencia práctica en el tipo de análisis que se requiere. Además, dada la alta variabilidad entre expertos que puede esperarse para algunos análisis especialmente difíciles, se suele requerir un consenso que tenga en cuenta las anotaciones de varios expertos [3, 4]. Estos factores suelen limitar el tamaño de los conjuntos de datos etiquetados que están disponibles en el ámbito de la imagen médica.

La escasez de datos etiquetados en el ámbito de la imagen médica puede paliarse siguiendo diferentes enfoques [18, 27]. En primer lugar, en el caso de las etiquetas globales de imagen, se pueden destilar etiquetas adicionales de los informes clínicos de pacientes existentes [18]. Sin embargo, este enfoque no puede aplicarse a las etiquetas a nivel de píxel, que son necesarias para tareas como la segmentación. Además, la anotación manual de etiquetas a nivel de píxel es especialmente difícil y laboriosa, lo que se refleja en el número significativamente menor de imágenes etiquetadas para este tipo de tareas [28, 29]. En segundo lugar, estrategias de aumento de datos son comúnmente utilizadas y representan una herramienta clave para lograr buenos resultados cuando los datos de entrenamiento son limitados [18]. Estas strategias pretenden simular nuevas muestras aplicando transformaciones de color y espaciales a las imágenes etiquetadas disponibles [30]. En este contexto, también hay un creciente interés en el desarrollo de métodos automáticos para la generación de muestras sintéticas utilizando RNPs [31]. Sin embargo, estos métodos presentan el riesgo de producir contenidos no plausibles en las imágenes [32]. Finalmente, un enfoque ampliamente extendido para el entrenamiento de las RNPs es el aprendizaje por transferencia [18, 27, 33]. En general, el aprendizaje por transferencia consiste en aprovechar los conocimientos adquiridos en el entrenamiento de una tarea para resolver otro problema relacionado. Este enfoque suele aplicarse de forma secuencial, preentrenando primero una RNP en una tarea auxiliar con un gran conjunto de datos etiquetados y, a continuación, refinando la red en una tarea objetivo con etiquetas limitadas. Sin embargo, también es posible aprovechar este enfoque entrenando simultáneamente ambas tareas. En este escenario multitarea, ambas tareas podrían beneficiarse de los datos de entrenamiento de la otra [34, 35]

Durante años, el enfoque habitual para el aprendizaje por transferencia en imagen médica ha sido el uso de un pre-entrenamiento totalmente supervisado realizado en un conjunto de datos masivo de imágenes naturales [36, 37] como ImageNet [24]. A pesar de la diferente naturaleza de las imágenes en este preentrenamiento, este enfoque ha demostrado facilitar el entrenamiento de numerosas tareas objetivo, independientemente del dominio de aplicación final [33]. Aun así, se podría argumentar que un preentrenamiento dentro del dominio de la aplicación final debería proporcionar representaciones de alto nivel más relevantes para la tarea objetivo, mejorando los resultados del aprendizaje por transferencia.

Recientemente, el aprendizaje autosupervisado ha surgido como una alternativa prometedora a los enfoques tradicionales totalmente supervisados para el aprendizaje por transferencia [38, 39]. En el paradigma autosupervisado, los objetivos de entrenamiento (o etiquetas) se derivan automáticamente de los datos de entrenamiento sin etiquetar. De este modo, se puede proporcionar una señal de supervisión a la red sin necesidad de realizar ningun etiquetado manual. Esto permite el preentrenamiento de una RNP utilizando imágenes del dominio de aplicación final. Los métodos autosupervisados existentes podrían dividirse, en términos generales, en tareas generativas o contrastivas [40]. La familia generativa autosupervisada se basa en la predicción de muestras ocultas de los datos o en la predicción de relaciones ocultas entre diferentes muestras de datos [40]. Por ejemplo, este tipo de aprendizaje autosupervisado puede realizarse mediante la predicción de regiones enmascaradas en una imagen de entrada [41, 42], la predicción de relaciones geométricas entre diferentes regiones candidatas en detección de objetos [43], o la predicción de la relación temporal entre diferentes fotogramas de un vídeo [44, 45]. Además, durante esta tesis doctoral, hemos propuesto una nueva alternativa autosupervisada consistente en la predicción de una modalidad de imagen médica complementaria [46, 47]. Con respecto a la familia autosupervisada contrastiva, el objetivo de entrenamiento es obtener una representación de alto nivel que maximice la similitud entre muestras de datos relacionadas [40, 39]. Estas muestras relacionadas suelen obtenerse aplicando técnicas estándar de aumento de datos a los datos brutos sin etiquetar. Este tipo de aprendizaje autosupervisado ha sido explorado recientemente en varios trabajos, que proponen diferentes arquitecturas de red y procedimientos de entrenamiento con el objetivo de aprovechar mejor el paradigma del aprendizaje contrastivo [48, 49].

B.2 Objetivos

La tesis doctoral que se presenta se centra en el desarrollo de nuevas metodologías de aprendizaje profundo para el análisis automático de imagen médica. En particular, se pretende aplicar las metodologías desarrolladas al análisis automático de imágenes de la retina. Los principales objetivos de la tesis doctoral se pueden resumir como sigue:

- Desarrollo de nuevas metodologías basadas en aprendizaje profundo para el análisis de imagen médica que reducen la necesidad de conjuntos masivos de datos etiquetados manualmente y pueden aplicarse a imágenes de alta resolución.
- Desarrollo de nuevas metodologías de análisis de imagen médica para mejorar la prevención y el diagnóstico de enfermedades oftálmicas y vasculares.

Asimismo, se definen los siguientes objetivos específicos de la tesis doctoral:

- Mejorar la detección y el análisis de estructuras anatómicas y patológicas en retinografía.
- Obtener un realce automático de la microvasculatura retiniana en retinografía.
- Explorar el uso de múltiples modalidades de imagen para los algoritmos desarrollados.
- Desarrollo de metodologías que no requieran conjuntos masivos de datos etiquetados manualmente.
- Desarrollo de metodologías que puedan aplicarse a imágenes de alta resolución.

B.3 Investigación y Discusión General

Esta sección ofrece al lector una visión general del trabajo de investigación incluido en la tesis doctoral. En particular, la sección ofrece un breve resumen y una discusión general de todas las publicaciones anexas que constituyen esta tesis por compendio de publicaciones. El trabajo de investigación incluido en la tesis por compendio comprende 4 artículos de revistas indexadas en JCR, 1 capítulo de libro y 4 artículos de conferencias internacionales. Atendiendo a su contenido y objetivos, estas publicaciones se organizan en 3 bloques diferentes: Parte I - Reconstructión Multimodal de Imágenes de la Retina, Parte II - Análisis de Estructuras de la Retina, Parte III - Diagnóstico Asistido por Ordenador.

B.3.1 Parte I - Reconstructión Multimodal de Imágenes de la Retina

En la práctica clínica moderna, es habitual el uso de diferentes modalidades de imagen que proporcionan visualizaciones complementarias de los mismos órganos o tejidos [2, 50, 13]. Las diferentes características visuales entre modalidades complementarias se deben principalmente al uso de diferentes dispositivos de captura o a contrastes inyectados que mejoran la visualización de ciertos tejidos. En este sentido, los clínicos deben elegir la modalidad de imagen más adecuada para cada caso. Aunque, en los casos más complejos, es habitual el uso de múltiples modalidades de imagen complementarias para el mismo paciente. Esto facilita la obtención de colecciones de imágenes multimodales. Sin embargo, los conjuntos de datos multimodales disponibles normalmente sólo se utilizan cuando las imágenes también están etiquetadas. En este sentido, hay varios ejemplos en la literatura de métodos automáticos que hacen una predicción basada en una entrada multimodal [51]. En este caso, los datos multimodales deben estar disponibles tanto para el entrenamiengo como para la inferencia. Además, los datos deben estar etiquetados para la fase de entrenamiento. Sin embargo, las diferencias entre modalidades complementarias representan una fuente de supervisión en sí mismas, sin necesidad de etiquetas realizadas manualmente. Por ejemplo, en esta tesis doctoral, hemos propuesto una nueva metodología de aprendizaje autosupervisado que consiste en la predicción de una modalidad de imagen a partir de otra [46]. Para resolver esta tarea, una RNP debe reconocer primero los diferentes elementos que componen la imagen de entrada, incluyendo diferentes estructuras anatómicas y patológicas. A continuación, la red neuronal debe aplicar la transformación más adecuada para cada uno de los elementos identificados y generar como salida la composición de todos los elementos transformados. Este complejo proceso requiere la capacidad de reconocer numerosos patrones específicos del dominio, así como tener un conocimiento de alto nivel del contenido de las imágenes. Por tanto, esta reconstrucción multimodal (RM) de modalidades de imagen complementarias puede utilizarse como una tarea auxiliar autosupervisada con fines de aprendizaje por transferencia. Además, la propia RM proporciona una estimación potencialmente valiosa de una modalidad de imagen adicional.

Durante esta tesis doctoral, exploramos la idea de la RM en el contexto del análisis de imágenes de la retina. En particular, nuestro objetivo era el desarrollo de métodos novedosos para el análisis de retinografía, por lo que esta modalidad de imagen se utiliza como entrada en el método de RM propuesto. Como modalidad de imagen objetivo, utilizamos la AF, una modalidad complementaria que proporciona una visualización mejorada de la microvasculatura retiniana. En este caso, la inyección de un contraste intravenoso produce un cambio drástico en la apariencia de las diferentes estructuras anatómicas y patológicas en las imágenes.

La metodología de RM propuesta se basa en el uso de datos multimodales pareados, concretamente pares retinografía-AF donde ambas imágenes corresponden al mismo ojo. Estos datos pareados se pueden obtener fácilmente debido a que la retinografía también suele estar disponible cuando se obtiene una AF para un determinado paciente [2]. Para aprovechar mejor las imágenes pareadas, los pares de imágenes multimodales se alinean para establecer una correspondencia pixel a pixel entre las modalidades. Esto facilita el entrenamiento de una RNP en la RM al permitir el uso de métricas pixel a pixel como función de error.

Para el alineamiento de los pares de imágenes multimodales, propusimos una nueva metodología de registro multimodal en [52]. La metodología propuesta es un enfoque híbrido que combina métodos de registro basados en puntos de referencia y en patrones de intensidad. En la primera parte del método, se utilizan como puntos de referencia los cruces y las bifurcaciones del árbol vascular. La detección de estos puntos de referencia se realiza adaptando el método de Ortega et al. [53] al escenario multimodal. A continuación, el registro basado en puntos de referencia se realiza comparando los puntos correspondientes entre las imágenes y estimando una transformación rígida entre ellas. Para estimar la transformación más adecuada, descartando puntos atípicos, se utiliza un algoritmo RANSAC. En la segunda parte del método, se aplica a las imágenes una transformación laplaciana multiescala (LMS). Esta transformación convierte ambas modalidades de imagen a un espacio de imagen común en el que la microvasculatura de la retina está realzada. Esta representación común para ambas modalidades permite el uso directo de métricas de similitud entre las imágenes. En este caso concreto, utilizamos la correlación cruzada normalizada. El registro basado en la intensidad se realiza estimando la transformación espacial que maximiza la similitud entre ambas imágenes. En este caso, se utilizan tanto transformaciones rígidas como deformables.

La metodología para la RM usando imágenes multimodales pareadas y registradas fue presentada en [46]. Esta metodología se basa en el uso de una red neural convolucional estándar. En particular, adoptamos la arquitectura U-Net [54] que es comúnmente utilizada para el análisis de imagen médica. Para el entrenamiento de la red, exploramos diferentes funciones de error. En particular, consideramos las métricas L1 y L2, que han sido utilizadas previamente en varios problemas de características similares. Además de estas métricas, también exploramos el uso de la métrica de similitud estructural SSIM [55]. Esta es una métrica de similitud propuesta originalmente para la evaluación de la calidad de las imágenes. Esta métrica considera la intensidad, el contraste y las diferencias estructurales entre las imágenes. Para ello, SSIM se calcula utilizando un conjunto de estadísticas locales para cada posición de píxel. Por ejemplo, se incluye la media para medir la intensidad, la varianza para el contraste y la covarianza para la estructura de la imagen. Estas medidas proporcionan un análisis más completo de las diferencias entre las imágenes, lo que puede superar algunas de las limitaciones de L1 y L2.

La propuesta de MR se exploró y probó inicialmente utilizando un conjunto de datos público de 59 pares retinografía-AF. Posteriormente, se realizó un análisis más exhaustivo de la metodología utilizando un conjunto de datos ampliado que incluía 59 pares de imágenes adicionales facilitados por un hospital local. El conjunto de datos adicional incluye varios ejemplos de lesiones patológicas graves e imágenes de menor calidad, lo que permite comprobar la solidez de la metodología. Este análisis exhaustivo, tanto para el registro multimodal como para la RM, fue presentado en [56]. Adicionalmente, en este trabajo se evaluó el reconocimiento de la microvasculatura retiniana utilizando directamente la predicción de AF. Esta evaluación se realizo utilizando diferentes conjuntos de datos con etiquetas de la vasculatura a nivel de píxel.

Durante el desarrollo de esta tesis doctoral, diferentes autores han propuesto varios métodos para realizar transformaciones imagen-a-imagen. Estos métodos han estado típicamente centrados en el realismo de las imágenes generadas, dejando en un segundo plano la precisión estructural y semántica de los resultados. Por ello, es común el uso de redes generativas adversarias (RGAs) [57, 58], que hoy en día representan el enfoque de referencia para la generación de imágenes realistas [59]. Sin embargo, las RGAs también presentan el riesgo de alucinar estructuras inexistentes, lo que es más probable que ocurra cuando los patrones de imagen en el conjunto de datos de entrenamiento están muy desequilibrados [60]. Sin embargo, una ventaja importante de algunos enfoques basados en RGAs es que permiten el aprendizaje de una transformación imagen-a-imagen sin la necesidad de datos de entrenamiento pareados [61]. Esto es clave en muchos dominios de aplicación con imágenes naturales porque las muestras pareadas son difíciles de obtener. En cambio, en imagen médica, las colecciones de imágenes pareadas son más fáciles de obtener debido al uso común de modalidades complementarias en la práctica clínica. En cualquier caso, aprovechar completamente los datos pareados también requiere realizar con éxito un registro multimodal de las imágenes, que puede fallar en los escenarios más complejos, ya sea por la presencia de patologías graves o por la baja calidad de las imágenes. Por estas razones, también exploramos el uso de metodos no pareados basados en RGAs para la RM de imágenes de la retina.

En cuanto al uso de métodos no pareados basados en RGAs para la RM de imágenes de la retina, presentamos un estudio completo comparando metodologías pareadas y no pareadas en [62, 63]. En este caso, para la metodología no pareada, adoptamos el método CycleGAN [61]. Los resultados muestran que la alternativa no pareada produce muestras generadas más realistas y, además, también mejora el reconocimiento de algunas lesiones pequeñas en las imágenes. Sin embargo, hay algunas imprecisiones estructurales entre las imágenes de entrada y las generadas. Por lo tanto, aunque el uso de RGAs puede proporcionar algunas ventajas, el uso de datos no pareados no es suficiente para garantizar la consistencia estructural y semántica de las imágenes generadas.

B.3.2 Parte II - Análisis de Estructuras de la Retina

El análisis de las diferentes estructuras anatómicas de la retina juega un papel destacado en el diagnóstico y seguimiento de numerosas enfermedades [64]. Por ejemplo, las lesiones patológicas pueden aparecer alrededor de ciertas regiones anatómicas que deben ser identificadas adecuadamente para proporcionar un diagnóstico. Además, algunos trastornos oculares producen directamente cambios morfológicos en la anatomía de la retina. En estos casos, es conveniente detectar y caracterizar las estructuras retinianas afectadas para valorar los efectos de la enfermedad [10, 9, 11].

A grandes rasgos, las principales estructuras anatómicas de la retina son la microvasculatura, el disco óptico y la fóvea [8]. La microvasculatura retiniana está implicada en el estudio de varias enfermedades oftálmicas y sistémicas. En este sentido, la retina es el único órgano del cuerpo humano que permite el estudio del sistema vascular in vivo y sin procedimientos invasivos [65]. La principal tarea relativa al análisis de la vasculatura es la segmentación de los vasos sanguíneos. Hoy en día, esta tarea puede resolverse fácilmente utilizando las modernas RNPs. Sin embargo, la segmentación de los vasos más pequeños en las imágenes sigue siendo un reto. Además, el etiquetado manual de la microvasculatura es una tarea especialmente tediosa debido al elevado número de vasos pequeños y al bajo contraste en algunas regiones de las imágenes. Además de la importancia de la microvasculatura para fines de diagnóstico, el árbol vascular de la retina también se suele aprovechar para otras aplicaciones. Por ejemplo, los cruces y bifurcaciones de los vasos sanguíneos en la retina se utilizan habitualmente como puntos de referencia para los algoritmos de registro de imágenes o para los métodos de verificación de identidad [52, 53].

En cuanto al disco óptico, esta estructura retiniana es especialmente importante para el diagnóstico de glaucoma. De hecho, un biomarcador ampliamente extendido para la evaluación del glaucoma, la relación copa-disco, puede obtenerse únicamente a partir del análisis morfológico del disco óptico y sus componentes internos [66]. En particular, el disco óptico puede dividirse en dos subregiones diferentes, la copa óptica y el borde neurorretiniano. En la literatura, numerosos trabajos han abordado la segmentación automatizada de estas dos regiones, con el objetivo de facilitar el diagnóstico del glaucoma mediante el uso de biomarcadores morfológicos [66]. Además, la localización o segmentación del disco óptico también se utiliza comúnmente como un procedimiento intermedio dentro de metodologías más complejas con fines de diagnóstico o para el análisis de otras estructuras de la retina [67]. Del mismo modo, la localización de la fóvea también se utiliza comúnmente como parte de metodologías más complejas. En particular, la identificación de la región foveal (o macular) es de gran interés para el diagnóstico de varias enfermedades que conducen al desarrollo de diferentes lesiones en esa zona, como por ejemplo DMAE o edema macular diabético [10].

En el contexto de esta tesis doctoral, se utilizó la localización y segmentación de las estructuras anatómicas de la retina para demostrar las ventajas de la RM propuesta en aprendizaje por transferencia [47]. En este sentido, para realizar con éxito la RM, una RNP debe aprender diferentes patrones retinianos de bajo y alto nivel. Así, utilizando la RM como tarea de pre-entrenamiento, este conocimiento específico del dominio puede ser aprovechado para diferentes tareas objetivo centradas en el análisis de la anatomía de la retina. Exploramos esta idea en [47], donde la RM se utilizó como tarea de pre-entrenamiento auto-supervisada para la segmentación de los vasos sanguíneos, la detección de la fóvea, y la segmentación y detección del disco óptico.

La metodología presentada en [47] se basa en una arquitectura U-Net [54], que es un algoritmo de referencia para la segmentación de los vasos sanguíneos y la localización de la fóvea [68, 69]. Todas las tareas se entrenaron siguiendo el mismo procedimiento, incluyendo la arquitectura de red, el aumento de datos y los hiperparámetros de optimización. La única diferencia entre las tareas es la formulación del objetivo de entrenamiento y la función de error. La segmentación de los vasos sanguíneos y del disco óptico se realiza como una clasificación binaria a nivel de pixel, utilizando la entropía cruzada como función de error [70]. En cuanto a la localización del disco óptico y la fóvea, realizamos una regresión de un mapa de distancia al punto objetivo, donde el valor de cada píxel depende de la distancia a la ubicación del objetivo [69]. En concreto, el mapa se construye calculando las distancias euclideas y después aplicando una función tangente hiperbólica a los valores obtenidos. Esto da como resultado un mapa de distancias con una pendiente más pronunciado cerca de la ubicación del objetivo y que se aplana en las regiones más lejanas. Para evaluar el método de aprendizaje por transferencia propuesto, realizamos experimentos utilizando diferentes cantidades de datos de entrenamiento etiquetados, que van desde una sola muestra de entrenamiento hasta todo el conjunto de entrenamiento. Los resultados obtenidos demuestran que el pre-entrenamiento de RM contribuye a las diferentes tareas, mejorando significativamente los resultados cuando los datos etiquetados disponibles para el entrenamiento son escasos.

Además de los experimentos mencionados anteriormente, en [71], también probamos el uso del pre-entrenamiento de RM para la segmentación del disco óptico y la copa óptica. Para abordar esta tarea, hemos seguido una metodología similar a la utilizada para la segmentación de vasos sanguíneos y disco óptico en [47]. La principal diferencia es que, en este caso, la segmentación se aborda como una clasificación multiclase a nivel de píxel. Concretamente, se consideran tres clases, la copa óptica, el borde neurorretiniano y el fondo. Después, el disco óptico se define como la suma de la copa y el borde. En este caso, los resultados experimentales también muestran que el pre-entrenamiento en RM mejora los resultados de la tarea de segmentación tanto para el disco óptico como para la copa óptica.

En cuanto a la microvasculatura de la retina, también exploramos nuevas alternativas para segmentar los vasos sanguíneos utilizando RNPs sin datos etiquetados. En este sentido, propusimos en [72] un nuevo método para la segmentación autosupervisada de los vasos de la retina que está motivado por dos desarrollos anteriores. En primer lugar, en [52], propusimos una transformación LMS que realza significativamente la microvasculatura retiniana tanto para la retinografía como para la AF. En este caso, se obtiene un mejor mapa vascular para la FA debido al contraste inyectado que ya realza la vasculatura en esta modalidad. En segundo lugar, en [46], proponemos el método de RM que genera y estima la AF a partir de retinografía, resaltando así los vasos sanguíneos en las imágenes. Finalmente, en [72], combinamos estos dos métodos para mejorar aún más el realce de la microvasculatura retiniana en las imágenes. En concreto, la metodología consiste en entrenar una RNP en la predicción del LMS de la AF utilizando la retinografía como entrada. De este modo, la red aprende a producir una representación altamente mejorada de los vasos sanguíneos directamente a partir de la retinografía y sin utilizar ninguna etiqueta anotada manualmente.

Por último, en relación con el análisis de la anatomía de la retina, también exploramos la detección de los cruces y bifurcaciones de los vasos vasculares de la retina [73]. En este caso, los métodos anteriores en la literatura normalmente se basaban en un extenso procesamiento ad-hoc, incluso cuando se utilizaban RNPs. Además, estos métodos previos normalmente separaban el problema en dos tareas diferentes, la detección de los puntos de interés en los vasos y su posterior clasificación entre cruces y bifurcaciones [74]. En este contexto, en [73], propusimos un método para detectar e identificar simultáneamente los cruces y bifurcaciones en un solo paso utilizando RNPs. En particular, la tarea de detección se formuló como una regresión de mapa de distancias que combina multiples puntos de interés en el mismo mapa. La ubicación precisa de cada cruce o bifurcación viene dada por los máximos locales en el mapa. Para proporcionar una heurística adecuada para el aprendizaje de la regresión del mapa, los valores del mapa se reducen progresivamente en los píxeles vecinos a cada punto de interés. Además, exploramos dos alternativas diferentes para generar los mapas objetivo, utilizando un kernel convolucional Gaussiano o uno de tangente hiperbólica radial (Radial Tanh). La diferenciación entre cruces y bifurcaciones se realiza mediante la predicción simultánea de dos mapas independientes. Los resultados experimentales muestran que tanto el kernel Gaussiano como el Radial Tanh proporcionan resultados similares cuando se ajusta adecuadamente la escala del kernel. Sin embargo, la alternativa Radial Tanh es más robusta a estos cambios, proporcionando un rendimiento más estable. Además, el método propuesto supera significativamente a los métodos anteriores tanto en la detección como en la identificación de los cruces de vasos y bifurcaciones.

B.3.3 Parte III - Diagnóstico Asistido por Ordenador

El aprendizaje profundo representa una herramienta fundamental para los sistemas DAO modernos. En este sentido, las RNPs han mejorado significativamente los resultados que se podían conseguir con métodos tradicionales para el diagnóstico de numerosas enfermedades [75]. Por ejemplo, en oftalmología, se han aplicado con éxito métodos basados en deep learning para el diagnóstico de DMAE, glaucoma o retinopatía diabética entre otras enfermedades [75]. Sin embargo, el éxito de estos enfoques está fuertemente ligado a la disponibilidad de grandes conjuntos de datos etiquetados para el entrenamiento de las RNPs [18]. En este contexto, durante esta tesis doctoral, presentamos una nueva metodología de aprendizaje por transferencia para sistemas DAO utilizando la RM previamente propuesta [76]. La idea es aprovechar el conocimiento específico del dominio que una RNP adquiere de los datos multimodales no etiquetados durante el entrenamiento de la RM. Sin embargo, en este caso, la aplicación final es el diagnóstico de varias enfermedades de la retina, es decir, diferentes tareas de clasificación de imagen. Este tipo de aplicación presenta diferentes requisitos técnicos, como por ejemplo la arquitectura de red, que hacen necesaria una metodología de aprendizaje por transferencia diferente.

En esta tesis doctoral, el enfoque de aprendizaje por transferencia propuesto para CAD se aplica al diagnóstico de DMAE y glaucoma. Estos son dos importantes trastornos oculares que afectan a diferentes zonas de la retina y que provocan una importante pérdida de visión si no se tratan. En concreto, la DMAE es un trastorno ocular degenerativo que afecta a la mácula, que representa la zona que rodea a la fóvea en la retina. Esta enfermedad se caracteriza por la presencia de diferentes estructuras o lesiones patológicas en esta zona, como drusas, exudados o anomalías epiteliales entre otras. Por ello, el diagnóstico se suele realizar analizando el fondo de ojo en busca de estas estructuras patológicas [10]. En cambio, el glaucoma se caracteriza por un aumento de la presión intraocular que produce daños en diferentes tejidos y estructuras retinianas, como la cabeza del nervio óptico. En este sentido, el glaucoma se puede diagnosticar analizando las imágenes del fondo de ojo en busca de cambios morfológicos en el disco óptico, como la reducción del borde neurorretiniano y el aumento de la copa óptica [9].

La metodología de aprendizaje de transferencia para el CAD de retina fue presentada en [76]. La metodología propuesta se adapta a cada enfermedad centrando el análisis en la región de interés (RDI) que se requiere según los criterios clínicos. En particular, se recorta una RDI cuadrada alrededor de la fóvea y el disco óptico para el diagnóstico de DMAE y glaucoma, respectivamente. La detección de la fóvea y el disco óptico se realiza automáticamente siguiendo el enfoque que propusimos anteriormente en [47]. Las RDIs extraídas se utilizan para la tarea objetivo de clasificación de imagen, así como para el pre-entrenamiento de la RM utilizando datos multimodales no etiquetados. Al igual que en [47], el pre-entrenamiento de RM se realiza utilizando una arquitectura de red U-Net. Sin embargo, para la clasificación de imágenes, el diseño típico de la red consiste principalmente en un encoder convolucional seguido de algunas capas de neuronas totalmente conectadas para realizar la predicción final. Por tanto, en este caso, sólo se reutiliza el encoder de la red pre-entrenada para las tareas objetivo de clasificación. Una cuestión adicional que debe considerarse, en relación con la arquitectura de red, es el efecto de las conexiones encoder-decoder de U-Net en el método de aprendizaje por transferencia propuesto. En este sentido, aunque las conexiones encoder-decoder facilitan el entrenamiento de la red, también hacen posible que cierta información relevante nunca llegue a las últimas capas del encoder. En la metodología propuesta, en la que sólo se reutiliza el encoder de la red pre-entrenada para las tareas objetivo, esto podría tener un efecto perjudicial en los resultados de aprendizaje por transferencia. Esta cuestión ha sido estudiada en [76]. Los resultados obtenidos muestran que, en algunos casos, el uso de todas las conexiones encoder-decoder puede comprometer el rendimiento de la tarea objetivo. Sin embargo, la eliminación de todas las conexiones también presenta un efecto perjudicial debido a la dificultad para realizar con éxito el pre-entrenamiento de la RM. Por lo tanto, los resultados más sólidos se consiguen siguiendo un enfoque intermedio. Por último, la metodología propuesta se validó comparando su rendimiento con el entrenamiento de la red desde cero y con un preentrenamiento en ImageNet. Los resultados muestran que la propuesta tiene un impacto positivo en el rendimiento de las diferentes tareas en el contexto de DAO para imágenes de la retina.

B.4 Conclusiones Generales

El análisis de las imágenes de fondo de ojo, como la retinografía, es un paso clave en la prevención, el diagnóstico y el seguimiento de numerosos trastornos oculares. En los últimos años, existe un creciente interés en el desarrollo de herramientas automáticas para el análisis de estas imágenes. Estas herramientas automáticas ayudan a los clínicos a proporcionar diagnósticos más fiables y facilitan la realización de programas preventivos de salud.

En esta tesis doctoral, hemos presentado varios desarrollos metodológicos para mejorar el análisis automático de imágenes de fondo de ojo utilizando técnicas de aprendizaje profundo. Las RNPs han demostrado ofrecer un rendimiento notable en numerosos problemas de visión y representan el enfoque predilecto para el análisis automatizado de imágenes médicas. En este contexto, la falta de datos de entrenamiento etiquetados representa una de las principales limitaciones para la aplicación exitosa de métodos basados en aprendizaje profundo en imagen médica. Teniendo esto en cuenta, hemos propuesto un nuevo paradigma para el entrenamiento de RNPs de forma auto-supervisada utilizando datos visuales multimodales sin etiquetar. Esta propuesta aprovecha los pares de imágenes multimodales que están comúnmente disponibles en oftalmología. El método presentado permite la predicción de imágenes de AF a partir de la retinografía y puede ser utilizado como preentrenamiento para cualquier tarea de objetivo realizada en la retinografía.

Para aprovechar los datos emparejados multimodales, primero desarrollamos una nueva metodología para el registro multimodal de imágenes de la retina. En particular, presentamos un enfoque híbrido que consiste en etapas de registro basadas en puntos de referencia y en patrones de intensidad. Esta metodología permite la construcción de conjuntos de datos multimodales con imágenes pareadas y alineadas, que posteriormente se utilizan para el entrenamiento de las RNPs en la RM. Además, también exploramos el uso de datos multimodales no pareados para realizar la RM.

Nuestros experimentos demostraron que el uso de datos pareados y alineados resulta ventajoso.

Teniendo en cuenta los resultados anteriores, exploramos el uso de la RM como pre-entrenamiento para diferentes tareas en retinografía. En primer lugar, abordamos la segmentación y localización de diferentes estructuras anatómicas en la retina, que son un paso inicial común en numerosos procedimientos de análisis de imágenes de la retina. En particular, nuestros experimentos se centraron en la microvasculatura retiniana, la fóvea y el disco óptico, que representan las principales estructuras o regiones anatómicas en las imágenes del fondo del ojo. Esta experimentación muestra que el enfoque de aprendizaje por transferencia propuesto reduce la cantidad de datos etiquetados que se necesitan para lograr resultados satisfactorios en todas las tareas. Este es un importante resultado que indica que la propuesta puede facilitar la aplicación de algoritmos de aprendizaje profundo para nuevos problemas con datos etiquetados limitados. Además, el mismo enfoque de aprendizaje por transferencia también ha demostrado ser ventajoso para la segmentación del disco óptico y la copa óptica, lo que es útil para el diagnóstico de glaucoma.

En cuanto al uso de la RM como pre-entrenamiento para tareas de clasificación, propusimos una metodología de aprendizaje de transferencia para sistemas de DAO. En particular, abordamos el diagnóstico de dos importantes trastornos oculares como son DMAE y glaucoma. El diagnóstico de estas dos enfermedades requiere tipos de análisis muy diferentes, por lo que proporcionan escenarios complementarios para una evaluación robusta de nuestra propuesta. Los resultados muestran que el método de aprendizaje por transferencia propuesto, utilizando pares de imágenes multimodales sin etiquetar, es ventajoso para el diagnóstico de estas enfermedades. Además, en general, el método proporciona un rendimiento más robusto que otras alternativas, como el pre-entrenamiento totalmente supervisado en el conjunto de datos ImageNet.

Para proporcionar una entendimiento más completo del fondo de ojo, también abordamos la detección e identificación de los cruces y bifurcaciones en los vasos sanguíneos. En este caso, propusimos una metodología que permite aprovechar mejor las ventajas de las RNPs para el análisis de imágenes. En este sentido, además de superar significativamente a métodos anteriores, nuestra propuesta proporciona un procedimiento más sencillo que evita cualquier procesamiento ad-hoc de los datos.

Por último, en lo que respecta a la anatomía de la retina y, en particular, a la microvasculatura retiniana, también propusimos un nuevo método para la segmentación de los vasos sanguíneos utilizando etiquetas generadas automáticamente. Este método aprovecha otros desarrollos realizados durante esta tesis doctoral, así como la disponibilidad de pares de imágenes retinográficas-AF no etiquetadas.

En resumen, en esta tesis doctoral hemos propuesto diferentes metodologías para realizar un análisis completo del fondo de ojo y reducir la necesidad de conjuntos masivos de datos etiquetados para el entrenamiento de las RNPs. En este sentido, dado el éxito de los enfoques de aprendizaje por transferencia propuestos utilizando datos multimodales no etiquetados, en futuros trabajos consideramos extender esta idea a aplicaciones adicionales. Por ejemplo, sería interesante explorar este tipo de técnicas multimodales auto-supervisadas para la detección y caracterización de diferentes lesiones o el diagnóstico de otras enfermedades como la retinopatía diabética. Estos trabajos pueden ir acompañados de desarrollos técnicos adicionales para mejorar aún más el paradigma propuesto. Además, también consideramos extender el paradigma propuesto a otras áreas médicas en las que la imagen multimodal es habitual. En este caso, también sería posible aprovechar los datos visuales 3D que son comunes en otras áreas médicas. Otra futura dirección de investigación que consideramos es explorar diferentes paradigmas de aprendizaje por transferencia, por ejemplo, aplicando aprendizaje multitarea. A diferencia del enfoque de pre-entrenamiento y refinamiento, el aprendizaje multitarea permite que ambas tareas saquen ventaja de los datos etiquetados disponibles para cada una. Así, en este caso, la necesidad de conjuntos masivos de datos etiquetados también podría reducirse combinando diferentes tareas supervisadas con objetivos complementarios.

B.5 Estructura de la Tesis

Esta tesis está estructurada por capítulos según se indica a continuación. El capítulo I presenta una breve introducción a la tesis doctoral. En primer lugar, este capítulo proporciona la motivación y el contexto para el trabajo de investigación. A continuación, se describen claramente los principales objetivos de la tesis doctoral. Por último, se presenta una breve discusión sobre el trabajo de investigación de esta tesis doctoral. Esta discusión pretende dar consistencia y coherencia a las diferentes publicaciones que componen esta tesis. El capítulo 2 incluye la descripción detallada de las metodologías y la experimentación para la RM de imágenes de la retina utilizando datos multimodales no etiquetados. El capítulo 3 incluye la descripción detallada de las metodologías y la experimentación para el análisis de las estructuras anatómicas en la retina. El capítulo 4 presenta la metodología de aprendizaje de transferencia propuesta para los sistemas CAD de la retina, incluyendo la experimentación y el análisis de los resultados obtenidos.

Bibliography

[1] *Medical imaging : technology and applications.* Devices, circuits, and systems, CRC Press/Taylor & Francis Group, 2014 - 2014.

[2] Cole, E. D., Novais, E. A., Louzada, R. N., and Waheed, N. K., "Contemporary retinal imaging techniques in diabetic retinopathy: a review," *Clinical & Experimental Ophthalmology*, vol. 44, no. 4, pp. 289–299, 2016.

[3] Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., and Webster, D. R., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[4] Niemeijer, M., van Ginneken, B., Cree, M. J., Mizutani, A., Quellec, G., Sanchez, C. I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., and Abramoff, M. D., "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 185–195, 2010.

[5] Vujosevic, S., Aldington, S. J., Silva, P., Hernández, C., Scanlon, P., Peto, T., and Simó, R., "Screening for diabetic retinopathy: new perspectives and challenges," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 4, pp. 337–347, 2020.

[6] Ting, D. S., Peng, L., Varadarajan, A. V., Keane, P. A., Burlina, P. M., Chiang, M. F., Schmetterer, L., Pasquale, L. R., Bressler, N. M., Webster, D. R., Abramoff, M., and Wong, T. Y., "Deep learning in ophthalmology: The technical and clinical considerations," *Progress in Retinal and Eye Research*, vol. 72, p. 100759, 2019.

[7] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., and Pattichis, C. S., "Ai in medical imaging informatics: Current challenges and future directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[8] Abràmoff, M. D., Garvin, M. K., and Sonka, M., "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[9] Weinreb, R. N., Aung, T., and Medeiros, F. A., "The pathophysiology and treatment of glaucoma: a review.," *JAMA*, vol. 311, no. 18, pp. 1901–11, 2014.

[10] AREDS Research Group, "The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the age-related eye disease study report number 6," *American Journal of Ophthalmology*, vol. 132, no. 5, pp. 668–681, 2001.

[11] Heng, L. Z., Comyn, O., Peto, T., Tadros, C., Ng, E., Sivaprasad, S., and Hykin, P. G., "Diabetic retinopathy: pathogenesis, clinical grading, management and future developments," *Diabetic Medicine*, vol. 30, pp. 640–650, 2013.

[12] Wong, T. Y., Klein, R., Klein, B. E., Tielsch, J. M., Hubbard, L., and Nieto, F., "Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality," *Survey of Ophthalmology*, vol. 46, no. 1, pp. 59–80, 2001.

[13] Lim, G., Bellemo, V., Xie, Y., Lee, X. Q., Yip, M. Y. T., and Ting, D. S. W., "Different fundus imaging modalities and technical factors in ai screening for diabetic retinopathy: a review," *Eye and Vision*, vol. 7, 2020.

[14] Sengupta, S., Sindal, M. D., Besirli, C. G., Upadhyaya, S., Venkatesh, R., Niziol, L. M., Robin, A. L., Woodward, M. A., and Newman-Casey, P. A., "Screening for vision-threatening diabetic retinopathy in south india: comparing portable non-mydriatic and standard fundus cameras and clinical exam," *Eye*, p. 375–383, 2018.

[15] Stolte, S. and Fang, R., "A survey on medical image analysis in diabetic retinopathy," *Medical Image Analysis*, vol. 64, p. 101742, 2020.

[16] Sengupta, S., Singh, A., Leopold, H. A., Gulati, T., and Lakshminarayanan, V., "Ophthalmic diagnosis using deep learning with fundus images – a critical review," *Artificial Intelligence in Medicine*, vol. 102, p. 101758, 2020.

[17] Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., and Fu, H., "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, vol. 69, p. 101971, 2021.

[18] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[19] Rawat, W. and Wang, Z., "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," *Neural Computation*, vol. 29, pp. 2352–2449, 09 2017.

[20] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., and Garcia-Rodriguez, J., "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[21] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning.* Adaptive computation and machine learning, MIT Press, 2016.

[22] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S., "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27 – 48, 2016.

[23] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds.), vol. 25, Curran Associates, Inc., 2012.

[24] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F., "Imagenet: A large-scale hierarchical image database.," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[25] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results." http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[26] Weese, J. and Lorenz, C., "Four challenges in medical image analysis from an industrial perspective," *Medical Image Analysis*, vol. 33, pp. 44–49, 2016. 20th anniversary of the Medical Image Analysis journal (MedIA).

[27] Cheplygina, V., de Bruijne, M., and Pluim, J. P., “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[28] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., and Kang, H., “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” *Information Sciences*, vol. 501, pp. 511 – 522, 2019.

[29] Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y. H., Lee, Y. C., Jung, S.-H., Li, Z., Sui, X., Wu, J., Li, X., Zhou, T., Toth, J., Baran, A., Kori, A., Chennamsetty, S. S., Safwan, M., Alex, V., Lyu, X., Cheng, L., Chu, Q., Li, P., Ji, X., Zhang, S., Shen, Y., Dai, L., Saha, O., Sathish, R., Melo, T., Araújo, T., Harangi, B., Sheng, B., Fang, R., Sheet, D., Hajdu, A., Zheng, Y., Mendonça, A. M., Zhang, S., Campilho, A., Zheng, B., Shen, D., Giancardo, L., Quellec, G., and Mériaudeau, F., “Idrid: Diabetic retinopathy – segmentation and grading challenge,” *Medical Image Analysis*, vol. 59, p. 101561, 2020.

[30] Bloice, M. D., Roth, P. M., and Holzinger, A., “Biomedical image augmentation using Augmentor,” *Bioinformatics*, vol. 35, pp. 4522–4524, 04 2019.

[31] Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., and Campilho, A., “End-to-end adversarial retinal image synthesis,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2018.

[32] Cohen, J., Luck, M., and Honari, S., “Distribution matching losses can hallucinate features in medical image translation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 09 2018.

[33] Morid, M. A., Borjali, A., and Del Fiol, G., “A scoping review of transfer learning research on medical image analysis using imagenet,” *Computers in Biology and Medicine*, vol. 128, p. 104115, 2021.

[34] Caruana, R., “Multitask learning,” *Machine Learning*, 1997.

[35] Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., and Van Gool, L., “Multi-task learning for dense prediction tasks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[36] Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M., "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[37] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[38] Jing, L. and Tian, Y., "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[39] Ohri, K. and Kumar, M., "Review on self-supervised image recognition using deep neural networks," *Knowledge-Based Systems*, vol. 224, p. 107090, 2021.

[40] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J., "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge & Data Engineering*, pp. 1–1, jun 2021.

[41] Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A., "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[42] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D., "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, p. 101539, 2019.

[43] Oh, C., Ham, B., Kim, H., Hilton, A., and Sohn, K., "Ocean: Object-centric arranging network for self-supervised visual representations learning," *Expert Systems with Applications*, vol. 125, pp. 281 – 292, 2019.

[44] Wei, D., Lim, J. J., Zisserman, A., and Freeman, W. T., "Learning and using the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., and Zhuang, Y., "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Hervella, A. S., Rouco, J., Novo, J., and Ortega, M., "Retinal image understanding emerges from self-supervised multimodal reconstruction," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[47] Hervella, A. S., Rouco, J., Novo, J., and Ortega, M., "Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction," *Applied Soft Computing*, p. 106210, 2020.

[48] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M., "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., eds.), vol. 33, pp. 21271–21284, Curran Associates, Inc., 2020.

[49] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E., "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., eds.), vol. 33, pp. 22243–22255, Curran Associates, Inc., 2020.

[50] Alipour, S. H. M., Rabbani, H., and Akhlaghi, M. R., "Diabetic retinopathy grading by digital curvelet transform," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[51] Hua, C. H., Kim, K., Huynh-The, T., You, J. I., Yu, S. Y., Le-Tien, T., Bae, S. H., and Lee, S., "Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[52] Hervella, A. S., Rouco, J., Novo, J., and Ortega, M., "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," *Procedia Computer Science*, vol. 126, pp. 97 – 104, 2018.

[53] Ortega, M., Penedo, M. G., Rouco, J., Barreira, N., and Carreira, M. J., "Retinal verification using a feature points-based biometric pattern," *EURASIP Advances in Signal Processing*, vol. 2009, Mar 2009.

[54] Ronneberger, O., P.Fischer, and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[55] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[56] Hervella, A. S., Rouco, J., Novo, J., and Ortega, M., “Self-supervised multimodal reconstruction of retinal images over paired datasets,” *Expert Systems with Applications*, p. 113674, 2020.

[57] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems (NIPS) 27*, pp. 2672–2680, 2014.

[58] Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., and Zheng, Y., “Recent progress on generative adversarial networks (gans): A survey,” *IEEE Access*, vol. 7, pp. 36322–36333, 2019.

[59] Wang, L., Chen, W., Yang, W., Bi, F., and Yu, F. R., “A state-of-the-art review on image synthesis with generative adversarial networks,” *IEEE Access*, vol. 8, pp. 63514–63537, 2020.

[60] Wolterink, J. M., Dinkla, A. M., Savenije, M. H. F., Seevinck, P. R., van den Berg, C. A. T., and Išgum, I., “Deep MR to CT Synthesis Using Unpaired Data,” in *Simulation and Synthesis in Medical Imaging* (Tsaftaris, S. A., Gooya, A., Frangi, A. F., and Prince, J. L., eds.), (Cham), pp. 14–23, Springer International Publishing, 2017.

[61] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference onComputer Vision (ICCV)*, 2017.

[62] Hervella, Á. S., Rouco, J., Novo, J., and Ortega, M., “Deep multimodal reconstruction of retinal images using paired or unpaired data,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019.

[63] Álvaro S. Hervella, Rouco, J., Novo, J., and Ortega, M., “Chapter 15 - multimodal reconstruction of retinal images over unpaired datasets using cyclical generative adversarial networks,” in *Generative Adversarial Networks for Image-to-Image Translation* (Solanki, A., Nayyar, A., and Naved, M., eds.), pp. 347–376, Academic Press, 2021.

[64] Besenczi, R., Tóth, J., and Hajdu, A., "A review on automatic analysis techniques for color fundus photographs," *Computational and Structural Biotechnology Journal*, vol. 14, pp. 371–384, 2016.

[65] Patton, N., Aslam, T. M., MacGillivray, T., Deary, I. J., Dhillon, B., Eikelboom, R. H., Yogesan, K., and Constable, I. J., "Retinal image analysis: Concepts, applications and potential," *Progress in Retinal and Eye Research*, vol. 25, no. 1, pp. 99–127, 2006.

[66] Sarhan, A., Rokne, J., and Alhajj, R., "Glaucoma detection using image processing techniques: A literature review," *Computerized Medical Imaging and Graphics*, vol. 78, p. 101657, 2019.

[67] Niemeijer, M., Abràmoff, M. D., and van Ginneken, B., "Fast detection of the optic disc and fovea in color fundus photographs," *Medical Image Analysis*, vol. 13, no. 6, pp. 859–870, 2009. Includes Special Section on Computational Biomechanics for Medicine.

[68] Gao, X., Cai, Y., Qiu, C., and Cui, Y., "Retinal blood vessel segmentation based on the gaussian matched filter and u-net," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, 2017.

[69] Meyer, M. I., Galdran, A., Mendonça, A. M., and Campilho, A., "A pixel-wise distance regression approach for joint retinal optical disc and fovea detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., eds.), (Cham), pp. 39–47, Springer International Publishing, 2018.

[70] Mookiah, M. R. K., Hogg, S., MacGillivray, T. J., Prathiba, V., Pradeepa, R., Mohan, V., Anjana, R. M., Doney, A. S., Palmer, C. N., and Trucco, E., "A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification," *Medical Image Analysis*, vol. 68, p. 101905, 2021.

[71] Hervella, A. S., Ramos, L., Rouco, J., Novo, J., and Ortega, M., "Multi-modal self-supervised pre-training for joint optic disc and cup segmentation in eye fundus images," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[72] Hervella, A. S., Rouco, J., Novo, J., and Ortega, M., "Self-supervised deep learning for retinal vessel segmentation using automatically generated labels

from multimodal data," in *International Joint Conference on Neural Networks (IJCNN)*, 2019.

[73] "Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images," *Computer Methods and Programs in Biomedicine*, vol. 186, p. 105201, 2020.

[74] Pratt, H., Williams, B. M., Ku, J. Y., Vas, C., McCann, E., Al-Bander, B., Zhao, Y., Coenen, F., and Zheng, Y., "Automatic detection and distinction of retinal vessel bifurcations and crossings in colour fundus photography," *Journal of Imaging*, vol. 4, no. 1, 2018.

[75] Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., and Fu, H., "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, vol. 69, p. 101971, 2021.

[76] Álvaro S. Hervella, Rouco, J., Novo, J., and Ortega, M., "Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis," *Expert Systems with Applications*, vol. 185, p. 115598, 2021.

[77] CORE - Computing Research & Education, "portal.core.edu.au." `http://portal.core.edu.au/conf-ranks/`, 2019. Accessed: 2010-09-30.

[78] JCR - Journal Citation Reports, "jcr.incites.thomsonreuters.com." `https://jcr.incites.thomsonreuters.com`, 2019. Accessed: 2010-09-30.