Expert Systems with Applications 161 (2020) 113674

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

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

ARTICLE INFO

Article history:
Received 13 January 2020
Revised 27 April 2020
Accepted 16 June 2020
Available online 26 June 2020

Keywords:
Self-supervised learning
Eye fundus
Deep learning
Multimodal
Retinography
Angiography

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.

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, Van Gool, Williams, Winn, & Zisserman, 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 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

* Corresponding author at: CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain.
E-mail addresses: a.suarezh@udc.es (Á.S. Hervella), jrouco@udc.es (J. Rouco), jnovo@udc.es (J. Novo), mortega@udc.es (M. Ortega).

https://doi.org/10.1016/j.eswa.2020.113674
0957-4174/

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, Kadir, & Zisserman, 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, Orton, Collins, Doran, & Leach, 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, Valpola, Honkala, Berglund, & Raiko, 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, Courville, & Vincent, 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, Bilen, Gavves, & Gould, 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, Kreiman, & Cox, 2017). Although simpler classification tasks, detecting video sequences with shuffled frames (Misra, Zitnick, & Hebert, 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, Krähenbühl, Donahue, Darrell, & Efros, 2016), while in others, simpler tasks consisting in the prediction of relative patch positions (Doersch, Gupta, & Efros, 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, Isola, and Efros (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, Wu, McDermott, Freeman, & Torralba, 2016), depth (Wang, Wang, Wu, You, & Neumann, 2017), or motion information (Agrawal, Carreira, & Malik, 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.

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 self-supervised reconstruction of angiography from retinography is summarized in the diagram of Fig. 1. The multimodal reconstruction is performed using a U-Net fully convolutional neural network (Ronneberger, Fischer, & Brox, 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, Rabbani, & Akhlaghi, 2012) and from an additional private dataset. The alignment of the images is performed using the multimodal retinography–angiography registration algorithm proposed by Hervella, Rouco, Novo, and Ortega (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, Abramoff, Niemeijer, Viergever, & van Ginneken, 2004) and STARE (Hoover, Kouznetsova, & Goldbaum, 2000). Preliminary results of this work have been presented in Hervella, Rouco, Novo, and Ortega (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, Penedo, Rouco, Barreira, & Carreira, 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

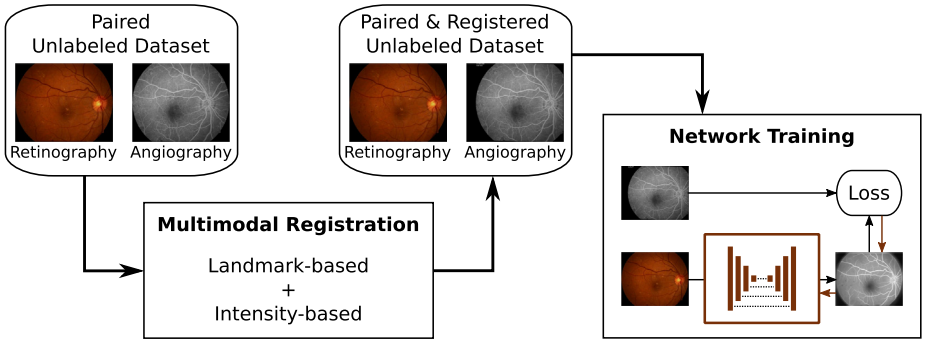


Fig. 1. Proposed self-supervised approach using unlabeled multimodal data. First, the paired multimodal dataset is registered. The resulting registered dataset is used to train a DNN in the multimodal reconstruction of angiography from retinography.

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 Fig. 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 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:

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

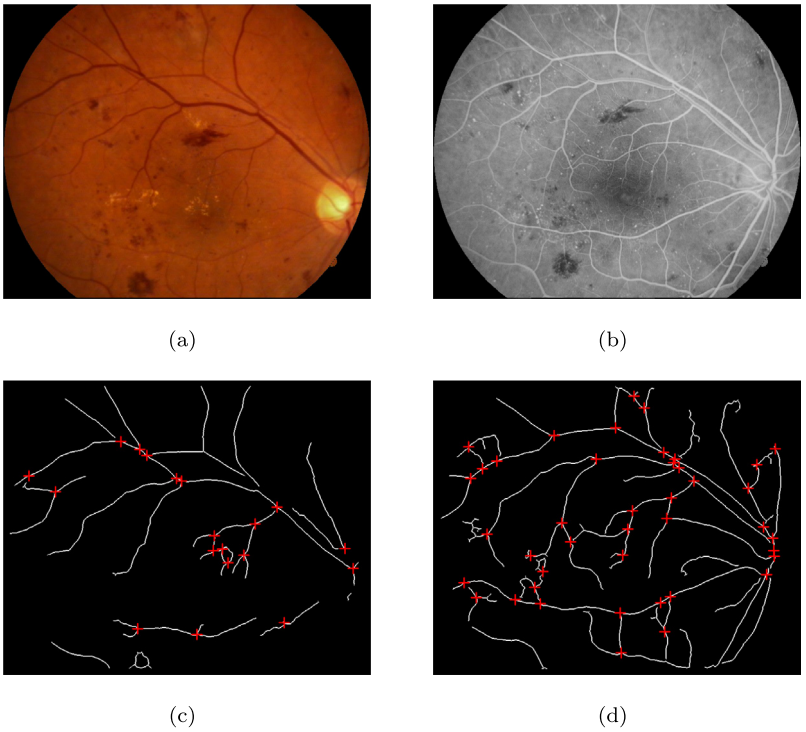


Fig. 2. Example of vessel tree and detected landmarks for a retinography–angiography pair from a diabetic retinopathy patient. (a) Retinography. (b) Angiography. (c) Vessel tree and landmarks from (a). (d) Vessel tree and landmarks from (b).

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:

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

where $\lceil \cdot \rceil_{\varnothing}$ 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. Fig. 3 depicts examples of multiscale Laplacian maps for the retinography and angiography in Fig. 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:

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

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 $\mathbf{a} \in \mathcal{A}$.

Fig. 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 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.

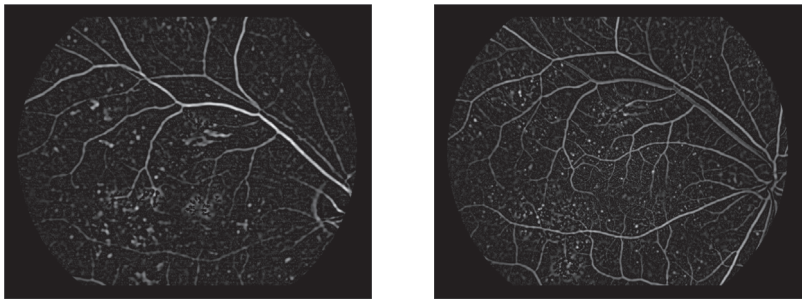
A neural network trained for multimodal reconstruction should, therefore, be able to recognize this relevant patterns. This recognition ability may be exploited in other applications of the same domain through transfer of multi-task learning approaches. Furthermore, the estimated transformation G can 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 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



(a) (b)

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

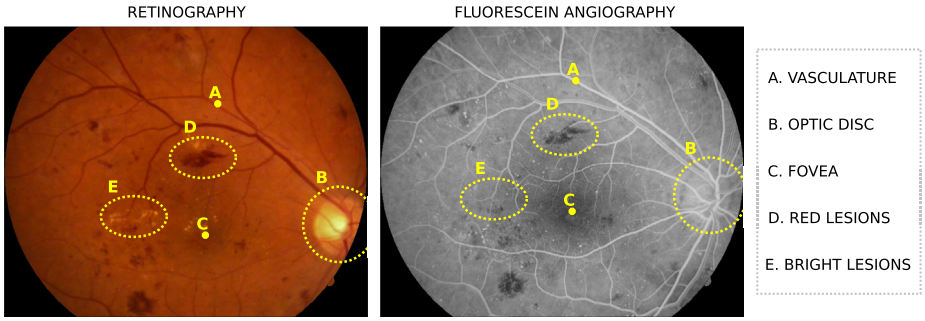


Fig. 4. Example of color retinography and fluorescein angiography from the same eye of a diabetic retinopathy patient. The appearance of the retinal structures, such as vasculature, optic disc, fovea, red lesions and bright lesions is different from one image modality to the other. The transformation between retinography and angiography requires the identification of these structures in the image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 Fig. 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 $\{(\mathbf{r},\mathbf{a})_1,\ldots,(\mathbf{r},\mathbf{a})_n\}$. 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 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 Fig. 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:

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

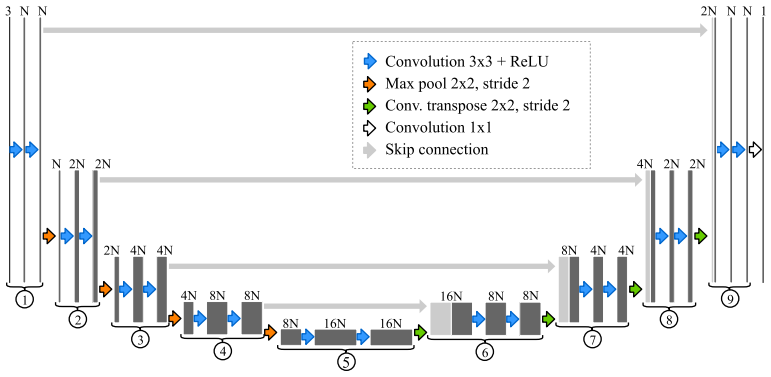


Fig. 5. U-Net architecture as implemented for the experiments of this work. The number of channels is indicated for each feature map. The numbers below identify the convolutional blocks.

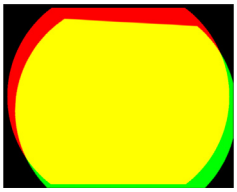


Fig. 6. Example of Multimodal ROI, in yellow, where multimodal data is available. The retinography comprises the red and yellow areas, whereas the angiography comprises the green and yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The third alternative is the optimization of the Structural Similarity (SSIM) index (Wang, Bovik, Sheikh, & Simoncelli, 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, Gallo, Frosio, and Kautz (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, Zhang, Ren, and Sun (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 and 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 $\alpha = 1e{-}4$, 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 Fig. 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 Eq. 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.

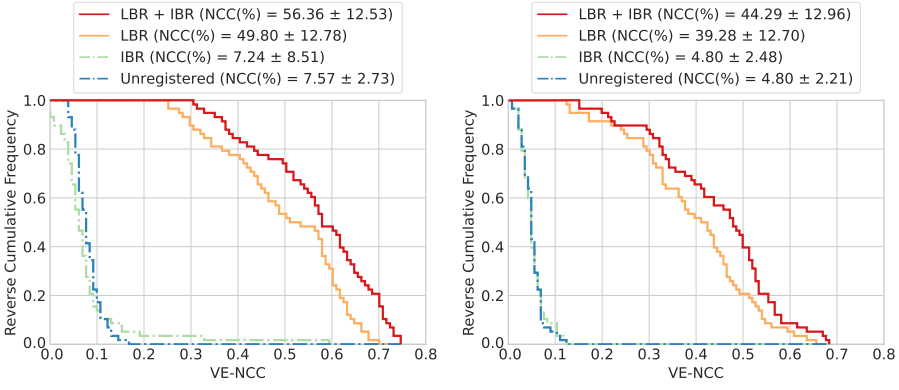
Fig. 7 depicts the reversed cumulative histograms for the VE-NCC before and after the multimodal registration in the training datasets. The plots also include the results of performing a registration with only the individual steps described in Section 2: the landmark-based registration (LBR) and the 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 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.

Fig. 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

Fig. 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 Fig. 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 (Fig. 4). On the



(a) (b)

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

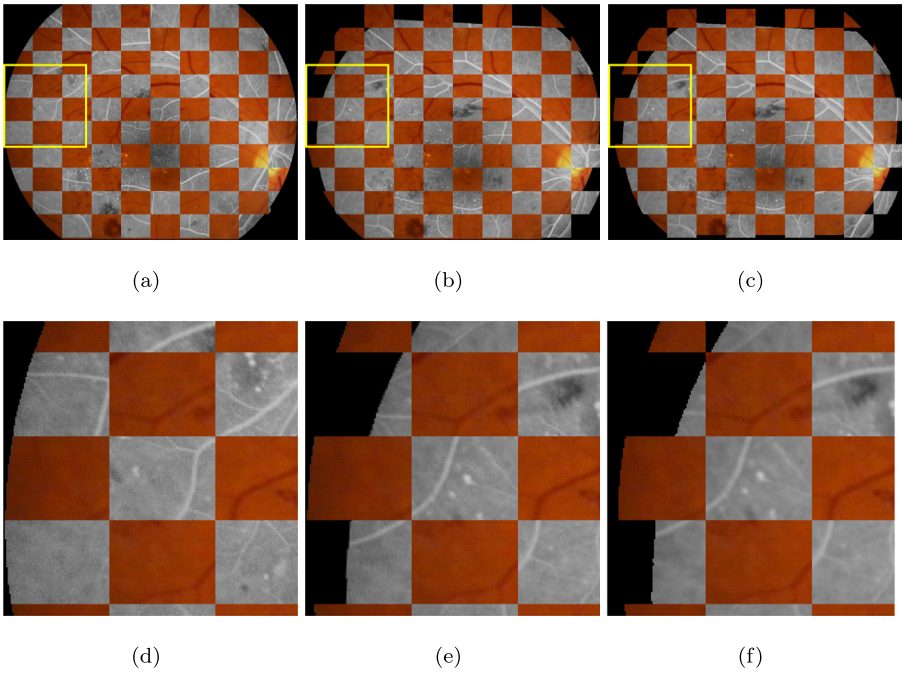


Fig. 8. Example of the multimodal registration for a retinography–angiography pair. (a) Before the registration. (b) After the initial registration (LBR). (c) After the refined registration (LBR+IBR). (d) Detail from (a). (e) Detail from (b). (f) Detail from (c).

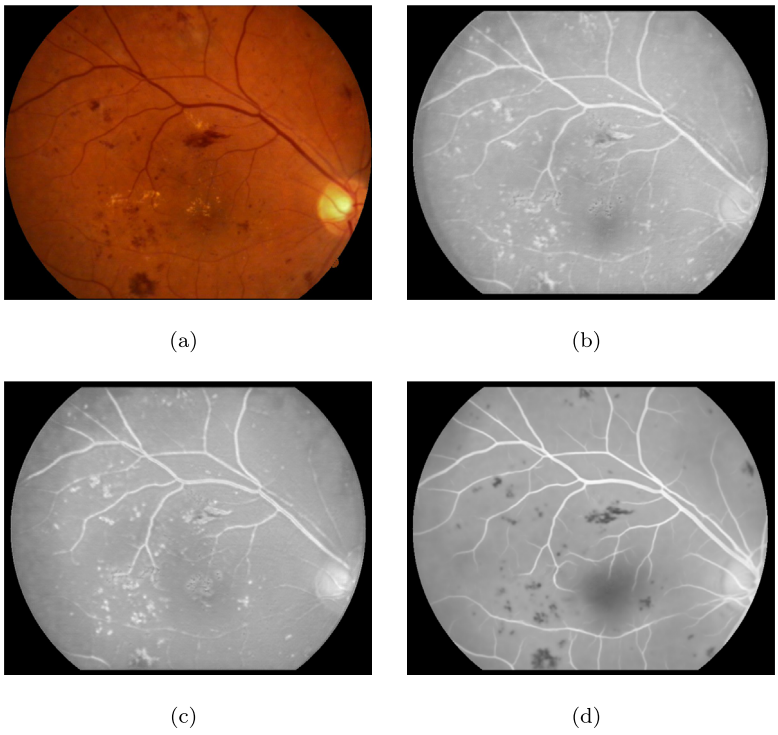


Fig. 9. Example of generated pseudo-angiographies. (a) Original retinography. (b) Using the L2 training loss. (c) Using the L1 training loss. (d) Using the SSIM training loss. L2 and L1 produce blurred images with similar appearance, whereas SSIM produces sharper images where the different retinal structures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

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 Fig. 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

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 loss	Validation 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**

The best result for each validation metric is highlighted in bold.

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 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 Fig. 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 Fig. 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. Fig. 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.

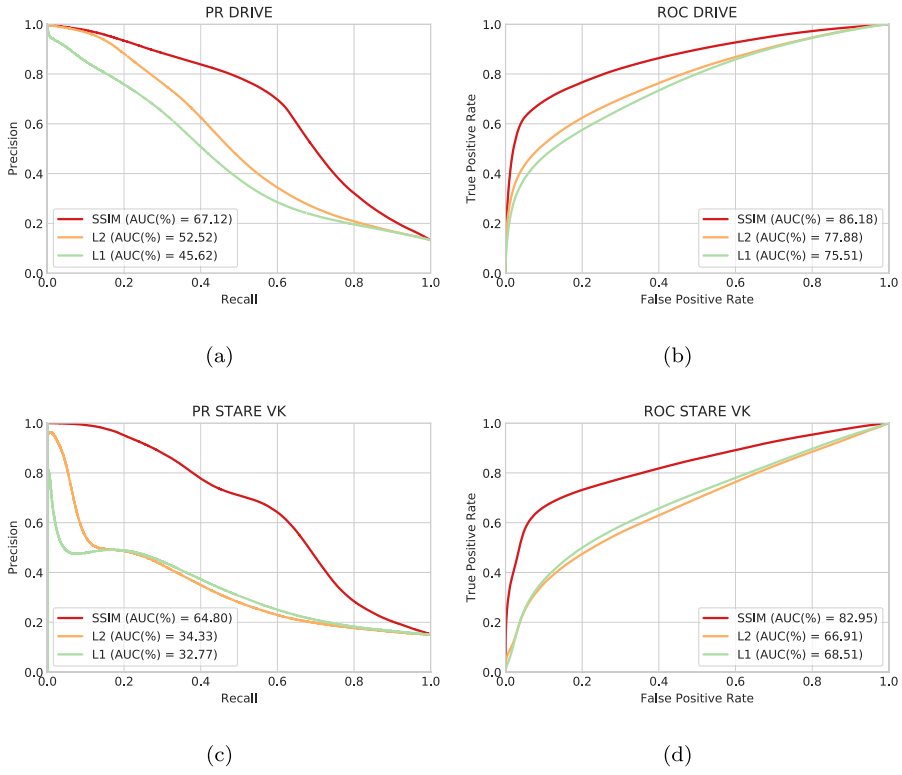


Fig. 10. Comparison of the different training losses. The graphics depict PR ((a), (c)) and ROC ((b), (d)) curves. (a)–(b) Using the DRIVE images as test set. (c)–(d) Using the STARE images as test set with the VK ground truth. The curves obtained for STARE AH are similar to those of figures (c) and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

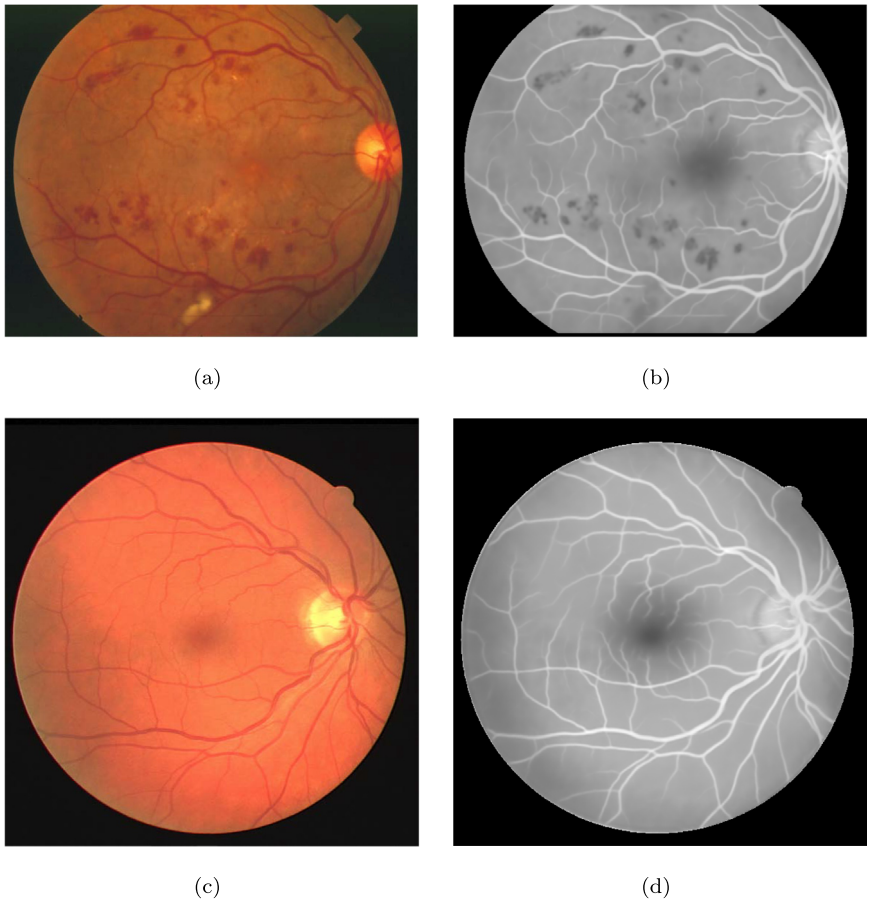


Fig. 11. Examples of generated pseudo-angiographies on images from the test datasets, using the SSIM model. (a) Retinography from the STARE dataset. (b) Generated pseudo-angiography from (a). (c) Retinography from the DRIVE dataset. (d) Generated pseudo-angiography from (c).

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 (Fig. 5) is used to control the size of the 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 Fig. 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 Fig. 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 Fig. 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.

5. Conclusions

The scarcity of annotated data in medical imaging motivates the development of solutions that target the successful training of

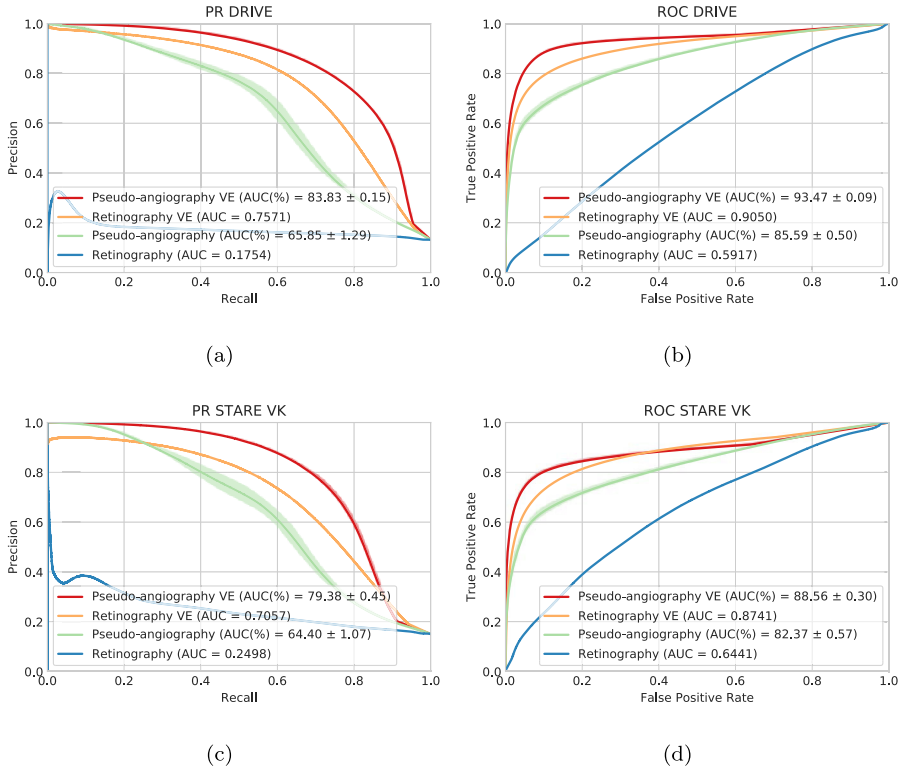


Fig. 12. Evaluation of the generated pseudo-angiography for the unsupervised recognition of vessel structures. (a)–(b) Using the DRIVE images as test set. (c)–(f) Using the STARE images as test set with the AH ground truth. The pseudo-angiography curves represent the mean and standard deviation over 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.

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	30 k	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	122 k	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	489 k	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	2 M	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	8 M	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	32 M	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	128 M	**66.03±0.94**	85.46±0.35	65.46±1.81	**87.56±0.47**	63.38±1.47	**82.38±0.46**

The best result for each metric and dataset is highlighted in bold.

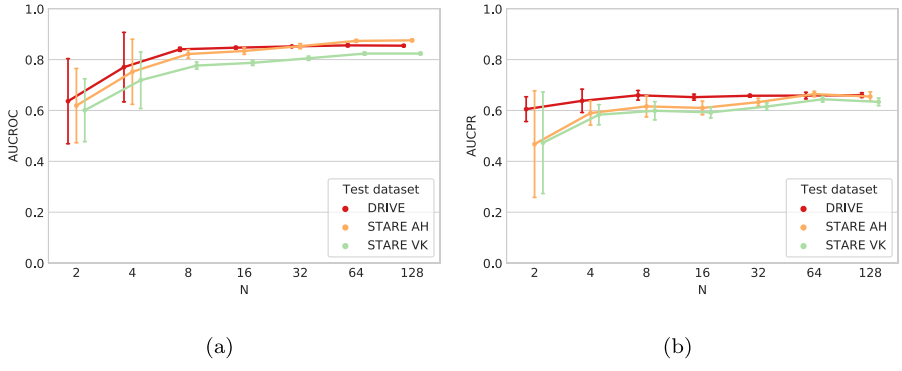


Fig. 13. Evaluation of the network size. (a) AUC-ROC with varying N. (b) AUC-PR with varying N. The plots represent the mean and standard deviation over five training repetitions. The increased network size improves the average results and reduces the variance. The improvement is higher for the more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

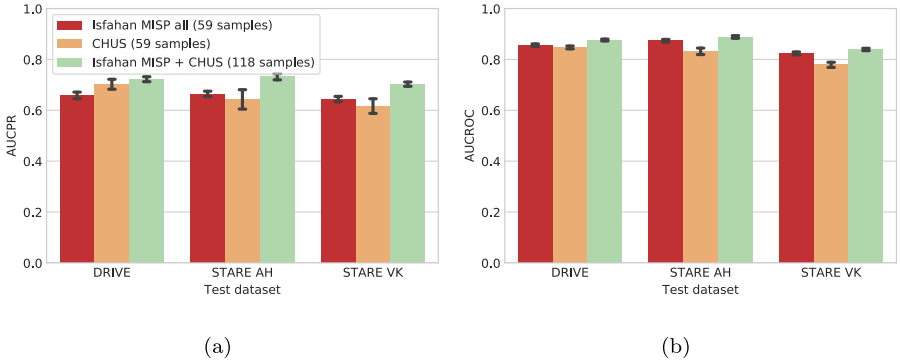


Fig. 14. Evaluation of additional training data. (a) AUC-PR. (b) AUC-ROC.

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.

CRediT authorship contribution statement

Álvaro S. Hervella: Methodology, Software, Validation, Writing - original draft, Visualization. **José Rouco:** Conceptualization, Validation, Writing - review & editing, Supervision. **Jorge Novo:** Conceptualization, Validation, Writing - review & editing, Supervision. **Marcos Ortega:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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%).

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*.

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., Abrámoff, 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). In *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.

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)*.

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)*.

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. (2015). ADNI 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., Fischer, P., & 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. In *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, 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.

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)*.

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.