Computers in Biology and Medicine 152 (2023) 106451

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Context encoder transfer learning approaches for retinal image analysis

Daniel I. Morís *, Álvaro S. Hervella, José Rouco, Jorge Novo, Marcos Ortega

Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain
Grupo VARPA, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, Xubias de Arriba, 84, 15006 A Coruña, Spain

ARTICLE INFO

Keywords:
Deep learning
Self-supervised learning
Transfer learning
Biomedical imaging
Eye fundus
Context Encoder

ABSTRACT

During the last years, deep learning techniques have emerged as powerful alternatives to solve biomedical image analysis problems. However, the training of deep neural networks usually needs great amounts of labeled data to be done effectively. This is even more critical in the case of biomedical imaging due to the added difficulty of obtaining data labeled by experienced clinicians. To mitigate the impact of data scarcity, one of the most commonly used strategies is transfer learning. Nevertheless, the success of this approach depends on the effectiveness of the available pre-training techniques for learning from little or no labeled data.

In this work, we explore the application of the Context Encoder paradigm for transfer learning in the domain of retinal image analysis. To this aim, we propose several approaches that allow to work with full resolution images and improve the recognition of the retinal structures. In order to validate the proposals, the Context Encoder pre-trained models are fine-tuned to perform two relevant tasks in the domain: vessels segmentation and fovea localization. The experiments performed on different public datasets demonstrate that the proposed Context Encoder approaches allow mitigating the impact of data scarcity, being superior to previous alternatives in this domain.

1. Introduction

The observation of the eye fundus is key for the diagnosis and treatment of important eye diseases, such as macular degeneration [1] or glaucoma [2]. Furthermore, the eye fundus observation is also relevant for the study of systemic diseases such as diabetes [3] or hypertension [4]. In this context, Computer Aided Diagnosis (CAD) methods [5] are very useful to support the work of the clinicians. The main focus of CAD systems in ophthalmology is the automated analysis of eye fundus images, using machine learning techniques. In that regard, in the last years, the use of deep learning algorithms has been widely increased due to the advantages they offer. For instance, these algorithms can be directly applied to raw data, without the necessity of using hand-engineered feature extraction methods [6].

Deep learning models have usually been trained in a supervised manner, an aspect that implies the necessity of annotated data, which is scarce in many domains [7]. Annotated data is difficult to retrieve because the manual labeling is a tedious and error-prone task that must be performed by field experts. In addition, the difficulty is greater for those cases where pixel-wise labeling is required, such is the case of the segmentation or localization of relevant structures in the images. This issue is even more critical in the case of biomedical imaging, where obtaining good quality images or proper ground truth data can be extremely challenging or even impossible in some problems.

The problem of data scarcity is well-known and, therefore, many efforts have been made to propose paradigms able to mitigate its negative consequences. In particular, it exists a kind of strategies with this aim, called as data augmentation [8]. A common strategy is to artificially augment the size of the original dataset using, e.g., classical data augmentation techniques. This data augmentation typically implies using random trivial transformations such as rotations, translations or color-intensity changes (among others). This allows to obtain new plausible images that are different from the original versions under the point of view of the deep models. Nevertheless, these transformations can be insufficient to represent the great variability of the biomedical imaging domains. Another alternative to augment the size of the original dataset is the generation of synthetic images. In the last years, Generative Adversarial Networks (GANs) have emerged as powerful deep learning architectures to generate novel synthetic images for specific domains [9]. The novel set of images generated with the GAN models can be added to the original dataset, hence artificially increases its dimensionality [10,11].

* Corresponding author at: Centro de Investigación CITIC, Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain.
E-mail addresses: daniel.iglesias.moris@udc.es (D.I. Morís), 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.compbiomed.2022.106451
Received 5 September 2022; Received in revised form 23 November 2022; Accepted 19 December 2022
Available online 22 December 2022

In contrast with the previous strategies, other paradigms aim to reuse the knowledge extracted from one task to complement the training of another task. This is the case of transfer learning [12], where the models are pre-trained in a complementary task before facing the final target task. We can find several relevant works in this scope. As reference, the work from [13] proposes the use of a model pre-trained on the ImageNet dataset to solve anomaly detection tasks in the domain of retinographic and brain imaging. Moreover, the contribution from [14] proposes a methodology of vessel segmentation using a transfer learning strategy with a fully-convolutional adaption of an AlexNet architecture. Finally, it is remarkable that the work of [15] proposes a novel architecture to solve the tasks of vessel segmentation and optic disk/cup segmentation, using a pre-trained model to extract bounding boxes and restrict the region of study as a preprocessing step in the latter case. Similarly, in multi-task learning, several tasks are trained simultaneously, in a way that each task can help to improve the performance of the other tasks [16]. However, additional tasks usually require additional annotated data and, therefore, the problem of data scarcity may still remain unsolved.

In order to further mitigate the issue of data scarcity, many efforts have been made to develop novel self-supervised learning tasks that can be used in transfer or multi-task learning settings [17]. In the self-supervised learning paradigm, the training process can be performed without the necessity of manually labeled data because the labels are automatically derived from the raw data. Some remarkable examples of self-supervised learning [18] can be seen in tasks like forecasting [19], colorization [20] or the Context Encoder [21].

The Context Encoder is based on the prediction of a masked region in an image using information from the surrounding pixels (*i.e.*, the context). This self-supervised learning strategy has demonstrated a great potential in generic domains (using natural images), obtaining an adequate performance for image inpainting and representation learning without manual supervision [21,22]. However, the Context Encoder paradigm offers an even greater potential for transfer learning in specific and restricted domains. This is due to the fact that images from generic domains can have a greater diversity, making it difficult to learn rich representations of the diverse image contents. In contrast, in specific and restricted domains, as the medical imaging modalities, the images always represent a similar reality and with a similar viewpoint, which facilitates the learning of useful high level abstractions of the data. Given this premise, some works have explored the application of the Context Encoder paradigm in particular biomedical imaging domains, such as brain MRI [23] or ultrasound imaging [24]. However, none of them have explored the application of transfer learning by pre-training solely with unlabeled images. In the same way, due to the limitations of the original Context Encoder methodology, none of these works use the original resolution of the images, that is needed for the effective analysis of some clinical cases.

Regarding the use of self-supervised learning models for transfer learning in retinal image analysis, the work from [25] uses a Multimodal Reconstruction pre-trained model (a model that converts retinographies to angiographies) to improve the performance of different target tasks related to the analysis of the retinal anatomy. However, while their method offers satisfactory results, it also presents important limitations. For instance, it requires having multimodal datasets of paired and registered images to train the model. These limitations could be avoided by using instead an alternative approach relying on a single image modality such as, e.g., Context Encoders, as it eases the construction of larger datasets for the self-supervised task, even including challenging scenarios that could lead the registration process to fail. However, in the case of the Context Encoder paradigm, previous methods are designed to work exclusively with low resolution images, hence not being valid for retinal image analysis. In that regard, in a preliminary work [26], we explored different approaches of Context Encoders for the reconstruction of eye fundus images at full resolution. However, despite the promising reconstruction results presented in this preliminary work, the potential for transfer learning in eye fundus images still remained unproved.

In this work, we study the application of the Context Encoder paradigm for transfer learning in full resolution eye fundus images. In order to deal with high resolution images, we propose 3 different variants of Context Encoder, considering both patch-wise and fully-convolutional image processing. Furthermore, and in contrast with previous works, we also explore the use of different loss functions for the Context Encoders. Our proposals allow exploiting the main advantage that the domain we are working on poses for transfer learning purposes, *i.e.*, the fact that all the images represent a similar reality, with the same structures observed from a similar viewpoint. In that regard, we evaluate our proposals in 2 different relevant downstream tasks, as is the case of vessels segmentation and fovea localization. We perform all the experiments using public datasets (in particular, Isfahan MISP, DRIVE and IDRiD) and compare the performance with a previous state-of-the-art self-supervised learning approach in eye fundus images. Overall, this methodology provides a method to work with high-resolution images and proves to be useful to mitigate the effect of data scarcity using single-modality datasets of eye-fundus images. Furthermore, it is remarkable that the great performance of the Context Encoder pre-trained model can be achieved with a relatively small retinographic dataset, as is the Isfahan MISP, composed of 59 images. Additionally, the results of this work could be also extrapolated to other medical imaging modalities and even other domains.

2. Methodology

An overview of the complete transfer learning methodology that is proposed in this work is depicted in Fig. 1. This methodology is divided into 2 parts. The first part consists in the Context Encoder pre-training using any of the three different variants that we propose. The second part reuses the domain-specific knowledge extracted from the previous pre-training phase to solve two important target tasks in eye fundus images: vessels segmentation and fovea localization.

2.1. Context Encoder pre-training

The Context Encoder [21] is a self-supervised learning paradigm that is based on predicting the content of a masked region in an image given only its surrounding pixels (*i.e.*, the context that encloses the masked region). This paradigm was initially proven in generic domains (using natural images) where the mask is solely applied in a single small region of the image. Therefore, the original methodology is meant to be applied to low resolution images (in particular, this resolution was of 128 × 128 pixels [21]). However, retinal images present a much higher resolution, an aspect that makes it impossible to directly apply the original Context Encoder approach. For this reason, we propose different alternatives to adapt the methodology, either in a patch-wise manner or in a fully-convolutional manner. In that regard, we propose a single Patch-Wise approach, denoted as PW-CE, that represents the most straightforward application of the original paradigm. On the other side, we proposed two different fully-convolutional approaches, that apply a Global Masking pattern to images. The first fully-convolutional alternative applies a CheckerBoard (CB) pattern and is denoted as GM-CE (CB). The second alternative applies instead a Center–Surround (CS) pattern, which omits less information from the image (i.e., provides more context to the network). This second alternative is denoted as GM-CE (CS). Examples of both global masking patterns can be seen in Fig. 2. For each variant, the training procedure as well as the way of generating fully reconstructed examples are explained in detail below. The reason to propose this type of masking patterns is because they allow creating regular grids on the image that completely exploit the full image size.

Patch-Wise Context Encoder (PW-CE). In order to adapt the original method to images of higher resolution, in this first approach the

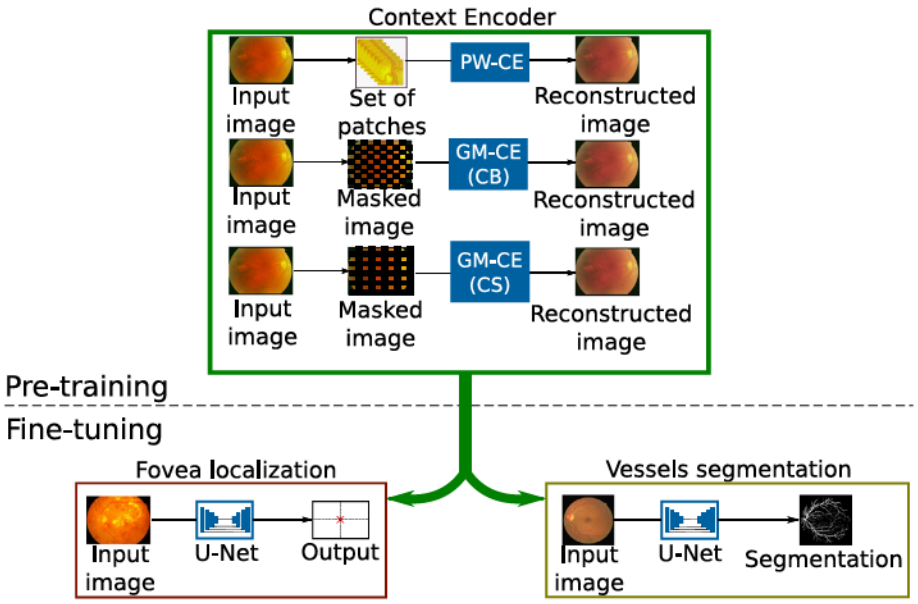
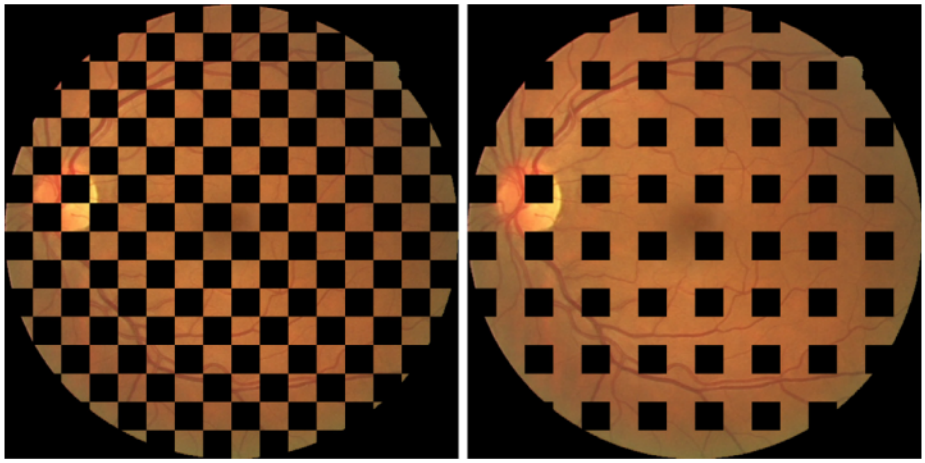


Fig. 1. Main overview of the proposed methodology, which is divided into 2 parts. The first part is a pre-training using 3 different Context Encoder approaches, and the second part is a fine-tuning in the final tasks of fovea localization and vessels segmentation.



(a) (b)

Fig. 2. Representative examples of the masking patterns used for the GM-CE approach. (a) Checkerboard pattern (GM-CE (CB)). (b) Center surround pattern (GM-CE (CS)).

images are processed in a patch-wise fashion. In that regard, we use the same patch size as in [26], given its proven suitability for this problem. A schematic representation of the training procedure for the PW-CE approach is depicted in Fig. 3. The process is as follows: firstly, a sliding window of 96 × 96 pixels is displaced over the image, extracting all the non-overlapping patches from the image. Then, a central square omission mask of 32 × 32 pixels is applied on each patch. These patches, with the central regions omitted, are used as input to the generator network, which returns the fully reconstructed patches as output. Then, the output of the generator is compared against the corresponding original patches to obtain the reconstruction loss. The training is performed by minimizing this loss.

With regard to the image reconstruction method, that is depicted in Fig. 4, it is structured as follows: firstly, the input image is processed to obtain a set of patches. This processing is made with a sliding window with size 96 × 96 pixels and a stride of 32 pixels. Then, a central omission mask of 32 × 32 pixels is applied on each input patch with size 96 × 96 pixels. The patches with the central regions omitted are used as input to the Context Encoder trained model, obtaining output patches of 96 × 96 pixels too. Each output patch is then cropped to the central region of 32 × 32 pixels, which is the one that was actually reconstructed by the network. This is because, for the surrounding unmasked regions, the networks can directly apply an identity mapping from input to output. Finally, all the reconstructed central regions are placed in their original positions obtaining, in this way, a fully reconstructed eye fundus image.

Global Mask Context Encoder with Checkerboard Pattern (GM-CE (CB)). Instead of processing the image in a patch-wise fashion, in this second approach the masking pattern is applied directly over the whole image. This fully-resolution strategy allows learning from a global context that, given the recurrence of the reality represented by retinographies, contributes with a greater learning potential. In this way, the method can be adapted to high resolution images more effectively. As its name implies, the GM-CE (CB) applies a checkerboard omission pattern, where the omission regions size can be adjusted. However, in order to make a fair comparison with the PW-CE approach, the size of the omission mask is also of 32 × 32 pixels. An example of this masking pattern can be seen in Fig. 2(a). Given this masking pattern, the training process is defined as can be seen in Fig. 5. Firstly, the checkerboard pattern masking is applied on the input image. Then,

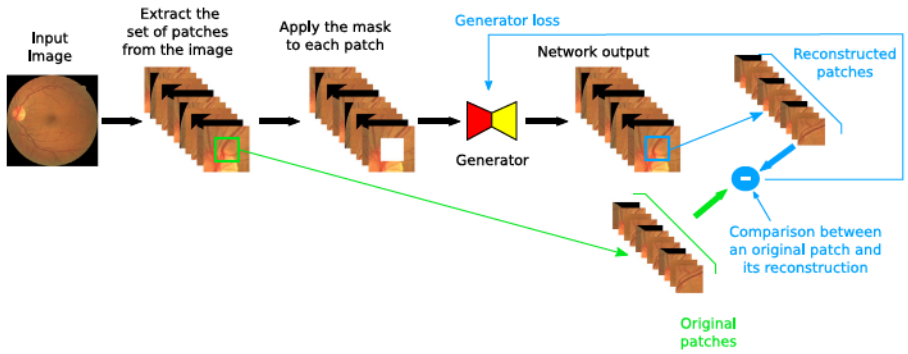


Fig. 3. Schematic description of the procedure for training process using the PW-CE approach.

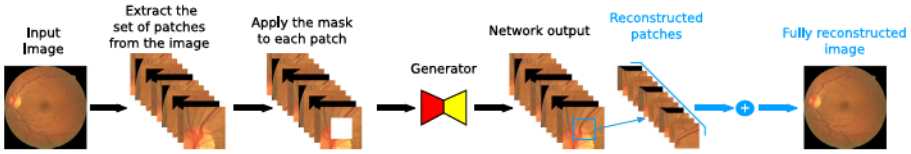


Fig. 4. Schematic description of the procedure for image reconstruction using the PW-CE approach.

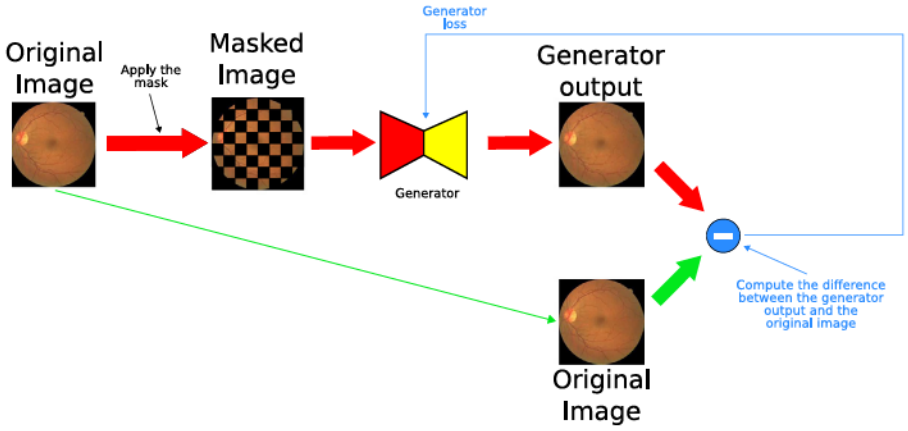


Fig. 5. Description of the training process followed to train the reconstruction model in the case of the GM-CE (CB) approach.

this masked image is used as input of the generator network, which returns the reconstructed image as output. Finally, the generator output is compared against the input image, computing the reconstruction loss. The training is performed by minimizing this loss. Additionally, as a data augmentation strategy, a random offset is applied online to the omission mask of each image during the training process.

The image reconstruction procedure for this approach is depicted in Fig. 6. Firstly, the checkerboard masking pattern is applied on the input image. Given that the network only truly reconstructs one half of the information using the checkerboard pattern, it is necessary to follow two different pathways to obtain a fully reconstructed image. In this way, the described strategy combines the outputs of the network obtained after masking the input image with the original global mask and its inverted version.

Global Mask Context Encoder with Center–Surround Pattern (GM-CE (CS)). In this third approach, the masking pattern is also directly applied over the whole image. This global masking pattern was proposed with a similar assumption as in the previous case, *i.e.*, to exploit the fact that all retinographies are always representing a very similar reality. However, in this case, we use a center–surround masking pattern, which can be seen in Fig. 2(b). The motivation of presenting this other masking pattern is to provide more contextual information to the network, which is achieved by reducing the amount of omitted information. Similarly, as in the previous case, the size of the omitted regions is also set to 32×32 pixels. The training process is depicted in Fig. 7. Firstly, the input image is masked with the center–surround pattern. Then, this masked image is used as the input of the generative model. After that, the output of the generative model is compared against the corresponding original version of the image, computing the reconstruction loss. The training is performed by minimizing this loss. In this particular case, the model only reconstructs one quarter of the information, as the model should apply an identity mapping for the remaining regions of the sample. Moreover, similarly as with the GM-CE (CB) approach, a random offset is applied online to the mask of each image during the training process as a data augmentation strategy.

With regard to the image reconstruction procedure of this approach, that is depicted in Fig. 8, the procedure is very similar to the previous approach. The main difference is that, in this case, instead of following 2 different pathways, the procedure follows 4 different pathways. This is due to the fact that, instead of 2 possible masking combinations as it was the case of the checkerboard pattern, the center–surround pattern has 4 different possibilities. Then, for each pathway, the input image is masked with the corresponding version of the pattern. Then, this masked image is used as the input of the generator model, whose outputs only have a quarter of the information actually reconstructed. To obtain only the reconstructed information, the inverted version of the initial mask is applied to the output image. Once the 4 quarters

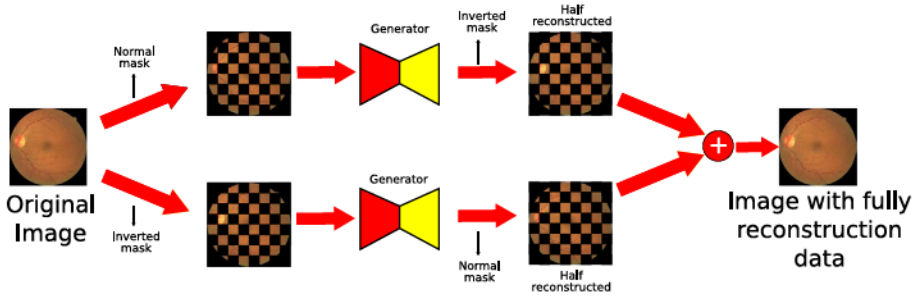


Fig. 6. Schematic description of the procedure for image reconstruction using the GM-CE (CB) approach.

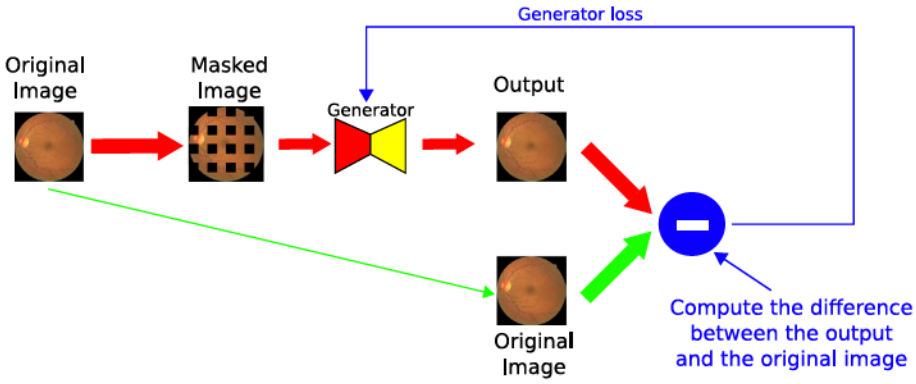


Fig. 7. Schematic description of the training process of the Context Encoder model using the GM-CE (CB) approach.

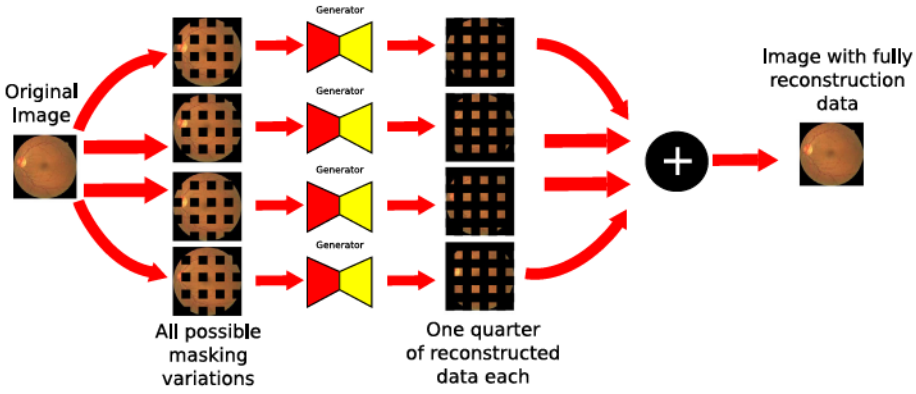


Fig. 8. Schematic description of the procedure for image reconstruction using the GM-CE (CS) approach.

of the truly reconstructed information are obtained, they are finally merged together.

Training loss. For the purposes of this work, we follow the same approach of the original Context Encoder work [21], that defines a reconstruction loss for the transfer learning experiments. In this case, the definition of an adversarial loss is unnecessary, as realistic image inpainting is out of scope for the methodology herein proposed. To perform a more exhaustive analysis of the problem, we explore two different reconstruction loss functions: Mean Squared Error (MSE), which is used in the original Context Encoder [21], and Structural Similarity (SSIM) [27], which we propose as an alternative. It is remarkable that, to the best of our knowledge, this is the first work that proposes the use of the SSIM loss to train the Context Encoder. Given these peculiarities of our contribution, the key points, and the motivation of the use of each loss function are extensively explained below.

With regard to MSE, this loss function first calculates the square of the differences between the expected output and the output predicted by the model. Then, the mean value of the squared differences is computed. In this case, the MSE loss is used because it has demonstrated its suitability for the Context Encoder in previous works [21,26]. The MSE loss can be expressed as is stated in Eq. (1):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2, \tag{1}$$

where N refers to the number of pixels, $\hat{Y}$ refers to the output predicted by the network per color channel and Y to the expected output per color channel. The final MSE loss is then calculated as the sum of the losses for each of the 3 color channels.

However, despite the MSE function being often the preferred choice to solve regression problems, it is not necessarily the most adequate alternative when dealing with images. In that regard, we propose as alternative the use of SSIM, which was specifically tailored to quantify the similarity between two images. It was also demonstrated in previous works that SSIM can improve the quality of the reconstructed eye-fundus images [28]. The SSIM loss considers three different components: luminance, contrast and structure. The computation of the metric is performed using a sliding window approach. Besides the difference between pixel values, the windows allow taking into account

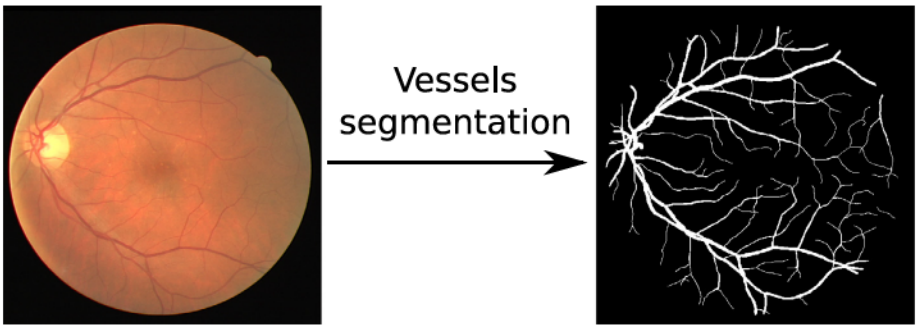


Fig. 9. Graphical description of the vessel segmentation procedure. The white pixels refer to points that belong to a vessel structure, and the black pixels refer to any other kind of eye fundus elements.

some additional useful information from the local neighborhood of each pixel. Defining a set of arbitrary windows X that are compared against another set of corresponding arbitrary windows Y, the expression of SSIM can be seen in Eq. (2):

$$SSIM(X,Y) = \frac{1}{N}\sum_{x\in X, y\in Y}\frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2}$$

where μ_x refers to the mean value of the x window, μ_y refers to the mean value of the y window, σ_x and σ_y refer to the variance of the x window and y window respectively, σ_{xy} refers to the covariance between the x window and the y window, C_1 and C_2 are 2 constants used to stabilize the division and N refers to the number of overlapping windows used on the image to compute the loss. Once again, similarly as in the previous case of the MSE loss, the expression shown in the equation refers to an individual color channel, so the final value of the SSIM loss is computed as the sum of the losses of the 3 channels. It is important to note that SSIM is actually a metric that measures the similarity between 2 images and, therefore, the higher the value, the higher the similarity. In consequence, the training must be performed by maximizing SSIM. In order to do that, we use the expression $1 - SSIM(X,Y)$ as loss function.

2.2. Vascular segmentation

The vascular segmentation is the first task that was selected to demonstrate the potential of the paradigms that are being proposed in this work. This task aims at determining the location and the extent of the blood vessels that can be seen in the retinographic images, as is depicted in Fig. 9. This task is performed as a prediction of pixel-level labels. In this way, each pixel can be classified into 2 different classes, depending on if it belongs to the retinal vasculature (positive class) or if it belongs to the non-vascular retinal background (negative class). Regarding the training process, the training data consists of retinographies and their corresponding segmentation ground truth maps. This can be denoted as $\{(R,Y)_1, \ldots, (R,Y)_N\}$ where R refers to the retinography and Y refers to its corresponding ground truth. The target of the training process is to obtain a transformation mapping that assigns to each pixel the probability of belonging to the blood vessels. To train this model, the cross-entropy loss is used, that compares the output of the network and the actual class at the ground truth. This loss function can be seen in Eq. (3):

$$\mathcal{L}_S(\hat{Y},Y) = -\sum_{i=1}^{N} Y_i \log(\hat{Y}_i) + (1 - Y_i)(\log(1 - \hat{Y}_i)), \tag{3}$$

where Y refers to the corresponding ground truth map and $\hat{Y}$ refers to the network output.

2.3. Localization of the fovea placement

Fovea localization is the second task that was selected to prove the potential of the approaches that were proposed in this work. The aim of the fovea localization task is to detect the precise location of this relevant eye fundus structure on an image, as can be seen in Fig. 10. To approach this task, the network is trained to obtain a distance map in the same way as defined in [25], where the value of each pixel denotes the distance to the target pixel that is used as reference. In order to build the ground truth distance maps, first the distance between each pixel and the target location is computed with the Euclidean norm as can be seen in Eq. (4):

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

where d_T refers to the distance map, x_i and y_i refer to the coordinates of each pixel in the image, while x_T and y_T refer to the coordinates of the target pixel. However, there are inconveniences related with using this distance map expression as it is, because it encourages the model to accurately predict the relative position of all the pixels in the image, including those far from the target. This may compromise the accurate localization of the target point. In order to avoid this issue, the distance map is modified to measure with greater precision the distance of the pixels that are closer to the target location and saturate for the more distant pixels. To do so, the location map is defined as:

$$y_L = 1 + tanh(-d_T\frac{\pi}{\beta}), \tag{5}$$

where y_L denotes the location map itself, $tanh$ refers to the hyperbolic tangent function, d_T refers to the previously defined distance map and β to the saturation distance. This factor of saturation distance allows defining the point from where the location map saturates, hence remaining close to a constant value. In this particular case, we decided to set β to the approximate value of the optic disk radius. Finally, the model expected to perform the fovea localization is trained with the loss computed when comparing the network output against the ground truth location map y_L, using the mean square error (MSE) function.

2.4. Deep network architecture

For the purposes of this work, we choose the U-Net architecture proposed by [29] due to its demonstrated capabilities dealing with biomedical image analysis problems [28,30,31]. In particular, we use the exact same architecture as proposed in the original U-Net work, thus keeping the same layers and channels settings. The only aspect that is different from the original work is the last layer of the network, as each target task has its specific setting in this regard. A graphical description of this deep architecture is shown in Fig. 11. Overall, this deep network architecture is an encoder–decoder that allows to obtain an output image with the same width and height as the input image. Additionally, it is worth noting that the U-Net is often trained from scratch, unlike other architectures that are designed to use different

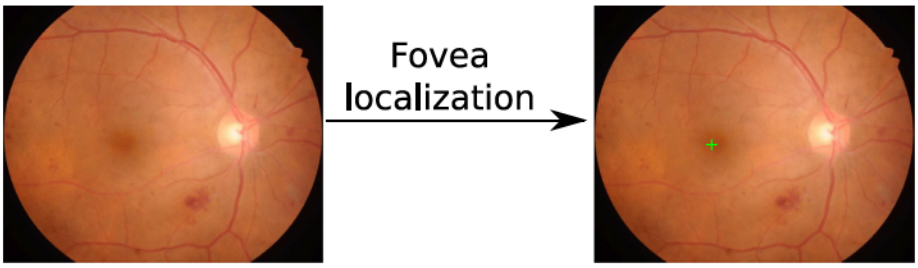


Fig. 10. Graphical description of the fovea localization task. The cross indicates the placement of this relevant eye fundus structure.

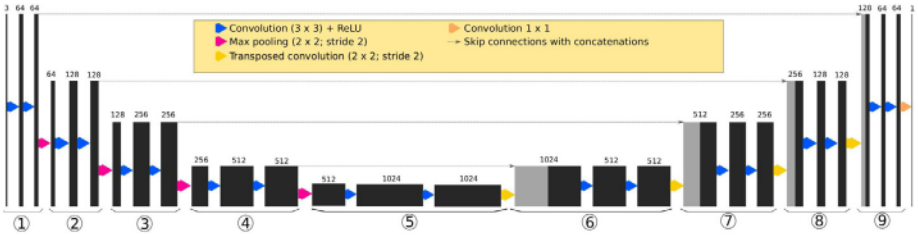


Fig. 11. Schematic representation of the U-Net architecture.

backbones pre-trained on the ImageNet dataset. It should be noted that our proposal allows pre-training both the encoder and the decoder rather than only the encoder as it happens with the previously mentioned ImageNet pre-training.

The encoder is similar to the VGG structure. In particular, it is composed of 4 blocks that are structured in 2 convolutional layers with a kernel size of 3×3 followed by a ReLU activation function and the max pooling layer. Each block in this part of the network reduces the resolution of the original image by a factor of 2. This forces the model to learn the most important features from the images. The resolution reduction is achieved with max pooling layers with a kernel size of 2×2 and a stride of 2. The decoder is composed of 4 different blocks that are structured in 2 convolutional layers with the same settings as in the encoder, followed by a transposed convolution layer with a kernel of 2×2 and a stride of 2. The transposed convolutions allow recovering the resolution of the input image. Each block increases the encoder output resolution by a factor of 2. The last block of the architecture is composed of 2 convolutional layers with the same characteristics as in the encoder and the decoder, followed by a last convolution with kernel of size 1×1. In this last layer, the activation function depends on the target task to be solved. In particular, for the case of the vessel segmentation, the considered activation function is the sigmoid function while for the case of the fovea localization this activation function will be the linear function.

The U-Net also includes the concept of skip connections. During the downsampling process, the model loses the spatial information. This compromises the upsampling part of the network, as it loses the track of the precise location of the extracted features. To avoid this situation, the U-Net architecture includes the above-mentioned skip connections, which concatenate feature maps from the earlier encoder layers to the later decoder layers. This allows to keep track of precise spatial localizations.

2.5. Network training details

Regarding the training details, the Adam algorithm is chosen to optimize the weights of the networks [32], using the following first-order and second-order momentum values, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively. In the case of the Context Encoder training, the learning rate is set to an initial value of $\alpha = 1e-4$ while for both the fovea localization and vessels segmentation, this value is initially set to $\alpha = 1e-5$. Moreover, we use a scheduler that reduces the learning rate by a factor of 10 if the validation loss stops improving during more than 2500 training steps. After this learning rate reduction, the training process is finished when the validation loss stops improving again. In all cases, the mini-batch size is set to 1 image at full resolution. In the particular case of the PW-CE approach, this means that each mini-batch has as many patches as it corresponds with a full resolution image. With respect to the network initialization in the case of training from scratch, the method proposed by [33] is used.

In order to avoid overfitting, we use dropout and several techniques of classical data augmentation with the same details as stated in the work of [25]. This applies both for the pre-training phase and for the target tasks. In particular, we add dropout layers with a probability of $p = 0.2$. These dropout layers are added at the end of the convolutional blocks 2, 3, 4, 5 and 6 with the same numeric labeling as can be seen in Fig. 11. Furthermore, the mentioned data augmentation techniques include spatial and color transformations. With regard to the spatial augmentation strategies, we use random affine transformations, in particular, shearing, scaling, and rotation. Talking about the color augmentation strategies, we use a random linear transformation, which is applied on the HSV representation of the processed image.

3. Experimental settings

3.1. Datasets

The experimental validation was performed using 3 of the most representative publicly available retinographies datasets. They are described in detail below:

- **Isfahan MISP dataset** [34]: The Isfahan MISP dataset is composed of 59 retinography–angiography image pairs, where 50 images are used for training and 9 images are used for validation. Despite the fact that angiographies are out of interest for this work, this dataset was chosen to make a fair comparison with the state-of-the-art Multimodal Reconstruction pre-training [25]. The images have a resolution of 720×576 pixels. This dataset can be divided into 2 different classes. The first class is composed of images obtained from patients that were diagnosed with diabetic retinopathy, and corresponds with half of the dataset. The second class is composed of images from patients without evidences of diabetic retinopathy and corresponds with the other half of the dataset. The retinographies in this dataset are used to train the

U-Net models for Context Encoding, that will be then used as pre-trained models of the target tasks. This means that training process is performed without considering the available manual labels.

- **Digital Retinal Images for Vessel Extraction (DRIVE) dataset** [35]: The DRIVE dataset is composed of 40 images (having 20 images for training and 20 images for test) and was specifically designed for the development of vessels segmentation methods. These images were obtained as the result of a diabetic retinopathy program from the Netherlands. Each image has a resolution of 584 × 565 pixels. This dataset is used to solve the task of vessels segmentation.
- **Indian Diabetic Retinopathy Image Dataset (IDRiD)** [36]: IDRiD is a publicly available dataset composed of 516 retinographic images (with 413 images for training and 103 for test), that can be divided into 2 different classes. The first class contains retinal images that can present diabetic retinopathy (DR), diabetic macular edema (DME) or both of them. The second class is composed of normal retinal images, *i.e.*, samples without evidences of both of the previously mentioned pathologies. These images have a resolution of 4288 × 2848 pixels. This dataset is used to solve the task of fovea localization. It is important to note that, in order to make a fair comparison with the other methods of the state-of-the-art, it was necessary to rescale the images to 858 × 570 pixels in this dataset, the same resolution that is used in [25].

3.2. Experiments and alternative methods

Overall, the experimental validation is designed to assess the benefits of the Context Encoder pre-training in scenarios that are affected by data scarcity. To do so, we conducted several experiments with a variable amount of training images in the target tasks. In particular, we train with only 1 sample, with half of the training set and with the whole training set, having the same holdout test set in the 3 cases. Moreover, it is important to point out that the training process with 1 image is repeated 5 times (with 5 different images) in order to avoid the impact that the arbitrary image selection can make in such scenario. Therefore, in these cases, we report the mean values of the 5 repetitions. For the remaining cases, we assume that the number of samples is sufficient to be representative of the whole dataset. In these cases, the training process is only performed once.

In both the experiments with 1 image and with half the training set, the remaining images of the training set are always used as validation subset for the early stopping and learning rate schedule. Meanwhile, in the experiments with the whole training set, there is no validation subset. In this case, we apply the early stopping and learning rate reduction points resulting from the experiments with half of the training set, so that the number of times each image is seen by the network remains the same.

Following the described settings, we perform a comparison among the 3 proposed Context Encoder paradigms and 2 different loss functions, making a total of 6 different scenarios. Moreover, we also perform a comparison of our proposed method with 2 reference approaches: training from scratch and training from a Multimodal Reconstruction pre-trained model. The reason to make a comparison against training from scratch is to assess the improvement that the pre-training implies over a baseline approach, while the reason to compare our method with the Multimodal Reconstruction is to contrast the performance of the proposed pre-training strategy against a relevant pre-training alternative in the literature. With regard to this latter case, it is important to take into account that Multimodal Reconstruction is one of the most powerful pre-training strategies present in the state-of-the-art for retinal image analysis. Finally, it should be noted that, in order to perform a fair comparison among all the methodologies in our experiments, we used the same random training subsets for the refinement of the downstream tasks.

- **Training the models from scratch (random initialization):** In this case, the network is trained from scratch, with the labeled data of each target task (vessels segmentation and fovea localization).
- **Multimodal Reconstruction pre-trained model:** For this scenario, the model is trained to convert a retinographic image to its corresponding angiography. To do so, it is necessary to have a dataset of pairs retinography–angiography that must be properly aligned. Finally, the pre-trained model can be fine-tuned to solve the target tasks.
- **Context-Encoder pre-trained models:** The models are trained to solve the task of inpainting in retinographic images, using each of the 3 proposed masking approaches (PW-CE, GM-CE (CB) and GM-CE (CS)). This allows the models to be pre-trained without the necessity of additional manually annotated data. Then, these pre-trained models are fine-tuned to solve the target task.

It is important to remark that both the Multimodal Reconstruction and the Context-Encoder models are trained with the Isfahan MISP dataset, making a total of 50 images for training and 9 images for validation. In order to make a comparison in equal conditions, the deep network architecture and the training strategy were the same in all cases. However, there are slight differences between vascular segmentation and fovea localization with regard to the output layer of the network as well as to the training loss as specified in Sections 2.2 and 2.3.

3.3. Evaluation metrics

With regard to the pre-training task, first we perform a qualitative evaluation of the images that were reconstructed by the Context Encoder approaches. In addition, we also perform a quantitative analysis, using 3 different reconstruction loss metrics that compare the original images with their corresponding reconstructed versions. These 3 loss metrics are the following ones: MAE (Mean Absolute Error, *i.e.*, the mean value of the absolute differences between the reconstructed image and its correspondent target image), MSE and 1-SSIM. MAE and MSE were considered as both are usual metrics in the state-of-the-art to measure differences (in particular, the absolute differences and the squared differences, respectively) while SSIM is also considered due to the fact that this metric better represents the structural similarity between 2 given images.

On the other hand, to evaluate the 2 target tasks (vessels segmentation and fovea localization) we have taken into account the usual strategies that are considered in the state-of-the-art such as in [25]. As segmentation and localization are different tasks, different evaluation metrics are used for each case. For vessels segmentation, we use both the Area Under the Curve of the ROC curve (AUC-ROC) and the Precision–Recall curve (AUC-PR) that measure how accurately the U-Net model is able to segment these structures. The reason to use both metrics is to assess the performance of the segmentation more precisely. For the problem of vessels segmentation, where the number of negative samples is much higher than the number of positive samples (as there are much more fundus points than vessels points) the difference between the ROC curve and the PR curve can be important, as the PR curve can be more sensitive for this kind of unbalanced problems. Other reason to use these metrics instead of other usual metrics like Dice score is because AUC-ROC and AUC-PR allow evaluating all the possible operation points of the method. This avoids the problematic of choosing a specific operation point that could lead to non-comparable operation points between different methods. With regard to the fovea localization, we define the metric of distance error, which is obtained computing the Euclidean distance between the ground truth and the predicted location. Thus, the quantitative evaluation can be done directly comparing these values.

Table 1
Reconstruction loss results for the 3 approaches, and the 2 considered training losses. The shadowed cells correspond to the best performance using MSE and 1-SSIM as the loss function, respectively, considering different reconstruction loss functions.

Approach	Training loss	Reconstruction metric		
		MAE	MSE	1-SSIM
PW-CE	MSE	0.1249 ± 0.0925	0.0241 ± 0.0322	0.3704 ± 0.0168
GM-CE (CB)	MSE	0.1252 ± 0.0823	0.0224 ± 0.0295	0.3068 ± 0.0310
GM-CE (CS)	MSE	0.1261 ± 0.0849	0.0231 ± 0.0295	0.3043 ± 0.0285
PW-CE	SSIM	0.1214 ± 0.0947	0.0237 ± 0.0329	0.3591 ± 0.0234
GM-CE (CB)	SSIM	0.1097 ± 0.0866	0.0195 ± 0.0306	0.2562 ± 0.0364
GM-CE (CS)	SSIM	0.1124 ± 0.0860	0.0200 ± 0.0292	0.2667 ± 0.0328

4. Experimental results and discussion

4.1. Context Encoder pre-training results

Firstly, with regard to the image reconstruction process, we perform a qualitative analysis. Some representative reconstructed samples can be seen in Fig. 12 where the most representative differences are highlighted with boxes. There, it can be seen that, independently of the used approach and loss function, the image reconstruction process obtains satisfactory results, with reconstructed retinographies that present the most important structures of the eye-fundus, as is the case of the optic disk, the macula or the vessels tree. The major differences between methods can be found in vessel structures (specially with regard to small vasculature) and other structures like bright or dark retinal lesions.

On the other hand, Table 1 depicts the quantitative analysis of the reconstructed images represented by means of the reconstruction loss metrics. From these results, some conclusions can be extracted. Firstly, given the losses obtained using the MAE and MSE metrics, it can be seen that the performance among the 3 approaches is very similar, being the GM-CE approaches slightly better than PW-CE. However, when using the 1-SSIM metric, the differences between PW-CE and both GM-CE approaches are much noticeable. Meanwhile, between the 2 GM-CE approaches themselves, the performances are very similar, even tough GM-CE (CS) is slightly better when taking 1-SSIM as the reconstruction metric. In that regard, it can be seen that the mean value of 1-SSIM is 0.3704 for PW-CE using MSE as the loss function while the mean values of the same reconstruction metric are 0.3068 and 0.3043 for GM-CE (CB) and GM-CE (CS), respectively, meaning an important drop in terms of reconstruction loss. This improvement is even more noticeable when using SSIM as the loss function for the training. Particularly, it can be seen that the mean value of 1-SSIM for PW-CE is 0.3591 while for GM-CE (CB) and GM-CE (CS) is 0.2562 and 0.2667, respectively. Then, it can be concluded that the images reconstructed using the GM-CE approaches present a better structural similarity with respect to their corresponding original versions. In the same way, it can be concluded that using the SSIM loss function for the training improves the quality of the reconstructed images for all the proposed approaches. In general, this demonstrates that GM-CE strategies take advantage of the global context to better reconstruct the main eye-fundus structures. These results are even better when training with the SSIM loss function, reinforcing the idea that SSIM is a more appropriate loss function than MSE for image reconstruction.

4.2. Transfer learning results

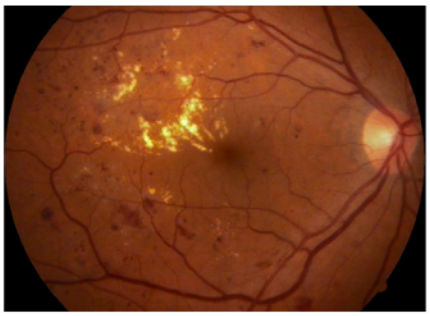
The results of the vascular segmentation are depicted in Table 2. Firstly, as expected, the overall performance of the segmentation models tend to increase as the amount of images also increases. This can be seen in terms of AUC-ROC but, most notably, in terms of AUC-PR. In fact, none of the Context Encoder approaches achieve an AUC-PR higher than 90% when fine-tuning with only 1 image. The experiments with 10 images demonstrate a greater performance with an AUC-PR always higher than 90% except in 2 cases. Finally, with 20 images, the performance increases slightly and all the approaches obtain AUC-PR values higher than 90%.

Other important aspect that can be extracted from the results of vascular segmentation is that the PW-CE approach using the SSIM loss always obtains the highest value of AUC-ROC and AUC-PR for each number of images (1, 10, and 20). In the best case, the PW-CE with SSIM loss trained with 20 images obtains a value of 97.94% of AUC-ROC and a 91.17% of AUC-PR. The graphic results of the vessel segmentation using the best approach (the PW-CE approach training with the SSIM loss) can be seen in Fig. 13. The examples depicted show that the model is able to perform a satisfactory segmentation with only 1 sample, but the model output continues to improve as the number of samples is higher, especially in the smallest vessels structures. Particularly, training with 1 image is enough to find the main vessel structures, but the smallest vessels are more well-defined when training with a higher amount of images.

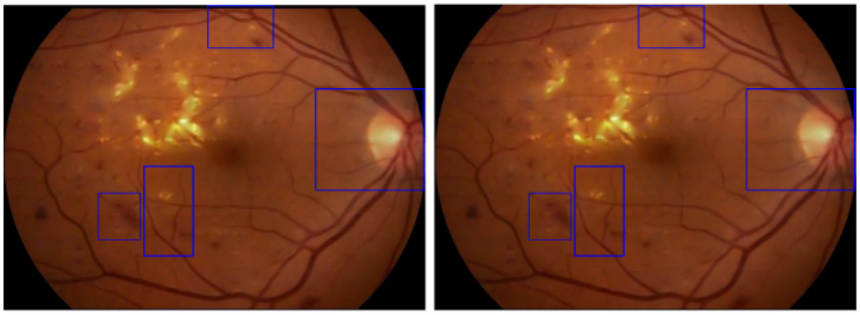
Overall, it can be seen that training with the SSIM loss implies a better performance than training with the MSE loss, as values of AUC-ROC and AUC-PR are slightly greater with this training loss for each approach. Moreover, the PW-CE is consistently better than both GM-CE approaches when training with MSE and SSIM. Finally, it is remarkable that the performance of the two GM-CE approaches is very similar, as none of them is consistently better than the other.

The results obtained for the vessel segmentation can be explained due to 2 main reasons. Firstly, the PW-CE happened to be advantageous for the vessel segmentation task. This is probably due to the fact that this approach encourages the network to rely on the local context, an aspect that is known to be important for this task. Secondly, using the SSIM loss should be advantageous because SSIM was specifically designed for computing the structural differences between images. Thus, in relative terms, SSIM should give more importance to the vascular structures than to the uniform background.

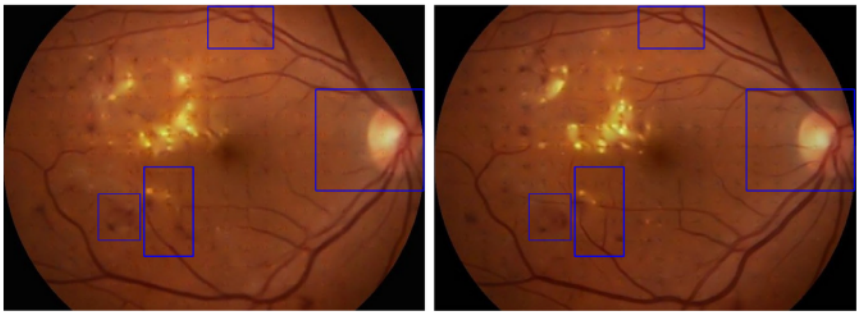
The results of the fovea localization can be seen in Table 3. As expected, the mean distance error of the model decreases as the number of training images increases. In general terms, it can be seen that the best performance obtained when training with only 1 image is achieved by the GM-CE (CB) approach using the SSIM loss function, with a mean distance error value of 75.11. In the case of training with 200 images and 413 images, the best performance is achieved by the PW-CE approach using the MSE loss, with the mean distance error values of 15.92 and 14.76, respectively. Apart from that, none of the Context Encoder approaches nor any of the training losses provide a consistently better performance than others. In that regard, given the good results that are achieved by PW-CE, it seems that the analysis of the local context may be enough to successfully localize the fovea in most of the images. However, it must also be considered that the GM-CE approaches may not be necessarily taking advantage of all the available global context. Additionally, in comparison to vessels segmentation, the good results that are achieved using MSE loss indicate that the fine structural details are not as important for the localization of the fovea. Finally, Fig. 14 depicts the evolution of the model output using the best approach with 413 images (PW-CE and the loss MSE) with respect to the number of training samples. There, it can be seen that, as this number increases, the distance map gets more precise. In fact, for both



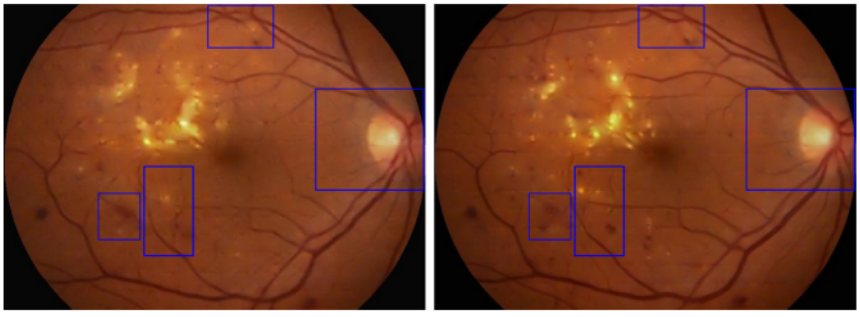
(a) Original image.



(b) PW-CE with MSE loss. (c) PW-CE with SSIM loss.



(d) GM-CE (CB) with MSE loss. (e) GM-CE (CB) with SSIM loss.



(f) GM-CE (CS) with MSE loss. (g) GM-CE (CS) with SSIM loss.

Fig. 12. Examples of images reconstructed by the Context Encoder approaches. The regions with the most significant differences are highlighted with blue boxes.

Table 2
Quantitative results obtained for the vascular segmentation task in terms of AUC-ROC and AUC-PR. The results of the best approach (PW-CE with SSIM loss) correspond with the shadowed cells.

Approach	Training loss	AUC-ROC (%)			AUC-PR (%)		
		1 image	10 images	20 images	1 image	10 images	20 images
PW-CE	MSE	96.38	97.58	97.87	87.65	90.27	90.99
GM-CE (CB)	MSE	95.86	97.43	97.78	86.60	89.85	90.79
GM-CE (CS)	MSE	96.11	97.45	97.77	87.11	89.93	90.75
PW-CE	SSIM	96.94	97.73	97.94	88.75	90.54	91.17
GM-CE (CB)	SSIM	96.81	97.67	97.88	88.63	90.44	91.06
GM-CE (CS)	SSIM	96.75	97.67	97.90	88.35	90.41	91.06

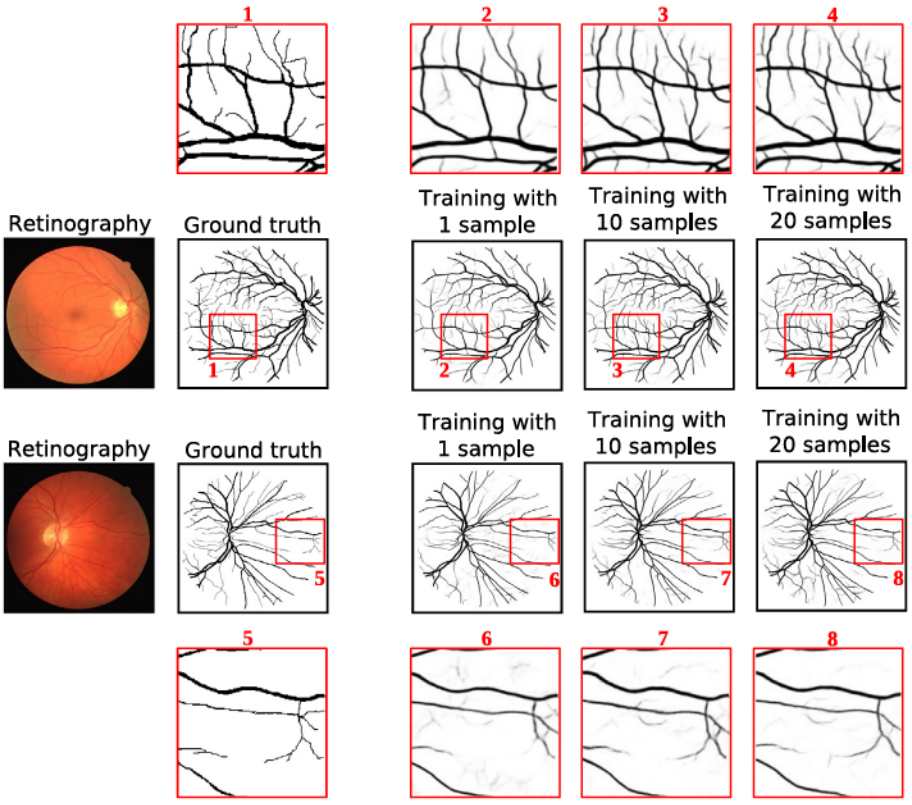


Fig. 13. Results of the vessel segmentation for 2 example retinographies training with a progressive number of samples. Some of the regions with the most notable differences are highlighted with boxes.

Table 3
Quantitative results of the fovea localization task, given the number of samples used for the training process. The shadowed cells point out the best result for each amount of used images.

Approach	Training loss	Mean distance error		
		1 image	200 images	413 images
PW-CE	MSE	89.47	15.92	14.76
GM-CE (CB)	MSE	80.21	20.52	18.85
GM-CE (CS)	MSE	82.19	16.54	17.51
PW-CE	SSIM	86.91	17.80	18.20
GM-CE (CB)	SSIM	75.11	19.26	15.69
GM-CE (CS)	SSIM	85.33	17.50	16.53

retinographic images, the distance map shows noticeable background noise that tends to decrease as the number of training images increases. It is remarkable that the bottom image also shows a considerable amount of false candidate detections, that also tend to disappear as the number of training images increases. This is due to the fact that this second image shows a great amount of dark lesions that can be confused with the macular area by the model.

In general, the overall idea that can be extracted from the results of both transfer learning tasks is that the PW-CE approach performs better in almost all cases. However, all the 3 approaches of Context Encoding demonstrate a very similar performance. Therefore, even if the performance achieved by the GM-CE approaches seems to be slightly lower in many cases, this is compensated with the advantages that this kind of approaches offer. In particular, the fully-convolutional fashion allows applying the methodology to full resolution retinographies without the necessity of patch-wise processing, an aspect that simplifies both the training and the inference processes. Finally, a point that can be made regarding the loss function is that the two losses achieve a similar overall performance for the task of the fovea localization. However,

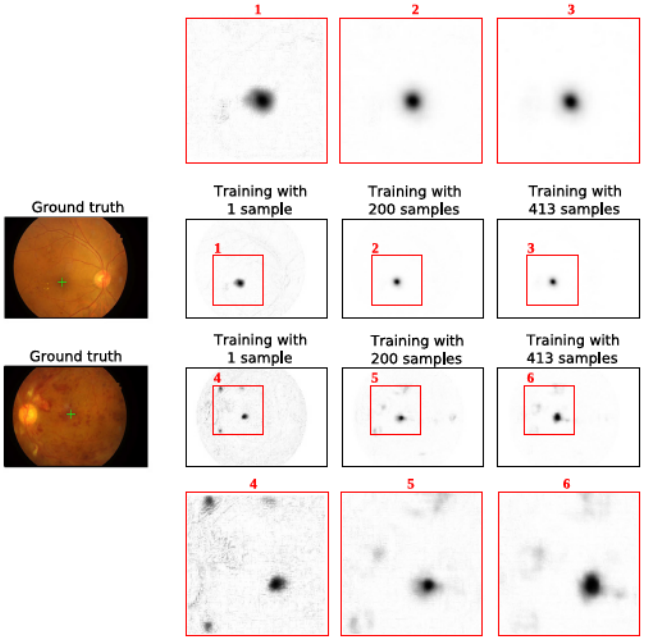


Fig. 14. Results of the fovea localization that depict the evolution of the distance maps as the number of training samples increases. Some remarkable regions of the images where notable difference can be seen are highlighted with boxes.

SSIM always achieves a higher performance for the task of vessels segmentation.

Table 4
Comparison of the method herein proposed with the best configuration against the random initialization and the Multimodal Reconstruction for the vessel segmentation task. The highlighted cells remark, for each amount of images, which is the best obtained performance.

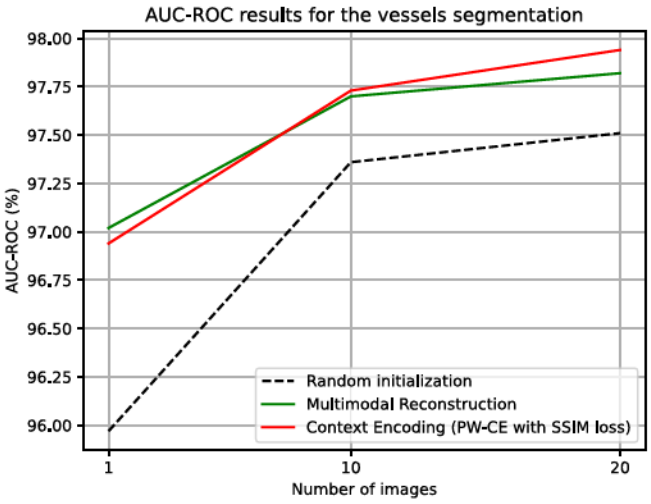
Method	AUC-ROC (%)			AUC-PR (%)		
	1 image	10 images	20 images	1 image	10 images	20 images
Random initialization	95.97	97.36	97.51	86.72	89.83	90.44
Multimodal Reconstruction [25]	97.02	97.70	97.82	89.14	90.52	91.02
Ours	96.94	97.73	97.94	88.75	90.54	91.17

4.3. Comparison with alternative methodologies

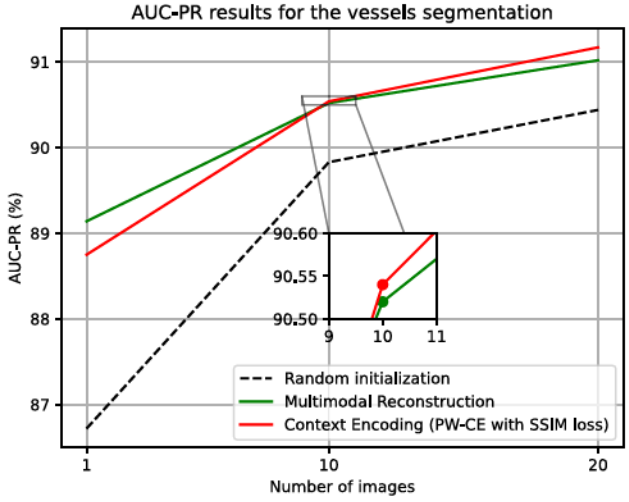
As part of the experimental validation, we also compare the performance obtained by our method in vascular segmentation and fovea localization against training from random initialization and from the Multimodal Reconstruction pre-training proposed by [25]. This latter work was considered for the comparison given that the nature of the proposed pre-training is very similar to the one that is proposed here, (as it uses a self-supervised learning task to pre-train the models and the same network architecture). It is important to note that, for our method, we are taking as reference the PW-CE paradigm training with the SSIM loss as this is the scenario with the best overall performance. The reason to compare this methodology with those 2 scenarios (random initialization and Multimodal Reconstruction) is to contrast the results obtained against a baseline approach without any kind of pre-training (random initialization) and a pre-training approach (Multimodal Reconstruction) that represents one of the most powerful strategies of the retinal image analysis state-of-the-art.

Regarding the vessel segmentation, the comparison can be seen in Table 4, where the best performance is obtained by our method in terms of both AUC-ROC and AUC-PR while using the best configuration of Context Encoding. Additionally, Fig. 15 depicts graphically the evolution of the performance in both AUC-ROC and AUC-PR with regard to the number of images used to refine the model. There, it is clearly noticeable that the performance of the random initialization is considerably lower in comparison with both transfer learning approaches. In that regard, both the Context Encoding and the Multimodal Reconstruction offer very similar results, being the Multimodal Reconstruction slightly better when training with 1 image, but slightly worse for 10 and 20 images for both metrics. It is important to note that, in general, despite the fact PW-CE with SSIM loss is the approach with the best results, the GM-CE approaches show a very similar performance, considerably better than random initialization and close to the performance obtained by the Multimodal Reconstruction approach.

With respect to the fovea localization, the comparison is shown in Table 5. It can be seen that the Context Encoding pre-training provides a considerable improvement with respect to the random initialization, as the mean distance error decreases from 22.47 to 14.76 when training with 413 images. In the same way, there is also a slight improvement when comparing our method against the Multimodal Reconstruction. Moreover, Fig. 16 depicts the performance improvement with respect to the number of images, showing a very similar scenario to the vascular segmentation. Firstly, the mean distance error of the random initialization is always noticeably higher, specially when training with only 1 image. In the case of the transfer learning approaches, it can be seen that, when training with 1 image, the performance of Context Encoding is closer to random initialization than to Multimodal Reconstruction. However, when training with 200 and 413 images, the Context Encoder experiences a notorious improvement, producing a slightly better performance than the Multimodal Reconstruction for both cases. In a similar line as in the vessel segmentation, the GM-CE approaches obtain similar results to Multimodal Reconstruction and the PW-CE approach trained with MSE loss (the best Context Encoder approach for this task).



(a)



(b)

Fig. 15. Comparison among the results obtained by the random initialization, the Multimodal Reconstruction and the best approach of Context Encoding for vessels segmentation (PW-CE with SSIM loss). (a) AUC-ROC results. (b) AUC-PR results.

With these results, it can be seen that the Context Encoder clearly improves the performance in comparison with a random initialization approach. Additionally, in comparison with the Multimodal Reconstruction, the previous state-of-the-art transfer learning methodology for retinal images that represents one of the most powerful approaches currently available in the field, the performances are very similar. However, the Context Encoder offers important advantages, mainly

Table 5
Comparison of the method proposed in this work given the best configuration with the random initialization results and the method of Multimodal Reconstruction for the fovea localization task. The highlighted cells show the best performance obtained for each amount of images.

Method	Mean distance error		
	1 image	200 images	413 images
Random initialization	113.25	36.38	22.47
Multimodal Reconstruction [25]	43.23	16.45	14.86
Ours	89.47	15.92	14.76

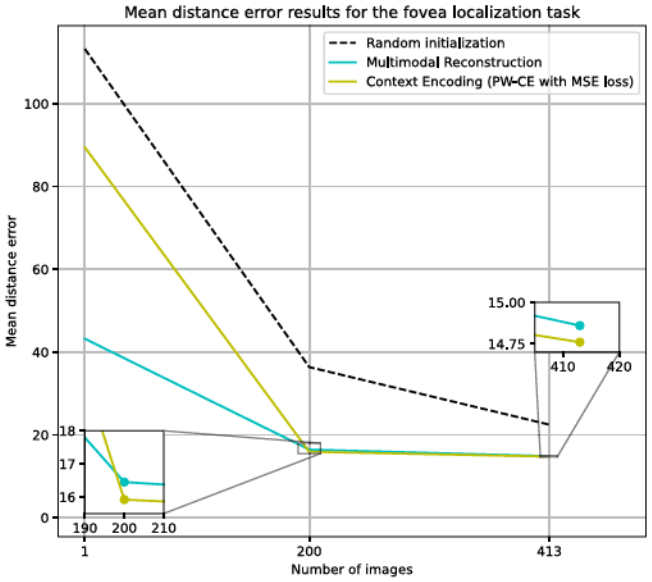


Fig. 16. Results of mean distance error obtained for the fovea localization task. In this case, the lower the values, the better the results.

due to the wider availability of single-modality eye fundus datasets. In contrast, the Multimodal Reconstruction requires more complex datasets composed of simultaneous multimodal images obtained from the same patient, an aspect that makes the data gathering much more challenging. Another limitation of this pre-training strategy is that the images of the same patient must be paired and registered, an aspect that makes it necessary to develop more complex methodologies (including the challenging task of registering images of different modalities rather than images of the same modality). Therefore, despite all the Multimodal Reconstruction strengths, the proposed Context Encoder transfer learning approach emerges as a powerful alternative strategy to mitigate the problem of data scarcity in this biomedical imaging domain. In that regard, we would like to point out that the methodology herein proposed could be adapted conveniently to perform other tasks as well, such as image classification and disease diagnosis. The potential of our proposal for these applications could be explored in future works.

5. Conclusions

Despite the great capabilities that deep learning algorithms provide, data scarcity is still a very common issue that limits their application in biomedical image domains. In that regard, in this work, we propose the application of the Context Encoder paradigm for transfer learning in retinal imaging. To this end, we propose 3 different approaches to adapt the original low-resolution Context Encoder methodology to the full resolution eye fundus images. The first approach considers a patch-wise processing, while the remaining 2 approaches consider a fully-convolutional processing. Additionally, in contrast to previous works, we also propose the use of the SSIM index for the loss function of the Context Encoder. This has the potential to facilitate the learning of the different structures in the images. Finally, we aim at demonstrating the advantages of the proposed approaches by solving, using transfer learning, two relevant tasks in retinal image analysis: vessel segmentation and fovea localization.

To validate the proposals, we conducted an exhaustive experimentation on different publicly available datasets. Overall, the results demonstrate that all the proposed Context Encoder approaches are able to recognize the most relevant structures of the eye fundus images without the necessity of manually labeled data. The results also demonstrate that the knowledge extracted from this self-supervised task can be reused to improve the performance of 2 relevant finalist tasks in the context of retinographic images: vessels segmentation and fovea localization. These tasks were chosen as they are often used as reference in the state-of-the-art, but this pre-training could be generalized to solve any kind of task within the application domain. Thus, the methodology herein proposed emerges as a powerful strategy to mitigate the data scarcity issue using single-image modality datasets for pre-training. This is a more powerful strategy in comparison with the previous state-of-the-art approach of Multimodal Reconstruction, as in this latter case paired and registered multimodal data is needed. Finally, it must be noticed that the developments made in this work could be extrapolated to other relevant tasks in the field as, for example, optic disk segmentation and localization, pathology detection, and even to different biomedical imaging domains where the analysis of full resolution images is also required. This also applies to other relevant experiments, such as training with bigger datasets to evaluate how this impacts on performance. Additionally, different paradigms of self-supervised learning, other types of pre-training and other architectures could be studied in independent works to compare their performance with the Context Encoder. These ideas should be explored in future works.

CRediT authorship contribution statement

Daniel I. Morís: Methodology, Software, Validation, Writing – original draft, Visualization. **Álvaro S. Hervella:** Methodology, Software, Validation, Writing – review & editing, Visualization, Supervision. **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 research was funded by Instituto de Salud Carlos III, Government of Spain, DTS18/00136 research project; Ministerio de Ciencia e Innovación y Universidades, Government of Spain, RTI2018-095894-B-I00 research project; Ministerio de Ciencia e Innovación, Government of Spain through the research project with reference PID2019-108435RB-I00; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, Spain through the predoctoral grant contract ref. ED481A 2021/196 and postdoctoral grant contract ref. ED481B-2022-025; and Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; Axencia Galega de Innovación (GAIN), Spain, Xunta de Galicia, grant ref. IN845D 2020/38; CITIC, Centro de Investigación de Galicia, Spain ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, Spain, through the ERDF (80%) and Secretaría Xeral de Universidades (20%). Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] L.S. Lim, P. Mitchell, J.M. Seddon, F.G. Holz, T.Y. Wong, Age-related macular degeneration, Lancet 379 (9827) (2012) 1728–1738, http://dx.doi.org/10.1016/S0140-6736(12)60282-7.

[2] D.A. Lee, E.J. Higginbotham, Glaucoma and its treatment: a review, Am. J. Health-Syst. Pharma. 62 (7) (2005) 691–699, http://dx.doi.org/10.1093/ajhp/62.7.691.

[3] O. Faust, R.A. U., E.Y.K. Ng, K.-H. Ng, J.S. Suri, Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review, J. Med. Syst. 36 (1) (2010) 145–157, http://dx.doi.org/10.1007/s10916-010-9454-7.

[4] S. Chatterjee, S. Chattopadhya, M. Hope-Ross, P. Lip, Hypertension and the eye: changing perspectives, J. Hum. Hypertens. 16 (10) (2002) 667–675, http://dx.doi.org/10.1038/sj.jhh.1001472.

[5] K. Doi, Computer-aided diagnosis in medical imaging: historical review, current status and future potential, Comput. Med. Imaging Graph. 31 (4–5) (2007) 198–211, http://dx.doi.org/10.1016/j.compmedimag.2007.02.002.

[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.

[7] M.A. Bansal, D.R. Sharma, D.M. Kathuria, A systematic review on data scarcity problem in deep learning: Solution and applications, ACM Comput. Surv. (2021) http://dx.doi.org/10.1145/3502287.

[8] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48, http://dx.doi.org/10.1186/s40537-019-0197-0.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: An overview, IEEE Signal Process. Mag. 35 (1) (2018) 53–65, http://dx.doi.org/10.48550/arXiv.1710.07035.

[10] D.I. Morís, J. de Moura, J. Novo, M. Ortega, Cycle generative adversarial network approaches to produce novel portable chest X-Rays images for Covid-19 diagnosis, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 1060–1064, http://dx.doi.org/10.1109/ICASSP39728.2021.9414031.

[11] D.I. Morís, J. de Moura, J. Novo, M. Ortega, Data augmentation approaches using cycle-consistent adversarial networks for improving COVID-19 screening in portable chest X-ray images, Expert Syst. Appl. 185 (2021) 115681, http://dx.doi.org/10.1016/j.eswa.2021.115681.

[12] Q. Yang, Y. Zhang, W. Dai, S.J. Pan, Transfer Learning, Cambridge University Press, 2020, http://dx.doi.org/10.1017/9781139061773.

[13] X. Li, H. Yang, Z. Lin, P. Krishnaswamy, Transfer learning with joint optimization for label-efficient medical image anomaly detection, in: J. Cardoso, et al. (Eds.), Interpretable and Annotation-Efficient Learning for Medical Image Computing, Springer International Publishing, Cham, 2020, pp. 146–154, http://dx.doi.org/10.1007/978-3-030-61166-8_16.

[14] Z. Jiang, H. Zhang, Y. Wang, S.-B. Ko, Retinal blood vessel segmentation using fully convolutional network with transfer learning, Comput. Med. Imaging Graph. 68 (2018) 1–15, http://dx.doi.org/10.1016/j.compmedimag.2018.04.005.

[15] S. Zhang, H. Fu, Y. Yan, Y. Zhang, Q. Wu, M. Yang, M. Tan, Y. Xu, Attention guided network for retinal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'2019), Springer, 2019, pp. 797–805, http://dx.doi.org/10.48550/arXiv.1907.12930.

[16] Y. Zhang, Q. Yang, An overview of multi-task learning, Nat. Sci. Rev. 5 (1) (2018) 30–43, http://dx.doi.org/10.48550/arXiv.1706.05098.

[17] D. Hendrycks, M. Mazeika, S. Kadavath, D. Song, Using self-supervised learning can improve model robustness and uncertainty, 2019, http://dx.doi.org/10.48550/arXiv.1906.12340, arXiv.

[18] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 4037–4058, http://dx.doi.org/10.1109/TPAMI.2020.2992393.

[19] C. Raman, H. Hung, M. Loog, Social processes: Self-supervised forecasting of nonverbal cues in social conversations, 2021, http://dx.doi.org/10.48550/arXiv.2107.13576, arXiv.

[20] C. Xiao, C. Han, Z. Zhang, J. Qin, T.-T. Wong, G. Han, S. He, Example-based colourization via dense encoding pyramids, Comput. Graph. Forum 39 (1) (2020) 20–33, http://dx.doi.org/10.1111/cgf.13659.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544, http://dx.doi.org/10.48550/arXiv.1604.07379.

[22] Y. Zeng, J. Fu, H. Chao, B. Guo, Learning pyramid-context encoder network for high-quality image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1486–1494, http://dx.doi.org/10.1109/CVPR.2019.00158.

[23] K. Armanious, V. Kumar, S. Abdulatif, T. Hepp, S. Gatidis, B. Yang, ipA-MedGAN: Inpainting of arbitrary regions in medical imaging, in: 2020 IEEE International Conference on Image Processing, ICIP, 2020, pp. 3005–3009, http://dx.doi.org/10.1109/ICIP40778.2020.9191207.

[24] S.-Y. Hu, S. Wang, W.-H. Weng, J. Wang, X. Wang, A. Ozturk, Q. Li, V. Kumar, A.E. Samir, Weakly supervised context encoder using DICOM metadata in ultrasound imaging, 2020, http://dx.doi.org/10.48550/arXiv.2003.09070.

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

[26] D.I. Morís, Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, Context encoder self-supervised approaches for eye fundus analysis, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–8, http://dx.doi.org/10.1109/IJCNN52387.2021.9533567.

[27] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for neural networks for image processing, 2015, http://dx.doi.org/10.48550/arXiv.1511.08861, arXiv.

[28] Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal image understanding emerges from self-supervised multimodal reconstruction, in: MICCAI, 2018, http://dx.doi.org/10.1007/978-3-030-00928-1_37.

[29] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015, http://dx.doi.org/10.48550/arXiv.1505.04597.

[30] Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Syst. Appl. 161 (2020) 113674, http://dx.doi.org/10.1016/j.eswa.2020.113674.

[31] J. Morano, Á.S. Hervella, N. Barreira, J. Novo, J. Rouco, Multimodal transfer learning-based approaches for retinal vascular segmentation, 2020, http://dx.doi.org/10.48550/arXiv.2012.10160.

[32] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv.

[33] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034, http://dx.doi.org/10.1109/ICCV.2015.123.

[34] S.H.M. Alipour, H. Rabbani, M.R. Akhlaghi, Diabetic retinopathy grading by digital curvelet transform, Comput. Math. Methods Med. 2012 (2012) http://dx.doi.org/10.1155/2012/761901.

[35] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, B. van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509, http://dx.doi.org/10.1109/TMI.2004.825627.

[36] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau, Indian diabetic retinopathy image dataset (idrid), 2018, http://dx.doi.org/10.21227/H25W98.