Context encoder self-supervised approaches for eye fundus analysis

Daniel I. Morís[1,2] and **Álvaro S. Hervella**[1,2,*] and **José Rouco**[1,2] and **Jorge Novo**[1,2] and **Marcos Ortega**[1,2]

Abstract.

The broad availability of medical images in current clinical practice provides a source of large image datasets. In order for these datasets to be useful in the training of deep neural networks, it is necessary to provide annotations associated to each image. However, the image annotation is a tedious, time consuming and error prone process that requires the participation of experienced specialists.

In this work, we propose different complementary context encoder self-supervised approaches to learn relevant characteristics for the restricted medical imaging domain of retinographies. In particular, we propose a patch-wise approach, inspired in the previous proposal of broad domain context encoders, and a complementary fully convolutional approach. These approaches take advantage of the restricted application domain to learn the relevant features of the eye fundus, situation that can be extrapolated to many medical imaging issues.

Different representative experiments were conducted in order to evaluate the performance of the trained models, demonstrating the suitability of the proposed approaches in the understanding of the eye fundus characteristics. The proposed self-supervised models can serve as reference to support other domain-related issues through transfer or multi-task learning paradigms, like the detection and evaluation of the retinal structures or anomaly detections in the context of pathological analysis.

1 INTRODUCTION

The human eye is one of the most complex anatomical parts of the body. In particular, the main relevance of the retinal observation is related with common diseases like glaucoma [14] or age-related macular degeneration (AMD) [4] which are among the main causes of blindness. Additionally, many systemic diseases, as hypertension [10] or diabetes [7] are also frequently studied in the eye fundus as they directly affect the retinal structures. In the analysis of all those systemic and eye diseases, the early detection and treatment is crucial to avoid or mitigate their outcome. In this context, the support of Computer Aided Diagnosis (CAD) [9] systems is extremely useful.

The analysis of the retina is typically based on the inspection of ophthalmic image modalities that graphically represent the eye fundus. The automatization of this analysis using machine learning techniques is the usual target of current CAD systems [12]. In this context, the use of deep learning models has been increased during the last years given their advantages when dealing with raw signal data (including medical imaging) [19]. These models represent a powerful alternative to the classic machine learning approaches that require the processing of raw data using hand-engineered feature extraction methods. However, in spite of these advantages, the training of deep learning models is usually supervised and requires a significantly larger amount of labeled data. Despite the wide availability of raw image data, the corresponding annotations are difficult to retrieve, specially in domains like the biomedical environment where the manual labeling is a tedious and error-prone process that must be performed by experienced specialists. This potential data scarcity represents an important limitation for the application of deep learning techniques in many biomedical applications.

Data scarcity is a well-known limitation in deep learning, existing several strategies to mitigate its impact. The most commonly used method is data augmentation [11]. It consists in artificially increasing the size of an image dataset using direct image transformations as random rotations, random color intensity variations, pixels translations, etc. Other approaches like transfer learning or multi-task learning techniques aim to take advantage of labeled data for complementary tasks. Transfer learning [22] implies the reuse of knowledge extracted from an already learned task to optimize the posterior learning on other related task (which is actually the target task). Similarly, multi-task learning [5] consists in training the model to perform both tasks simultaneously. Both techniques are based on the hypothesis that if there is a problem of data scarcity with a target task, the model training can be supported with the learning of complementary related tasks, helping to increase the final target performance. However, the labeled data for appropriate complementary tasks may also be scarce on many application domains. This motivated the recent proposal of different self-supervised learning paradigms [1]. It is considered self-supervised as it is not necessary to support the training process with manual labeling. Instead, the target labels are automatically derived from the unlabeled data. We can find many applications of self-supervised learning models as, for reference, colorization [24], autoencoders [3] or future predictors [13]. In particular, in ophthalmology, some applications can be found as the multimodal prediction of fluorescein angiographies from classical retinographies [6] [18].

In self-supervised learning, the use of context encoders [15] represent a powerful strategy where an omitted region in an image is reconstructed using information from the context. Context encoders have proved their potential in generic domains [17], being able to generate images with genuine appearance and being also capable of learning relevant features from the analyzed pictures. As result, context encoding has demonstrated its effectiveness as pre-training task to improve the performance of other classification, detection and segmentation related tasks [17]. However, the application of context en-

[1] Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain.
[2] VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña, Spain.
* Corresponding author. Email: a.suarezh@udc.es

coders have been limited to applications in broad domain images, where the diversity of structures that may be present is difficult to be covered.

In this context, biomedical imaging represents a more restricted and plausible scenario for context encoding. Image modalities of a particular specialty always capture and represent the same domain and structures of the patients. Thus, it should be possible to predict the image contents from the context with a higher accuracy and level of detail. If context encoding demonstrated its potential from generic domains, its capacity can be better exploited in the medical environment to even allow the exploitation of specific semantics of the application domain.

In this work, we propose different context encoder approaches using self-supervised learning for the analysis in a particular medical domain, ophthalmology, for the comprehension of the eye fundus. With this aim, two alternative strategies are presented:

Patch-wise context encoder. This first approach is inspired in the original context encoder proposal for generic domain [17], adapted to analyze the target retinographies. In line with the original proposal, and in order to deal with the much higher image resolution, the retinographies are sampled in patches of a defined size where the central regions are omitted. This forces the system to learn to reconstruct the omitted data only using its immediate local neighborhood.

Fully-convolutional global mask context encoder. Given the limited context of analysis to generate each omitted region of the previous approach, and considering the restricted context of analysis of this medical image modality (as many others), an alternative global approach is also proposed and studied, using directly the whole image as the input of the network. In addition to the efficiency improvement in terms of time, the fully-convolutional global mask context encoder uses context from the whole image in order to get the reconstruction data. This could help to improve the performance of the model as it is well-known that, in medical domains, using the global context from the images is a desirable characteristic for any CAD procedure in the analysis of many diseases. This is related, once again, with the idea that medical images from a particular environment reflect always the same reality (in our case, the eye fundus). The hypothesis is that using a global context could bring a better effectiveness from the model.

Successfully trained models of either approach present the potential to improve the performance of other related tasks with the support of transfer learning or multi-task learning techniques, or can be used to model normal images in the context of anomaly detection [21].

2 MATERIALS AND METHODS

In this work, we propose two approaches allowing to provide retinal pattern understanding through context encoding. The first approach, named patch-wise context encoder (PW-CE), is inspired in a previous proposal in generic domain [17] and consists in the application of the paradigm in a patch-wise fashion. In the second approach, instead, we propose a novel strategy allowing to perform context encoding in a fully-convolutional fashion with the whole images thanks to the application of global masks following a specific local omission pattern. This alternative approach is denoted as fully-convolutional global mask context encoder (GM-CE).

2.1 Patch-wise context encoder

Context encoders [17] were originally proposed using a Generative Adversarial Network (GAN) architecture [8], composed of a generator and a discriminator, that is trained to fill in a missing portion of the input image according to the surrounding context.

The use of a GAN model allows to integrate a point-wise reconstruction loss of the generator with the adversarial loss derived from the discriminator, which can evaluate the plausibility of the overall aspect of the generated image. The used training loss for the generator network $\mathbf{g}$ is a linear combination of adversarial and reconstruction losses defined as

$$\mathcal{L} = \lambda_{rec}\, \mathcal{L}_{rec} + \lambda_{adv}\, \mathcal{L}_{adv}\ , \qquad (1)$$

where λ_{rec} and λ_{adv} are parameters weighting the importance of the reconstruction $\mathcal{L}_{rec}$ and adversarial $\mathcal{L}_{adv}$ losses, respectively. The used reconstruction loss for each image is defined as

$$\mathcal{L}_{rec} = ||\mathbf{M} \odot [\mathbf{x} - \mathbf{g}((1-\mathbf{M}) \odot \mathbf{x})]||_2^2\ , \qquad (2)$$

where $\odot$ denotes the element-wise product operation, $\mathbf{M}$ a binary omission mask, $\mathbf{x}$ an input image, $\mathbf{g}(\cdot)$ the generator network and $||\cdot||_2^2$ the L_2 norm. Differently, the adversarial loss $\mathcal{L}_{adv}$ for the image $\mathbf{x}$ is given by

$$\mathcal{L}_{adv} = \log\left(\mathbf{d}(\mathbf{x}) + \log\left(1 - \mathbf{d}(\mathbf{g}(1-\mathbf{M}) \odot \mathbf{x})\right)\right)\ , \qquad (3)$$

where $\mathbf{d}(\cdot)$ denotes the discriminator network. The training discriminator network $\mathbf{d}$, contrary to the network $\mathbf{g}$, is performed so that the adversarial loss $\mathcal{L}_{adv}$ is maximized.

The proposed network architecture, based in [17], is composed of a generator and a discriminator, both detailed in Figure 1. It works with input images of fixed size (128×128) and outputs of 64×64. The input images are masked with a central omission square of 32×32 pixels. This omission mask is smaller than the one in the original paper (64×64) to alleviate the prediction complexity of the problem as there will be a surrounding context with trivial reconstruction. In addition, this allows the discriminator to evaluate the generated output conditioned by this trivial context. The use of low resolution and fixed size images in the original approach is not appropriate for medical imaging applications, which usually require a much higher resolution. Additionally, it would not be appropriate to predict the whole center part of a medical image, as the important features are usually related to fine details of local structures. Thus, in order to apply the prior work on the retinal image context, we propose the use of a patch-wise processing approach.

Figure 2 depicts the general scheme of the proposed training strategy. First, each image in the training and validation stages is split into a dense set of non-overlapping image patches of size 128×128. This set of image patches is then used for training as originally proposed, by applying the omission mask in the center region of size 32×32.

Additionally, we propose a patch-wise reconstruction approach for the test images that allows to generate synthesized images at full resolution. In this case, as depicted in Figure 3, the test images are split into a dense set of 128×128 size patches. However, in this case, the patches are overlapped by using a stride of a quarter the size (32 pixels on each dimension). Then, the trained generator network is used to generate 64×64 images from each masked input patch, from which only the central 32×32 regions are used, and placed back into their original image position. This gives rise to a full resolution image that is completely synthesized by the network from the local context.

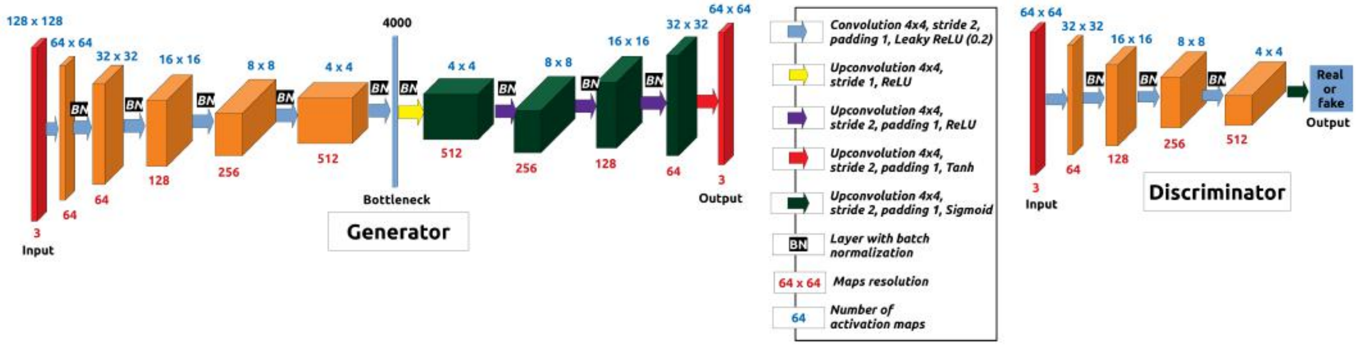

128 x 128
64 x 64
32 x 32
16 x 16
8 x 8
4 x 4
4000
4 x 4
8 x 8
16 x 16
32 x 32
64 x 64
BN
3
Input
64
64
128
256
512
Bottleneck
512
256
128
64
3
Output
Generator
Convolution 4x4, stride 2, padding 1, Leaky ReLU (0.2)
Upconvolution 4x4, stride 1, ReLU
Upconvolution 4x4, stride 2, padding 1, ReLU
Upconvolution 4x4, stride 2, padding 1, Tanh
Upconvolution 4x4, stride 2, padding 1, Sigmoid
BN Layer with batch normalization
64 x 64 Maps resolution
64 Number of activation maps
64 x 64
32 x 32
16 x 16
8 x 8
4 x 4
Real or fake
Output
3
Input
64
128
256
512
Discriminator

Figure 1. Architectures used in the PW-CE approach. Example with an input patch size of 128 × 128 for the generator and 64 × 64 for the discriminator.

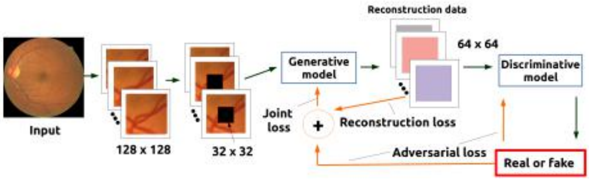


Figure 2. Training process of the PW-CE approach.

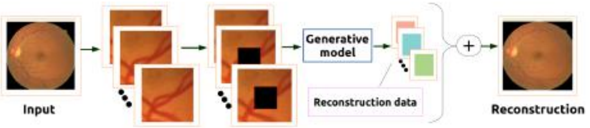
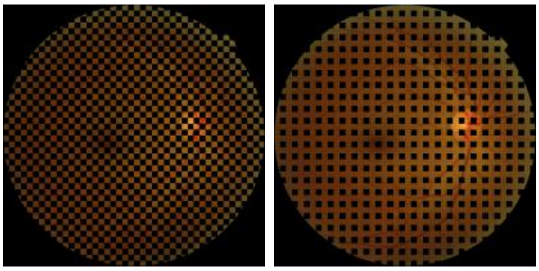


Figure 3. Generation of the reconstruction images using the PW-CE approach.

2.2 Fully convolutional global mask context encoder

The PW-CE approach has several a priori limitations. On one hand, the size of the image patches and allowed omission masks is fixed in the network architecture design, and requires specific processing for the training dataset. However, the omission and context sizes are linked to the scale of the relevant information that the network can encode and predict. Thus, this scale parameter should probably be adjusted for each application. On the other hand, despite that the local structures are important in medical imaging, these images are characterized for usually having a normalized viewpoint, with a relevant global context. However, the local analysis provided by the PW-CE can not account for this relevant global context information. To solve these issues, an alternative approach is proposed named fully convolutional global mask context encoder (GM-CE), based on the use of fully-convolutional neural networks (FCN) and global masks with local omission patterns.

In relation to the omissions, we propose two alternative global masks following two local omission patterns: checkerboard (CB) and center surround patterns (CS). Examples of these masks are illustrated in Figure 4 and the training process for this approach is depicted in Figure 5. These masks allow to adjust the size of the local omitted regions. Additionally, they can be easily shifted spatially as a data augmentation technique, allowing to account for a wider diversity of local regions. Note that this same effect could be achieved using the PW-CE approach, but it would imply the extraction of dense overlapping patches from the images, which is not practical in computational terms. The main difference between both mask pattern reside on the amount of local context that is available to predict each of the omitted parts. The local omission patterns (*i.e.* the masking squares) are of size 32 × 32 in this work for comparison purposes with the previous approach.



(a) (b)

Figure 4. Examples of masked images for the GM-CE approach. (a) Checkerboard pattern mask (CB). (b) Center surround pattern mask (CS).

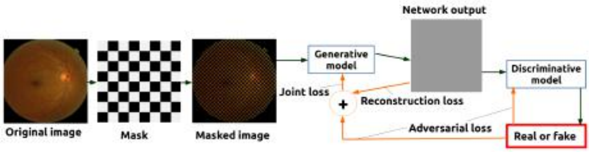


Figure 5. Example of the training process using a GM-CE (CB) approach.

Regarding the network architecture, we propose minor modifications to the PW-CE network in 1 so that it is fully convolutional with equal input and output sizes. To that end, an additional upconvolutional layer is added at the end of the generator, being completely symmetric. Likewise, an additional convolutional layer with stride 2 is added to the discriminator to increase the receptive field to 128 × 128. Additionally, to provide a scalar output for the discrimi-

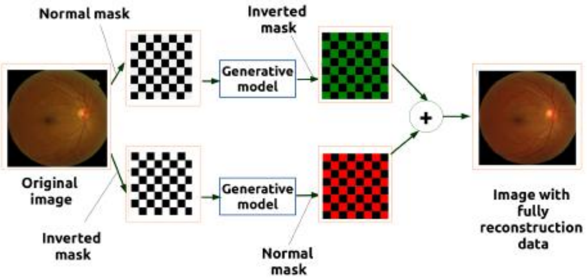


Figure 6. Image generation procedure using GM-CE (CB).

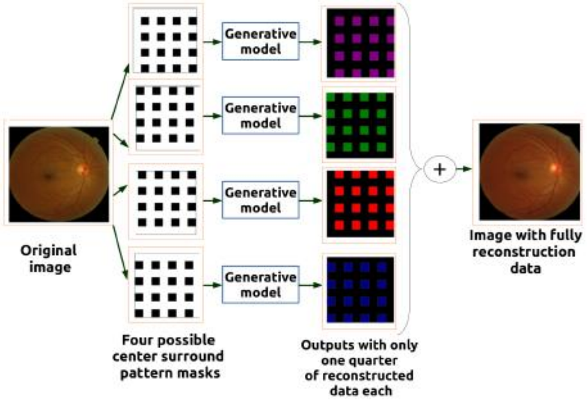

Generative model
Generative model
Generative model
Generative model
Original image
+
Image with fully reconstruction data
Four possible center surround pattern masks
Outputs with only one quarter of reconstructed data each

Figure 7. Image generation procedure using GM-CE (CS).

nator, a global average pooling is added as the last layer.

The generation of completely synthetic images is simplified using these models. The reconstruction procedures for the case of GM-CE (CB) using the checkerboard pattern, and the GM-CG (CS) using the center surround pattern, are depicted in the schemes of Figures 6 and 7, respectively. The overall idea is to combine the outputs of the trained generator network using the input image masked with the original global mask and all its complementary versions. This allows to generate fully synthetic images based on the context by only using two inference calls to the trained generator, in the case of the GM-CE (CB), and four inference calls, in the case of the GM-CE (CS).

2.3 Retinal image datasets

In order to test and validate the proposed context encoder approaches, two different public retinal image datasets of reference were used: DRIVE [20] and MESSIDOR [2]. The MESSIDOR dataset has been established to help studies on CAD software also for diabetic retinopathy, containing $1,200$ eye fundus images from three different departments, presenting resolutions that vary from $1,440 \times 960$ to $2,304 \times 1,536$ pixels. Instead, the DRIVE dataset contains 40 retinographies from a diabetic retinopathy program with a resolution of 565×584. From the 40 cases, 7 present signs of mild early diabetic retinopathy, and the remaining 33 are from healthy subjects.

In this work, we aim at providing a preliminary validation of the proposed approaches with a subset of the MESSIDOR dataset, along with images from the DRIVE dataset. Specifically, we selected a random subset of 160 healthy and pathological images from MESSIDOR for training. Also, a complementary test set of 40 images from MESSIDOR was used. The DRIVE images were used to corroborate the reached conclusions on an independent set. The eye fundus ROI was identified in all the images to omit the background, homogenizing their resolutions to a convenient ROI size of $1,408 \times 1,408$ (multiple of 128×128). This allows to obtain a homogeneous size of the retinal structures for all the varying input resolutions. This standardized ROI size dataset is used for both the PW-CE and GM-CE approaches. In the case of the GM-CE approach, the images are directly used as the network inputs. On the other hand, for the PW-CE approach, the images from the training and validation sets are split into 128×128 patches. Using the indicated patch size, 121 non-overlapping samples are obtained from each retinography, which results in $19,360$ patches for training and $4,840$ patches for validation.

Related with the training processes, the PW-CE is trained with a mini-batch size of 121 patches. On the other hand, the GM-CE are trained with a mini-batch size of 1 image. Additionally, in this case, an online data augmentation is performed in the form of random horizontal flips, which transform right eye images into left eye appearance, and vice versa. Moreover, the global masks are applied with horizontal and vertical random shifts of up to 32 pixels..

2.4 Evaluation of the context encoder approaches

In order to evaluate the capabilities of the proposed approaches, we performed different complementary quantitative and qualitative analyses. Firstly, we studied the model errors during the training process. Given that the training with a GAN model is more difficult than a standard network due to the use of the two losses combined, its evolution can be better understood graphically. Secondly, with respect to the reconstruction stage, fully reconstructed retinography examples are provided to offer a direct and graphical idea about the performance of the approaches. Finally, two additional quantitative analyses are provided based on global reconstruction errors and their corresponding reconstruction error maps. The global reconstruction errors are calculated using as reference the complete reconstructed retinographies and their corresponding original retinographies. For each image, the differences are computed using three error metrics: L_1 and L_2 and SSIM [23]. The first two integrate pixel-wise differences along the images. SSIM, instead, integrate local statistics (computed in 7×7 image windows) to evaluate the structural similarity between image patches. These reconstruction errors are averaged for the entire test set to obtain global metrics. In addition, we calculated an L_2 error map for each retinography by locally computing the squared difference on each image position, and locally integrating them in 8×8 size windows. This allows to evaluate the image patterns achieving a higher reconstruction error.

3 RESULTS

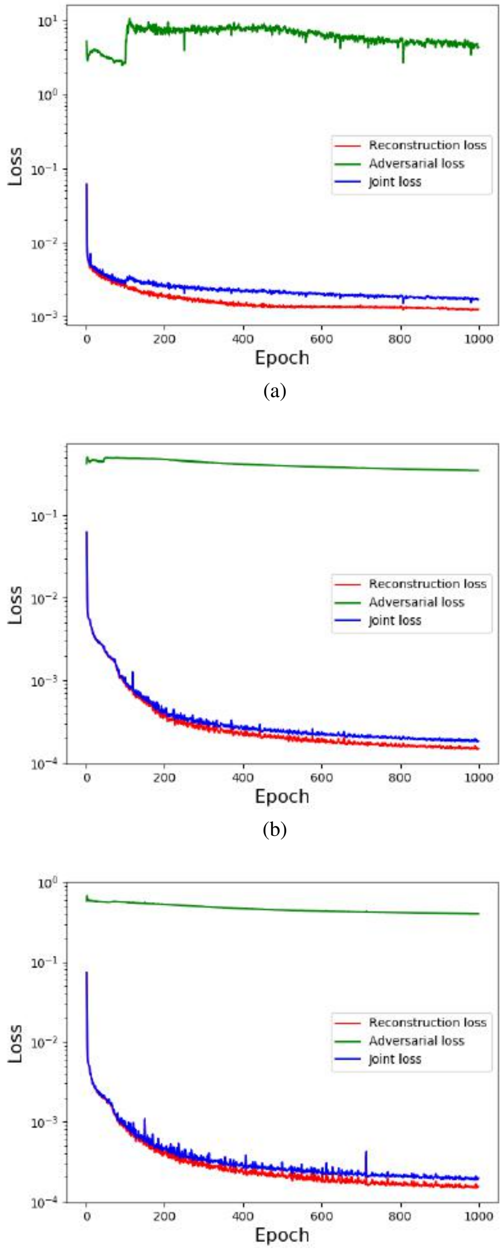
All the experiments to analyze and validate the proposed approaches were performed in the three proposed cases, *i.e.* the patch-wise context encoder PW-CE model, along with the fully-convolutional global mask context encoder GM-CE for the cases of checkerboard pattern GM-CE (CB) and center-surround pattern GM-CE (CS). The Adam optimization algorithm [16] was used to train all the compared networks. The learning rate was set to $\alpha = 2e - 4$, and the decay rates to $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Meanwhile, the weighting values of the loss in Eq. 1 are $\lambda_{rec} = 0.9999$ and $\lambda_{adv} = 0.0001$ for reconstruction and adversarial losses, respectively. All the networks were trained for a fixed number of $1,000$ epochs, although from the examination of the training curves, and the generated images, it was observed that the GM-CE approaches could have been benefit from further training, while the PW-CE approach almost stalled.

With respect to the loss during the training process, we can see representative examples of the evolution of each approach in Figure 8. In both approaches, the adversarial loss (discriminative model loss) is significantly higher than both the reconstruction loss and the joint loss. A relevant aspect that can be observed in the PW-CE approach is the significant oscillation of the adversarial loss. Probably, this is because the network inputs are considerably different among them as they are patches from any part of the retinal fundus. This represents a very complex scenario for the discriminative model as the criterion of what is realistic or non-realistic seems to be difficult to find. Finally, as expected, the joint loss and the reconstruction loss are very similar, being the joint loss slightly higher. On the other hand, the oscillation of the adversarial loss in both GM-CE is less significant than in the PW-CE approach. This is probably because the discriminator inputs can take advantage of the global context to smooth the existent local differences. Due to this, it is easy for the discriminative model to find the criterion to understand what is realistic and what is non-realistic.

Table 1 shows the global reconstruction error obtained for the proposed approaches in the MESSIDOR test set. It is observed that the PW-CE approach obtains higher reconstruction error compared to the GM-CE approaches, regardless of the evaluation metric. The GM-CE approaches present a similar performance between them, considering their standard deviations. The most significant differences between the PW-CE and GM-CE, are in terms of SSIM. This is a quantitative indication of the higher reconstruction accuracy for the GM-CE models regarding the structural information, *i.e.* the appearance of the relevant local shapes matches the original image better, regardless of the similarity of the independent pixel values.

Considering completely independent images from both the training and test sets, some representative fully-reconstructed retinographies from the DRIVE set are depicted in Figure 9. In particular, we have selected one healthy (first row), and two pathological cases, one of them presenting a pathological optic disc (second row), and the other with bright retinal lesions (third row). Additionally, the corresponding error maps for the first example are presented in Figure 10. Globally, it is observed that the main structures of the eye fundus can be easily recognized in the reconstructed images, regardless of the approach. In fact, the reconstructed images generally show that the appearance is very similar to the original retinographies. Specifically, those relevant structures are the main vasculature, the optic disc or the macula. Additionally, many secondary small vasculature is also recognized, being a complex identification case for any computational method. We would like to remark that this microvasculature plays an important role in the analysis of the retinal microvasculature that is associated with many diseases.

By visual inspection, the direct comparison between both paradigms allows to identify some important differences. Regarding the vascular tree, it is observed that the GM-CE approaches provide more coherent and continuous vessels. The PW-CE reconstructs much less small vessels, which are sometimes inconsistent or with a staircase appearance. This same staircase reconstruction pattern is specially observed at the ROI boundary reconstruction. This is specially appreciated comparing the error maps of Figure 10, and clearly visible in the details of Figure 11. This is due to the independence among the patches during the reconstruction, as well as the significantly lower shift sampling variability during the training phase.



(a)

(b)

(c)

Figure 8. Evolution of the training loss. (a) PW-CE. (b) GM-CE (CB). (c) GM-CE (CS).

Regarding the optic disc, the overall intensity reconstruction does not present significant differences among the compared approaches. However, it should be noticed that there is a more complex vascular pattern in this region that, once again, is better handled by the GM-CE approaches, despite that the reconstruction is not accurate in some cases (see the first row example in Figure 9, and the second row details in Figure 11, where there is an horizontal vessel that is missed in the reconstruction). In this case, the better appearance of the vessel tree is due to structural consistency. The PW-CE approach shows split vessels with incoherent orientation, while the global appearance of the GM-CE results looks like a regular vessel tree pattern. This is an indication that the GM-CE approaches are able to learn higher

Table 1. Average and standard deviation reconstruction errors on the MESSIDOR test set for the three tested models.

	L1 ($\times 10^{-2}$)	L2 ($\times 10^{-2}$)	1-SSIM
PW-CE	3.8799 ± 0.5688	0.3068 ± 0.0894	0.0782 ± 0.0102
GM-CE(CB)	2.2415 ± 1.2391	0.1329 ± 0.1233	0.0355 ± 0.0115
GM-CE(CS)	2.4862 ± 1.3864	0.1581 ± 0.1658	0.0332 ± 0.0111

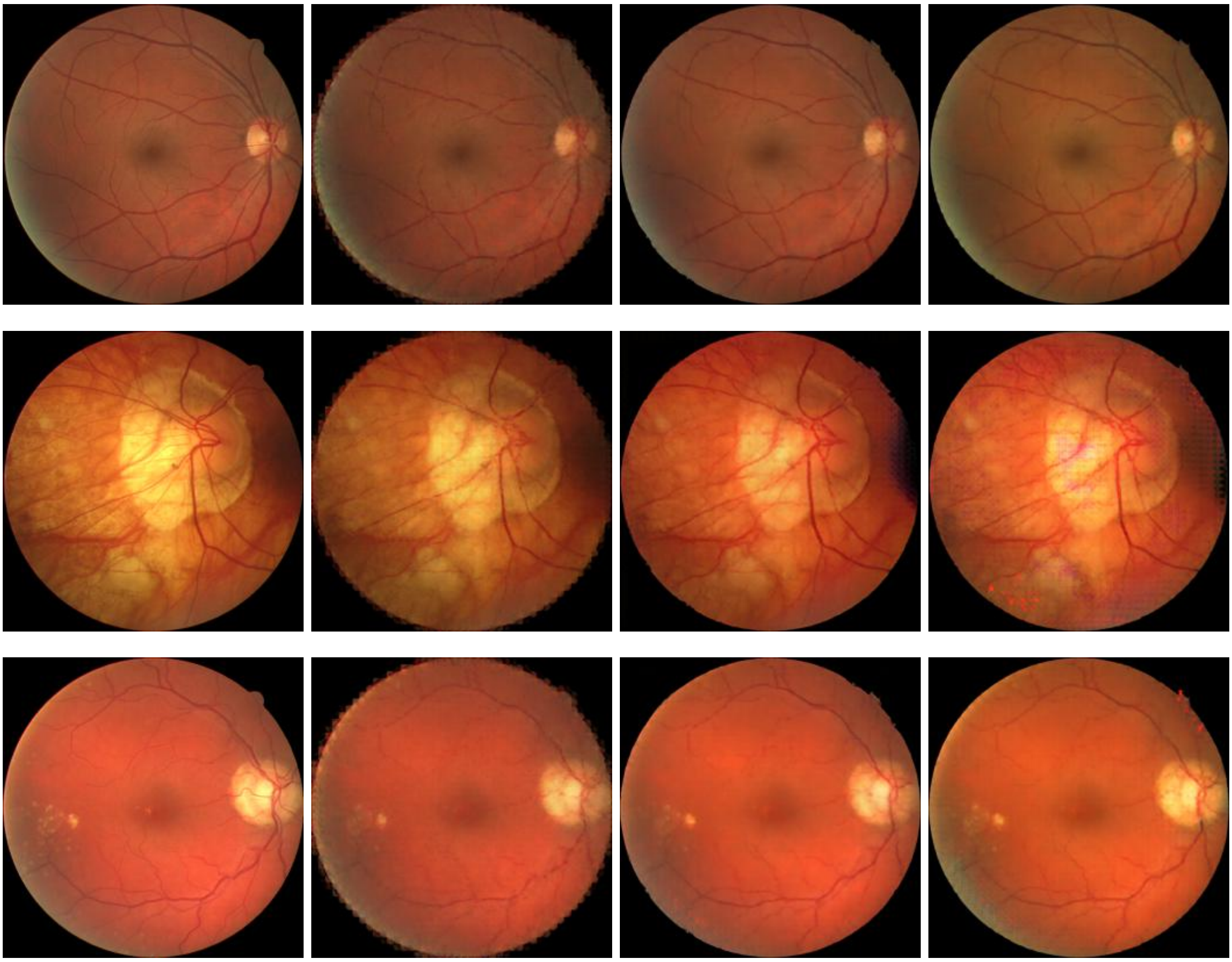


Figure 9. Example fully-reconstructed retinographies from the DRIVE test dataset. Each row corresponds to a normal retinography, a retinography with pathological optic disc, and a retinography bright retinal lesions, respectively. The first column is the original image. The subsequent columns correspond to PW-CE, GM-CE (CB) and GM-CE (CS) results, respectively.

level semantics from the eye fundus contents. Another example of this semantic enforcement of learnt retinal patterns is observed in the boundaries of the pathological optic disc in the second row example of Figure 9, and the detail images in the third row of Figure 11. In this case, we can appreciate the generation of artificial vascular structures surrounding the pathology. This is symptomatic of the anomalous complex pathological patterns of this image not being sufficiently represented in the training set. Due to this, the GM-CE networks are not able to deal with them accurately, when compared to the input image. Instead, the network opts to provide a semantically coherent reconstruction with respect to the learnt normal retinal patterns, which is that those anomalous boundaries corresponded to vessels. Thus the provided generated image is coherent.

Another indication of the more semantically coherent retinal reconstruction obtained by the GM-CE approaches can be found in the retinal background. The PW-CE approach presents a naive performance regarding the reconstructed colors, by simply limiting the inference to mimicking the surrounding context color. Nevertheless,

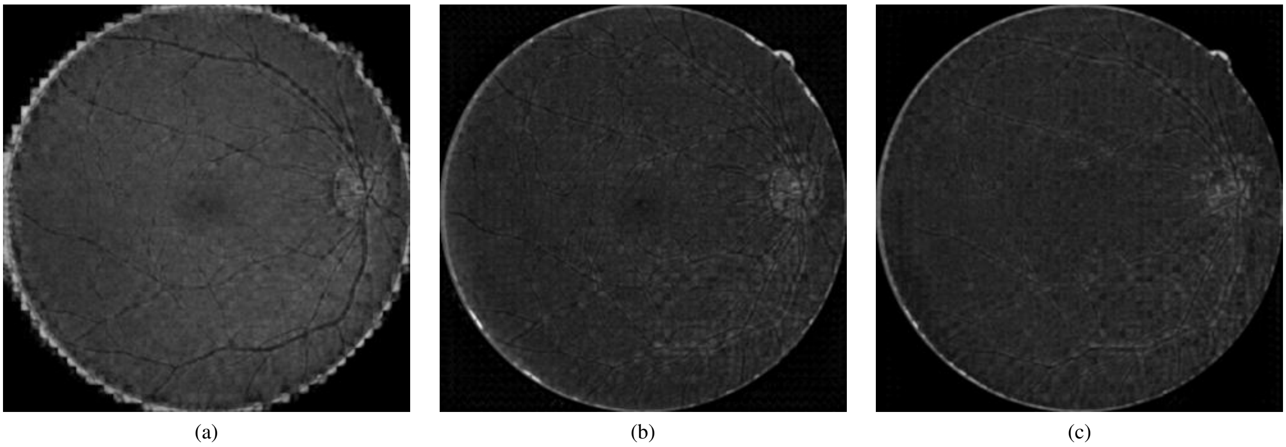


Figure 10. L2 error maps for the example in the first row of Figure 9. All the images are gamma corrected with $\gamma = 1/3$ for enhanced visibility, and normalized using the same transformation. (a) PW-CE. (b) GM-CE (CB). (c) GM-CE (CS).

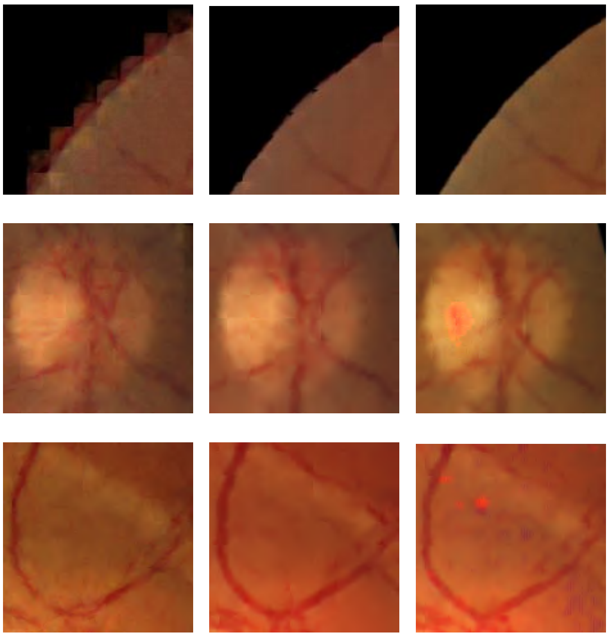


Figure 11. Reconstruction example details. The columns correspond to the PW-CE, GM-CE (CB) and GM-CE (CS) models, respectively.

both GM-CE approaches tend to artificially change the whole retina color patterns to those that were learnt during the training. This is specially visible in anomalous color appearances, like in the second row example of Figure 9. This is a symptom, once again, that these models were capable of learning, as normal, the specific retinal background color patterns that were presented during the training stage.

Finally, it is observed that the smallest pathological lesions in the third example of Figure 9 were removed by all the compared approaches, while the larger ones were preserved. This is an indication of the relevance of the omission mask scale, as already discussed in Section 2.2. It should be noticed that the experimentation with different omission scales is simplified for the case of GM-CE approaches, as this scale is neither fixed by the network architecture design nor requiring an *ad-hoc* dataset preprocessing.

4 CONCLUSION

This work presents different context encoder approaches using self-supervised learning taking advantage of a restricted domain as is frequent in medical image analysis. In particular, the different approaches face the analysis of the eye fundus using the widely spread and frequently used retinographies. Thus, a patch-wise approach, inspired in a previous proposal of general scope, is presented as well as two fully-convolutional global mask approaches, with a great potential of global context analysis given the recurrent eye fundus observation of this application.

Representative experiments were conducted to validate and demonstrate the suitability and potential of the proposed context encoder approaches in the restricted domain of ophthalmological images. In particular, in this issue, the approaches were capable of fully reconstructing the target retinographies, clearly representing the main structures of the eye fundus as the fovea, the optic disc or the retinal vessels tree. The presented results evidenced that the patch-wise approach was able to recognize and reconstruct the relevant local patterns as appeared in the input images. However, the global mask approaches, additionally, recognized more general patterns and reconstructed higher level structures with a higher global consistency, and according to the learnt semantics of the retinographies during the training.

We would like to emphasize that these self-supervised context encoder approaches present the potential of being used in different transfer learning or multi-task learning settings or to model normal retinographies for anomaly detection.

ACKNOWLEDGEMENTS

This work is supported by Instituto de Salud Carlos III, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project, and by Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the DPI2015-69948-R and RTI2018-095894-B-I00 research projects. The authors of this work also receive financial support from the ERDF and European Social Fund (ESF) of the EU, and Xunta de Galicia through Grupo de Referencia Competitiva, ref. ED431C 2016-047, and the predoctoral grant contract ref. ED481A-2017/328.

REFERENCES

[1] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, ‘Self-supervised learning for medical image analysis using image context restoration’, *Medical Image Analysis*, **58**, (2019).

[2] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J. Klein, ‘Feedback on a publicly distributed database: the messidor database’, *Image Analysis & Stereology*, **33**(3), 231–234, (2014).

[3] G. E. Hinton and R. R. Salakhutdinov, ‘Reducing the dimensionality of data with neural networks’, *Science 2006*, **313**, 504–507, (2006).

[4] L. F. H. Zimbrón, R. Z. Alvarado, L. O. De la Paz, R. V. Montoya, E. Zenteno, R. G. Cañizo, H. Q. Mercado, and R. G. Salinas, ‘Age-related macular degeneration: New paradigms for treatment and management of amd’, *Oxidative Medicine and Cellular Longevity*, **2018**, 14 pages, (2018).

[5] H. Harutyunyan, H. Khachatrian, D. Kale, and A. Galstyan, ‘Multitask learning and benchmarking with clinical time series data’, *Scientific Data*, **6**, (2019).

[6] A. S. Hervella, J. Rouco, J. Novo, and M. Ortega, ‘Retinal image understanding emerges from self-supervised multimodal reconstruction’, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 321–328, (2018).

[7] E. J. Duh, J. K. Sun, and A. W. Stitt, ‘Diabetic retinopathy: current understanding, mechanisms, and treatment strategies’, *JCI Insight*, (2017).

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘Generative adversarial nets’, *Neural Information Processing Systems (NeurIPS)*, **3**, (2014).

[9] S. Junji, L. Qiang, D. Appelbaum, and K. Doi, ‘Computer-aided diagnosis and artificial intelligence in clinical imaging’, *Seminars in nuclear medicine*, **41**(4), 449–462, (2011).

[10] A. K. Deb, S. Kaliaperumal, V. A. Rao, and S. Sengupta, ‘Relationship between systemic hypertension, perfusion pressure and glaucoma: A comparative study in an adult indian population’, *Indian J Ophthalmol*, **62**, 917–922, (2014).

[11] A. Mikołajczyk and M. Grochowski, ‘Data augmentation for improving deep learning in image classification problem’, *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 117–122, (2018).

[12] Y. Morales, R. Nuñez, J. Suarez, and C. O. T. Moreno, ‘Digital tool for detecting diabetic retinopathy in retinography image using gabor transform’, *Journal of Physics Conference Series*, **792**, (2017).

[13] S. N. Aakur and S. Sarkar, ‘A perceptual prediction framework for self supervised event segmentation’, *Computer Vision and Pattern Recognition (CVPR)*, 1197–1206, (2019).

[14] R. N. Weinreb, T. Aung, and F. A. Medeiros, ‘The pathophysiology and treatment of glaucoma’, *Journal of the American Medical Association*, **311**, 1901–1911, (2014).

[15] M. Noroozi and P. Favaro, ‘Unsupervised learning of visual representations by solving jigsaw puzzles’, *ECCV 2016*, (2016).

[16] D. P. Kingma and J. Ba, ‘Adam: A method for stochastic optimization’, *3rd International Conference for Learning Representations*, (2017).

[17] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, ‘Context encoders: Feature learning by inpainting’, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2536–2544, (2016).

[18] A. S. Hervella, J. Rouco, J. Novo, and M. Ortega, ‘Deep multimodal reconstruction of retinal images using paired and unpaired data’, *2019 International Joint Conference on Neural Networks (IJCNN)*, (2019).

[19] D. Shen, G. Wu, and H. Suk, ‘Deep learning in medical image analysis’, *Annual Review of Biomedical Engineering*, **19**, 221–248, (2017).

[20] J. J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, ‘Ridge based vessel segmentation in color images of the retina’, *IEEE Transactions on Medical Imaging*, **23**, 501–509, (2004).

[21] S. Sutradhar, J. Rouco, and M. Ortega, ‘Blind-spot network for image anomaly detection: A new approach to diabetic retinopathy screening’, *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN’19*, (2019).

[22] L. Torrey and J. Shavlik, ‘Handbook of research on machine learning applications and trends: Algorithms, methods and techniques’, *University of Wisconsin*, **Chapter 11**, (2009).

[23] Z. Wang, A. C. Bovik, R. Sheikh, and E. P. Simoncelli, ‘Image quality assessment: from error visibility to structural similarity’, *IEEE Transactions on Image Processing*, **13**(4), (2004).

[24] R. Zhang, P. Isola, and A. A. Efros, ‘Colorful image colorization’, *European Conference on Computer Vision (ECCV)*, 649–666, (2016).