Contents lists available at ScienceDirect

Applied Soft Computing

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

End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images

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

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

ARTICLE INFO

Article history:
Received 24 January 2021
Received in revised form 13 December 2021
Accepted 16 December 2021
Available online 24 December 2021

Keywords:
Multi-task learning
Medical imaging
Glaucoma diagnosis
Optic disc and cup segmentation
Cup-to-disc ratio
Deep learning
Retinal imaging

ABSTRACT

The automated analysis of eye fundus images is crucial towards facilitating the screening and early diagnosis of glaucoma. Nowadays, there are two common alternatives for the diagnosis of this disease using deep neural networks. One is the segmentation of the optic disc and cup followed by the morphological analysis of these structures. The other is to directly address the diagnosis as an image classification task. The segmentation approach presents the advantage of using pixel-level labels with precise morphological information for training. However, while this detailed training feedback is not available for the classification approach, in this case the image-level labels may allow the discovery of additional non-morphological cues that are also relevant for the diagnosis.

In this work, we propose a novel multi-task approach for the simultaneous classification of glaucoma and segmentation of the optic disc and cup. This approach can improve the overall performance by taking advantage of both pixel-level and image-level labels during the network training. Additionally, the segmentation maps that are predicted together with the diagnosis allow the extraction of relevant biomarkers such as the cup-to-disc ratio. The proposed methodology presents two relevant technical novelties. First, a network architecture for simultaneous segmentation and classification that increases the number of shared parameters between both tasks. Second, a multi-adaptive optimization strategy that ensures that both tasks contribute similarly to the parameter updates during training, avoiding the use of loss weighting hyperparameters.

To validate our proposal, an exhaustive experimentation was performed on the public REFUGE and DRISHTI-GS datasets. The results show that our proposal outperforms comparable multi-task baselines and is highly competitive with existing state-of-the-art approaches. Additionally, the provided ablation study shows that both the network architecture and the optimization approach are independently advantageous for multi-task learning.

1. Introduction

Glaucoma represents the leading cause of irreversible vision loss in the world [1]. This eye condition is characterized by an increased intra-ocular pressure that produces damage to different retinal tissues. However, it usually remains asymptomatic until there is noticeable vision loss, hence being typically diagnosed at advanced stages [2]. This motivates the development of automated methods that may facilitate the diagnosis at early stages, when the onset of vision loss can be prevented [2].

In order to diagnose glaucoma, ophthalmologists can perform visual field tests [2] and measure parameters such as the intra-ocular pressure or corneal thickness [3]. Additionally, important signs of the disease, such as optic disc deformations, peripapillary atrophy, or retinal nerve fiber layer defects, can be directly observed using color retinography, a widely available retinal imaging technique. Among these signs, the deformation of the optic disc has been the most widely studied for the early diagnosis of glaucoma [4]. The optic disc represents the head of the optic nerve and it can be divided into two regions: the optic cup and the neuroretinal rim. As reference, Fig. 1 depicts a representative example of retinography, including a detailed view of the optic disc region. An extensively studied change in the optic disc of glaucomatous eyes is the enlargement of the cup and the reduction of the rim [4].

The most common approach for the automated diagnosis of glaucoma is the segmentation of the optic disc and the optic cup followed by the extraction of relevant biomarkers [5]. For instance, the vertical Cup-to-Disc Ratio (CDR) has been extensively studied as a means of diagnosing and studying the progression of glaucoma [5]. However, in recent years, several works have

[*] Corresponding author at: Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain.
E-mail address: a.suarezh@udc.es (Á.S. Hervella).

https://doi.org/10.1016/j.asoc.2021.108347

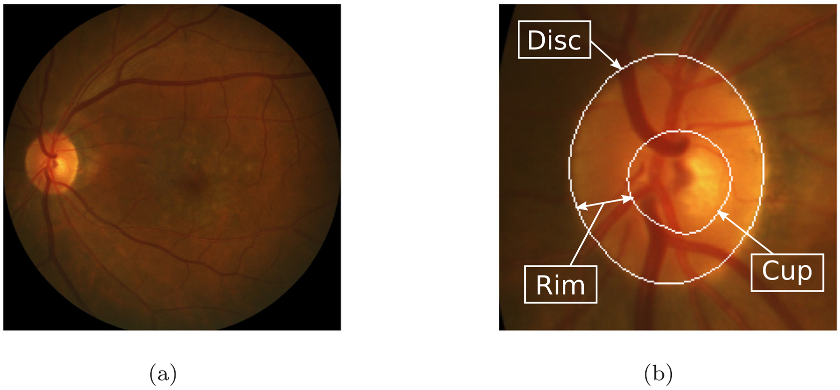


Fig. 1. (a) Retinography and (b) detailed view of the optic disc.

also explored the diagnosis of glaucoma as a classification task directly performed over the raw retinographies using Deep Neural Networks (DNNs) [6]. This approach can exploit additional cues besides the optic disc and cup morphology and typically offers a superior performance. However, it requires a larger amount of annotated training images and lacks the interpretability of the morphological biomarkers.

In this work, we propose a novel multi-task learning approach for the diagnosis of glaucoma. In particular, we propose to simultaneously learn the optic disc and cup segmentation and the glaucoma classification using the same DNN. This way, the segmentation can benefit from the high level representations learned from the image-level classification labels while the classification can exploit the cues and detailed feedback provided by the pixel-level segmentation annotations. In order to successfully apply this approach, we propose novel alternatives in different key areas regarding multi-task learning. First, regarding the network architecture, we propose an alternative to effectively perform both pixel-level and image-level predictions while sharing most of the learned representations. Second, regarding the joint optimization of both tasks, we propose a multi-adaptive optimization strategy that avoids the use and fine-tuning of task-balancing hyperparameters. Finally, the complete methodology as well as the different components that we propose are extensively validated on different public datasets presenting complementary clinical scenarios. In this regard, we analyze the performance of the proposed methodology for both glaucoma classification and optic disc and cup segmentation, including the extraction of a relevant biomarker such as the CDR.

The rest of the manuscript is structured as follows. Section 2 presents a discussion of related works regarding the automated analysis of glaucoma and multi-task learning. The proposed methodology is described in Section 3, including (Section 3.1) the proposed multi-task neural network as well as (Section 3.2) the training approach and proposed optimization strategy. Section 4 describes the experimentation setup, whereas the results of our experiments are presented and discussed in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

2.1. Glaucoma diagnosis

In the literature, numerous works have approached the automated diagnosis of glaucoma from color retinography. Most of the existing works are focused on the analysis of the optic disc and optic cup [4], although there are examples of methods focused on other structures or lesions as well, such as the retinal nerve fiber layer [7] or the possible peripapillary atrophy [8]. Regarding the different techniques that are used, existing approaches can be roughly divided into deep learning and classical/non-deep learning approaches [4,6]. A comprehensive analysis of classical approaches can be found in Sarhan et al. [4]. Regarding the deep learning-based approaches, a brief discussion of relevant works is presented hereafter.

The diagnosis of glaucoma has been addressed as an image classification task using DNNs in several works [6,9,10]. In these cases, it is common to use state-of-the-art network architectures that have already been demonstrated to be successful for image classification in other application domains [6]. Additionally, these networks are typically pre-trained on a large-scale annotated dataset of natural images, such as ImageNet [9–11], although the pre-training in an optic disc segmentation task was also explored [9]. The importance of the optic disc for the study of glaucoma is usually considered by restricting the input domain to this particular region of the image [10,11]. However, although this approach may facilitate the training of the network, it can potentially discard some valuable cues for the diagnosis of the disease. Moreover, it makes the methods reliant on an additional processing step for the localization or segmentation of the optic disc. For those reasons, in the proposed methodology, the whole unprocessed retinal image is used as input to the network. Then, the particular relevance of the optic disc region is considered by incorporating the segmentation of the optic disc and cup as a complementary task within the same neural network.

Another approach for the diagnosis of glaucoma using DNNs is the segmentation of the optic disc and cup followed by the extraction of relevant biomarkers, typically the CDR. In this case, it is also common to use state-of-the-art network architectures and pre-training the networks on large-scale annotated datasets of natural images [12–14]. Although there are also several examples of custom network architectures being proposed [14,15]. Additionally, the input domain is usually restricted by cropping the optic disc region [12,14–16], which facilitates the segmentation but requires the previous localization of the optic disc.

Given that the optic cup is an internal part of the optic disc, there is a spatial overlap between both regions in the image (see Fig. 1). Consequently, there exist two possible alternatives for formulating the segmentation task. One is the multi-label segmentation [12,15], where the task is split into two binary problems (disc vs. background and cup vs. background). The other, which is the one followed in this work, is the multi-class segmentation [14,16,17]. In this case, the network must predict the most likely of three mutually exclusive classes: background, neuroretinal rim, and optic cup. Then, the optic disc can be recovered by adding the cup and rim predictions. Alternatively, Jiang et al. [13] approach the segmentation as a minimal bounding box prediction. In this case, taking advantage of the expected shape for the optic disc and cup, the segmented regions are defined as the inscribed ellipses in the predicted bounding boxes. The expected morphology of the optic disc and cup is also exploited in other works by refining the network predictions in a post-processing stage [12,14–16], e.g., filling holes [12] or computing the convex hull [16]. However, the methodology proposed in this work avoids such assumptions about the expected morphology of the segmentation output. Thus, in our case, both the segmentation and classification are learned end-to-end, from the raw image to the final predictions, avoiding additional processing stages.

In contrast to other works that focus on the segmentation and perform the morphological analysis at a later stage, Zhao and Li [18] propose to directly estimate several morphological indices together with the segmentation in a multi-task setting. Then, the predicted indices can be used to elaborate a diagnosis or risk index for glaucoma. Nevertheless, in this case, the analysis is restricted to the morphological features of the optic disc. Differently, we propose a multi-task methodology that directly estimates the diagnosis together with the segmentation, hence integrating into the same network all the different features that may be useful for the study of glaucoma. This presents the potential of improving the predicted diagnosis as well as the predicted segmentations and the morphological biomarkers that may be extracted from them.

2.2. Multi-task learning

Several prior works have specifically focused on addressing the particular challenges of multi-task learning, such as the network architecture design and the optimization procedure. A comprehensive analysis of these aspects can be found in Vandenhende et al. [19].

Regarding the network architecture, most of the existing proposals are focused on the combination of multiple pixel-level prediction tasks, such as is e.g. semantic segmentation [19,20]. However, the combination of image-level and pixel-level prediction tasks, such as classification and segmentation, presents additional challenges due to the different output characteristics of both tasks. In this case, the most common approach in the literature is to apply those network designs that are known to be successful for single-task learning, i.e. encoder and encoder–decoder architectures for image-level and pixel-level prediction tasks, respectively. Thus, the typical approach is an encoder–decoder network where the pixel-level output is obtained from the decoder and the image-level output from the encoder [21,22]. However, this design limits the number of learned representations that can be shared between tasks. This motivates the network design that is proposed in this work, which aims at providing an increased number of shared representations between the tasks.

With regards to the optimization procedure, challenges arise due to the uneven back-propagated gradients that are provided by the different tasks [23]. In those cases, the training will be biased towards the task that provides stronger gradients, i.e. stronger supervisory signals. This issue is even more accentuated in cases of conflicting feedback among tasks, further penalizing those tasks with weaker back-propagated gradients. The common approach to face these issues is the use of loss weighting hyperparameters that aim at balancing the feedback provided by the training losses [19]. Usually, these hyperparameters are set empirically by the researchers, which requires extensive experimentation to obtain the best performance [24,25]. However, there are also several proposals to estimate the adequate hyperparameters during the network training, attending to different criteria [19,23]. For instance, Kendall et al. [26] propose to increase the weight of those tasks with lower inherent uncertainty, whereas Guo et al. [27] increases the weight of the tasks that are performing worse. In this vein, DWA [14] takes into account the decreasing rate of the individual task losses. In contrast to previous alternatives, in this work we propose a task-balancing approach that does not use loss weighting hyperparameters. Instead, following a multi-adaptive optimization, our proposal directly balances the individual parameter updates during training, regardless of the original balance among the supervisory signals.

Instead of balancing the training feedback, other proposals have focused on mitigating the conflicting supervision. This is usually addressed by directly manipulating the gradients such that the number of conflicting cases is reduced [28,29]. In that regard, we hypothesize that an adequate balance will also reduce the negative impact of the conflicting cases.

3. Methodology

The proposed multi-task methodology for classification and segmentation in the context of the glaucoma diagnosis is divided into two main areas: (1) the network architecture, and (2) the training and optimization strategy. The procedure and proposals for each of these areas are described in detail below.

3.1. Multi-task neural network

Regarding the network architecture, the aim of this work is to provide a viable and effective alternative for applying multi-task learning to the problem at hand. To that end, we first consider, as base network, the segmentation-guided architecture proposed in the work of Fu et al. [9]. This base architecture is an encoder–decoder network where the segmentation predictions are obtained from the decoder output and the classification predictions from the encoder output (Fig. 2(a)), which is a common setting in multi-task learning [21,22]. However, in this case, only part of the learned representations is shared between tasks. In this regard, we propose an alternative design based on the idea of sharing most of the learned representations, allowing to further exploit the complementary feedback during the training.

The proposed multi-task network, depicted in Fig. 2(b), mainly consists of an encoder–decoder structure where the predictions for both the segmentation and classification are derived from the decoder. Thus, both encoder and decoder are shared between tasks. In order to adequately generate the classification predictions, a classification head shaped as a mini-encoder is added after the main network. This classification head presents a minimal number of parameters and reduced capacity. Thus, the relevant representations must still be learned in the main network shared between tasks. A detailed diagram of the proposed multi-task network is depicted in Fig. 3. For the encoder–decoder part, we adopt the U-Net architecture [30], which is a well-proven and commonly used neural network [31,32]. Moreover, the encoder of U-Net is that of VGG-B [33], which also represents a well-proven baseline for image classification. The main characteristic of U-Net is the use of skip connections between the encoder and the decoder. These connections concatenate feature maps from the encoder with those of the same spatial resolution in the decoder. This way, the decoder not only receives the high level representations of the encoder output but a complete collection of multi-scale representations. This facilitates the prediction of pixel-level details in the segmentation task. In contrast to this, classification tasks typically benefit from the availability of high level and low spatial resolution representations near the network output, as in an encoder network. Motivated by this and following the idea of U-Net, the proposed multi-task network presents skip

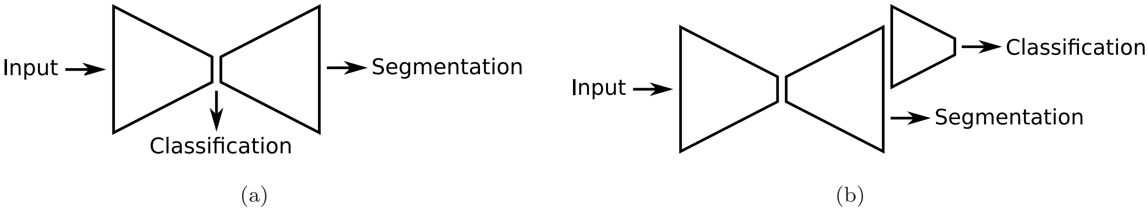


Fig. 2. Diagrams of multi-task neural networks. (a) Base network. (b) Proposed network.

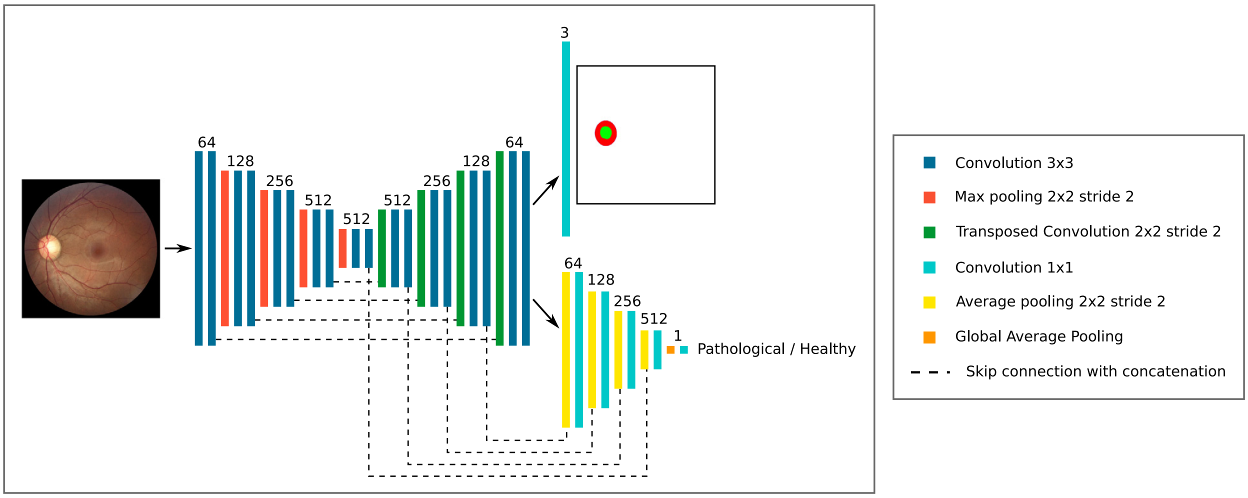

64
128
256
512
512
512
256
128
64
3
64
128
256
512
1
Pathological / Healthy
🟦 Convolution 3x3
🟥 Max pooling 2x2 stride 2
🟩 Transposed Convolution 2x2 stride 2
🟦 Convolution 1x1
🟨 Average pooling 2x2 stride 2
🟧 Global Average Pooling
- - Skip connection with concatenation

Fig. 3. Detailed diagram of the proposed multi-task neural network. The blocks of the main encoder and decoder follow the structure of U-Net [30] and VGG [33]. The numbers above the different blocks indicate the number of output channels, which are used as input to the next block. The concatenations in the classification head are performed after the downsampling (average pooling), whereas the concatenations in the decoder are performed after the upsampling (transposed convolution). Additionally, each convolutional layer, excluding the last ones, is followed by a ReLU activation function. The last layer before the segmentation output uses a softmax function whereas the last layer before the classification output uses a sigmoid.

connections between the decoder and the classification head. During the training, the network learns to combine these multi-scale representations in the most adequate way, as it does for the segmentation.

The objective of the classification head in the proposed multi-task network is to simultaneously perform an aggregation of multi-scale representations and a progressive reduction of the spatial dimensions. In order to achieve this with the minimum parameters, 1 × 1 convolutions are used for the aggregation of multiple channels and average pooling for the reduction of the spatial dimensions. As depicted in Fig. 3, the last 64 feature maps from U-Net are used as input to the classification head. These feature maps are downsampled and concatenated with those other of the same resolution in the decoder of U-Net. Then, a 1 × 1 convolution is applied. This process is repeated at 4 different spatial resolutions, matching those in the decoder of U-Net. Then, global average pooling is used to reduce the feature maps to a fixed size of 1 × 1 regardless of the input image size. The final prediction is obtained using a 1 × 1 convolution, which, in this case, is equivalent to a fully connected layer. Using this configuration, the classification head only accounts for a 2.89% of the network parameters. Finally, in order to adequately generate the network predictions, the last segmentation layer uses a multi-class softmax activation function whereas the last classification layer uses sigmoid.

3.2. Multi-task training

In the proposed multi-task approach, classification and segmentation are simultaneously trained to take advantage of the complementary feedback provided by the image-level and pixel-level labels. The training for each of these tasks is formulated according to the common approaches in the literature. In particular, the diagnosis of glaucoma is approached as a binary classification and the training is conducted using cross-entropy as the loss function. Thus, the loss for the classification task is defined as:

$$\mathcal{L}^C = -\, y_c \, log(p_c) - (1 - y_c) \, log(1 - p_c) \qquad (1)$$

where p_c denotes the classification output of the network and y_c the corresponding ground truth label.

The joint segmentation of the optic disc and cup is approached as a multi-class semantic segmentation [17]. In particular, the optic disc region is split into its two constituent parts, the optic cup and the neuroretinal rim (see Fig. 1). This way, the network is trained in a three-class segmentation problem (background, cup, and rim). Then, the optic disc can be recovered from the network output by adding the cup and rim predictions. The training is conducted using cross-entropy as the loss function. Thus, the loss for the segmentation task is defined as:

$$\mathcal{L}^S = -\frac{1}{N} \sum_{n}^{N} \sum_{k}^{K} y_s^{k,n} \, log(p_s^{k,n}) \qquad (2)$$

where $\mathbf{p_s}$ denotes the segmentation output of the network, $\mathbf{y_s}$ the corresponding ground truth label, N the number of pixels in the image, and K the number of classes, which is 3 in this case.

Finally, the multi-task training is conducted by simultaneously optimizing the classification and segmentation losses. In general, the simultaneous optimization of complementary tasks represents a complex issue due to different factors. For instance, the involved tasks may present a different level of difficulty, resulting in different learning dynamics. Also, the use of different loss functions or task-specific layers can affect the relative magnitude of gradients back-propagated to the shared layers of the network. These situations can make that one task prevail over the other, resulting in an unbalanced training and not successfully exploiting the complementary multi-task feedback. To overcome these challenges, we propose a novel multi-adaptive optimization approach that aims at providing a well-balanced training between tasks without the use of loss weighting hyperparameters.

Nowadays, adaptive gradient-based optimization algorithms, such as Adam [34] or RMSprop [35], are commonly used for the training of DNNs. These adaptive algorithms provide an online per-parameter tuning of the learning rate, resulting in an smaller effective learning rate for frequently updated parameters and a larger one for the most infrequent. This is performed by dividing the global learning rate by a function of past gradients for each parameter. Thus, as reference, the effective per-parameter learning rate is defined as:

$$\eta^t_{\theta_i} = \frac{\eta}{f(g^t_{\theta_i}, g^{t-1}_{\theta_i}, \ldots, g^0_{\theta_i})} \tag{3}$$

where η denotes the global learning rate, $\eta^t_{\theta_i}$ the effective learning rate for parameter θ_i at time t, and $g^t_{\theta_i}$ the gradient component for parameter θ_i at time t. The form of the function f represents the main difference among the different adaptive algorithms that have been proposed in the literature.

Considering that the parameter updates are typically computed as the product of the effective learning rates by the gradients, the adaptive algorithms implicitly provide a normalization of the gradients at the parameter level. Particularly, the gradient component for each parameter is always divided by a function of accumulated past gradients at that same parameter. This implicit normalization can be taken advantage of in multi-task learning for ensuring a balanced training without using any additional hyperparameter. In the proposed approach, this is achieved by decoupling the gradients of both tasks and computing task-specific per-parameter learning rates, only considering the past gradients of each specific task. Thus, the gradients of each task are normalized independently, resulting in a balanced contribution of the different tasks to the final parameter updates. In the case of the classification/segmentation multi-task training, the task-specific effective learning rates for a parameter θ_i at time t are obtained as:

$$\eta^t_{\theta_{i,C}} = \frac{\eta}{f(g^t_{\theta_{i,C}}, g^{t-1}_{\theta_{i,C}}, \ldots, g^0_{\theta_{i,C}})} \tag{4}$$

$$\eta^t_{\theta_{i,S}} = \frac{\eta}{f(g^t_{\theta_{i,S}}, g^{t-1}_{\theta_{i,S}}, \ldots, g^0_{\theta_{i,S}})} \tag{5}$$

where $\eta^t_{\theta_{i,C}}$ and $\eta^t_{\theta_{i,S}}$ denote the effective learning rates for the classification and segmentation tasks, respectively, and $g^t_{\theta_{i,C}}$ and $g^t_{\theta_{i,S}}$ the gradient components for the classification and segmentation tasks at time t, respectively. This results in normalized task-specific parameters updates that avoid the imbalance due to the potentially different magnitude of the gradients between tasks. Also, the task-specific effective learning rates could arguably provide additional benefits in cases of different learning speeds among tasks. Finally, the global parameters update is performed as:

$$\theta^{t+1}_i = \theta^t_i + \Delta\theta^t_{i,C}(\eta^t_{\theta_{i,C}}) + \Delta\theta^t_{i,S}(\eta^t_{\theta_{i,S}}) \tag{6}$$

where $\Delta\theta^t_{i,C}$ and $\Delta\theta^t_{i,S}$ denote the parameter updates due to the classification and segmentation tasks, respectively. These parameter updates, which depend on the task-specific effective learning rates, are obtained by applying the particular formulation of the desired adaptive algorithm.

Fig. 4 depicts a flowchart for the application of the multi-adaptive optimization. In practice, the proposed approach can be applied by using two independent instances of the desired adaptive optimization algorithm, one for each task, and following the procedure depicted in Fig. 4(b). In this work, we use Adam [34] as optimization algorithm, which is commonly used for both segmentation and classification tasks [10,36]. In the case of Adam, the function f in Eqs. (3), (4), and (5) is the square root of the exponential moving average of squared gradients. Additional details of the algorithm are described in the work of Kingma and Ba [34]. Regarding the computational requirements, the proposed approach requires additional operations to compute the task-specific gradients and parameter updates. Particularly, whereas the number of operations in the forward pass is kept the same, the number of operations in the backward pass is duplicated. Additionally, in terms of memory footprint, the proposed approach requires to keep in memory an additional set of values, including the gradients and the state of the function f for each task. In practice, the exact increase in execution time and memory footprint depends on the particular experimental setup being used. For the experiments in this work, this information is detailed in Section 4.2.

4. Experimental setup

4.1. Datasets

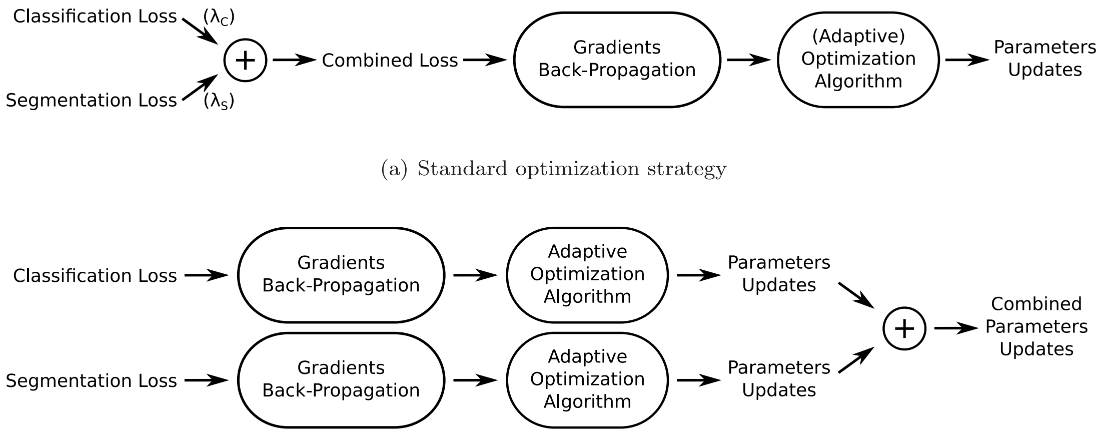
For the experiments of this work, we used the public datasets REFUGE [32] and DRISHTI-GS [37]. The first dataset is named after the REFUGE Challenge [32] and consists of 1200 annotated retinographies, of which 121 correspond to glaucomatous eyes. The retinographies in this dataset are centered on the macula and present a size of 1634 × 1634 or 2124 × 2056 pixels. The default split of the dataset consists of 400 images for test and 800 for training and validation.

The DRISHTI-GS dataset consists of 101 annotated retinographies, of which 70 correspond to glaucomatous eyes. The retinographies in this dataset are centered on the optic disc and present a size of 2896 × 1944 pixels. The default split of the dataset consists of 51 images for test and 50 for training and validation.

These two datasets provide complementary clinical scenarios, representing the two most common configurations regarding the capture of the images (centered at the macula or the optic disc). This allows for a more robust evaluation of the proposed methodology.

4.2. Training details

For the experiments in this work, the networks are pre-trained using the self-supervised multimodal reconstruction approach that was proposed in [36]. Instead of requiring manually annotated data, this self-supervised pre-training relies on the availability of paired multimodal retinal images, namely retinography and fluorescein angiography. Additionally, these image pairs do not need to be related to the specific application at hand (e.g. glaucoma diagnosis), which facilitates the gathering and reuse of the same small set of multimodal images for any target application in retinography [36,38]. In particular, we use the publicly available multimodal dataset provided by Alipour et al. [39], consisting of 59 image pairs. The pre-training methodology is described in detail in [40]. The multimodal reconstruction consists in the prediction of fluorescein angiography from retinography. This represents a pixel-level prediction task, such as the segmentation, hence the pre-training is applied to the main encoder and decoder in the networks. Each retinography–angiography pair in the dataset is registered using the methodology described


Classification Loss
(λC)
Segmentation Loss
(λS)
+
Combined Loss
Gradients Back-Propagation
(Adaptive) Optimization Algorithm
Parameters Updates
Classification Loss
Gradients Back-Propagation
Adaptive Optimization Algorithm
Parameters Updates
Segmentation Loss
Gradients Back-Propagation
Adaptive Optimization Algorithm
Parameters Updates
+
Combined Parameters Updates

(a) Standard optimization strategy

(b) Multi-adaptive optimization strategy

Fig. 4. Flowchart for the application of the multi-adaptive approach in the classification/segmentation multi-task training. For reference, a flowchart of the standard approach is also included. λ_C and λ_S in (a) indicate optional weighting hyperparameters.

in [40]. Then, the network is trained using the negative Structural Similarity as loss function [40]. The optimization is performed with the Adam algorithm [34] with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and batch size of 1 image. The learning rate is initially set to $\alpha = 1e - 4$ and is reduced by a factor of 10 each time the validation loss ceases to improve for 50 epochs. Finally, early stopping is applied with patience of 100 epochs. 9 of the 50 image pairs are used as validation data.

For the multi-task training of the target tasks, we use the Adam algorithm [34] with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and batch size of 1 image. In order to apply a learning rate schedule and early stopping, 25% of the training data is used as validation subset. In the case of the REFUGE dataset, the initial learning rate is set to $\alpha = 1e - 5$. For each task, this learning rate is reduced to $\alpha = 1e - 6$ after 10 epochs without improvement in the validation data. Finally, after 20 epochs without any improvement for both tasks, the training is stopped. These values are set by taking as reference previous works in the literature [36] as well as monitoring the learning curves. In the case of DRISHTI-GS, we follow a common approach in the literature that consists in fine-tuning a network previously trained in a larger dataset [15,16]. Thus, the experiments for DRISHTI-GS are performed by fine-tuning the networks previously trained on the REFUGE dataset. The fine-tuning is performed on all the training data at a constant learning rate of $\alpha = 1e{-}6$ during 1000 epochs, which was deemed enough for convergence in all the cases. The adequate number of training epochs was empirically set by performing some initial experiments with 25% of the training data as validation subset.

In all the experiments, the images are rescaled so that the width of the visible retina in the images is always 720 pixels. In order to avoid overfitting, we apply data augmentation consisting of random spatial transformations of the input images and the target segmentation maps. In particular, we use affine transformations including scaling and rotation components, which are commonly used in medical imaging [41,42]. We use a scaling factor of 2^S where S is uniformly distributed in the range $[-0.25, 0.25]$. Likewise, the rotations are uniformly distributed in the range $[-45°, 45°]$.

The proposed methodology and the experiments in this work were implemented in Python 3, using the open source framework PyTorch for the parts specific to deep learning. In particular, we used PyTorch 1.1.0 with CUDA 10.0. The training of the neural networks was performed in GPU using an NVIDIA GTX 1070 with memory size of 8 GB. Regarding the computational requirements, using the experimental setup herein described, the proposed multi-adaptive optimization strategy increases the memory footprint from 4051 MB to 4473 MB and the execution time during training from 398 ms to 526 ms per iteration on average.

4.3. Evaluation methods

The evaluation is performed following the common methods in the related literature [32]. In particular, the classification task is evaluated using Receiver Operator Characteristic (ROC) curves, which plot the sensitivity and specificity for different decision thresholds. Additionally, we also compute the Area Under Curve (AUC) for ROC, which is commonly used to summarize the performance into a single value.

The segmentation task is evaluated independently for the optic disc and the optic cup. For this purpose, the predicted optic disc is recovered from the optic cup and the neuroretinal rim predictions described in Section 3.2. Then, each segmentation is evaluated using Precision–Recall curves and the Dice coefficient. The Precision–Recall analysis allows the study of the performance at different decision thresholds whereas the Dice coefficient is typically computed for a single threshold. In this regard, the binary prediction for the Dice evaluation is directly obtained by assigning to each pixel the class with the highest likelihood in the network output.

Additionally, the extraction of morphological biomarkers from the optic disc and cup segmentation is also evaluated. This complementary evaluation is based on the vertical Cup-to-Disc Ratio (CDR), which is a broadly extended biomarker for the diagnosis of glaucoma. Similarly to the Dice coefficient, the CDR is computed for a particular threshold only. The CDR is defined as:

$$CDR = \frac{OC_{height}}{OD_{height}} \tag{7}$$

where OC_{height} and OD_{height} denote the height of the optic cup and optic disc, respectively. To perform the evaluation, the CDR is measured in both the predicted segmentations and the ground truth annotations. Then, the CDR error, δ_{CDR}, is measured as the absolute difference between the predicted and ground truth CDR.

The Dice coefficient and the δ_{CDR} are computed for each image and then the mean and standard deviation of the measures are reported. Additionally, given the expected high inter-dataset variability regarding these measures [32], a Wilcoxon signed-rank test is used to evaluate whether the differences between the distributions are statistically significant [43].

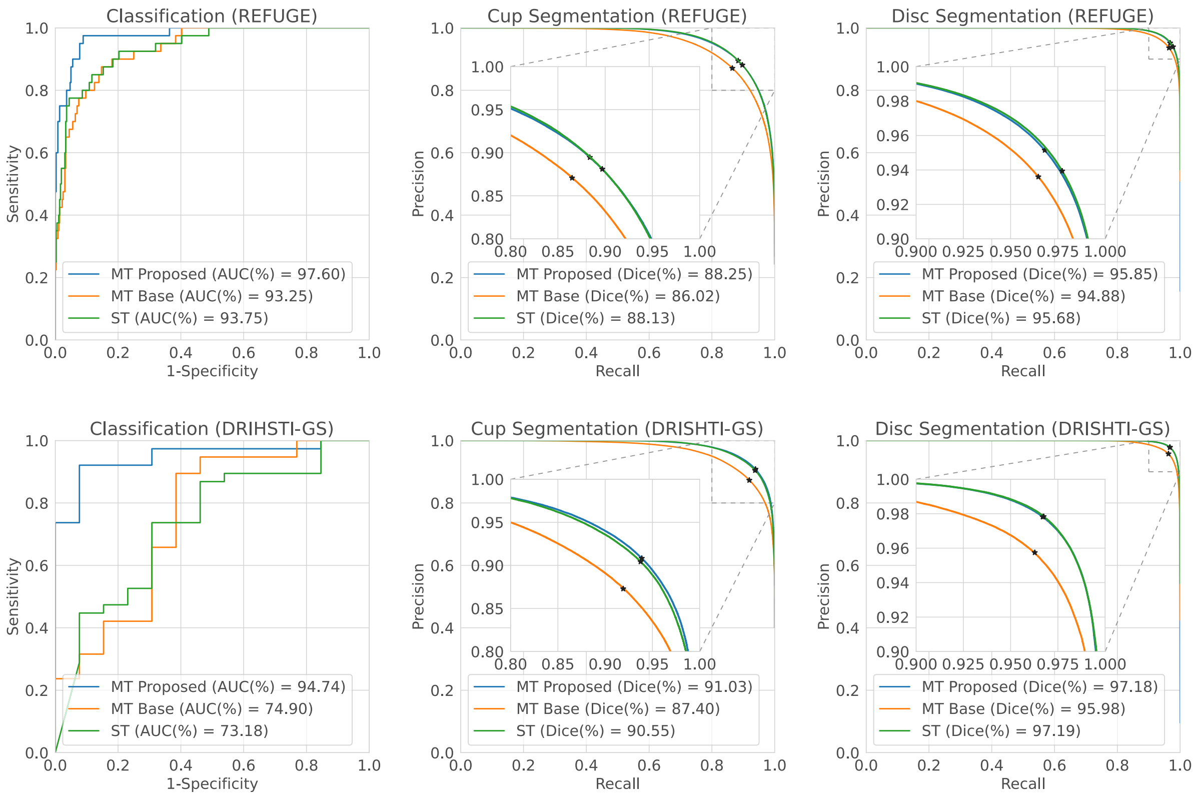


Fig. 5. Quantitative results and comparison of different methods for the (1st row) REFUGE and (2nd row) DRISHTI-GS datasets. MT denotes multi-task and ST single-task. In the segmentation results, black markers indicate the operating point at which the Dice coefficient is computed.

4.4. Alternative approaches in the experimentation

In order to comprehensively evaluate the proposed methodology, we perform several experiments including common alternative approaches. In particular, the experiments include the following alternatives:

- MT Proposed: The proposed multi-task methodology consisting of the Proposed network (Fig. 3) and the Multi-Adaptive (M-Ada) optimization (Fig. 4(b))
- MT Base: The multi-task baseline consisting of a Standard network (Fig. 2(a)) and the Standard optimization (Fig. 4(a)). This approach employs loss weighting hyperparameters for balancing the tasks. The optimal values are empirically found by grid search in the REFUGE dataset. In particular, the average performance on the validation set is used as selection criteria.
- ST: The single-task baseline consisting of an encoder network for classification and an encoder–decoder network for segmentation.

Additionally, we perform an ablation study considering all the possible combinations of network architectures and optimization strategies. In this case, we also tested two variants of the Standard optimization:

- S-GS: The optimal loss weighting hyperparameters found by grid search in MT Base.
- S-EW: Equal weighting of the two tasks. This is a more naive variant that does not require additional experimentation. Thus, in contrast to S-GS, presents a reduced training budget similar to that of M-Ada.

All the experiments are performed using the same encoder and decoder components described in Section 3.1.

5. Results

5.1. Experimental results

Fig. 5 and Table 1 depict the main results of our experiments in the REFUGE and DRISHTI-GS datasets. In general, it is observed that MT Proposed offers the best overall performance in both datasets. This demonstrates that MT Proposed successfully leverages the complementary multi-task feedback provided during the training. Meanwhile, ST results in the second best overall alternative. This shows that the use of specialized networks for each particular task can outperform a multi-task setting.

In comparison to MT Base, the greatest improvement when applying MT Proposed is obtained for the classification task. Additionally, regarding the segmentation task, the greatest improvement is obtained for the optic cup. Both these results indicate that the harder problems benefit more from the MT Proposed approach. In the first case, it must be noticed that the image-level annotations provide significantly less feedback than the pixel-level counterparts for training a DNN. Thus, for the same number of annotated images, the classification task is expected to be a harder problem. In the second case, the boundary of the optic cup is significantly less defined than the outer boundary of the optic disc. Additionally, in comparison to the optic disc, the optic cup morphology is more affected in the glaucomatous eyes, which increases the variability of this structure. For these reasons, the segmentation of the optic cup is a harder problem, which is also reflected in the obtained results.

Finally, considering the datasets, the greatest improvement when applying MT Proposed is obtained for DRISHTI-GS. Similarly to the previous analysis, we argue that this is a consequence of the DRISHTI-GS being a harder scenario, especially for the classification task. This is mainly due to the limited number of training

Table 1
Quantitative results and comparison of different methods for the REFUGE and DRISHTI-GS datasets. MT denotes multi-task and ST single-task. The best result for each evaluation metric is always highlighted in bold. For cup segmentation, disc segmentation, and cup-to-disc ratio, asterisks denote whether the difference with respect to the best result is statistical significant (* denotes p value < 0.05; ** p value < 0.01; and *** p value < 0.001).

Method	Classification AUCROC (%)	Cup segmentation Dice (%)	Disc segmentation Dice (%)	Cup-to-disc ratio δ_{CDR}
REFUGE				
MT Proposed	**97.60**	**88.25** ± **5.96**	**95.85** ± **1.92**	0.0373 ± 0.0313
MT Base	93.25	$86.02 \pm 7.94^{***}$	$94.88 \pm 2.43^{***}$	$0.0438 \pm 0.0394^{***}$
ST	93.75	88.13 ± 5.80	$95.68 \pm 1.91^{***}$	**0.0367** ± **0.0312**
DRISHTI-GS				
MT Proposed	**94.74**	**91.03** ± **6.08**	97.18 ± 1.37	**0.0413** ± **0.0481**
MT Base	74.90	$87.40 \pm 9.11^{***}$	$95.98 \pm 2.70^{***}$	$0.0747 \pm 0.0747^{**}$
ST	73.18	90.55 ± 6.73	**97.19** ± **1.70**	0.0447 ± 0.0431

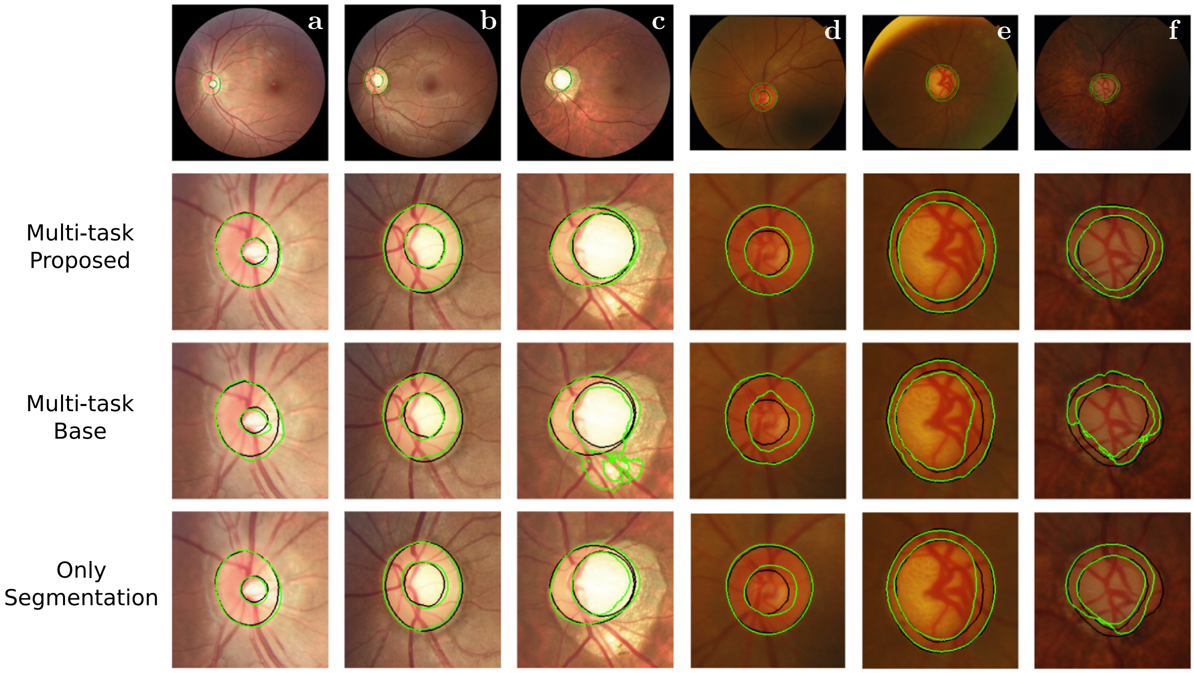


Fig. 6. Examples of predicted optic disc and cup regions for images of the (a,b,c) REFUGE and (d,e,f) DRISHTI-GS test sets. The first row depicts the input images and each of the next rows depicts the corresponding results for a different method. For a better appreciation, the results are depicted on a cropped square around the optic disc region. However, all the methods generate full-sized predictions. Green lines indicate the boundaries of the predicted regions whereas black lines indicate the boundaries of the ground truth regions.

images and the significantly higher proportion of pathological samples in this dataset. In this regard, it must be noticed that, as explained in Section 4.2, the results in DRISHTI-GS were obtained after fine-tuning the networks previously trained in the REFUGE dataset. However, there is a significant domain gap between datasets due to the different spatial positioning of the retina and the general visual appearance. In this case, MT Proposed was able to better adapt the previously acquired knowledge to the new scenario with limited annotations.

Fig. 6 depicts examples of predicted optic disc and cup segmentations from the test sets of REFUGE and DRISHTI-GS. The examples show that MT Proposed and ST produce the most accurate results across the different scenarios. The performance of the different approaches is closer in the scenarios that, a priori, are less complex. This is the case of examples (a) and (b), which correspond to healthy retinas and display adequate illumination. In contrast, example (c), despite having a similar illumination, presents significant peripapillary atrophy, a lesion that changes the general appearance of the optic disc region. Additionally, there is a limited number of samples of this kind of lesion in the training data. Another challenging case is that of example (f). In this example, the optic disc boundary seems well-defined but the visual characteristics of the image are very different from the other examples. In these more challenging scenarios, MT Proposed clearly outperforms MT Base.

5.2. Ablation study of the proposed methodology

The results of the ablation study on the REFUGE and DRISHTI-GS datasets are depicted in Fig. 7 and Table 2. First of all, the results show that the best performance is always achieved by MT Proposed, i.e the M-Ada optimization and the Proposed network applied together. However, the individual applications of these two components does not always produce an overall improvement with respect to the baseline approach. On the one hand, the addition of the M-Ada optimization always improves or, at least, keeps a similar performance with respect to the Standard counterpart. However, on the other hand, the addition of the Proposed network architecture does not always provide an improvement. Particularly, when using the Standard optimization strategies, the Base network achieves a better segmentation performance in several cases. Nevertheless, the Proposed network is always

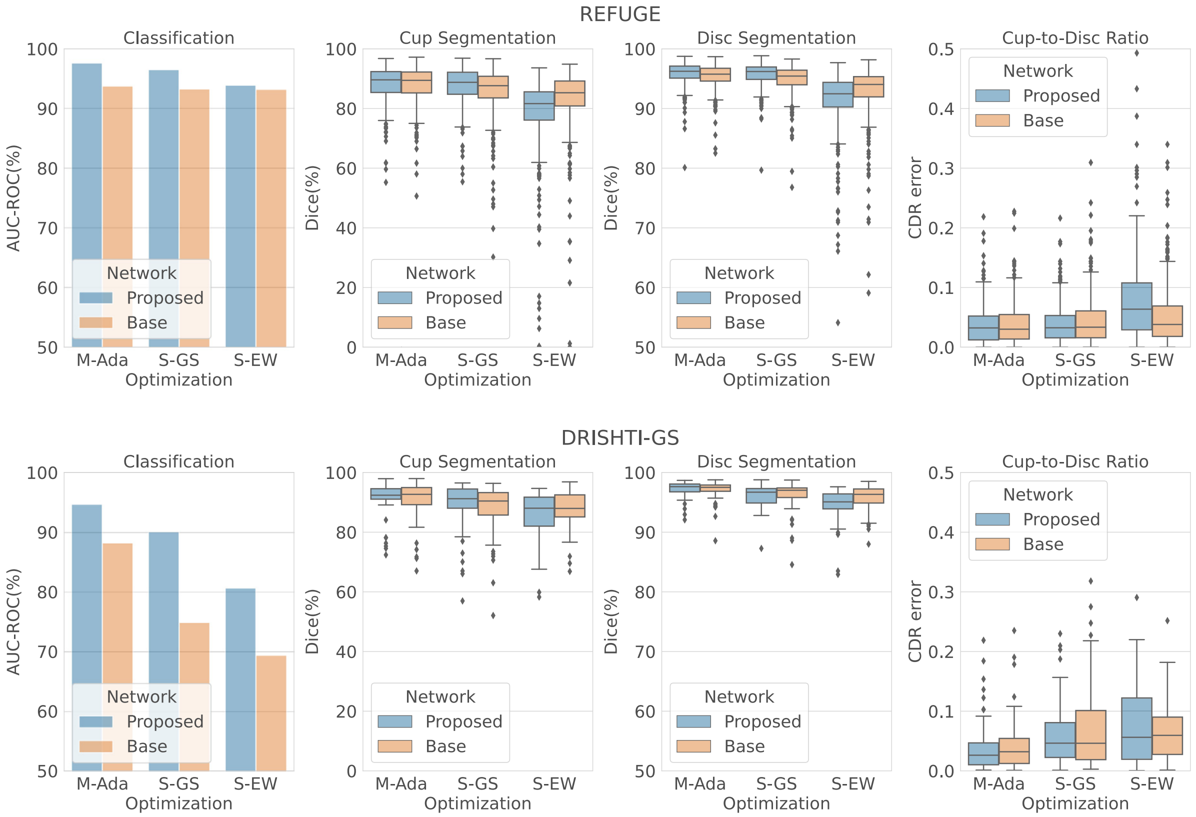


Fig. 7. Ablation study of the proposed methodology in the (1st row) REFUGE and (2nd row) DRISHTI-GS datasets. M-Ada denotes Multi-Adaptive, S-GS denotes Standard with Grid Search, and S-EW denotes Standard with Equal Weights.

Table 2
Ablation study of the proposed methodology in the (1st row) REFUGE and (2nd row) DRISHTI-GS datasets. M-Ada denotes Multi-Adaptive, S-GS denotes Standard with Grid Search, and S-EW denotes Standard with Equal Weights. The best result for each evaluation metric is always highlighted in bold. For cup segmentation, disc segmentation, and cup-to-disc ratio, asterisks denote whether the difference with respect to the best result is statistical significant (* denotes p value < 0.05; ** p value < 0.01; and *** p value < 0.001).

Optimization	Network	Classification AUCROC (%)	Cup segmentation Dice (%)	Disc segmentation Dice (%)	Cup-to-disc ratio δ_{CDR}
REFUGE					
Proposed (M-Ada)	Proposed Base	**97.60** 93.75	**88.25** ± **5.96** 87.94 ± 6.05***	**95.85** ± **1.92** 95.39 ± 2.04***	**0.0373** ± **0.0313** 0.0381 ± 0.0332
Standard (S-GS)	Proposed Base	96.50 93.25	87.68 ± 6.05*** 86.02 ± 7.94***	95.72 ± 1.88*** 94.88 ± 2.43***	0.0392 ± 0.0313* 0.0438 ± 0.0394**
Standard (S-EW)	Proposed Base	93.87 93.18	78.11 ± 14.49*** 82.63 ± 12.4***	91.44 ± 4.91*** 93.0 ± 4.32***	0.0778 ± 0.0664*** 0.0517 ± 0.0491***
DRISHTI-GS					
Proposed (M-Ada)	Proposed Base	**94.74** 88.26	**91.03** ± **6.08** 90.31 ± 7.05	**97.18** ± **1.37** 96.95 ± 1.73	**0.0413** ± **0.0481** 0.0429 ± 0.048
Standard (S-GS)	Proposed Base	90.08 74.90	88.7 ± 8.56*** 87.4 ± 9.11***	96.1 ± 2.01*** 95.98 ± 2.70***	0.0622 ± 0.0554*** 0.0747 ± 0.0747***
Standard (S-EW)	Proposed Base	80.67 69.43	85.15 ± 8.81*** 87.21 ± 6.69***	94.58 ± 2.94*** 95.68 ± 2.20***	0.0794 ± 0.0691*** 0.0665 ± 0.0507***

the best alternative for the classification task, regardless of the optimization approach.

In order to better analyze the differences between architectures and their interaction with the optimization approaches, Fig. 8 and Table 3 show the effect of varying the loss weighting hyperparameters of the Standard optimization in the performance of each network. It is observed that the Proposed network is able to reach a better performance, especially in the classification task. However, this network is also significantly more sensitive to the balance between tasks. For instance, the segmentation performance degrades drastically when the weight of the classification loss is increased. In that regard, the best segmentation results are achieved at a very low weight for the classification loss. However, the best classification results are also achieved at a low weight for this task. Thus, the Proposed network is able to simultaneously achieve a successful performance in the classification and the segmentation. Additionally, these results indicate that the advantage of the Proposed network in the classification is not due to the mere increase of layers that are available for that task, but to the increased availability of relevant representations that were successfully learned for the segmentation. In that regard, when combined with an adequate balance between tasks, the Proposed network allows to further

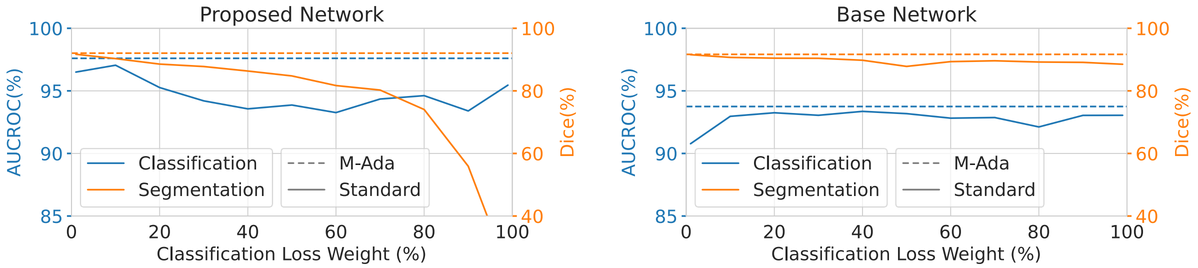


Fig. 8. Effect of the loss weighting hyperparameters in the classification and segmentation performance when using a Standard optimization strategy. The performance when using M-Ada (without loss weighting hyperparameters) is included as a reference point. The segmentation results correspond to the average performance between the optic disc and the optic cup.

Table 3
Effect of the loss weighting hyperparameters in the classification and segmentation performance when using a Standard optimization strategy. The performance using M-Ada is included as a reference. The segmentation results correspond to the average performance between the optic disc and the optic cup. Bold denotes the best results overall for each task and network. Underline denotes the best results of the Standard approach for each task and network.

Task	Metric	Standard at a given classification loss weight											M-Ada
		1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%	
Proposed network													
Classification	AUCROC (%)	96.50	97.05	95.26	94.20	93.56	93.87	93.26	94.35	94.62	93.40	95.46	**97.60**
Segmentation	Avg. Dice (%)	91.70	90.31	88.55	87.78	86.34	84.77	81.69	80.26	74.04	55.92	22.65	**92.05**
Base network													
Classification	AUCROC (%)	90.78	92.97	93.25	93.05	93.36	93.18	92.82	92.87	92.12	93.04	93.05	**93.75**
Segmentation	Avg. Dice (%)	91.56	90.69	90.45	90.41	89.79	87.82	89.34	89.61	89.22	89.11	88.50	**91.66**

Table 4
State-of-the-art comparison for the REFUGE dataset.

Methods	Classification AUCROC (%)	Cup segmentation Avg. Dice (%)	Disc segmentation Avg. Dice (%)	CDR Avg. δ_{CDR}
Wang et al. [12] (CUHKMED)[a]	96.44	88.26	**96.02**	0.0450
Orlando et al. [32] (VRT)	**98.85**	86.00	95.32	0.0525
Orlando et al. [32] (Masker)	95.24	**88.37**	94.64	0.0414
Orlando et al. [32] (SDSAIRC)	98.17	83.15	94.36	0.0674
Orlando et al. [32] (BUCT)	93.48	87.28	95.25	0.0456
Ours (Multi-task Proposed)	97.60	88.25	95.85	**0.0373**

[a] The classification results are reported in Orlando et al. [32].

take advantage of the multi-task training, providing an overall better performance than the baseline alternative.

5.3. State-of-the-art comparison

In this section, we provide a comparison of the proposed multi-task approach against relevant works in the literature. In this regard, it must be noticed that our proposal is the first to successfully take advantage of the multi-task learning paradigm for segmentation and classification in the context of glaucoma diagnosis. Thus, the state-of-the-art works that are included in the comparison either are focused on one of the tasks or address both tasks but applying different specific approaches.

The comparison for the REFUGE dataset is depicted in Table 4. In the comparison, we include the three top performing teams for classification and segmentation from the REFUGE challenge [32]. Nevertheless, it must be considered that, in general, these methods are finely tuned to obtain the best performance in the challenge. Thus, it is common the use of different ad-hoc processing stages and network ensembles [32]. In contrast, we propose an end-to-end multi-task learning approach without any ad-hoc processing. Regarding the results in Table 4, it is observed that our multi-task approach offers a very competitive performance for both the classification and the segmentation tasks. In this regard, there are a pair of methods that produce better results in classification but, coincidentally, the corresponding segmentation methods have a worse performance, especially for the optic cup.

Regarding the segmentation, our approach practically equals the best results, for both the optic cup and the optic disc. Additionally, our approach provides the lowest mean CDR error, which indicates the best clinical interpretability of the predicted segmentations by means of the CDR. Considering these results, our approach is arguably the most well-balanced between classification and segmentation. In this sense, it can be seen in Table 4 that only one team in the challenge, CUHKMED, is placed top three for both classification and segmentation.

The comparison for the DRISHTI-GS dataset is depicted in Table 5. In this case, most of the works focus only on the classification or the segmentation. In this regard, our approach is again the only one consisting of a single neural network that is trained end-to-end for both tasks. It is observed that our multi-task approach produces the best results for classification and a very competitive performance for segmentation, including the best results for the optic cup. Additionally, in comparison with our method, the state-of-the-art works that obtain better results for the optic disc present worse results for the optic cup. Hence, our approach provides a more balanced performance regarding the segmentation.

Regarding the CDR, our approach provides again the lowest δ_{CDR}, which indicates the best clinical interpretability of the predicted segmentations by means of the CDR. In this regard, the

Table 5
State-of-the-art comparison for the DRISHTI-GS dataset.

Methods	Classification AUCROC (%)	Cup segmentation Avg. Dice (%)	Disc segmentation Avg. Dice (%)	CDR Avg. δ_{CDR}
Shankaranarayana et al. [16]	–	84.8	96.3	0.1045
Yu et al. [15]	–	88.77	97.38	–
Liu et al. [14]	–	89	**98**	–
Wang et al. [12]	85.83	90.1	97.4	0.048
Sreng et al. [10]	92.06	–	91.73	-
Ours (Multi-task Proposed)	**94.74**	**91.03**	97.18	**0.0413**

classification results reported by Wang et al. [12] are directly obtained by using the CDR as risk index, which is a common approach in the literature. For comparison, we performed the same evaluation using the CDR as risk index for glaucoma diagnosis. The obtained result is 89.88 AUCROC(%). Therefore, our lower δ_{CDR} also translates to a significantly better separability of the healthy and pathological cases. Nevertheless, the result obtained by the classification output of the multi-task network is even better by a significant margin. This also happens for the REFUGE dataset, where the result obtained by the CDR-based classification is 94.18 AUCROC(%), again lower than the result of the classification output in the multi-task network. This indicates that the prediction of the network is based on additional relevant information besides the optic disc and optic cup morphology (at least as measured by the CDR).

6. Discussion

6.1. Main results and advantages

The results presented in Section 5.1 demonstrate that the proposed methodology is advantageous for simultaneous classification and segmentation in the context of glaucoma diagnosis. In this regard, the proposed methodology always outperforms the standard multi-task methodology that is used as baseline and significantly improves the classification performance with respect to the single-task approach. In particular, the results indicate that our proposal is particularly advantageous in the most challenging scenarios, such as the classification task or the smaller DRISHTI-GS dataset. These satisfactory results are achieved due to the combination of the two main technical novelties in our methodology. First, the Proposed network provides a significant increase in the number of shared representations, which seems to be particularly beneficial for the classification task. Then, the M-Ada optimization ensures a balanced training between tasks while avoiding any additional hyperparameter tuning. The conducted experiments demonstrate that these two approaches are advantageous for multi-task learning independently of each other.

The comparison with the state-of-the-art also shows that the proposed methodology is highly competitive with previous approaches that are specialized for either classification or segmentation. Additionally, besides the rescaling of the images, our methodology does not involve any pre-processing or post-processing stage. In the state-of-the-art, however, it is common the use of pre-processing stages, including in several cases the previous extraction of the optic disc region. This unnecessarily complicates the methodology by adding the optic disc detection [12,14–16] or segmentation [10] as an additional step. Additionally, the detection failure rates of this additional step are usually not integrated in the reported performances. Also, several works post-process the predicted segmentations to adequate the output to the expected optic disc and cup morphology [12,14–16]. In comparison to all these alternatives, our methodology consists of a single neural network that simultaneously learns to predict the final classification and segmentation from the raw images.

6.2. Optimization and network architecture analyses

In Section 5.2, we provide a detailed ablation study of the proposed optimization approach and network architecture. The results show that, while the M-Ada optimization is always advantageous for multi-task learning, the Proposed network requires an adequate balance between tasks to outperform the Base network. In our methodology, this adequate balance is automatically achieved by using the M-Ada optimization.

The success of the M-Ada optimization and the higher sensitivity of the Proposed network to the balance between tasks can be explained by a combination of different factors. First, it must be considered that the parameter updates during the network training are directly obtained from the back-propagated gradients. In this regard, following a Standard optimization strategy, those tasks that provide relatively stronger gradients will have a larger influence in the direction of training. Second, adaptive optimization algorithms, such as Adam, rely on the recent history of past gradients to tune the effective learning rates for each parameter (Eq. (3)). In the Standard optimization procedure, the gradients of the different tasks are integrated together before computing the gradients history and the effective learning rates (Fig. 4). Thus, in the shared layers, the optimization of each task is also directly influenced by the past gradients of the other tasks. These effects are more pronounced in the Proposed network because both encoder and decoder are shared between tasks.

Another factor that should be considered is the effect of the skip connections. One of the roles of the skip connections is to facilitate the back-propagation of gradients to the bottom layers of the network. Besides the skip connections between encoder and decoder, the Proposed network presents several skip connections to the classification head. This facilitates the back-propagation of the classification gradients. Thus, relative to the Base network, the Proposed architecture makes it easier for the classification task to dominate the training of the shared layers, which comprise both encoder and decoder in this network.

The previous issues, which are evident when using the S-EW optimization strategy, are partially mitigated by the S-GS alternative and corrected by the proposed M-Ada optimization. In this regard, M-Ada usually provides the best performance, for both Proposed and Base networks. This may be explained by the fact that S-GS can only act on the balance between tasks at a single point (the network output), whereas M-Ada acts at each layer of the network. Thus, the latter approach presents a higher capacity to manipulate the balance between tasks. In particular, M-Ada allows the simultaneous application of independent corrective actions at different layers (and parameters) of the network. In contrast, the approaches that are based on the use of loss weighting hyperparameters, such as S-GS, necessarily assume that a single corrective action is enough for balancing the tasks throughout the network. The results obtained in Section 5.2 show that this is not necessarily the case. Additionally, this issue may be accentuated by the presence of skip connections that explicitly provide different paths for the back-propagation of the gradients of each task. In this scenario, the results show that M-Ada is the most successful of the studied approaches.

6.3. Limitations and future works

Besides the improved performance, the proposed methodology also presents different computational requirements in comparison to the baseline multi-task approach. First, the Proposed network has additional operations due to the mini-encoder classification head with skip connections. However, this classification head has been precisely designed to minimize the increase in the number of parameters. Second, the M-Ada optimization requires additional memory as well as additional operations to compute the task-specific gradients and parameter updates. Thus, the computational time for each training experiment is extended in comparison to the use of fixed loss weighting hyperparameters. However, in practice, a grid search is typically used to find the optimum hyperparameters. In that case, our proposal still provides a significant reduction of the computational time due to the fact that only a single training experiment is required. Nevertheless, future works could explore different alternatives to further improve the computational efficiency of the proposed approach. Also, regarding the M-Ada optimization, in this work we provide an extensive comparison against the use of fixed loss weighting hyperparameters, which represents the most commonly used alternative for multi-task learning. However, in the literature, there are also additional approaches that allow the automated estimation of the loss weighting hyperparameters during a single training experiment. The proposed M-Ada optimization can be seen as a novel alternative to these approaches and presents the potential of being successfully applied in different application domains and multi-task scenarios. Similarly, the Proposed network also shows remarkable potential for being advantageous in other multi-task settings combining pixel-level and image-level prediction tasks. In that regard, future works could explore the application of our proposals in other domains and tasks as well as perform extended comparisons with related approaches in the literature.

7. Conclusions

In this work, we propose a novel methodology that allows the simultaneous classification of glaucoma and segmentation of the optic disc and cup in retinal images. Our proposal presents two main novelties regarding multi-task learning. First, we propose a network design that shares most of the layers and learned representations between the classification and segmentation tasks. Second, we propose a multi-adaptive optimization approach that provides a well-balanced multi-task training without using loss weighting hyperparameters.

The proposed methodology is exhaustively validated on two different public datasets, taking into consideration the diagnosis of glaucoma, the segmentation of the optic disc and cup, and the extraction of a relevant biomarker such as the CDR. The obtained results show that, in general, the proposed methodology outperforms comparable multi-task and single-task alternatives. Additionally, the proposal is competitive with the best state-of-the-art approaches, which are specialized for each task and typically rely on additional ad-hoc processing. Finally, we also provide a detailed ablation study and analysis of the methodology. This analysis demonstrates that both the proposed network architecture and the optimization procedure are advantageous for multi-task learning regardless of each other. Consequently, both proposals could be individually considered for other applications of multi-task learning in future works.

CRediT authorship contribution statement

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

Declaration of competing interest

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

Acknowledgments

This work is supported by Instituto de Salud Carlos III, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project; Ministerio de Ciencia e Innovación, Government of Spain, through the RTI2018-095894-B-I00 and PID2019-108435RB-I00 research projects; Axencia Galega de Innovación (GAIN), Spain, Xunta de Galicia, through grant ref. IN845D 2020/38; Xunta de Galicia and the European Social Fund (ESF) of the EU through the predoctoral contract ref. ED481A-2017/328; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, Spain, through Grupos de Referencia Competitiva, grant ref. ED431C 2020/24. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Cultura, Educación e Universidade, Xunta de Galicia, Spain, through the ERDF (80%) and Secretaría Xeral de Universidades (20%), Spain. Funding for open access charge: Universidade da Coruña/CISUG, Spain

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.asoc.2021.108347.

References

[1] Y. Tham, X. Li, T.Y. Wong, H.A. Quigley, T. Aung, C. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040, Ophthalmology 121 (2014) 2081–2090, http://dx.doi.org/10.1016/j.ophtha.2014.05.013.

[2] R.N. Weinreb, T. Aung, F.A. Medeiros, The pathophysiology and treatment of glaucoma: A review, JAMA (ISSN: 0098-7484) 311 (18) (2014) 1901–1911, http://dx.doi.org/10.1001/jama.2014.3192.

[3] S. Shah, A. Chatterjee, M. Mathai, S.P. Kelly, J. Kwartz, D. Henson, D. McLeod, Relationship between corneal thickness and measured intraocular pressure in a general ophthalmology clinic, Ophthalmology (ISSN: 0161-6420) 106 (11) (1999) 2154–2160, http://dx.doi.org/10.1016/S0161-6420(99)90498-0.

[4] A. Sarhan, J. Rokne, R. Alhajj, Glaucoma detection using image processing techniques: A literature review, Comput. Med. Imaging Graph. (ISSN: 0895-6111) 78 (2019) 101657, http://dx.doi.org/10.1016/j.compmedimag.2019.101657.

[5] Y. Hagiwara, J.E.W. Koh, J.H. Tan, S.V. Bhandary, A. Laude, E.J. Ciaccio, L. Tong, U.R. Acharya, Computer-aided diagnosis of glaucoma using fundus images: A review, Comput. Methods Programs Biomed. (ISSN: 0169-2607) 165 (2018) 1–12, http://dx.doi.org/10.1016/j.cmpb.2018.07.012.

[6] A.C. Thompson, A.A. Jammal, F.A. Medeiros, A review of deep learning for screening, diagnosis, and detection of glaucoma progression, Transl. Vis. Sci. Technol. (ISSN: 2164-2591) 9 (2) (2020) 42, http://dx.doi.org/10.1167/tvst.9.2.42.

[7] F.A. Medeiros, A.A. Jammal, A.C. Thompson, From machine to machine: An OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs, Ophthalmology (ISSN: 0161-6420) 126 (4) (2019) 513–521, http://dx.doi.org/10.1016/j.ophtha.2018.12.033.

[8] Y. Chai, H. Liu, J. Xu, A new convolutional neural network model for peripapillary atrophy area segmentation from retinal fundus images, Appl. Soft Comput. (ISSN: 1568-4946) 86 (2020) 105890, http://dx.doi.org/10.1016/j.asoc.2019.105890.

[9] H. Fu, J. Cheng, Y. Xu, C. Zhang, D.W.K. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Trans. Med. Imaging 37 (11) (2018) 2493–2501, http://dx.doi.org/10.1109/TMI.2018.2837012.

[10] S. Sreng, N. Maneerat, K. Hamamoto, K.Y. Win, Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images, Appl. Sci. (ISSN: 2076-3417) 10 (14) (2020) http://dx.doi.org/10.3390/app10144916.

[11] J.J. Gómez-Valverde, A. Antón, G. Fatti, B. Liefers, A. Herranz, A. Santos, C.I. Sánchez, M.J. Ledesma-Carbayo, Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning, Biomed. Opt. Exp. 10 (2) (2019) 892–913, http://dx.doi.org/10.1364/BOE.10.000892.

[12] S. Wang, L. Yu, X. Yang, C. Fu, P. Heng, Patch-based output space adversarial learning for joint optic disc and cup segmentation, IEEE Trans. Med. Imaging 38 (11) (2019) 2485–2495, http://dx.doi.org/10.1109/TMI.2019.2899910.

[13] Y. Jiang, L. Duan, J. Cheng, Z. Gu, H. Xia, H. Fu, C. Li, J. Liu, Jointrcnn: A region-based convolutional neural network for optic disc and cup segmentation, IEEE Trans. Biomed. Eng. 67 (2) (2020) 335–343, http://dx.doi.org/10.1109/TBME.2019.2913211.

[14] Q. Liu, X. Hong, S. Li, Z. Chen, G. Zhao, B. Zou, A spatial-aware joint optic disc and cup segmentation method, Neurocomputing (ISSN: 0925-2312) 359 (2019) 285–297, http://dx.doi.org/10.1016/j.neucom.2019.05.039.

[15] S. Yu, D. Xiao, S. Frost, Y. Kanagasingam, Robust optic disc and cup segmentation with deep learning for glaucoma detection, Comput. Med. Imaging Graph. (ISSN: 0895-6111) 74 (2019) 61–71, http://dx.doi.org/10.1016/j.compmedimag.2019.02.005.

[16] S.M. Shankaranarayana, K. Ram, K. Mitra, M. Sivaprakasam, Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation, IEEE J. Biomed. Health Inf. 23 (4) (2019) 1417–1426, http://dx.doi.org/10.1109/JBHI.2019.2899403.

[17] A.S. Hervella, L. Ramos, J. Rouco, J. Novo, M. Ortega, Multi-modal self-supervised pre-training for joint optic disc and cup segmentation in eye fundus images, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 961–965, http://dx.doi.org/10.1109/ICASSP40776.2020.9053551.

[18] R. Zhao, S. Li, Multi-indices quantification of optic nerve head in fundus image via multitask collaborative learning, Med. Image Anal. (ISSN: 1361-8415) 60 (2020) 101593, http://dx.doi.org/10.1016/j.media.2019.101593.

[19] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, L. Van Gool, Multi-task learning for dense prediction tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1, http://dx.doi.org/10.1109/TPAMI.2021.3054719.

[20] S. Vandenhende, S. Georgoulis, L. Van Gool, Mti-net: Multi-scale task interaction networks for multi-task learning, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 527–543.

[21] C. Playout, R. Duval, F. Cheriet, A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images, IEEE Trans. Med. Imaging 38 (10) (2019) 2434–2444, http://dx.doi.org/10.1109/TMI.2019.2906319.

[22] A. Amyar, R. Modzelewski, H. Li, S. Ruan, Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation, Comput. Biol. Med. (ISSN: 0010-4825) 126 (2020) 104037, http://dx.doi.org/10.1016/j.compbiomed.2020.104037.

[23] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, O.H. Elibol, A comparison of loss weighting strategies for multi task learning in deep neural networks, IEEE Access 7 (2019) 141627–141632, http://dx.doi.org/10.1109/ACCESS.2019.2943604.

[24] S. Graham, Q.D. Vu, S.E.A. Raza, A. Azam, Y.W. Tsang, J.T. Kwak, N. Rajpoot, Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. (ISSN: 1361-8415) 58 (2019) 101563, http://dx.doi.org/10.1016/j.media.2019.101563.

[25] X. Wang, H. Chen, A.-R. Ran, L. Luo, P.P. Chan, C.C. Tham, R.T. Chang, S.S. Mannil, C.Y. Cheung, P.-A. Heng, Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning, Med. Image Anal. (ISSN: 1361-8415) 63 (2020) 101695, http://dx.doi.org/10.1016/j.media.2020.101695.

[26] A. Kendall, Y. Gal, R. Cipolla, Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[27] M. Guo, A. Haque, D.-A. Huang, S. Yeung, L. Fei-Fei, Dynamic Task Prioritization for Multitask Learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[28] Z. Chen, J. Ngiam, Y. Huang, T. Luong, H. Kretzschmar, Y. Chai, D. Anguelov, Just pick a sign: Optimizing deep multitask models with gradient sign dropout, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 2039–2050.

[29] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, C. Finn, Gradient surgery for multi-task learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 5824–5836.

[30] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[31] J. Morano, A.S. Hervella, N. Barreira, J. Novo, J. Rouco, Multimodal transfer learning-based approaches for retinal vascular segmentation, in: ECAI 2020 - 24th European Conference on Artificial Intelligence, in: Frontiers in Artificial Intelligence and Applications, vol. 325, 2020, pp. 1866–1873, http://dx.doi.org/10.3233/FAIA200303.

[32] J.I. Orlando, H. Fu, J. Barbosa Breda, K. van Keer, D.R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, J. Lee, X. Li, P. Liu, S. Lu, B. Murugesan, V. Naranjo, S.S.R. Phaye, S.M. Shankaranarayana, A. Sikka, J. Son, A. van den Hengel, S. Wang, J. Wu, Z. Wu, G. Xu, Y. Xu, P. Yin, F. Li, X. Zhang, Y. Xu, H. Bogunović, REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, Med. Image Anal. (ISSN: 1361-8415) 59 (2020) 101570, http://dx.doi.org/10.1016/j.media.2019.101570.

[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

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

[35] T. Tieleman, G. Hinton, Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude, in: COURSERA: Neural Networks for Machine Learning, 2012.

[36] A.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.

[37] J. Sivaswamy, S. Krishnadas, A. Chakravarty, A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis, JSM Biomed. Imag. Data Pap. 2 (1) (2015).

[38] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis, Expert Syst. Appl. 185 (2021) 115598, http://dx.doi.org/10.1016/j.eswa.2021.115598.

[39] 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.

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

[41] M.D. Bloice, P.M. Roth, A. Holzinger, Biomedical image augmentation using augmentor, Bioinformatics (ISSN: 1367-4803) 35 (21) (2019) 4522–4524, http://dx.doi.org/10.1093/bioinformatics/btz259.

[42] A.S. Hervella, J. Rouco, J. Novo, M.G. Penedo, M. Ortega, Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Comput. Methods Programs Biomed. 186 (2020) 105201, http://dx.doi.org/10.1016/j.cmpb.2019.105201.

[43] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83, http://dx.doi.org/10.2307/3001968.