Multidataset Incremental Training for Optic Disc Segmentation

Javier Civit-Masot[1](✉), Antonis Billis[2], MJ Dominguez-Morales[1], Saturnino Vicente-Diaz[1], and Anton Civit[1]

[1] Escuela Superior de Ingenieria Informatica, University of Seville, Seville, Spain
{jcivit,mdominguez,satur,civit}@atc.us.es
[2] Lab of Medical Physics, Aristotle University of Thessaloniky, Thessaloniky, Greece
ampillis@med.auth.gr

Abstract. When convolutional neural networks are applied to image segmentation results depend greatly on the data sets used to train the networks. Cloud providers support multi GPU and TPU virtual machines making the idea of cloud-based segmentation as service attractive. In this paper we study the problem of building a segmentation service, where images would come from different acquisition instruments, by training a generalized U-Net with images from a single or several datasets. We also study the possibility of training with a single instrument and perform quick retrains when more data is available. As our example we perform segmentation of Optic Disc in fundus images which is useful for glaucoma diagnosis. We use two publicly available data sets (RIM-One V3, DRISHTI) for individual, mixed or incremental training. We show that multidataset or incremental training can produce results that are similar to those published by researchers who use the same dataset for both training and validation.

Keywords: Deep learning · Eye fundus image segmentation · Multiple dataset training · Incremental training · Glaucoma

1 Introduction

Glaucoma is a disabling decease that can lead to blindness in about 2 to 5% of the cases and sight impairment in 10% of the cases [19]. Although Loss of vision can occur even with the best treatment, correct therapy and follow-up will stabilize the majority of patients with glaucoma.

The key to detection and management of glaucoma is understanding how to examine the optic disc (OD) [4]. The OD is an oval ‘plughole’ down which the retinal nerve fibres descend through a sheet known as the lamina cribrosa. The retinal nerve fibres are then bundled together to form the optic nerve. The optic cup (OC) is the white, cup-like area in the center of the optic disc. The tissue between the border of the cup and the disc is the neuroretinal rim. This tissue consists mainly of nerve fibers with some glial cells and is usually pink. Most normal discs are more vertically oval and their cup more horizontally oval. A typical retina fundus image is shown in Fig. 1.

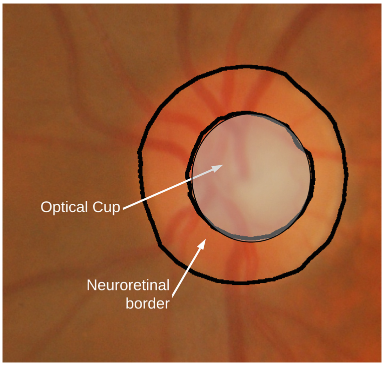

Fig. 1. Optic disc and cup

Several indicators are used to aid the diagnosis of glaucoma from fundus eye images. The cup to disc ratio (CDR) [18] which is the rate between the diameters of the optic disc and cup is the most widely used. In the mean CDRs of the glaucoma and normal eyes were 0.65 ± 0.13 and 0.39 ± 0.15, respectively allowing CDR to be used as a diagnostic aid. Another diagnostic approach is based on the rule based on the shape of the neuroretinal RIM. According to this rule in normal eyes, the thickness of the neuroretinal rim along the cardinal meridians of the OD decreases in the order inferior (I) $>$ superior (S) $>$ nasal (N) $>$ temporal (T) [9]. In any case accurate OC/OD segmentation is required to be able to apply these techniques. This segmentation is an error prone process even for expert ophthalmologists specially in typical work overloaded scenarios.

Thus machine learning (ML) approaches are attractive for fundus image segmentation. Segmentation methodologies are based in three possible approaches [24]: Form matching based on random forests, support vector machines or K-means [14]; techniques based on transformations and active contours [3] oe Deep learning-based methods [1,21,22,26].

There are two scenarios for using image segmentation tools in a clinical set up. In the first one the tool is marketed by the provider of the image acquisition instrument. In this case we can train using images captured with the instrument linked to the tool. In a second scenario the segmentation is implemented as a service and has to be able to segment images from different clinics acquired with different instruments. Some approaches have been proposed for combined dataset training e.g. [5], however, they have not been applied, to medical image segmentation. Some previous works [1,21] have used different data sets but they train and test with each set independently. In [8] authors use several datasets but training is always performed with a combined dataset and, thus, it does not show the influence of performing single or multidataset training.

The objective of this paper is to study the problems that we would face to implement fundus segmentation as a service. For this purpose we will first find a suitable architecture for fundus image segmentation. Then we will train the system using a single data set and use it to predict over images from that set and images acquired with other instruments. After this we will use a mixed training dataset combining images from the different datasets, train our net with it and use it to make predictions on the different testing datasets. Finally we will study the realistic approach for clinical practice, i.e., to train with what is available and later do retraining when more data becomes available.

2 Materials and Methods

Implementation. Selection of Architectures for Segmentation. U-Net is a widely used fully convolutional neural network introduced in [20]. It has been widely used in image segmentation applications including several works related to ophthalmologic images including OD segmentation [21], retinal vessel segmentation [16] and diabetic retinopathy diagnosis [2]. In this work tools were developed to generalize U-Net models to allow rapid implementation in cloud-based GPU and TPU [10] architectures. We use a Keras [7] Tensorflow 2.1 implementation on the Google Collaboratory Python notebook environment.

In this section we will establish a methodology to select a generalized U-Net that correctly segments the OD. For this purpose we will train using combined datasets leaving the comparison of this approach with other alternatives for later sections.

Our networks are optimized versions of the U-Net proposed in [21]. Among the modifications are the use of a different image generator that produces the larger image batches for TPU training by means of static and dynamic data augmentation [27]. Also, to be able to modify our U-Net structure without recoding, we use a highly parameterizable U-Net recursive model. With this model we can change the depth and width of the net, the possibility of batch normalization, the use of upsampling or transpose convolution and the width ratio between successive layers known as increment ratio (IR). IR [13] is widely used as an effective pruning method.

We select network with the smallest IR and, thus, the one with less trainable parameters when two networks produce similar results. Although we train and make predictions in the web pruning improves time and reduces operating costs. The reduction of the initial network width and its depth are alternatives that we also explore.

We use 96 image batches for both training and testing, and train for 15 epoch using 100 training steps and 30 test steps per time. We use an Adam optimization algorithm with a learning rate of 0.0008. These values have proven suitable for TPU and GPU based training and provide good results with training times of less than 30 min for the TPU implementation.

Regarding the data sets, we use the publicly available RIM-ONE v3 and DRISHTI fundus image data sets. RIM ONE-v3 [12], form the MIAG group at the University of La Laguna (Spain), consists of 159 images labeled by expert ophthalmologists for OD/OC. DRISHTI-GS [23], from Aravind Eye Hospital, Madurai (India), consists of 101 images also labeled for OC/OD. In Fig. 2 we can see that the images that come from both data sets have very different characteristics.

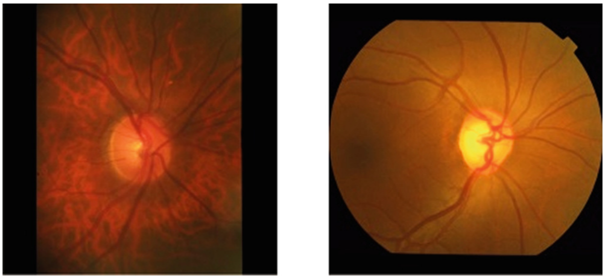

Fig. 2. Images RIM ONE (left) and DRISHTI (right) datasets

Figure 3 shows the used segmentation methodology. We start by trimming and resizing the images in the data sets by removing a 10% border on all edges of the image to reduce the black borders. After this, we resize the images to 128×128 pixels and perform a limited clip contrast histogram equalization. After, we split the data sets. For each set, we use 75% of the images for training and 25% for validation. Next, we perform, for each data set, static data augmentation by creating images with modified brightness and different contrast parameters. Later, we merge the data from the different sets. In the merging process, we perform data replication and random combustion to provide longer vectors as input for our dynamic image generators. Image generators [6] increase the data by performing random rotations, moving, zooming and flipping over the images of the extended merged data set.

Dice coefficient [11] is used to estimate the similarity between the correct and predicted disc. This figure of merit, also known as F1 score, is widely used and allows us to compare our results with those from other works. Dice coefficient is defined as:

$$DC = \frac{2TP}{2TP + FP + FN} \tag{1}$$

In this equation TP indicates true positives, FP false positives, and FN false negatives.

2.1 Instrument Based Versus Cloud-Based Segmentation

To show the results obtained when training with data from a single instrument or when training as a service, we will first train the system using a single data set and use it to predict images from other sets. After this we will use last section's mixed data set to train our network and use it to make the same predictions. In this section we use a generalized 6-layer U-Net which had good results in the previous section. It has only 40 channels in the first layer and the layer IR is 1.1. Given that we also resize the images in the sets to 128×128, the number of trainable parameters is less than 1 million.

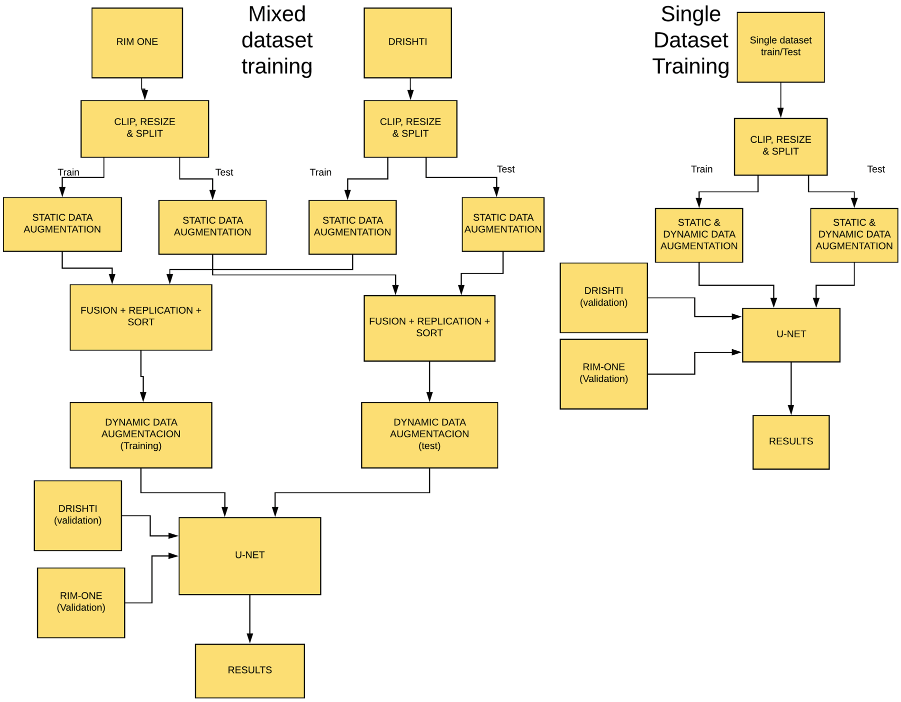

Fig. 3. Segmentation methodology

2.2 Study of the Viability of Incremental Learning

An interesting alternative, essential to implement cloud services, is to train initially with the available data and later modify the weights incrementally as we have data from new instruments. In this section, we will first train the system using a single data set and make predictions as stated above. Subsequently we will perform a short retraining (3 epochs) using images from the other data set and study the influence on the results. Thus, we test the feasibility of an incremental training using the resources and networks that we would deploy to implement a web service for segmentation of fundus images. This methodology is different from that used in other papers (e.g., [1,21,22,26]), where data from a single source are used for both training and testing.

We compare our work with those works that use deep learning for OD segmentation and use the DRISHTI or RIM-ONE data set. Zilly et al. [26] use a light three-layer CNN including sophisticated pre and postprocessing and apply it independently to both data sets. Sevastopolsky [21] uses a very light U-Net and provides results for RIM ONE. Al-Bander [1] uses a heavily modified, dense U-Net and provides results for both data sets. Shankaranarayana [22] uses a residual U-Net and provides results for RIM ONE.

3 Results

3.1 Network Selection

Regarding disk segmentation (Table 1), for our experiments we initially used a network that is very similar to the original U-Net: 5 stages and default dropout rates (0.3). We use transpose convolution since direct subsampling is not currently compatible with TPUs.

Table 1. Disk segmentation results for different network architectures.

D/W/IR	Train/Test	Best/Worst/Std.	RRP	MTP
5/32/1.5	0.94/0.91	0.99/0.69/0.07	95	3.5
5/40/1.2	0.90/0.79	0.98/0.64/0.09	95	1.1
6/40/1.3	0.95/0.91	0.98/0.64/0.09	96	3.3
6/40/1.1	0.95/0.91	0.97/0.59/0.09	95	.9
7/40/1.2	0.95/0.92	0.98/0.61/0.11	97	2.6
7/64/1.3	0.96/0.94	0.99/0.62/0.08	97	14

Table 1 shows Dice coefficients for train and for test sets for various evaluated U-Net alternatives. The first row of the table defines the main parameters of the architecture, that is, the network depth (D), the number of filters in the first layer (W), and the IR. As an example, 6/40/1.1 means that we use a generalized 6-layer U-Net with 40 channels in the first layer and an IR of 1.1. This network is shown in Fig. 4.

In addition to this base case we provide data for pruned networks where we try to obtain the same or greater performance with a smaller number of parameters. To achieve this goal, we decrease the IR while increasing the number of filters in the first layer, the depth of the network or both. The MTP column in Table 1 shows the millions of trainable parameters in the network. For each network architecture, we provide the mean Dice coefficient for the training and test sets, the Dice coefficient for the best and worst predicted images in the test set and the Dice standard deviation over the test set.

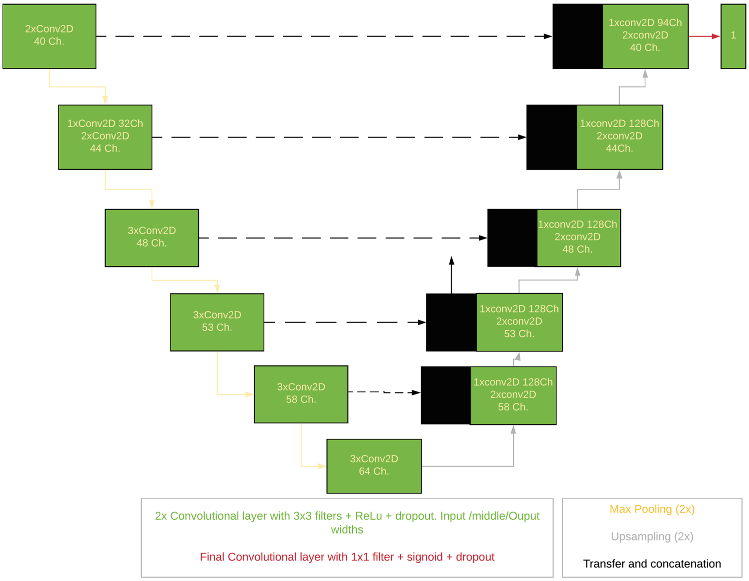

Fig. 4. Generalized 6 layer 6/40/1.1 U-Net architecture

Real disk shapes are not circular, but are approximately elliptical. In general, the ratio of the horizontal cup and disc diameters is larger than that of the diameters in the horizontal direction [17], however, in most of the works on the subject including all those mentioned in Table 2 the CDR is calculated using the mean diameters of the OC and OD. Although there are several possible interpretations of the average diameter, they have small differences with the real fundus images. In this paper we consider that the average radius of the OD is the square root of its area divided by π. We define a new parameter (Radio Ratio-RRP parameter) that is very useful for estimating the accuracy of the CDR. This parameter is defined as the percentage of test images for which the predicted disk radius has less than 10% error compared to the average radius of the ground truth. As an example, for the deepest network in the table, for 97% of the images our estimate of the radius has an error below 10% (RRP = 97).

We can see that deep networks with few parameters such as 6/40/1.1, which has less than 1M trainable parameters, achieve good results for disk segmentation. In this specific case, it achieves a RRP of 95. We will use this architecture for all the experiments in the rest of this paper. As a reference, we include in the Table 1 a very wide and deep network (7/64/1.3) with more than 14 million parameters. Although the performance of this network is better than in any other case, the small improvement does not justify the additional complexity of the network.

3.2 Disk Segmentation with Multiple Datasets

We want to discover how our system behaves when training with the combined data set and compare these results with those obtained when a single data set (i.e. RIM ONE or DRISHTI) is used to train the system. We will also compare the results with those obtained by other researchers. Table 2 shows Disk segmentation results for the three scenarios in this section. In the first two we train using a single data set and validate using the part of that data set that is not used for training and the other data set, while in the last scenario we train and validate with a mixed data set. Our scenarios are as follows:

- 75% of the DRISHTI data set is used for training and validation is done first with the rest of DRISHTI and then with the RIM ONE data set.
- 75% of the RIM ONE data set is used for training and validation is carried out with the rest of the RIM ONE data set and the DRISHTI data set.
- 75% of a mixed data set is used to train the networks and then we validate with the rest of the mixed data set.

We can see in Table 2 that, with our 6-layer network in the scenarios in which we train with a single data set the results when testing with images of that data set are good with Dice coefficients greater than 0.98 (DRISHTI) and 0.96 (RIM1). However, when we validate these networks with the other set, the results are below 0.66 or even 0.50 in some cases. In the third scenario in which we train with a mixed data set, we get results that are more similar when we test with images that come from both data sets. In this case, we obtain a Dice of 0.96 for the DRISHTI test subset and 0.87 for the RIM ONE subset.

Table 2. Multiple/single set dice coefficients.

Author	DRI	RIM
Zilly et al. [26]	0.97	-
Al-Bander [1]	0.95	0.90
Sevastopolsky [21]	-	0.94
Shankaranarayana et al. [22]	-	0.98
Drishti trained	0.98	0.50
RIM trained	0.66	0.97
Multi-dataset	0.96	0.87

In Table 2 we also results of previously referenced works which have trained and tested with each data set independently. Thus, they are related to our first two scenarios but they never test a network trained with one data set with images from another. Although we use networks with a small number of parameters, when we train with a single set we obtain results similar to those obtained by other papers. When training with DRISHTI, we obtained a Dice value of 0.98.

This value is slightly higher than 0.97 [26]. In the RIM ONE trained case, we get a dice value of 0.97 which compares well with 0.98 [1].

The most significant results in the Table 2 come from what can't be obtained in other studies, i.e., when we train with a dataset and predict using data from another source. In this case, we always get poor prediction results. This demonstrates that it is not feasible to create a service using training data captured with a single acquisition device. We also see in Table 2 that when you train with a combined data set, the network produces good results for both data sets, although not as good as when the training and prediction sets are part of the same set of data.

Regarding the clinically significant RRP parameter (Table 3), in the first two scenarios almost all radios for the test data are predicted with less than 10% error when compared to the segmentation done by ophthalmologists. However, the prediction for the other data set is much worse and, in some cases, we never get errors below 10%. This situation improves significantly when we train with a mixed data set.

Table 3. Radio ratio parameter.

	DRI	RIM
Drishti trained	100	38
RIM ONE trained	62	100
Multi-dataset	100	82

3.3 Incremental Training Results

We want to find out how our system behaves, when training with one set and then retraining lightly with some data from the other, and see if the results similar to those obtained when a single set of data is used (i.e. RIM ONE or DRISHTI) to train the system. Table 4 shows the results of disk segmentation for our two cases. On the first train, we use only DRISHTI data and validate using remaining of that data set and RIM ONE. In the second scenario, we make a brief retrain (3 epochs) using RIM ONE and the data set. Our scenarios are defined as follows:

- 75% of DRISHTI is used for training and validation is carried out first with the rest DRISHTI and then with the complete RIM ONE.
- 75% of RIM ONE is used to retrain the network and then we validate with the test part of both sets.

We can see in Table 4 that when we train with DRISHTI the tests with images from that same data set obtain very good Dice values. Specifically, we obtain an average Dice of 0.98 (DRISHTI) but only 0.64 (RIM1). The situation is worse than it seems as in the worst case for some RIM images the segmentation does not produce any pixels.

Table 4. Segmentation dice with retraining

Author	DRI	RIM
Zilly et al. [26]	0.97	-
Al-Bander [1]	0.95	0.90
Sevastopolsky [21]	-	0.94
Shankaranarayana et al. [22]	-	0.98
Drishti trained	0.98	0.64
RIM retrained	0.89	0.80

When we retrain the network with the other data set, the Dice values are 0.89 (DRISHTI) and 0.80 (RIM). For the worst case, we get a Dice of 0.69. Therefore, we can see that with a light retraining, the network can quickly learn the specific characteristics of the second data set. In Table 4 we include results of the other papers analyzed before. When we train with a single set of data, we obtain results for that set that are similar to those obtained by other papers. When training with the DRISHTI data set, we obtained a dice value of 0.98 for OD segmentation. This value is slightly above 0.97 [26]. As in the previous section the most significant results in Table 4 come from what is not available from other studies. The results obtained when we do a quick retraining show that, in this case, we get good prediction results for all test images.

4 Conclusions

We have been able to demonstrate that through the use of data from different data sets, adequate image preprocessing and significant data augmentation, we have been able to perform disk segmentation obtaining results with a performance similar to those obtained by other authors who use a single set for training and testing.

The use of a generalized configurable U-net recursive model allows us to easily train and test any U-Net configuration without having to make any changes to the code. This allows great flexibility for testing different architectures. We have tested networks with 4 to 7 layers, from 32 to 92 input channels, and with IR from 1.1 to 2.0. The number of parameters has varied from 0.9 to 44 million. Several U-Net architectures have been proven suitable for disk segmentation.

We have shown that by performing a rapid retraining with data from a new data set, and by preprocessing images and performing data augmentation, we can implement disk segmentation with performance equivalent to that reported by researchers using a single set for both training and testing.

4.1 Future Work

There are many possibilities to expand this work in the future. Among other possible topics, it would be interesting to implement modifications to always produce disc shapes that are acceptable to ophthalmologists. The automation of architecture configuration parameters and the use of other CNN architectures in parallel for the direct detection of glaucoma.

This work shows the importance of retraining by adding new sources to the segmentation system. In a real clinical service scenario, we would have to start training the network with the initially available data and retrain it when new images become available. The possibility of improving the network architecture by including residual blocks [25] or the combination of these blocks and a conventional U-Net [15] has proven effective in several segmentation applications and could potentially improve the performance of our process. The robustness of these networks when analyzing images from many different instruments remains an open problem for the future.

References

1. Al-Bander, B., Williams, B., Al-Nuaimy, W., Al-Taee, M., Pratt, H., Zheng, Y.: Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. Symmetry **10**(4), 87 (2018)
2. Aujih, A., Izhar, L., Mériaudeau, F., Shapiai, M.I.: Analysis of retinal vessel segmentation with deep learning and its effect on diabetic retinopathy classification. In: 2018 International Conference on Intelligent and Advanced System (ICIAS), pp. 1–6. IEEE (2018)
3. Bhat, S.H., Kumar, P.: Segmentation of optic disc by localized active contour model in retinal fundus image. In: Smart Innovations in Communication and Computational Sciences, pp. 35–44. Springer (2019)
4. Bourne, R.R.: The optic nerve head in glaucoma. Community Eye Health **19**(59), 44 (2006)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
6. Chollet, F.: Building powerful image classification models using very little data. Keras Blog (2016)
7. Chollet, F.: Deep Learning with Python, 1st edn. Manning Publications Co., Greenwich (2017)
8. Civit-Masot, J., Luna-Perejón, F., Vicente-Díaz, S., Rodríguez Corral, J.M., Civit, A.: TPU cloud-based generalized U-net for eye fundus image segmentation. IEEE Access **7**, 142379–142387 (2019). https://doi.org/10.1109/ACCESS.2019.2944692
9. Das, P., Nirmala, S., Medhi, J.P.: Diagnosis of glaucoma using CDR and NRR area in retina images. Netw. Model. Anal. Health Inform. Bioinform. **5**(1), 3 (2016)
10. Dean, J., Patterson, D., Young, C.: A new golden age in computer architecture: empowering the machine-learning revolution. IEEE Micro **38**(2), 21–29 (2018)
11. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
12. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: an open retinal image database for optic nerve evaluation. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. IEEE (2011)

13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Kim, S.J., Cho, K.J., Oh, S.: Development of machine learning models for diagnosis of glaucoma. PLoS One **12**(5), e0177726 (2017)
15. Kim, S., Bae, W., Masuda, K., Chung, C., Hwang, D.: Fine-grain segmentation of the intervertebral discs from MR spine images using deep convolutional neural networks: BSU-Net. Appl. Sci. **8**(9), 1656 (2018)
16. Lian, S., Li, L., Lian, G., Xiao, X., Luo, Z., Li, S.: A global and local enhanced residual u-net for accurate retinal vessel segmentation. IEEE/ACM Trans. Comput. Biol. Bioinform. (2019)
17. Lingam, C.L., Mansberger, S., Miglior, S., Paranhos, A., Pasquale, L.R., Susanna Jr., R., Wang, N.: 4. risk factors (ocular). Diagnosis of Primary Open Angle Glaucoma: WGA consensus series-10, vol. 10, p. 127 (2017)
18. MacIver, S., MacDonald, D., Prokopich, C.L.: Screening, diagnosis, and management of open angle glaucoma. Can. J. Optom. **79**(1), 5–71 (2017)
19. Quigley, H.A., Broman, A.T.: The number of people with glaucoma worldwide in 2010 and 2020. Br. J. Ophthalmol. **90**(3), 262–267 (2006)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
21. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of U-net convolutional neural network. Pattern Recogn. Image Anal. **27**(3), 618–624 (2017)
22. Shankaranarayana, S.M., Ram, K., Mitra, K., Sivaprakasam, M.: Joint optic disc and cup segmentation using fully convolutional and adversarial networks. In: Fetal, Infant and Ophthalmic Medical Image Analysis, OMIA 2017, pp. 168–176. Springer International Publishing (2017)
23. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S.: Drishti-GS: retinal image dataset for Optic Nerve Head (ONH) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 53–56. IEEE (2014)
24. Thakur, N., Juneja, M.: Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. Biomed. Signal Process. Control **42**, 162–189 (2018)
25. Xiuqin, P., Zhang, Q., Zhang, H., Li, S.: A fundus retinal vessels segmentation scheme based on the improved deep learning U-net model. IEEE Access **7**, 122634–122643 (2019). https://doi.org/10.1109/ACCESS.2019.2935138
26. Zilly, J., Buhmann, J.M., Mahapatra, D.: Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Comput. Med. Imaging Graph. **55**, 28–41 (2017)
27. Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V.: Learning data augmentation strategies for object detection. arXiv preprint arXiv:1906.11172 (2019)