Stack-U-Net: Refinement Network for Image Segmentation on the Example of Optic Disc and Cup

Artem Sevastopolsky[1,2], Stepan Drapak[1,3], Konstantin Kiselev[1], Blake M. Snyder[4,5], and Anastasia Georgievskaya[1,6]

[1] Youth Laboratories Ltd., Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Moscow, Russia
[3] Lomonosov Moscow State University, Moscow, Russia
[4] University of Colorado Denver School of Medicine, Aurora, CO, USA
[5] Francis I. Proctor Foundation, University of California San Francisco, San Francisco, CA, USA
[6] Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Moscow, Russia

Abstract. In this work, we propose a special cascade network for image segmentation, which is based on the U-Net networks as building blocks and the idea of the iterative refinement. The model was mainly applied to achieve higher recognition quality for the task of finding borders of the optic disc and cup, which are relevant to the presence of glaucoma. Compared to a single U-Net and the state-of-the-art methods for the investigated tasks, very high segmentation quality has been achieved without a need for increasing the volume of datasets. Our experiments include comparison with the best-known methods on publicly available databases DRIONS-DB, RIM-ONE v.3, DRISHTI-GS, and evaluation on a private data set collected in collaboration with University of California San Francisco Medical School. The analysis of the architecture details is presented, and it is argued that the model can be employed for a broad scope of image segmentation problems of similar nature.

1 Introduction

Medical image segmentation is a special computer vision area that is very important for many real-life applications. It is a natural extension of object classification and localization, necessary to fully understand the contents of an image. Nowadays, methods of deep learning provide the state-of-the-art results on many tasks of image processing, including the semantic and instance segmentation.

In many cases of biomedical applications, a small number of objects is to be found, but, on the other hand, often only small datasets can be acquired, class imbalance is present, and very high recognition quality and robustness is required [1].

In this work we intend to provide a new approach to the medical image segmentation tasks, which is based on well-known and highly-performing U-Net

[2] convolutional neural network (CNN) of encoder-decoder style. The latter is used as a basic block for a cascade of networks employed as the main model proposed. We refer to the neural network built as Stack-U-Net.

Compared to many other approaches of building the cascade of refinement networks, the one proposed in this work does not depend on the structure of the task and can be straightforwardly applied to many applications of image segmentation, image-to-image translation, etc.

Despite the linear growth of the number of parameters with the number of blocks, we observe that the model leads to the rate of overfitting similar to the original U-Net and only provides a noticeable quality gap. We consider this a consequence of regularly placed bottlenecks — the first layers of each basic network. This way, the basic models, conditioned by an input image, are only working to refine the output of the previous basic models.

The main purpose of the development of the described model was to provide an end-to-end solution for optic disc and cup segmentation with higher quality than the most of the known solutions. Relative size of these two organs is one of the most valuable factors determining the presence of glaucoma — the second leading cause of blindness all over the world. Segmentation of the optic disc and cup is a very time-consuming task currently performed only by the professionals. As stated in [3], according to a research, full segmentation of optic disc and cup requires about eight minutes per eye for a skilled grader. Solutions for automated analysis and assessment of glaucoma can be very valuable in various situations, such as mass screening and medical care in countries with significant lack of qualified experts.

2 Related work

The idea of the cascade network is present in a large number of various computer vision works. However, the information passed between sub-networks in a cascade is usually chosen differently and is sometimes implied by the structure of a solved problem.

The paper [4] applies a cascade multi-path refinement network by augmenting ResNet [5] pretrained on ImageNet [6] with RefineNet blocks, which take the output of ResNet's intermediate layers as an input and are organized in a decoder-like topology. Cascades of up to 4 2-scale RefineNet's are compared for the semantic segmentation problem. Similar approach is proposed in [7] for the task of instance-aware semantic segmentation: the first sub-network finds box instances (ROIs), they are fed to another sub-network which outputs a binary segmentation mask, and the mask is fed to another sub-network which segments separate instances.

In [8] a cascade of two U-Net's is applied for the liver and lesion segmentation in CT images as a model backbone, which is followed by 3D Conditional Random Field. Followed by the fact that the lesions are smaller regions inside the liver, the cascade is applied as follows: the first U-Net segments the liver, then its localized ROI is passed to a second U-Net. It is experimentally shown in the

work that the Dice score can be improved this way by 20% compared to a single U-Net. The same approach is applied in [9] for the segmentation of the optic disc and the optic cup, as the latter is smaller than the optic disc and is always inside of it.

There is a number of works that apply cascade of neural networks in a fashion more similar to our proposed idea. For instance, in [10] a well-known DeepPose method for human pose estimation is proposed, which is based on a cascade of regressors, iteratively refining each other. The first basic network localizes all the "skeleton" joints on an input image, and all the subsequent basic networks are refining previously found joints locations, conditioned by sub-images cropped by joints areas found. The work [11] follows a close approach for the face landmarks detection, but also benefits an idea of applying recurrent neural network (RNN): the weights of all basic networks, starting from the second one, are shared, and the whole model is trained as the RNN.

3 Stack-U-Net

As a preprocessing, unsupervised Contrast-Limited Adaptive Histogram Equalization (CLAHE) [12] is applied in order to bring the brightness characteristics closer across all the dataset.

The presented cascade model, which we refer to as Stack-U-Net, is depicted on Fig. 2. It consists of basic blocks, and each of them follows the encoder-decoder architecture similar to U-Net [2], depicted on Fig. 1. We consider 2 kinds of basic blocks: U-Net and Res-U-Net. They both feature skip connections (shown gray on the Fig. 1), linking layers of the encoder and decoder, which are of very high importance. Compared to the conventional U-Net, Res-U-Net also features residual connections (shown dashed light-brown on Fig. 1). All the basic blocks except the last one, end with 32 feature maps, which are stacked with the input image by long skip connections (shown dashed light-brown on the Fig. 2). The latter provide an additional information to the next basic block, so that it refines the previous features by directly accessing colors from the input image. One can notice that Stack-U-Net with Res-U-Net blocks allows for relatively more efficient gradient propagation in terms of information, as it preserves an identity mapping [13,4] between input and output without any intermediate layers.

As a loss function, we use $l(A, B)$:

$$l(A, B) = -\log d(A, B), \text{where:}$$
$$d(A, B) = \frac{2\sum\limits_{i,j} a_{ij} b_{ij}}{\sum\limits_{i,j} a_{ij}^2 + \sum\limits_{i,j} b_{ij}^2},$$

where $A = (a_{ij})_{i=1}^{H}{}_{j=1}^{W}$ is a predicted output map, containing probabilities that each pixel belongs to the foreground, and $B = (b_{ij})_{i=1}^{H}{}_{j=1}^{W}$ is a correct binary output map.

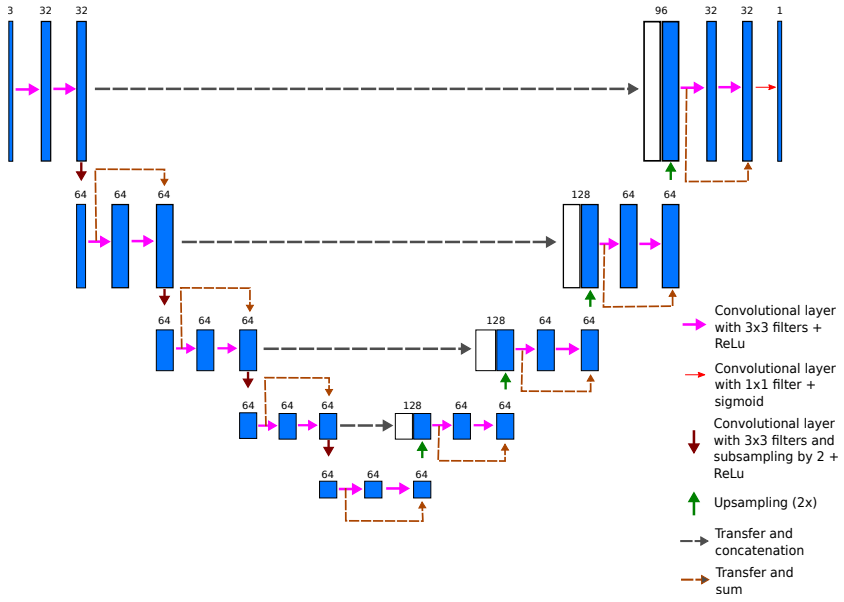

Fig. 1: Res-U-Net architecture — a basic block of the Stack-U-Net model. Another possible basic block is U-Net, which is the same module without residual connections marked light-brown dashed lines.

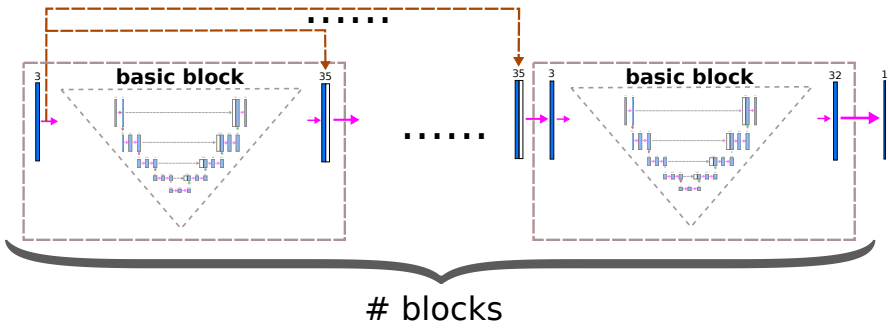

Fig. 2: Stack-U-Net — the main proposed model

$d(A, B)$ is a real-valued extension of Dice score for binary images $Dice(A, B) = \frac{2|A \cap B|}{|A|+|B|}$. Along with Dice score, we report the Intersection-over-Union score values: $IOU(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are defined as above. During the training, data augmentation was used to enlarge the training set by artificial ex-

amples. Images were subject to random rotations, zooms, shifts, flips and affine shears. Adam optimization method with learning rate of 10^{-5} was used.

4 Experiments

For experiments, we used the following datasets:

1. DRIONS-DB [14] — publicly available 110 color eye fundus images without cropping with annotation of the optic disc borders.
2. RIM-ONE v.3 [15] — publicly available 159 color eye fundus images with cropping (image side is approximately 5 times larger than the optic nerve diameter) with annotation of the optic disc and cup borders. Version 3 is the actual version.
3. DRISHTI-GS [16,17] — publicly available 50 color eye fundus images without cropping with annotation of the optic disc and cup borders.
4. UCSF-DB — private dataset of 963 color eye fundus images of 238 people without cropping, kindly provided by University of California, San Francisco (UCSF) Medical School, US and collected for optic disc and cup annotation tasks. For each photo, annotation of the optic disc and cup borders were prepared by 3 annotators. Final annotations were acquired as pixel-wise average of 3 masks for each of the 2 organs. Images were cropped by an optic disc area (with gap of 20 pixels from each side) based on the ground truth annotations.

For UCSF-DB dataset, several images of the same person were put either in train set altogether or in validation set altogether.

The comparison with the best found methods for the described public databases is presented in Table 1 and Table 2. We were unable to reproduce the results of other state-of-the-art methods. Evaluation on the large UCSF-DB dataset is presented in Table 4, which also contains a score of human annotator vs. another human annotator averaged by all pairs of annotators.

	DRIONS-DB		RIM-ONE v.3		DRISHTI-GS	
	IOU	Dice	IOU	Dice	IOU	Dice
Stack-U-Net (15 ResU-Net blocks)	**0.92**	0.96	0.91	0.95	**0.95**	**0.97**
Stack-U-Net (15 U-Net blocks)	0.90	0.95	**0.92**	**0.96**	0.94	**0.97**
U-Net [9]	0.89	0.94	0.89	0.95	0.90	0.95
Maninis et al. 2016 [18]	0.88	**0.97**	0.89	**0.96**	—	—
Zilly et al. 2017 [19]	—	—	0.89	0.94	0.91	**0.97**

Table 1: Results for **optic disc** segmentation. "—" indicates that the result is not reported.

	DRISHTI-GS		RIM-ONE v.3	
	IOU	Dice	IOU	Dice
Stack-U-Net (15 ResU-Net blocks)	0.80	**0.89**	0.73	**0.84**
Stack-U-Net (15 U-Net blocks)	0.77	0.86	0.72	0.83
U-Net with cropping by OD region [9]	0.75	0.85	0.69	0.82
Zilly et al. 2017 [19]	0.85	0.87	**0.80**	0.82
Zilly et al. 2015 [20]	**0.86**	0.83	—	—

Table 2: Results for **optic cup** segmentation. "—" indicates that the result is not reported.

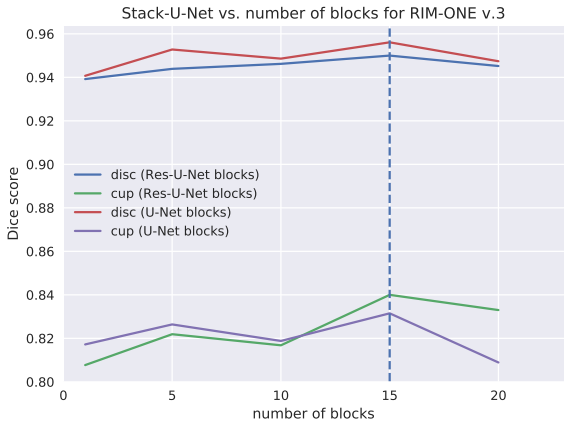

Fig. 3: Stack-U-Net performance w.r.t. the number of basic blocks.

	RIM-ONE v.3			
	Disc		Cup	
	IOU	Dice	IOU	Dice
Stack-U-Net (15 Res-U-Net blocks) w/ skip	**0.91**	**0.95**	**0.73**	**0.84**
Stack-U-Net (15 Res-U-Net blocks) w/o skip	0.90	0.94	0.72	0.83
Stack-U-Net (15 U-Net blocks) w/ skip	**0.92**	**0.96**	0.72	0.83
Stack-U-Net (15 U-Net blocks) w/o skip	0.91	0.95	**0.74**	**0.85**

Table 3: Comparison of the cascade model with and without long skip connections linking input image with the first layer of each basic block.

We observe that the model with 15 blocks works better than with the lower and higher number of blocks, regardless of the block type (Fig. 3). Skip connections typically enhance the results by a small extent, except for the case of Stack-U-Net with 15 U-Net blocks without skip connections (Table 3).

	UCSF-DB			
	Disc		Cup	
	IOU	Dice	IOU	Dice
Stack-U-Net (15 Res-U-Net blocks)	**0.92**	**0.96**	0.73	0.84
Stack-U-Net (15 U-Net blocks)	**0.92**	**0.96**	**0.74**	**0.85**
U-Net	**0.92**	0.94	0.73	0.84
Mean Human-vs.-Human	0.81	0.87	0.53	0.66

Table 4: Results on UCSF-DB large private dataset.

Visual comparison of the best and worst cases for the best-performing networks on each task for RIM-ONE v.3 database can be made based on Fig. 4.

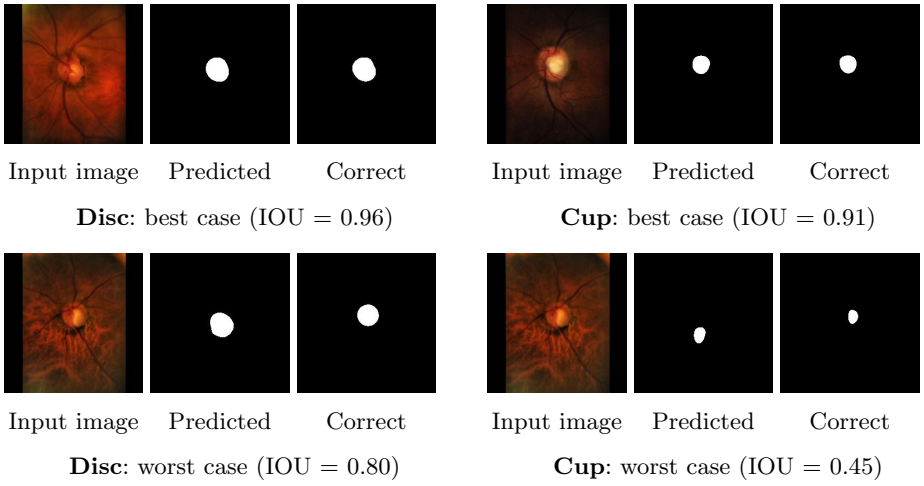

Fig. 4: The best and the worst cases of the algorithm performance on RIM-ONE v.3 database for the respective best models:
for optic disc — with Stack-U-Net with 15 U-Net blocks,
for optic cup — with Stack-U-Net with 15 Res-U-Net blocks.

5 Discussion

We present the model for image segmentation based on a stack of the well-known U-Net models. Each model in a cascade refines the result of the previous one, directly accessing the colors from an input image. For the task of optic disc and optic cup segmentation on eye fundus image, which requires a solution for the reliable glaucoma detection, we report high results, and the model outperforms existing solutions by a large number of benchmarks.

Linear increase of the number of parameters and of the time of the forward / backward pass remains a drawback, and, together with the observed quality gap, it especially motivates the further research.

Acknowledgment

Blake M. Snyder was supported in part by the Doris Duke Charitable Foundation through a grant supporting the Doris Duke International Clinical Research Fellows Program at the University of California San Francisco School of Medicine. Blake M. Snyder is a Doris Duke International Clinical Research Fellow.

References

1. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 565–571
2. 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, Springer (2015) 234–241
3. Lim, G., Cheng, Y., Hsu, W., Lee, M.L.: Integrated optic disc and cup segmentation with deep learning. In: Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on, IEEE (2015) 162–169
4. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 211–252
7. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3150–3158
8. Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., et al.: Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2016) 415–423
9. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. Pattern Recognition and Image Analysis **27**(3) (2017) 618–624
10. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1653–1660
11. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4177–4187

12. Szeliski, R.: Computer vision: algorithms and applications. Springer Science & Business Media (2010)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, Springer (2016) 630–645
14. Carmona, E.J., Rincón, M., García-Feijoó, J., Martínez-de-la Casa, J.M.: Identification of the optic nerve head with genetic algorithms. Artificial Intelligence in Medicine **43**(3) (2008) 243–259
15. Fumero, F., Alayón, S., Sanchez, J., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: An open retinal image database for optic nerve evaluation. In: Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on, IEEE (2011) 1–6
16. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomedical Imaging Data Papers **2**(1) (2015) 1004
17. 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: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, IEEE (2014) 53–56
18. Maninis, K.K., Pont-Tuset, J., Arbeláez, P., Van Gool, L.: Deep retinal image understanding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2016) 140–148
19. Zilly, J., Buhmann, J.M., Mahapatra, D.: Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Computerized Medical Imaging and Graphics **55** (2017) 28–41
20. Zilly, J.G., Buhmann, J.M., Mahapatra, D.: Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images. In: International Workshop on Machine Learning in Medical Imaging, Springer (2015) 136–143