UOLO - Automatic Object Detection and Segmentation in Biomedical Images

Teresa Araújo[1,2(✉)], Guilherme Aresta[1,2(✉)], Adrian Galdran[1], Pedro Costa[1], Ana Maria Mendonça[1,2], and Aurélio Campilho[1,2]

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
{tfaraujo,guilherme.m.aresta,adrian.galdran,pvcosta}@inesctec.pt
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
{amendon,campilho}@fe.up.pt

Abstract. We propose UOLO, a novel framework for the simultaneous detection and segmentation of structures of interest in medical images. UOLO consists of an object segmentation module which intermediate abstract representations are processed and used as input for object detection. The resulting system is optimized simultaneously for detecting a class of objects and segmenting an optionally different class of structures. UOLO is trained on a set of bounding boxes enclosing the objects to detect, as well as pixel-wise segmentation information, when available. A new loss function is devised, taking into account whether a reference segmentation is accessible for each training image, in order to suitably backpropagate the error. We validate UOLO on the task of simultaneous optic disc (OD) detection, fovea detection, and OD segmentation from retinal images, achieving state-of-the-art performance on public datasets.

Keywords: Detection · Segmentation · Biomedical images
Eye fundus images · Convolutional neural networks

1 Introduction

Detection and segmentation of anatomical structures are central medical image analysis tasks since they allow to delimit Regions-Of-Interest (ROI), create landmarks and improve feature collection. In terms of segmentation, Deep Fully-Convolutional (FC) Neural Networks (NNs) achieve the highest performance on a variety of images and problems. Namely, U-Net [1] has become a reference model – its autoencoder structure with skip connections enables the propagation from the encoding to the decoding part of the network, allowing a more robust multi-scale analysis while reducing the need for training data.

Similarly, Deep Neural Networks (DNNs) have become the technique of choice in many medical imaging detection problems. The standard approach is to use

T. Araújo and G. Aresta—Authors contributed equally to this work.

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 165–173, 2018.
https://doi.org/10.1007/978-3-030-00889-5_19

networks pre-trained on large datasets of natural images as feature extractors of a detection module. For instance, Faster-R CNN [2] uses these features to identify ROIs via a specialized layer. ROIs are then pooled, rescaled and supplied to a pair of Fully-Connected NNs responsible for adjusting the size and label the bounding boxes. Alternatively, YOLOv2 [3] avoids the use of an auxiliary ROI proposal model by directly using region-wise activations from pre-trained weights to predict coordinates and labels of ROIs.

When a ROI has been identified, the segmentation of an object contained on it becomes much easier. For this reason, the combination of detection and segmentation models into a single method is being explored. For instance, Mask-R CNN [4] extends Faster-R CNN with the addition of FC layers after its final pooling, enabling a fine segmentation without a significant computational overhead. In this architecture, the segmentation and detection modules are decoupled, *i.e.* the segmentation part is only responsible for predicting a mask, which is then labeled class-wise by the detection module. However, despite the high performance achieved by Mask-R CNN in computer vision, its application to medical image analysis problems remains limited. This is due to the large requirement of data annotated at a pixel level, which is usually not available in medical applications.

In this paper we propose UOLO (Fig. 1), a novel architecture that performs simultaneous detection and segmentation of structures of interest in biomedical images. UOLO harvests the best of its individual detection and segmentation modules to allow robust and efficient predictions even when few training data is available. Moreover, training UOLO is simple since the entire network can be updated during back-propagation. We experimentally validate UOLO on eye fundus images for the joint task of fovea (FV) detection, optic disc (OD) detection, and OD segmentation, where we achieve state-of-the-art performance.

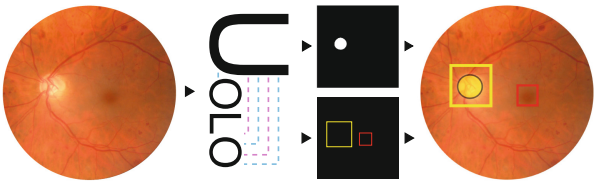

Fig. 1. Using UOLO for fovea detection and optic disc detection and segmentation.

2 UOLO Framework

2.1 Object Segmentation Module

For object segmentation we consider an adapted version of the U-Net network presented in [1]. U-Net is composed of FC layers organized on an auto-encoder scheme, which allows to obtain an output of the same size of the input, thus

enabling pixel-wise predictions. Skip connections between the encoding and decoding parts are used for avoiding the information loss inherent to encoding. The model's upsampling path includes a large number of feature channels with the aim of propagating the multi-scale context information to higher resolution layers. Ultimately, the segmentation prediction results from the analysis of abstract representations of the images from multiple scales, with the majority of the relevant classification information being available on the decoder portion of the network due to the skip connections. We modify the network by adding batch normalization after each convolutional layer, and replacing the pooling layers by convolutions with stride. The soft intersection over union (IoU) is used as loss:

$$\mathcal{L}_{\text{U-Net}} = 1 - \text{IoU} = 1 - \frac{\sum I_t \circ I_p}{\sum (I_t + I_p) - \sum I_t \circ I_p}, \tag{1}$$

where I_t and I_p are the ground truth mask and the soft prediction mask, respectively, and $\circ$ is the Hadamard product.

2.2 Object Detection Module

For object detection we take inspiration from YOLOv2 [3], a network composed of: (1) a DNN that extracts features from an image (F_{YOLO}); (2) a feature interpretation block that predicts both labels and bounding boxes for the objects of interest (D_{YOLO}). YOLOv2 assumes that every image's patch can contain an object of size similar to one of various template bounding boxes (or *anchors*) computed *a priori* from the objects' shape distribution in the training data.

Let the output of F_{YOLO} be a tensor F of shape $S \times S \times N$, where S is the dimension of the spatial grid and N is the number of maps. F_{YOLO} convolves and reshapes F into Y, a tensor of shape $S \times S \times A \times (C + 5)$, where A is the number of anchors, C is the number of object classes, and 5 is the number of variables to be optimized: center coordinates x and y, width w, height h, and the confidence c (how likely is the bounding box to be an object) of the bounding boxes. For each anchor A_k in Y, the value of each feature map element $m_{i,j}$ is responsible for adjusting a property of the predicted bounding box $\hat{b}$,

$$\begin{aligned} (\hat{b}_x, \hat{b}_y) &= (\sigma(\hat{x}) + x_{i,j,k}, \sigma(\hat{y}) + y_{i,j,k}) \\ (\hat{b}_w, \hat{b}_h) &= (w_{i,j,k} e^{\hat{w}}, h_{i,j,k} e^{\hat{h}}) \\ \text{confidence} &= \sigma(\hat{c}) \end{aligned} \tag{2}$$

where σ is a sigmoid function. YOLOv2 is trained by optimizing the loss function:

$$\mathcal{L}_{\text{YOLO}} = \lambda_1 \mathcal{L}_{\text{centers}} + \lambda_2 \mathcal{L}_{\text{dimensions}} + \lambda_3 \mathcal{L}_{\text{confidence}} + \lambda_4 \mathcal{L}_{\text{classes}} \tag{3}$$

where λ_i are predefined weighting factors, $\mathcal{L}_{\text{centers}}$, $\mathcal{L}_{\text{dimensions}}$ and $\mathcal{L}_{\text{confidence}}$ are mean squared errors, and $\mathcal{L}_{\text{classes}}$ is the cross-entropy loss. Each loss term penalizes a different error: (1) $\mathcal{L}_{\text{centers}}$ penalizes the error in the center position of the cells; (2) $\mathcal{L}_{\text{dimensions}}$ penalizes the incorrect size, *i.e.* height and width, of the bounding box; (3) $\mathcal{L}_{\text{confidence}}$ penalizes the incorrect prediction of a box presence; (4) $\mathcal{L}_{\text{classes}}$ penalizes the misclassification of the objects.

2.3 UOLO for Joint Object Detection and Segmentation

UOLO framework for object detection and segmentation is depicted in Fig. 2, where the segmentation module itself is used as a feature extraction module, adopting the role of F_{YOLO}, and serving as input for the localization module D_{YOLO}. The intuition behind this design is that the abstract representation learned by the decoding part of U-Net contains multi-scale information that can be useful not only to segment objects, but also to detect them. In addition, the class of objects that UOLO can detect is not limited to those for which segmentation ground-truth is available.

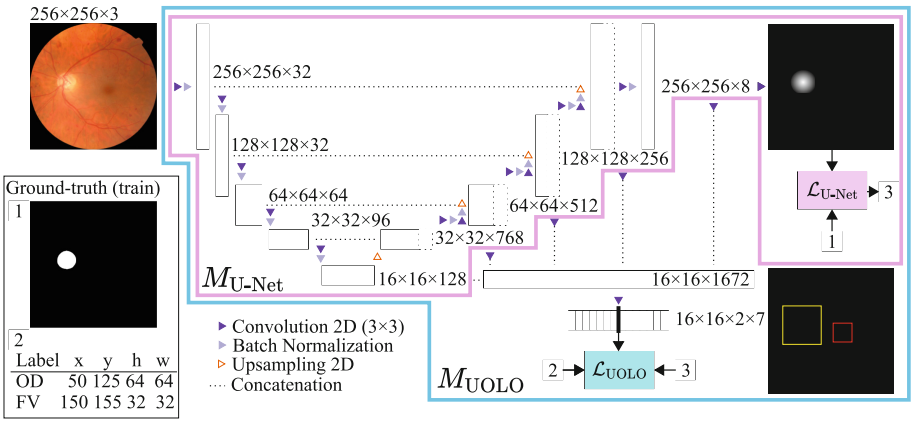

Fig. 2. UOLO framework, nesting an U-Net responsible for segmentation and feature extraction for an YOLOv2-based detector. $M_{\text{U-Net}}$: U-net part; M_{UOLO}: full UOLO.

Let $M_{\text{U-Net}}$ be an U-Net-like network that, given pairs of images and binary masks, can be trained for performing segmentation by minimizing $\mathcal{L}_{\text{U-Net}}$ (Eq. 1). $M_{\text{U-Net}}$ has a second output corresponding to the concatenation of the down-sampled decoding maps with its bottle neck (last encoder layer). The resulting tensor corresponds to a set of multi-scale representations of the original image that are supplied to the object detection block D_{YOLO}, which, by its turn, can be optimized via $\mathcal{L}_{\text{YOLO}}$, defined in Eq. 3. D_{YOLO} and $M_{\text{U-Net}}$ are then merged by concatenation into M_{UOLO}, a single model that can be optimized by minimizing the addition of the corresponding loss functions:

$$\mathcal{L}_{\text{UOLO}} = \mathcal{L}_{\text{YOLO}} + \mathcal{L}_{\text{U-Net}} \tag{4}$$

Thanks to the straightforward definition of the loss function in Eq. (4), M_{UOLO} can be trained with a simple iterative scheme detailed in Algorithm 1. In essence, $\mathcal{L}_{\text{U-Net}}$ is updated only when segmentation information is available. However, a global weight update is performed at every step based on the prediction error backpropagation. Furthermore, the outlined training scheme allows

Algorithm 1. Loss computation scheme of UOLO. $M_{\text{U-Net}}$: U-net part from the UOLO model; M_{UOLO}: full UOLO model; b_{det}: batches of images with objects' bounding boxes ground truth; b_{seg}: batches of images with segmentation ground truth.

$\mathcal{L}_{\text{U-Net}} \leftarrow 1$
for each training step **do**
  $M_{\text{UOLO}} \leftarrow \texttt{train}(M_{\text{UOLO}}, b_{\text{det}}, \mathcal{L}_{\text{UOLO}})$ {train on n_{det} batches from b_{det}, backpropagating $\mathcal{L}_{\text{UOLO}}$};
  $\texttt{update}(\mathcal{L}_{\text{YOLO}})$
  $M_{\text{U-Net}} \leftarrow \texttt{train}(M_{\text{U-Net}}, b_{\text{seg}}, \mathcal{L}_{\text{U-Net}})$ {train on n_{seg} batches from b_{seg}, backpropagating $\mathcal{L}_{\text{U-Net}}$}
  $\texttt{update}(\mathcal{L}_{\text{U-Net}})$
  $\mathcal{L}_{\text{UOLO}} \leftarrow \mathcal{L}_{\text{YOLO}} + \mathcal{L}_{\text{U-Net}}$

for a different number of strong (pixel-wise) and weak (bounding boxes) annotations, easing its application to medical images.

3 Experiments and Results

3.1 Datasets and Experimental Details

We test UOLO on 3 public eye fundus datasets with healthy and pathological images: (1) Messidor [5] has 1200 images (1440 × 960, 2240 × 1488 and 2304 × 1536 pixels, 45° field-of-view (FOV)), 1136 having ground truth (GT) for OD segmentation and FV centers[1]; (2) IDRID[2] training set has 413 images (4288 × 2848 pixels, 50° FOV) with OD and FV centers and 54 with OD segmentation; (3) DRIVE [6] has 40 images (768 × 584 pixels, 45° FOV) with OD segmentation[3].

All images are cropped around the FOV (determined via Otsu's thresholding) and resized to 256 × 256 pixels. The side of the square GT bounding boxes is set to 32 and 64 for the FV and OD following their relative size in the image. For training, n_{det} and n_{seg} (Algorithm 1) are set to 256 and 32, respectively. Online data augmentation, a mini-batch size of 8, and the Adam optimizer (learning rate of 1e–4) were used for training, while 25% of the data was kept for validation. The bounding box with highest confidence for each class is kept. The predicted soft segmentations are binarized using a threshold of 0.5.

The OD segmentation is evaluated with IoU and Sorensen-Dice coefficient overlap metrics. The detection is evaluated in terms of mean euclidean distance (ED) between the prediction and the GT. We also evaluate ED relatively to the OD radius, $\bar{D}$ [7,8]. Finally, detection success, S_{1R}, is assessed using the maximum distance criteria of 1 OD radius.

[1] http://www.uhu.es/retinopathy.
[2] https://idrid.grand-challenge.org/, available since January 20, 2018.
[3] https://sites.google.com/a/uw.edu/src/useful-links.

3.2 Results and Discussion

We evaluate UOLO both inter and intra-dataset-wise. For inter-dataset experiments, UOLO was trained on Messidor and tested in the other datasets whereas for intra-dataset studies stratified 5-fold cross-validation was used. We do not extensively optimize the training parameters to verify how robust UOLO is when dealing with segmentation and detection simultaneously. Table 1 shows the results of UOLO for the OD detection and segmentation and FV detection tasks, Table 2 compares our performance with state-of-the-art methods and Fig. 3 shows two prediction examples in complex detection and segmentation cases.

UOLO achieves equal or better performance in comparison to the state-of-the-art on both detection and segmentation tasks (IoU 0.88 ± 0.09 on Messidor) in a single step prediction. Furthermore, the proposed network is robust even in inter-dataset scenarios, maintaining both segmentation and detection performances. This indicates that the abstract representations learned by UOLO are

Table 1. UOLO performance on optic disc (OD) detection and segmentation and fovea (FV) detection. n: number of training images for detection and segmentation.

Datasets		**n**		**OD seg.**		**OD det.**		**FV det.**	
Train	**Test**	seg.	det.	IoU	Dice	$\bar{D}$	S_{1R}	$\bar{D}$	S_{1R}
Messidor		680	680	0.88	0.93	0.111	99.74	0.121	99.38
Messidor		100	680	0.87	0.93	0.114	99.74	0.114	97.89
IDRID		30	280	0.88	0.93	0.095	99.79	0.288	93.78
Messidor	IDRID	852	852	0.84	0.91	0.138	99.78	0.403	89.06
Messidor	DRIVE	852	852	0.82	0.89	0.171	97.50	-	-

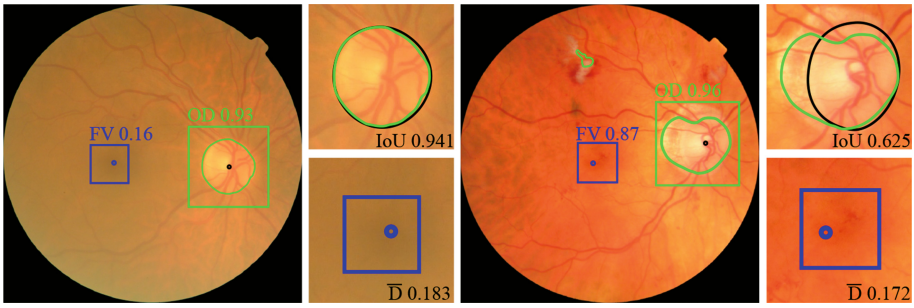

Fig. 3. Examples of results of UOLO on Messidor images. Green curve: segmented optic disc (OD), green and blue boxes: predicted OD and FV locations, respectively; black curve: ground truth OD segmentation; black and blue dots: ground truth OD and FV locations, respectively. The object detection confidence is shown next to each box. IoU (intersection over union) and normalized distance ($\bar{D}$) values are also shown. (Color figure online)

highly effective for solving the task at hands. It is worth noting that our segmentation and detection performances do not alter significantly even when UOLO is trained with only 15% of the pixel-wise annotated images. This means that UOLO does not require a significant amount of pixel-wise annotations, easing its application on the medical field, where these are expensive to obtain.

Our results also suggest that UOLO is capable of using multi-scale information (*eg.* relative position to the OD or vessel tree) to perform predictions. For instance, Fig. 3 shows UOLO's output for two Messidor images, illustrating that the network is capable of detecting the FV in a low contrast scenario. On the other hand, the segmentation and detection processes are not completely interdependent, as expected from the proposed training scheme, since the network segments OD confounders outside the detected OD region. Another advantage of UOLO is that these segmentation errors are easily correctable by limiting the pixel-wise predictions to the found OD region. Unlike hand-crafted feature-based methods, UOLO does not require an extensive parameter tunning and it is simple to extend to different applications.

We also evaluate U-Net ($\mathcal{M}_{\text{U-Net}}$, Fig. 2) for OD segmentation and YOLOv2 (with a pretrained Inceptionv3 as feature extractor) for OD and FV detection (Table 2). The training conditions were set as in UOLO. UOLO segmentation performance is practically the same as U-Net, whereas the detection drops slightly when comparing with YOLOv2, mainly for OD detection. However, one has to consider the trade-off between computational burden and performance, since UOLO network has 23 347 063 parameters, whereas U-Net has 15 063 985 and YOLOv2 has 21 831 470, being that for training U-Net and YOLO a total of 36 895 455 parameters have to be optimized (60% increase).

Table 2. State-of-the-art for OD detection and segmentation and FV detection.

(a) OD segmentation

Dataset	Messidor		DRIVE	
Method	IoU	Dice	IoU	Dice
UOLO	0.88	0.93	0.82	0.89
U-Net	0.88	0.93	0.81	0.88
[9]	0.91	-	-	-
[10]	0.89	0.94	-	-
[11]	0.84	-	0.81	-
[12]	0.82	-	0.72	-
[13]	-	-	0.82	-

(b) OD and FV detection

Task	**OD det.**				**FV detection**	
Dataset	Messidor		DRIVE		Messidor	
Method	ED	S_{1R}	ED	S_{1R}	ED	S_{1R}
UOLO	9.40	99.74	8.13	97.5	10.44	99.38
YOLOv2	6.86	100	7.20	97.5	9.01	100
[14]	-	97	-	-	-	96.6
[8]	-	-	-	-	16.09	98.24
[7]	-	-	-	-	20.17	98.24
[15]	-	98.83	-	-	-	-
[16]	23.17	99.75	15.57	100	34.88	99.40
[10]		99.75	-	-	-	-

4 Conclusions

We presented UOLO, a novel network that performs joint detection and segmentation of objects of interest in medical images by using the abstract representations learned by U-Net. Furthermore, UOLO can detect objects from a different class for which segmentation ground-truth is available.

We tested UOLO for simultaneous fovea detection and optic disk detection and segmentation, achieving state-of-the-art results. This network can be trained with relatively few images with segmentation ground-truth and still maintain a high performance. UOLO is also robust to inter-dataset settings, thus showing great potential for applications in the medical image analysis field.

Acknowledgements. T. Araújo is funded by the FCT grant SFRH/BD/122365/2016. G. Aresta is funded by the FCT grant SFRH/BD/120435/2016. This work is funded by the ERDF European Regional Development Fund, Operational Programme for Competitiveness and Internationalisation - COMPETE 2020, and by National Funds through the FCT - project CMUP-ERI/TIC/0028/2014.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
2. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Patt. Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
3. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. arXiv (2016)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv (2017)
5. Decenciere, E., Zhang, X., Cazuguel, G.: Feedback on a publicly distributed image database: the Messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)
6. Staal, J., Niemeijer, M., Viergever, M.A., Ginneken, B.V.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
7. Gegundez-Arias, M.E., Marin, D., Bravo, J.M., Suero, A.: Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques. Comput. Med. Imaging Graph. **37**(5–6), 386–393 (2013)
8. Aquino, A.: Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features. Comput. Biol. Med. **55**, 61–73 (2014)
9. Dai, B., Wu, X., Bu, W.: Optic disc segmentation based on variational model with multiple energies. Patt. Recogn. **64**, 226–235 (2017)
10. Dashtbozorg, B., Mendonça, A., Campilho, A.: Optic disc segmentation using the sliding band filter. Comput. Biol. Med. **56**, 1–12 (2015)
11. Roychowdhury, S., Koozekanani, D.D., Kuchinka, S.N., Parhi, K.K.: Optic disc boundary and vessel origin segmentation of fundus images. IEEE J. Biomed. Health Inform. **20**(6), 1562–1574 (2016)

12. Morales, S., Naranjo, V., Angulo, U., Alcaniz, M.: Automatic detection of optic disc based on PCA and mathematical morphology. IEEE Trans. Med. Imaging **32**(4), 786–796 (2013)
13. Salazar-Gonzalez, A., Kaba, D., Li, Y., Liu, X.: Segmentation of blood vessels and optic disc in retinal images. IEEE J. Biomed. Health Inform. **18**(6), 1874–1886 (2014)
14. Al-Bander, B., Al-Nuaimy, W., Williams, B.M., Zheng, Y.: Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. Biomed. Sig. Process. Control **40**, 91–101 (2018)
15. Aquino, A., Gegúndez-arias, M.E., Marín, D.: Detecting the optic disc boundary in digital fundus feature extraction techniques. IEEE Trans. Med. Imaging **29**(11), 1860–1869 (2010)
16. Kamble, R., Kokare, M., Deshmukh, G., Hussin, F.A., Mériaudeau, F.: Localization of optic disc and fovea in retinal images using intensity based line scanning analysis. Comput. Biol. Med. **87**, 382–396 (2017)