2019 IEEE International Conference on Multimedia and Expo (ICME)

SEMANTIC-EMBEDDING AND SHAPE-AWARE U-NET FOR ULTRASOUND EYEBALL SEGMENTATION

Fanchao Lin,Chuanbin Liu,Hongtao Xie,Zheng-Jun Zha,Yongdong Zhang

School of Information Science and Technology,
University of Science and Technology of China, Hefei 230026, China
{lfc1995,lcb592}@mail.ustc.edu.cn,{htxie,zhazj,zhyd73}@ustc.edu.cn

ABSTRACT

Segmentation of eyeball region from ultrasound images is a new research direction for the diagnosis of ophthalmic diseases. Despite the advantages of convenience and cheapness, ultrasound images bring more noise and fuzzy contour compared with other medical images. Existing methods fail to give a segmentation with reasonable eyeball shape, especially when the contour is ambiguous. In this paper, we propose a novel framework based on convolutional neural network, named semantic-embedding and shape-aware U-Net, to deal with the segmentation in blurred images. A signed distance field is used as label instead of the traditional binary mask label to add shape prior in network. The applying of semantic embedding modules fuses semantic information between different stages of the network. Experimental results show that our method improves the ability to segment image with blurred edges and outperforms existing methods in the accuracy of segmentation.

Index Terms— semantic embedding, SDF label, feature fusion, ultrasound image segmentation, eyeball reconstruction

1. INTRODUCTION

High myopia is a common disease in modern society, which can cause fundus abnormalities like posterior scleral staphyloma, retinal atrophy and even retinal detachment. To prevent the disease or make correct diagnosis, comprehensive observation of the eyeball morphology is necessary. Semantic segmentation of eyeball region can be used for eyeball model reconstruction, the measurement of axial length and volume of macular fovea. Thus it is very meaningful to make segmentation of eyeball for the study of eyeball morphology and diagnosis of ophthalmic diseases.

There has been method proposed for the segmentation of pig eyeball in micro-CT images [1]. Though CT detection has the advantage of high resolution and low noise, it is harmful to human body and cannot be a routine choice. There are also works for the Retinal Vessel Image Segmentation [2], but they are not comprehensive enough to reflect the overall situation of the whole eyeball. In this paper, however, we propose method for the segmentation of eyeball in ultrasound images. Taking the advantage of cheapness, convenience and the ability to reflect the whole eyeball, segmentation of eyeball ultrasound images is thought to be a better choice for diagnosis.

With the recent development of deep learning, fully convolutional network (FCN) [3] has become the main framework for medical image segmentation task. U-net [4], as one of the variants of FCN, is proved to be an excellent network for segmentation in ultrasound images. The encoder-decoder structure and skip connections in U-net make it able to integrate features at different stages. Chen at al. [5] use a cascade structure to bridge two U-nets together, but they do not maintain the shape of segmentation results. Al Arif at al. [6] train their network with a loss function that computes the error in the shape domain, where an singular value decomposition (SVD) of matrix is needed as well as a high time cost. Though many approaches based on U-net achieve good performance in their own tasks, they are not designed for images with ambiguous and discontinuous boundary like eyeball ultrasound images. The main shortcoming of them is the lack of efficient shape prior and fusion of semantics due to the limitation of U-net itself.

Aiming at the correct segmentation in blur images, semantic fusion and shape prior are necessary to be taken into account. The semantic fusion helps to detect the right class and position and the shape prior gives a reasonable constraint to the appearance of segmentation results. In our work, the shape information is learnt more efficiently using a signed distance field (SDF) label. Besides, a semantic embedding module (SEM) is applied to embed semantic information between stages for a better understanding of the whole image. Making use of them, as is shown in Fig. 1, we present a new method, semantic-embedding and shape-aware U-net (SS-Unet), for the eyeball segmentation. We demonstrate that the proposed approach outperforms the state-of-the-art method, particularly on more challenging cases with a blur contour.

The contributions of this paper are three-fold:

- Applying a semantic embedding module to U-net for

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00158

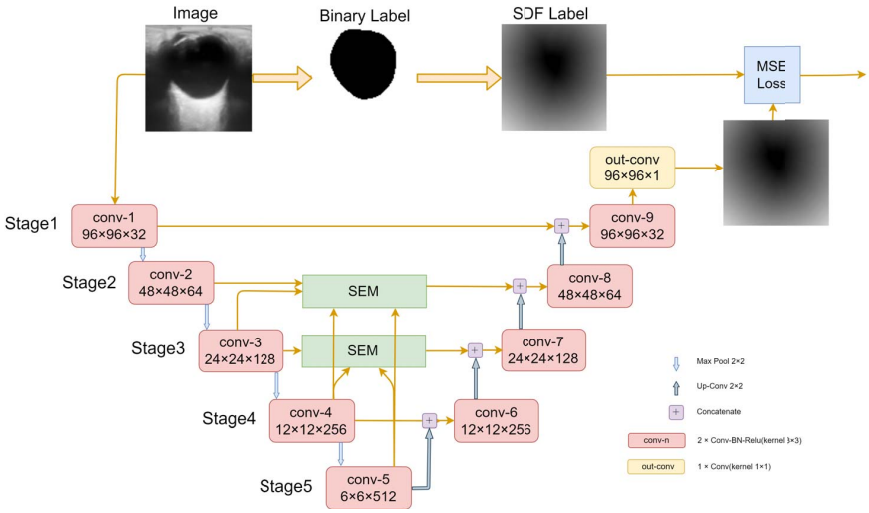


Fig. 1. The overall architecture of our approach

better fusion of semantics and achieve higher segmentation accuracy.

- The using of SDF label instead of mask label for training to help the network learn about shape information.
- To the best of our knowledge, this is the first work using a CNN-based model to make automatic segmentation for eyeball ultrasound images.

The rest of this paper is organized as follows. Section 2 reviews works related to the CNN-based segmentation in medical imaging. Section 3 details the proposed segmentation approach. Experimental results are shown and analyzed in Section 4. Conclusions are made in Section5.

2. RELATED WORK

In recent years, inspirational renovations have been witnessed in medical image segmentation, due to the development of convolutional neural networks (CNN) based methods [7] [8] [9] [10]. As one of the earliest works, the emergence of FCN makes it possible to make dense prediction end to end. After that, more and more CNN-based methods are used to solve the segmentation task.

2.1. U-net structure

Comparing with nature images, the eyeball ultrasound images have more noise and relatively small amount of data. U-net has the characteristic of multi-scale feature fusion and can be trained on a small dataset, so we choose the U-net structure as a basic backbone. The framework of U-net can be seen as an encoder-decoder architecture. The encoder path reduces the spatial dimension of an input image to a smaller version and gets advanced semantic information as well as boarder receptive field. The decoder path resumes the spatial resolution to the original size and is combined with stages from encoder path by skip connection. With simple structure and fewer parameters, U-net shows great capability of segmenting different target objects in different medical image modalities.

2.2. Shape prior

Due to the inherent noise and low contrast in ultrasound images, accurate segmentation is hard to be achieved especially on regions with blur boundary. To overcome the difficulties, it is critical to add shape prior knowledge to network and constrain the output segmentation to be a reasonable shape.

Unet-S [11] adds an extra term in the loss function to penalize points which are split incorrectly. SP-net [5] adds a shape predictor branch at the end of the U-net so that the network can learn to predict the shape parameters b after an singular value decomposition(SVD) process. However, solving the SVD of a matrix needs a time complexity of $O(n^3)$ and cost too much time when both the number and size of the image is very large. Noticing that the signed distance field (SDF) mentioned in SP-net, which assign values to points according to their distance to the boundary curve, is already a good representation of shape information, we directly use the SDF of original mask as label in our approach. By this means, the calculation of SVD is avoided, while the networks ability to split regions based on shape prior is still maintain.

2.3. Feature fusion

Making efficient use of image features is important in segmentation tasks.A classic approach is to build an en encoder-

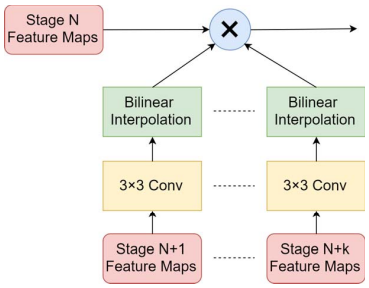
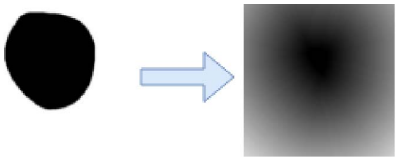


Fig. 2. Semantic embedding module(SEM)



Binary label SDF label

Fig. 3. the original mask label(left) and its SDF matrix label(right). For convenience, elements of the SDF matrix is added with a value of 80 and visualized as an image.

decoder structure like U-net and fuse feature maps at different levels. The encoder path aims to extract features and the decoder way is to resume resolution for the convenience of segmentation. On the way to extract information, feature maps at low level have high resolution and little global semantic, which means more details and noise at the same time. As the network goes deeper, the feature maps maintain more advanced semantic information, but their size becomes too small to resume a segmentation map.

Though U-net has skip connections to combine features together, we argue that it lacks a direct way for the flow of semantics. Zhang et al. [12] add semantic embedding branch on the skip connection path for the segmentation of nature images. Following them, we also use modules to embed semantic information from high-level stages to low-level stage, thus the semantic information can be combined with high resolution low-level feature closely.

In this paper, we proposed a semantic-embedding and shape-aware U-net for better fusion of semantic information and effective shape constraint of the segmentation results.

3. METHOD

In this section, we introduce the proposed semantic embedding and shape-aware U-net. Before the concatenation between two sides of the U-shaped structure, the semantic embedding module is added to fuse semantics from high-level stages to low-level stage. Furthermore, the shape-aware of input images is achieved by using an SDF label for training. The overall framework of our method is illustrated in Fig. 1 for easy reference.

3.1. Overall framework

The proposed network can be viewed as 3 sections: 1) forward feature extracting path (left), 2) semantic fusion path (middle) and 3) path to resume resolution (right).

The forward feature extracting path has 5 stages, each of which has 2 convolution blocks, one to increase feature maps and one for further extracting features. Every convolution block contains a convolution layer with kernel size of 3×3, a batch normalization layer and an activation layer using ReLU function. Max-pooling is done following the two convolution blocks to reduce the size of feature maps as well as enlarge the respective field.

3.2. Semantic embeded U-net

The feature fusion path consists of several semantic embedding modules (SEM) between stages. Each SEM takes feature maps from a relatively low-level stage and all higher-level feature maps as input. Firstly, The feature maps from higher stages are passing through a convolution layer with kernel size 3×3 to reduce the number of channels by half and further refine the semantics. Then a bilinear layer is applied to resume the resolution, so that both the map size and the number of channels equal to the low stage. Finally, feature maps from every stage are multiplied element to element as output for follow-up concatenation. The structure of SEM is illustrated in Fig. 2.

Low-level features mainly contain local information such as edge and texture. While it can be helpful to recover details of segmentation, the small respective field and little semantic information make it hard to reflect well about the entire object. Higher-level feature with advanced semantic information is necessary for a right detection and location. The SEM uses operations of bilinear interpolation and multiplication to integrate together the local information and semantics without increasing the number of channels. Through such a module, different from the original skip connection, the features and semantics are combined in a more direct way. Thus the correct global recognition and accurate local details can be aligned at the same time.

The path to resume resolution also consists of 5 stages, each stage upsamples the feature maps through 2 convolution layers. The feature maps after upsampling are concatenated with features maps output by SEM accordingly. At the end of the path is a convolution layer with kernel of size 1×1 to get a one-channel map as prediction.

Table 1. Ablation study of the proposed method

Metrics	DSC			pA(%)			PCD			FF(%)
Methods	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	
U-net [4]	0.965	0.953	0.014	97.86	97.01	1.81	1.202	5.745	7.122	38.10
U-net+SEM	0.972	0.959	0.013	98.32	97.65	1.88	1.090	4.087	5.449	33.93
U-net+SDF	0.975	0.963	0.011	98.76	98.67	1.46	0.674	1.183	2.503	8.33
SS-Unet(our method)	**0.976**	**0.963**	**0.011**	**98.83**	**98.81**	**1.08**	**0.665**	**0.891**	**1.315**	**5.95**

3.3. Shape-aware SDF label

In order to introduce an effective shape prior to the model, the signed distance function (SDF) matrix of original binary label is used rather than the binary label itself for training. Function of the signed distance field $\varphi(P)$ can be written as follows:

$$\varphi(P) = \begin{cases} d(P,\ C) & if\ P \in \Omega_{outside} \\ -d(P,\ C) & if\ P \in \Omega_{inside} \end{cases} \tag{1}$$

where P is a point in image, C is the boundary curve according to the binary mask label, D(P, C) means the shortest distance between the point and the boundary curve, $\Omega_{outside}$ and $\in\ \Omega_{inside}$ are regions inside and outside the boundary, respectively.

Fig. 3 shows the conversion form a mask label to an SDF label. The original mask label is a binary image with background as white and the eyeball region as black. While we mainly concern about the boundary of the eyeball, the pixel points inside the boundary are given a value same as the boundary points in original binary label, so that they contribute to the loss equally in training. Owing to the fact that pixels around the edge are rather less than those inside the edge or in the background, we may get a result with high score mainly due to the correct classification inside the eyeball region, but not the boundary accuracy, which we really concern about.

We use SDF label as the shape prior to represent the relationship between boundary pixels and other pixels based on their spatial position. If we directly use a network to predict the boundary of the eyeball, it is likely to get discontinuous curves due to the ambiguous or discontinuous part of original ultrasound images. Applying the SDF matrix as label, we can still use a segmentation framework while letting the network learn to decide the boundary actually. In this way, we can emphasize the pixels around the boundary and get successive boundary curve from the segmentation result as well.

4. EXPERIMENTS

4.1. Experimental setup

The datasets are collected from the Beijing Tongren Hospital, consisting of 668 ultrasound images based on 6 different eyeballs. We take 500 images for training and 168 for evaluation. Data augmentation is applied by horizontal flipping, random cropping and rotation from -5° to 5° so that we get 4290 images as training set finally.

Table 2. Comparison of methods using SDF label

Metrics(mean)	DSC	pA(%)	PCD	FF(%)
W-net [5]	0.950	96.51	5.495	35.71
U-net [4]	0.953	97.01	5.745	38.10
SP-net [6]	0.958	98.01	1.707	14.29
W-net+SDF	0.961	98.60	1.333	**6.55**
U-net+SDF	**0.963**	**98.67**	**1.183**	8.33

Experiments are performed using the Pytorch library. AdamGrad optimizer and mean squared error(MSE) loss function are used.

To evaluate the segmentation results, performance of pixel level and curve level are both taken into account. For pixel level, the prediction results are compared with the ground truth to compute the number of pixels detected as true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Then two metrics: Dice similarity coefficient (DSC) and Pixel-wise accuracy (pA) are calculated as follow:

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

$$pA = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

For evaluation of curve level, we use point-to-curve distance (PCD) which computes the mean pixel distance between predicted curve and the ground truth, and fit failure rate which is the rate of results with a PCD over than 2 pixels.

4.2. Experimental results

In this section, experiments are conducted to evaluate the performance of the proposed SS-Unet. Overall results of ablation study are listed in Table 1.

4.2.1. Evaluation of using SDF label

In order to validate the generality of the SDF label, we use it both in U-net [4] and W-net [5] . Experimental results are

Table 3. Results of using SEM at different stages

Stage1	Stage2	Stage3	Stage4	PCD
				1.183
	✓	✓		**0.891**
✓	✓	✓		1.159
	✓	✓	✓	1.092
✓	✓	✓	✓	1.085

shown in Table 2. About half of data in our test set are images without an eyeball region or with unclear eyeball regions, which the raw U-net or W-net fails to detect. The networks trained with the SDF label can learn about the shape and position correctly and will not mistake background for the eyeball region as they used to be. Thus the point-to-curve distance is reduced significantly.

We also compare the performance with SPNet [6], which processes the SDF matrix with singular value decomposition (SVD) to extract shape parameters rather than use it as label. As shown in Table 2, just applying SDF matrix as label can get better result than further transforming it using SVD as SP-Net. The SVD itself is a way to extract the primary information of a matrix, which cannot provide information more than the original matrix. Using directly the SDF label to represent the shape of region, our network avoids the complex calculation of SVD and still apperceive the shape information.

4.2.2. Ablation study on SEM

By using the semantic embedding module between stages, we enhance the ability of network to fix the right position and further improve the segmentation accuracy. To examine the effectiveness of SEM, we select several subsets of combined stages and use SEM at these subsets to retrain the whole system.

Results are shown in Table 3. It can be seen that using SEM can achieve higher performance, but it is not the more the better. Using SEM at stage 2 and 3 can achieve the best performance, but it leads to performance degradation when expanding to stage 1 and 4. From these results, we can infer that there are semantic gaps between stages. The information in stage 1 contains many low-level features and is hard to fuse directly with stage 4 or 5, which has the highest level semantics. On the other hand, the stage 4 and stage 5 are concatenated directly, thus it is redundant to use SEM between them. Based on the above reasons, we only use SEM in stage 2 and 3 in the proposed SS-Unet.

4.2.3. Overall results and discussion

The comparison between the original U-net and networks after adding different part of our method is illustrated in Fig. 4. Our proposed method has significant improvement in blur images by using the SDF label and achieves more accurate positioning through the SEM. The overall quantitative evaluation results of all the 168 test patches are reported at Table 1. SEM can reduce the fit failure rate by about 4.2%, the using of SDF label can greatly reduce it to 8.33% and the combination of them can further improve the performance.

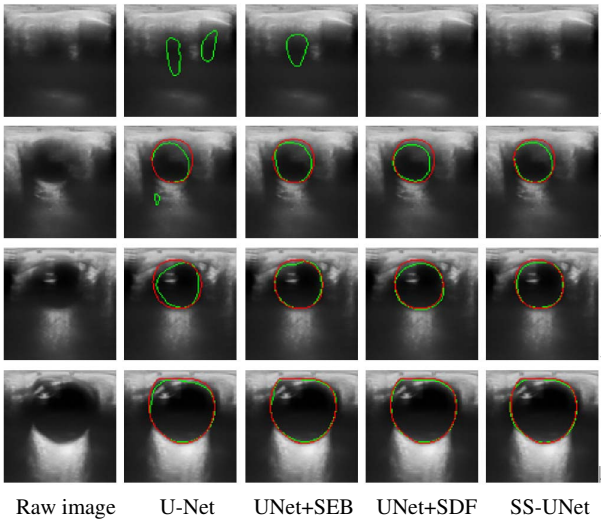


Fig. 4. Comparison between different methods. The first row is an image without an eye region, the 2nd to 3th rows are images with blurred boundary and the 4th row is an image with clear boundary. Prediction and the ground-truth are in green and red, respectively.

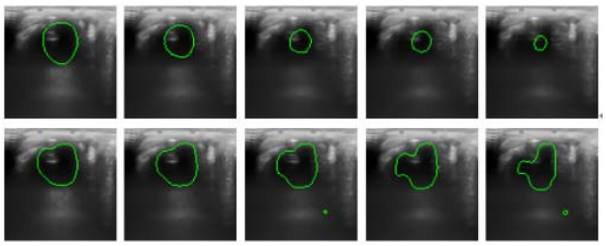


Fig. 5. Segmentation results from a continuous sequence of images at the bottom of eyeball. The top column belongs to SS-Unet and the bottom column belongs to U-net [4].

As can be seen in Fig. 5, the original U-net tends to produce big unregular segmentation in blurred images, which is unreasonable actually. The result of our proposed method has a transversion from large to small, which is a reasonable tendency when the ultrasound sections are collected from two poles of the eyeball.

The results of 3D reconstruction are shown in Fig. 6. Using the proposed method, we can get the segmentation result needed for 3D reconstruction in an end-to-end way without

(a) Ground-truth (b) U-net [4] (c) SS-Unet

Fig. 6. Results of 3D reconstruction

manually selecting part of results.

5. CONCLUSIONS

This paper presents a new method, SS-Unet, for the eyeball region segmentation in ultrasound images. Through the using of SDF label instead of the binary mask label, the neural network is trained to be shape-aware and predicts eyeball region reasonably. Furthermore, the advantages of local information at low stages and semantics at high stages complement each other by adding semantic embedding modules to the network. Experimental results demonstrate that the proposed SS-Unet has strong ability for segmenatation in blur images and is superior to previous state-of-the-art approaches.

6. ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), National Defense Science and Technology Fund for Distinguished Young Scholars (2017-JCJQ-ZQ-022), the National Nature Science Foundation of China (61525206, 61771468,61622211,61620106009),the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209),and the Fundamental Research Funds for the Central Universities (WK2100100030) .

7. REFERENCES

[1] Takaaki Sugino, Holger R Roth, Masahiro Oda, and Kensaku Mori, "Fully convolutional network-based eyeball segmentation from sparse annotation for eye surgery simulation model," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pp. 118–126. Springer, 2018.

[2] Avijit Dasgupta and Sonam Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 248–251.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[5] Wanli Chen, Yue Zhang, Junjun He, Yu Qiao, Yifan Chen, Hongjian Shi, and Xiaoying Tang, "W-net: Bridged u-net for 2d medical image segmentation," *arXiv preprint arXiv:1807.04459*, 2018.

[6] SM Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh, "Spnet: Shape prediction using a fully convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 430–439.

[7] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang, "Automated pulmonary nodule detection in ct images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[8] Hongtao Xie, Zhendong Mao, Yongdong Zhang, Han Deng, Chenggang Yan, and Zhineng Chen, "Double-bit quantization and index hashing for nearest neighbor search," *IEEE Transactions on Multimedia*, 2018.

[9] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang, "Convolutional attention networks for scene text recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*.

[10] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun, Jianlong Tan, and Yongdong Zhang, "Attention and language ensemble for scene text recognition with convolutional sequence modeling," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 248–256.

[11] SM Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, 2017, pp. 12–24.

[12] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Dazhi Cheng, and Jian Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," *arXiv preprint arXiv:1804.03821*, 2018.