Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation †

Balamurali Murugesan⋆, Kaushik Sarveswaran⋆, Sharath M Shankaranarayana,
Keerthi Ram, Jayaraj Joseph and Mohanasankar Sivaprakasam

Abstract— **Image segmentation is a primary task in many medical applications. Recently, many deep networks derived from U-Net has been extensively used in various medical image segmentation tasks. However, in most of the cases, networks similar to U-net produce coarse and non-smooth segmentations with lots of discontinuities. To improve and refine the performance of U-Net like networks, we propose the use of parallel decoders which along with performing the mask predictions also perform contour prediction and distance map estimation. The contour and distance map aid in ensuring smoothness in the segmentation predictions. To facilitate joint training of three tasks, we propose a novel architecture called Psi-Net with a single encoder and three parallel decoders (thus having a shape of Ψ), one decoder to learn the segmentation mask prediction and other two decoders to learn the auxiliary tasks of contour detection and distance map estimation. The learning of these auxiliary tasks helps in capturing the shape and the boundary information. We also propose a new joint loss function for the proposed architecture. The loss function consists of a weighted combination of Negative Log Likelihood and Mean Square Error loss. We have used two publicly available datasets: 1) Origa dataset for the task of optic cup and disc segmentation and 2) Endovis segment dataset for the task of polyp segmentation to evaluate our model. We have conducted extensive experiments using our network to show our model gives better results in terms of segmentation, boundary and shape metrics.**

I. INTRODUCTION

Image segmentation is the process of delineating structures of importance from an image. Identifying these structures in the medical image finds application in many medical procedures. To state some of them: 1) segmentation of optic cup and disc in the retinal fundus image is useful in glaucoma screening, 2) segmentation of polyp in colonoscopy image is helpful in cancer diagnosis, 3) segmentation of the organ, bones benefit surgery planning and 4) segmentation of lung nodules in chest Computed Tomography aids physicians to differentiate malignant lesions from benign lesions. In recent years, deep learning networks [1] are widely used in medical image segmentation, and the most commonly used deep learning network is UNet [2].

TABLE I

SUMMARY OF PROS AND CONS OF THE MODELS.

	[2]	[3]	[4]	Ours
Shape information	x	✓	✓	✓
Class imbalance	x	x	✓	✓
Smooth boundary	x	x	✓	✓
Multiple object instances	✓	✓	x	✓

UNet [2] is an encoder-decoder type of network which takes an image as input and outputs a pixel-wise classification probability score with cross-entropy as its loss function. This network has set new state of the art results for different medical image segmentation tasks. But there are some drawbacks with the architecture type, and loss functions used. For instance, the encoder block of the network undersamples the input through max-pooling layers which results in loss of spatial information. Similarly, having pixel-wise classification alone as a loss function produces uneven mask boundaries and outliers. In addition to this, the loss function doesn't take shape information into account which can help in performance improvement. Also, using cross-entropy as a loss function introduces class imbalance problem for images in which background dominates the object of interest which is very common in the medical images. To overcome the above-mentioned issues, multiple works have been reported in the literature [1]. In that, the architecture and loss functions followed by [3] and [4] are of our interest. Both these works use a similar architecture with a single encoder and two parallel decoders. The decoders are used for mask and contour prediction in [3] whereas in [4] it is used for estimating mask and distance map. Contour and distance map estimation act as regularizers to mask prediction. Shape information is imposed through contour and distance map in [3] and [4]. Class imbalance problem is mitigated in [4] through its joint classification and regression approach while it will still be an issue in [3] because of both decoders acting as classifiers. The boundaries obtained using [4] are smooth and the segmentation has reduced outliers compared to [2], [3]. But in multi-instance object segmentation cases, an object of smaller size can be treated as an outlier resulting in unsatisfactory segmentation. The summary of the above discussion are shown in Table I.

The main contributions of our paper are as follows:

- We propose a novel multi-task network Psi-Net with a single encoder and three decoders (architecture with shape Ψ). The decoders are used to learn three different

⋆ Contributed equally

† https://github.com/Bala93/Multi-task-deep-network/

Balamurali Murugesan and Mohanasankar Sivaprakasam are with Indian Institute of Technology Madras (IITM), India and Healthcare Technology Innovation Centre (HTIC), IITM, India (email: balamurali@htic.iitm.ac.in)

Kaushik Sarveswaran is with Indian Institute of Information Technology Design & Manufacturing Kancheepuram (IIITDM), India and HTIC, IITM, India

Keerthi Ram and Jayaraj Joseph are with HTIC, IITM, India

Sharath M Shankaranarayana is with Zasti, India

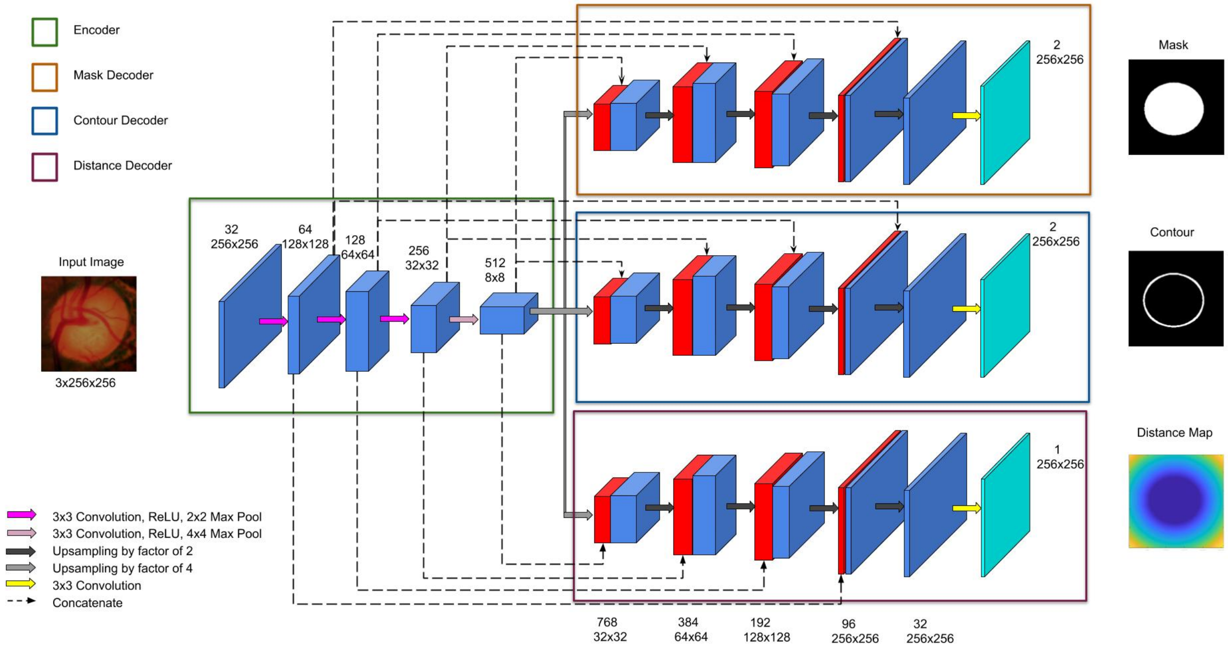

Fig. 1. Psi-Net: Proposed architecture for segmentation with a single encoder and three decoders.

tasks in parallel. The mask prediction is the primary task while the contour detection and distance map estimation are auxiliary tasks. These additional tasks are used to regularize the mask prediction path to produce a refined mask with smooth boundaries.

- We propose a novel joint loss function to handle the three different tasks together. The joint loss function consists of a combination of Negative Log Likelihood (NLL) for mask, Negative Log Likelihood (NLL) for contour, and Mean Square Error (MSE) for distance.
- We qualitatively compared our results with [2], [3] and [4]. The following evaluation metrics are used to perform a quantitative comparison:
  - *Segmentation metrics* : Jaccard and Dice coefficients.
  - *Shape similarity metrics* : Hausdorff distance
  - *Boundary metrics* : Segmentation evaluation around boundaries using trimap method.

  The comparative study showed that our network performed better than others in all the evaluation metrics producing a better mask with smooth boundaries.

II. METHODOLOGY

A. Architecture

The architecture Psi-Net is a UNet-like encoder-decoder network, with one contracting encoder path on the left and three expansive structurally similar decoder paths on the right. The shape of the architecture resembles the mathematical symbol Ψ. The encoder path consists of repeated downsampling operations which halves the size of feature map at each stage. Each downsampling operation is preceded by a convolution operation with kernel size 3x3 and stride 1, which is followed by a Rectified Linear Unit (ReLU) activation. Each decoder block is symmetric to the encoder, and at each decoder layer, the features from the corresponding encoder layer are concatenated which helps in retaining multi-scale features. The final convolutional layer in the encoder is upsampled by a factor of 4 and given as input to the decoder blocks.

Each decoder block is trained for a different tasks - Mask segmentation, contour extraction and distance map estimation. The former two are pixel-wise classification tasks while the latter is a regression task. The blocks are identical in structure until the last layer, where a 3x3 convolution is applied, and the number of output channels is 1 in the distance decoder block and is equal to the number of input classes in the other two blocks. The outline of proposed network is shown in Fig. 1.

B. Loss Function

The loss function consists of three components - Negative Log Likelihood (NLL) loss for mask and contour decoder blocks, and Mean Square Error (MSE) loss for the distance decoder block. Mask prediction is regularized by both contour and distance map predictions. The total loss is given by

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{contour} + \lambda_3 \mathcal{L}_{distance} \qquad (1)$$

where $\lambda_1, \lambda_2, \lambda_3$ are scaling factors.

The individual losses are formulated below.

1) Mask:

$$\mathcal{L}_{mask} = \sum_{\boldsymbol{x}\,\epsilon\,\Omega} log\, p_{mask}(\boldsymbol{x}; l_{mask}(\boldsymbol{x})) \tag{2}$$

$\mathcal{L}_{mask}$ denotes the pixel-wise classification error. $\boldsymbol{x}$ is the pixel position in image space Ω. $p_{mask}(\boldsymbol{x}; l_{mask})$ denotes the predicted probability for true label l_{mask} after softmax activation function.

2) Contour:

$$\mathcal{L}_{contour} = \sum_{\boldsymbol{x}\,\epsilon\,\Omega} log\, p_{contour}(\boldsymbol{x}; l_{contour}(\boldsymbol{x})) \tag{3}$$

$\mathcal{L}_{contour}$ denotes the pixel-wise classification error. $p_{contour}(\boldsymbol{x}; l_{contour})$ denotes the predicted probability for true label $l_{contour}$ after softmax activation function.

3) Distance:

$$\mathcal{L}_{distance} = \sum_{\boldsymbol{x}\,\epsilon\,\Omega} (\hat{D}(\boldsymbol{x}) - D(\boldsymbol{x}))^2 \tag{4}$$

$\mathcal{L}_{distance}$ denotes the pixel-wise mean square error. $\hat{D}(\boldsymbol{x})$ is the estimated distance map after sigmoid activation function while $D(\boldsymbol{x})$ is the ground-truth distance map.

III. EXPERIMENTS AND RESULTS

A. Dataset and Pre-processing

1) Dataset Description: We validated our proposed segmentation approach for the following two applications:

1) **Optic cup and disc segmentation** : We use ORIGA dataset [5] for the task of optic disc and cup segmentation. The dataset consists of 650 color fundus image with ground truth segmentations for optic disc and cup. The color fundus images are of dimension 256 × 256. Ellipse fit is applied to output segmentation mask.
2) **Polyp segmentation** : We also use Polyp segmentation dataset from MICCAI 2018 Gastrointestinal Image ANalysis (GIANA) [6]. The dataset consists of 912 images with ground truth masks. The dataset is split into 70% for training and 30% for testing. The images are center cropped and resized to 256 × 256.

2) Preprocessing: The dataset contains only segmentation mask. But for training our model, we need ground truth contour and distance map. The contour map is obtained by estimating the boundaries of connected components. These boundaries are subsequently dilated with a disk filter of radius 5. The distance map is obtained by applying an euclidean distance transform to the mask. The final distance map will contain zeros in the mask region, with the rest of the pixels denoting the shortest distance between that pixel and the mask boundary.

B. Implementation Details

All the models are implemented using PyTorch. Models are trained for 150 epochs using Adam optimizer, with a learning rate of 1e-4 and batch size 4. Experiments have been conducted with NVIDIA GeForce GTX 1060 GPU - 6GB RAM.

C. Evaluation metrics

In this section, A corresponds to the output of the method and B to the actual ground truth.

1) Segmentation evaluation: Jaccard index and Dice similarity score are the most commonly used evaluation metrics for segmentation. Jaccard index (also known as intersection over union, IoU) is defined as the size of the intersection divided by the size of the union of the sample sets, and it is calculated as follows:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{6}$$

2) Shape Similarity: The shape similarity is measured by using the Hausdorff distance between the shape of segmented object and that of the ground truth object, defined as

$$H(A, B) = max\left\{ \underset{x\epsilon A}{sup}\ \underset{y\epsilon B}{inf} ||x-y||,\ \underset{y\epsilon B}{sup}\ \underset{x\epsilon A}{inf} ||x-y|| \right\} \tag{7}$$

D. Results and Discussion

Some of the abbreviations which will be used in this section are Encoder (Enc), Decoder (Dec), Mask (M), Contour (C) and Distance (D). The results of the proposed network (1Enc 3Dec MCD) is compared with the following networks.

- A network (1Enc 1Dec M) [2] with a single encoder and a decoder having NLL as loss function for mask prediction.
- A network (1Enc 2Dec MC) [3] with a single encoder and two decoders having NLL as loss function for both mask and contour estimation.
- A network (1Enc 2Dec MD) [4] with a single encoder and two decoders having NLL as loss function for mask and MSE as loss function for distance map estimation.

1) Standard Evaluation: From Table II it can be seen that the network 1Enc 3Dec MCD has shown better performance in Dice and Jaccard compared to the networks 1Enc 1Dec M, 1Enc 2Dec MC and 1Enc 2Dec MD. This improvement in performance can be attributed to the use of two auxiliary regularizers, in the form of contour detection and distance map estimation, as opposed to a single regularizer in 1Enc 2Dec MC and 1Enc 2Dec MD. Both the networks 1Enc 2Dec MC and 1Enc 2Dec MD use shape information for mask refinement. While 1Enc 2Dec MD provides smooth boundaries compared to 1Enc 2Dec MC, it has a drawback in handling multiple object instances which is not an issue in 1Enc 2Dec MC. Since both these networks complement one another, combining these models brings the best result. The segmentation of polyp is relatively difficult when compared to optic cup and disc segmentation because of its large variations in size and shape. From Table II, it is evident that our network shows substantial improvement in performance for polyp segmentation compared to optic cup and disc segmentation.

TABLE II
COMPARISON OF SEGMENTATION AND SHAPE METRICS.

Architecture	Cup			Disc			Polyp		
	Dice	Jaccard	Hausdorff	Dice	Jaccard	Hausdorff	Dice	Jaccard	Hausdorff
1Enc 1Dec M [2]	0.8655	0.7712	14.832	0.9586	0.9215	8.802	0.8125	0.7323	24.133
1Enc 2Dec MC [3]	0.8715	0.7803	14.775	0.9646	0.9324	8.992	0.8151	0.7391	22.737
1Enc 2Dec MD [4]	0.8723	0.7807	14.814	**0.9665**	**0.9358**	9.538	0.8283	0.7482	22.686
1Enc 3Dec MCD (Ours)	**0.8745**	**0.7848**	**14.541**	**0.9665**	**0.9358**	**7.268**	**0.8462**	**0.7721**	**21.143**

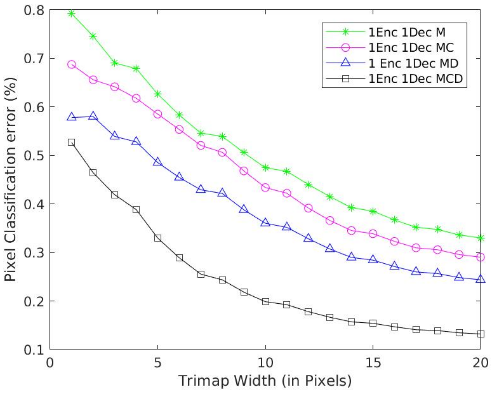

Fig. 2. Percent of misclassified pixels within trimaps of different widths.

2) Shape Similarity: Along with better segmentation, the network should also produce segmentation maps which are similar to ground truth masks regarding shape [3]. This shape similarity is obtained by Hausdorff distance. From Table II, it is clear that our network does well in capturing shape information compared to other networks. Also, sorting the Hausdorff distance helps in coming to the following inferences: 1) the addition of auxiliary tasks does help in preserving shape. 2) the auxiliary task of distance map estimation captures the shape better than the contour extraction.

3) Segmentation around boundaries: In the above paragraphs, we have mentioned that our network produces segmentation masks with smooth boundaries. Smooth boundaries indicate a better segmentation around the boundary. We evaluated the segmentation accuracy around boundary with the method adopted in [7]. Specifically, we count the relative number of misclassified pixels within a narrow band (trimap) surrounding actual object boundaries, obtained from the accurate ground truth images. As can be seen in Figure 2, our method has less error for trimaps of different widths.

4) Qualitative comparison: The qualitative comparison of our network 1Enc 3Dec MCD with 1Enc 1Dec M, 1Enc 2Dec MC and 1Enc 2Dec MD can be seen in Fig. 3. To better appreciate the improvement of our model we have shown only the polyp dataset outputs. The mask predicted by our network and 1Enc 2Dec MD is smooth without outliers compared to the mask predicted by the networks 1Enc 1Dec M and 1Enc 2Dec MC. This is depicted in the first two rows of the figure. In the third row of the figure, it can be seen that the network 1Enc 1Dec MD fails in case of multi-instance object segmentation while our network performs well as that of 1Enc 2Dec MC. The fourth row shows a case where our network outperforms the other networks.

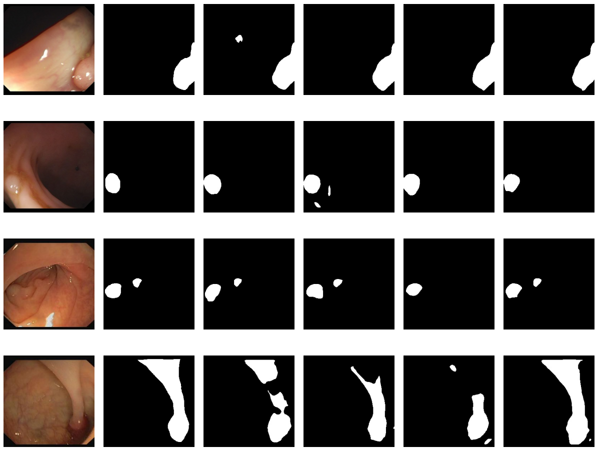

Fig. 3. From left to right : Image, Ground truth mask, Predicted mask of [2], [3], [4] and Ours.

IV. CONCLUSION

In this paper, we have introduced a network called Psi-Net with a single encoder and three parallel decoders. The three decoders are used for mask prediction, contour extraction and distance map estimation respectively. We have also introduced a joint loss function to optimize the proposed network. We have shown that this kind of architecture preserves shape well with better boundary outputs and improved segmentation performance.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[3] H. Chen, X. Qi, L. Yu, and P. Heng, "DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016, pp. 2487–2496.

[4] C. Tan, L. Zhao, Z. Yan, K. Li, D. Metaxas, and Y. Zhan, "Deep multi-task and task-specific feature learning network for robust shape preserved organ segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Apr 2018, pp. 1221–1224.

[5] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 3065–3068.

[6] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, “A benchmark for endoluminal scene segmentation of colonoscopy images,” *Journal of healthcare engineering*, vol. 2017, 2017.

[7] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011, pp. 109–117.