Learning and Incorporating Shape Models for Semantic Segmentation

H. Ravishankar, R. Venkataramani(✉), S. Thiruvenkadam, P. Sudhakar, and V. Vaidya

GE Global Research, Bangalore, India
rahul.Venkataramani@ge.com

Abstract. Semantic segmentation has been popularly addressed using Fully convolutional networks (FCN) (e.g. U-Net) with impressive results and has been the forerunner in recent segmentation challenges. However, FCN approaches do not necessarily incorporate local geometry such as smoothness and shape, whereas traditional image analysis techniques have benefitted greatly by them in solving segmentation and tracking problems. In this work, we address the problem of incorporating shape priors within the FCN segmentation framework. We demonstrate the utility of such a shape prior in robust handling of scenarios such as loss of contrast and artifacts. Our experiments show $\approx 5\%$ improvement over U-Net for the challenging problem of ultrasound kidney segmentation.

1 Introduction

Segmentation from medical volumes can get quite challenging depending on modality and anatomy. Traditional approaches such as active contours have handled the ill-posed nature of the segmentation problem using linear/non-linear models of shape (e.g. [4,6]). Recently, fully convolutional networks (FCN) have been successfully applied to 2D/3D medical image segmentation [13], optic flow [7], restoration [2], etc. While FCNs have success in bringing contexts into learning, there are a few drawbacks which recent works have tried to address. Firstly, local geometry such as smoothness and topology are not reliably and explicitly captured. Secondly, there is noticeable need for enough of representative training data to intrinsically model the foreground, background, shape, and the contextual interactions of above entities. With limited training data, failure modes of FCNs are hard to interpret or improve upon.

Motivated by traditional approaches, we propose to augment the FCN framework with prior shape information. The advantage of explicitly modeling shape within FCN is two fold: (1) we notice that generalization to appearance deviations from the training data is much better and (2) data augmentation strategies is essential for robust performance of FCNs. Especially for medical data, it is quite hard to come up with realistic appearance variations to enable FCN to

The first two authors contributed equally.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part I, LNCS 10433, pp. 203–211, 2017.
DOI: 10.1007/978-3-319-66182-7_24

handle scenario such as low contrast and artifacts. With the shape model decoupled, it is much easier to build data augmentation strategies for the class of shapes to capture invariances which can in turn boost prediction performance. We demonstrate the efficacy of our approach on the difficult problem of kidney anatomy segmentation from 2-D ultrasound B-mode images.

In summary, the key contributions of our paper are as follows:

(1) Learning a non-linear shape model and projection of arbitrary masks to the shape manifold space. We also discuss two novel data augmentation strategies to implement a shape convolution auto encoder.
(2) Incorporating the shape model explicitly in a FCN formulation through a novel loss function that penalizes deviation of the predicted segmentation mask from a learnt shape model.
(3) Demonstration of superiority of the proposed approach by as much as $\approx 5\%$ dice overlap with negligible increase in overall network complexity ($<\approx 1\%$).

2 Related Work

With limited training data, failure modes of FCNs are hard to interpret or improve upon. In a recent work [9], we have shown increased robustness to FCNs by explicit, joint modeling of appearance and shape through parallel networks tied together using weight sharing or novel loss functions.

Additionally, incorporating geometric characteristics (e.g., shape and smoothness of a particular object) of the images is critical when solving image-wide prediction problems such as segmentation, optic flow, etc. In [1], the authors address the problem of local geometry by imposing smoothness and topology priors for a multi-labelling problem of histology segmentation. For 3D shape segmentation, the authors in [11] combine outputs of multiple FCNs, which are label confidences, via a surface projection layer, which are processed through a surface-based conditional random field for consistent labelling. Another body of work concerns learning of shape priors using deep networks that are subsequently used in a classical fashion within a variational framework. In [3,5], shape priors are learnt using deep Boltzmann machines but used in a variational formulation for image segmentation and image completion tasks correspondingly. In [14], a segmentation network is proposed where a pre-trained *analysis* network is used to obtain image features which are then passed through a FCN to obtain global segmentation masks. These global masks are then refined by using the weights from the low-level layers of the *analysis* network.

In our work, we accomplish shape-prior influenced segmentation by employing two CNNs in a cascade. The key differences of our work are as follows: (1) incorporating shape regularization through an elegant formulation inside FCN and not as a post-processing step on label confidences or incomplete shapes like [5,11]. The motivation for the proposed method is that the output of FCN may not lie on the manifold of true shapes, and hence they need to be projected onto the correct manifold. This projection is realized by the auto-encoder (AE), and it implicitly provides a shape prior during training. During the test time,

the segmentation results are directly obtained from the output of the FCN. (2) a generic formulation that can be appended to other geometry or topology priors [1] (3) realization of shape regularization using a simple network which is trained using two interesting data augmentation strategies. In the next section, we provide a reasoning for such an approach.

3 Our Approach

FCNs are extensions of CNNs for pixel wise predictions (e.g., [12,13])that essentially have heirarchical deconvolution layers that work on CNN feature maps to give an "image" output. Each of these deconvolution layers have connections with the respective convolution layers in order to preserve fine detail while upsampling. FCNs have the utility of bringing spatial context into the predictions with a significant advantage of being really fast for pixel predictions being just feed forward operations. In standard FCN formulations such as U-Net [13], given training examples of pairs of images and segmentations masks I_k, S_k, $k = 1, 2, \ldots, N$, the framework learns a predictor $\hat{S}_w[\cdot]$ defined by parameters w that minimizes the training loss, e.g., $RMSE := \frac{1}{N}\sum_{k=1}^{N} |S_k - \hat{S}_w[I_k]|^2$.

In this work, we modify the above loss to incorporate a shape prior. While there are many choices for linear/non-linear representations for a segmentation shape prior [6], we use *convolutional autoencoders* (CAE) (e.g. used for de-noising [8], human motion modeling, [10]) for shape representation to enable easy integration with existing FCN implementations.

Denote as $\mathcal{M}$, the underlying space composed of valid shapes as defined by the ground truth training masks S_k, $k = 1, 2, \ldots, N$. Suppose that we are able to learn a p-dimensional shape projection (*encoder*) E and a (*decoder*) R. Note that for the purpose of being able to plug-in to a segmentation framework, the projection E should be able to take any arbitrary shape S and project it to a valid representation on M. Thus, the composition with the decoder R, i.e. $(R \circ E)[S]$ is the projection of S onto a valid shape on $\mathcal{M}$. One can see $R \circ E$ playing the role of a *convolutional de-noising autoencoder* (CDAE) [8] within a segmentation loss function. Denoting $\hat{S}_k = \hat{S}_w[I_k]$, we modify the loss as:

$$L[w] = \frac{1}{N}\sum_{k=1}^{N} |\hat{S}_k - (R \circ E)[\hat{S}_k]|^2 + \lambda_1 |E[S_k] - E[\hat{S}_k]|^2 + \lambda_2 |S_k - \hat{S}_k|^2. \quad (1)$$

The first term drives the predicted shape $\hat{S}_k$ to lie close to the shape space $\mathcal{M}$ by minimizing the projection error. The second term drives the distance between the encoded representations of the ground truth mask and the predicted mask. The last term tries to preserve variations in the ground truth shape from the learnt shape space $\mathcal{M}$. In vanilla implementations of FCN such as U-Net, since the loss function is based on Euclidean distance, the network parameters have to predict a complex transformation from the input image to a high dimensional shape. Thus there is a need for enough representative training data to intrinsically model appearance, shape, and the contextual interactions of above entities. In the proposed approach, a good part of the network complexity is borne by the autoencoder since the distance between the predicted shape $\hat{S}_k$ and the ground truth S_k is based on the encoded representations (Fig. 1).

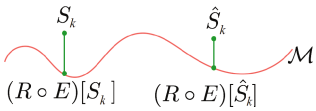


Fig. 1. Projection onto the shape space M

4 Architectures

In this section, we explain neural network models built to realize our formulation in (1). We build a cascade of two FCNs - one for segmentation and one for shape regularization as shown in Fig. 2. Segmentation network operates on the input image, while shape regularization network constraints the predicted shape to be in the manifold $\mathcal{M}$ defined by the training shapes.

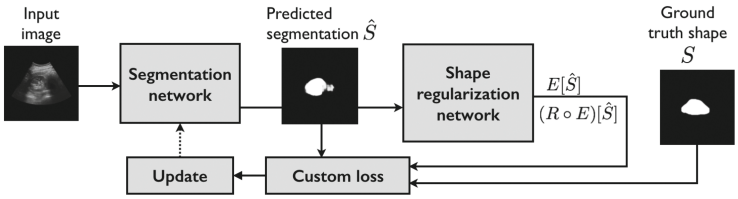


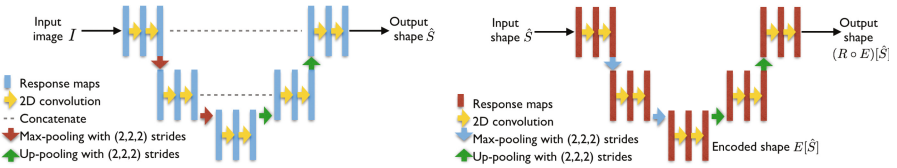
Fig. 2. Cascade network - SR-UNet

4.1 Segmentation Network

Our segmentation network is the vanilla U-Net architecture [13] (shown in Fig. 3a), which has become one of the most successful and popular CNN architecture for medical image segmentation. U-Net is nothing but a FCN with analysis-synthesis blocks, and skip-level connections between responses from layers of the analysis arm to the synthesis arms as shown in Fig. 3a.

4.2 Shape Regularization Network

The objective of this network is to operate on incomplete, under/over segmentation shape masks and force them to conform to the manifold of training shapes. We propose the use of a convolutional auto encoder to realize shape regularization as shown in Fig. 3b. The shape regularization network contains shape encoder and decoder blocks, which project the incomplete shapes into latent representations using compositions of convolutions and non-linear mappings. We hypothesize that the encoder would provide a concise, compact latent space representation that would not be affected by the errors in input shape from which the decoder block can accurately reconstruct the completed shape. There are no skip-level connections between the encoder and decoder blocks unlike the U-Net.



(a) Segmentation network - U-Net architecture

(b) Shape regularization network

Fig. 3. Network architectures

The information flow between different parts of the cascade network is shown in Fig. 2. Outputs of the shape completion network from the encoder and reconstruction layers - $E[\hat{S}_k]$ and $(R \circ E)[\hat{S}_k]$ affect the first two terms in (1), while segmentation network output contributes to the third term. The shape regularization network is pre-trained separately on noisy augmented shapes (Sect. 4.3), which is then plugged into the cascade architecture. It updates the segmentation network through (1), producing a shape regularized U-Net (SR-UNet).

4.3 Implementation Details

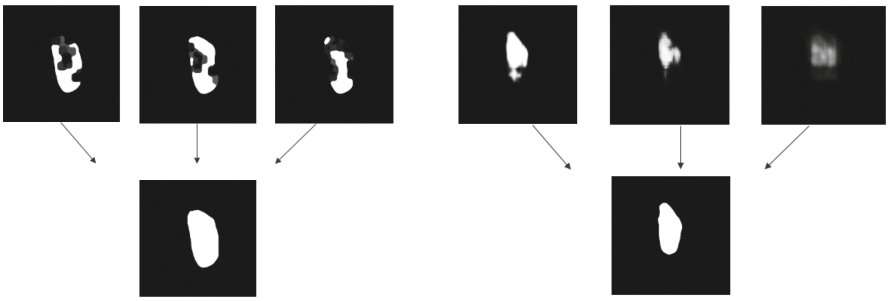
Our segmentation network consists of convolutional and up/downsampling layers, totally 10 in number, equally distributed between the two arms of the U-Net. The total number of trainable parameters is $\approx 14 \times 10^6$ and we use ReLUs and batch normalization as activation units and for regularization respectively. Intuitively, we expect the shape completion network to be simpler and hence, we build a convolutional auto encoder with $\approx 12 \times 10^3$ trainable parameters, contributing to a network complexity increase of less than 1% compared to the standard implementation. Typical λ values in (1) were around 0.5 and not much difference in performance was noted with variation around these values. We next describe the pre-training of shape completion network.

4.4 Data Augmentation for Shape Regularization Network

For the shape regularization network to achieve shape completion, it has to be trained with inaccurate shapes as input and ground truth masks as the output. We pursue two data segmentation strategies for creating these incomplete shapes:

(a) **Random corruption of shapes** We use a corruption kernel of high but random mean intensity and roll it across the shape on random seed locations and erode. We repeat this multiple times and create multiple instances of corrupted shapes as shown in Fig. 4a.

(b) **Intermediate U-Net predictions** We sample the U-Net predictions only on the training images at different epochs before convergence and treat the inaccurate predictions as the input to shape completion network Fig. 4b. The idea is to make the CAE learn to complete the failure modes of U-Net.



(a) Random corruption of shapes (b) Intermediate U-Net predictions

Fig. 4. Data augmentation strategies for shape CAE training

5 Kidney Segmentation from U/S B-Mode Images

Accelerated clinical work-flow using automated methods of detection and segmentation offer many advantages like operator independence, improved clinical outcomes, etc. Automated kidney segmentation from 2D ultrasound longitudinal scans is challenging due to many reasons - variability in kidney shape, size and orientation, acquisition scan plane differences, variability in the internal regions (renal sinus) and influence of adjacent structures like diaphragm, liver and fat layers. Presence of any pathology or abnormality can severely modify the observed texture which can further be compounded by ultrasound issues like shadow artifacts, speckle, sensitivity to spurious scatterers, etc. Also, automated algorithms are expected to work across different scan protocols with images from different probes, varying acquisition or reconstruction settings.

Data. The goal of this experiment is to demonstrate the robustness and generalization properties of the our approach over the state-of-the-art U-Nets. The data-set consists a total of 231 B-mode images obtained from two different scanning sites with varying acquisition settings. The images contain cases of varying challenges - pathology, shadow artifacts, incomplete kidney acquisition, other abnormalities and contains images from adult and pediatric subjects. We use 100 images for training and the remaining images for testing our method. The results show the competitive advantages of our algorithm on 131 images.

6 Results

We use Dice coefficient to compare our results with expert annotated ground truth. We refer to the results of our shape regularized FCN as SR-UNet_1 and SR-UNet_2, which corresponds to results of two different data augmentation strategies of random corruption and noisy U-Net predictions respectively (without extensive experiments with hyper parameters). In Table 1, we see that SR-UNet_1,2 improving Dice overlap by 4–5%, a significant improvement on a

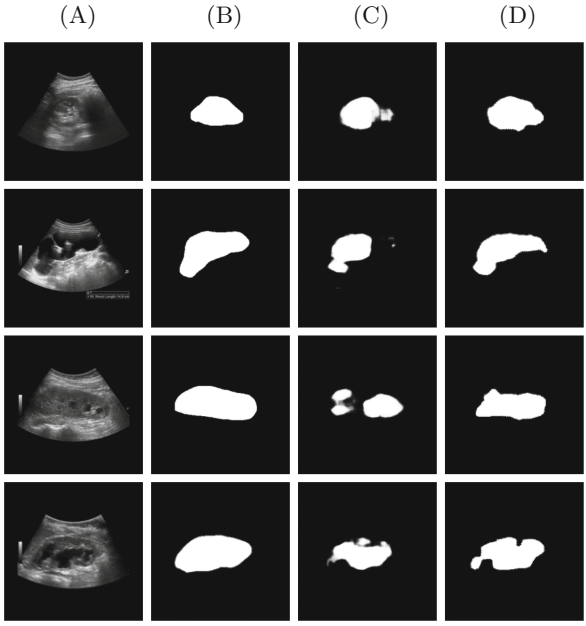


Fig. 5. (A) Input ultrasound images (B) Ground truth masks for segmentation (C) Segmentation masks predicted by U-Net (D) Segmentation masks predicted by U-Net with shape regularization - Proposed approach

Table 1. Average Dice overlap on 131 test images.

	Vanilla U-Net	SR-UNet_1	SR-UNet_2
Average dice	79.29%	**83.48**%	**83.95**%

challenging problem. Unsurprisingly, shape completion network built using noisy U-Net predictions is better as it explicitly works on failure modes but interestingly, synthetic data augmentation is equally powerful. More importantly, Fig. 5 illustrates how SR-UNet is able to complete complex structures even in the presence of significant pathology. For example in row 1, a shadow artifact has removed nearly all information from the right side of the kidney. Nevertheless, the cascaded network is able to arrive at a solution close to ground truth. Similarly in row 3, the presence of cysts disrupts conventional U-Net while SR-UNet is able to get a much more accurate result. Also in rows 2 and 4, a large portion of kidney is affected by abnormality which affects U-Net segmentation, while our method produces a near perfect segmentation in row 2 and an improvement in row 4. We would like to highlight that our novel shape regularization approach is generic and can be incorporated into any semantic segmentation neural network. We have chosen to compare our method with U-Net which is a popular representative technique for medical image segmentation.

While few would argue that U-Net has proved extraordinarily effective on a range of medical image analysis problems, our results indicate that at least in limited data scenarios U-Net can struggle with shape, particularly when textural and local information is unavailable due to pathology. A related undesirable characteristic is the tendency to produce disconnected small islands. While other techniques such as carefully engineered post-processing can also address these issues, we feel that our approach provides a natural and robust way to integrate desirable shape characteristics into the learning process of a deep neural network.

7 Discussion

Shape priors, when incorporated into the training loss of a neural network, can significantly improve prediction results, as demonstrated by our U/S kidney segmentation experiments. Though some cases can be really challenging, we feel that our contribution is an important step in the use of FCNs in clinical settings where meaningful and interpretable outputs are a necessity. Also, extension of shape priors to 3D segmentation is a straightforward task in our formulation. While we used a convolutional auto-encoder to obtain shape prior, alternatives such as Boltzmann machines, linear shape dictionaries, etc., can be explored. Also, shape is just one of the geometric attributes of anatomical objects and much more meaningful priors (e.g., texture, size, etc.) can be embedded into training objectives to achieve robustness and stability of neural networks.

References

1. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 460–468. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_53
2. Chaudhury, S., Roy, H.: Can fully convolutional networks perform well for general image restoration problems? CoRR abs/1611.04481 (2016)
3. Chen, F., Yu, H., Hu, R., Zeng, X.: Deep learning shape priors for object segmentation. In: Proceedings of CVPR, pp. 1870–1877, June 2013
4. Cremers, D., Rousson, M., Deriche, R.: A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. IJCV **72**(2), 195–215 (2007)
5. Eslami, S.M., Heess, N., Williams, C.K., Winn, J.: The shape boltzmann machine: a strong model of object shape. Int. J. Comput. Vis. **107**(2), 155–176 (2014)
6. Etyngier, P., Segonne, F., Keriven, R.: Shape priors using manifold learning techniques. In: Proceedings of ICCV, pp. 1–8 (2007)
7. Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: learning optical flow with convolutional networks. arXiv preprint (2015). arXiv:1504.06852
8. Gondara, L.: Medical image denoising using convolutional denoising autoencoders. arXiv preprint (2016). arXiv:1608.04667

9. Ravishankar, H., Thiruvenkadam, S., Venkataramani, R., Vaidya, V.: Joint deep learning of foreground, background and shape for robust contextual segmentation. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 622–632. Springer, Cham (2017). doi:10.1007/978-3-319-59050-9_49
10. Holden, D., Saito, J., Komura, T., Joyce, T.: Learning motion manifolds with convolutional autoencoders. In: SIGGRAPH Asia Tech. Briefs, p. 18. ACM (2015)
11. Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S.: 3D shape segmentation with projective convolutional networks. CoRR abs/1612.02808 (2016)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of CVPR, pp. 3431–3440 (2015)
13. 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). doi:10.1007/978-3-319-24574-4_28
14. Safar, S., Yang, M.H.: Learning shape priors for object segmentation via neural networks. In: Proceedings of ICIP, pp. 1835–1839 (2015)