Joint Deep Learning of Foreground, Background and Shape for Robust Contextual Segmentation

Hariharan Ravishankar(✉), S. Thiruvenkadam, R. Venkataramani, and V. Vaidya

GE Global Research, Bangalore, India
hariharan.ravishankar@ge.com

Abstract. Encouraged by the success of CNNs in classification problems, CNNs are being actively applied to image-wide prediction problems such as segmentation, optic flow, reconstruction, restoration etc. These approaches fall under the category of fully convolutional networks [FCN] and have been very successful in bringing contexts into learning for image analysis. In this work, we address the problem of segmentation from medical images. Segmentation or object delineation from medical images/volumes is a fundamental step for subsequent quantification tasks key to diagnosis. Semantic segmentation has been popularly addressed using FCN (e.g. U-NET) with impressive results and has been the fore runner in recent segmentation challenges. However, there are a few drawbacks of FCN approaches which recent works have tried to address. Firstly, local geometry such as smoothness and shape are not reliably captured. Secondly, spatial context captured by FCNs while giving the advantage of a richer representation carries the intrinsic drawback of overfitting, and is quite sensitive to appearance and shape changes. To handle above issues, in this work, we propose a hybrid of generative modeling of image formation to jointly learn the triad of foreground (F), background (B) and shape (S). Such generative modeling of F, B, S would carry the advantages of FCN in capturing contexts. Further we expect the approach to be useful under limited training data, results easy to interpret, and enable easy transfer of learning across segmentation problems. We present ∼8% improvement over state of art FCN approaches for US kidney segmentation and while achieving comparable results on CT lung nodule segmentation.

1 Introduction

Convolutional neural networks (CNNs) [7,10,14,18] have proven to be very successful in a wide range of visual tasks such as classification, recognition, characterization, tracking and segmentation. CNNs provide effective models for above vision learning tasks by incorporating spatial context and weight sharing between pixels across several hierarchical layers. Currently, CNNs are being actively applied to image-wide prediction problems such as segmentation [16],

The first two authors contributed equally.

M. Niethammer et al. (Eds.): IPMI 2017, LNCS 10265, pp. 622–632, 2017.
DOI: 10.1007/978-3-319-59050-9_49

optic flow [8], reconstruction [12], restoration [5] etc. These approaches fall under the category of fully convolutional networks [FCN] and have been very successful in bringing contexts into learning for image analysis. Models based on FCN have now been applied successfully to various 2D/3D medical image segmentation problems (e.g. U-NET, [17]). FCNs have a few drawbacks which recent works have tried to address. Firstly, local geometry such as smoothness and topology are not reliably captured. Secondly, there is noticeable need for enough of representative training data to learn the multiple entities: foreground, background, shape, and the contextual interactions of above entities. With limited or not enough training data, failures are hard to interpret and it is not easy to handpick training data that can improve performance. Finally, it is hard to transfer weights learnt from FCN to new problems since the above entities are abstractly tied to each other for the current problem. The problem of local geometry was addressed recently in [2] imposing smoothness and topology priors for a multi-labeling problem of histology segmentation. Next, the problem of overfitting is tackled in [6], using parameter reduction due to very Deep networks with skip level connections, motivated by ResNets [9].

In this work, we propose an alternative refinement to the FCN framework compared to the above enhancements [2,9]. Segmentation and motion tracking using foreground/background modeling has a rich history e.g. [4,13] using DCNN, see survey [3] for traditional approaches. For example, in [13], a multistage FCN is proposed to integrate appearance and motion cues for crowd segmentation. Both appearance filters and motion filters are pre-trained stage-by-stage and then jointly optimized to give improved accuracy. Inspired by the above class of methods, we propose a novel approach to enhance FCN segmentation using a generative modeling of the triad of F, B and S. There are three distinct advantages from the proposed FCN framework: Firstly, by modeling the appearance F, B, challenging scenarios such as non-linear shading effects, artifacts, and loss of contrast are factored out leaving the learning and prediction of S more robust. Secondly, domain specific tuning of networks (e.g. data augmentation, complexity of the network) corresponding to F, B and S makes it easier to control the number of parameters and hence over fitting. Finally, weights corresponding to either of F, B and S models can be easily transferred across applications. We look at a few innovative FCN network architectures and loss functions to achieve the above. Broadly speaking, we consider three parallel FCN networks, each modeling one of F, B and S. Analogous to multi-task learning (e.g. [11,15]), the models are jointly learnt using weight sharing and through a novel loss function that ties the outputs together. Figure 1 shows the input image and also the predicted foreground, background texture and segmented shape from a longitudinal ultrasound B-mode scan of adult kidney.

For our experiments, we consider the applications of kidney segmentation from 2-D ultrasound images and 3-D CT lung nodule segmentation. Both the applications are clinically relevant and have varying challenges as explained in the Results section. We also present quantitative comparisons of our results with U-NET [17] on the above data. We show that we outperform U-NET by almost 8% on the kidney segmentation problem while achieving marginally higher and comparable results on lung nodule segmentation.

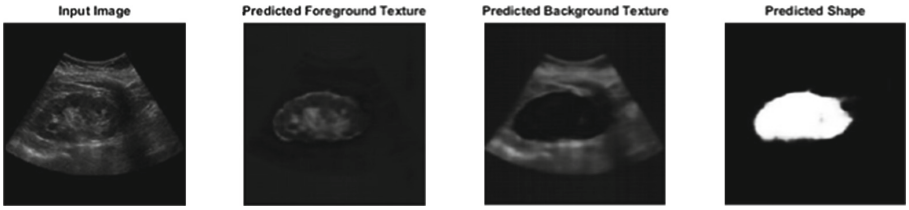

Fig. 1. (a) Input ultrasound B-mode kidney image (b) Synthesized foreground texture map (c) Synthesized background texture map (d) Predicted segmentation mask

2 Methods

CNNs provide effective models for several vision learning tasks by incorporating weight sharing between pixels across several hierarchical layers. The last layer is a fully connected layer whose outputs are used for regression/classification tasks. For image analysis tasks such as segmentation, one could use CNN in a sliding window fashion to predict the current pixel to be in the object or not. But such an approach has the disadvantage of being too slow and not being able to capture spatial context during pixel predictions.

Extending CNNs for pixel wise predictions are FCNs (e.g. [16,17] that essentially have hierarchical 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 to be able to preserve fine detail while upsampling. FCNs have the advantage of being really fast for pixel predictions being just feed forward operations along with the added utility of bringing spatial context into the predictions. In standard FCN formulations such as U-Net [17], given training examples of pairs of images and segmentations masks $I_k, S_k, k = 1, 2, \ldots, N$, the framework learns a predictor $\hat{S}_w[.]$ 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 our segmentation work, we extend FCNs to jointly model appearance (F and B) and shape (S). We learn the triad of predictors $\hat{F}_{w_1}[.], \hat{B}_{w_2}[.], \hat{S}_{w_3}[.], \hat{S}_{w_3} \in [0,1]$ that minimize the following possibilities for the training loss, FBS_1 and FBS_2. Analogous to multi-task learning (e.g. [11,15]), FBS_1 can be seen to tie F, B, S together with shared weights while FBS_2 ties the models together using a loss function that mimics image formation.

In FBS_1, we seek *shared* parameters w_1, w_2, w_3 and $\hat{S}_{w_3} \in [0,1]$ to minimize:

$$E_{FBS_1}[w_1, w_2, w_3] = \frac{1}{N}\sum_{k=1}^{N} |\hat{F}_{w_1}[I_k] - S_k.I_k|^2 + |\hat{B}_{w_2}[I_k] - (1 - S_k).I_k|^2 + |\hat{S}_{w_3}[I_k] - S_k|^2 + E_{smth}[\hat{S}_{w_3}[I_k]] \quad (1)$$

The first two terms learn the foreground and background predictors respectively. Note that without sharing of weights between w_1, w_2, w_3, the first 3 terms are independent of each other and the shape predictor is no longer benefited by

the appearance predictors and the appearance predictors in turn could just learn the identity mapping. Thus weight sharing is critical for the above formulation. E_{smth} is a smoothness prior (e.g. TV norm) on the shape predictor.

For ease of notation, we write e.g. $\hat{S}_{w_3}[I_k] = \hat{S}_k$. We look at a second formulation FBS_2 mimicing the image formation model. We seek parameters w_1, w_2, w_3 and $\hat{S}_{w_3} \in [0, 1]$ to minimize:

$$E_{FBS_2}[w_1, w_2, w_3] = \frac{1}{N}\sum_{k=1}^{N} |I - (\hat{S}_k\hat{F}_k + (1 - \hat{S}_k)\hat{B}_k)|^2 + |I - (S_k\hat{F}_k + (1 - S_k)\hat{B}_k)|^2 + |\hat{S}_k - S_k|^2 + E_{smth}[\hat{S}_k] \tag{2}$$

The first term is the image formation model that ties the predictors $\hat{F}, \hat{B}, \hat{S}$. By itself, this term would not make sense since we have 3 unknowns. The 3rd term seeks a shape predictor $\hat{S}$ given the ground truth masks $S_k, k = 1, 2, \ldots, N$. With just the first and third terms, in the absence of a good initial guess for the weights w_1 and w_2, or w_3, it would be difficult to converge to good predictors $\hat{F}, \hat{B}, \hat{S}$. Thus, we add the second term; since we know the ground-truth mask S_k, we can use this to derive the foreground/background predictors as shown.

In both FBS_1 and FBS_2, the predictor $\hat{S}$ is influenced by the predictions of $\hat{F}, \hat{B}$ because of shared weights (FBS_1) and the choice of loss function (FBS_2). Consequently, the proposed approach is more robust to choice of training data due to complementarity of the foreground/background/shape predictors. In competing FCN methods such as U-Net, enough of training data is needed to abstract the foreground/background texture, the shape, and relations of texture with the shape. As seen in the above illustrative example on simulated data we created to study FCNs (Fig. 2), U-net has not been able to complete the shape (Green: Ground truth, Red: Result) in regions of poor contrast or complex background. FBS_1 has been able to complete the shape since the foreground and background texture models have been jointly learnt with shape.

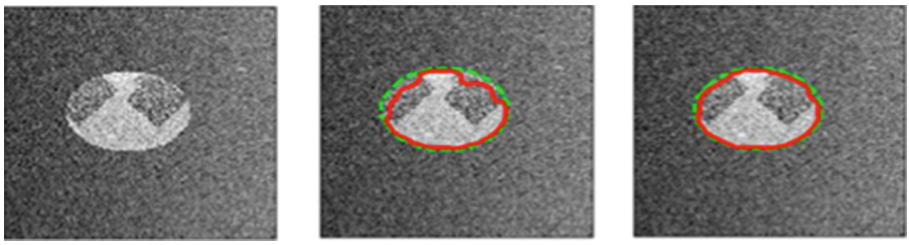

Fig. 2. Synthetic study example with contrast variation, weak edges and noise. (a) Input image (b) output of U-net (c) output of FBS_1. We see that results are better with joint appearance/shape modeling (FBS_1) (Color figure online)

3 Architectures

In this section, we explain how we realize the formulations described in previous sections using interesting FCN architectures. The vanilla U-NET architecture is shown in Fig. 3, which has become one of the most successful and popular approaches for medical image segmentation. U-NET is essentially a FCN with encoder-decoder blocks, with skip-level connections between responses from layers of the analysis arm to the synthesis arms as shown in Fig. 3.

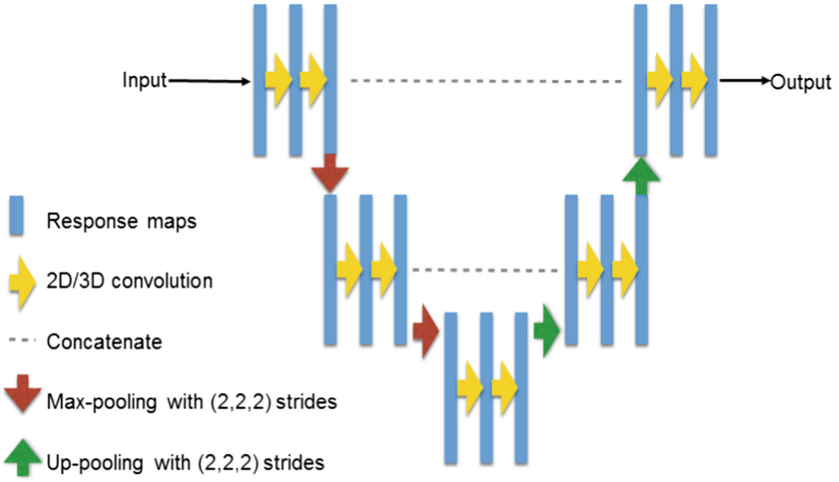

Fig. 3. U-NET architecture

3.1 Shared Weights Architecture

This architecture is an extension of U-NET with multiple outputs (Fig. 4). We proceed in the spirit of multi-task CNN learning [11,15], where FCNs are trained to simultaneously predict F, B and S based on our formulations from Eqs. (1) and (2). The intuition is that sharing weights for joint texture and shape prediction can lead to better generalization and robustness.

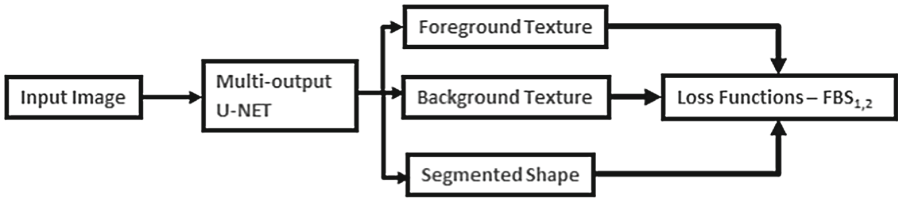

Fig. 4. Shared weights architecture

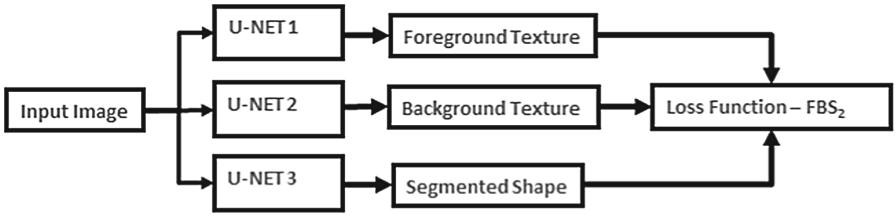

Fig. 5. Parallel architecture

3.2 Parallel Architecture

As discussed in Sect. 2, the weights corresponding to the three terms can be different and the parallel architecture in Fig. 5 attempts to model this scenario. The motivation for such an architecture can be attributed to the following reasons - (a) Even though U-NET attempts to get hierarchical, non-linear abstractions for texture and shape in an implicit fashion, the properties of these terms are very different and hence intuitively makes sense to model them using parallel networks. (b) This model also allows us to distribute the number of weights depending upon complexity of the term, for instance, background being a high entropy entity can be modeled using more weights than foreground. (c) Transferability across problems - depending upon tissue characteristics or background properties for similar, related problems, weights from the arms of relevance can be selectively transferred. We note that such an architecture may not be suitable for FBS_1 formulation as there is no single binding term that can jointly influence learning for F, B, S terms, with the independent parallel network implementation. However, this architecture is a natural fit for FBS_2 formulation, as it explicitly encodes the image synthesis model while allowing the flexibility to model the terms separately.

3.3 Implementation Details

We would like to point out that foreground and background texture modeling is a pixel-wise regression problem, while shape masks prediction is a binary classification problem. Hence, the output activation units of Figs. 4 and 5 have to be one of tanh, RelU or linear units for the two texture outputs and a sigmoid function for the shape term respectively. Note that terms on the FBS_1 formulation from Eq. (1) can have independent optimization metrics - the texture regression terms can be optimized for variants of L^2 norm and the binary shape term can be optimized using binary cross entropy as done in the baseline U-NET model. In our implementation, we used tanh units and mean-squared error for output activation and optimization metric for texture regression, respectively.

We experimented with different variants of dropout including (a) vanilla dropout - randomly drop inputs and (b) spatial dropout - the new variant of dropout tailored specifically for convolutional neural networks for zeroing entire feature maps. Our best results were obtained using spatial dropout, the results

of which we have reported. All the implementations in this paper also used Batch Normalization -which can be seen as a regularization technique applied on different layers to maintain their mean activation close to 0 and standard deviation close to 1. Finally, for a fair comparison, we have ensured that the total number of weights is roughly same across different implementations (Sect. 4 contains the details). In the remainder of the paper, we would refer to implementation and performances of $FBS_{1,2}$ using shared weight architecture as $FBS^a_{1,2}$ and implementation of FBS_2 using parallel weight architecture as FBS^b_2.

4 Experiments and Results

In this section, we establish the efficacy of our approach on two challenging medical imaging segmentation problems. One is anatomy segmentation − kidney segmentation from 2-D ultrasound B-mode images and another problem is 3-D Lung Nodule segmentation from CT images.

4.1 Lung Nodule Segmentation from 3-D CT

Lung cancer contributes to a large proportion of cancer related fatalities. Like other cancer types, early detection of nodules through screening procedures is critical for treatment planning and recovery. Recently, low dose CT (LDCT) scan has emerged as the standard procedure for lung cancer screening. In addition to the clinical relevance, technical challenges like wide contrast variation and lack of clear shape or appearance features make this a challenging problem, which were amongst the reasons to choose lung nodule segmentation from 3-D LDCT images as one application for the proposed method.

Data. Lung Image Database Consortium (LIDC-IDRI) [1] contains a collection 1010 3-D LDCT volumes of patients with lung cancer. We work with a pre-selected subset of 93 volumes containing 267 lesions on which manual segmentations have been performed, of which 179 lesions were used for training, and remaining 88 was used for validation.

Table 1. Performance comparison for lung nodule in 3-D CT images

Architecture	Dice overlap on validation set in %
U-Net	65.54
$FBS1^a$	**66.68**

Performance. The main goal of this experiment was to demonstrate the applicability of our approach to 3-D problems and also to different modalities. We use Sect. 4.2 to illustrate nuances of our approach, exhaustive comparisons and intuitions towards generalization, transferability and other properties. For this problem, we implemented a baseline 6-layer deep U-Net architecture with

3-D convolution units. We also implemented Architectures FBS_1^a as explained in Sect. 3 for comparison. We use Dice overlap with ground truth as the performance comparison metric. Table 1 contains the performance comparison. It should be noted that performance is similar, while shared architecture implementation of formulation FBS_1 has slightly better performance achieving 1% more than U-NET. All the results are averaged over 5 runs. The comparative performance should not be surprising given that validation set also comes from the same population and the problems of overfitting are less critical. Even in such scenarios, explicit modeling of foreground and background texture adds value as shown by the marginal increase over vanilla U-NET.

4.2 Kidney Segmentation from U/S B-Mode Images

Automated methods for determining the morphology and size of kidney from 2-D or 3-D ultrasound images have many benefits - accelerated work flow, operator independence on measurements and improved clinical outcomes. However, automated kidney segmentation is extremely challenging due to large variability in kidney shape, weak boundaries and large variation in appearance of internal regions based on acquisition scan plane. Additionally, shape, size and texture of the kidney region could vary drastically depending on the age of the subject - adult or pediatric and healthy or diseased. Another important challenge for the segmentation algorithm is to work across different scan protocols – every site can have different probes, acquisition settings including depth, TGC, etc.

Data. The goal of this experiment is to demonstrate the robustness and generalization properties of the proposed approach over the state-of-the-art U-NETs. We consider two datasets of B-mode kidney images acquired from two different scanning sites, which we would refer to as Population 1 and Population 2, with 108 and 123 images respectively. Population 2 is significantly more difficult than Population 1 due to the presence of challenging subjects (healthy and non-healthy), larger age differences and varied probe and acquisition settings. We train on a subset of Population 1 (60 images) and validate it on the remainder of Population 1 (48 images) and the entire Population 2 (123 images). We compare performances for both the formulations $FBS_{1,2}$ and for all the different architecture implementations explained in previous sections.

Performance. We use Dice coefficient as the metric to compare our results with expert annotated ground truth. Figure 6 shows an illustrative example of a difficult ultrasound image for kidney segmentation. Multiple lines of shadow, deep fat layer, weak bottom edge and inconsistent kidney contrast are some factors that make this case challenging. Figure 6(c) shows the result our approach which achieves a dice overlap of 91% while U-NET result in Fig. 6(b) fails completely with dice overlap of 61%. Table 2 contains the aggregate results. All the results reported are averaged over five independent runs for every experiment. It should be noted that architecture FBS_1^a - shared weight architecture of FBS_1 outperforms U-NET by 8% difference. We would also like to highlight

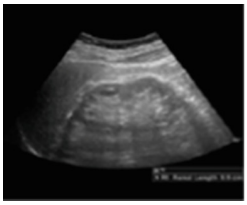
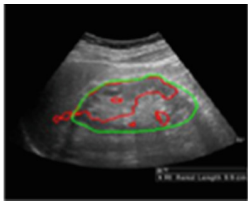
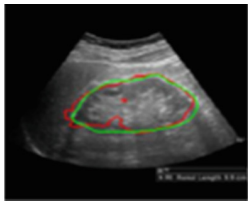

Fig. 6. Illustrative example on B-mode ultrasound kidney image. (a) Input image (b) U-NET segmentation result in red (ground truth - green) (c) FBS_1^a segmentation result in red (ground truth - green) (Color figure online)

Table 2. Performance comparison for kidney segmentation from U/S images

Architecture	Dice overlap with ground truth annotations in %		
	Validation set - population 1	Population 2	Mean over population 1 and 2
U-NET	75.90	62.18	66.03
FBS_1^a	**77.24**	**72.83**	**74.06**
FBS_2^a	75.02	66.28	68.74
FBS_2^b	68.98	59.89	62.44

that the difference in performance between U-NET and FBS_1^a on Population 1 is only 1.34%, however on more challenging, completely unseen Population 2, the improvement is 10%. This result clearly establishes the power of foreground, background modeling along with shape leading to better generalization and lesser over-fitting. Table 2 also shows that shared weight architecture for our second formulation - FBS_2^a also outperforms U-NET on Population 2 establishing the power of image synthesis formulations for segmentation. We mention that for FBS_2 in Eq. (2), we have not tuned the weights of relative contributions of different terms, which means that the cumulative cost could be dominated by the image synthesis terms than by shape error minimization term, explaining the lesser performance than FBS_1^a. Further, for the parallel architecture approach FBS_2^b, by enforcing the total number of parameters to be similar to U-Net and shared architectures (FBS_1^a, FBS_2^b), we have reduced the modeling capability of the S-arm by a factor of 3. This possibly explains the lesser performance of FBS_2^b. Improving the performance through better distribution of weights between the F, B, S arms along with better relative weighting of the terms, and demonstrating the value of transferability to related problems will be subject of our future work.

Finally, we show a palette of a few visual examples Fig. 7 for the kidney data. The second and third columns shown segmentation results from U-Net and FBS_1^a respectively for the input images (first column). The last two columns show the predicted foreground and background textures from FBS_1^a. The first two rows show examples where the proposed approach has done better than U-net while the third row shows over-segmentation in our approach.

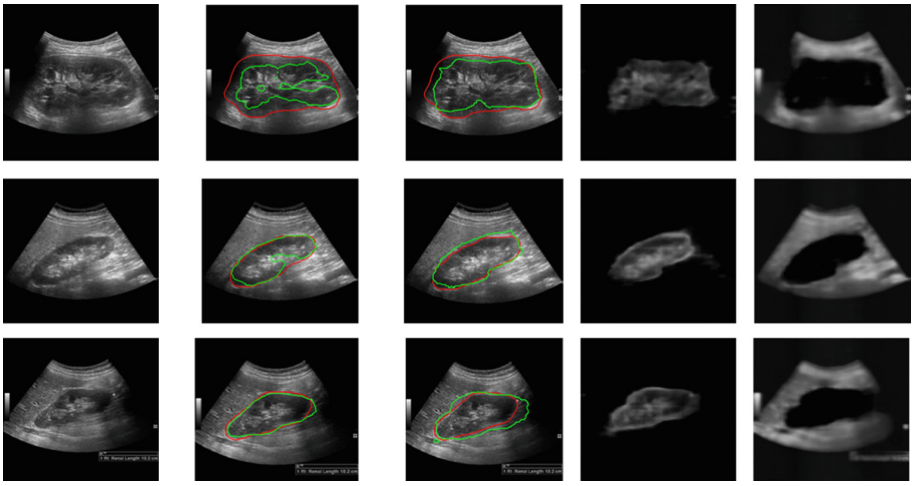

Fig. 7. A few examples on B-mode ultrasound kidney images. Columns: (a) Input image (b) U-NET segmentation result in green (ground truth - red) (c) FBS_1^a segmentation result in green (ground truth - red) (d) Synthesized foreground (e) Synthesized background (Color figure online)

5 Discussion

While U-NET has delivered impressive results on many challenging medical imaging segmentation tasks, it is still limited in its applicability in clinical applications because of lack of predictability in its output and correspondingly in its failure cases. We extend the FCN approaches by constructing a novel objective function which models texture and shape separately. The disentanglement of different properties allows us deeper insight into the model which in turn enables us to tune hyper-parameters and regularization approaches in a more meaningful manner. For instance, we could use shape regularizers only for the shape arm of the FBS_2^b architecture.

While a range of effective approaches for adding shape priors to traditional methods exist we find that integrating these methods into deep architectures poses new challenges. We are currently investigating adding shape priors to U-NET like architectures through dictionary learning approaches. Initial experiments have shown promise in enforcing smoothness of output and robustness of results. These characteristics are crucial in clinical applications where interpretability and failure modeling is crucial to technology acceptance.

References

1. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans

2. 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
3. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: an overview. Comput. Sci. Rev. **11**, 31–66 (2014)
4. Braham, M., Van Droogenbroeck, M.: Deep background subtraction with scene-specific convolutional neural networks. In: International Conference on Systems, Signals and Image Processing, 23–25 May 2016, Bratislava. IEEE (2016)
5. Chaudhury, S., Roy, H.: Can fully convolutional networks perform well for general image restoration problems? CoRR abs/1611.04481 (2016). http://arxiv.org/abs/1611.04481
6. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. CoRR abs/1608.04117 (2016). http://arxiv.org/abs/1608.04117
7. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
8. 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 arXiv:1504.06852 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
10. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. arXiv preprint arXiv:1502.06796 (2015)
11. Huang, Y., Wang, W., Wang, L., Tan, T.: Multi-task deep neural network for multi-label learning. In: 2013 IEEE International Conference on Image Processing, pp. 2897–2900. IEEE (2013)
12. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. CoRR abs/1611.03679 (2016). http://arxiv.org/abs/1611.03679
13. Kang, K., Wang, X.: Fully convolutional neural networks for crowd segmentation. arXiv preprint arXiv:1411.4464 (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in NIPS, pp. 1106–1114 (2012)
15. Li, X., Zhao, L., Wei, L., Yang, M., Wu, F., Zhuang, Y., Ling, H., Wang, J.: DeepSaliency: multi-task deep neural network model for salient object detection. CoRR abs/1510.05484 (2015). http://arxiv.org/abs/1510.05484
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. 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
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)