CGI2017 manuscript No.
(will be inserted by the editor)

Stacked Fully Convolutional Networks with Multi-Channel Learning: Application to Medical Image Segmentation

Lei Bi · Jinman Kim · Ashnil Kumar · Michael Fulham · Dagan Feng

Abstract The automated segmentation of regions of interest (ROIs) in medical imaging is the fundamental requirement for the derivation of high-level semantics for image analysis in clinical decision support (CDS) systems. Traditional segmentation approaches such as region-based depend heavily upon hand-crafted features and a priori knowledge of the user. As such, these methods are difficult to adopt within a clinical environment. Recently, methods based on fully convolutional networks (FCN) have achieved great success in the segmentation of general images. FCNs leverage a large labelled dataset to hierarchically learn the features that best correspond to the shallow appearance as well as the deep semantics of the images. However, when applied to medical images, FCNs usually produce coarse ROI detection and poor boundary definitions primarily due to the limited number of labelled training data and limited constraints of label agreement among neighboring similar pixels. In this paper, we propose a new stacked FCN architecture with multi-channel learning (SFCN-ML). We embed the FCN in a stacked architecture to learn the foreground ROI features and background non-ROI features separately, and then integrate these different channels to produce the final segmentation result. In contrast to traditional FCN methods, our SFCN-ML architecture enables the visual attributes and semantics derived from both the fore- and background channels to be iteratively learned and inferred. We conducted extensive experiments on three public datasets with a variety of visual challenges. Our results show that our SFCN-ML is more effective and robust than a routine FCN and its variants, and other state-of-the-art methods.

Keywords Fully convolutional networks (FCNs) · Segmentation · Regions of interest (ROI)

L. Bi, J. Kim(✉), A. Kumar, M. Fulham, D. Feng
School of Information Technologies, The University of Sydney, Australia (jinman.kim@sydney.edu.au)

M. Fulham
Department of Molecular Imaging, Royal Prince Alfred Hospital, Australia and Sydney Medical School, The University of Sydney, Australia

D. Feng
Med-X Research Institute, Shanghai Jiao Tong University, China

1 Introduction

Clinical decision support (CDS) systems are a major research area in medical imaging [1]. The basic concept is that the computer output is employed as a second opinion to assist physicians' image interpretation so as to improve diagnostic accuracy and reduce image reading time [2, 3]. Further, medical image segmentation is a fundamental requirement for a CDS system [4, 5]. The underlying objective is to partition the medical image into different anatomical structures, thereby separating the regions of interest (ROI), such as tumors from their background [6]. This fundamental need has motivated the development of numerous segmentation methods for medical images. However, traditional methods that use edges, regions and shape models, depend heavily on hand-crafted features and prior knowledge, which inhibit widespread application.

Deep learning methods based on fully convolutional networks (FCNs) have recently achieved great success in segmentation problems [7-9]. This success is primarily attributed to the ability of FCNs to leverage large datasets to derive a feature representation that combines low-level appearance information with high-level semantic information [7]. In addition, FCNs can be trained in an end-to-end manner for efficient inference, i.e., images are taken as inputs and the segmentation results are directly outputted. Many investigators have attempted to adapt FCNs to medical image segmentation [10-13]. However, there is a scarcity of annotated medical image training data due to the large cost and complicated acquisition procedures [10]. Consequently, without sufficient training data to cover all the variations in ROIs, e.g., lesions from different patients can have major differences in size/shape/texture, FCNs cannot provide accurate segmentation. In addition, FCNs have large receptive fields in the convolutional filters and hence produce coarse outputs of the ROI boundaries. They also lack smoothness constraints to encourage label agreement among similar neighboring pixels, and therefore it is difficult to produce a probability map with a consistent spatial appearance.

1.1 Related Work

Medical image segmentation methods can be grouped into four main categories: (1) semi-automated — where the segmentation is interactive; (2) un-supervised fully automated — where the segmentation is done without using training data (labels); (3) traditional supervised fully automated — where the segmentation proceeds using trained models (without deep learning techniques); and (4) deep learning based supervised fully automated — where segmentations uses a trained deep learning model.

Semi-automated methods require user initialization of the segmentation process, such as through seed selection [14-16] or contour placement [17, 18]. These seeds and contours can then been grown or morphed to the boundaries of the ROI according to predefined functions [18]. The manual initializations, however, are usually subjective, time-consuming and difficult to reproduce. As a consequence, such methods are unreliable for wide adoption in clinical environments.

Unsupervised fully automated medical image segmentation methods mainly focus on thresholding [19], energy functions [20, 21] and region merging [22]. Thresholding methods attempt to separate the ROI based on a threshold value, which is

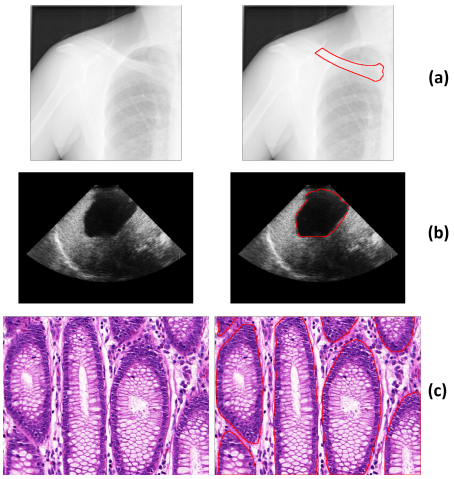

Fig. 1 Three types of medical images used in our evaluation arranged in rows. The left column shows the original image and the right column is the respective ground truth ROI annotations outlined in red. Top row (a) plain chest radiograph with annotated clavicle, (b) abdominal ultrasound (US) image with annotated liver lesion, and (c) histological image of colorectal cancer with annotated glands.

generally calculated by analyzing pre-defined image features e.g., an intensity histogram [19]. Methods based on energy function attempt to identify ROI boundaries by minimizing a well-defined cost (energy) function defined on image characteristics such as edges [23] and statistical distributions [24]. Region merging based methods recursively merge pixels or regions together in a hierarchical manner [25]. Unsupervised methods have a limited capacity to accurately segment challenging medical images, such as those with artifacts or where ROI boundary information is missing [4]. Thresholding based methods are further limited by the intensity distribution of the ROI and may fail if the distribution contains multiple peaks.

Traditional supervised fully automated medical image segmentation methods mainly focus on using shape models [26] and trained classifiers [27]. Methods based on shape models attempt to build a model (via atlas registration or deformation) together with prior knowledge of the ROI such as shape and appearance information [26, 28-31]. The built model is then applied to the image to segment the ROI based on local features e.g., intensity or texture features. Methods based on trained classifiers attempt to firstly extract pixel or regions features such as SIFT [32], HoG [33], texture features [34] and then use various classifiers, such as support vector machine [33], to separate the ROI from the background. Model based methods are usually limited to ROIs that have strong shape priors with relatively fixed locations, such as the liver and the lung fields. Therefore, these model based methods are not applicable to ROIs with various sizes and random locations e.g., tumors in different organs that spread beyond usual tissue boundaries, fractures and regions of infarction. Model based methods are further limited to segment multiple ROIs in an image simultaneously e.g., multiple lesions. These supervised methods also rely on using low-level features, which do not capture image-wide

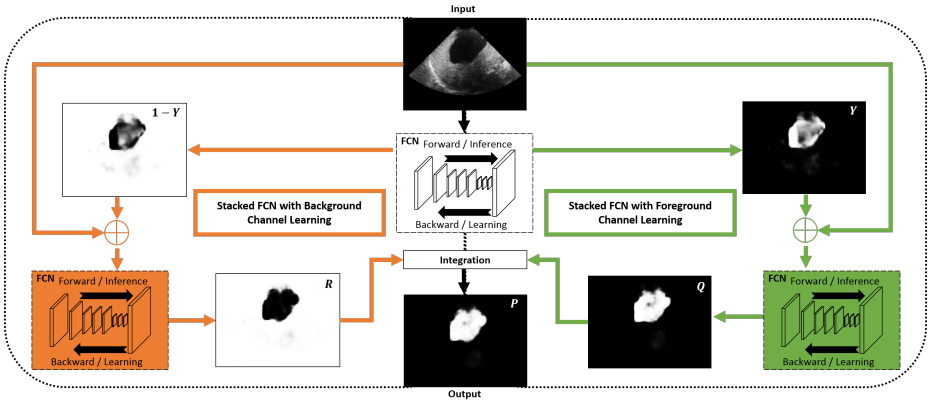

Fig. 2 Overview of our stacked fully convolutional networks with multi-channel learning method (SFCN-ML).

variations, and their performance relies on correctly tuning a large number of parameters and effective pre-processing techniques e.g., image de-noising.

The current deep learning based supervised fully automated segmentation methods mainly focus on using patches such as sliding windows [35-37] and super-pixels [38, 39], or adopt a FCN architecture. Patch-based methods try to predict if the patch or the center of the patch is inside the ROI; the patches are trained and inferred in a deep learning network. These patch-based methods are inefficient since accurate segmentation requires a prediction for every pixel in the image. In addition, the independent training and prediction of patches results in a loss of spatial context, meaning that the segmentation results usually lack consistency with coarse and inconsistent labeling of adjacent pixels. Recent work based on FCNs by Chen et al. [10] and Xu et al. [12], combined FCNs with contour (edge) features to constrain the boundary of the ROIs for accurate segmentation. Ronneberger et al. [40] proposed a U-net for image segmentation, where the FCN architecture was modified to combine different feature maps (intermediate results) produced at different layers to produce the final segmentation result, which thereby increase the smoothness. Other researchers have added graph models such as conditional random field (CRF) [11, 41], topology [13], graph cut [42] and level-sets [43]. However, these methods have limited capacity to refine the segmentation results. Compared with the supervised deep FCN architecture, the following refinements are usually based on unsupervised learning or priori knowledge, with low-level features. Such refinement procedures could dictate the final segmentation performance and ignore the FCN segmentation outcome. Furthermore, there is no feedback mechanism to guide the refinement, which is problematic with challenging ROIs. As an example, the aim of the CRF is to minimize the overall energy function and so it could refine the coarse outcome of the FCN. However, the CRF could also fail into a local minimum for inhomogeneous ROIs.

1.2 Our Contribution

We propose a new fully automated ROI segmentation method for medical images to overcome the challenges mentioned above. We have named it stacked fully convolutional networks with multi-channel learning (SFCN-ML). Our method improves on the state-of-the-art object detection method of FCN [7] that is specifically optimized for common visual attributes to medical image segmentation. In contrast to the coarse and noisy segmentation results of FCN, our method uses the previously estimated segmentation results to learn and refine the segmentation results across the stacked FCN architecture. Our method has the following contributions:

(1) The stacked FCN (SFCN) learns and predicts the segmentation iteratively and so minimizes the segmentation errors for challenging ROIs. During training, SFCN learns from the training data (training images and the manual annotations) and the estimated results derived from the previous SFCN iteration. The ability to learn from the previous iterations boosts the training data and also optimizes the learning of the ROI boundaries, which are usually difficult to segment. During prediction, the SFCN uses test (input) images and the estimated probabilities derived from previous iterations to gradually improve the segmentation accuracy.
(2) The stacked multi-channel learning refines the ROI segmentation in the context of the foreground and background. Existing methods mainly focus on segmenting the foreground area (directly segment the ROI) which is problematic when ROIs having inhomogeneous textures. Our method integrates the foreground and background segmentation results.
(3) Our method can be applied to a large variety of medical images and in this work we use examples from plain radiography, ultrasound and histology. These images were selected to provide a range of diversity including:
   (a) different types of ROIs - lesions and anatomical structures;
   (b) various ROI localizations - relatively fixed and random positions;
   (c) grayscale and color images;
   (d) ROIs with varying contrasts/textures and,
   (e) regular and irregular shapes/boundaries.

The rest of this paper is organized as follows: Section 2 describes our methods; Section 3 presents the experimental results on the three imaging datasets and includes the comparison of our method to the existing state-of-the-art methods and the conventional FCN architecture; the discussion is found in Section 4 and the conclusions are found in Section 5.

2 Methods and Materials

2.1 Overview of the Framework

The outline of our SFCN-ML method is shown in Fig. 2. The FCN component was applied to the input medical image to obtain a foreground and a background probability map. These probability maps together with the input medical image are then fed into two separated FCN components for foreground and background channel refinement. Finally, the refined foreground and background probability map are integrated to produce the final segmentation results.

2.2 Fully Convolutional Networks (FCN)

The traditional FCN architecture contains downsampling and upsampling parts [7]. The downsampling part has convolutional and max-pooling layers to extract high-level abstract information and has been widely used in convolutional neural networks (CNN) for image classification related tasks [44]. The upsampling part has convolutional and deconvolutional layers that upsample the feature maps to output the score masks [10].

Convolutional layers are defined on a translation invariance basis and have shared weights across different spatial locations. The input and the output of convolutional layers are feature maps and are calculated by convolving convolutional kernels:

$$f_s\left(X; W, b\right) = W *_s X + b \tag{1}$$

where X is the input feature map, W denotes the kernel, b is the bias, $*_s$ represents convolution operation with stride s. As a result, the resolution of the output feature map $f_s\left(X; W, b\right)$ is downsampled by a factor of s. Convolutional layers are usually interleaved with max-pooling layers. Max-pooling layers are a form of non-linear downsampling, which is usually used to further improve translation invariance and representation capability [45]. Max-pooling layers also have the ability to partition the input into non-overlapping sub-regions, which minimizes the computation cost of the upper layers and also reduces over-fitting [7]. The FCN network can be defined as:

$$Y = U_S\left(F_S\left(I; \theta\right); \varphi\right) \tag{2}$$

where Y is the output prediction (probability map of ROI), I is the input image, F_S denotes the feature map produced by the stacked convolutional layers with a list of stride S, U_S denotes the deconvolution layers that upsamples the feature map by a list of factors S to ensure both the output Y and input I have the same size (height and width). θ and φ are the learned parameters for convolutional and deconvolution layers. For training of FCN, the whole architecture can be defined as minimizing the overall loss between the predicted results and the ground truth annotation of the training data:

$$\arg\min_{\theta,\varphi} \sum \mathcal{L}(Y, Z|\theta, \varphi) \tag{3}$$

where $\mathcal{L}$ calculates the loss (per-pixel multinomial logistic loss) of the ground truth annotation Z and the predicted results. The FCN network parameters θ and φ can then be iteratively updated using stochastic gradient descent (SGD) [46] algorithm. For segmentation, FCN takes an image of arbitrary size and outputs a probability map of the same size that indicates the ROI area.

2.3 Stacked Fully Convolutional Networks with Multi-Channel Learning (SFCN-ML)

Our SFCN-ML embeds the foreground and background probability map produced at the previous FCN component for training and testing. The foreground and background channel learning can then be defined as:

$$\begin{cases} Q = U_S\left(F_S\left(I, Y; \theta_Q\right); \varphi_Q\right) \\ R = U_S\left(F_S\left(I, [1-Y]; \theta_R\right); \varphi_R\right) \end{cases} \tag{4}$$

Where Y is the output of equation (2), Q and R represent the refined foreground and background prediction (probability map), respectively. We trained Q and R separately based on equation (3) to get the convolutional parameters θ_Q, θ_R and deconvolution parameters φ_Q, φ_R. We also changed the loss function to only consider foreground or background to facilitate the fore- or background learning.

In general, foreground prediction Q and background prediction R are complementary to each other, in which foreground prediction could produce over-segmentation results with isolated regions, while background prediction could under-segment the ROI. Therefore, we produce the final segmentation map by integrating these two maps, and this can be defined as:

$$P = \gamma \cdot Q + (1-\gamma) \cdot [1-R] \tag{5}$$

Where γ is a constant weight that balances the importance of the foreground and the background prediction and we empirically set $\gamma = 0.6$ to favor the foreground prediction.

The final integrated probabilistic map was converted into a binary segmentation result via $\geqslant 0.5$ probability thresholding.

2.4 Materials

We selected three different datasets with a variety of visual characteristics to evaluate the effectiveness and robustness of our method.

(1) We used the chest radiograph anatomical structure segmentation (CRASS) set (Fig. 1(a)) [47] a public set of 299 posterior-anterior chest radiographs. The set was selected from a database containing images with a high rate of tuberculosis. All subjects were 15 years or older. The manually annotated clavicles were used as the ground truth. In this paper, we randomly selected 200 images as training set (∼67%) and the remaining 99 images were selected as the test set (∼33%).
(2) The second dataset was the SYSU-US dataset (Fig. 1(b)) [15] a public dataset of 23 ultrasound studies of the abdomen with lesions in the liver. Each patient had approximately 20 images (slices). Expert manually annotated lesions were used as the ground truth. We used 240 randomly selected images (50%) as the test set and the remaining images as the training set (240 images, 50%). Then we reversed the role of the two sets and averaged the results.
(3) The third dataset was the GlaS dataset (Fig. 1(c)) [48] – a public dataset consisting of 165 histological images of colorectal cancer. Expert manually annotated gland structures were used as the ground truth. The dataset was labelled as the training and the test set by the dataset provider which including 85 training images and 80 test images.

Table 1 Image datasets used in this paper.

Type		**CRASS**	**SYSU-US**	**GlaS**
Train	Image Number	200	240	85
	Percentage	67%	50%	52%
Test	Image Number	99	240	80
	Percentage	33%	50%	48%
Total	Image Number	299	480	165

A summary of each dataset is listed in Table 1.

2.5 Training SFCN-ML

There is a scarcity of medical image training data as noted previously when compared to general images [49, 50]. Research has suggested that the lack of training data can be alleviated by fine-tuning, where the lower layers of the fine-tuned network are more general filters (trained on general images) while those in the higher layers are more specific to the target problem [49, 51]. Therefore, we used the off-the-shelf MatConvNet [52] version of FCN trained on the PASCAL VOC 2011 dataset. To achieve more precise details of pixel prediction, we fine-tuned a stride-8 FCN architecture (FCN-8s) on the each of the training datasets. Data augmentation techniques including random crops and flips were used to improve robustness [51, 53]. For each FCN component, we fine-tuned the pre-trained FCN model separately (using the same training images and the same annotations, except for the background channel learning. We used the inverse of the annotation for the background channel learning). Each FCN component took about 6-8 hours to fine-tune over 200 epochs with a batch size of 50 on a 12GB Titan X GPU, with converged at about the 150th epoch. For the CRASS dataset, we first cropped out the top half of the chest radiograph. We then separated the cropped chest radiograph into left and right images for training and segmentation (see Fig. 1). For the GlaS dataset, there was an additional fourth input channel (the probability map), which meant we could not directly fine-tune the FCN-8s, which expected 3-channel inputs (usually RGB images). So we fine-tuned 3 FCN models separately and we replaced one of the RGB channels with a probability map produced by the initial FCN results. We rotated replacement of the three color channels of the 3 FCN models to cover all different replacement variations.

3 Experiments and Results

3.1 Evaluation Metrics

We used the most common segmentation evaluation metrics including: the dice similarity coefficient (Dice), Jaccard index (Jac.), sensitivity (Sen.), specificity (Spe.) and accuracy (Acc.), defined as:

$$Dice = \frac{2\,|GT \cap PR|}{|GT| + |PR|}, Jac. = \frac{|GT \cap PR|}{|GT \cup PR|} \tag{6}$$

Table 2 Segmentation results for the CRASS dataset.

CRASS	Dice	Jac.	Sen.	Spe.	Acc.
FCN	74.43	61.25	70.35	**99.53**	98.61
SFCN-FL	81.10	70.01	80.06	99.52	98.92
SFCN-BL	80.54	69.20	80.44	99.46	98.86
SFCN-ML	**81.40**	**70.47**	**80.68**	99.52	**98.93**
diff	*6.96*	*9.22*	*10.33*	*-0.01*	*0.32*

Table 3 Segmentation results for the SYSU-US dataset.

SYSU-US	Dice	Jac.	Sen.	Spe.	Acc.
FCN	71.62	58.65	68.04	**99.56**	98.67
SFCN-FL	79.94	68.66	82.45	99.44	98.97
SFCN-BL	79.95	68.35	**84.28**	99.35	98.95
SFCN-ML	**80.71**	**69.50**	84.01	99.42	**99.00**
diff	*9.09*	*10.86*	*15.97*	*-0.14*	*0.33*

$$Sen. = \frac{|TP|}{|TP| + |FN|}, Spe. = \frac{|TN|}{|TN| + |FP|} \tag{7}$$

$$Acc. = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{8}$$

Where GT denotes the ground truth, PR is the algorithm predicted segmentation result, TPs are the true positive pixels (ROIs), TNs are the true negative pixels (background), FPs are the false positive pixels and FNs are the false negative pixels. We also calculated the pixel-level receiver operating characteristic (ROC) curve and the precision-recall (PR) curve for additional comparisons. ROC and PR curves are widely used for segmentation related problems on general images [54].

3.2 Results from CRASS, SYSU-US and GlaS Datasets

We first compared our method to the state-of-the-art FCN method (also trained with a stride-8 FCN architecture) on all different datasets. We also measured the performance when only using foreground or background channel learning, denoted by SFCN-FL and SFCN-BL. Tables 2, 3, 4 and Fig. 3 show the segmentation results on the 3 datasets. Our SFCN-ML method had the best overall performance across all the different measurements and improved the existing FCN method with a margin of $\sim$2 — 11% in the Jaccard measure. Also the multi-channel learning method consistently performed better than the single channel learning approaches. For Tables 2-4, the values in bold are the best results; $diff$= difference between our SFCN-ML and FCN.

Table 4 Segmentation results for the GlaS dataset.

GlaS	Dice	Jac.	Sen.	Spe.	Acc.
FCN	87.21	78.24	84.49	90.84	88.31
SFCN-FL	88.66	80.50	86.29	**91.65**	89.47
SFCN-BL	88.58	80.37	**86.73**	90.74	89.38
SFCN-ML	**88.70**	**80.56**	86.37	91.63	**89.50**
diff	*1.49*	*2.32*	*1.88*	*0.79*	*1.19*

Table 5 Comparison with the state-of-the-art segmentation methods on the SYSU-US dataset. Super. = supervised methods; Semi. = semi-automated methods; Deep = deep learning methods.

SYSU-US	Jac.	Super.	Semi.	Deep
CVAC	33.68	✘	✘	✘
LS	33.88	✘	✘	✘
STF	57.04	✔	✘	✘
SnapCut	62.26	✘	✔	✘
CM	65.71	✘	✔	✘
FCN	58.65	✔	✘	✔
SFCN-FL	68.66	✔	✘	✔
SFCN-BL	68.35	✔	✘	✔
SFCN-ML	**69.50**	✔	✘	✔

3.3 Comparison with the state-of-the-art methods

We compared our methods with the state-of-the-art methods on the SYSU-US and GlaS datasets. The SYSU-US dataset was recently used in the evaluation of a number of state-of-the-art methods [15]. The comparison methods were as follows: (1) CVAC [55] — Chan-Vese active contour based segmentation; (2) LS [56] — level-set based segmentation; (3) STF [57] — semantic texton forest; (4) SnapCut [58] — segmentation using localized classifiers; (5) CM [15] — inference with collaborative model; and (6-8) FCN, SFCN-FL and SFCN-BL. The results for methods (1-5) were reported in [15] and the attributes of each method are listed in Table 5. The results in Table 5 shows that our SFCN-ML provided a major improvement on existing methods for ultrasound ROIs.

For the GlaS dataset, we compared our SFCN-ML method with the recent published methods (reported in [13]), which are also based on FCNs. These methods were: (i) FCN-32s [7] — a FCN with a 32 stride; (ii) DeepLab [8] — a FCN with a conditional random field (CRF) as the post refinement; (iii) CRF-RNN [9] — CRF as recurrent neural networks (RNN), which embed CRFs inside the FCN architecture; (iv) FCN+SM [13] — FCN with a smoothness term for refinement; (v) FCN+SM+TP [13] — a FCN with a smoothness and topology term for refinement. To make our results comparable, we followed the evaluation methods

Table 6 Segmentation results of our method compared with the state-of-the-art methods on the GlaS dataset.

GlaS	DiceObj	Acc.
FCN-32s	70.00	80.00
DeepLab	69.00	78.00
CRF-RNN	42.00	73.00
FCN+SM	78.00	**90.00**
FCN+SM+TP	**80.00**	86.00
FCN	74.99	88.31
SFCN-FL	78.14	89.47
SFCN-BL	76.68	89.38
SFCN-ML	78.26	89.50

reported by Sirinukunwattana et al [48] with a DiceObj value which measures the gland level Dice for evaluation and is defined as:

$$DiceObj = \frac{1}{2}\left[\begin{array}{l}\sum_{i=1}^{N_{GT}} \mu_i Dice\left(GT_i, PR_i\right) + \\ \sum_{j=1}^{N_{PR}} \delta_i Dice\left(GT_j, PR_j\right)\end{array}\right] \tag{9}$$

Where N_{GT} and N_{PR} denotes the number of glands in the ground truth and the algorithm predicted results. The first term reflects how well each ground truth gland overlaps with the algorithm segmented gland, and the second term reflects how well the algorithm segmented gland overlaps its ground truth gland. μ_i and δ_i are defined to put less weights on small regions (on both algorithm segmented gland and ground truth gland) and can be calculated as:

$$\mu_i = \frac{|GT_i|}{\sum_{k=1}^{N_{GT}} |GT_k|}, \delta_i = \frac{|PR_i|}{\sum_{t=1}^{N_{PR}} |PR_t|} \tag{10}$$

Table 6 shows that our method without adding any post-refinement techniques has better performance when compared to DeepLab and CRF-RNN and it performs competitively when compared with the recent published methods FCN+SM and FCN+SM+TP.

4 Discussion

Our findings show that our method has higher segmentation accuracy than state-of-the-art FCN methods for ROI segmentation. In general, FCNs can be employed to segment ROIs on medical images. However, due to the lack of label agreement and consistency constraints, FCNs usually generate poor boundary definitions with the production of many isolated regions (see Fig 4ii). In comparison, our methods provide improved segmentation results with an average increase of ∼2—11% in the Jaccard measure. We attribute these benefits to the stacked FCN architecture that constrains and refines the ROI definitions. Both the foreground (SFCN-FL) and the background channel learning (SFCN-BL) methods are able to

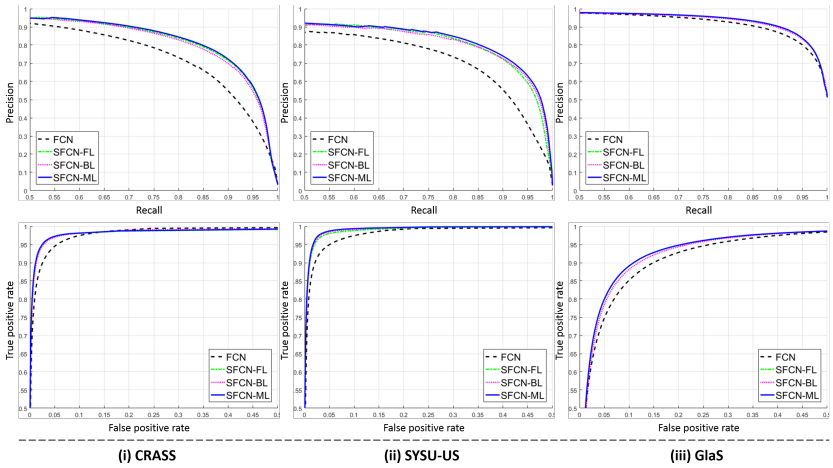

Fig. 3 Precision-recall (PR, top) and receiver operating characteristic (ROC, bottom) curves of the different methods on the CRASS (i), SYSU-US (ii) and GlaS (iii) datasets.

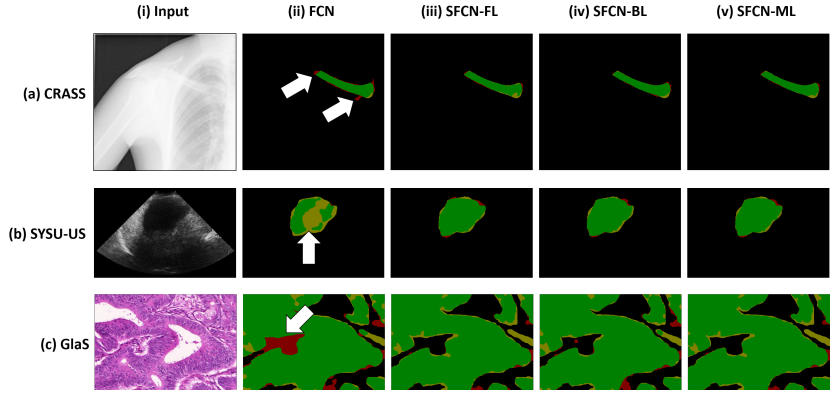

Fig. 4 Segmentation results from 3 examples. (i) input images from CRASS (a), SYSU-US (b) and GlaS (c) dataset, (ii-v) segmentation results of FCN and our SFCN-FL, SFCN-BL and SFCN-ML methods. The colors represent true positive (green), true negative (black), false positive (red) and false negative (yellow) pixels.

refine the boundary definitions and achieved similar performance. The combined model (SFCN-ML), takes advantages of each result and provides overall better segmentation results.

Our results from Table 5 indicate that, when compared with traditional methods such as CVAC and LS, FCN can achieve higher segmentation accuracy (∼25% increase). We attribute this to FCNs combining deep semantic information in the upper layers and shallow appearance information in the lower layers in a hierarchical manner, so it can encode image-wide location and semantic information. The marginal improvement (∼1.5%) of FCN over the STF method was likely due to the pre- and post-processing techniques.

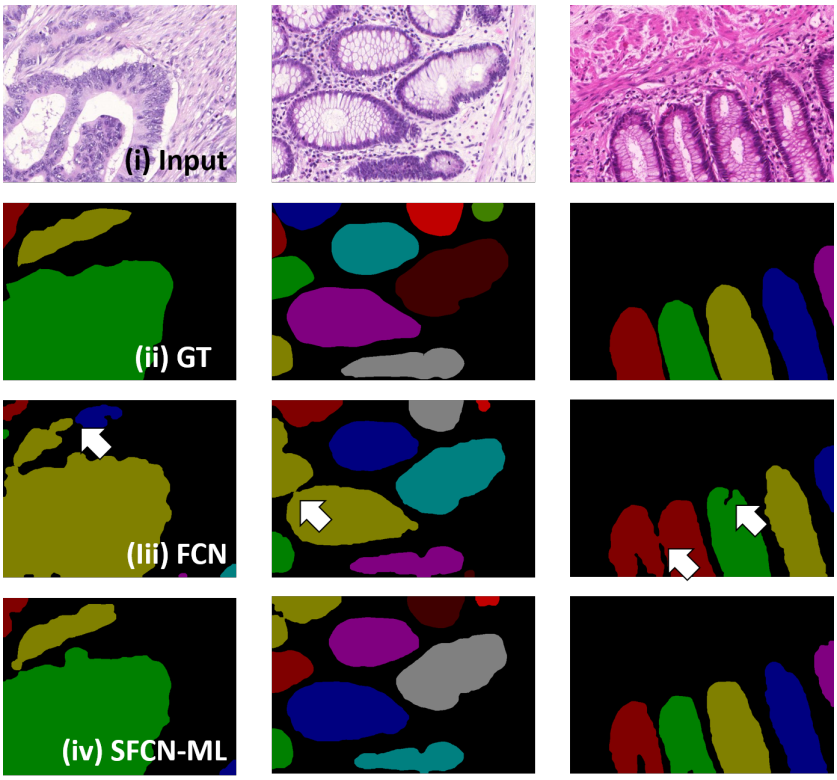

Fig. 5 Segmentation results from 3 example GlaS studies. (i) input images, (ii) ground truth annotation (GT), (iii, iv) segmentation results from FCN and our SFCN-ML methods. Different colors represent different gland regions.

As expected, the semi-automated methods CM and SnapCut methods produced the best results due to refinements from user interactions. User interactions usually carry a priori knowledge and enable detection and refinement of the segmentation from user selected regions, and therefore perform better than the FCNs that do not use any refinements. Nevertheless, due to the use of low-level features, these methods are unable to refine challenging regions e.g., low-texture difference to the background and inhomogeneous lesions. In contrast, the proposed SFCN-FL, SFCN-BL and SFCN-ML methods refined the boundary definitions of the challenging lesions based on using high-level semantic features and learned errors in a stacked architecture, and therefore improved the Jaccard measure by ∼5—7%. We suggest that our methods could be improved with additional information from user interactions.

Table 6 shows our methods when compared to recent published FCN based methods. The lower performance of FCN-32s was due to the large stride (32) used. The large stride size usually resulted in more coarse outputs. The improvement of FCN+SM and FCN+SM+TP over DeepLab and CRF-RNN was likely due to the graph model used. FCN+CM and FCN+SM+TP used a dedicated graph model

for gland segmentation, while the graph model used in DeepLab and CRF-RNN was optimized for general images.

When compared with FCN, our SFCN-ML method has the benefit of correctly separating different glands, where FCN usually over- or under-separated the glands, as shown by the example in Fig. 5 (indicated by arrows).

While FCN+CM and FCN+CM+TP perform slightly better than our method (0.5% in accuracy measure), their reliance on dataset specific graph models may limit their generalizability. In addition, both FCN+CM and FCN+CM+TP require careful tuning and do not often produce stable DiceObj and accuracy results, which may limit their adoption in clinical environments. In contrast, our method is dataset agnostic and can achieve similar results without adding refinement techniques. This suggests that our method may further benefit through the addition of graph models.

5 Conclusions

In this paper, we outline a new FCN-based method to automatically segment ROIs in medical images. Our method obtained accurate segmentation by combining the important visual characteristics of the foreground and background channels, which were then iteratively learned and gradually inferred from the stacked FCN architecture. Our method leverages the capacity of FCNs to segment ROIs without using any pre- or post- processing techniques e.g., filtering, de-noising and graph models. Our experiments on three datasets across disparate imaging modalities show that our method had higher segmentation accuracy compared to conventional FCNs and the state-of-the-art methods. In the future, we will investigate adapting our method to other datasets, and how our methods could be improved by user interaction and pre- and post-processing techniques.

References

1. M. A. Musen, et al., "Clinical decision-support systems," Biomedical informatics, Springer, 643-674 (2014).
2. K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," The British journal of radiology (2014).
3. L. Bi, et al., "Automatic detection and classification of regions of FDG uptake in whole-body PET-CT lymphoma studies," Computerized Medical Imaging and Graphics (2016).
4. X. Chen, et al., "Medical image segmentation by combining graph cuts and oriented active appearance models," IEEE Transactions on Image Processing, 21 (2012).
5. E. Ahn, et al., "Saliency-based Lesion Segmentation via Background Detection in Dermoscopic Images," IEEE Journal of Biomedical and Health Informatics (2017).
6. B. N. Li, et al., "Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation," Computers in biology and medicine, 41, 1-10 (2011).
7. J. Long, et al., "Fully convolutional networks for semantic segmentation," in Proc. IEEE CVPR, 3431-40 (2015).
8. L.-C. Chen, et al., "Semantic image segmentation with deep convolutional nets and fully connected crfs," in Proc. ICLR (2015).
9. S. Zheng, et al., "Conditional random fields as recurrent neural networks," in Proc. IEEE ICCV, 1529-37 (2015).
10. H. Chen, et al., "DCAN: Deep contour-aware networks for object instance segmentation from histology images," Medical Image Analysis, (2016).
11. Q. Dou, et al., "3d deeply supervised network for automatic liver segmentation from ct volumes," in Proc. MICCAI, 149-157 (2016).

12. Y. Xu, et al., "Gland instance segmentation by deep multichannel side supervision," MICCAI, 496-504 (2016).
13. A. BenTaieb, et al., "Topology Aware Fully Convolutional Networks for Histology Gland Segmentation," in Proc. MICCAI, 460-8 (2016).
14. F. Paulano, et al., "3D segmentation and labeling of fractured bone from CT images," The Visual Computer, 30, 939-48, (2014).
15. L. Lin, et al., "Inference with collaborative model for interactive tumor segmentation in medical image sequences," IEEE transactions on cybernetics, 46, 2796-2809, (2016).
16. L. Bi, et al., "Cellular automata and anisotropic diffusion filter based interactive tumor segmentation for positron emission tomography," IEEE EMBC, 5453-5456 (2013).
17. G. T. ONeill, et al., "Segmentation of cam-type femurs from CT scans," The Visual Computer, 28, 205-218 (2012).
18. M. Silveira, et al., "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," IEEE Journal of Selected Topics in Signal Processing (2009).
19. M. Emre Celebi, et al., "Lesion border detection in dermoscopy images using ensembles of thresholding methods," Skin Research and Technology, 19, e252-e258 (2013).
20. C. Li, et al., "Minimization of region-scalable fitting energy for image segmentation," IEEE transactions on image processing, 17, 1940-1949 (2008).
21. C. Li, et al., "Distance regularized level set evolution and its application to image segmentation," IEEE Transactions on image processing, 19, 3243-3254 (2010).
22. A. Wong, et al., "Automatic skin lesion segmentation via iterative stochastic region merging," IEEE Transactions on Information Technology in Biomedicine (2011).
23. K. Somkantha, et al., "Boundary detection in medical images using edge following algorithm based on intensity gradient and texture gradient features," IEEE transactions on biomedical engineering, 58, 567-573, (2011).
24. A. Roy, et al., "JCLMM: A Finite Mixture Model for Clustering of Circular-Linear data and its application to Psoriatic Plaque Segmentation," Pattern Recognition, (2017).
25. R. Nock et al., "Statistical region merging," IEEE Transactions on pattern analysis and machine intelligence, 26, 1452-1458, (2004).
26. E. M. van Rikxoort, et al., "Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus," Medical Image Analysis, 14, 39, (2010).
27. Y. Song, et al., "Similarity guided feature labeling for lesion detection," in Proc. MICCAI, 284-291 (2013).
28. C. Li, et al., "Automated PET-guided liver segmentation from low-contrast CT volumes using probabilistic atlas," Computer methods and programs in biomedicine (2012).
29. L. Bi, et al., "Automatic Descending Aorta Segmentation in Whole-Body PET-CT Studies for PERCIST-Based Thresholding," in Proc. DICTA, 1-6 (2012).
30. C. Li, et al., "Joint probabilistic model of shape and intensity for multiple abdominal organ segmentation from volumetric CT images," IEEE journal of biomedical and health informatics, 17, 92-102, (2013).
31. I. Isgum, et al., "Multi-atlas-based segmentation with local decision fusionApplication to cardiac and aortic segmentation in CT scans," IEEE Transactions on Medical Imaging, 28, 1000-1010, (2009).
32. L. Bi, et al., "Multi-stage Thresholded Region Classification for Whole-Body PET-CT Lymphoma Studies," in Proc. MICCAI, 569-576 (2014).
33. Y. Song, et al., "A multistage discriminative model for tumor and lymph node detection in thoracic images," IEEE transactions on Medical Imaging, 31, 1061-1075, (2012).
34. C. Lartizien, et al., "Computer-Aided Staging of Lymphoma Patients With FDG PET/CT Imaging Based on Textural Information," IEEE Journal of Biomedical and Health Informatics, 18, 946-955, (2014).
35. D. Ciresan, et al., "Deep neural networks segment neuronal membranes in electron microscopy images," in Proc. NIPS, 2843-2851 (2012).
36. K. H. Cha, et al., "Urinary bladder segmentation in CT urography using deeplearning convolutional neural network and level sets," Medical physics, 43, 1882-1896, (2016).
37. M. Havaei, et al., "Brain tumor segmentation with deep neural networks," Medical image analysis, 35, 18-31, (2017).
38. H. R. Roth, et al., "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in Proc. MICCAI, 556-564 (2015).
39. A. Farag, et al., "A Bottom-up Approach for Pancreas Segmentation using Cascaded Superpixels and (Deep) Image Patch Labeling," IEEE Transactions on Image Processing, 26, 386-399, (2017).

40. O. Ronneberger, et al., "U-net: Convolutional networks for biomedical image segmentation," MICCAI (2015).
41. H. Fu, et al., "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in Proc. IEEE ISBI, 698-701 (2016).
42. F. Lu, et al., "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," International Journal of Computer Assisted Radiology and Surgery, 1-12, (2016).
43. P. Hu, et al., "Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets," International Journal of Computer Assisted Radiology and Surgery, 1-13, (2016).
44. K. Chatfield, et al., "Return of the devil in the details: Delving deep into convolutional nets," BMVC (2014).
45. L. Wang, et al., "Saliency detection with recurrent fully convolutional networks," in Proc. ECCV, 825-841 (2016).
46. J. Dean, et al., "Large scale distributed deep networks," in NIPS, 1223-1231 (2012).
47. L. Hogeweg, et al., "Clavicle segmentation in chest radiographs," Medical image analysis, 16, 1490-1502 (2012).
48. K. Sirinukunwattana, et al., "A stochastic polygons model for glandular structures in colon histology images," IEEE transactions on medical imaging, 2366-78 (2015).
49. H.-C. Shin, et al., "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE transactions on medical imaging, (2016).
50. J. Deng, et al., "Imagenet: A large-scale hierarchical image database," in Proc. CVPR, 248-255, (2009).
51. A. Kumar, et al., "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," IEEE Journal of Biomedical and Health Informatics (2016).
52. A. Vedaldi et al., "Matconvnet: Convolutional neural networks for matlab," in Proc. ACM MM, 689-692 (2015).
53. A. Krizhevsky, et al., "Imagenet classification with deep convolutional neural networks," in NIPS, 1097-1105 (2012).
54. X. Li, et al., "Contextual hypergraph modeling for salient object detection," in Proc. ICCV, 3328-3335 (2013).
55. T. F. Chan and L. A. Vese, "Active contour and segmentation models using geometric PDEs for medical imaging," Geometric methods in bio-medical image processing (2002).
56. Y. Zhang, et al., "Medical image segmentation using new hybrid level-set method," in Proc. MEDIVIS, 71-76 (2008).
57. J. Shotton, et al., "Semantic texton forests for image categorization and segmentation," in Proc. CVPR, 1-8 (2008).
58. X. Bai, et al., "Video snapcut: robust video object cutout using localized classifiers," in ACM Transactions on Graphics (2009).