This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2624198, IEEE Transactions on Image Processing

A Bottom-up Approach for Pancreas Segmentation using Cascaded Superpixels and (Deep) Image Patch Labeling

Amal Farag, Le Lu*, *Senior Member, IEEE*, Holger R. Roth, Jiamin Liu, Evrim Turkbey, Ronald M. Summers*

***Abstract*—Robust organ segmentation is a prerequisite for computer-aided diagnosis (CAD), quantitative imaging analysis, pathology detection and surgical assistance. For organs with high anatomical variability (e.g., the pancreas), previous segmentation approaches report low accuracies, compared to well studied organs, such as the liver or heart. We present an automated bottom-up approach for pancreas segmentation in abdominal computed tomography (CT) scans. The method generates a hierarchical cascade of information propagation by classifying image patches at different resolutions and cascading (segments) superpixels. The system contains four steps: 1) decomposition of CT slice images into a set of disjoint boundary-preserving superpixels; 2) computation of pancreas class probability maps via dense patch labeling; 3) superpixel classification by pooling both intensity and probability features to form empirical statistics in cascaded random forest frameworks; and 4) simple connectivity based post-processing. Dense image patch labeling is conducted using two methods: efficient random forest classification on image histogram, location and texture features; and more expensive (but more accurate) deep convolutional neural network classification, on larger image windows (i.e., with more spatial contexts). Over-segmented $2D$ CT slices by the Simple Linear Iterative Clustering (SLIC) approach are adopted through model/parameter calibration and labeled at the superpixel level for positive (pancreas) or negative (non-pancreas or background) classes.**

The proposed method is evaluated on a dataset of 80 manually segmented CT volumes, using six-fold cross-validation. Its performance equals or surpasses other state-of-the-art methods (evaluated by "leave-one-patient-out"), with a Dice coefficient of 70.7% and Jaccard Index of 57.9%. In addition, the computational efficiency has improved significantly, requiring a mere $6 \sim 8$ minutes per testing case, versus ≥ 10 hours for other methods. The segmentation framework using deep patch labeling confidences is also more numerically stable, as reflected in the smaller performance metric standard deviations. Finally, we implement a multi-atlas label fusion (MALF) approach for pancreas segmentation using the same dataset. Under six-fold cross-validation, our bottom-up segmentation method significantly outperforms its MALF counterpart: $70.7 \pm 13.0\%$ versus $52.51 \pm 20.84\%$ in Dice coefficients.

***Index Terms*—Abdominal Computed Tomography (CT), Deep Convolutional Neural Networks, Dense Image Patch Labeling, Cascaded Random Forest, Pancreas Segmentation.**

Amal Farag, Le Lu, Holger R. Roth, Jiamin Liu and Ronald M. Summers are with Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA. Le Lu and Ronald M. Summers, are also with Clinical Image Processing Service, Department of Radiology and Imaging Sciences, NIH.

Amal Farag is currently affiliated with Kentucky Imaging and Technologies, Louisville, Kentucky; Holger R. Roth is currently with Nagoya University, and Evrim Turkbey was a member of Clinical Image Processing Service, NIH and now is with School of Medicine, Johns Hopkins University, 733 N Broadway, Baltimore, Maryland. Contact e-mails: {le.lu, rms}@nih.gov.

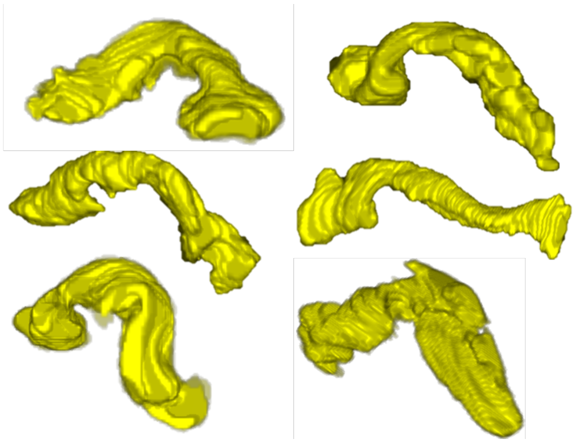

Fig. 1. 3D manually segmented volumes of six pancreases from six patients. Notice the variations in shape and size.

I. Introduction

Segmentation of abdominal organs such as the spleen, liver and pancreas in abdominal computed tomography (CT) scans is a crucial task in computer aided diagnosis (CAD), for quantitative and qualitative analysis and for surgical assistance. Pancreas segmentation, in particular, is a critical component in CAD systems that perform quantitative imaging analysis of diabetic patients or pancreatic cancer detection. At present, high accuracy levels (e.g., Dice coefficients $> 90\%$) have been reported for automatic segmentation methods on CT of other organs, such as the kidneys [1], lungs [2], heart [3] and liver [4]. However, automatic pancreas segmentation remains a challenge. Very few single or multi-organ setups have been designed for this problem. The pancreas exhibits high anatomical variability in terms of shape and size. Furthermore, its location in the abdominal cavity varies from patient to patient. The degree of boundary contrast depends on the amount of visceral fat surrounding the pancreas. These factors, along with many others, make pancreas segmentation very complex. The manually segmented 3D volumes of patient pancreases in Fig. 1 illustrate the variations and challenges mentioned. From the above observations, we conclude that automated pancreas segmentation should follow a different approach than current organ segmentation methods, which often use statistical shape models.

This paper proposes a new, bottom-up method using image and (deep) patch-level labeling confidences for pancreas seg-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2624198, IEEE Transactions on Image Processing

mentation. The method is applied to 80 single phase CT patient volumes. The method is designed to improve the segmentation accuracy of *highly deformable organs*, like the pancreas, by leveraging the *middle-level representation of image segments*. First, all 2D slices of an input patient abdominal CT scan are over-segmented into superpixels. Second, the superpixels are classified into two semantic classes (pancreas and non-pancreas) using multi-stage feature extraction and a random forest (RF) classification process on the image and (deep) patch-level confidence maps, pooled at the superpixel level. Two cascaded random forest superpixel classification frameworks are presented and compared. Fig. 2 presents the first proposed framework. Fig. 5 illustrates the modularized flow charts of both frameworks. All experiments are conducted under six-fold cross-validation. Our system's speed is about two orders of magnitude greater than the speed of the atlas registration based approaches to process a new testing case [5], [6], [7], [8], [9], [10]. Furthermore, our results equal or surpass those from other state-of-the-art methods (evaluated by ”leave-one-patient-out”), with a Dice coefficient of 70.7% and Jaccard Index of 57.9%. Under six-fold cross-validation, our bottom-up segmentation method significantly outperforms its “multi-atlas registration and joint label fusion” (MALF) counterpart (based on our implementation borrowed from [11], [12]): Dice coefficients $70.7 \pm 13.0\%$ versus $52.51 \pm 20.84\%$. Additionally, another bottom-up supervoxel based multi-organ segmentation without registration in 3D abdominal CT images is also investigated [13] in a similar spirit, for demonstrating this methodological synergy.

II. Previous Work

Organ segmentation methods can be divided into two categories: top-down and bottom-up methods. In top-down methods, a-priori knowledge such as atlas(es) and/or shape models of the organ are generated and incorporated into the framework via learning based shape model fitting [1], [3], [4] or volumetric image registration [7], [8], [10]. In bottom-up methods, segmentation is performed by local image similarity grouping and growing [2] or pixel, superpixel/supervoxel based labeling [14]. In general, top-down methods are suitable for organs which can be modeled well by statistical shape models [1], [3]; whereas, bottom-up methods are more effective for highly non-Gaussian shaped [14] or pathological organs.

Most previous work on pancreas segmentation from CT images adopts top-down methods [5], [6], [7], [8], [9], [10]. [5] is not comparable to the others, as it uses three-phase contrast enhanced CT images, as opposed to single phase CT images.

- Shimizu et. al [5] use three-phase contrast enhanced CT data, which are first registered together for a particular patient and later registered to a reference patient by landmark-based deformable registration. The spatial support area of the abdominal cavity is reduced by segmenting the liver, spleen and three main vessels used to interpret the pancreas's location (i.e., splenic, portal and superior mesenteric veins). Coarse-to-fine pancreas segmentation is performed by using generated patient-specific probabilistic atlas guided segmentation, followed by intensity-based classification and post-processing. This method was validated on 20 multi-phase datasets, and yielded a Jaccard index of 57.9%.
- Okada et. al [6] perform multi-organ segmentation by combining inter-organ spatial interrelations with probabilistic atlases. The approach incorporates various a-priori knowledge into the model that includes the shape representations of seven organs. The method was validated on 28 abdominal contrast-enhanced CT datasets, and yielded an overall volume overlap of Dice index 46.6% for the pancreas.
- Chu et. al [8] present an automated multi-organ segmentation method based on spatially-divided probabilistic atlases. The algorithm consists of image-space division and a multi-scale weighting scheme to handle the large differences among patients in organ shape and position in local areas. Their experimental results show that the liver, spleen, pancreas and kidneys can be segmented with Dice similarity indices of 95.1%, 91.4%, 69.1%, and 90.1%, respectively, using 100 annotated abdominal CT volumes.
- Wolz et. al [7] use a multi-organ segmentation approach that combines hierarchical weighted subject-specific atlas-based registration and patch-based segmentation. Segmentation predictions are post-processed using optimized graph-cuts with a learned intensity model. The method yields state-of-the-art results, with the mean pancreas segmentation Dice coefficients of 69.6% on 150 patients and 58.2% on a sub-population of 50 patients.
- Recent work by Wang et. al [10] presents a patch-based label propagation approach that uses relative geodesic distances. This method can be considered a preview for bottom-up segmentation, where affine registration between dataset and atlases were conducted followed by refinement, which uses the patch-based segmentation to reduce misregistrations and instances of high anatomical variability. The approach, evaluated on 100 abdominal CT scans, yielded a mean Dice coefficient of 65.5% for pancreas segmentation.

Many of the atlas based approaches [5], [6], [7], [8], [9], [10] conduct experiments under a “leave-one-patient-out” or “leave-one-out” (LOO) scheme, for up to N=150 patients. In the clinical setting, leave-one-out based dense volume registration (from all other N-1 patients as atlas templates) and label fusion may be computationally impractical (10+ hours per testing case). More importantly, it does not scale up easily to large datasets.

The proposed bottom-up approach is far more efficient in memory and computation speed than the multi-atlas registration framework [10], [7], [8], [9], [5], [6]. It is evaluated on 80 manually segmented CT patient volumes, under six-fold cross-validation (CV). The method's results equal or surpass those of state-of-the-art methods (even under “leave-one-patient-out”, or LOO), with Dice coefficient 70.7% and Jaccard Index 57.9%. Strict quantitative comparison is not possible, as experiments cannot be performed on the same

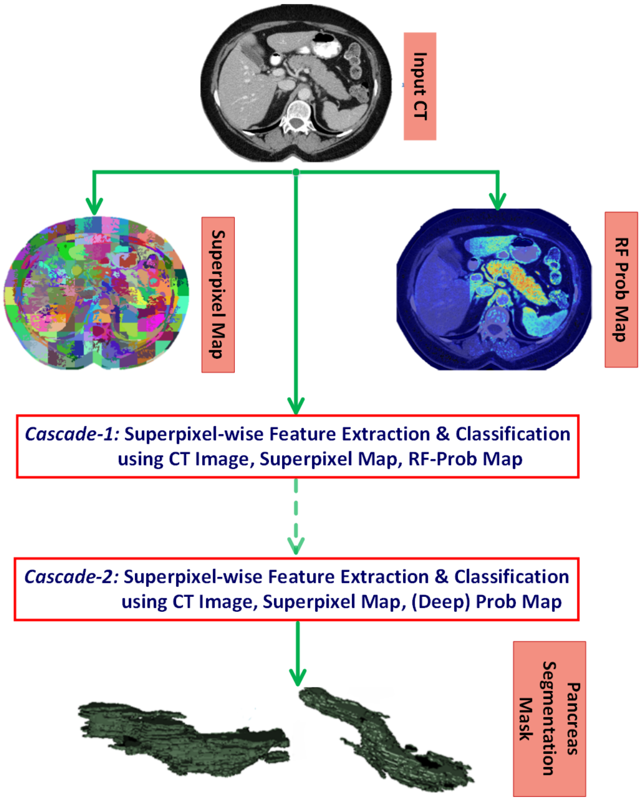
Input CT
Superpixel Map
RF Prob Map
Cascade-1: Superpixel-wise Feature Extraction & Classification using CT Image, Superpixel Map, RF-Prob Map
Cascade-2: Superpixel-wise Feature Extraction & Classification using CT Image, Superpixel Map, (Deep) Prob Map
Pancreas Segmentation Mask

Fig. 2. Overall pancreas segmentation framework via cascaded superpixel classification and dense image patch labeling. The dot-arrow flow represents the second cascade (Cascade-2) of random forest based superpixel classification which is built upon the same information channels of CT image intensity, superpixel maps and patch labeling probability maps as cascade-1. The difference is that the patch labeling maps in Cascade-2 can be generated by classifying hand-crafted image features (as the only option in Cascade-1) or feed-forwarding deep convolutional neural networks. Cascade-2 runs on the small subset of superpixels that pass Cascade-1 (i.e., without being rejected as background).

datasets [1]. We instead implement a multi-atlas label fusion (MALF) pancreas segmentation approach [11], [12] using our datasets. Under six-fold CV, our bottom-up segmentation method clearly outperforms its MALF counterpart: Dice coefficients $70.7 \pm 13.0\%$ versus $52.51 \pm 20.84\%$.

Superpixel based representation for pathological region detection and segmentation has been recently studied in [15] using MRI datasets. Under any problem representation, the visual feature extraction and classification framework [15] differ significantly from those in our approach. Our bottom-up image parsing method, namely the usage of superpixels as an intermediate level image representation, is partially inspired by similar approaches in PASCAL semantic segmentation [16] and scene labeling [17]. However, very different technical and algorithmic components are required, due to the pancreas's small size and the high unbalance between pancreas and non-pancreas tissues. Indeed, the pancreas generally occupies less than 1% of space in a CT scan. We here use a "Cascaded Superpixels" representation to address this domain-specific challenge prevalent in medical image analysis.

III. Methods

This section presents the steps in our algorithm (Sec. III-A and Sec. III-B). A diagram is provided in Fig. 2. The method's extensions exploiting sliding-window CNN based dense image patch labeling and framework variations are described in Sec. III-C and Sec. III-D.

A. Boundary-preserving Over-segmentation

Over-segmentation denotes the segmentation or decomposition of images (or more generally grid graphs) into smaller perceptually meaningful regions, "superpixels". Within a superpixel, pixels are similar in color, texture, intensity, etc. and generally align with image edges. Hence, superpixels need not form rectangular patches; they may be irregular in shape and size. In the computer vision literature, numerous approaches have been proposed for superpixel segmentation [18], [19], [20], [21], [22]. Superpixel methods fall under two main categories: graph-based (e.g., Simple Linear Iterative Clustering or SLIC [18], entropy rate [20] and [21]) and gradient ascent methods (e.g., watershed [22] and mean shift [23]). We select three graph-based superpixel algorithms (SLIC [18], [19], efficient graph-based [21] and Entropy rate [20]) and one gradient ascent method (watershed [22]), due to their remarkable speed, efficiency, and accuracy. The algorithms in [21], [22] are relatively fast, at $O(MlogM)$ complexity, where M is the number of pixels or voxels in the image or grid graph. Mean shift [23] and normalized cuts [24] are $O(M^2)$ and $O(M^{\frac{3}{2}})$, respectively. Simple linear iterative clustering (SLIC) [18] is both fast and memory efficient. Fig. 3 shows sample superpixel results using the SLIC approach. Original CT slices and cropped, enlarged pancreas superpixel regions are also illustrated. From our evaluation, the SLIC approach yields high *boundary recalls* (within the distance range of [1,6] pixels from the semantic pancreas ground-truth boundary annotation), thus indicating that most superpixels were located in the same 3×3 neighborhood as ground truth pancreas boundaries. The watershed approach provided the least promising results for usage in the pancreas, due to the lack of conditions in the approach, to utilize boundary information in conjunction with intensity information as implemented in graph-based approaches. The superpixel number range per axial image is constrained $\in [100, 200]$ to make a good trade-off on superpixel dimensions or sizes.

The *overlap ratio* r is defined as the percentage of pixels/voxels inside each superpixel that are annotated as pancreas in the ground-truth pancreas mask. We can obtain binary pancreas segmentation results by thresholding on r, namely choosing a value τ such that $r > \tau$ indicates a pancreas superpixel. When $\tau = 0.50$, the mean (standard deviation) Dice coefficient is $81.2\% \pm 3.3\%$. This is called the "Oracle" segmentation accuracy, since computing r requires knowledge of the ground-truth segmentation. "Oracle" is also the upper bound for our superpixel labeling and classification framework. The Dice coefficient of $81.2 \pm 3.3\%$ is significantly higher and more numerically stable (in standard deviation) than those from

[1] Our annotated pancreas segmentation datasets are publicly available at https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT, to ease future comparisons.

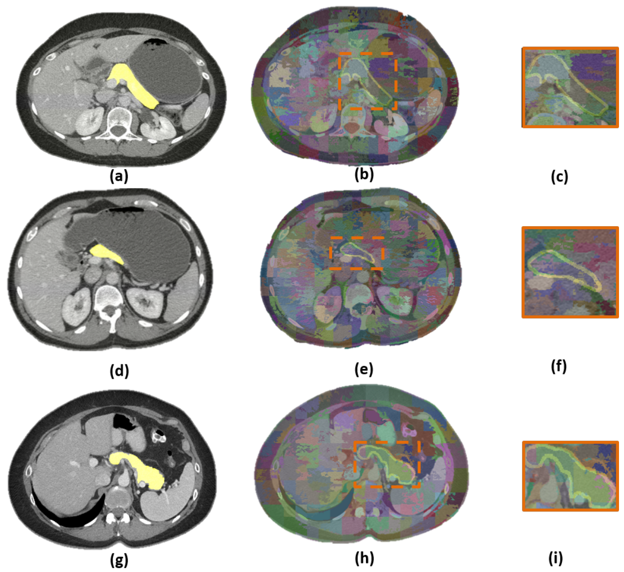

Fig. 3. Sample superpixel generation results from the SLIC method [18]. The first column depicts different slices from different patient scans with the ground-truth pancreas segmentation in yellow (a, d & g). The second column depicts the over-segmentation results with the pancreas contours superimposed on the image (b, e & h). Last, (c) (f) and (i) show zoomed-in areas of the pancreas superpixel results from (b) (e) and (h).

previous state-of-the-art methods [10], [7], [8], [9], [5]. The choices of SLIC and $\tau = 0.50$ have been calibrated using the training CT scans. We empirically find that there is no need to change the superpixel generation method, its parameters and the threshold $\tau = 0.50$ in each round of cross-validation. Our superpixel calibration procedure generalizes well to all of our datasets. Voxel-level pancreas segmentation can be propagated from superpixel-level classification and further improved by efficient narrow-band level-set based curve evolution [25], or the learned intensity model based graph-cut [7].

Mid-level visual representations like superpixels have been widely adopted in recent *PASCAL Visual Object Classes* or *VOC* semantic segmentation challenges [16] from computer vision. Generally speaking, a diverse set of plausible segments are expected per image to be fully exploited by later employed probabilistic models [26]. For many patients (especially those with low body fat percentages) in our study, the degree of contrast for pancreas boundaries can be much weaker than that for boundaries in *PASCAL VOC* images (in which the challenge rather lies in the clutter among multi-class objects).

Although superpixels may be extended to supervoxels, we maintain superpixel representation, due to the potential boundary leakage problem of supervoxels that may severely deteriorate pancreas segmentation in multiple CT slices. Object level pancreas boundaries in abdominal CT images can appear very weakly, especially for patients with low body fat percentages. On the other hand, as in Sec. III-B, image patch based appearance descriptors and statistical models are well developed for 2D images (in both computer vision and medical imaging). 3D statistical texture descriptors and models to accommodate supervoxels can be computationally expensive, with no significant improvement in performance [14], [27].

B. Patch-level Visual Feature Extraction and Classification via Random Forest: P^{RF}

In this work, a total of 46 patch-level image features depicting the pancreas and background are extracted. Each patient's 3D abdominal body region is first segmented and identified using a standard table-removal procedure, in which all extra-body voxels are removed.

1), To describe the texture, we adopt the Dense Scale-Invariant Feature transform (dSIFT) approach [28], extending the SIFT descriptor [29] in several aspects. The publicly available VLFeat implementation of the dSIFT is employed [28]. The descriptors are densely and uniformly extracted from image grids separated by 3 pixels. The patch centers correspond to the green points on the original image slice. From the center coordinates, the dSIFT is computed with $[2x2]$ bins and bin size 6 pixels, which results in a 32 dimensional texture descriptor for each image patch. The image patch size is fixed at 25×25, as a trade-off between computational efficiency and description power. Image patch sizes ranging from 15 to 35 pixels are tested using a small sub-sampled dataset for classification, as we later describe. Stable performance statistics are observed and quantitative experimental results using the default patch size of 25×25 pixels are reported.

2), A second feature group using the pixel intensity histograms of the ground-truth pancreas and the surrounding tissues is built on the class-conditional probability density functions (PDF). A kernel density estimator (KDE[2]) is created using image intensities from the training patient CT scans. The KDE represents the CT intensity distributions of the positive (pancreas) $\{X^+\}$ and negative (non-pancreas) classes $\{X^-\}$ of pixels. The positive sample set comprises all pixels residing inside pancreas. However, since negative pixels greatly outnumber positive pixels, only 5% of negative pixels from each CT scan (by random resampling) is considered. Let $\{X^+\} = \left(h_1^+, h_2^+, \cdots, h_n^+\right)$ and $\{X^-\} = \left(h_1^-, h_2^-, \cdots, h_m^-\right)$ where h_n^+ and h_m^- represent the pixel intensities of positive and negative samples from the training patient scans. The kernel density estimators $f^+(X^+) = \frac{1}{n}\sum_{i=1}^{n} K\left(X^+ - X_i^+\right)$ and $f^-(X^-) = \frac{1}{m}\sum_{j=1}^{m} K\left(X^- - X_j^-\right)$ are computed, where $K()$ is assumed to be a Gaussian kernel with optimal computed bandwidth 3.039 for this data. Kernel sizes or bandwidth may be selected automatically using a 1D Likelihood-based search, as provided by the used KDE toolkit. The normalized likelihood ratio is computed as a function of CT intensity in the range of $H = [0 : 1 : 4095]$. Thus, the probability of a superpixel being classified as pancreas is formulated as: $y^+ = \frac{(f^+(X^+))}{(f^+(X^+)+f^-(X^-))}$. This function is converted into a pre-computed look-up table over $H = [0 : 1 : 4095]$ (the full CT intensity range), which allows for very efficient $O(1)$ access time.

3), The third feature group contains object boundary-preserving intensity features. Features are computed as follows. The same KDE response statistics within the intersected sub-regions $\acute{P} \in P$ (where P presents a generic image patch) are extracted, by first utilizing the KDE probability

[2] http://www.ics.uci.edu/~ihler/code/kde.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2624198, IEEE Transactions on Image Processing

response maps above and later the superpixel CT masks described in Sec. III-A, as underlying supporting masks to each image patch. The idea is that an image patch, P, may be divided into more than one superpixel. This set of statistics is calculated with respect to the most representative superpixel (the superpixel covering the patch's center pixel).

4), The final two features for each axial slice (in the patient volumes) are the normalized relative x-axis and y-axis positions $\epsilon[0, 1]$, computed at each image patch center against the segmented body region (self-normalized [3] for patients with different body mass indices). After concatenating all features, a total of 46 image patch-level features per superpixel are used to train a random forest (RF) classifier C_p. Image patch labels are obtained by directly borrowing the class information of their patch center pixels, based on the manual segmentation.

To resolve the data unbalance issue between the positive (pancreas) and negative (non-pancreas) class patches (during RF training), the sample weights for both classes are normalized so that the sum of the sample weights for each class reaches the same constant (i.e., assuming a uniform prior). Pancreas class patch samples are weighted much more heavily than non-pancreas background class instances, since they are more rare. A similar sample re-weighting scheme is exploited in the regression forest for anatomy localization [30]. RF training is conducted under six-fold cross-validation. In summary, SIFT and its variations, e.g., D-SIFT have proven to be informative, especially through spatial pooling or packing [31]. A wide range of pixel-level correlations and visual information per image patch is also captured by the remaining 14 features. Both high classification specificity and recall have been obtained in cross-validation using a Random Forest implementation with 50 trees and minimum leaf size set to 150 (i.e., using the $treebagger(\bullet)$ function in Matlab).

C. Patch-level Classification via Deep Convolutional Neural Network: P^{CNN}

In this work, we use a Convolutional Neural Network (CNN) for binary image patch classification. The model architecture is composed of five convolutional layers with max-pooling and two fully-connected layers containing DropOut [32] connections, with probability 0.5. Our CNN model ends with a final two-way softmax classification layer for 'pancreas' and 'non-pancreas' classes (refer to Fig. 4). In testing, no DropOut operation is necessary. Modern GPU acceleration allows efficient training and run-time evaluation of deep CNN models. We use the publicly available *cuda-convnet2*[4] which is a simplified version of AlexNet [33]. In each round of cross-validation, approximately $0.6 \sim 0.8$ million image patches are used for training (obtained using resampling and data augmentation [34]).

To extract dense image patch response maps, we use a straight-forward sliding window approach that extracts 2.5D image patches composed of axial, coronal and sagittal planes at all image positions. Since the deep CNN architecture can encode large scale image patches (even the whole 224×224 pixel images [33]) very efficiently, hand crafted image features are no longer required. In this paper, the dimensions of the image patches for CNN training are 64×64 pixels, which is significantly larger than 25×25 from Sec. III-B. The larger spatial context is expected to improve the accuracy in patch labeling. For efficiency purposes, we extract patches at every ℓ^{th} pixel for the CNN feed-forward evaluation and later apply the nearest neighbor interpolation[5] to estimate the values at skipped pixels. We denote the CNN model generated probability maps as P^{CNN}.

The computational cost of deep CNN patch labeling per patch (via a sliding window) is still higher than the cost in Sec. III-B. In practice, dense patch labeling by P^{RF} runs exhaustively at 3 pixel intervals, while P^{CNN} is only evaluated at pixel locations that pass the first stage of a cascaded random forest superpixel classification framework. This process is detailed in Sec. III-D, where at the initial layer of the cascade, C^1_{SP} is operated at a high recall (close to 100%) but low specificity mode to minimize the false negative rate (FNR). The other important reason for doing so is to greatly reduce the training unbalance for P^{CNN} in C^3_{SP}. After this initial pruning, the ratio of non-pancreas to pancreas superpixels decreases from > 100 to ~ 5.

D. Superpixel-level Feature Extraction, Cascaded Classification and Pancreas Segmentation

In this section, we train three different superpixel-level random forest classifiers: C^1_{SP}, C^2_{SP} and C^3_{SP}. These classifiers are used to form two cascaded RF classification frameworks (F-1, F-2), as shown in Fig. 5. The superpixel labels are inferred from the overlap ratio r (defined in Sec. III-A). If $r \geq 0.5$, the superpixel is labeled as positive, while if $r \leq 0.2$, the superpixel is labeled as negative. The few remaining superpixels for which $0.2 < r < 0.5$ (a relatively very small portion/subset of all superpixels, i.e., less than 3%) are considered ambiguous and are excluded from training.

Training C^1_{SP} utilizes both the original CT image slices (I^{CT} in Fig. 5) and the probability response maps (P^{RF}) via the hand-crafted feature based patch-level classification (described in Sec.III-B). The 2D superpixel supporting maps (computed in Sec.III-A) are used for feature pooling and extraction on a superpixel level. The CT pixel intensity/attenuation numbers and the per-pixel pancreas class probability response values (from dense patch labeling of P^{PF} or P^{CNN} later) within each superpixel are treated as two empirical unordered distributions. Thus our superpixel classification problem is converted as modeling the difference between empirical distributions of positive and negative classes. We compute 1) simple statistical features of the 1st-4th order statistics such as mean, std, skewness, kurtosis [35] and 2) histogram-type features of eight percentiles $(20\%, 30\%, \ldots, 90\%)$, per distribution in intensity or P^{RF} channel, respectively. Once

[3] The axial reconstruction CT scans in our study either have largely varying ranges or extend in the z-axis. If some anatomical landmarks, such as the bottom plane of liver, the center of kidneys, can be provided automatically, the anatomically normalized z-coordinate positions for superpixels can be computed and used as an additional spatial feature for RF classification.

[4] https://code.google.com/p/cuda-convnet2

[5] In our empirical testing, simple nearest neighbor interpolation seems sufficient, due to the high quality of the deep CNN probability predictions.

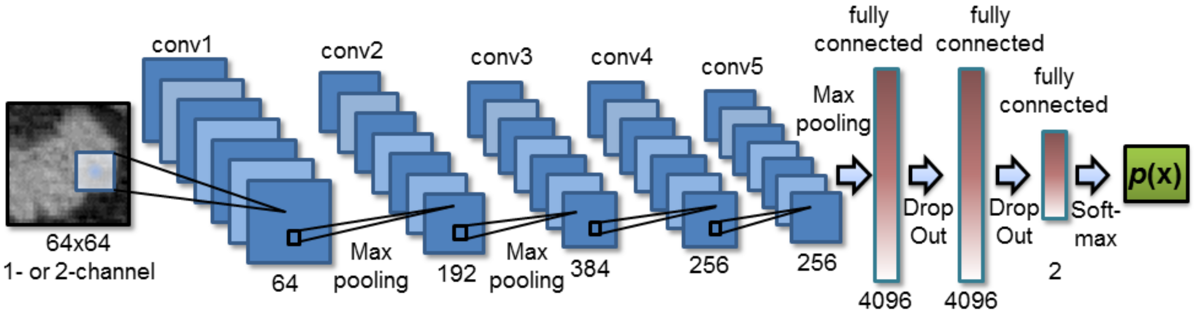

Fig. 4. The proposed CNN model architecture is composed of five convolutional layers with max-pooling and two fully-connected layers with DropOut [32] connections. A final 2-way softmax layer gives a probability $p(x)$ of "pancreas" and "non-pancreas" per data sample (or image patch). The number and model parameters for the convolutional filters and neural network connections for each layer are as indicated. The convolutional filter sizes are 9×9; 5×5; 5×5; 3×3; 3×3 and filter numbers are 64, 192, 384, 256, 256 for the five convolutional layers, respectively.

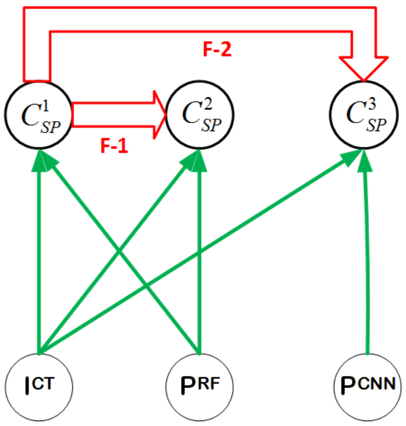

Fig. 5. The flow-chart of input channels and component classifiers to form the overall frameworks 1 (F-1) and 2 (F-2). I^{CT} indicates the original CT image channel; P^{RF} represents the probability response map by RF based patch labeling in Sec. III-B and P^{CNN} from deep CNN patch classification in Sec. III-C, respectively. Superpixel level random forest classifier C^1_{SP} is trained with all positive and negative superpixels in I^{CT} and P^{RF} channels; C^2_{SP} and C^3_{SP} are learned using only "hard negatives" and all positives, in the $I^{CT} \bigcup P^{RF}$ or $I^{CT} \bigcup P^{CNN}$ channels, respectively. Frameworks F-1 and F-2 are cascaded models formed from $C^1_{SP} \mapsto C^2_{SP}$, or $C^1_{SP} \mapsto C^3_{SP}$. Note that F-1 and F-2 share the first layer of classification cascades, in order to coarsely prune $\sim$ 96% initial superpixels using both intensity and P^{RF} features (refer to Fig. 8 **Left**).

concatenated, the resulted 24 features for each superpixel instance is fed to train random forest classifiers.

Due to the highly unbalanced quantities between foreground (pancreas) superpixels and background (the rest of CT volume) superpixels, a two-tiered cascade of random forests are exploited to address this type of rare event detection problem [36]. In a cascaded classification, C^1_{SP} once trained is applied exhaustively on scanning all superpixels in an input CT volume. Based on the receiver operating characteristic (ROC) curves in Fig. 8 (Left) for C^1_{SP}, we can safely reject or prune 97% negative superpixels while maintaining nearly $\sim 100\%$ recall or sensitivity. The remaining 3% of negatives, often referred as "hard negatives" [36], along with all positives, are employed to train the second C^2_{SP} in the same feature space. The combination of C^1_{SP} and C^2_{SP} is referred to as Framework 1 (F-1) in the following sections.

Similarly, we can train a random forest classifier C^3_{SP} by replacing C^2_{SP}'s feature extraction dependency on the P^{RF} probability response maps, with the deep CNN patch classification maps of P^{CNN}. The same 24 statistical moments and percentile features per superpixel, from two information channels I^{CT} and P^{CNN}, are extracted to train C^3_{SP}. Note that the CNN model that produces P^{CNN} is trained with the image patches sampled from only "hard negative" and positive superpixels (aligned with the second-tier RF classifiers C^2_{SP} and C^3_{SP}). For simplicity, P^{RF} is only trained once with all positive and negative image patches. This will be referred to as Framework 2 (F-2) in the following sections. F-1 only use P^{RF}, while F-2 depends on both P^{RF} and P^{CNN} (with minor additional computational cost).

The flow chart of frameworks 1 (F-1) and 2 (F-2) is illustrated in Fig. 5. The two-level cascaded random forest classification hierarchy is found empirically to be sufficient (although a deeper cascade is possible) and is implemented to obtain F-1: C^1_{SP} and C^2_{SP}, or F-2: C^1_{SP} and C^3_{SP}. Only intensity based statistical features for C^1_{SP} produce significantly inferior results. On the other hand, C^3_{SP} can be learned using all three available information channels of $I^{CT} \bigcup P^{RF} \bigcup P^{CNN}$ that will result in 36 superpixel-level features. Based on our initial empirical evaluation, $I^{CT} \bigcup P^{CNN}$ is as sufficient as using all three channels. Hence, $I^{CT} \bigcup P^{CNN}$ seems to be the optimal feature channel configuration considering both the classification effectiveness and model complexity. The coupling of P^{CNN} into C^3_{SP} consistently shows better segmentation results than P^{RF} for C^2_{SP}; whereas, P^{CNN} is not powerful enough to be used alone.

The binary 3D pancreas volumetric mask is obtained by stacking the binary superpixel labeling outcomes (after C^2_{SP} in F-1 or C^3_{SP} in F-2) for each 2D axial slice, followed by 3D connected component analysis. By assuming the overall pancreas connectivity of its 3D shape, the largest 3D connected component is kept as the final segmentation. The binarization thresholds of random forest classifiers in C^2_{SP} and C^3_{SP} are calibrated using data in the training folds in 6-fold cross-validation, via a simple grid search. In [37], standalone *Patch*-ConvNet dense probability maps (without any post-processing) are processed for pancreas segmentation after using (F-1) as an initial cascade. The corresponding pancreas segmentation performance is not as accurate as (F-1) or (F-2). This finding is in analogy to [38], [39] where hand-crafted features need to be combined with deep image features to improve pulmonary nodule detection in chest CT scans or chest pathology detection using X-rays. Refer to Sec. IV-B and Sec. IV-C for detailed comparison and numerical results. Recent computer vision work also demonstrate the performance improvement when combining hand-crafted and deep image features for image segmentation [40] and video action recognition [41].

IV. Data and Experimental Results

A. Imaging Data

Our method is evaluated on a dataset of 80 3D abdominal portal-venous contrast enhanced CT scans ($\sim$ 70 seconds after intravenous contrast injection) acquired from 53 male and 27 female subjects. 17 of the subjects are healthy kidney donors who received abdominal CT scans prior to nephrectomy. The remaining 63 patients are randomly selected by a radiologist from the Picture Archiving and Communications System (PACS) from a population exhibiting neither major abdominal pathologies nor pancreatic cancer lesions. The CT datasets are obtained from the National Institutes of Health Clinical Center. Subject age ranges from 18 to 76 years, with a mean of 46.8 ± 16.7. Scan resolution is 512×512 pixels (varying pixel sizes) with slice thickness ranging from $1.5 - 2.5$ mm on Philips and Siemens MDCT scanners. The tube voltage is 120 kVp. Manual, ground-truth pancreas segmentation masks for all 80 cases have been provided by a medical student and have been revised by a radiologist.

B. Results on Pancreas Segmentation Accuracy

All experiments are conducted under six-fold cross-validation, as described in Sec. III-B and Sec. III-D. Four evaluation metrics are employed: Dice similarity coefficient, Jaccard index, volumetric recall, and volumetric precision. The Dice similarity index measures the overlap between two sample sets, $DSC = 2(|A \cap B|)/(|A| + |B|)$ where A and B refer to the algorithm output and manual ground-truth 3D pancreas segmentation, respectively. The Jaccard index (JI) is another statistic used to compute similarities between the segmentation result against the reference standard, $JI = (|A \cap B|)/(|A \cup B|)$.

Fig. 6 (d) and (h) show sample illustrative slices from different patients. Response maps are computed for the image patch-level classification and dense labeling. High probability values corresponding to the pancreas are represented by the red color regions (the background is blue). The response maps (denoted as P^{RF}) are highly informative. The most interesting observation is that the relative x and y positions, when included as features, allow for clearer spatial separation of positive and negative regions, via internal RF feature thresholding. The trained RF classifier is able to recognize negative class patches, such as liver, vertebrae and muscle using spatial location cues. Fig. 6(d,h) contains implicit vertical and horizontal decision boundary lines, unlike Fig. 6(c,g). This demonstrates the superior descriptive and discriminative power of the feature descriptor on image patches (P and $\acute{P}$), relative to single pixel intensities. Organs with similar CT values are suppressed significantly in the patch-level response maps.

Cascade ROC Analysis: Fig. 8 contains the ROC curves for 6-fold cross-validation of the two-tiered superpixel-level classifiers C^1_{SP} and C^2_{SP}, C^3_{SP} to assemble our frameworks F-1 and F-2, respectively. The red plots are constructed from the average sensitivity and specificity per superpixel. In the blue plots, each superpixel is weighted by its size (e.g., numbers of pixels and pixel sizes) prior to computing sensitivity and specificity. The Area Under the Curve (AUC) values of C^2_{SP} are noticeably lower than the C^1_{SP} AUC values (0.884 versus 0.997). This indicates that C^2_{SP} is much harder to train, since it uses "hard negatives" as negative samples that are classified positively by C^1_{SP}. Random Forest classifiers with $50 \sim 200$ trees are evaluated, with similar empirical performances. In C^3_{SP}, the dense patch-level image labeling (in the second level of cascade) is conducted by a Deep Convolutional Neural Network (i.e., $Patch$-ConvNet) to generate P^{CNN} (Refer to Fig. 7). Three examples of dense CNN based image patch labeling are demonstrated in Fig. 9. The AUC value of C^3_{SP} obtained by swapping the probability response maps from P^{RF} to P^{CNN} does improve to 0.931, compared to 0.884 using C^2_{SP} in the pixel-weighted volume metric. This demonstrates the performance benefit of using CNN for dense patch labeling (Sec. III-C) versus hand-crafted image features (Sec. III-B). See Fig. 8 (**Right**) and (**Middle**), respectively.

Quantitative Pancreas Segmentation Accuracy: Pancreas segmentation is evaluated with respect to the total number of patient scans used for training and testing. Using our F1 framework on 40, 60 and 80 (i.e., 50%, 75% and 100% of the total 80 datasets) patient scans, Dice, JI, Precision and Recall are computed under six-fold cross-validation. Table I provides results from the usage of image patch-level features and multi-level classification (i.e., performing C^1_{SP} and C^2_{SP} on I^{CT} and P^{RF}) and shows the effect of the patient population size on performance. Steady improvements of $\sim 4\%$ in the Dice coefficient and $\sim 5\%$ for the Jaccard index are observed, from 40 to 60, and 60 to 80 patients. Fig. 10 contains some sample final pancreas segmentation results from the 80 patient execution (i.e., Test 3 in Table I) for two different patients. The results are divided into three categories: good, fair and poor. The good category refers to Dice coefficients above 90% (of 15 patients), fair to $50\% \leq Dice \geq 90\%$ (49 patients) and poor to Dice $< 50\%$ (16 patients).

Six-fold CV is employed once more, to allow for direct comparison on segmentation accuracy using the same four metrics (i.e., Dice, JI, precision and recall) for framework F-1 versus F-2. Table I indicates that F-2 increases the Dice coefficient by 2%. However, the main improvement lies in the minimum values (i.e., the lower performance bound) for each of the metrics. Usage of deep patch labeling prevents the case of no pancreas segmentation while keeping slightly higher mean precision and recall values. Furthermore, F-2 decreases the standard deviations by $\sim 50\%$ (from 25.6% to 13.0% in Dice; and 25.4% to 13.6% in JI). Note that F-1's standard deviation ranges are similar to those from the previous methods [10], [7], [8], [9], [5] and F-2 significantly improves upon all methods. From Fig. 2 and Fig. 6, one can

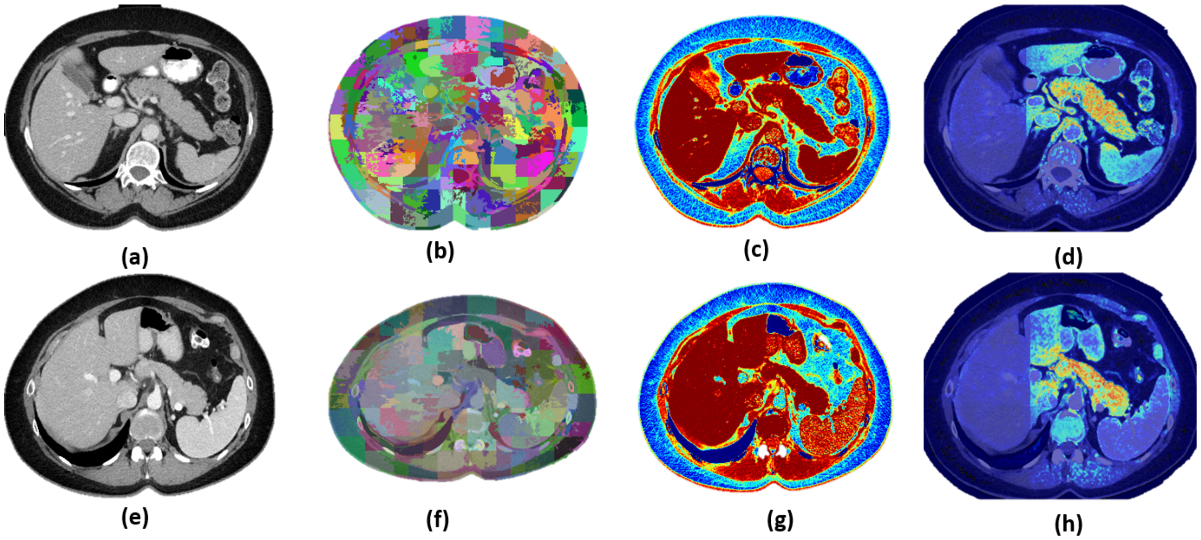

Fig. 6. Two sample slices from different patients are shown in (a) and (e). The corresponding superpixels segmentation (b,f), KDE probability response maps (c, g) and RF patch-level probability response maps (d, h) are shown. In (c,g) and (d,h), red represents regions with the highest probabilities. In (d,h), purple represents areas considered as insignificant, due to very small probabilities.

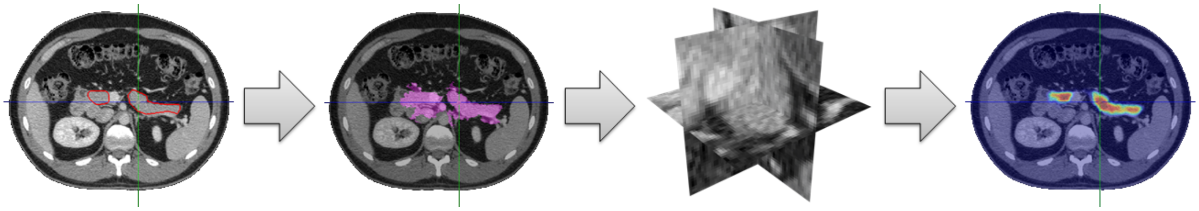

Fig. 7. Axial CT slice of a manual (gold standard) segmentation of the pancreas. From *Left* to *Right*, there are the ground-truth segmentation contours (in red); the RF based coarse segmentation $\{S_{\mathrm{RF}}\}$; a 2.5D input image patch to CNN and the CNN deep patch labeling CNN result.

infer that including the relative x-axis and y-axis positions as features reduced the overall false negative rates. F-2's consistent improvement over F-1 implies that CNN based dense patch labeling yields more promising results (Sec. III-C) than the conventional hand-crafted image features and random forest patch classification alone (Sec. III-B). Fig. 11 depicts an sample patient for which F-2's Dice score increases F-1's score by 18.6% (from 63.9% to 82.5%). In this particular case, the proximity of the stomach and duodenum to the pancreas head is especially challenging for F-1, and can only be detected by the CNN. The surface-to-surface overlays compare both frameworks to the ground truth manual segmentation.

F-1's performance is comparable to that of other state-of-the-art pancreas segmentation methods, while F-2's performance slightly but consistently surpasses others', even under six-fold cross-validation instead of the "leave-one-patient-out" (LOO) used in [10], [7], [8], [9], [5], [6]. Note that our results are not directly or strictly comparable with [10], [7], [8], [9], [5], [6], as different datasets are used for evaluation. Under six-fold CV, our bottom-up segmentation method can significantly outperform the "multi-atlas and label fusion" (MALF) from [11], [12], on the pancreas segmentation dataset studied in this paper. Details are provided later in this section. Table II compares Dice, JI, precision and recall results, between our methods of F-1, F-2 and other approaches, in multi-atlas registration and label fusion based multi-organ segmentation [10], [7], [8], [9], [6] and multi-phase single organ (i.e., pancreas) segmentation [5]. Previous numerical results derive from the publications [10], [7], [8], [9], [5], [6]. We choose the best result out of the different parameter configurations in [8]. Based on 80 CT datasets, our results are comparable and slightly better than those from the recent state-of-the art work [10], [7], [8], [9], [5]. For example, F-1 yields Dice coefficients of 68.8%±25.6% and F-2 70.7%±13.0% (6-fold CV), versus 69.6% ± 16.7% in [7], 65.5% in [9], 65.5% ± 18.6% in [10] and 69.1% ± 15.3% in [8] (LOO).

Standalone $Patch$-ConvNet dense probability maps can be smoothed and thresholded for pancreas segmentation as reported in [37] where Dice coefficients 60.9 ± 10.4% are achieved. When only 12 features are extracted from P^{CNN} maps for C^3_{SP}, the final pancreas segmentation accuracy drops to 64.5±12.3% in Dice scores, compared to F-1 (68.8±25.6%) and F-2 (70.7 ± 13.0%) in Table I. Similarly, recent work [38], [39] observe that deep CNN image features should be combined with hand-crafted features to improve performance in computer-aided detection tasks.

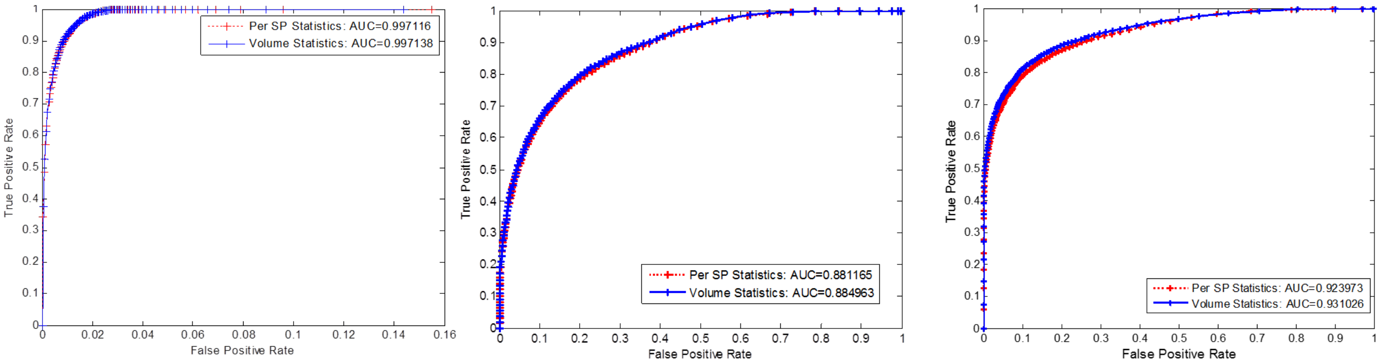

Fig. 8. ROC curves to analyze the superpixel classification results, in a two layer cascade of RF classifiers: (**Left**) the first layer classifier, C^1_{SP} and (**Middle**) the second layer classifier, C^2_{SP}; (**Right**) the alternative second layer classifier, C^3_{SP}. Red plots are constructed from the average sensitivity and specificity per superpixel. For the blue plots, superpixels are weighted by their size (e.g., numbers of pixels and pixel sizes) prior to computing sensitivity and specificity. All ROC curves shown are obtained by averaging the validation ROCs through 6-fold of cross-validation. No training ROC plots are demonstrated. Qualitative comparison of the training versus validation ROC curves at the each round of cross-validation indicates good generality of the random forest classifiers C^1_{SP}, C^2_{SP} and C^3_{SP}.

Method	N	DSC (%)	JI (%)	Precision (%)	Recall (%)
F-1	40	60.4 ± 22.3 [2.0, 96.4]	46.7 ± 22.8 [0, 93.0]	55.6 ± 29.8 [1.2, 100]	80.8 ± 21.2 [4.8, 99.8]
F-1	60	64.9 ± 22.6 [0, 94.2]	51.7 ± 22.6 [0, 89.1]	70.3 ± 29.0 [0, 100]	69.1 ± 25.7 [0, 98.9]
F-1	80	68.8 ± 25.6 [0, 96.6]	57.2 ± 25.4 [0, 93.5]	71.5 ± 30.0 [0, 100]	72.5 ± 27.2 [0, 100]
F-2	80	70.7 ± 13.0 [24.4, 85.3]	57.9 ± 13.6 [13.9, 74.4]	71.6 ± 10.5 [34.8, 85.8]	74.4 ± 15.1 [15.0, 90.9]

TABLE I
EVALUATION OF VARYING THE PATIENT POPULATION SIZE USING FRAMEWORK 1, WITH FOUR METRICS OF DICE, JI, PRECISION AND RECALL. MEAN, STANDARD DEVIATION, LOWER AND UPPER PERFORMANCE RANGES ARE REPORTED. FRAMEWORK 1 (F-1) IS COMPARED TO FRAMEWORK 2 (F-2) IN 80 PATIENTS.

C. Quantitative Comparison on Pancreas Segmentation Methods

We exploit two variants of bottom-up pancreas segmentation that propagate information from image patches to (segments) superpixels. Both frameworks are conducted under six-fold cross-validation. Six-fold CV is more difficult than the "leave-one-out" (LOO) criterion in [10], [7], [8], [9], [5], as less patients are used for training, and more separate patient scans are used for testing. In fact, under LOO, [7] experiences a notable decline in performance from using 149 patients to using 49 patients in training. In particular, the mean Dice coefficient decreases from $69.6\% \pm 16.7\%$ to $58.2\% \pm 20.0\%$. This indicates that under six-fold cross-validation, the multi-atlas fusion approaches [10], [6], [7], [8], [9], [5] may achieve lower segmentation accuracies than reported. At 40 patients, our Dice coefficient using framework 1 is 2.2% higher than the reported Dice coefficient from [7] using 50 patients (60.4% versus 58.2%). Compared to using $N-1$ patient datasets directly in the memory for multi-atlas registration methods, our learned models are more compactly encoded into a series of patch and superpixel-level random forest classifiers and into a CNN classifier for patch labeling. Computational efficiency also improves significantly, in the order of $6 \sim 8$ minutes per testing case (using a mix of Matlab and C implementation, $\sim 50\%$ time for superpixel generation), compared to others requiring 10 hours or more. The segmentation framework (F-2) using deep patch labeling confidences is also more numerically stable, with no complete failure case and with remarkably lower standard deviations.

Comparison to R-CNN and its variations [42], [37]: In previous work [42], we chose to employ the recently developed "Regional CNN" (R-CNN) for pancreas segmentation, due to its high performance on superpixel classification [43]. Our simple R-CNN implementation reported notably worse results (Dice coefficient $62.9\% \pm 16.1\%$) than our F-2 framework (Dice $70.7 \pm 13.0\%$), which may be due to the R-CNN's training scheme. In particular, R-CNN [43] is not an "end-to-end" trainable deep learning system. First, R-CNN uses pre-trained or fine-tuned CNNs to extract image features for superpixels. Afterwards, the computed deep image features are classified by support vector machine models.

Recent work [37] extends the region-based convolutional neural networks (R-CNN) designed for semantic image segmentation [43], [16] to pancreas segmentation. In [37], 1) we exploit multi-level deep convolutional networks, which sample a set of bounding boxes covering each image superpixel at increasing spatial scales [44]; 2) our best performing model is a stacked R^2-ConvNet, which operates in the joint space of CT intensities and $Patch$-ConvNet dense probability maps, similar to F-2. [37] yields a mean (standard deviation) Dice coefficient of $71.8 \pm 10.7\%$ in four-fold cross-validation (which is slightly better than $70.7 \pm 13.0\%$ from F-2 using the same dataset). However, [37] cannot be directly trained and tested on the raw CT scans as in this paper, due to the large non-pancreas to pancreas superpixel ratio in the dataset. Therefore, given an input abdomen CT, an initial set of superpixel regions is first generated or filtered by a coarse cascading process of operating the random forests based pancreas segmentation (similar to F-1), at low classification thresholds. Over 96% of the original 3D abdominal CT scan space is discarded for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2016.2624198, IEEE Transactions on Image Processing

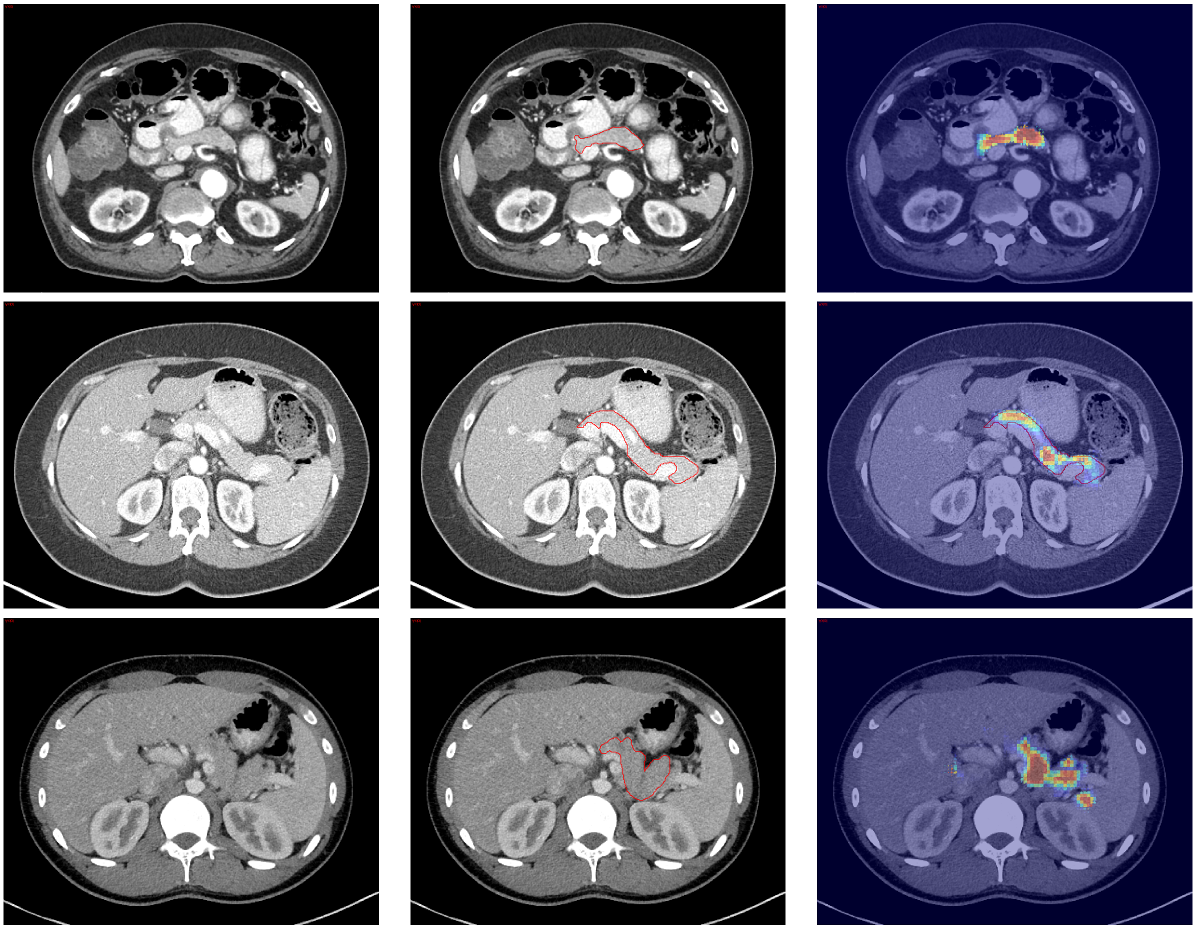

Fig. 9. Three examples of deep CNN based image patch labeling probability response maps per row. Red indicates stronger pancreas class response and blue presents weaker response. From Left, Center to Right are the original CT image, CT image with annotated pancreas contour in red, and CNN response map overlaid on the CT image.

the next step (see Fig. 8 **Left**). For pancreas segmentation, these pre-labeled superpixels serve as regional candidates with high sensitivity (>97%) but low precision (generally called Candidate Generation or CG process). The resulting initial DSC is 27% on average. Next, [37] evaluates several CNN variants for segmentation refinement (or pruning). F-2 performs comparably to the R-CNN extended for pancreas segmentation [37] and is able to run without using F-1 to generate pre-selected superpixel candidates (which nevertheless is required by [42], [37]).

Comparison to MALF (under six-fold CV): To easily compare our method to the previously well studied "multi-atlas and label fusion" (MALF) approaches, we implement a MALF solution for pancreas segmentation using the publicly available C++ code bases [11], [12]. We again use **six-fold cross-validation**, instead of "leave-one-patient-out" (LOO) as in [10], [7], [8], [9], [5], [6]. Specifically, each atlas in the training folds is registered to every target CT image in the testing fold, with the fast free-form deformation algorithm developed in NiftyReg [11]. Cubic B-Splines are used to deform a source image to optimize an objective function based on the normalized mutual information and a bending energy term. Grid spacing along the three axes is set to 5 mm. The weight of the bending energy term is 0.005 and the normalized mutual information is partitioned into 64 bins. The optimization is performed at three coarse-to-fine levels, with a maximum of 300 iterations per level. More details can be found in [11]. The registrations are used to warp the pancreas in the atlas set (66, or 67 atlases) to the target image. Nearest-neighbor interpolation is employed, since the labels are binary images. For each pixel in the target image, each atlas provides an opinion about the label. The probability of any pixel belonging to the pancreas x in the target U is given by $\hat{L}(x) = \sum_{i=1}^{n} \omega_i(x) L_i(x)$, where $L_i(x)$ is the warped i-th pancreas atlas, $\omega_i(x)$ is a weight assigned to the i-th atlas at location x with $\sum_{i=1}^{n} \omega_i(x) = 1$, and n is the number of atlases. In our six-fold cross-validation experiments, $n = 66$ or 67. We adopt the joint label fusion algorithm [12], which estimates voting weights $\omega_i(x)$ by simultaneously considering the pairwise atlas correlations and local image appearance similarities at x. More details about capturing the probability that different atlases produce the same label error at location x via a dependency matrix formulation are provided in [12]. The final binary pancreas segmentation label or map $L(x)$ in target can be computed by thresholding on $\hat{L}(x)$. The MALF segmentation yields a Dice coefficient of $52.51 \pm 20.84\%$, in the range of $[0\%, 80.56\%]$. This value is noticeably lower than the mean Dice score of $58.2\% \sim 69.6\%$ reported in [10],

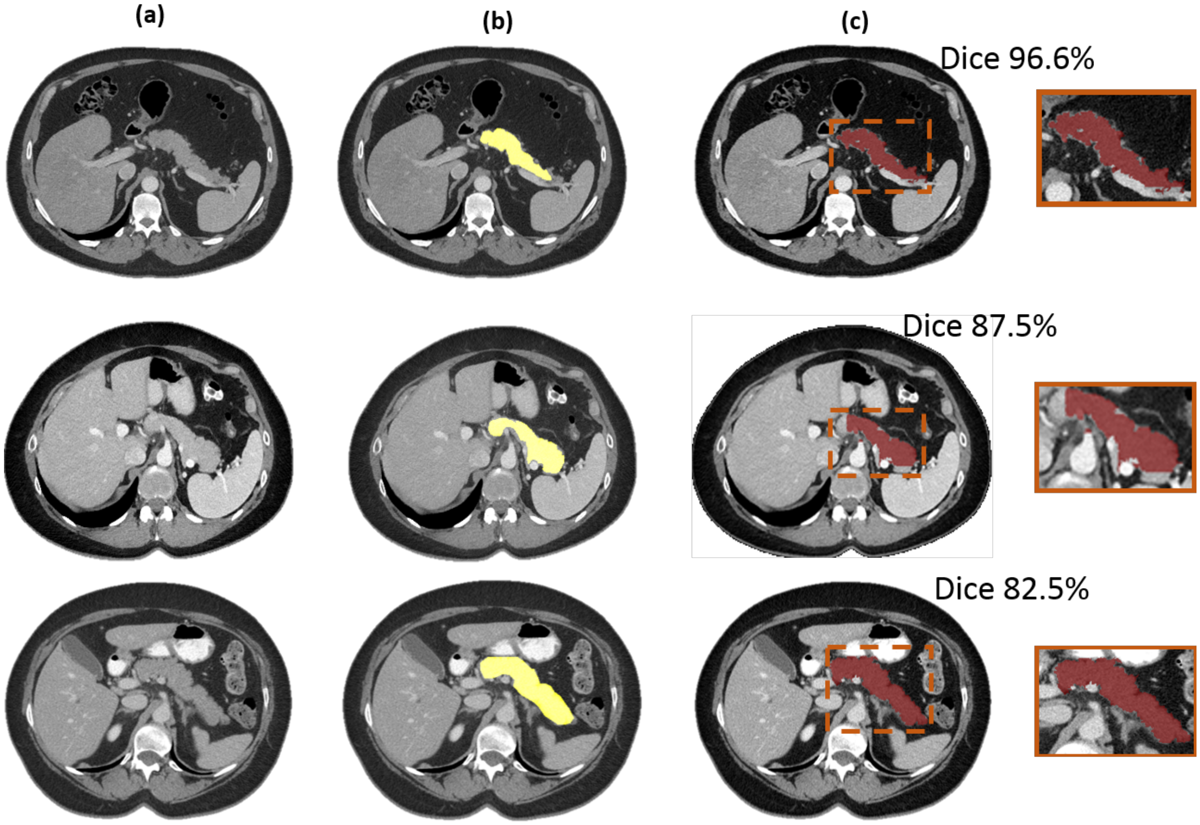

Fig. 10. Pancreas segmentation results with the computed Dice coefficients for one good (**Top Row**) and two fair (**Middle, Bottom Rows**) segmentation examples. Sample original CT slices for both patients are shown in (**Left Column**) and the corresponding ground truth manual segmentation in (**Middle Column**) are in yellow. Final computed segmentation regions are shown in red in (**Right Column**), with Dice coefficients for the volume above each slice. The zoomed-in areas of the slice segmentation in the orange boxes are shown to the right of the image.

[7], [8], [9], [5], [6] under the "leave-one-patient-out" (LOO) protocol for MALF methods. The performance deterioration of MALF from LOO (equivalent to 80-fold CV) to six-fold CV is consistent with the finding that the MALF segmentation accuracy drops from 69.6% to 58.2% when only 49 atlases are available instead of 149 [7].

Furthermore, $\sim$ 33.5 days are required to complete the MALF six-fold cross-validation experiments using a Windows server; whereas, the proposed bottom-up superpixel cascade approach requires $\sim$ 9 hours for 80 cases (6.7 minutes per patient scan on average). In summary, when using six-fold cross-validation on the same dataset, our bottom-up segmentation method significantly outperforms its MALF counterpart: $70.7 \pm 13.0\%$ versus $52.51 \pm 20.84\%$ in Dice coefficients, and is also $\sim$ 90 times faster. Converting our Matlab/C++ implementation into pure C++ should further reduce speed by a factor of $2 \sim 3$.

V. Conclusion

In this paper, we present a fully-automated bottom-up approach for pancreas segmentation in abdominal computed tomography (CT) scans. The proposed method generates a hierarchical cascade of information propagation by classifying image patches at different resolutions and pooling multi-channel feature information at (segments) superpixels. Our algorithm first decomposes CT slice images into a set of disjoint boundary-preserving superpixels; computes pancreas class probability maps via dense patch labeling; classifies superpixels by generating image features from both intensity and probability information to be fed into cascaded random forests; and finally enforces a simple spatial connectivity based post-processing. The dense image patch labeling is performed by an efficient random forest classifier (P^{RF}) on hand-crafted image histogram, location and texture features or by a deep convolutional neural network classification (P^{CNN}) using larger image windows (equivalent to having more spatial context). Hierarchical cascade classification may serve the core for other important medical imaging applications, including anatomical organ detection and parsing [45], and general computer-aided cancer detection and diagnosis [46], [47].

To overcome the low image boundary contrast issue in superpixel generation prevalent in medical imaging, efficient supervised edge learning techniques may be utilized to artificially "enhance" the strength of semantic object-level boundary curves in 2D or surfaces in 3D. For example, one may couple or integrate the structured random forests [48] or Holistically-nested neural network [49] based edge detection into a new image segmentation framework (MCG: Multiscale Combinatorial Grouping) [27] which permits a user-customized image gradient map. This new approach may generate superpixels that adequately preserve most boundaries, including very weak semantic object boundaries (in image

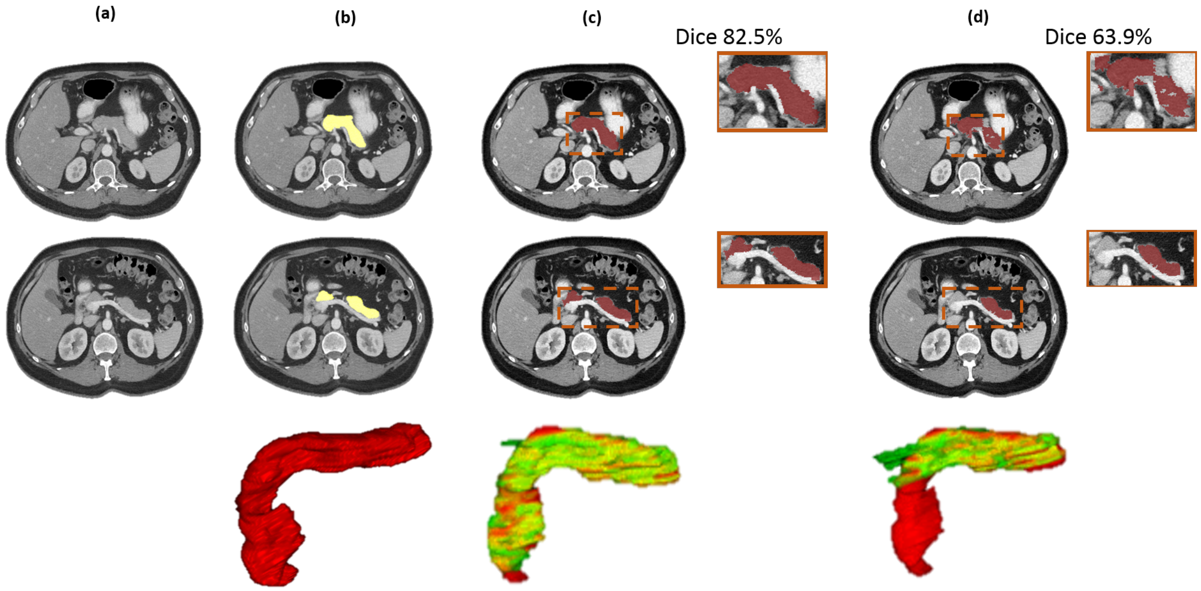

Fig. 11. Examples of pancreas segmentation results using F-1 and F-2, with the computed Dice coefficients for one patient. Original CT slices for the patient are shown in **Column (a)**, and the corresponding ground truth manual segmentations in **Column (b)** is yellow. Final computed segmentations using F-2 and F-1 are shown in red in **Columns (c,d)** with Dice coefficients for the volume above the first slice. The zoomed-in areas of the slice segmentation in the orange boxes are shown to the right of the images. Their surface-to-surface distance map overlaid on the ground truth mask is provided in **Columns (c,d) Bottom**, and the corresponding ground truth segmentation masks in **Column (b) Bottom** are in red. Red indicates higher differences, and green smaller distances.

Reference	N	DSC (%)	JI (%)	Precision (%)	Recall (%)	Protocol
[5]	20	-	57.9	-	-	LOO
[6]	28	-	46.6	-	-	LOO
[7]	150	69.6 ± 16.7	55.5 ± 17.1	67.9 ± 18.2	74.1 ± 17.1	LOO
[7]	50	58.2 ± 20.0	43.5 ± 17.8	-	-	LOO
[9]	100	65.5	49.6	70.7	62.9	LOO
[10]	100	65.5 ± 18.6	-	-	-	LOO
[8]	100	69.1 ± 15.3	54.6	-	-	LOO
[42]	82	68.0 ± 10.0	-	-	-	60/2/20
[37]	82	71.8 ± 10.7	-	-	-	CV-4
Framework 1	80	68.8 ± 25.6	57.2 ± 25.4	71.5 ± 30.0	72.5 ± 27.2	CV-6
Framework 2	80	70.7 ± 13.0	57.9 ± 13.6	71.6 ± 10.5	74.4 ± 15.1	CV-6
MALF	80	52.5 ± 20.8	38.1 ± 18.3	-	-	CV-6

TABLE II

COMPARISON OF F-1 AND F-2 UNDER SIX-FOLD CROSS-VALIDATION (CV-6) TO THE RECENT STATE-OF-THE-ART METHODS [10], [7], [8], [9], [5], [6] UNDER LOO, REGIONAL CNN METHODS [42], [37] AND OUR IMPLEMENTATION OF "MULTI-ATLAS AND LABEL FUSION" (MALF) USING PUBLICLY AVAILABLE C++ CODE BASES [11], [12] UNDER THE SAME SIX-FOLD CROSS-VALIDATION. [42] IS EVALUATED USING A HARD SPLIT OF (60/2/20) FOR THE TRAINING, VALIDATION AND TESTING DATASETS, RESPECTIVELY. FOUR-FOLD CROSS-VALIDATION (CV-4) IS ADOPTED IN [37]. THE PROPOSED BOTTOM-UP PANCREAS SEGMENTATION METHODS OF F-1 AND F-2 SIGNIFICANTLY OUTPERFORM THEIR MALF COUNTERPART: 68.8 ± 25.6% (F-1), 70.7 ± 13.0% (F-2) VERSUS 52.51 ± 20.84% IN DICE COEFFICIENTS (MEAN±STD).

gradients) and subsequently prevent segmentation leakage.

In future work, it may be necessary to further examine subconnectivity processes for the pancreas segmentation framework that considers the spatial relationships of splenic, portal and superior mesenteric veins with the pancreas.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center. We thank Andrew Dwyer, MD at National Institutes of Health Clinical Center for critical review of the manuscript.

REFERENCES

[1] R. Cuingnet, R. Prevost, D. Lesage, L. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3d ct images using random forests," in *MICCAI*, 2012, pp. 66–74.

[2] A. Mansoor, U. Bagci, Z. Xu, B. Foster, K. Olivier, J. Elinoff, A. Suffredini, J. Udupa, and D. Mollura, "A generic approach to pathological lung segmentation," *IEEE Trans. on Medical Imaging*, vol. 33, no. 12, pp. 2293–2310, 2014.

[3] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3d cardiac ct volumes using marginal space learning and steerable features," *IEEE Trans. on Medical Imaging*, vol. 27, no. 11, pp. 1668–1681, 2008.

[4] H. Ling, S. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *IEEE Conf. CVPR*, 2008, pp. 1–8.

[5] A. Shimizu, T. Kimoto, H. Kobatake, S. Nawano, and K. Shinozaki, "Automated pancreas segmentation from three-dimensional contrast-enhanced computed tomography," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 5, no. 1, pp. 85–98, 2010.

[6] T. Okada, M. Linguraru, Y. Yoshida, M. Hor, R. Summers, Y. Chen, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation of ct images based on hierarchical spatial modeling of organ interrelations," in *Abdominal Imaging - Computational and Clinical Applications*, 2012, pp. 173–180.

[7] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. on Medical Imaging*, vol. 32, no. 7, pp. 1723–1730, 2013.

[8] C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, and K. Mori, "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3d abdominal ct images," in *MICCAI*, vol. 2, 2013, pp. 165–172.

[9] R. Wolz, C. Chu, K. Misawa, K. Mori, and D. Rueckert, "Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases," in *MICCAI*, vol. 1, 2012, pp. 10–17.

[10] Z. Wang, K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert, "Geodesic patch-based segmentation," in *MICCAI*, vol. 1, 2014, pp. 666–673.

[11] M. Modat, J. McClelland, and S. Ourselin, "Lung registration using the niftyreg package," in *Medical Image Analysis for the Clinic-A Grand Challenge*, 2010, pp. 33–42.

[12] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2012.

[13] V. Zografos, B. Menze, and F. Tombari, "Hierarchical multi-organ segmentation without registration in 3d abdominal ct images," in *MICCAI Medical Computer Vision Workshop*, 2015.

[14] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua, "Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features," *IEEE Trans. on Medical Imaging*, vol. 31, no. 2, pp. 474–486, 2012.

[15] D. Mahapatra, P. Schuffler, J. Tielbeek, J. Makanyanga, J. Stoker, S. Taylor, F. Vos, and J. Buhmann, "Automatic detection and segmentation of crohn's disease tissues from abdominal mri," *IEEE Trans. on Medical Imaging*, vol. 32, no. 12, pp. 2332–2348, 2013.

[16] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[17] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 329–349, 2013.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. on Pat. Ana. Mach. Intel.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Proc. Forum Bildverarbeitung*, 2012, pp. 1–12.

[20] M. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *IEEE Conf. CVPR*, 2011, pp. 2099–2104.

[21] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[22] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. on Pat. Ana. Mach. Intel.*, vol. 13, no. 6, pp. 583–598, 1991.

[23] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pat. Ana. Mach. Intel.*, vol. 24, no. 5, pp. 603–619, 2002.

[24] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," *IEEE Proc. on CVPR*, 2005.

[25] Y. Shi and W. Karl, "A real-time algorithm for the approximation of level-set-based curve evolution," *IEEE Trans. Image Processing*, vol. 17, no. 5, pp. 645–656, 2008.

[26] P. Yadollahpour, D. Batra, and G. Shakhnarovich, "Discriminative re-ranking of diverse segmentations," in *IEEE Conf. CVPR*, 2013, pp. 2099–2104.

[27] P. Arbelez, J. Pont-Tuset, J. Barron, F. Marqus, and J. Malik, "Multiscale combinatorial grouping," in *IEEE Conf. CVPR*, 2014, pp. 328–335.

[28] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[29] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Medical Image Analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

[31] A. Gilinsky and L. Zelnik-Manor, "Siftpack: a compact representation for efficient sift matching," in *IEEE Conf. ICCV*, 2013, pp. 777–784.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[34] H. Roth, L. Lu, A. Seff, K. Cherry, S. Wang, J. Liu, E. Turkbey, and R. Summers, "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," in *MICCAI*, 2014, pp. 520–527.

[35] R. Groeneveld and G. Meeden, "Measuring skewness and kurtosis," *The Statistician*, vol. 33, pp. 391–399, 1984.

[36] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[37] H. Roth, L. Lu, A. Farag, H. Shin, J. Liu, E. Turkbey, and R. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI*, 2015, pp. 556–564.

[38] B. van Ginneken, A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 286–289.

[39] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 294–297.

[40] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *IEEE Conf. on CVPR Workshop*, 2015.

[41] Z. Lan, S. Yu, M. Lin, B. Raj, and A. Hauptmann, "Handcrafted local features are convolutional neural networks," in *arXiv:1511.05045*, 2015.

[42] H. Roth, A. Farag, L. Lu, E. Turkbey, J. Liu, and R. Summers, "Deep convolutional networks for pancreas segmentation in ct imaging," in *SPIE Conf. Medical Imaging*, 2015, pp. 1–8.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and semantic segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence, to appear*, 2015.

[44] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *IEEE Conf. on CVPR*, 2015, pp. 3376–3385.

[45] L. Lu, A. Barbu, M. Wolf, J. Liang, L. Bogoni, M. Salganicoff, and D. Comaniciu, "Simultaneous detection and registration for ileo-cecal valve detection in 3d ct colonography," in *European Conference on Computer Vision*, 2008, pp. 465–478.

[46] L. Lu, J. Bi, M. Wolf, and M. Salganicoff, "Effective 3d object detection and regression using probabilistic segmentation features in ct images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1049–1056.

[47] L. Lu, P. Devarakota, S. Vikal, D. Wu, Y. Zheng, and M. Wolf, "Computer aided diagnosis using multilevel image features on large-scale evaluation," in *Springer, Medical Computer Vision*, 2013.

[48] P. Dollr and L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. on Pat. Ana. Mach. Intel.*, vol. 37, 2015.

[49] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015, pp. 1395–1403.

Amal Farag received her BSc, MEng and PhD in Electrical and Computer Engineering from the University of Louisville. Her research domain is multimodality image analysis using variational and statistical methods. After her PhD, Dr. Farag received a two-year postdoctoral Imaging Sciences Training Program (ISTP) fellowship from the National Institutes of Health, at the Imaging Biomarkers and CAD Lab, Clinical Center with Dr. M. Summers in the Radiology and Imaging Sciences Department. Research focused on image analysis/understanding to segment occluded and highly variable organ using image parsing and machine learning. In the past two years, Dr. Farag joined Kentucky Imaging Technologies (KIT) as a biomedical imaging scientist working on computer-assisted diagnosis of colon and lung cancers. Dr. Farag is author of 35 peer-reviewed manuscripts, including five journals. Her theoretical contribution is on deformable object modeling and recognition. Her practical domain spans computer-assisted interventions and immersive visualization.

Le Lu is a staff scientist in Department of Radiology and Imaging Sciences, National Institutes of Health (NIH) Clinical Center (CC), Bethesda, Maryland since 2013. His research is focused on medical image understanding and semantic parsing to fit into new clinical practices, especially in the areas of preventive early cancer detection/diagnosis and developing precise novel imaging bio-markers, via large scale imaging protocols and statistical (deep) learning principles. He worked on various core R&D problems in colonic polyp and lung nodule CADx systems, and vessel, bone imaging at Siemens Corporate Research and Siemens Healthcare from Oct. 2006 until Jan. 2013, and his last post was a senior staff scientist. He is the (co-)inventor of 16 US/International patents, 30 inventions and has authored over 80 peer-reviewed papers. He received his Ph.D. in Computer Science from Johns Hopkins University in May 2007. He won the Mentor of the Year award (staff scientist/staff clinician category) at NIH in 2015 and the best summer intern mentor award from NIH-CC in 2013. He served as a program committee member for MICCAI 2015, 2016 and serves as an area chair for IEEE CVPR 2017. He has mentored 22 interns, students or postdoctoral fellow trainees at Siemens and NIH (who have won several internal and international research awards and fellowships).

Holger R. Roth is an assistant professor (research) at Mori Laboratory, Nagoya University, Japan, since August 2016. From 2013 to 2016, he worked at the Imaging Biomarkers and CAD Lab at the National Institutes of Health Clinical Center, Bethesda, USA, as a postdoctoral fellow with Dr. Ronald M. Summers and Dr. Le Lu. Since then, his research efforts are focusing on the application of deep learning methodologies for medical image analysis, multi-scale computational anatomy, and computer-aided detection and diagnosis. From 2009 until 2013, he was a PhD student at the Centre for Medical Image Computing at University College London, UK, under the supervision of Prof. David Hawkes and Dr. Jamie McClelland. His PhD thesis was specialised in the registration of highly deformable organs such as the colon and linking it to colonoscopy videos. Until Oct. 2016, he has published 38 peer-reviewed major conference papers and journal articles in MICCAI, ISBI, IEEE Trans. Medical Imaging, IEEE Trans. on Image Processing, Medical Physics and Journal of Radiology, etc. His total citation number reaches 302 by Google Scholar.

Jiamin Liu received her PhD in Bioengineering from the University of Pennsylvania in 2006, where she developed a patented automated technique for optimal boundary detection in medical images. Dr. Liu holds an MS and BS in Electronic Engineering from Beijing Normal University in China. She is currently a Staff Scientist in the Imaging Biomarkers and Computer-Aided Diagnosis (CAD) Laboratory, Radiology and Imaging Sciences, at the National Institutes of Health. Her research interests include medical image analysis in general with a focus on segmentation, registration and computer-aided detection. She is an associate editor of Medical Physics and reviewer for IEEE TMI, IEEE TBME, Medical Physics, Pattern Recognition Letters, and MICCAI.

Evrim Turkbey is a PET/CT fellow in Department of Radiology and Radiological Sciences, Johns Hopkins University, Baltimore, Maryland. Her research focused on image quantification in cardiovascular diseases and oncology to better characterize sub-clinical findings of cardiovascular diseases and early detection of cancer. She completed her diagnostic radiology residency at the Department of Radiology, Hacettepe University Scholl of Medicine in 2007. She has worked as a post doctoral fellow at MRI division of Johns Hopkins University and at the department of Radiology and Imaging Sciences, National Institutes of Health between 2007 and 2014. She completed her nuclear medicine residency at the Johns Hopkins University in 2016. She is a researcher of MESA (Multi-ethnic Study of Atherosclerosis) and EDIC (the Epidemiology of Diabetes Interventions and Complications) studies. She has authored more than 40 peer reviewed papers. Dr. Turkbey was awarded the Trainee Research Prize in 2011 RSNA Scientific Assembly and Annual Meeting.

Ronald M. Summers received the B.A. degree in physics and the M.D. and Ph.D. degrees in Medicine/Anatomy & Cell Biology from the University of Pennsylvania. He completed a medical internship at the Presbyterian-University of Pennsylvania Hospital, Philadelphia, PA, a radiology residency at the University of Michigan, Ann Arbor, MI, and an MRI fellowship at Duke University, Durham, NC. In 1994, he joined the Diagnostic Radiology Department at the NIH Clinical Center in Bethesda, MD where he is now a tenured Senior Investigator and Staff Radiologist. He is currently Chief of the Clinical Image Processing Service and directs the Imaging Biomarkers and Computer-Aided Diagnosis (CAD) Laboratory. In 2000, he received the Presidential Early Career Award for Scientists and Engineers, presented by Dr. Neal Lane, President Clintons science advisor. In 2012, he received the NIH Directors Award, presented by NIH Director Dr. Francis Collins. His research interests include deep learning, virtual colonoscopy, CAD and development of large radiologic image databases. His clinical areas of specialty are thoracic and abdominal radiology and body cross-sectional imaging. He is a member of the editorial boards of the journals Radiology, Journal of Medical Imaging and Academic Radiology. He is a program committee member of the Computer-aided Diagnosis section of the annual SPIE Medical Imaging conference. He has co-authored over 400 journal, review and conference proceedings articles and is a co-inventor on 14 patents. An avid mentor, he has supervised over 80 trainees.