Medical Image Analysis 45 (2018) 94–107

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation☆

Holger R. Roth*, Le Lu*, Nathan Lay, Adam P. Harrison, Amal Farag, Andrew Sohn, Ronald M. Summers*

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Clinical Image Processing Service, Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA

ARTICLE INFO

Article history:
Received 23 August 2017
Revised 28 November 2017
Accepted 30 January 2018
Available online 1 February 2018

Keywords:
Fully convolutional networks
Holistically nested networks
Deep learning
Medical imaging
Computed tomography
Pancreas segmentation

ABSTRACT

Accurate and automatic organ segmentation from 3D radiological scans is an important yet challenging problem for medical image analysis. Specifically, as a small, soft, and flexible abdominal organ, the pancreas demonstrates very high inter-patient anatomical variability in both its shape and volume. This inhibits traditional automated segmentation methods from achieving high accuracies, especially compared to the performance obtained for other organs, such as the liver, heart or kidneys. To fill this gap, we present an automated system from 3D computed tomography (CT) volumes that is based on a two-stage cascaded approach—pancreas localization and pancreas segmentation. For the first step, we localize the pancreas from the entire 3D CT scan, providing a reliable bounding box for the more refined segmentation step. We introduce a fully deep-learning approach, based on an efficient application of holistically-nested convolutional networks (HNNs) on the three orthogonal axial, sagittal, and coronal views. The resulting HNN per-pixel probability maps are then fused using pooling to reliably produce a 3D bounding box of the pancreas that maximizes the recall. We show that our introduced localizer compares favorably to both a conventional non-deep-learning method and a recent hybrid approach based on spatial aggregation of superpixels using random forest classification. The second, segmentation, phase operates within the computed bounding box and integrates semantic mid-level cues of deeply-learned organ *interior* and *boundary* maps, obtained by two additional and separate realizations of HNNs. By integrating these two mid-level cues, our method is capable of generating boundary-preserving pixel-wise class label maps that result in the final pancreas segmentation. Quantitative evaluation is performed on a publicly available dataset of 82 patient CT scans using 4-fold cross-validation (CV). We achieve a (mean ± std. dev.) Dice similarity coefficient (DSC) of 81.27 ± 6.27% in validation, which significantly outperforms both a previous state-of-the art method and a preliminary version of this work that report DSCs of 71.80 ± 10.70% and 78.01 ± 8.20%, respectively, using the same dataset.

1. Introduction

Pancreas segmentation in computed tomography (CT) challenges current computer-aided diagnosis (CAD) systems. While automatic segmentation of numerous other organs in CT scans, such as the liver, heart or kidneys, achieves good performance with Dice similarity coefficients (DSCs) of $>90\%$ (Wang et al., 2014c; Chu et al., 2013; Wolz et al., 2013), the pancreas' variable shape, size, and location in the abdomen limits segmentation accuracy to $<73\%$ DSC being reported in the literature (Wolz et al., 2013; Chu et al., 2013; Tong et al., 2015; Okada et al., 2015; Farag et al., 2014; Roth et al., 2015). Examples of pancreas as seen in CT are shown in Fig. 1. Previous pancreas segmentation work (Wolz et al., 2013; Chu et al., 2013; Tong et al., 2015; Okada et al., 2015) are all based on performing volumetric multiple atlas registration (Modat et al., 2010; Avants et al., 2009, 2011) and executing robust label fusion methods (Wang et al., 2013; Bai et al., 2013; Wang et al., 2014a) to optimize the per-pixel organ labeling process. This type of organ segmentation strategy is widely used for many organ segmentation problems, such as the brain (Wang et al., 2013; 2014a), heart (Bai et al., 2013), lung (Murphy et al., 2011), and pancreas (Wolz et al., 2013; Chu et al., 2013; Tong et al., 2015; Okada et al., 2015). These methods can be referred as a *top-down* model fitting approach, or more specifically, MALF (Multi-Atlas

☆ The content of this article is covered by US Provisional Patent Applications of 62/345,606# and 62/450,681#.

* Corresponding authors.

E-mail addresses: rothhr@mori.m.is.nagoya-u.ac.jp, h.roth@ucl.ac.uk (H.R. Roth), le.lu@nih.gov (L. Lu), rms@nih.gov (R.M. Summers).

https://doi.org/10.1016/j.media.2018.01.006
1361-8415/© 2018 Published by Elsevier B.V.

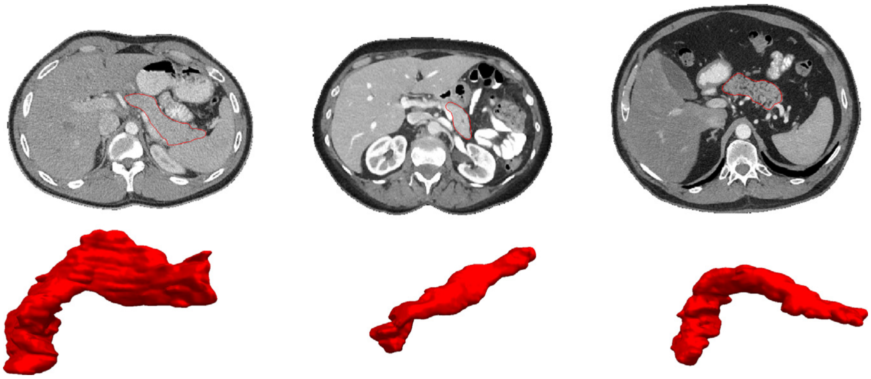


Fig. 1. Examples of variations in appearance, shape, and size of the pancreas as seen in contrast enhanced CT after removal of the image background by masking the patient's body. Manual ground truth annotations are shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Registration & Label Fusion). Another group of top-down frameworks (Ecabert et al., 2008; Zheng et al., 2008; Ling et al., 2008) leverages statistical model detection, e.g., generalized Hough transform (Ecabert et al., 2008) or marginal space learning (Zheng et al., 2008; Ling et al., 2008), for organ localization; and deformable statistical shape models for object segmentation (Cootes et al., 1994). However, due to the intrinsic huge 3D shape variability of the pancreas, statistical shape modeling has not been applied for pancreas segmentation.

Recently, a new *bottom-up* pancreas segmentation representation has been proposed in Farag et al. (2014), which uses dense binary image patch labeling confidence maps that are aggregated to classify image regions, or superpixels (Felzenszwalb and Huttenlocher, 2004; Pont-Tuset et al., 2017; Girshick et al., 2016), into pancreas and non-pancreas label assignments. This method's motivation is to improve segmentation accuracy of highly deformable organs, such as the pancreas, by leveraging mid-level visual representations of image segments. The segmentation performance was further advanced by our prior work Roth et al. (2015), which proposed a probabilistic bottom-up approach using a set of multi-scale and multi-level deep convolutional neural networks (CNNs) applied in a sliding window fashion on local image patches in order to capture the complexity of pancreas appearance in CT images. The resulting system improved upon the performance of Farag et al. (2014) with a reported DSC of $71.8 \pm 10.7\%$ against $68.8 \pm 25.6\%$. Compared to the MALF based pancreas segmentation work (Wolz et al., 2013),Chu et al. (2013),Tong et al. (2015) and Okada et al. (2015) that are evaluated using "leave-one-patient-out" (LOO) protocol, the bottom-up approaches using superpixel representation (Farag et al., 2014; Roth et al., 2015) have reported comparable or higher DSC accuracy measurements, under more challenging 6-fold or 4-fold cross-validation[1] Comparing the two bottom-up approaches, the usage of deep CNN models has noticeably improved the performance stability, which is evident by the significantly smaller standard deviation (Roth et al., 2015) than all other top-down or bottom-up works (Farag et al., 2014; Wolz et al., 2013; Chu et al., 2013; Tong et al., 2015; Okada et al., 2015).

Deep CNNs have successfully been applied to many high-level tasks in medical imaging, such as recognition and object detection (Yan et al., 2015). The main advantage of CNNs comes from the fact that end-to-end learning of salient feature representations for the task at hand is more effective than hand-crafted features with heuristically tuned parameters (Zheng et al., 2015). Similarly, CNNs demonstrate promising performance for pixel-level labeling problems, e.g., semantic segmentation in recent computer vision and medical imaging analysis work, e.g., fully convolutional neural networks (FCN) (Long et al., 2015), DeepLab (Chen et al., 2014), SegNet (Badrinarayanan et al., 2017) and U-Net (Ronneberger et al., 2015).

Another important type of networks are "holistically-nested networks" (HNN) (Xie and Tu, 2015) which combine FCNs with deep supervision (Lee et al., 2015) in order to get enhanced performance at different scales. Note that the HNN architecture was first proposed under the name "holistically-nested edge detection" (HED) as a deep learning based general image edge detection method, but has also been shown to work effectively for other semantic segmentation tasks (Roth et al., 2016a; Harrison et al., 2017b)

These approaches have all garnered significant improvements in performance over previous methods by applying state-of-the-art CNN-based image classifiers and representation to the semantic segmentation problem in both domains.

Semantic organ segmentation involves assigning a label to each pixel in the image. On one hand, features for classification of single pixels (or patches) play a major role, but on the other hand, factors such as edges, i.e., organ boundaries, appearance consistency, and spatial consistency, could greatly impact the overall system performance (Zheng et al., 2015). Furthermore, there are indications of semantic vision tasks requiring hierarchical levels of visual perception and abstraction (Xie and Tu, 2015). As such, generating rich feature hierarchies for both the interior and the boundary of the organ could provide important "mid-level visual cues" for semantic segmentation. Subsequent spatial aggregation of these mid-level cues then has the prospect of improving semantic segmentation methods by enhancing the accuracy and consistency of pixel-level labeling.

A preliminary version of this work appears as Roth et al. (2016a), where we demonstrate that a two-stage bottom-up localization and segmentation approach can improve upon the state of the art. In this work, the major extension is that we describe an improved pancreas localization method by replacing the initial super-pixel based one, with a new general deep learning based approach. This methodological component is designed to optimize or maximize the pancreas spatial recall criterion while reducing the non-pancreas volume as much as possible. Specifically, we generate the per-pixel pancreas class probability maps (or "heat maps") through an efficient combination of holistically-nested convolutional networks (HNNs) in the three orthogonal axial, sagittal, and coronal CT views. We fuse the three HNN outputs to produce a 3D bounding box covering the underlying, yet latent in testing, pancreas volume by nearly 100%. In addition, we

[1] As discussed in Shin et al. (2016), LOO can be considered as an extreme case of M-fold cross-validation with $M = N$ when N patient datasets are available for experiments. When M is decreasing and significantly smaller than N, M-fold cross-validation (CV) becomes more challenging since there are less data for training and more patient cases on testing.

show that exactly the same HNN model architecture can be effective for the subsequent pancreas segmentation stage by integrating both deeply learned boundary and appearance cues. This also results in a simpler overall pancreas localization and segmentation system using HNNs only, rather than the previous hybrid setup involving non-deep- and deep-learning method components (Roth et al., 2016a). Lastly, our current method reports an overall improved DSC performance compared to the preliminary work Roth et al. (2016a) and the sliding-window approach in Roth et al. (2015): DSC of 81.14 ± 7.3% versus 78.0 ± 8.2% and 71.8 ± 10.7% (Roth et al., 2015), respectively. In summary, our main novelties and contributions are proposing and validating new problem presentations on integrating deeply-learned organ interior and boundary cues (extended from Roth et al., 2015); a new robust and efficient multi-view pancreas segmentation fusion; and importantly, a generic coarse-to-fine localization and segmentation framework for organ segmentation (improved upon Roth et al., 2016a).

The proposed two-stage process essentially performs *3D spatial aggregation and assembling* on the HNN-produced per-pixel pancreas probability maps that run on 2D axial, coronal, and sagittal CT planes. This process operates exhaustively for pancreas localization and selectively for pancreas segmentation. Therefore, this work inherits a hierarchical and compositional visual representation of computing 3D object information aggregated from 2D image slices or parts, in a similar spirit of Roth et al. (2014), Farabet et al. (2013) and Lu et al. (2008). Alternatively, there are recent studies on directly using 3D convolutional neural networks for liver, brain segmentation (Dou et al., 2016; Chen et al., 2016a) and volumetric vascular boundary detection (Merkow et al., 2016). Due to CNN memory restrictions, these 3D CNN approaches adopt padded sliding windows or volumes to process the original CT scans, such as 96 × 96 × 48 segments (Merkow et al., 2016), 160 × 160 × 72 subvolumes (Dou et al., 2016) and 80 × 80 × 80 windows (Chen et al., 2016a), which may cause segmentation discontinuities or inconsistencies at overlapped window boundaries. We argue that learning shareable lower-dimensional 2D CNN models may be more generalizable and handle the "curse-of-dimensionality" issue better than their fully 3D counterparts, especially when used to parse complex 3D anatomical structures, e.g., lymph node clusters (Nogues et al., 2016; Roth et al., 2016b) and the pancreas (Roth et al., 2015; 2016a). Analogous examples of comparing compositional multi-view 2D CNNs versus direct 3D deep models can be found in other computer vision problems: 1) video based action recognition where a two-stream 2D CNN model (Simonyan and Zisserman, 2014a), capturing the image intensity and motion cues, significantly improves upon the 3D CNN method (Karpathy et al., 2014); 2) the advantageous performance of multi-view CNNs over volumetric CNNs in 3D Shape Recognition (Su et al., 2015). The rest of this paper is organized as follows. We describe the technical motivation and details of the proposed approach in Section 2. Experimental results and comparison with related work are addressed in Section 3. We conclude the paper, and with extended discussion, in Section 4.

2. Methods

In this work, we present a two-phased approach for automated pancreas **localization** and **segmentation**. The pancreas localization step aims to robustly compute a bounding box which, at the desirable setting, should cover the entire pancreas while pruning the high majority volumetric space from any input CT scan without any manual pre-processing. The second stage of pancreas segmentation incorporates deeply learned organ interior and boundary mid-level cues with subsequent spatial aggregation, focusing only on the properly zoomed or cascaded pancreas location and spatial extents that are generated after the first phase. In Section 2.1 we introduce the HNN model that proves effective for both stages. Afterwards, we focus on localization in Section 2.2, which discusses and contrasts a conventional approach to localization with newer CNN-based ones—a hybrid and a fully deep-learning approach. We show how the latter approach, which relies on HNNs, provides a simple, yet state-of-the-art, localization method. Importantly, it relies on the same HNN architecture as the later segmentation step. With localization discussed, we explain our segmentation approach in Section 2.3, which relies on combining semantic mid-level cues produced from HNNs. Our approach to organ segmentation is based on simple, reproducible, yet effective, machine-learning principles. In particular, we demonstrate the most effective configuration of our system is simply composed of cascading and aggregating outputs from six HNNs trained at three orthogonal views and two spatial scales. No multi-atlas registration or multi-label fusion techniques are employed. Fig. 2 provides a flowchart depicting the makeup of our system.

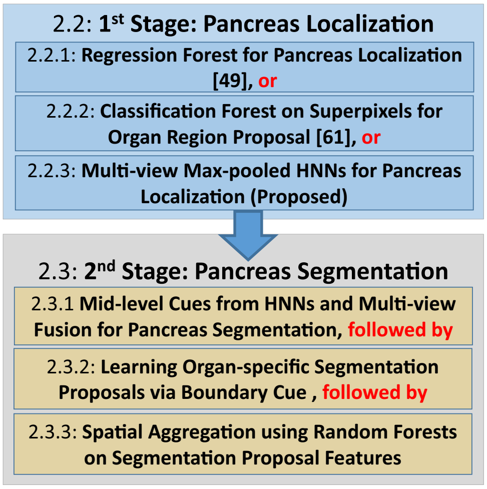

2.2: 1st Stage: Pancreas Localization
2.2.1: Regression Forest for Pancreas Localization [49], or
2.2.2: Classification Forest on Superpixels for Organ Region Proposal [61], or
2.2.3: Multi-view Max-pooled HNNs for Pancreas Localization (Proposed)
2.3: 2nd Stage: Pancreas Segmentation
2.3.1 Mid-level Cues from HNNs and Multi-view Fusion for Pancreas Segmentation, followed by
2.3.2: Learning Organ-specific Segmentation Proposals via Boundary Cue , followed by
2.3.3: Spatial Aggregation using Random Forests on Segmentation Proposal Features

Fig. 2. Flowchart of the proposed two-stage pancreas localization and segmentation framework with the corresponding section numbers where each method is described. Section 2.2.1 and Section 2.2.2 describe alternative means of bottom-up organ localization and are compared to the proposed method (Section 2.2.3). The remaining modules are part of our proposed pancreas segmentation approach (Section 2.3).

2.1. Learning mid-level cues via holistically-nested networks for localization and segmentation

In this work, we use the HNN architecture, to learn the pancreas' interior and boundary image-labeling maps, for both localization and segmentation. Object-level interior and boundary information are referred to as mid-level visual cues. HNN has been used successfully for extracting "edge-like" structures like blood vessels in 2D retina images (Fu et al., 2016). We however would argue and validate that it can serve as a suitable deep representation to learn general raw pixel-in and label-out mapping functions, e.g., to perform binary or even multi-labeled semantic image segmentation tasks. HED is developed from FCN (Long et al., 2015) and enhanced using the multi-scale deep supervision at each convolutional layers (Lee et al., 2015). The loss layer of HED is formulated as the same of FCN, performing the per-pixel classification cross-entropy loss. This cross-entropy loss can be flexibly defined as discerning samples of different classes: boundary versus non-boundary (Xie and Tu, 2015; Fu et al., 2016), object (i.e., interior mask) or non-object regions (Long et al., 2015; Hou et al., 2016; Harrison et al., 2017a). Specifically, similar HED-based CNN architectures are used for successfully detecting the saliency map of objects in images

(Hou et al., 2016). Harrison et al. (2017a) proposes and validates an improved version of HED for direct pathological lung segmentation in CT images, via progressively adding more gradient flows for better network regularization and performance. We use these principles to segment the interior of organs. Last, the boundary loss and segmentation loss can be generally formulated together with a multi-task neural network, such as Chen et al. (2016b). In this paper, we exploit to integrate deeply learned boundary and segmentation cues via explicit optimization of organ-specific object proposals. Another possible approach is integrating the boundary information channel within the structured prediction image segmentation framework (e.g., conditional random field (Fu et al., 2016) or boundary neural field (Nogues et al., 2016)).

HNN is designed to address two important issues: (1) training and prediction on the whole image end-to-end, i.e, holistically, using a per-pixel labeling cost; and (2) incorporating multi-scale and multi-level learning of deep image features (Xie and Tu, 2015) via auxiliary cost functions at each convolutional layer. HNN computes the image-to-image or pixel-to-pixel prediction maps from any input raw image to its annotated labeling map, building on fully convolutional neural networks (Long et al., 2015) and deeply-supervised nets (Lee et al., 2014). The per-pixel labeling cost function (Long et al., 2015; Xie and Tu, 2015) makes it feasible that HNN/FCN can be effectively trained using only several hundred annotated image pairs as opposed to using thousands of small image patches to train a per-image cost function as commonly done in sliding window based CNN approaches (Roth et al., 2015). This enables the automatic learning of rich hierarchical feature representations and contexts that are critical to resolve spatial ambiguity in the segmentation of organs. The network structure is initialized based on an ImageNet pre-trained VGGNet model (Simonyan and Zisserman, 2014b). It has been shown that fine-tuning CNNs pre-trained on general image classification tasks is helpful for low-level tasks, e.g., edge detection (Xie and Tu, 2015). Furthermore, we can utilize pre-trained edge-detection networks (trained on BSDS500 (Xie and Tu, 2015)) to segment organ-specific boundaries.

Network formulation. Our training data $S^{I/B} = \{(X_n, Y_n^{I/B}), n = 1, \dots, N\}$ where X_n denotes cropped axial CT images X_n (see Figs. 4 and 5 for details the cropping procedure), rescaled to within $[0, \dots, 255]$ with a soft-tissue window of $[-160, 240]$ HU in order to allow fine-tuning from other models that have been trained on natural images with this intensity range. $Y_n^I \in \{0, 1\}$ and $Y_n^B \in \{0, 1\}$ denote the binary ground truths of the interior and boundary map of the pancreas, respectively, for any corresponding X_n. Each image is considered holistically and independently as in Xie and Tu (2015). The network is able to learn features from these images alone from which interior and boundary prediction maps can be produced, which we denote as **HNN-I** and **HNN-B**, respectively.

HNN can efficiently generate multi-level image features due to its deep architecture. Furthermore, multiple stages with different convolutional strides can capture the inherent scales of organ edge/interior labeling maps. However, due to the difficulty of learning such deep neural networks with multiple stages from scratch, we use the pre-trained network provided by Xie and Tu (2015) and fine-tuned to our specific training data sets $S^{I/B}$. We use the HNN network architecture with 5 stages, including strides of 1, 2, 4, 8 and 16, respectively, and with different receptive field sizes as suggested by the authors.[2]

In addition to standard CNN layers, a HNN network has M side-output layers as shown in Fig. 3. These side-output layers are also realized as classifiers in which the corresponding weights are $\boldsymbol{w} = (\boldsymbol{w}^{(1)}, \dots, \boldsymbol{w}^{(M)})$. For simplicity, all standard network layer parameters are denoted as $\boldsymbol{W}$. Hence, the following objective function can be defined[3]:

$$\mathcal{L}_{\text{side}}(\boldsymbol{W}, \boldsymbol{w}) = \sum_{m=1}^{M} \alpha_m l_{\text{side}}^{(m)}(\boldsymbol{W}, \boldsymbol{w}^m). \qquad (1)$$

Here, l_{side} denotes an image-level loss function for side-outputs, computed over all pixels in a training image pair X and Y. The term α_m allows to weight different side-output layers against in each other. We found $\alpha_m = 1$ to work fine in practice. Note that, in HNNs, the side-output layers are connected to the standard CNN layers via deconvolutional layers and their parameters are fixed to perform bilinear interpolation (see Fig. 6). This approach is identical to the use of deconvolutional layers for upsampling in FCNs (Long et al., 2015) and allows the computation of a cross-entropy loss function for each side-output layer by comparing to the ground truth per-pixel label. All HNN layers are path-connected and their parameters can be updated during training via backpropagation (Xie and Tu, 2015). Because of the heavy bias towards negatively labeled pixels in the ground truth data, (Xie and Tu, 2015) introduces a strategy to automatically balance the loss between positive and negative classes via a per-pixel class-balancing weight β. This offsets the imbalances between edge/interior $(y = 1)$ and non-edge/exterior $(y = 0)$ samples. Specifically, a class-balanced cross-entropy loss function can be used in Eq. (1) with j iterating over the spatial dimensions of the image:

$$l_{\text{side}}^{(m)}(\boldsymbol{W}, \boldsymbol{w}^{(m)}) = -\beta \sum_{j \in Y_+} \log Pr\Big(y_j = 1 | X; \boldsymbol{W}, \boldsymbol{w}^{(m)}\Big) - (1 - \beta) \sum_{j \in Y_-} \log Pr\Big(y_j = 0 | X; \boldsymbol{W}, \boldsymbol{w}^{(m)}\Big). \qquad (2)$$

Here, β is simply $|Y_-|/|Y|$ and $1 - \beta = |Y_+|/|Y|$, where $|Y_-|$ and $|Y_+|$ denote the ground truth set of *negatives* and *positives*, respectively. In contrast to Xie and Tu (2015), where β is computed for each training image independently, we use a constant balancing weight computed on the entire training set. This is because some training slices might have no positives at all and otherwise would be ignored in the loss function. The class probability $Pr(y_j = 1 | X; W, w^{(m)}) = \sigma(a_j^{(m)}) \in [0, 1]$ is computed on the activation value at each pixel j using the sigmoid function $\sigma(.)$. Now, organ edge/interior map predictions $\hat{Y}(m)_{\text{side}} = \sigma(\hat{A}(m)_{\text{side}})$ can be obtained at each side-output layer, where $\hat{A}(m)_{\text{side}} \equiv \{a_j^{(m)}, j = 1, \dots, |Y|\}$ are activations of the side-output of layer m. Finally, a “weighted-fusion” layer is added to the network that can be simultaneously learned during training. The loss function at the fusion layer L_{fuse} is defined as

$$\mathcal{L}_{\text{fuse}}(\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h}) = Dist\Big(Y, \hat{Y}_{\text{fuse}}\Big), \qquad (3)$$

where $\hat{Y}_{\text{fuse}} = \sigma\Big(\sum_{m=1}^{M} h_m^{\hat{A}_{\text{side}}}\Big)$ with $h = (h_1, \dots, h_M)$ being fusion weights that are learned within a convolutional layer with kernel size 1×1 that is applied to the concatenated side output layers. *Dist*(., .) is a distance measure between the fused predictions and the ground truth label map. We use cross-entropy loss for this purpose. Hence, the following objective function can be minimized via standard stochastic gradient descent and back propagation as in Xie and Tu (2015):

$$(\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h})^{\star} = \text{argmin}(\mathcal{L}_{\text{side}}(\boldsymbol{W}, \boldsymbol{w}) + \mathcal{L}_{\text{fuse}}(\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h})) \qquad (4)$$

Testing phase. Given image X, we obtain both interior (**HNN-I**) and boundary (**HNN-B**) predictions from the models' side output layers and the weighted-fusion layer as in Xie and Tu (2015):

$$\Big(\hat{Y}_{\text{fuse}}^{I}, \hat{Y}_{\text{side}}^{I_1}, \dots, \hat{Y}_{\text{side}}^{I_M}\Big) = \textbf{HNN-I}(X, (\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h})^{\star}) \qquad (5)$$

[2] https://github.com/s9xie/hed.

[3] We follow the notation of (Xie and Tu, 2015).

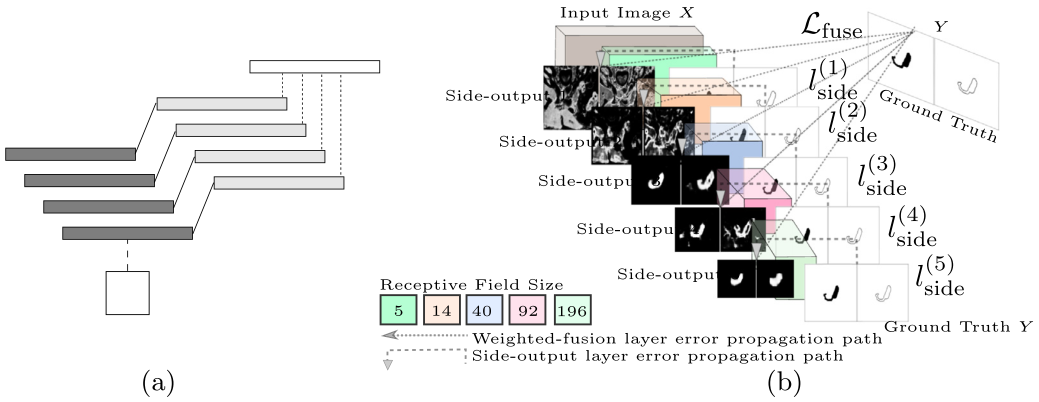


Fig. 3. Schematics of (a) a holistically-nested network, in which multiple side outputs are added, and (b) the **HNN-I/B** network architecture for both interior (left images) and boundary (right images) detection pathways. We highlight the error back-propagation paths to illustrate the deep supervision performed at each side-output layer after the corresponding convolutional layer. As the side-outputs become smaller, the receptive field sizes get larger. This allows HNN to combine multi-scale and multi-level outputs in a learned weighted fusion layer. The ground truth images are inverted for aided visualization (Figures adapted from Xie and Tu, 2015 with permission).

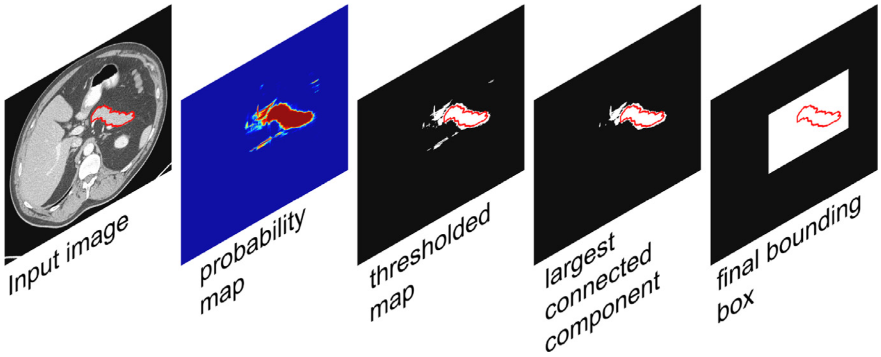


Fig. 4. Candidate bounding box region generation pipeline (left to right). Gold standard pancreas in red. We start from CT images are tightly cropped around the patient's based on a simple threshold and connected component analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

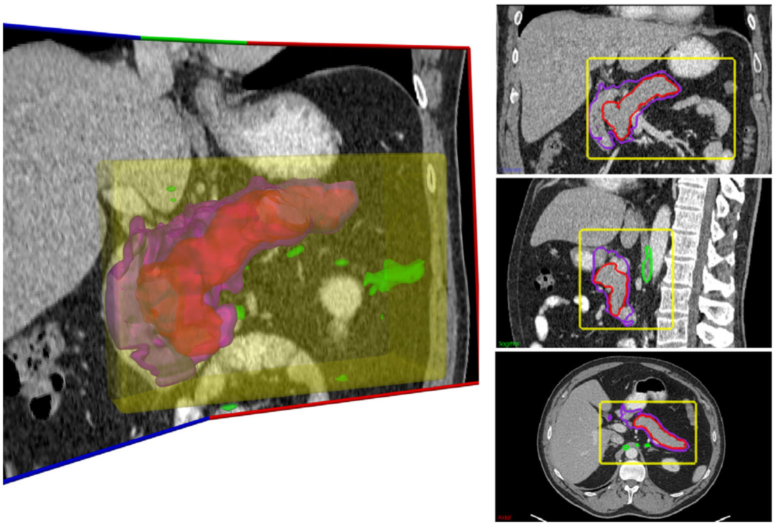


Fig. 5. Candidate bounding box region generation. Gold standard pancreas in red, blobs of ≥0.5 probabilities in green, the selected largest 3D connected component in purple, the resulting candidate bounding box that is used for tightly cropping the images is shown in yellow (no margin has been added). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\left(\hat{Y}^{B}_{\text{fuse}}, \hat{Y}^{B_1}_{\text{side}}, \ldots, \hat{Y}^{B_M}_{\text{side}}\right) = \textbf{HNN-B}(X, (\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h})^{\star}) \qquad (6)$$

Here, $\textbf{HNN-I/B}(\cdot)$ denotes the interior/boundary prediction maps estimated by the CNN networks using the optimized parameters $(\boldsymbol{W}, \boldsymbol{w}, \boldsymbol{h})^{\star}$.

2.2. Pancreas localization

Segmentation performance can be enhanced if irrelevant regions of the CT volume are pruned out. Conventional organ localization methods using random forest regression (Criminisi et al., 2013; Lay et al., 2013), which we explain in Section 2.2.1, may

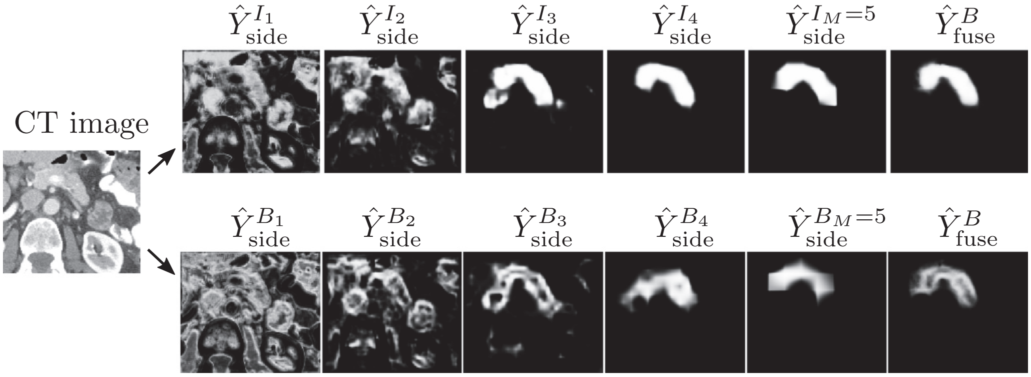

CT image
$\hat{Y}^{I_1}_{\text{side}}$
$\hat{Y}^{I_2}_{\text{side}}$
$\hat{Y}^{I_3}_{\text{side}}$
$\hat{Y}^{I_4}_{\text{side}}$
$\hat{Y}^{I_M=5}_{\text{side}}$
$\hat{Y}^{B}_{\text{fuse}}$
$\hat{Y}^{B_1}_{\text{side}}$
$\hat{Y}^{B_2}_{\text{side}}$
$\hat{Y}^{B_3}_{\text{side}}$
$\hat{Y}^{B_4}_{\text{side}}$
$\hat{Y}^{B_M=5}_{\text{side}}$
$\hat{Y}^{B}_{\text{fuse}}$

Fig. 6. The side-output and fused prediction maps from **HNN-I**: $\hat{Y}^{I/B_1}_{\text{side}}$, $\hat{Y}^{I/B_M}_{\text{side}}$, and $\hat{Y}^{I/B}_{\text{fuse}}$ applied for pancreas segmentation, after the localization stage. As can be seen, the side-outputs from deeper within the network become more semantically coherent to the pancreas. Each side output is upsampled to the same size via bilinear interpolation implemented via deconvolutional layers.

not guarantee that the regressed organ bounding box contains the targeted organ with extremely high sensitivities on the pixel-level coverage. In Section 2.2.2 we outline a superpixel based approach (Farag et al., 2014), based on hand-crafted and CNN features, that is able to provide improved performance. While this is effective, the complexity involved motivates our own development of a simpler and more accessible newly proposed multi-view HNN fusion based procedure. This is explained in Section 2.2.3. The output of the localization method will later feed into a more detailed and accurate segmentation method combining multiple mid-level cues from HNNs as illustrated in Fig. 2.

2.2.1. Regression forest

Object localization by regression has been studied extensively in the literature including (Criminisi et al., 2013; Cuingnet et al., 2012; Lay et al., 2013). The general idea is to predict an offset vector $\Delta x \in \mathbb{R}^3$ for a given image patch $I(x)$ centered about $x \in \mathbb{R}^3$. The predicted object position is then given as $x + \Delta x$. This is repeated for many examples of image patches and then aggregated to produce a final predicted position. Aggregation can be done with non-maximum suppression on prediction voting maps, mean aggregation (Criminisi et al., 2013), cluster medoid aggregation (Cuingnet et al., 2012), and the use of local appearance with discriminative models to accept or reject predictions (Lay et al., 2013). The pancreas can be localized by regression due to their locations in the body in correlation to other anatomical structures. The objective is to predict bounding boxes $(x_{\text{center}}, \Delta x_{\text{lower}}, \Delta x_{\text{upper}}) \in \mathbb{R}^{3\times 3}$ where x_{center} is the center of the pancreas and $x_{\text{center}} + \Delta x_{\text{lower}}$ and $x_{\text{center}} + \Delta x_{\text{upper}}$ are the lower and upper corner of the pancreas bounding box respectively. The addition of the extra three parameters follows from the observation that the center of the bounding box is not necessarily the center of the localized object. The pancreas Regression Forest predicts $(\Delta x,\ \Delta x_{\text{lower}},\ \Delta x_{\text{upper}})$ for a given image patch $I(x)$. This produces pancreas bounding box candidates of the form $(x + \Delta x, \Delta x_{\text{lower}}, \Delta x_{\text{upper}})$. We additionally use a discriminative model to accept or reject predictions $x + \Delta x$ as in Lay et al. (2013). Finally, accepted predictions are aggregated using non-maximum suppression over probability scores and then the bounding boxes are ranked by the count of accepted predictions within the box. The box with the highest count of predictions is kept as the final prediction. Training regression forest in this work uses 100 mm × 100 mm × 100 mm Haar features and is constrained to the maximum depth of 10. The decision criteria optimization is based on a variance reduction gain objective. The discriminative model used to accept or reject candidates is a sliding-window detector based on a two level cascade of Random Forest. It is trained independently of the Regression Forest and employs Haar features with multi-scale window sizes.

2.2.2. Random forest on superpixels

As a form of initialization, we alternatively employ a previously proposed method based on random forest (RF) classification (Farag et al., 2014; Roth et al., 2015) using both hand-crafted and deep CNN derived image features to compute a candidate bounding box regions. We only operate the RF labeling at a low probability threshold of > 0.5 which is sufficient to reject the vast amount of non-pancreas from the CT images. This initial candidate generation is sufficient to extract bounding box regions that nearly surround the pancreases completely in all patient cases with $\sim 97\%$ recall. All candidate regions are computed during the testing phase of cross-validation (CV) as in Roth et al. (2015). As we will see next, candidate generation can be done even more efficiently by using the same HNN architectures, which are based on convolutional neural networks. The technical details of HNNs were described in Section 2.1.

2.2.3. Multi-view aggregated HNNs

Alternatively to the candidate region generation process described in Section 2.2.2 that uses hybrid deep and non-deep learning techniques, we employ **HNN-I** (interior, see Section 2.1) as a building block for pancreas localization, inspired by the effectiveness of HNN being able to capture the complex pancreas appearance in CT images (Roth et al., 2016a). This enables us to drastically discard large negative volumes of the CT scan, while operating **HNN-I** on a conservative probability threshold of >=0.5 that retains high sensitivity/recall ($> 99\%$). The constant balancing weight on β during training **HNN-I** is critical in this step since the high majority of CT slices have empty pancreas appearance and are indeed included for effective training of **HNN-I** models, in order to successfully suppress the pancreas probability values from appearing in background. Furthermore, we perform a largest connected-component analysis to remove outlier "blobs" of high probabilities. To get rid of small incorrect connections between high-probability blobs, we first perform an erosion step with radius of 1 voxel, and then select the largest connected-component, and subsequently dilate the region again (Fig. 4). **HNN-I** models are trained in axial, coronal, and sagittal planes in order to make use of the multi-view representation of 3D image context. Empirically, we found a max-pooling operation across the 3D models to give the highest sensitivity/recall while still being sufficient to reject the vast amount of non-pancreas from the CT images (see Table 2). One illustrative example is demonstrated in Fig. 5. This initial candidate generation is sufficient to extract bounding box regions that completely surround the pancreases with nearly 100% recall. All candidate regions are computed during the testing phase of cross-validation (CV) with the same split as in Roth et al. (2015). Note that this candidate region proposal is a crucial step for further processing. It removes “easy” non-pancreas tissue from further analysis and allows **HNN-I** and **HNN-B** to focus on the more difficult distinction of pancreas versus its surrounding tissue. The fact that we can use exactly the same HNN model architecture for both stages though is noteworthy.

2.3. Pancreas segmentation

With pancreas localized, the next step is to produce a reliable segmentation. Our segmentation pipeline consists of three steps. We first use HNN probability maps to generate mid-level boundary and interior cues. These are then used to produce superpixels, which are then aggregated together into a final segmentation using RF classification.

2.3.1. Combining mid-level cues via HNNs

We now show that organ segmentation can benefit from multiple mid-level cues, like organ interior and boundary predictions. We investigate deep-learning based approaches to independently learn the pancreas’ interior and boundary mid-level cues. Combining both cues via learned spatial aggregation can elevate the overall performance of this semantic segmentation system. Organ boundaries are a major mid-level cue for defining and delineating the anatomy of interest. It could prove to be essential for accurate semantic segmentation of an organ.

2.3.2. Learning organ-specific segmentation proposals

Multiscale combinatorial grouping (MCG) (Pont-Tuset et al., 2017) is one of the state-of-the-art methods for generating segmentation object proposals in computer vision. Inside MCG, a multi-resolution image pyramid is used to perform the hierarchical segmentation at each scale independently. These hierarchies are then aligned and combined into a single multi-scale segmentation hierarchy. Last, the combinatorial space of these regions is explored using an efficient grouping algorithm in order to produce a ranked list of region proposals (Pont-Tuset et al., 2017). Different from other generation image segmentation methods, MCG utilizes the supervisedly learned image boundary cue to better capture more semantic object-level edges, instead of raw image gradients. We utilize this approach and publicly available code,[4] to generate organ-specific superpixels based on the learned boundary predication maps **HNN-B**. Superpixels are extracted via continuous oriented watershed transform at three different scales, denoted ($\hat{Y}^{B_2}_{\text{side}}, \hat{Y}^{B_3}_{\text{side}}, \hat{Y}^{B}_{\text{fuse}}$), supervisedly learned by **HNN-B**. This allows the computation of a hierarchy of superpixel partitions at each scale, and merges superpixels across scales, thereby efficiently exploring their combinatorial space (Pont-Tuset et al., 2017). This allows MCG to group the merged superpixels toward object proposals.

The later **HNN-B** layers and the fused layer are more semantically coherent at object level than the early ones that are mainly focusing on low level feature, like edges – thus, this would result in too many and too finer-scaled superpixel candidates if early layers are used. In our empirical evaluation, layer 2 and 3 are the suitable trade-off to capture some mid-level image gradients combined with the more semantic fused layer (see Fig. 6). On the other hand, we find that the first two levels of object MCG proposals are sufficient to achieve ∼88% DSC while keeping moderate numbers of superpixel proposals (see Table 4 and Fig. 7). The segmentation upper-bound performance is obtained by the optimally computed superpixel labels using their spatial overlapping ratios against the segmentation ground truth map. All *merged* superpixels $\mathcal{S}$ from the first two levels are used for the subsequent spatial aggregation step. Note that **HNN-B** can only be trained using axial slices where the manual annotation was performed. Note that the pancreas boundary maps in coronal and sagittal views can display strong artifacts because the manual annotation was only performed on axial slices. Hence, we compute the pancreas-specific superpixels using only the axial **HNN-B**.

Note that superpixels obtained from unsupervised image cues (which is the normal process on generating superpixels or image regions) are used for pancreas segmentation in an early work (Farag et al., 2014). Furthermore, four different image segmentation methods are evaluated and compared in Farag et al. (2014), including Watersheds (Vincent and Soille, 1991), SLIC (Achanta et al., 2012), efficient graph-based partitioning (Felzenszwalb and Huttenlocher, 2004) and Entropy rate (Liu et al., 2011). Assuming that the pancreas segmentation ground-truth masks are known as an “Oracle”, we can assign the optimal labels for superpixels to compute the upper-bounded pancreas segmentation results. Based on the AVGDIST (average surface distance) metric on comparing segmentation results, the optimal performance of the SLIC (Achanta et al., 2012) generated superpixels are 1.06 mm, and AVGDIST=1.69 mm for Felzenszwalb and Huttenlocher (2004). From Table 4, our organ-specific supervisedly trained superpixels or segmentation proposals have the upper-bound segmentation limit at AVGDIST=0.16 mm which is significantly smaller than 1.06 mm (Achanta et al., 2012) or 1.69 mm (Felzenszwalb and Huttenlocher, 2004). This observation clearly demonstrates the performance benefit of employing the supervised boundary cues (encoded by **HNN-B**) into the procedure of image superpixel generation.

2.3.3. Spatial aggregation with random forest

We use the superpixel set $\mathcal{S}$ generated previously to extract features for spatial aggregation via random forest classification.[5] Within any superpixel $s \in \mathcal{S}$ we compute simple statistics including the 1st–4th order moments (mean, variance, skewness, kurtosis), and 8 percentiles $[20\%, 30\%, \ldots, 90\%]$ on the CT intensities, and multi-view **HNN-I**s and **HNN-B** responses. Additionally, we compute the mean *x*, *y*, and *z* coordinates normalized by the range of the 3D candidate region (Section 2.2.3). This results in 87 features describing each superpixel and are used to train a RF classifier on the training positive or negative superpixels at each round of 4-fold CV. Empirically, we find 50 trees to be sufficient to model our feature set. A final 3D pancreas segmentation is simply obtained by stacking each slice prediction back into the original CT volume space. No further post-processing is employed.

This complete pancreas segmentation model is denoted as **HNN-RF**. Note that this model differs from the model employed in Roth et al. (2016a) in that it aggregates information across all three orthogonal planes of the images and not just the axial plane as in Roth et al. (2016a).

3. Experimental results

3.1. Data

Manual tracings of the pancreas for 82 contrast-enhanced abdominal CT volumes are provided by a publicly available dataset[6] (Roth et al., 2015), for the ease of comparison. Our experiments are conducted on random splits of 82 patients into four folds of

[4] https://www.github.com/jponttuset/mcg.

[5] Using MATLAB’s TreeBagger() class.

[6] https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU.

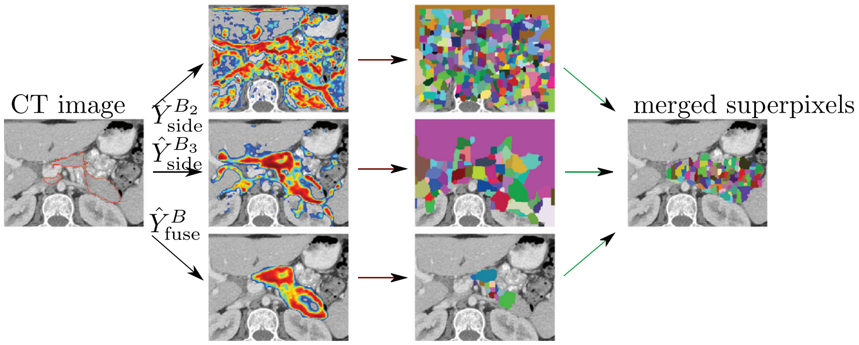

CT image
$\hat{Y}^{B_2}_{\text{side}}$
$\hat{Y}^{B_3}_{\text{side}}$
$\hat{Y}^{B}_{\text{fuse}}$
merged superpixels

Fig. 7. Multiscale combinatorial grouping (MCG) (Pont-Tuset et al., 2017) on three different scales of learned boundary predication maps from **HNN-B**: $\hat{Y}^{B_2}_{\text{side}}$, $\hat{Y}^{B_3}_{\text{side}}$, and $\hat{Y}^{B}_{\text{fuse}}$ using the original CT image on far left as input (with ground truth delineation of pancreas in red). MCG computes superpixels at each scale and produces a set of merged superpixel-based object proposals. We only visualize the boundary probabilities whose values are greater than .10. From this illustrative example, the boundary maps from all three **HNN-B** levels are needed and combined to produce a hierarchy of MCG superpixels that achieve high recalls on preserving the pancreas boundaries. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

20, 20, 21, 21 patients, respectively. In each round of 4-fold cross-validation (CV-4), we employ three folds of data as training of both segmentation cascades, and the left fold for testing. This means that each patient case is exactly seen and validated once in testing (of total 4 rounds). CV-4 is used throughout in this section, unless otherwise mentioned.

3.2. Training

The HNN architecture and hyper parameters like learning rates (lr) remain fixed for all models used in this work. We fine-tune from the initial models (VGGnet or BSDS500) with a relatively small $lr = 1e-6$ using a stepwise weight decay by a factor 10 for every 10,000 iterations as suggested by Xie and Tu (2015). Training is stopped at 100,000 iterations over the training image slices. Training time on a NVIDIA TITAN X GPU card took around 9–12 h for all models. In testing, the runtime is in the order of 2–3 min for each case applying the cascaded **HNN-I** models. **HNN-RF** processing adds another 5 minutes of processing time. Our implementation and models are available for download.[7]

3.3. Evaluation

We perform extensive quantitative evaluation on different configurations of our method and compare to the previous state-of-the-art work with in-depth analysis.

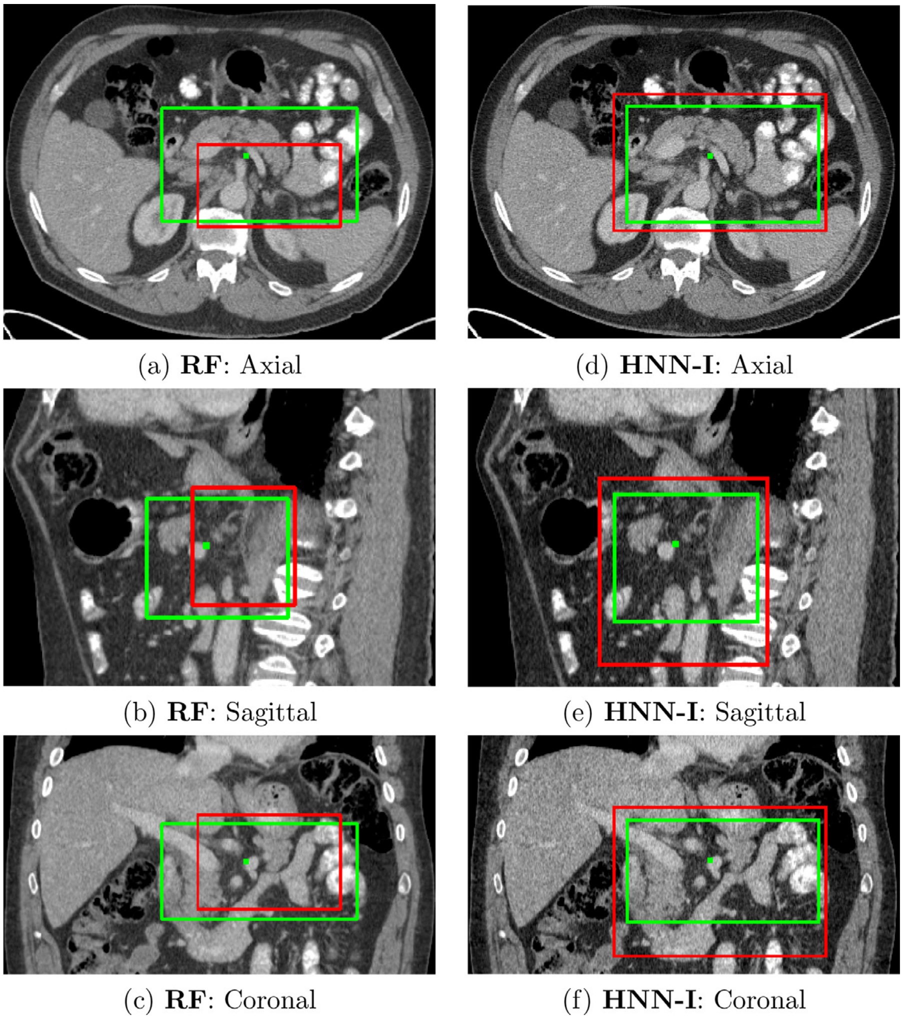
3.3.1. Localization

From our empirical study, the candidate region bounding box generation based on multi-view max-pooled **HNN-I**s (Section 2.2.3) or previous hybrid methods (Section 2.2.2 (Farag et al., 2014)) works comparably in terms of addressing the requirement to produce spatially-truncated 3D regions that maximally cover the pancreas in the pixel-to-pixel level and reject as much as possible the background spaces. An average reduction of absolute volume of 90.36% (range [80.45%–96.26%]) between CT scan and candidate bounding box is achieved during this step, while keeping a mean recall of 99.93%, ranging [94.54%–100.00%]. Table 1 shows the test performance of pancreas localization and bounding box prediction using regression forests in DSC and average Euclidean distance against the gold standard bounding boxes. The distance errors are significantly lower for our proposed prediction scheme with $p < 0.001$ (Wilcoxon signed rank test). As illustrated in Fig. 9, regression forest based localization generates 16 out of 82 bounding boxes that lie below 60% in the pixel-to-pixel recall against the ground-truth pancreas masks. Nevertheless we obtain nearly 100% recall for all scans (except for two cases $\geq 94.54\%$) through the multi-view max-pooled **HNN-I**s. An example of detected pancreas can be seen in Fig. 8.

3.3.2. HNN spatial aggregation for pancreas segmentation

The interior HNN models trained on the axial (AX), coronal (CO) or sagittal (SA) CT images in Section 2.2.3 can be straightforwardly used to generate pancreas segmentation masks. We exploit different spatial aggregation or pooling functions on the AX, CO, and SA viewed **HNN-I** probability maps, denoted as **AX, CO, SA** (any single view **HNN-I** probability map simply used); **mean(AX,CO), mean(AX,SA), mean(CO,SA)** and **mean(AX,CO,SA)** (element-wise mean of two or three view **HNN-I** probability maps); **max(AX,CO,SA)** (element-wise maximum of three view **HNN-I** probability maps); and finally **meanmax(AX,CO,SA)** (element-wise mean of the maximal two scores from three view **HNN-I** probability maps). After the optimal thresholding calibrated using the training folds on these pooled **HNN-I** maps, the resulting binary segmentation masks are further refined by 3D connected component process and simple morphological operations (as in Section 2.2.3). Table 2 demonstrates the DSC pancreas segmentation accuracy performance by investigating different spatial aggregation functions. We observe that the element-wise multi-view (mean or max) pooling operations on **HNN-I** probabilities maps generally outperform their single view counterparts. **max(AX,CO,SA)** performs slightly better than **mean(AX,CO,SA)**. The configuration of **meanmax(AX,CO,SA)** produces the most superior performance in mean DSC which may behave as a robust fusion function by rejecting the smallest probability value and averaging the remained two **HNN-I** scores per pixel location. After the pancreas localization stage, we train a new set of multi-view **HNN-I**s with the spatially truncated scales and extents. This serves a desirable "Zoom Better to See Clearer" effect for deep neural network segmentation models (Xia et al., 2016) where cascaded **HNN-I**s only focus on discriminating or parsing the remained organ candidate regions. Similarly, DSC [%] pancreas segmentation accuracy results of various spatial aggregation or pooling functions on AX, CO, and SA viewed **HNN-I** probability maps (trained in the second cascaded stage) are shown in Table 3. We find consistent empirical observations as above when comparing multi-view HNN pooling operations. The **meanmax(AX,CO,SA)** operation again reports the best mean DSC performance at 81.14% which is increased significantly from 76.79% in Table 2 with $p < 0.001$ (Wilcoxon signed rank test). We denote this system configuration as $\mathbf{HNN}_{\text{meanmax}}$. This result validates our two staged pancreas segmentation framework of proposing candidate region generation for organ localiza-

[7] https://github.com/rsummers11/CADLab/tree/master/panreas_hnn.


(a) **RF**: Axial

(d) **HNN-I**: Axial

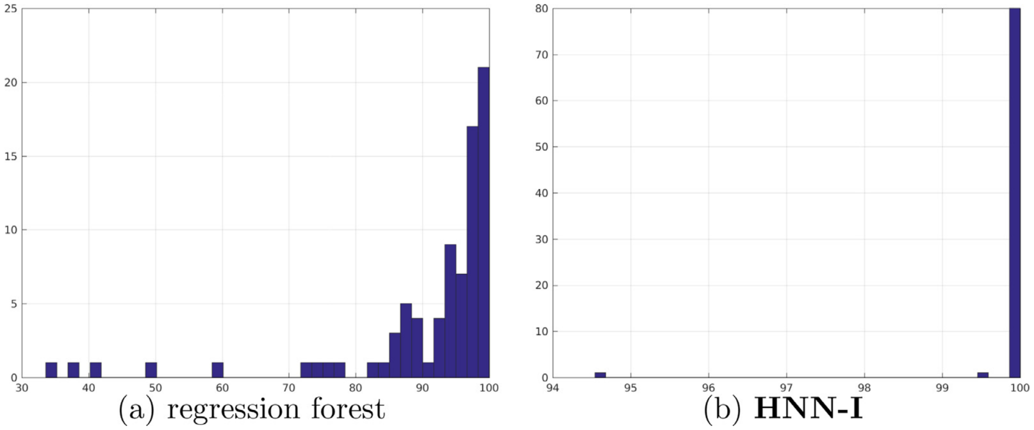
(b) **RF**: Sagittal

(e) **HNN-I**: Sagittal

(c) **RF**: Coronal

(f) **HNN-I**: Coronal

Fig. 8. An example for comparison of regression forest (RF, a–c) and **HNN-I** (d–f) for pancreas localization. Green and red boxes are ground truth and detected bounding boxes respectively. The green dot denotes the ground truth center. This case demonstrates a case in the 90th percentile in RF localization distance and serves as a representative of poorly performing localization. In contrast, **HNN-I** includes all of the pancreas with nearly 100% recall in this case. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) regression forest

(b) **HNN-I**

Fig. 9. Histogram plots (Y-Axis) of regression forest based bounding boxes (a) and **HNN-I**'s generated bounding boxes (b) in recalls (X-axis) covering the ground-truth pancreas masks in 3D. Note that Regression Forest produces 16 out of 82 bounding boxes that lie below 60% in pixel-to-pixel recall while **HNN-I** produces 100% recalls, except for two cases $\geq 94.54\%$.

Table 1
Test performance of pancreas localization and bounding box prediction using regression forests (RF) in Dice and average Euclidean distance against the gold standard bounding boxes, in 4-fold cross validation. The percentages state to the *k*-percentile for each metric. For comparison, the same metrics are shown for the bounding boxes generated by **HNN-I** (as described in Section 2.2.3). The *p*-values for testing significant difference between both methods of localization are shown (Wilcoxon signed rank test). Statistically significant improvement is shown in *italic* with $p < 0.05$.

Metric	Method	Mean	Std	10%	50%	90%	Min	Max	*p*-value
Dist.	**RF**	14.9	9.4	6.4	11.7	29.3	2.8	48.7	
(mm)	**HNN-I**	*9.4*	*5.2*	*4.5*	*8.4*	*16.0*	*1.9*	*27.7*	*< 0.001*
Dice	**RF**	71.2	10.7	56.2	73.4	83.1	33.1	91.5	
(%)	**HNN-I**	71.5	11.5	56.6	71.5	85.5	38.3	95.2	0.994

Table 2
Four-fold cross-validation: DSC [%] pancreas segmentation performance of various spatial aggregation functions on AX, CO, and SA viewed **HNN-I** probability maps in the *candidate region generation stage* (the best results in **bold**). Performances that are significantly different from AX are shown in *italic* ($p < 0.05$, Wilcoxon signed rank test).

DSC	Mean	Std	Min	Max	*p*-value
AX	73.46	11.63	1.88	85.97	n/a
CO	*70.19*	***9.81***	***39.72***	*83.84*	*< 0.001*
SA	72.42	11.26	14.00	84.92	0.051
mean(AX,CO)	*74.65*	*11.21*	*5.08*	*86.87*	*0.002*
mean(AX,SA)	*75.08*	*12.29*	*2.31*	*86.97*	*< 0.001*
mean(CO,SA)	73.70	11.40	18.96	86.64	0.293
mean(AX,CO,SA)	*75.07*	*12.08*	*4.26*	*87.19*	*< 0.001*
max(AX,CO,SA)	*75.67*	*10.32*	*16.11*	*87.65*	*< 0.001*
meanmax(AX,CO,SA)	***76.79***	*11.07*	*8.97*	***88.03***	*< 0.001*

Table 3
Four-fold cross-validation: DSC [%] pancreas segmentation performance of various spatial aggregation functions on AX, CO, and SA viewed **HNN-I** probability maps in the *second cascaded stage* (the best results in **bold**). Performances that are significantly different from AX are shown in *italic* ($p < 0.05$, Wilcoxon signed rank test).

DSC	Mean	Std	Min	Max	*p*-value
AX	78.99	7.70	44.25	88.69	n/a
CO	*76.16*	*8.67*	*45.29*	*88.11*	*< 0.001*
SA	*76.53*	*9.35*	*40.60*	*88.34*	*< 0.001*
mean(AX,CO)	79.02	7.96	42.64	88.82	0.285
mean(AX,SA)	79.29	8.21	42.32	89.38	0.096
mean(CO,SA)	77.61	8.92	44.14	89.11	0.057
mean(AX,CO,SA)	*80.40*	*7.30*	*45.18*	*89.11*	*< 0.001*
max(AX,CO,SA)	*80.55*	***6.89***	***45.66***	*89.92*	*< 0.001*
meanmax(AX,CO,SA)	***81.14***	*7.30*	*44.69*	***89.98***	*< 0.001*

tion followed by “Zoomed” deep HNN models to refine segmentation. Table 4 shows the improvement from the **meanmax**-pooled **HNN-I**s (i.e., **HNN**$_{\text{meanmax}}$) to the **HNN-RF** based spatial aggregation, using DSC and average minimum surface-to-surface distance (AVGDIST). The average DSC is increased from 81.14% to 81.27%, However, this improvement is not statistically significantly with $p > 0.05$ using Wilcoxon signed rank test. In contrast, using dense CRF (DCRF) optimization (Chen et al., 2014) (with **HNN-I** as the unary term and the pairwise term depending on the CT values) as a means of introducing spatial consistency does not improve upon **HNN-I** noticeably as shown in Roth et al. (2016a). Comparing to the performance of previous state-of-the-art methods (Roth et al., 2015; 2016a) at mean DSC scores of 71.4% and 78.01% respectively, both variants of **HNN**$_{\text{meanmax}}$ and **HNN-RF** demonstrate superior quantitative segmentation accuracy in DSC and AVGDIST metrics. We have the following two observations. 1, The main performance gain compared to Roth et al. (2016a) (similar to **HNN**$_{\text{AX}}$ in Table 3) is found by the multi-view aggregated HNN pancreas segmentation probability maps (e.g., **HNN**$_{\text{meanmax}}$), which also serve in **HNN-RF**. 2, The new candidate region bounding box generation method (Section 2.2.3) works comparably to the hybrid technique (Section 2.2.2 (Farag et al., 2014; Roth et al., 2015; 2016a)) based on our empirical evaluation. However the proposed pancreas localization via multi-view max-pooled HNNs greatly simplified our overall pancreas segmentation system which may also help the generality and reproducibility. The variant of **HNN**$_{\text{meanmax}}$ produces competitive segmentation accuracy but merely involves evaluating two sets of multi-view **HNN-I**s at two spatial scales: whole CT slices or truncated bounding boxes. There is no need to compute any hand-crafted image features (Farag et al., 2014) or train other external machine learning classifiers. As shown in Fig. 9, the conventional organ localization framework using regression forest (Criminisi et al., 2013; Lay et al., 2013) does not address well the purpose of candidate region generation for segmentation where extremely high pixel-to-pixel recall is required since it is mainly designed for organ detection. In Table 5, the quantitative pancreas segmentation performance of two method variants, **HNN**$_{\text{meanmax}}$, **HNN-RF** spatial aggregation, are evaluated using four metrics of DSC (%), Jaccard Index (%) (Levandowsky and Winter, 1971), Hausdorff distance (HDRFDST [mm]) (Rockafellar and Wets, 2005) and

Table 4
Four-fold cross-validation: The DSC [%] and average surface-to-surface minimum distance (AVGDIST [mm]) performance of Roth et al. (2015), Roth et al. (2016a), **HNN**$_{\text{meanmax}}$, **HNN-RF** spatial aggregation, and optimally achievable superpixel assignments (*italic*). Best performing method in **bold**.

DSC	Roth et al. (2015)	Roth et al. (2016a)	**HNN**$_{\text{meanmax}}$	HNN-RF	*Opt.*
Mean	71.42	78.01	81.14	**81.27**	*87.67*
Std	10.11	8.20	7.30	**6.27**	*2.21*
Min	23.99	34.11	44.69	**50.69**	*81.59*
Max	86.29	88.65	**89.98**	88.96	*91.71*
AVGDIST	Roth et al. (2015)	Roth et al. (2016a)	**HNN**$_{\text{meanmax}}$	**HNN-RF**	***Opt.***
Mean	1.53	0.60	0.43	**0.42**	*0.16*
Std	1.60	0.55	0.32	**0.31**	*0.04*
Min	0.20	0.15	**0.12**	0.14	*0.10*
Max	10.32	4.37	**1.88**	2.26	*0.26*

Table 5
Four-fold cross-validation: The quantitative pancreas segmentation performance results of our two method variants, **HNN**$_{\text{meanmax}}$, **HNN-RF** spatial aggregation, in four metrics of DSC (%), Jaccard Index (%), Hausdorff distance (HDRFDST [mm]), and AVGDIST [mm]. Best performing methods are shown in **bold**. Note that there is no statistical significance when comparing the performance by two variants in three measures of DSC, JACARD, and AVGDIST, except for HDRFDIST with $p < 0.001$ (Wilcoxon Signed Rank Test). This indicates that **HNN-RF** may be more robust than **HNN**$_{\text{meanmax}}$ in the worst case scenario.

	DSC		Jaccard		HDRFDST		AVGDIST	
	HNN$_{\text{meanmax}}$	HNN-RF	**HNN**$_{\text{meanmax}}$	HNN-RF	**HNN**$_{\text{meanmax}}$	HNN-RF	**HNN**$_{\text{meanmax}}$	HNN-RF
Mean	81.14	**81.27**	68.82	**68.87**	22.24	**17.71**	0.43	**0.42**
Std	7.30	**6.27**	9.27	**8.12**	13.90	**10.40**	0.32	**0.31**
Median	**82.98**	82.75	**70.92**	70.57	18.03	**14.88**	0.32	0.32
Min	44.69	**50.69**	28.78	**33.95**	5.83	**5.20**	**0.12**	0.14
Max	**89.98**	88.96	79.52	**80.12**	79.52	**69.14**	**1.88**	2.26

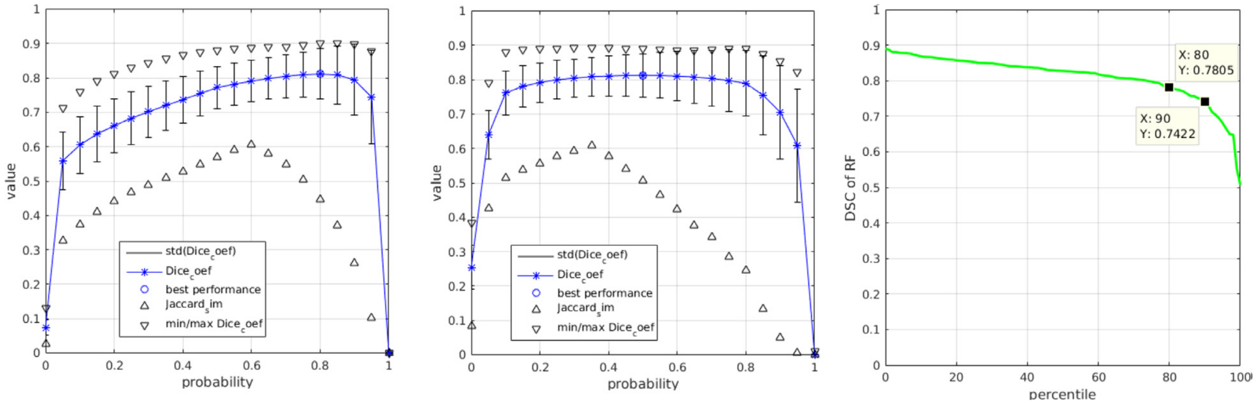


Fig. 10. Average DSC performance as a function of pancreas probability using **HNN**$_{\text{meanmax}}$ (left) and spatial aggregation via **RF** (middle) for comparison. Note that the DSC performance remains much more stable after **RF** aggregation with respect to the probability threshold. The percentage of total cases that lie above a certain DSC with **RF** are shown (right): 80% of the cases have a DSC of 78.05%, and 90% of the cases have a DSC of 74.22% and higher.

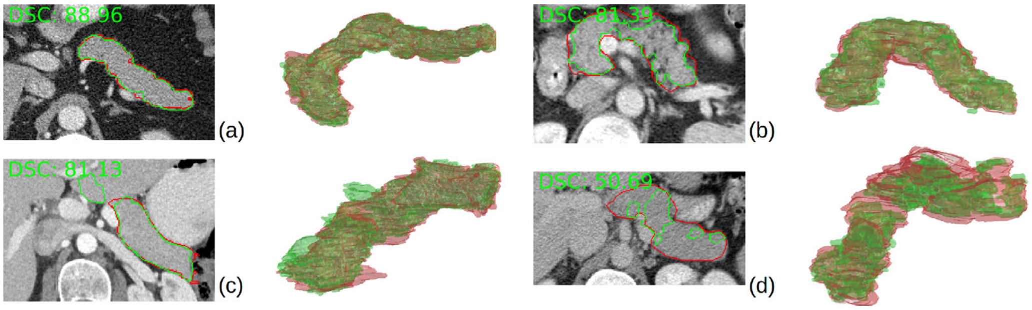


Fig. 11. Examples of our **HNN-RF** pancreas segmentation results (green) comparing with the ground-truth annotation (red). The best performing case (a), two cases with DSC scores close to the data set mean (b,c) and the worst case are shown (d). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

AVGDIST [mm]. Note that there is no statistical significance when comparing the performance of two variants in three measures of DSC, JACARD, and AVGDIST, except for HDRFDIST with $p < 0.001$ under Wilcoxon signed rank test. Since Hausdorff distance represents the maximum deviation between two point sets or surfaces, this observation indicates that **HNN-RF** may be more robust than **HNN**$_{\text{meanmax}}$ in the worst case scenario (see also the more stable pancreas segmentation behavior of **HNN-RF** with respect to the different probability thresholds shown in Fig. 10).

Pancreas segmentation on illustrative patient cases are shown in Fig. 11. Furthermore, we applied our trained **HNN-I** model on a different CT data set[8] with 30 patients, and achieve a mean DSC of 62.26% without any re-training on the new data cases, but if we average the outputs of our 4 **HNN-I** models from cross-validation, we achieve 65.66% DSC. This demonstrates that **HNN-I** may be generalizable in the cross-dataset evaluation providing further fine-tuning. Last, we collected an additional dataset of 19 unseen CT scans using the same patient data protocol (Roth et al., 2015, 2016a). Here, **HNN**$_{\text{meanmax}}$ achieves a mean DSC of 81.2%.

4. Discussion

To the best of our knowledge, our result comprises the state-of-the-art average DSC in testing folds under 4-fold CV evaluation metric. Strict comparison to other methods (except for Roth et al., 2015; Roth et al., 2016a) is not directly possible due to different datasets utilized. Our holistic segmentation approach with multi-view pooling and spatial aggregation advances the current state-of-the-art quantitative performance to an average DSC of 81.27% in testing. Previous notable results for CT images range from ∼68% to ∼78% (Wolz et al., 2013; Chu et al., 2013; Tong et al., 2015; Okada et al., 2015; Oda et al., 2016), all under the “leave-

[8] 30 training data sets at https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

Table 6
Comparison to recent literature on pancreas segmentation in CT imaging under cross-validation (CV-n with n folds), leave-one-out-validation (LOOV) or a hard training/validation/testing split. We limit this comparison to studies that employed the same dataset (Farag et al., 2017; Zhou et al., 2017) or used similar datasets in terms of patient numbers (N).

Method	N	DSC (%)	Jaccard (%)	Protocol
Wolz et al. (2012)	100 (CT)	65.5	49.6	LOOV
Wang et al. (2014b)	100 (CT)	65.5 ± 18.6	–	LOOV
Wolz et al. (2013)	150 (CT)	69.6 ± 16.7	55.5 ± 17.1	LOOV
Wolz et al. (2013)	50 (CT)	58.2 ± 20.0	43.5 ± 17.8	LOOV
Chu et al. (2013)	100 (CT)	69.1 ± 15.3	54.6	LOOV
Tong et al. (2015)	150 (CT)	71.1 ± 14.7	–	LOOV
Oda et al. (2016)	147 (CT)	75.1 ± 15.4	–	LOOV
Karasawa et al. (2017)	150 (CT)	78.5 ± 14.0	66.3 ± 15.5	LOOV
Roth et al. (2017)	150 (CT)	82.2 ± 10.2	–	281/50/150
Farag et al. (2014)	80 (CT)	68.8 ± 25.6	57.2 ± 25.4	CV-6
Farag et al. (2017)	80 (CT)	70.7 ± 13.0	57.9 ± 13.6	CV-6
Roth et al. (2015)	82 (CT)	71.8 ± 10.7	–	CV-4
Zhou et al. (2017)	82 (CT)	82.4 ± 5.7	–	CV-4
Roth et al. (2017) (same data)	82 (CT)	76.8 ± 9.4	–	CV-4
Cai et al. (2016)	78 (MRI)	76.1 ± 8.7	–	CV-3
Ours (**HNN**$_{\text{meanmax}}$)	82 (CT)	81.14 ± 7.30	68.82 ± 9.27	CV-4
Ours (**HNN-RF**)	82 (CT)	81.27 ± 6.27	68.87 ± 8.12	CV-4

one-patient-out" (LOO) cross-validation scheme. In particular, DSC drops from 68% (150 patients) to 58% (50 patients) as reported in Wolz et al. (2013). Our methods also perform with the better statistical stability, i.e., comparing 7.3% or 6.27% versus 18.6% (Wang et al., 2014c), 15.3% (Chu et al., 2013) in the standard deviation of DSC scores. The minimal DSC values are 44.69% with **HNN**$_{\text{meanmax}}$ and 50.69% for **HNN-RF** whereas Wang et al. (2014c), Chu et al. (2013), Wolz et al. (2013) and Roth et al. (2015) all report patient cases with DSC $< 10\%$. Recent work that explores the direct application of 3D convolutional filters with fully convolutional architectures also shows promise (Çiçek et al., 2016; Merkow et al., 2016; Milletari et al., 2016). It has to be established whether 2D or 3D implementations are more suited for certain tasks. There is some evidence that deep networks representations with direct 3D input suffer from the *curse-of-dimensionality* and are more prone to overfitting (Roth et al., 2016b; Su et al., 2015; 2016). In fact, we directly compare to 3D-U-Net applied in a cascaded fashion as described in Roth et al. (2017) on the same dataset and can only achieve 76.8 ± 9.4 [43.7, 89.4] % Dice score in testing when training from scratch using the same 4-fold cross-validation split (see Table 6). This means that volumetric object detection might require more training data and might suffer from scalability issues. However, proper hyper-parameter tuning of the CNN architecture and enough training data (including data augmentation) might help eliminate these problems. In the mean time, spatial aggregation in multiple 2D views (as proposed here) might be a very efficient (and computationally less expensive) way of diminishing the curse-of-dimensionality. Furthermore, using 2D views has the advantage that networks trained on much larger databases of natural images (e.g. *ImageNet, BSDS500*) can be used for fine-tuning to the medical domain. It has been shown that transfer learning is a viable approach when the medical imaging data set size is limited (Shin et al., 2016; Tajbakhsh et al., 2016). 3D CNN approaches often adopt padded spatially-local sliding volumes to parse any CT scan, e.g., $96 \times 96 \times 48$ (Merkow et al., 2016), $160 \times 160 \times 72$ (Dou et al., 2016) or $80 \times 80 \times 80$ (Chen et al., 2016a), which may cause the segmentation discontinuity or inconsistency at overlapped window boundaries. Ensemble of several neural networks trained with random configuration variations is found to be advantageous comparing a single CNN model in object recognition (Simonyan and Zisserman, 2014b; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014a). Our pancreas segmentation method can be indeed considered as ensembles of multiple correlated HNN models but good complementary information gain since they are trained from orthogonal axial, coronal or sagittal CT views.

Additionally, as studied in Kamnitsas et al. (2017b), for problems of segmenting small but variable objects, such as brain lesions, micro-brain bleeding, lung nodules and so on, 2D CNN can be uncertain to train and recognize where 3D CNN may be the default choice of CNN architecture. We do not mean that 2D CNNs will always perform better than 3D CNNs but would argue that 2D CNN approaches, when properly used, serve as valid and competitive alternatives. Especially, spatial aggregation of 3D information via multiple-view robust fusion using 2D CNNs performs comparatively well to employing 3D CNNs directly, while being substantially easier to train (requiring less patient data) and computationally efficient (and less GPU memory intensive). Note that the network width (i.e., the convolutional kernel numbers) in HNN is markedly wider than the 3D CNN architecture adopted in Kamnitsas et al. (2017b). The recent work of Roth et al. (2017) exploits cascaded 3D CNNs (fully convolutional architectures implemented via 3D U-Net (Çiçek et al., 2016)) for pancreas segmentation and reports a very comparable performance of $82.2 \pm 10.2\%$ DSC on a similar data set (Roth et al., 2015), only after having to use a much larger training set of > 300 annotated pancreas CT scans. As a comparable 2D cascaded CNN approach, Zhou et al. (2017) also performs on par ($82.4 \pm 5.7\%$ DSC) to our approach. Cai et al. (2016) reports the pancreas segmentation result of $76.1 \pm 8.7\%$ DSC on 78 MRI pancreas volumes under 3-fold cross-validation. For a comprehensive comparison to recent literature on pancreas segmentation, see Table 6.

The state-of-the-art reported results for several computer-aided detection problems are still based on fusing multi-view 2D CNNs (Roth et al., 2016b; Arindra et al., 2016). Last, the segmentation performance drop when directly applying the deep models trained on the public dataset (Roth et al., 2015) to the patient data of another medical center indicates the importance of cross-center and cross-protocol (i.e., patient recruiting protocols) deep transfer learning (Kamnitsas et al., 2017a).

5. Conclusion

In conclusion, we present a holistic deep CNN approach for pancreas localization and segmentation in abdominal CT scans, exploiting multi-view spatial pooling and combining interior and boundary mid-level cues. The robust fusion of **HNN**$_{\text{meanmax}}$ aggregating on interior holistically-nested networks (**HNN-I**) alone already achieve good performance at DSC of 81.14% ± 7.30% in 4-fold CV. The other method variant **HNN-RF** incorporates the organ boundary responses from the **HNN-B** model and significantly improves the worst case pancreas segmentation accuracy in Hausdorff distance ($p < 0.001$). The highest reported DSCs of 81.27% ± 6.27% is achieved, at the computational cost of 2–3 min, not hours as in Wang et al. (2014c), Chu et al. (2013) and Wolz et al. (2013). Because of the data-driven manner of how our deep learning based approach learns features for organ segmentation, it could be generalizable to other segmentation problems with large variations such as pathological organs (Harrison et al., 2017b).

Acknowledgment

This work was supported by the Intramural Research Program of National Institutes of Health Clinical Center. The authors would like to thank Prof. Kensaku Mori of Nagoya University, Japan, for providing the computational resources for the comparison experiment with Roth et al. (2017).

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11).

Arindra, A., Setio, A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S., Wille, M., Naqibullah, M., Snchez, C., van Ginneken, B., 2016. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging 35 (5), 1160–1169.

Avants, B., Tustison, N., Song, G., Cook, P., Klein, A., Gee, J., 2011. A reproducible evaluation of ants similarity metric performance in brain image registration. NeuroImage 54(3):2033–2044.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ants). Insight J.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell.

Bai, W., Shi, W., O'Regan, D., Tong, T., Wang, H., Jamil-Copley, S., Peters, N., Rueckert, D., 2013. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images. IEEE TMI 32(7):1302–1315.

Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q., 2016. Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks. In: MICCAI, 2, pp. 442–450.

Chen, H., Dou, Q., Yu, L., Heng, P., 2016a. Voxresnet: deep voxelwise residual networks for volumetric brain segmentation. arXiv preprint arXiv:1608.05895.

Chen, H., Qi, X., Yu, L., Heng, P., 2016. Dcan: deep contour-aware networks for accurate gland segmentation. IEEE CVPR.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Nimura, Y., Rueckert, D., Mori, K., 2013. Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. In: MICCAI, pp. 165–172.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. Springer, pp. 424–432.

Cootes, T.F., Hill, A., Taylor, C.J., Haslam, J., 1994. Use of active shape models for locating structures in medical images. Image Vis. Comput. 12 (6), 355–365.

Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K., 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. MedIA 17 (8), 1293–1303.

Cuingnet, R., Prevost, R., Lesage, D., Cohen, L., Mory, B., Ardon, R., 2012. Automatic detection and segmentation of kidneys in 3D ct images using random forests. In: MICCAI, pp. 66–74.

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P., 2016. 3D deeply supervised network for automatic liver segmentation from ct volumes. MICCAI, arXiv preprint arXiv:1607.00582.

Ecabert, O., Peters, J., Schramm, H., Lorenz, C., von Berg, J., Walker, M., Vembar, M., Olszewski, M., Subramanyan, K., Lavi, G., Weese, J., 2008. Automatic model-based segmentation of the heart in ct images. IEEE TMI 27(9):1189–1201. doi:10.1109/TMI.2008.918330.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. IEEE PAMI 35(8):1915–1929.

Farag, A., Lu, L., Roth, H.R., Liu, J., Turkbey, E., Summers, R.M., 2017. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. IEEE Trans. Image Process. 26 (1), 386–399.

Farag, A., Lu, L., Turkbey, E., Liu, J., Summers, R.M., 2014. A bottom-up approach for automatic pancreas segmentation in abdominal CT scans. MICCAI Abdominal Imaging Workshop.

Felzenszwalb, P., Huttenlocher, D., 2004. Efficient graph-based image segmentation. Int. J. Comp. Vis. 59, 167–181.

Fu, H., Xu, Y., Wong, D.W.K., Liu, J., 2016. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on. IEEE, pp. 698–701.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE PAMI 38 (1).

Harrison, A., Xu, Z., George, K., Lu, L., Summers, R., Mollura, D., 2017. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. MICCAI. Springer.

Harrison, A.P., Xu, Z., George, K., Lu, L., Summers, R.M., Mollura, D.J., 2017. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. MICCAI. Springer.

Hou, Q., Cheng, M.-M., Hu, X.-W., Borji, A., Tu, Z., Torr, P., 2016. Deeply supervised salient object detection with short connections. arXiv preprint arXiv:1611.04849.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., amd D. Rueckert, A.C., Glocker, B., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging, pp. 597–609.

Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. MedIA 36, 61–78.

Karasawa, K., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Chu, C., Zheng, G., Rueckert, D., Mori, K., 2017. Multi-atlas pancreas segmentation: atlas selection based on vessel structure. Med Image. Anal. 39, 18–28.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In: IEEE CVPR, pp. 1725–1732.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. NIPS.

Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K., 2013. Rapid multi-organ segmentation using context integration and discriminative models. In: Information Processing in Medical Imaging, pp. 450–462.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2014. Deeply-supervised nets. arXiv:1409.5185.

Lee, C.-Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. AISTATS.

Levandowsky, M., Winter, D., 1971. Distance between sets. Nature 234 (5), 34–35.

Ling, H., Zhou, S., Zheng, Y., Georgescu, B., Suehling, M., Comaniciu, D., 2008. Hierarchical, learning-based automatic liver segmentation. In: IEEE CVPR, pp. 1–8.

Liu, M.-Y., Tuzel, O., Ramalingam, S., Chellappa, R., 2011. Entropy rate superpixel segmentation. In: CVPR. IEEE, pp. 2097–2104.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE CVPR, pp. 3431–3440.

Lu, L., Barbu, A., Wolf, M., Liang, J., Comaniciu, D., 2008. Accurate polyp segmentation for 3D ct colonography using multi-staged probabilistic binary learning and compositional model. IEEE CVPR.

Merkow, J., Kriegman, D., Marsden, A., Tu, Z., 2016. Dense volume-to-volume vascular boundary detection. MICCAI, arXiv preprint arXiv:1605.08401.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV). IEEE, pp. 565–571.

Modat, M., McClelland, J., Ourselin, S., 2010. Lung registration using the niftyreg package. In: Medical Image Analysis for the Clinic-A Grand Challenge, pp. 33–42.

Murphy, K., van Ginneken, B., et al., 2011. Evaluation of registration methods on thoracic ct: the empire10 challenge. IEEE TMI 30 (11), 1901–1920.

Nogues, I., Lu, L., Wang, X., Roth, H., Bertasius, G., Lay, N., Shi, J., Tsehay, Y., Summers, R., 2016. Automatic lymph node cluster segmentation using holistically-nested networks and structured optimization. MICCAI.

Oda, M., Shimizu, N., Karasawa, K., Nimura, Y., Kitasaka, T., Misawa, K., Fujiwara, M., Rueckert, D., Mori, K., 2016. Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In: MICCAI. Springer, pp. 556–563.

Okada, T., Linguraru, M.G., Hori, M., Summers, R.M., Tomiyama, N., Sato, Y., 2015. Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors. MedIA 26 (1), 1–18.

Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J., 2017. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (1), 128–140.

Rockafellar, R., Wets, R., 2005. Variational analysis. Nature.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI. Springer, pp. 556–564.

Roth, H.R., Lu, L., Farag, A., Sohn, A., Summers, R.M., 2016. Spatial aggregation of holistically-nested networks for automated pancreas segmentation. MICCAI. Springer.

Roth, H.R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., Summers, R.M., 2016. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans. Med. Imaging 35 (5), 1170–1181.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: MICCAI. Springer, pp. 520–527.

Roth, H. R., Oda, H., Hayashi, Y., Oda, M., Shimizu, N., Fujiwara, M., Misawa, K., Mori, K., 2017. Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:1704.06382.

Shin, H., Roth, H., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learnings. IEEE TMI 35 (5), 1285–1298.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576.

Simonyan, K., Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition. ICLR 2015 arXiv:1409.1556.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition. IEEE ICCV.

Su, H., Qi, C., Niessner, M., Dai, A., Yan, M., Guibas, L., 2016. Volumetric and multi--view cnns for object classification on 3D data. IEEE CVPR.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE TMI 35 (5), 1299–1312. doi:10.1109/TMI.2016.2535302.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. MedIA 23 (1), 92–104.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13, 583–598.

Wang, H., Suh, J., Das, S., Pluta, J., Craige, C., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion. IEEE PAMI.

Wang, L., Shi, F., Li, G., Gao, Y., Lin, W., Gilmore, J., Shen, D., 2014. Segmentation of neonatal brain mr images using patch-driven level sets. NeuroImage 84 (1), 141–158.

Wang, Z., Bhatia, K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K., Rueckert, D., 2014. Geodesic patch-based segmentation. In: MICCAI, 1, pp. 666–673.

Wang, Z., Bhatia, K.K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K., Rueckert, D., 2014. Geodesic patch-based segmentation. In: MICCAI, pp. 666–673.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. TMI 32 (9), 1723–1730.

Wolz, R., Chu, C., Misawa, K., Mori, K., Rueckert, D., 2012. Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases. In: MICCAI, 1, pp. 10–17.

Xia, F., Wang, P., Chen, L.-C., Yuille, A. L., 2016. Zoom better to see clearer: human and object parsing with hierarchical auto-zoom net. ECCV, arXiv preprint arXiv:1607.00582.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. In: IEEE ICCV, pp. 1395–1403.

Yan, Z., Zhan, Y., Peng, Z., Liao, S., Zhou, X.S., 2015. Bodypart recognition using multi-stage deep learning. In: Information Processing in Medical Imaging. Springer, pp. 449–461.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks. In: IEEE ICCV, pp. 1529–1537.

Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., 2008. Four-chamber heart modeling and automatic segmentation for 3D cardiac ct volumes using marginal space learning and steerable features. IEEE TMI 27 (11), 1668–1681.

Zhou, Y., Xie, L., Shen, W., Fishman, E., Yuille, A., 2017. Pancreas segmentation in abdominal ct scan: a coarse-to-fine approach. arXiv preprint, MICCAI arXiv:1612.08230.