Chapter 1
Pancreas Segmentation in CT and MRI via Task-Specific Network Design and Recurrent Neural Contextual Learning

Jinzheng Cai, Le Lu, Fuyong Xing and Lin Yang

Abstract Automatic pancreas segmentation in radiology images, e.g., computed tomography (CT), and magnetic resonance imaging (MRI), is frequently required by computer-aided screening, diagnosis, and quantitative assessment. Yet, pancreas is a challenging abdominal organ to segment due to the high inter-patient anatomical variability in both shape and volume metrics. Recently, convolutional neural networks (CNN) have demonstrated promising performance on accurate segmentation of pancreas. However, the CNN-based method often suffers from segmentation discontinuity for reasons such as noisy image quality and blurry pancreatic boundary. In this chapter, we first discuss the CNN configurations and training objectives that lead to the state-of-the-art performance on pancreas segmentation. We then present a recurrent neural network (RNN) to address the problem of segmentation spatial inconsistency across adjacent image slices. The RNN takes outputs of the CNN and refines the segmentation by improving the shape smoothness.

J. Cai · L. Yang (✉)
J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA
e-mail: lin.yang@bme.ufl.edu

J. Cai
e-mail: jimmycai@ufl.edu

L. Lu
PAII Inc., Bethesda Research Lab, 6720B Rockledge Drive Ste 410, Bethesda, MD 20817, USA
e-mail: le.lu@paii-labs.com; lelu@cs.jhu.edu

Johns Hopkins University, Baltimore, MD, USA

F. Xing
Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
e-mail: fuyong.xing@ucdenver.edu

L. Lu et al. (eds.), *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-13969-8_1

1.1 Introduction

Detecting unusual volume changes and monitoring abnormal growths in pancreas using medical images is a critical yet challenging task for computer-aided diagnosis (CAD). This would require to delineate pancreas from its surrounding tissues in radiology images, e.g., computed tomography (CT), and magnetic resonance imaging (MRI) scans. The accurate segmentation of pancreas delivers more reliable and quantitative representations than the cross section diameter measurement and it may facilitate the producing of segmentation-based biomarkers, such as volumetric measurements and 3D shape/surface signatures. Moreover, automated rapid and accurate segmentation of pancreas on the scale of processing thousands of image scans may facilitate new protocols, findings, and insights for clinical trials. On the other hand, manual pancreas segmentation is very expensive and sometimes even intractable on the dataset at a very large scale. To fulfill this practical and important demand, many efforts have been investigated to significantly boost the segmentation performance in both CT and MRI modalities.

One major group on the automatic pancreas segmentation in CT images is based on top-down multi-atlas registration and label fusion (MALF) [11, 17, 26, 27]. Due to challenges from the high deformable shape and vague boundaries of the pancreas in CT scans from various patients, the reported segmentation accuracy (measured in Dice Similarity Coefficient or DSC) is limited in the range from 69.6 ± 16.7% [27] to 78.5 ± 14.0% [11, 17] under leave-one-patient-out (LOO) evaluation protocol. On the other hand, bottom-up deep CNN-based pancreas segmentation work [2, 8, 19–21, 30] have revealed promising results and steady performance improvements, e.g. from 71.8 ± 10.7% [19], 78.0 ± 8.2% [20], to 81.3 ± 6.3% [21] evaluated using the same NIH 82-patient CT dataset [6, 19] under fourfold cross-validation (CV).

Deep learning-based approaches appear to demonstrate noticeably higher segmentation accuracies and numerically more stable results (significantly lower in standard deviation, or std). References [20, 21] are built upon the fully convolutional neural network (FCN) architecture [13] and its variant [28]. However, [20, 21] both rely on post-processing with random forest to further refine CNN's outputs, which cannot propagate errors back to the CNN model. Similarly, for pancreas segmentation on a 79-patient MRI dataset, [2] achieves 76.1 ± 8.7% in DSC, where graph-based result fusion is applied. Therefore, an end-to-end trainable deep learning model for pancreas segmentation may be more desirable to achieve superior results. Additionally, deep CNN-based bottom-up pancreas segmentation methods also have significant advantages on run-time computational efficiency, such as 2∼4 h [11] versus 2∼3 m [21] to process a new segmentation case.

1.2 Convolutional Neural Network for Pancreas Segmentation

With the goal of obtaining accurate segmentation for objects with complex shape outlines, many deep learning approaches have been proposed [4, 13, 18, 28, 29] to report good performances. Specifically, some of these models regularize deep learning output with appearance constraint that image pixels sharing similar color and location statistics would probably come from the same object category and this leads to conditional random field (CRF) post-processing that presented in [4, 29]. On the other hand, some methods propose to learn localized deep learning features. For instance, deep supervision is proposed in [12, 28] forcing all convolutional layers to learn effective and consistent low-level representations, e.g., edge and object boundary. Meanwhile, U-Net architecture [18] makes full use of low-level convolutional feature maps by projecting them back to the original image size. The dedicated backward propagation combines convolutional layer features of multiple scales, thereby boosting the accuracy of object segmentation.

1.2.1 Design of Network Architecture

Delineating the pancreas boundary from its surrounding anatomies in CT/MRI scans is challenging due to its complex visual appearance and ambiguous outlines. For example, Fig. 1.1 displays an example of the pancreas in CT and MRI images, where the pancreas shares similar intensity values with other soft tissues and its boundary is blurry where touching with other abdominal organs. These natures of pancreas segmentation inspire us to combine virtues from both of the holistically nested network (HNN) [28] and U-Net [18] and we name the combination P-Net as it is designed task specifically for pancreas segmentation. In Fig. 1.2, we visually depict the network architecture of the proposed P-Net and the most correlated networks, i.e. HNN and U-Net. The P-Net inherits the deep supervisions from HNN and the skip connections from U-Net for feature multi-scale aggregation.

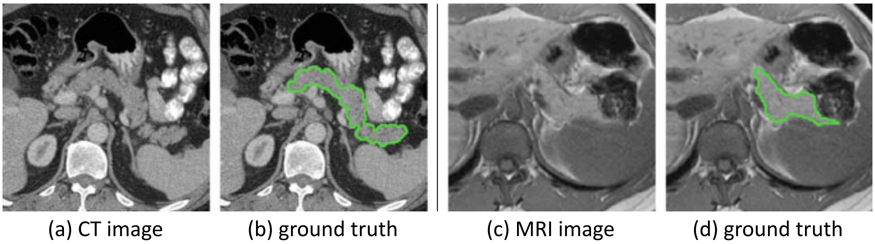

Fig. 1.1 Examples of pancreas images in CT and MRI. The ground truth pancreas boundaries are presented in (**b**) and (**d**) delineated in green. Best viewed in color

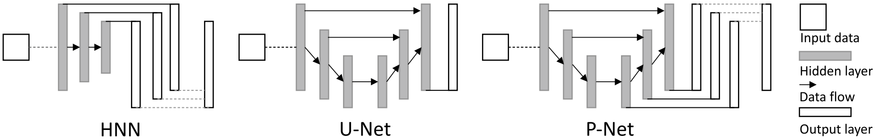

Fig. 1.2 Network architecture of several popular CNN models for pancreas segmentation

1.2.2 Design of Model Training Strategy

Apart from the network architecture, the proposed task-specific design also includes a strategy to train the P-Net from scratch. This is because the CT and MRI modalities demonstrate very different image statistics from natural images. Thus, a direct transfer of ImageNet [7] pretrained neural network to the medical domain could be suboptimal. Similarly, to transfer the model from other medical tasks can also be problematic for P-Net to achieve the optimal pancreas segmentation. During model training, we also observe gradient-vanishing problem occurs when fine-tune the network from pretrained models. Specifically, the top layers of the network will quickly capture the hierarchical (conceptual) appearance of pancreas but leaving lower layers not well tuned as the magnitudes of gradients (backpropagated from the training error) fastly decreased. To circumvent this problem, we propose to initialize P-Net from scratch and train the network layer by layer.

To present the training algorithm of P-Net in formal, we denote the number of steps for layer-by-layer training to be K. Then, the corresponding convolutional and deconvolutional layers can be represented as $\{C_1, \ldots, C_K\}$ and $\{D_1, \ldots, D_K\}$. We also denote standard network operations as up-sampling to be $Up(\#)$, concatenation to be $Concat([\cdot; \cdot])$, and dimension reduction to be $R(\#)$. For representation clarity, we use $\circ$ to denote the composition of two transformations and use $\prod$ for multiple transformations. We drop pooling operations between convolutional layers for simplicity.

First, we define the forward propagation as a combination of convolutional layers

$$F_k = \prod_{i=1}^{K+1-k} C_i. \tag{1.1}$$

Then in P-Net, the feature map is up-scaled step-by-step util it restores the size of the original input image. Specifically, when k equals 1, we have

$$B_k = F_k \circ D_k, \tag{1.2}$$

otherwise, the feature map process can be represented as

$$B_k = Concat([B_{k-1} \circ Up(2);\ F_k])\ \circ\ D_k. \tag{1.3}$$

At each scale (or step), the segmentation output is

$$\hat{Y}_k = B_k \circ Up\ (2^{K-k}) \circ R(1) \circ Sigmoid, \tag{1.4}$$

where the feature map B_k is first up-scaled by a factor of 2^{K-k} such that restores the size of the original input image. Its channel is then be reduced to 1 via $R(1)$, and then it passes through a sigmoid activation function *Sigmoid* to produce the segmentation output at scale (or step) k as $\hat{Y}_k$.

The output $\hat{Y}_k$ is a probability map on which the segmentation loss is measured as

$$\mathcal{L}_k = H(\hat{Y}_k, Y), \tag{1.5}$$

where Y is the ground truth label map and $H(\cdot, \cdot)$ is the loss function, e.g. cross-entropy loss. Thus, each unit module B_k has its own gradient back propagation path that starts at the loss $\mathcal{L}_k$ and ends at the input image. It introduces deep supervision to the bottom CNN layers and enables us to train P-Net from swallow to deep. More specifically, the training of P-Net starts at $k = 1$ and increase k by 1 when $\mathcal{L}_k$-plot converges. The final segmentation result $\hat{Y}$ is a weighted combination of the side outputs as

$$\hat{Y} = \sum_{i=1}^{K} \omega_i \hat{Y}_i, \tag{1.6}$$

and the overall objective for P-Net is,

$$\mathcal{L} = H(\hat{Y}, Y) + \sum_{i=1}^{K} \mathcal{L}_i, \tag{1.7}$$

where $K = 5$ delivers the best for pancreas segmentation. We summarize the training procedure in Algorithm 1 and visually depict the semantic illustration of P-Net structures at $k = 1$, $k = 2$, and $k = 5$, respectively, in Fig. 1.3.

Discussion: Although the network architecture of P-Net can be extended to process 3D inputs [5], we maintain the current 2D architecture in model training and inference because the 3D version can be computationally expensive while gaining

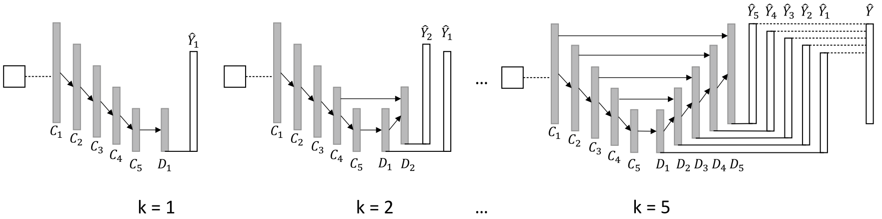
$\hat{Y}_1$
C_1 C_2 C_3 C_4 C_5 D_1
k = 1
$\hat{Y}_2$ $\hat{Y}_1$
C_1 C_2 C_3 C_4 C_5 D_1 D_2
k = 2
...
$\hat{Y}_5$ $\hat{Y}_4$ $\hat{Y}_3$ $\hat{Y}_2$ $\hat{Y}_1$ $\hat{Y}$
C_1 C_2 C_3 C_4 C_5 D_1 D_2 D_3 D_4 D_5
k = 5

Fig. 1.3 Semantic illustration of the P-Net training algorithm

no significant improvement in performance [30]. As a compromise, P-Net takes 3-connected slices as its input when given the segmentation ground truth mask of the middle slice. As explained in Sect. 1.3, P-Net is transformed into a lightweighted 3D model with RNN stacked to the end of it, which allows our model to capture 3D imaging information with minor extra computational loads. This is in a similar spirit to employ RNN to regulate, process and aggregate CNN activations for video classification and understanding [16].

```
Result: $\hat{Y}, \hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_K$
$K$=5, convolutions: $\{C_1, \ldots, C_K\}$, deconvolutions: $\{D_1, \ldots, D_K\}$;
Define: $F_k = \prod_{i=1}^{K+1-k} C_i$;
for $k$=1:K do
    if $k$==1 then
        $B_k = F_k \circ D_k$;
    else
        $B_k = Concat([B_{k-1} \circ Up(2); F_k]) \circ D_k$;
    end
    $\hat{Y}_k = B_k \circ Up(2^{K-k}) \circ R(1) \circ Sigmoid$;
    $\mathcal{L}_k = H(\hat{Y}_k, Y)$;
    Optimize $\sum_{i=1}^{k} \mathcal{L}_i$ until converge;
end
$\hat{Y} = \sum_{i=1}^{K} \omega_i \hat{Y}_i$;
$\mathcal{L} = H(\hat{Y}, Y)$;
Optimize $\mathcal{L} + \sum_{i=1}^{K} \mathcal{L}_i$ until converge;
```

Algorithm 1: The training algorithm of P-Net.

1.2.3 Design of Loss Functions

Loss functions compare the segmented volumetric image (e.g., $\hat{Y}$) with ground truth annotation (i.e., Y) and produces segmentation errors for model updating. Cross-entropy loss is one of the most popular loss functions that widely used for foreground–background segmentation. It is defined as

$$\mathcal{L}_{ce} = -\frac{1}{|Y|} \sum_{j \in Y} [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)], \tag{1.8}$$

where $|Y|$ is the cardinality (or size) of Y representing the number of voxels in the volumetric image. It can be observed in the formulation of $\mathcal{L}_{ce}$ that errors from every voxel are equally treated. However, it is common in medical volumes that the anatomy of interest occupies only a very small region of the image. Thus, a direct use of $\mathcal{L}_{ce}$ will probably result in the foreground regions to miss or only partially detect. To automatically balance the loss between foreground/background classes, a class-balanced cross-entropy is designed to remedy this problem. It is defined as

$$\mathcal{L}_{cbce} = -\frac{\beta}{|Y_+|}\sum_{j\in Y_+}\log \hat{y}_j - \frac{1-\beta}{|Y_-|}\sum_{j\in Y_-}\log(1-\hat{y}_j), \tag{1.9}$$

where a class-balancing weight β is introduced on a per-voxel term basis. Specifically, we define $\beta = |Y_-|/|Y|$ and $1-\beta = |Y_+|/|Y|$, where Y_+ and Y_- denote the foreground and background ground truth label sets, respectively.

Apart from $\mathcal{L}_{ce}$ and $\mathcal{L}_{cbce}$, many work directly optimize evaluation metrics, e.g. Jaccard index and Dice score. In terms of advantages, the Jaccard loss makes procedures of model training and testing consistent and helps to generate threshold-free probability maps. It is defined as

$$\begin{aligned}\mathcal{L}_{jac} = \mathcal{L}(\hat{Y}, Y) &= 1 - \frac{|Y_+ \bigcap \hat{Y}_+|}{|Y_+ \bigcup \hat{Y}_+|} \\ &= 1 - \frac{\sum_{j\in Y_+}(y_j \wedge \hat{y}_j)}{\sum_{j\in Y_-}(y_j \vee \hat{y}_j)} = 1 - \frac{\sum_{j\in Y_+}(1 \wedge \hat{y}_j)}{|Y_+| + \sum_{j\in Y_-}(0 \vee \hat{y}_j)}.\end{aligned} \tag{1.10}$$

Practically, $\hat{y}_j$ can be relaxed to the value of foreground probability $\in [0, 1]$ and $\mathcal{L}_{jac}$ is then be approximated by

$$\tilde{\mathcal{L}}_{jac} = 1 - \frac{\sum_{f\in Y_+}\min(1, \hat{y}_f)}{|Y_+| + \sum_{b\in Y_-}\max(0, \hat{y}_b)} = 1 - \frac{\sum_{f\in Y_+}\hat{y}_f}{|Y_+| + \sum_{b\in Y_-}\hat{y}_b}. \tag{1.11}$$

The model is then updated by gradient flows as

$$\frac{\partial \tilde{\mathcal{L}}_{jac}}{\partial \hat{y}_j} = \begin{cases} -\frac{1}{|Y_+|+\sum_{b\in Y_-}\hat{y}_b}, \text{ for } j \in Y_+ \\ \\ \frac{\sum_{f\in Y_+}\hat{y}_f}{(|Y_+|+\sum_{b\in Y_-}\hat{y}_b)^2}, \text{ for } j \in Y_- \end{cases} \tag{1.12}$$

Since the inequality $(\sum_{j\in Y_+}\hat{y}_j) < (|Y_+| + \sum_{j\in Y_-}\hat{y}_j)$ holds, the Jaccard loss $\mathcal{L}_{jac}$ assigns greater gradients to foreground pixels than the background ones, which intrinsically balances the foreground and background classes. It empirically works better than the cross-entropy loss $\mathcal{L}_{ce}$ and classed balanced cross-entropy loss $\mathcal{L}_{cbce}$ when segmenting small-sized objects, e.g., pancreas in CT/MRI images.

1.2.4 Experimental Results

1.2.4.1 Experimental Setup

Datasets and evaluation metrics We use two fully annotated pancreas datasets to validate the presented methods. The first one is the NIH-CT dataset [6, 19, 20] that is publicly available and contains 82 abdominal contrast-enhanced 3D CT scans. We organize an in-house MRI dataset [2] that consists of 79 abdominal T1-weighted MRI scans. We treat the CT and MRI datasets as two independent groups and repeat the experiment on both of them. Results from both groups are evaluated to validate the generalization of segmentation methods.

In image preprocessing, we use simple morphological operations to find the abdomen area and have it extracted from the whole image volume. To generate images for training, we use ground truth mask to allocate pancreas and then crop a 256×256 sub-image centered at the target pancreas region. The cropped image patch is then fed for model training. In the data inference phase, we scan testing images with the 256×256 scanning window and fuse outputs together to generate the final segmentation result.

Following the evaluation strategy in [2, 19, 20], we conduct *fourfold cross-validation (CV)* for the reported segmentation result. The set of used evaluation metrics includes the Dice similarity coefficient (DSC): $DSC = 2(|Y_+ \cap \hat{Y}_+|)/(|Y_+| + |\hat{Y}_+|)$, Jaccard index (JI): $JI = (|Y_+ \cap \hat{Y}_+|)/(|Y_+ \cup \hat{Y}_+|)$, foreground precision: $P = |\hat{Y}_+ \cap Y_+|/|\hat{Y}_+|$ and foreground recall: $R = |\hat{Y}_+ \cap Y_+|/|Y_+|$.

Network Implementation We implement HNN [20], U-Net [18], and the introduced P-Net for comparison. Especially, lower layers of HNN are transferred from ImageNet [7] pretrained VGG-16 model [24], and the U-Net is initiated from a ssTEM [18] pretrained model. We note that ssTEM has a very different image statistics from CT and MRI images. Thus, the HNN and U-Net are two baseline methods that fine-tuned from other domains and the proposed P-Net is first initialized with Xavier initialization [9] and then trained from scratch.

Hyperparameters are determined via model selection with the training set. Specifically, the training dataset is first split into a training subset for network parameter training and a validation subset for hyperparameter selection. Denote the training accuracy as Acc_t after model selection, we then combine the training and validation subsets together to further fine-tune the network until its performance on the validation subset converges to Acc_t. Also validated from the validation subset, we observe the P-Net architecture, which contains 5 unit modules with 64 output channels in each convolution/deconvolution layer produces the best empirical performance and meanwhile holds a compact model size. The learning rate (1e-4) together with other hyperparameters are all fixed for all sets so that changes observed in experiment results can be traced back to factors of interest.

1.2.4.2 CNN Comparison

Table 1.1 presents segmentation results of different CNN architectures that trained with $\mathcal{L}_{jac}$. Without loss of generality, we set the output threshold for all CNN outputs to 0.5. P-Net achieved the best performance on both of the CT and MRI datasets. Specifically, it marginally outperformed the HNN baseline by 3.7% and 4.8% Dice scores in CT and MRI segmentation, respectively. In comparison with the U-Net

Table 1.1 Comparison to different CNN architectures including P-Net, HNN [28] and U-Net [18]. Specifically, the P-Net is trained from scratch while HNN and U-Net are fine-tuned from pretrained models

	Models	DSC (%)	JI (%)	Precision	Recall
CT	HNN [28]	79.6 ± 7.7 [41.9, 88.0]	66.7 ± 9.40 [26.5, 78.6]	83.4 ± 6.5 [62.0, 94.9]	77.4 ± 11.6 [28.3, 92.6]
	U-Net [18]	79.7 ± 7.6 [43.4, 89.3]	66.8 ± 9.60 [27.7, 80.7]	81.3 ± 7.5 [49.6, 97.0]	79.2 ± 11.3 [38.6, 94.1]
	P-Net	**83.3 ± 5.6** **[59.0, 91.0]**	**71.8 ± 7.70** **[41.8, 83.5]**	**84.5 ± 6.2** **[60.7, 96.7]**	**82.8 ± 8.37** **[56.4, 94.6]**
MRI	HNN [28]	75.9 ± 10.1 [33.0, 86.8]	62.1 ± 11.3 [19.8, 76.6]	**84.4 ± 6.4** **[61.0, 93.5]**	70.6 ± 13.3 [20.7, 88.2]
	U-Net [18]	79.9 ± 7.30 [54.8, 90.5]	67.1 ± 9.50 [37.7, 82.6]	83.7 ± 6.9 [64.6, 94.6]	77.3 ± 10.3 [46.1, 94.8]
	P-Net	**80.7 ± 7.40** **[48.8, 90.5]**	**68.2 ± 9.64** **[32.3, 82.7]**	84.3 ± 7.6 [55.8, 95.8]	**78.3 ± 10.2** **[38.6, 95.0]**

baseline, P-Net presented 3.6% and 0.8% Dice scores improvements in CT and MRI segmentation, respectively.

1.2.4.3 Loss Function Comparison

Table 1.2 presents comparison results of the three losses, i.e., the cross-entropy loss $\mathcal{L}_{ce}$, the class-balanced cross-entropy loss $\mathcal{L}_{cbce}$ [28], and the Jaccard loss $\mathcal{L}_{jac}$, under fourfold cross-validation with the same P-Net segmentation model. On the CT dataset, $\mathcal{L}_{jac}$ outperformed $\mathcal{L}_{ce}$ and $\mathcal{L}_{cbce}$ by 0.5% and 0.2% Dice scores, respectively. On the MRI dataset, also achieved the best performance referring to the Dice score and Jaccard index. We then evaluate the stability of segmentation performance with various thresholds. The CNN network usually outputs probabilistic segmentation maps instead of binary masks and an appropriate probability threshold is required to obtain the final binarized segmentation outcomes. However, it is often nontrivial to find the optimal probability threshold in practice. Figure 1.4 visually depicts results of our analysis that the probability output maps from the Jaccard loss $\mathcal{L}_{jac}$ delivered the steadiest segmentation results referring to different output thresholds. Empirically, the Naïve $\mathcal{L}_{ce}$ assigns same penalties on positive and negative pixels so that the probability threshold should be around 0.5. Meanwhile, $\mathcal{L}_{cbce}$ gives a higher penalty on positive pixels (due to its scarcity) making the resulted optimal threshold at a relatively higher value. By contrast, $\mathcal{L}_{jac}$ pushes the foreground pixels to the probability of 1.0 while remaining to be strongly discriminative against the background pixels. Thus, the plateau around the optimal segmentation performance of $\mathcal{L}_{jac}$ would be much wider than $\mathcal{L}_{ce}$ and $\mathcal{L}_{cbce}$ so that it could perform stably in a wide range of thresholds, i.e., [0.05, 0.95] in our experiments.

Table 1.2 Comparision of loss functions: $\mathcal{L}_{ce}$, $\mathcal{L}_{cbce}$, and $\mathcal{L}_{jac}$ with P-Net

	Loss	mean ± stdv. [min, max]	
		Dice score (%)	Jaccard index (%)
CT	$\mathcal{L}_{ce}$	83.5 ± 5.6 [59.3, 91.1]	72.0 ± 7.70 [42.2, 83.6]
	$\mathcal{L}_{cbce}$	83.2 ± 5.7 [57.2, 90.3]	71.6 ± 7.80 [40.1, 82.4]
	$\mathcal{L}_{jac}$	**83.7 ± 5.4 [58.4, 90.4]**	**72.3 ± 7.50 [41.3, 82.4]**
MRI	$\mathcal{L}_{ce}$	80.0 ± 7.60 [50.7, 89.9]	67.3 ± 9.80 [34.0, 81.6]
	$\mathcal{L}_{cbce}$	**80.2 ± 7.20 [53.6, 90.5]**	**67.6 ± 9.50 [36.6, 82.7]**
	$\mathcal{L}_{jac}$	**80.2 ± 7.90 [51.2, 90.1]**	**67.6 ± 10.3 [34.4, 82.0]**

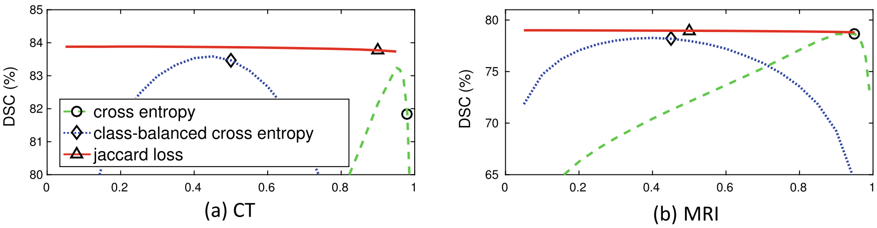

Fig. 1.4 Plot of the threshold versus Dice score (DSC): the proposed jaccard loss $\mathcal{L}_{jac}$ performs the steadiest across thresholds in the range of [0.05, 0.95] comparing to the cross-entropy loss $\mathcal{L}_{ce}$ and the class-balanced cross-entropy loss $\mathcal{L}_{cbce}$. The threshold that selected from validation dataset is marked as ○, ◇, and △ for losses $\mathcal{L}_{ce}$, $\mathcal{L}_{cbce}$, and $\mathcal{L}_{jac}$, respectively

1.3 Recurrent Neural Network for Contextual Learning

Previous work [2, 20, 30] perform deep 2D CNN segmentation on CT (or MRI) axial slices independently, not taking the correlation between neighboring images into consideration. Organ segmentation in 3D volumes can also be performed by directly taking cropped 3D sub-volumes as input to 3D CNNs [10, 14, 15]. However, even at the expense of being computationally expensive and prone-to-overfitting [30], the result of very high segmentation accuracy has not been reported for complexly shaped organs [14], or small anatomical structures [10]. Despite more demanding memory requirement, 3D CNN approaches deserve more investigation for future work. On the other hand, [3, 25] use hybrid CNN-RNN architectures to process/segment sliced CT (or MRI) images in sequence and present a promising direction to process CT and MRI segmentations. However, these methods do not apply spatial shape continuity constrain or regularization to enforce the segmentation consistency among successive slices. Thus, in this chapter, we present our own research for regulating pancreas segmentation across 3D slices with recurrent neural network.

1.4 Recurrent Neural Network

As discussed above, the P-Net processes pancreas segmentation with individual 2D image slices, delivering remarkable performance on the tested CT and MRI datasets. However, as shown in the first row of Fig. 1.5, the transition among the resulting CNN pancreas segmentation regions in the consecutive slices may not be smooth which often implies boarder failures of segmentations. Adjacent CT/MRI slices are expected to be correlated with each other thus segmentation results from successive slices need to be constrained for shape continuity. The model for 3D object segmentation is required to be able to detect and recover abnormally lost part inside slices (see $\hat{Y}_\tau$ in Fig. 1.5).

To achieve this, we concatenate a recurrent neural network (RNN) subnetwork to the output end of P-Net for modeling inter-slice shape continuity and regularization. The RNN is originally designed for sequential data analysis and thus naturally meets the need of processing the ordered image slices. Specifically, we slice the CT (or MRI) volume into an ordered sequence of 2D images and process to learn the segmentation shape continuity among neighboring image slices with a typical RNN architecture, the long short-term memory (LSTM) unit. However, the standard LSTM requires vectorized input, which will sacrifice the spatial information encoded in the output of CNN. To circumvent such problem, we utilize the convolutional-LSTM (C-LSTM) model [23] to preserve the 2D image segmentation layout by CNN. As shown in Fig. 1.5, H_τ and C_τ are the hidden state and cell output of C-LSTM in respective at the τth slice. The current cell output C_τ is computed based on both of the former cell hidden state $H_{\tau-1}$ and the current CNN output $\hat{Y}_\tau$. Then, H_τ will be calculated from C_τ and used to produce the next cell output $C_{\tau+1}$. Contextual information is propagated from slice τ to $\tau + 1$ through convolutional operations.

The strategy of the C-LSTM based context continuity learning is built upon an intuitive conditional probability assumption. Segmentation results of the former image slices are encoded in the cell hidden state $H_{\tau-1}$. Values of C_τ is decided by taking $H_{\tau-1}$ and $\hat{Y}_\tau$ together into consideration. If position p_i in $\hat{Y}_\tau$ is predicted as pancreatic tissue by the CNN model (e.g. P-Net), and the same position in $H_{\tau-1}$ are also encoded as pancreatic tissue, then with high confidence that position p_i should be a pancreatic pixel in C_τ, and vice versa. As a result, C-LSTM not only recovers missing segmentation parts but also outputs more confident probability maps than the original CNN subnetwork.

Formally, the C-LSTM unit is formulated,

$$i_\tau = \sigma(W_{yi} * \hat{Y}_\tau + W_{hi} * h_{\tau-1} + W_{ci} \odot c_{\tau-1} + b_i), \tag{1.13}$$

$$f_\tau = \sigma(W_{yf} * \hat{Y}_\tau + W_{hf} * h_{\tau-1} + W_{cf} \odot c_{\tau-1} + b_f), \tag{1.14}$$

$$c_\tau = f_\tau \odot c_{\tau-1} + i_\tau \odot \tanh(W_{yc} * \hat{Y}_\tau + W_{hc} * h_{\tau-1} + b_c), \tag{1.15}$$

$$o_\tau = \sigma(W_{yo} * \hat{Y}_\tau + W_{ho} * h_{\tau-1} + W_{co} \odot c_\tau + b_o), \tag{1.16}$$

$$h_\tau = o_\tau \odot \tanh(c_\tau), \tag{1.17}$$

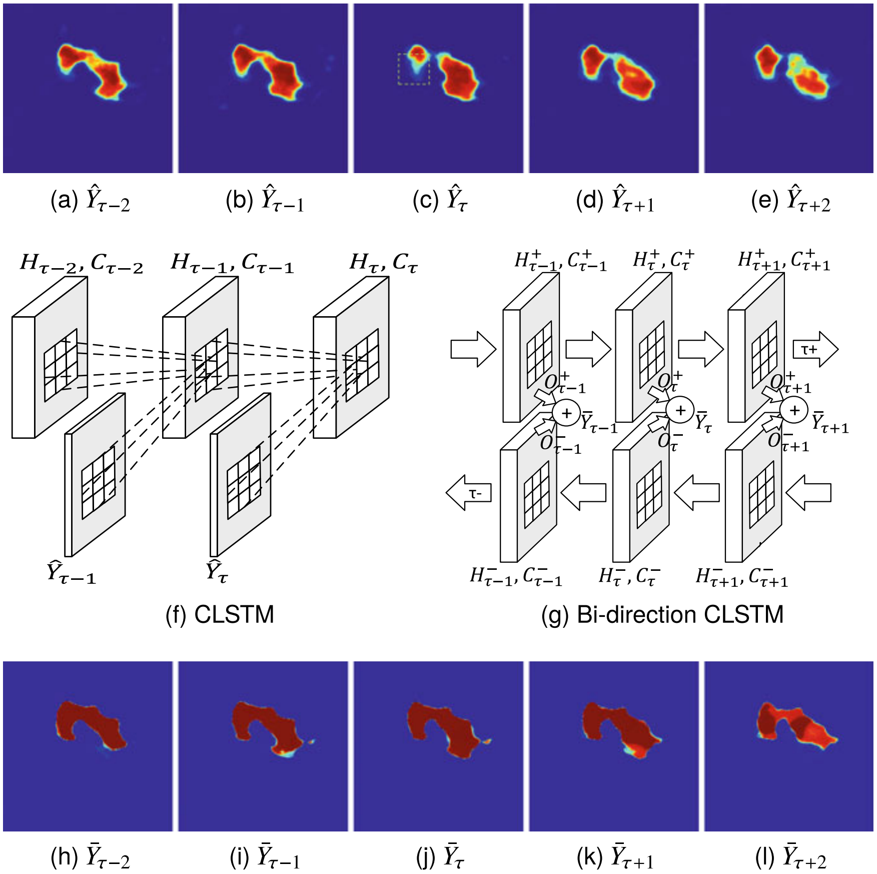

Fig. 1.5 The main construction units of the proposed RNN model and its input/output segmentation sequence. The sequence of CNN outputs is shown in the first row (**a**–**e**), is taken as the input of the bidirectional C-LSTM (**g**), which is an RNN architecture composed of two layers of C-LSTM (**f**) working in opposite directions. The third row (**h**–**l**) presents the corresponding output sequence, which is sharp and clean. Note that the missing pancreatic part in $\hat{Y}_\tau$ (**c**), in the green dashed box, is recovered by shape continuity modeling in $\bar{Y}_\tau$ (**j**). For visual clarity, we omit the input $\hat{Y}_{(\cdot)}$ in the bidirectional CLSTM (**g**), which is same as in (**f**)

where $*$ represents convolution operation, and $\odot$ denotes the Hadamard product. Gates i_τ, f_τ, o_τ are the input, forget, and output, respectively, following the original definition of C-LSTM. $W_{(\cdot)}$, and $b_{(\cdot)}$ are weights and bias in the corresponding C-LSTM unit that needs model optimization. Finally, $\sigma(\cdot)$ and $\tanh(\cdot)$ denote the sigmoid and hyperbolic tangent activation function, respectively.

1.4.1 Bidirectional Contextual Regularization

Next, we have the contextual learning extended to a bidirectional. For pancreas, as well as other organs, its shape in the current slice is constrained by slices from not only its former slices but also the followings. The contextual information to the input could be doubled if the shape regularization is taken in both directions leading to a further improvement. Two layers of C-LSTM are stacked working in two opposite directions as shown in Fig. 1.5. Then, outputs of the two layers, one in the τ^--direction and the other in the τ^+-direction, are combined as the final segmentation output,

$$\bar{Y}_\tau = \sum_{i \in \{\tau^-, \tau^+\}} \lambda^i o^i_\tau, \tag{1.18}$$

where i represents the τ^- and τ^+ directions, and λ^i is the learned weights when combining CLSTM outputs from both directions. Thus, the bidirectional design of shape continuity modeling permits to explicitly enforce the pancreas segmentation to be spatial smooth and higher order inter-slice regularized.

Lastly, we define the objective of contextual learning based on Jaccard loss as

$$\mathcal{L}_{rnn} = \sum_{\tau=1}^{T} \mathcal{L}_{jac}(\bar{Y}_\tau, Y), \tag{1.19}$$

where T is the length of image sequence processed in each unit, and we empirically set $T = 5$ in our experiments.

1.4.2 Experimental Results

Given outputs of P-Net as the best CNN-based segmentation results, the bidirectional RNN (BiRNN) subnetwork is then stacked to the output end of P-Net and trained end to end. In each direction, a one-layer CLSTM is implemented with one hidden state and 3×3 convolution filter kernels [23]. Particularly, the number of hidden state is set to 1 since our shape continuity learning is inherently simplified by processing only the output probability maps of CNN subnetwork. CNN output $\hat{Y}_\tau \in R^{d_1^1 \times d_1^2 \times 1}$, where d_1^1 and d_1^2 are the width and height of the input image, provides a highly compacted representation of the input CT/MRI image for shape learning. Thus, BiRNN with the hidden state $H_\tau \in R^{d_1^1 \times d_1^2 \times 1}$ is sufficient to model and capture the shape continuity regularization among CNN outputs. We notice that BiRNN cannot converge stably during model training when a larger hidden state is used. In addition, we attempt to employ BiRNN on the feature maps from CNN's intermediate layers. However, this causes the model training process failed to converge. Thus, we mainly focus on the current design of BiRNN, which emphasizes to learn the inter-slice shape continuity among the successive segmentation outputs of P-Net.

Table 1.3 Evaluate pancreas segmentation on the CT dataset. BiRNN refines output of P-Net providing better performance in both volume measurement (DSC) and surface reconstruction (HD)

Method	HD (mm)	DSC (%)
P-Net	0.61 ± 0.53 [0.15, 3.48]	83.3 ± 5.6 [59.0, 91.0]
BiRNN	**0.54 ± 0.53 [0.12, 3.78]**	**83.7 ± 5.1 [59.0, 91.0]**

We model the segmentation shape continuity as a higher order inter-slice regularization among the CT/MRI axial slices. The average physical slice thickness in CT and MRI are 2 mm and 7 mm, respectively. Thus, slight shape change occurs between two correct segmented slices. Given the measured Hausdorff distance (HD) of neighboring slices in ground truth, the mean ± standard deviation of shape changes in CT and MRI are 0.35 ± 0.94 mm and 2.69 ± 4.45 mm, respectively. The drastic shape changes in MRI volumes indicates that successive MRI image slices are actually more independent, so that in our implementation, shape continuity learning brings marginal but consistent performance improvement. The improvement in CT images is more evident. Specifically, we detect abnormal shape changes in the outputs of CNN and have them refined by BiRNN. We define abnormal shape change occurs between two neighboring CT when $HD\ (\hat{Y}_\tau, \hat{Y}_{\tau-1}) > 0.5$ mm, which is decided basing on the shape change statics in the CT dataset.

Table 1.3 illustrates performance with and without shape continuity learning, where BiRNN boost volume segmentation (i.e., DSC) by 0.4%. More importantly, the error for pancreatic surface reconstruction (i.e., HD) drops from 0.61 to 0.54 mm, improved by 11.5%. Figure 1.7 further shows the segmentation performance difference statistics, with or without contextual learning in subject-wise. In particular, those cases with low DSC scores are greatly improved by BiRNN.

Finally, Fig. 1.6 displays examples of output probability maps from all of the comparative methods, i.e., HNN [28], U-Net [18], P-Net and *P-Net+BiRNN*, where the latter one delivers the sharpest and clearest output on both CT and MRI datasets. More specifically, P-Net presents detailed results that recover the major part of the pancreas, where both HNN and U-Net suffer from significant lower segmentation recall. When observing the BiRNN outputs for CT and MRI, we find detailed pancreas parts in CT have been recovered via shape continuity learning and regularization, while in MRI, the BiRNN only outputs probability map with the same shape in P-Net's output, which is optimal when the inter-slice shape changes drastically in the MRI dataset. Thus, BiRNN would help refine pancreas segmentation with a smoothed surface in the situation that slice thickness of the 3D scans is reasonably small, e.g., <2 mm.

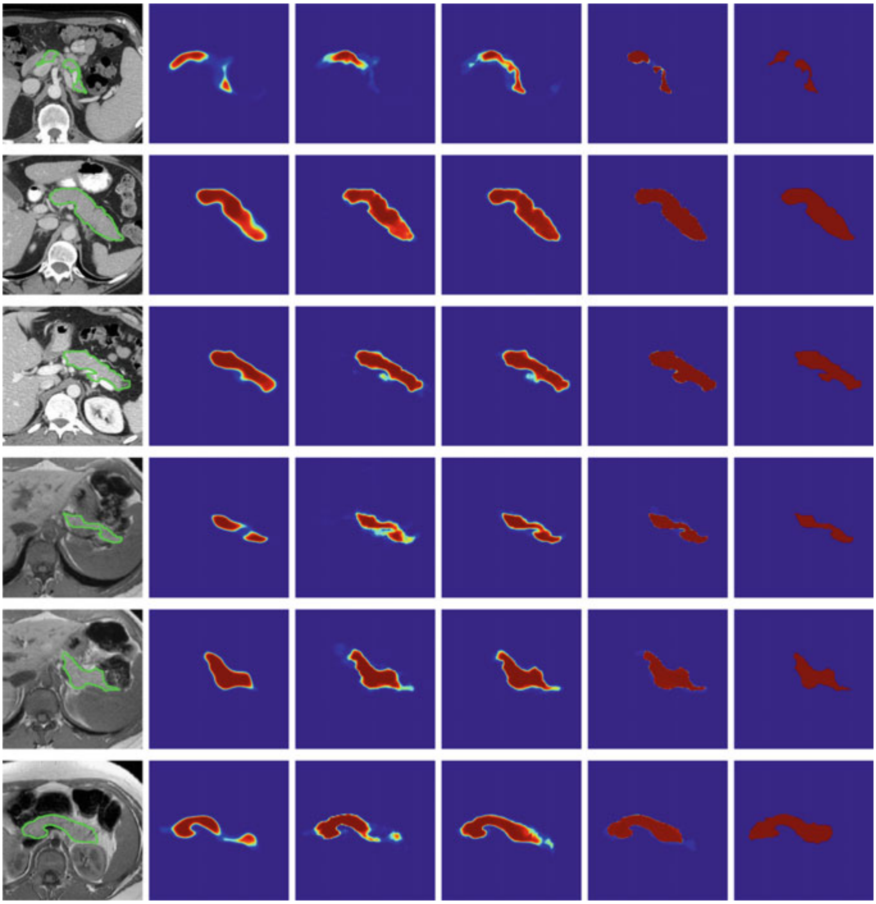

Fig. 1.6 Examples of output probability map: columns from left to right are the input CT/MRI image and results from HNN [28], U-Net [18], P-Net, and the full CNN-RNN (*P-Net+BiRNN*) model, and the ground truth. The CNN-RNN model delivers the most clear probability maps which preserve detailed pancreatic boundaries

1.5 State-of-the-Art Methods for Pancreas Segmentation

We compare selected state-of-the-art methods for pancreas segmentation. Dice score and Jaccard index are computed and reported in Table 1.4 under fourfold CV. The method *P-Net+BiRNN* performs the best on the CT dataset and P-Net achieves the best result on the MRI dataset. We notice that the current implementation of FCN 3D [22] is not as effective as its 2D segmentation counterparts, where *P-Net+BiRNN* outperforms FCN 3D by a large margin of 6.9% Dice score. The problem of segmenting 3D CT/MRI image volumes within a single inference is much more complex than the 2D CNN approaches where further network architecture exploration as well as

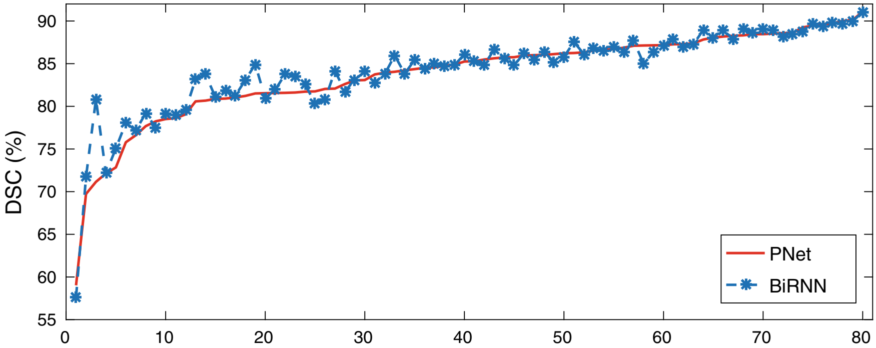

Fig. 1.7 Comparison of P-Net and *P-Net+ BiRNN* outputs for all 80 NIH-CT scans and the scans are sorted left to right using Dice scores of P-Net. Small fluctuations among the well- segmented cases (on the top right) possibly result from model updating, which can be omitted as noise

Table 1.4 Performance of the state-of-the-art methods for segmentation under fourfold CV. We show Dice score and Jaccard index in the form of *mean* ± *standard dev. [worst case, best case]*. The best results on CT and MRI are highlighted in bold

	Method	Dice score (%)	Jaccard index (%)
CT	3D FCN [22]	76.8 ± 9.4 [43.7, 89.4]	
	Roth et al. [20]	78.0 ± 8.2 [34.1, 88.6]	
	Roth et al. [21]	81.3 ± 6.3 [50.6, 88.9]	68.8 ± 8.12 [33.9, 80.1]
	Coarse-to-Fine [30]	82.3 ± 5.6 [62.4, 90.8]	
	CNN+RNN [1]	82.4 ± 6.7 [60.0, 90.1]	70.6 ± 9.00 [42.9, 81.9]
	P-Net	83.3 ± 5.6 [59.0, 91.0]	71.8 ± 7.70 [41.8, 83.5]
	P-Net+BiRNN	**83.7 ± 5.1 [59.0, 91.0]**	**72.3 ± 7.04 [41.8, 83.5]**
MRI	Graph-Fusion [2]	76.1 ± 8.70 [47.4, 87.1]	
	CNN+RNN [1]	80.5 ± 6.70 [59.1, 89.4]	67.9 ± 8.90 [41.9, 80.9]
	P-Net	**80.7 ± 7.40 [48.8, 90.5]**	**68.2 ± 9.64 [32.3, 82.7]**

more training images are typically required. This is also referred as *curse of dimensionality* in [19, 30]. In this scenario, we would argue that 2D network architectures may still be optimal for pancreas segmentation with large inter-slice thicknesses. We also note that our intuition of developing CNN-RNN combination is orthogonal to the principles of *coarse-to-fine* pancreas location and detection [21, 30]. Better performance may be achievable with the combination of both methodologies. Figure 1.8 visually depicts examples of reconstructed 3D segmentation results from the CT dataset.

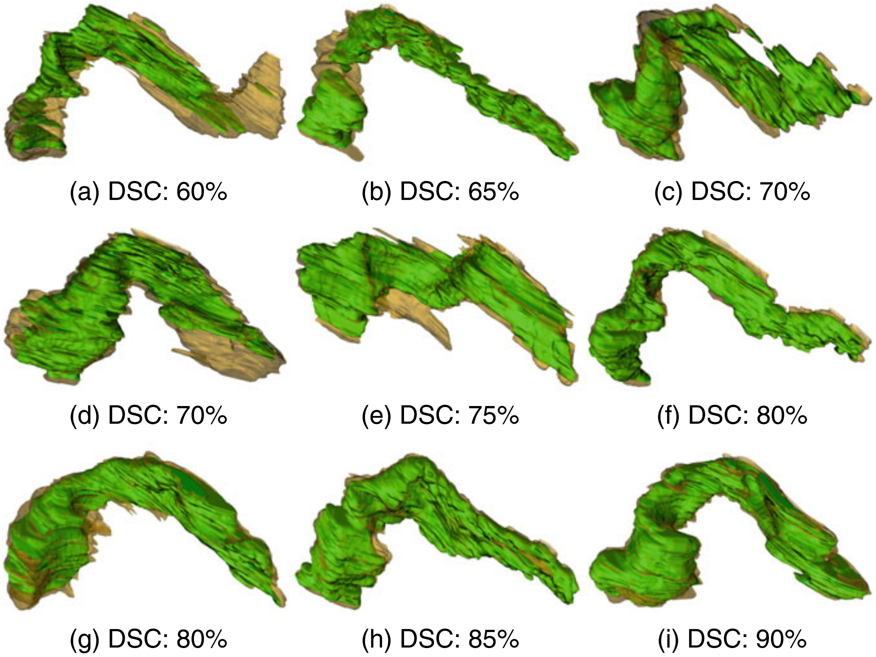
(a) DSC: 60%
(b) DSC: 65%
(c) DSC: 70%
(d) DSC: 70%
(e) DSC: 75%
(f) DSC: 80%
(g) DSC: 80%
(h) DSC: 85%
(i) DSC: 90%

Fig. 1.8 3D visualization of pancreas segmentation results where human annotation shown in yellow and computerized segmentation displayed in green. The DSC are 90%, 75%, and 60% for three examples from left to right, respectively

1.6 Summary

In this chapter, we present a novel CNN-RNN architecture for pancreas segmentation in CT and MRI scans via our tailor-made CNN module (P-Net) followed by a bidirectional C-LSTM (BiRNN). It is presented to regularize the segmentation results on individual image slices. The shape continuity regularization permits to enforce the pancreas segmentation spatial smoothness explicitly in the axial direction, in analogy to comprehending into videos by parsing and aggregating successive frames [16]. This may also share some similarity to the human doctor's way of reading radiology images. Combined with the proposed Jaccard loss function for model training to generate the threshold-free segmentation results, our quantitative pancreas segmentation result outperforms the previous state-of-the-art approaches [2, 20, 21, 30] on both CT and MRI datasets, with noticeable margins. Although the discussion focuses on pancreas segmentation in this chapter, the approaches would be generalizable to other organ segmentations in medical image analysis.

References

1. Cai J, Lu L, Xie Y, Xing F, Yang L (2017) Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function. In: MICCAI, pp 674–682. Springer
2. 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, pp 442–450. Springer
3. Chen J, Yang L, Zhang Y, Alber MS, Chen DZ (2016) Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: NIPS, pp 3036–3044
4. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI 40(4):834–848
5. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI, pp 424–432. Springer
6. Clark KW, Vendt BA, Smith KE, Freymann JB, Kirby JS, Koppel P, Moore SM, Phillips SR, Maffitt DR, Pringle M, Tarbox L, Prior FW (2013) The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057
7. Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE computer society conference on computer vision and pattern recognition (CVPR 2009), 20–25 June 2009, Miami, Florida, USA, pp 248–255
8. Farag A, Lu L, Roth HR, Liu J, Turkbey E, Summers RM (2017) A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. TMI 26(1):386–399
9. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: AISTATS, pp 249–256
10. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. MIA 36:61–78
11. 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. MIA 39:18–28
12. Lee C, Xie S, Gallagher PW, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: AISTATS
13. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: CVPR, pp 3431–3440
14. Merkow J, Marsden A, Kriegman DJ, Tu Z (2016) Dense volume-to-volume vascular boundary detection. In: MICCAI, pp 371–379
15. Milletari F, Navab N, Ahmadi S (2016) V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International conference on 3D vision, pp 565–571
16. Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: ICCV, pp 4694–4702
17. Oda M, Shimizu N, Karasawa K, Nimura Y, Kitasaka T, Misawa K, Rueckert D, Mori K (2016) Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In: MICCAI, pp 556–563. Springer
18. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: MICCAI, pp 234–241
19. Roth HR, Lu L, Farag A, Shin HC, Liu J, Turkbey EB, Summers RM (2015) Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI, pp 556–564. Springer
20. Roth HR, Lu L, Farag A, Sohn A, Summers RM (2016) Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: MICCAI, pp 450–451. Springer
21. Roth HR, Lu L, Lay N, Harrison AP, Farag A, Summers RM (2018) Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. MIA 45:94–107

22. Rotha HR, Odaa H, Zhoub X, Shimizua N, Yanga Y, Hayashia Y, Odaa M, Fujiwarac M, Misawad K, Moria K (2018) An application of cascaded 3D fully convolutional networks for medical image segmentation. ArXiv e-prints
23. Shi X, Chen Z, Wang H, Yeung D, Wong W, Woo W (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: NIPS, pp 802–810
24. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations, pp 1–14
25. Stollenga MF, Byeon W, Liwicki M, Schmidhuber J (2015) Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. In: NIPS, pp 2998–3006
26. Tong T, Wolz R, Wang Z, Gao Q, Misawa K, Fujiwara M, Mori K, Hajnal JV, Rueckert D (2015) Discriminative dictionary learning for abdominal multi-organ segmentation. MIA 23(1):92–104
27. 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
28. Xie S, Tu Z (2015) Holistically-nested edge detection. In: ICCV, pp 1395–1403
29. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr PH (2015) Conditional random fields as recurrent neural networks. In: ICCV, pp 1529–1537
30. Zhou Y, Xie L, Shen W, Fishman E, Yuille AL (2016) Pancreas segmentation in abdominal CT scan: a coarse-to-fine approach. http://arxiv.org/abs/1612.08230