Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function

Jinzheng Cai[1], Le Lu[3], Yuanpu Xie[1], Fuyong Xing[2], and Lin Yang[1,2(✉)]

[1] Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA
lin.yang@bme.ufl.edu
[2] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA
[3] Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA

Abstract. Deep neural networks have demonstrated very promising performance on accurate segmentation of challenging organs (*e.g.,* pancreas) in abdominal CT and MRI scans. The current deep learning approaches conduct pancreas segmentation by processing sequences of 2D image slices independently through deep, dense per-pixel masking for each image, without explicitly enforcing spatial consistency constraint on segmentation of successive slices. We propose a new convolutional/recurrent neural network architecture to address the contextual learning and segmentation consistency problem. A deep convolutional sub-network is first designed and pre-trained from scratch. The output layer of this network module is then connected to recurrent layers and can be fine-tuned for contextual learning, in an end-to-end manner. Our recurrent sub-network is a type of Long short-term memory (LSTM) network that performs segmentation on an image by integrating its neighboring slice segmentation predictions, in the form of a dependent sequence processing. Additionally, a novel segmentation-direct loss function (named Jaccard Loss) is proposed and deep networks are trained to optimize Jaccard Index (JI) directly. Extensive experiments are conducted to validate our proposed deep models, on quantitative pancreas segmentation using both CT and MRI scans. Our method outperforms the state-of-the-art work on CT [11] and MRI pancreas segmentation [1], respectively.

1 Introduction

Detecting unusual volume changes and monitoring abnormal growths in pancreas using medical images is a critical yet challenging diagnosis task. This would require to dissect pancreas from its surrounding tissues in radiology images (*e.g.,* CT and MRI scans). Manual pancreas segmentation is laborious, tedious, and sometimes prone to inter-observer variability. One major group of related work

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 674–682, 2017.
DOI: 10.1007/978-3-319-66179-7_77

on automatic pancreas segmentation in CT images are based on multi-atlas registration and label fusion (MALF) [8,15,16] under leave-one-patient-out evaluation protocol. Due to the high deformable shape and vague boundaries of pancreas in CT, their reported segmentation accuracy results (measured in Dice Similarity Coefficient or DSC) range from 69.6 ± 16.7% [16] to 75.1 ± 15.4% [8]. On the other hand, deep convolutional neural networks (CNN) based pancreas segmentation work [1,3,10–12,18] have revealed promising results and steady performance improvements, e.g., from 71.8 ± 10.7% [10], 78.0 ± 8.2% [11], to 81.3 ± 6.3% [12] evaluated using the same NIH 82-patient CT dataset https://doi.org/10.7937/K9/TCIA.2016.TNB1KQBU.

In comparison, deep CNN approaches appear to demonstrate the noticeably higher segmentation accuracy and numerically more stable results (significantly lower in standard deviation, or std) than their MALF counterparts. [11,12] are built upon the fully convolutional network (FCN) architecture [5] and its variant [17]. However, [11,12] are not completely end-to-end trained due to their segmentation post processing steps. Consequently, the trained models may be suboptimal. For pancreas segmentation on a 79-patient MRI dataset, [1] achieves 76.1 ± 8.7% in DSC.

In this paper, we propose a new deep neural network architecture with recurrent neural contextual learning for improved pancreas segmentation. All previous work [1,11,18] perform deep 2D CNN segmentation on either CT or MRI image or slice independently[1]. There is no spatial smoothness consistency constraints enforced among successive slices. We first follow this protocol by training 2D slice based CNN models for pancreas segmentation. Once this step of CNN training converges, inspired by sequence modeling for precipitation nowcasting in [13], a convolutional long short-term memory (CLSTM) network is further added to the output layer of the deep CNN to explicitly capture and constrain the contextual segmentation smoothness across neighboring image slices. Then the whole integrated CLSTM network can be end-to-end fine-tuned via stochastic gradient descent (SGD) until converges. The CLSTM module will modify the segmentation results produced formerly by CNN alone, by taking the initial CNN segmentation results of successive axial slices (in either superior or interior direction) into account. Therefore the final segmented pancreas shape is constrained to be consistent among adjacent slices, as a good trade-off between 2D and 3D segmentation deep models.

Next, we present a novel segmentation-direct loss function to train our CNN models by minimizing the jaccard index between any annotated pancreas mask and its corresponding output segmentation mask. The standard practice in FCN image segmentation deep models [1,5,11,17] use a loss function to sum up the cross-entropy loss at each voxel or pixel. Segmentation-direct loss function can

[1] Organ segmentation in 3D CT and MRI scans can also be performed by directly taking cropped 3D sub-volumes as input [4,6,7]. Even at the expense of being computationally expensive and prone-to-overfitting, the result of very high segmentation accuracy has not been reported for complexly shaped organs [6]. [2,14] use hybrid CNN-RNN architectures to process/segment sliced CT or MRI images in sequence.

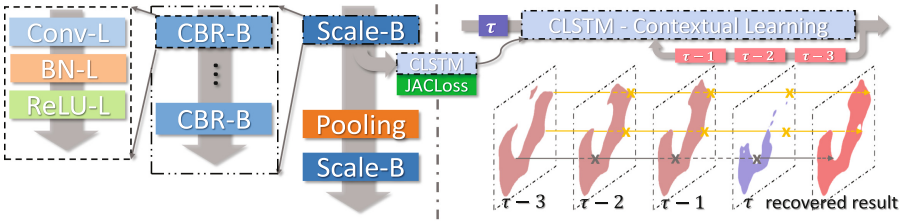

Fig. 1. Network architecture: Left is the CBR block (CBR-B) that contains convolutional layer (Conv-L), batch normalization layer (BN-L), and ReLU layer (ReLU-L). While, each scale block (Scale-B) has several CBR blocks and followed with a pooling layer. Right is the CLSTM for contextual learning. Segmented outcome at slice τ would be regularized by the results of slice $\tau - 3$, $\tau - 2$, and $\tau - 1$. For example, contextual learning is activated in regions with $\times$ markers, where sudden losses of pancreas areas occurs in slice τ comparing to consecutive slices.

avoid the data balancing issue during CNN training between the positive pancreas and negative background regions. Pancreas normally only occupies a very small fraction on each slice. Furthermore, there is no need to calibrate the optimal probability threshold to achieve the best possible binary pancreas segmentation results from the FCN's probabilistic outputs [1,5,11,17]. Similar segmentation metric based loss functions based on DSC are concurrently proposed and investigated in [7,18].

We extensively and quantitatively evaluate our proposed deep convolutional LSTM neural network pancreas segmentation model and its ablated variants using both a CT (82 patients) and one MRI (79 patients) dataset, under 4-fold cross-validation (CV). Our complete model outperforms 4% of DSC comparing to previous state-of-the-arts [1,11]. Although our contextual learning model is only tested on pancreas segmentation, the approach is directly generalizable to other three dimensional organ segmentation tasks.

2 Method

Simplifying Deep CNN Architecture: We propose to train deep CNN network from scratch and empirically observe that, for the specific application of pancreas segmentation in CT/MRI images, ImageNet pre-trained CNN models do not noticeably improve the performance. More importantly, we design our CNN network architecture specifically for pancreas segmentation where a much smaller CNN model than the conventional models [5,17]is found to be most effective. This model reduces the chance of over-fitting (against small-sized medical image datasets) and can speed up both training and inference. Our specialized deep network architecture is trained from scratch using pancreas segmentation datasets, without being first pre-trained using ImageNet [5,17] and then fine-tuned. It also outperforms the ImageNet fine-tuned conventional CNN models [5,17] from our empirical evaluation.

First, *Convolutional* layer is followed by *ReLU* and *Batch normalization* layers to form the basic unit of our customized network, namely the CBR block. Second, following the deep supervision principle proposed in [17], we stack several CBR blocks together with an auxiliary loss branch per block and denote this combination as Scale block. Figure 1 shows exemplar CBR block (CBR-B) and Scale block (Scale-B). Third, we use CBR block and Scale block as the building blocks to construct our tailored deep network, with each Scale block is followed with a pooling layer. Hyper parameters of the numbers of feature maps in convolutional layers, the number of CBR blocks in a Scale block, as well as the number of Scale blocks to fit into our network can be determined via a model selection process on a subset of training dataset (i.e., split for model validation).

2.1 Contextual Regularization

From above, we have designed a compact CNN architecture which can process pancreas segmentation on individual 2D image slices. However as shown in the first row of Fig. 2, the transition among the resulted CNN pancreas segmentation regions in consecutive slices may not be smooth, often implying that segmentation failure occurs. Adjacent CT/MRI slices are expected to be correlated to each other thus segmentation results from successive slices need to be constrained for shape consistence.

To achieve this, we concatenate long short-term memory (LSTM) network to the 2D CNN model for contextual learning, as a compelling architecture for sequential data processing. That is, we slice any 3D CT (or MRI) volume into a 2D image sequence and process to learn the segmentation contextual constraints among neighboring image slices with LSTM. Standard LSTM network requires the vectorized input which would sacrifice the spatial information encoded in the output of CNN. We therefore utilize the convolutional LSTM (CLSTM) model [13] to preserve the 2D image segmentation layout by CNN. The second row of Fig. 2 illustrates the improvement by enforcing CLSTM based segmentation contextual learning.

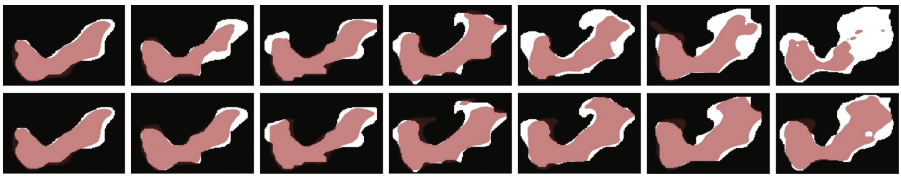

Fig. 2. NIH Case51: segmentation results with and without contextual learning are displayed in row 1 and row 2, respectively. Golden standards are displayed in white, and automatic outputs are rendered in red.

2.2 Jaccard Loss

We propose a new jaccard loss (JACLoss) for training neural network image segmentation model. To optimize JI (a main segmentation metric) directly in network training makes the learning and inference procedures consistent and generate threshold-free segmentation. JACLoss is defined as follows:

$$L_{jac} = 1 - \frac{|Y_+ \bigcap \hat{Y}_+|}{|Y_+ \bigcup \hat{Y}_+|} = 1 - \frac{\sum_{j \in Y} y_j \wedge \hat{y}_j}{\sum_{j \in Y} y_j \vee \hat{y}_j} = 1 - \frac{\sum_{f \in Y_+} (1 \wedge \hat{y}_f)}{|Y_+| + \sum_{b \in Y_-} (0 \vee \hat{y}_b)} \quad (1)$$

where Y and $\hat{Y}$ represent the ground truth and network predictions. Respectively, we have Y_+ and Y_- defined as the foreground pixel set and the background pixel set, and $|Y_+|$ is the cardinality of Y_+. Similar definitions are also applied to $\hat{Y}$. y_j and $\hat{y}_j \in \{0, 1\}$ are indexed pixel values in Y and $\hat{Y}$. In practice, $\hat{y}_j$ is relaxed to the probability number in range $[0, 1]$ so that JACLoss can be approximated by

$$\tilde{L}_{jac} = 1 - \frac{\sum_{f \in Y_+} \min(1, \hat{y}_f)}{|Y_+| + \sum_{b \in Y_-} \max(0, \hat{y}_b)} = 1 - \frac{\sum_{f \in Y_+} \hat{y}_f}{|Y_+| + \sum_{b \in Y_-} \hat{y}_b} \quad (2)$$

Obviously, L_{jac} and $\tilde{L}_{jac}$ are sharing the same optimal solution of $\hat{Y}$, with slight abuse of notation, we use L_{jac} to denote both. The model is updated by:

$$\frac{\partial L_{jac}}{\partial \hat{y}_j} = \begin{cases} -\frac{1}{|Y_+| + \sum_{b \in Y_-} \hat{y}_b}, \text{for } j \in Y_+ \\ -\frac{\sum_{f \in Y_+} \hat{y}_f}{(|Y_+| + \sum_{b \in Y_-} \hat{y}_b)^2}, \text{for } j \in Y_- \end{cases} \quad (3)$$

Since the inequality $\sum_{f \in Y_+} \hat{y}_f < (|Y_+| + \sum_{b \in Y_-} \hat{y}_b)$ holds by definition, the JACLoss assigns larger gradients to foreground pixels that intrinsically balances the foreground and background classes. It is empirically works better than the cross-entropy loss or the classed balanced cross-entropy loss [17] when segmenting small objects, such as pancreas in CT/MRI images. Similar loss functions are independently proposed and utilized in [7,18].

3 Experimental Results and Analysis

Datasets: Two annotated pancreas datasets are utilized for experiments. The first NIH-CT-82 dataset [10,11] is publicly available and contains 82 abdominal contrast-enhanced 3D CT scans. We obtain the second dataset UFL-MRI-79 from [1], with 79 abdominal T1-weighted MRI scans acquired under multiple controlled-breath protocol. For the case of comparison, 4-fold cross validation is conducted similar to [1,10,11]. Unlike [11], no sophisticated post processing is employed. We measure the quantitative segmentation results using dice similarity coefficient (DSC): $DSC = 2(|Y_+ \cap \hat{Y}_+|)/(|Y_+| + |\hat{Y}_+|)$, and jaccard index (JI): $JI = (|Y_+ \cap \hat{Y}_+|)/(|Y_+ \cup \hat{Y}_+|)$.

Network Implementation: Hyper-parameters are determined via model selection inside training dataset. The network that contains five Scale blocks with four CBR blocks in each Scale block produces the best empirical performance, while remaining with the compact model size (<3 million parameters). Training folds are first split into a training subset for network parameter training and a validation subset for learning hyper-parameters. Note the training accuracy as Acc_t after model selection. We then combine training and validation subsets to further fine-tune the network until its performance on validation subset converges to Acc_t. The average time for model training is ∼3 h on a single *GeForce GTX TITAN*.

Analysis on Contextual Regularization: We evaluate the proposed neural network architectures with and without contextual learning on both CT and MRI datasets. We first train a network of five Scale blocks with JACLoss. The number of output feature channels per convolutional layer is set as 64, and we name this network JAC-64. CLSTM contextual regularization is then applied on JAC-64's five Scale block outputs, forming a new extension of RNN-64. RNN-64 is initialized from JAC-64 and trained with enough SGD updates until convergence. We next investigate the performance impact on increasing the convolutional output channels from 64 to 128. Similarly, JAC-128 and RNN-128 are used to denote this variant and its contextually regularized version, respectively. From Table 1, RNN-enhanced deep models improve upon JAC-64/JAC-128 by 2.0% and 0.9% in mean DSC on NIH-CT-82. For UFL-MRI-79, RNN-64 achieves 1.8% mean DSC gain against JAC-64. RNN-128 and JAC-128 produce the best segmentation results comparably. Figure 3 further shows the segmentation performance difference statistics, with or without contextual learning. Especially, these cases with low DSC scores are greatly improved by contextual learning.

Analysis on Jaccard Loss: Figure 4 represents quantitative segmentation results of the three losses under 4-fold CV. JACLoss achieves the highest mean DSC, regardless of different segmentation thresholds. FCN or HNN outputs probabilistic image segmentation maps instead of binary masks. Thus an appropriate probability threshold is required to obtain the final binary segmentation outcomes. Naïve cross-entropy loss assigns the same penalty on positive and negative pixels so the probability threshold should be around 0.5. Its class-balanced version gives higher penalty scores on positive pixels (due to its scarcity), making the resulted "optimal" threshold at a relatively higher value. In contrast, JACLoss can push the foreground pixels to the probability of 1 while remains being strongly discriminative against background pixels.

Comparison with the State-of-the-Art Methods: Last, we compare our pancreas segmentation models (as trained above) with the state-of-the-art methods. Holistically-nested network [17] (HNN) is a CNN architecture that is originally proposed for semantic edge detection. HNN has been adapted for pancreas segmentation in [11] and proved with good performance. We also implement UNet [9] for universal medical image segmentation problems. As a reference that **HNN and UNet contains 10 times more parameters than**

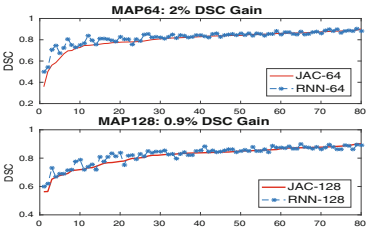

Fig. 3. 80 cases with/without contextual learning, and sorted left to right by DSC values of JAC-models with no contextual learning. Small fluctuations among good cases are normally resulted from model updating.

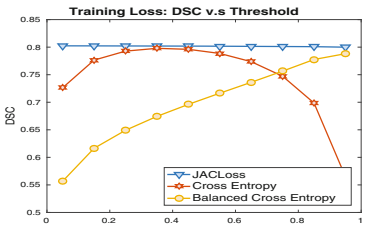

Fig. 4. Thresholded results of models that are trained with different loss functions. The proposed Jaccard loss (JACLoss) performs stable across thresholds in range $[0.05, 0.95]$.

JAC-64, we choose to fine-tune both networks from pre-trained models. Lower layers of HNN are transferred from VGG16 while UNet parameters are transferred from the snapshot released in [9]. Dice similarity coefficient (DSC), and Jaccard index (JI) results computed from their segmentation outputs are reported in Table 1, under the same 4-fold cross validation. RNN-128 performance best on CT-82, and JAC-128 achieves the best result on MRI-79. For self-contained content, results reported in [1,12,18] are also included in Table 1. Note that our method development is orthogonal to the principles of "coarse-to-fine" pancreas location and detection [12,18]. Better performance for pancreas segmentation may be achievable with the combination of both methodologies. Figure 5 displays exemplars of reconstructed segmentation results from NIH-CT-82 dataset.

Table 1. Comparison with the state-of-the-art methods under 4-fold cross validation: JAC and RNN represent networks trained with JACLoss and contextual regularization, respectively. −64 and −128 represent numbers of convolutional output channels. We show dice similarity coefficient (DSC), jaccard index (JI) as mean ± standard dev. [worst, best]. The best result on CT and MRI are reported by RNN-128 and JAC-128 with bold font.

Method	NIH-CT82		MRI-79	
	DSC(%)	JI(%)	DSC(%)	JI(%)
UNet [9]	79.7 ± 7.6 [43.4, 89.3]	66.8 ± 9.60 [27.7, 80.7]	79.9 ± 7.30 [54.8, 90.5]	67.1 ± 9.50 [37.7, 82.6]
HNN [17]	79.6 ± 7.7 [41.9, 88.0]	66.7 ± 9.40 [26.5, 78.6]	75.9 ± 10.1 [33.0, 86.8]	62.1 ± 11.3 [19.8, 76.6]
JAC-64	80.3 ± 9.0 [35.8, 90.2]	67.9 ± 10.9 [21.8, 82.1]	76.3 ± 12.9 [6.30, 88.8]	63.1 ± 14.0 [3.30, 79.9]
JAC-128	81.5 ± 7.2 [56.3, 90.1]	69.3 ± 9.50 [39.2, 82.0]	**80.5 ± 6.70** [59.1, 89.4]	**67.9 ± 8.90** [41.9, 80.9]
RNN-64	82.3 ± 6.7 [49.8, 90.2]	70.4 ± 8.60 [33.1, 82.2]	78.1 ± 9.40 [39.5, 90.0]	64.9 ± 11.4 [24.6, 81.8]
RNN-128	**82.4 ± 6.7 [60.0, 90.1]**	**70.6 ± 9.00** [42.9, 81.9]	80.4 ± 6.60 [58.9, 90.0]	67.7 ± 8.70 [41.8, 81.8]
Roth *et al.* [12]	81.3 ± 6.3 [50.6, 88.9]	68.8 ± 8.12 [33.9, 80.1]	–	–
Zhou *et al.* [18]	82.3 ± 5.6 [62.4, 90.8]	–	–	–
Cai *et al.* [1]	–	–	76.1 ± 8.7 [47.4, 87.1]	–

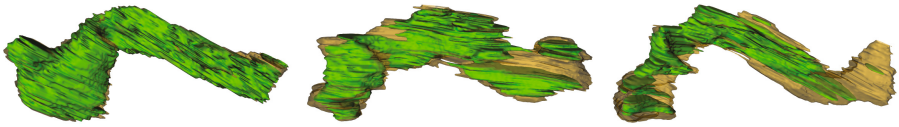

Fig. 5. 3D visualization of pancreas segmentation results: human annotation shown in golden and computerized segmentation displayed in green. The DSC are 90%, 75%, and 60% for three examples from left to right, respectively.

4 Conclusion

In this paper, we use a new deep neural network architecture for pancreas segmentation, via our tailor-made convolutional neural network followed by convolutional LSTM to regularize the segmentation results on individual image slices, unlike the independent process assumed in previous work [1,11,12,18]. The contextual regularization permits to enforce the pancreas segmentation spatial smoothness explicitly. Combined with the proposed JACLoss function for CNN training to generate threshold-free segmentation results, our quantitative pancreas segmentation results improve the previous state-of-the-art approaches [1,11] on both CT and MRI datasets.

References

1. Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q.: Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 442–450. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_51
2. Chen, J., Yang, L., Zhang, Y., Alber, M.S., Chen, D.Z.: Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. CoRR abs/1609.01006 (2016)
3. Farag, A., Lu, L., Roth, H.R., Liu, J., Turkbey, E., Summers, R.M.: A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. IEEE Trans. Image Process. **26**(1), 386–399 (2017)
4. Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. CoRR abs/1603.05959 (2016)
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE CVPR, pp. 3431–3440, June 2015
6. Merkow, J., Kriegman, D.J., Marsden, A., Tu, Z.: Dense volume-to-volume vascular boundary detection. CoRR abs/1605.08401 (2016)
7. Milletari, F., Navab, N., Ahmadi, S.: V-net: fully convolutional neural networks for volumetric medical image segmentation. CoRR abs/1606.04797 (2016)
8. Oda, M., et al.: Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 556–563. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_64

9. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28
10. Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M.: DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 556–564. Springer, Cham (2015). doi:10.1007/978-3-319-24553-9_68
11. Roth, H.R., Lu, L., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 451–459. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_52
12. Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Summers, R.M.: Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. CoRR abs/1702.00045 (2017)
13. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. CoRR abs/1506.04214 (2015)
14. Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. CoRR abs/1506.07452 (2015)
15. Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D.: Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. **23**(1), 92–104 (2015)
16. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans. Med. Imaging **32**(9), 1723–1730 (2013)
17. Xie, S., Tu, Z.: Holistically-nested edge detection. In: IEEE ICCV, pp. 1395–1403 (2015)
18. Zhou, Y., Xie, L., Shen, W., Fishman, E., Yuille, A.L.: Pancreas segmentation in abdominal CT scan: a coarse-to-fine approach. CoRR abs/1612.08230 (2016)