A Fixed-Point Model for Pancreas Segmentation in Abdominal CT Scans

Yuyin Zhou[1], Lingxi Xie[2](✉), Wei Shen[3],
Yan Wang[4], Elliot K. Fishman[5], Alan L. Yuille[6]

[1,2,3,4,6]The Johns Hopkins University, Baltimore, MD 21218, USA
[3]Shanghai University, Baoshan District, Shanghai 200444, China
[5]The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA
[1]zhouyuyiner@gmail.com [2]198808xc@gmail.com [3]wei.shen@t.shu.edu.cn
[4]wyanny.9@gmail.com [5]efishman@jhmi.edu [6]alan.l.yuille@gmail.com
http://bigml.cs.tsinghua.edu.cn/~lingxi/Projects/OrganSegC2F.html

Abstract. Deep neural networks have been widely adopted for automatic organ segmentation from abdominal CT scans. However, the segmentation accuracy of some small organs (*e.g.*, the pancreas) is sometimes below satisfaction, arguably because deep networks are easily disrupted by the complex and variable background regions which occupies a large fraction of the input volume. In this paper, we formulate this problem into a fixed-point model which uses a predicted segmentation mask to shrink the input region. This is motivated by the fact that a smaller input region often leads to more accurate segmentation. In the training process, we use the ground-truth annotation to generate accurate input regions and optimize network weights. On the testing stage, we fix the network parameters and update the segmentation results in an iterative manner. We evaluate our approach on the NIH pancreas segmentation dataset, and outperform the state-of-the-art by more than 4%, measured by the average Dice-Sørensen Coefficient (DSC). In addition, we report 62.43% DSC in the worst case, which guarantees the reliability of our approach in clinical applications.

1 Introduction

In recent years, due to the fast development of deep neural networks [4][10], we have witnessed rapid progress in both medical image analysis and computer-aided diagnosis (CAD). This paper focuses on an important prerequisite of CAD [3][13], namely, automatic segmentation of small organs (*e.g.*, the pancreas) from CT-scanned images. The difficulty mainly comes from the high anatomical variability and/or the small volume of the target organs. Indeed researchers sometimes design a specific segmentation approach for each organ [1][9].

Among different abdominal organs, pancreas segmentation is especially difficult, as the target often suffers from high variability in shape, size and location [9], while occupying only a very small fraction (*e.g.*, $< 0.5\%$) of the entire CT volume. In such cases, deep neural networks can be disrupted by the background

region, which occupies a large fraction of the input volume and includes complex and variable contents. Consequently, the segmentation result becomes inaccurate especially around the boundary areas.

To alleviate this, we apply a fixed-point model [5] using the predicted segmentation mask to shrink the input region. With a relatively smaller input region (*e.g.*, a bounding box defined by the mask), it is straightforward to achieve more accurate segmentation. At the training stage, we fix the input regions generated from the ground-truth annotation, and train two deep segmentation networks, *i.e.*, a coarse-scaled one and a fine-scaled one, to deal with the entire input region and the region cropped according to the bounding box, respectively. At the testing stage, the network parameters remain unchanged, and an iterative process is used to optimize the fixed-point model. On a modern GPU, our approach needs around 3 minutes to process a CT volume during the testing stage. This is comparable to recent work [8], but we report much higher accuracy.

We evaluate our approach on the NIH pancreas segmentation dataset [9]. Compared to recently published work [9][8], our average segmentation accuracy, measured by the Dice-Sørensen Coefficient (DSC), increases from 78.01% to 82.37%. Meanwhile, we report 62.43% DSC on the worst case, which guarantees reasonable performance on the particularly challenging test samples. In comparison, [8] reports 34.11% DSC on the worst case and [9] reports 23.99%. Meanwhile, our approach can be applied to segmenting other organs or tissues, especially when the target is very small, *e.g.*, the pancreatic cyst [13].

2 Approach

2.1 Deep Segmentation Networks

Let a CT-scanned image be a 3D volume $\mathbf{X}$ of size $W \times H \times L$ and annotated with a ground-truth segmentation $\mathbf{Y}$ where $y_i = 1$ indicates a foreground voxel. Consider a segmentation model $\mathbb{M} : \mathbf{Z} = \mathbf{f}(\mathbf{X}; \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ denotes the model parameters, and the loss function is written as $\mathcal{L}(\mathbf{Z}, \mathbf{Y})$. In the context of a deep segmentation network, we optimize $\mathcal{L}$ with respect to the network weights $\boldsymbol{\Theta}$ by gradient back-propagation. As the foreground region is often very small, we follow [7] to design a DSC-loss layer to prevent the model from being heavily biased towards the background class. We slightly modify the DSC of two voxel sets $\mathcal{A}$ and $\mathcal{B}$, $\mathrm{DSC}(\mathcal{A}, \mathcal{B}) = \frac{2 \times |\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}| + |\mathcal{B}|}$, into a loss function between the ground-truth mask $\mathbf{Y}$ and the predicted mask $\mathbf{Z}$, *i.e.*, $\mathcal{L}(\mathbf{Z}, \mathbf{Y}) = 1 - \frac{2 \times \sum_i z_i y_i}{\sum_i z_i + \sum_i y_i}$. Note that this is a "soft" definition of DSC, and it is equivalent to the original form if all z_i's are either 0 or 1. The gradient computation is straightforward: $\frac{\partial \mathcal{L}(\mathbf{Z}, \mathbf{Y})}{\partial z_j} = -2 \times \frac{y_j \left(\sum_i z_i + \sum_i y_i\right) - \sum_i z_i y_i}{\left(\sum_i z_i + \sum_i y_i\right)^2}$.

We use the 2D fully-convolutional network (FCN) [6] as our baseline. The main reason for not using 3D models is the limited amount of training data. To fit a 3D volume $\mathbf{X}$ into a 2D network $\mathbb{M}$, we cut it into a set of 2D slices. This is obtained along three axes, *i.e.*, the coronal, sagittal and axial views.

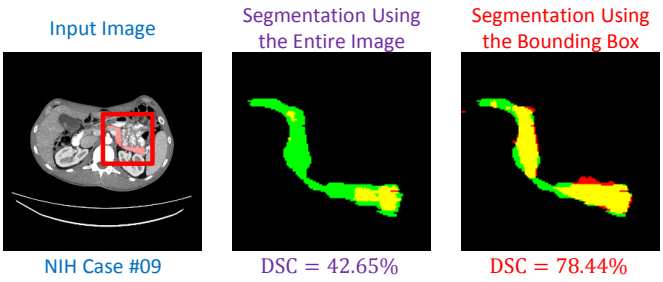

Fig. 1. Segmentation results with different input regions (best viewed in color), either using the entire image or the bounding box (the red frame). Red, green and yellow indicate the prediction, ground-truth and overlapped pixels, respectively.

We denote these 2D slices as $\mathbf{x}_{\mathrm{C},w}$ $(w = 1, 2, \ldots, W)$, $\mathbf{x}_{\mathrm{S},h}$ $(h = 1, 2, \ldots, H)$ and $\mathbf{x}_{\mathrm{A},l}$ $(l = 1, 2, \ldots, L)$, where the subscripts C, S and A stand for "coronal", "sagittal" and "axial", respectively. We train three 2D-FCN models $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$ to perform segmentation through three views individually (images from three views are quite different). In testing, the segmentation results from three views are fused via majority voting.

2.2 Fixed-Point Optimization

The pancreas often occupies a very small part (*e.g.*, $< 0.5\%$) of a CT volume. It was observed [9] that deep segmentation networks such as FCN [6] produce less satisfying results when detecting small organs, arguably because the network is easily disrupted by the varying contents in the background regions. Much more accurate segmentation can be obtained by using a smaller input region around the region-of-interest. A typical example is shown in Figure 1.

This inspires us to make use of the predicted segmentation mask to shrink the input region. We introduce a transformation function $r(\mathbf{X}, \mathbf{Z}^\star)$ which generates the input region given the current segmentation $\mathbf{Z}^\star$. We rewrite the model as $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star); \boldsymbol{\Theta})$, and the loss function is $\mathcal{L}(\mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star); \boldsymbol{\Theta}), \mathbf{Y})$. Note that the segmentation mask ($\mathbf{Z}$ or $\mathbf{Z}^\star$) appears in both the input and output of $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star); \boldsymbol{\Theta})$. This is a fixed-point model, and we apply the approach described in [5] for optimization, *i.e.*, finding a steady-state solution for $\mathbf{Z}$.

In training, the ground-truth annotation $\mathbf{Y}$ is used as the input mask $\mathbf{Z}^\star$. We train two sets of models (each set contains three models for different views) to deal with different input sizes. The *coarse-scaled* models are trained on those slices on which the pancreas occupies at least 100 pixels (approximately 25mm^2 in a 2D slice, our approach is not sensitive to this parameter) so as to prevent the model from being heavily impacted by the background. For the *fine-scaled* models, we crop each slice according to the minimal 2D box covering the pancreas, add a frame around it, and fill it up with the original image

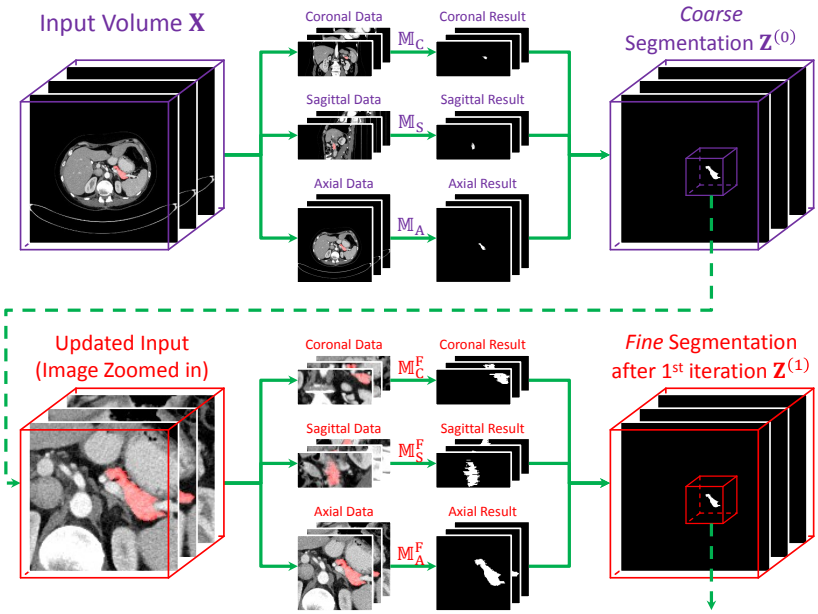

Fig. 2. Illustration of the testing process (best viewed in color). Only one iteration is shown here. In practice, there are at most 10 iterations.

data. The top, bottom, left and right margins of the frame are random integers sampled from $\{0, 1, \ldots, 60\}$. This strategy, known as data augmentation, helps to regularize the network and prevent over-fitting.

We initialize both networks using the FCN-8s model [6] pre-trained on the PascalVOC image segmentation task. The coarse-scaled model is fine-tuned with a learning rate of 10^{-5} for 80,000 iterations, and the fine-scaled model undergoes 60,000 iterations with a learning rate of 10^{-4}. Each mini-batch contains one training sample (a 2D image sliced from a 3D volume).

In testing, we use an iterative process to find a steady-state solution for $\mathbf{Z} = \mathbf{f}(r(\mathbf{X}, \mathbf{Z}^\star); \boldsymbol{\Theta})$. At the beginning, $\mathbf{Z}^\star$ is initialized as the entire 3D volume, and we compute the *coarse* segmentation $\mathbf{Z}^{(0)}$ using the *coarse-scaled* models. In each of the following T iterations, we slice the predicted mask $\mathbf{Z}^{(t-1)}$, find the smallest 2D box to cover all predicted foreground pixels in each slice, add a 30-pixel-wide frame around it (this is the mean value of the random distribution used in training), and use the *fine-scaled* models to compute $\mathbf{Z}^{(t)}$. The iteration terminates when a fixed number of iterations T is reached, or the the similarity between successive segmentation results ($\mathbf{Z}^{(t-1)}$ and $\mathbf{Z}^{(t)}$) is larger than a given threshold R. The similarity is defined as the inter-iteration DSC, namely $d^{(t)} = \mathrm{DSC}\big(\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\big) = \frac{2 \times \sum_i z_i^{(t-1)} z_i^{(t)}}{\sum_i z_i^{(t-1)} + \sum_i z_i^{(t)}}$. The testing stage is illustrated in Figure 2 and described in Algorithm 1.

Algorithm 1 Fixed-Point Model for Segmentation

Input: the testing volume $\mathbf{X}$, coarse-scaled models $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$, fine-scaled models $\mathbb{M}_{\mathrm{C}}^{\mathrm{F}}$, $\mathbb{M}_{\mathrm{S}}^{\mathrm{F}}$ and $\mathbb{M}_{\mathrm{A}}^{\mathrm{F}}$, threshold R, maximal rounds in iteration T.
Initialization: using $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$ to generate $\mathbf{Z}^{(0)}$ from $\mathbf{X}$;
for $t = 1, 2, \ldots, T$ **do**
    Using $\mathbb{M}_{\mathrm{C}}^{\mathrm{F}}$, $\mathbb{M}_{\mathrm{S}}^{\mathrm{F}}$ and $\mathbb{M}_{\mathrm{A}}^{\mathrm{F}}$ to generate $\mathbf{Z}^{(t)}$ from $\mathbf{Z}^{(t-1)}$;
    if $\mathrm{DSC}\left(\mathbf{Z}^{(t-1)}, \mathbf{Z}^{(t)}\right) \geqslant R$ **then break**;
    end if
end for
Output: the final segmentation $\mathbf{Z}^{\star} = \mathbf{Z}^{(t)}$.

3 Experiments

3.1 Dataset and Evaluation

We evaluate our approach on the NIH pancreas segmentation dataset [9], which contains 82 contrast-enhanced abdominal CT volumes. The resolution of each CT scan is $512 \times 512 \times L$, where $L \in [181, 466]$ is the number of sampling slices along the long axis of the body. The slice thickness varies from 0.5mm–1.0mm. Following the standard cross-validation strategy, we split the dataset into 4 fixed folds, each of which contains approximately the same number of samples. We apply cross validation, *i.e.*, training the model on 3 out of 4 subsets and testing it on the remaining one. We measure the segmentation accuracy by computing the Dice-Sørensen Coefficient (DSC) for each sample. This is a similarity metric between the prediction voxel set $\mathcal{Z}$ and the ground-truth set $\mathcal{Y}$, with the mathematical form of $\mathrm{DSC}(\mathcal{Z}, \mathcal{Y}) = \frac{2 \times |\mathcal{Z} \cap \mathcal{Y}|}{|\mathcal{Z}| + |\mathcal{Y}|}$. We report the average DSC score together with the standard deviation over 82 testing cases.

3.2 Results

We first evaluate the baseline (coarse-scaled) approach. Using the coarse-scaled models trained from three different views (*i.e.*, $\mathbb{M}_{\mathrm{C}}$, $\mathbb{M}_{\mathrm{S}}$ and $\mathbb{M}_{\mathrm{A}}$), we obtain 66.88%±11.08%, 71.41%±11.12% and 73.08%±9.60% average DSC, respectively. Fusing these three models via majority voting yields 75.74%±10.47%, suggesting that complementary information is captured by different views. This is used as the starting point $\mathbf{Z}^{(0)}$ for the later iterations.

To apply the fixed-point model for segmentation, we first compute $d^{(t)}$ to observe the convergence of the iterations. After 10 iterations, the average $d^{(t)}$ value over all samples is 0.9767, the median is 0.9794, and the minimum is 0.9362. These numbers indicate that the iteration process is generally stable.

Now, we investigate the fixed-point model using the threshold $R = 0.95$ and the maximal number of iterations $T = 10$. The average DSC is boosted by 6.63%, which is impressive given the relatively high baseline (75.74%). This verifies our hypothesis, *i.e.*, a fine-scaled model depicts a small organ more accurately.

Method	Mean DSC	# Iterations	Max DSC	Min DSC
Roth *et.al*, MICCAI'2015 [9]	71.42 ± 10.11	–	86.29	23.99
Roth *et.al*, MICCAI'2016 [8]	78.01 ± 8.20	–	88.65	34.11
Coarse Segmentation	75.74 ± 10.47	–	88.12	39.99
After 1 Iteration	82.16 ± 6.29	1	**90.85**	54.39
After 2 Iterations	82.13 ± 6.30	2	90.77	57.05
After 3 Iterations	82.09 ± 6.17	3	90.78	58.39
After 5 Iterations	82.11 ± 6.09	5	90.75	62.40
After 10 Iterations	82.25 ± 5.73	10	90.76	61.73
After $d_t > 0.90$	82.13 ± 6.35	1.83 ± 0.47	**90.85**	54.39
After $d_t > 0.95$	$\mathbf{82.37} \pm 5.68$	2.89 ± 1.75	**90.85**	**62.43**
After $d_t > 0.99$	82.28 ± 5.72	9.87 ± 0.73	90.77	61.94
Best among All Iterations	82.65 ± 5.47	3.49 ± 2.92	90.85	63.02
Oracle Bounding Box	83.18 ± 4.81	–	91.03	65.10

Table 1. Segmentation accuracy (measured by DSC, %) reported by different approaches. We start from initial (coarse) segmentation $\mathbf{Z}^{(0)}$, and explore different terminating conditions, including a fixed number of iterations and a fixed threshold of inter-iteration DSC. The last two lines show two upper-bounds of our approach, *i.e.*, "Best of All Iterations" means that we choose the highest DSC value over 10 iterations, and "Oracle Bounding Box" corresponds to using the ground-truth segmentation to generate the bounding box in testing. We also compare our results with the state-of-the-art [9][8], demonstrating our advantage over all statistics.

We also summarize the results generated by different terminating conditions in Table 1. We find that performing merely 1 iteration is enough to significantly boost the segmentation accuracy (+6.42%). However, more iterations help to improve the accuracy of the worst case, as for some challenging cases (*e.g.*, Case #09, see Figure 3), the missing parts in coarse segmentation are recovered gradually. The best average accuracy comes from setting $R = 0.95$. Using a larger threshold (*e.g.*, 0.99) does not produce accuracy gain, but requires more iterations and, consequently, more computation at the testing stage. In average, it takes less than 3 iterations to reach the threshold 0.95. On a modern GPU, we need about 3 minutes on each testing sample, comparable to recent work [8], but we report much higher segmentation accuracy (82.37% vs. 78.01%).

As a diagnostic experiment, we use the ground-truth (oracle) bounding box of each testing case to generate the input volume. This results in a 83.18% average accuracy (no iteration is needed in this case). By comparison, we report a comparable 82.37% average accuracy, indicating that our approach has almost reached the upper-bound of the current deep segmentation network.

We also compare our segmentation results with the state-of-the-art approaches. Using DSC as the evaluation metric, our approach outperforms the recent published work [8] significantly. The average accuracy over 82 samples increases remarkably from 78.01% to 82.37%, and the standard deviation decreases from 8.20% to 5.68%, implying that our approach are more stable. We also

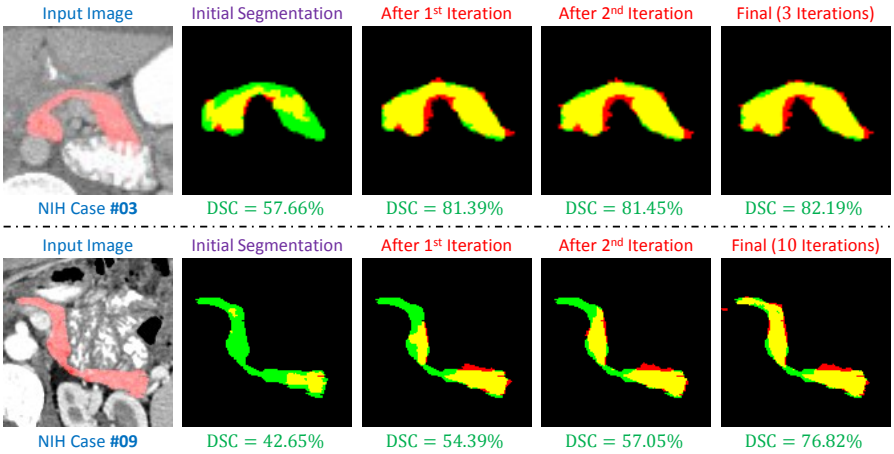

Fig. 3. Examples of segmentation results throughout the iteration process (best viewed in color). We only show a small region covering the pancreas in the axial view. The terminating condition is $d^{(t)} \geqslant 0.95$. Red, green and yellow indicate the prediction, ground-truth and overlapped regions, respectively.

implement a recently published coarse-to-fine approach [12], and get a 77.89% average accuracy. In particular, [8] reported 34.11% for the worst case (some previous work [2][11] reported even lower numbers), and this number is boosted considerably to 62.43% by our approach. We point out that these improvements are mainly due to the fine-tuning iterations. Without it, the average accuracy is 75.74%, and the accuracy on the worst case is merely 39.99%. Figure 3 shows examples on how the segmentation quality is improved in two challenging cases.

4 Conclusions

We present an efficient approach for accurate pancreas segmentation in abdominal CT scans. Motivated by the significant improvement brought by small and relatively accurate input region, we formulate a fixed-point model taking the segmentation mask as both input and output. At the training stage, we use the ground-truth annotation to generate a smaller input region, and train both coarse-scaled and fine-scaled models to deal with different input sizes. At the testing stage, an iterative process is performed for optimization. In practice, our approach often comes to an end after 2–3 iterations.

We evaluate our approach on the NIH pancreas segmentation dataset with 82 samples, and outperform the state-of-the-art by more than 4%, measured by the Dice-Sørensen Coefficient (DSC). Most of the benefit comes from the first iteration, and the remaining iterations only improve the segmentation accuracy by a little (about 0.3% in average). We believe that our algorithm can achieve an even higher accuracy if a more powerful network structure is used. Meanwhile,

our approach can be applied to other small organs, *e.g.*, spleen, duodenum or a lesion area in pancreas [13]. In the future, we will try to incorporate the fixed-point model into an end-to-end learning framework.

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and NSFC No. 61672336. We thank Dr. Seyoun Park and Zhuotun Zhu for their enormous help, and Weichao Qiu, Cihang Xie, Chenxi Liu, Siyuan Qiao and Zhishuai Zhang for instructive discussions.

References

1. Al-Ayyoub, M., Alawad, D., Al-Darabsah, K., Aljarrah, I.: Automatic Detection and Classification of Brain Hemorrhages. WSEAS Transactions on Computers 12(10), 395–405 (2013)
2. Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Nimura, Y., Rueckert, D., Mori, K.: Multi-organ Segmentation based on Spatially-Divided Probabilistic Atlas from 3D Abdominal CT Images. International Conference on Medical Image Computing and Computer-Assisted Intervention (2013)
3. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P., Larochelle, H.: Brain Tumor Segmentation with Deep Neural Networks. Medical Image Analysis 35, 18–31 (2017)
4. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (2012)
5. Li, Q., Wang, J., Wipf, D., Tu, Z.: Fixed-Point Model For Structured Labeling. International Conference on Machine Learning (2013)
6. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. Computer Vision and Pattern Recognition (2015)
7. Milletari, F., Navab, N., Ahmadi, S.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. International Conference on 3D Vision (2016)
8. Roth, H., Lu, L., Farag, A., Sohn, A., Summers, R.: Spatial Aggregation of Holistically-Nested Networks for Automated Pancreas Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (2016)
9. Roth, H., Lu, L., Farag, A., Shin, H., Liu, J., Turkbey, E., Summers, R.: DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (2015)
10. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (2015)
11. Wang, Z., Bhatia, K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K., Rueckert, D.: Geodesic Patch-based Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (2014)
12. Zhang, Y., Ying, M., Yang, L., Ahuja, A., Chen, D.: Coarse-to-Fine Stacked Fully Convolutional Nets for Lymph Node Segmentation in Ultrasound Images. IEEE International Conference on Bioinformatics and Biomedicine (2016)
13. Zhou, Y., Xie, L., Fishman, E., Yuille, A.: Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CT Scans. International Conference on Medical Image Computing and Computer-Assisted Intervention (2017)