

# Robust Abdominal Organ Segmentation Using Regional Convolutional Neural Networks

Måns Larsson<sup>1</sup>, Yuhang Zhang<sup>1</sup>, and Fredrik Kahl<sup>1,2</sup>

<sup>1</sup> Chalmers University of Technology, Gothenburg, Sweden,

{mans.larsson, alven, fredrik.kahl}@chalmers.se

<sup>2</sup> Centre for Mathematical Sciences, Lund University, Lund, Sweden

**Abstract.** A fully automatic system for abdominal organ segmentation is presented. As a first step, an organ localization is obtained via a robust and efficient feature registration method where the center of the organ is estimated together with a region of interest surrounding the center. Then, a convolutional neural network performing voxelwise classification is applied. The convolutional neural network consists of several full 3D convolutional layers and takes both low and high resolution image data as input, which is designed to ensure both local and global consistency. Despite limited training data, our experimental results are on par with state-of-the-art approaches that have been developed over many years. More specifically the method is applied to the MICCAI2015 challenge “Multi-Atlas Labeling Beyond the Cranial Vault” in the free competition for organ segmentation in the abdomen. It achieved the best results for 3 out of the 13 organs with a total mean Dice coefficient of **0.757** for all organs. Top scores were achieved for the gallbladder, the aorta and the right adrenal gland.

**Keywords:** Medical Image Analysis, Segmentation, Convolutional Neural Networks

## 1 Introduction

Segmentation is a key problem in medical image analysis, and an automated method for organ segmentation can be crucial for numerous applications in medical research and clinical care such as computer aided diagnosis and surgery assistance. The high variability of the shape and position of abdominal organs makes segmentation a challenging task. Previous work done on segmentation of abdominal organs includes, among others, multi-atlas methods [17], patch-based methods [16], and methods based on a probabilistic atlas [11, 2]. These techniques achieve great results for several abdominal organs but struggle with the segmentation of organs where the anatomical variability is large such as the gallbladder.

During the last few years, deep convolutional neural networks have shown great performance and achieved state of the art results in many computer vision applications [9, 10]. This fact can be partly attributed to the constant increase in available computing power, most notably GPU computing solutions, and the availability of large annotated datasets. In the field of medical image analysis this development has led to an increased interest in methods based on deep convolutional neural networks with promising results

[4, 13]. Recently, more intricate deep learning approaches have been proposed in the field of image analysis. Ronneberger et al [12] proposed the U-Net, a network based on the idea of “fully convolutional networks” [10]. This work was extended to a network utilizing 3D convolutional filters by Çiçek et al [3], a similar network structure where also proposed by Brosch et al [1]. Another example is Kamnitsas et al [8] that used an eleven layer deep 3D convolutional network for brain lesion segmentation with good results.

In this paper, instead of designing larger and more complicated network structures we propose a two step method, simplifying the task the convolutional neural network need to solve. In summary, an automatic system for the segmentation of abdominal organs in contrast enhanced CT images is presented. Our main contribution is to show that despite limited training data (compared to other successful deep learning approaches), it is possible to design a system that achieves on par with state-of-the-art for very challenging segmentation tasks with high anatomical variability. Another contribution is that we develop a computationally efficient framework that allows for a fully integrated 3D approach to multi-organ segmentation. To our knowledge, this is the first attempt on using 3D CNNs for multi-organ abdominal segmentation.

The presented method is trained and test on the MICCAI2015 challenge “Multi-Atlas Labeling Beyond the Cranial Vault” [18] where it achieved state of the art results in the free competition for organ segmentation in the abdomen. To this date, our method gives the best results for 3 out of the 13 organs.

## 2 Proposed Solution

Our system segments each organ independently and can be divided into three steps:

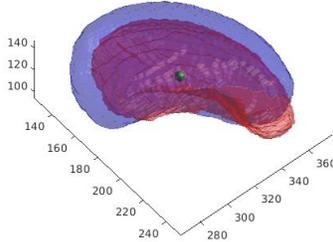
1. Localization of region of interest using a multi-atlas approach.
2. Voxelwise binary classification using a convolutional neural network.
3. Postprocessing by thresholding and removing all positive samples except the largest connected component.

Each step will now be described in detail.

### 2.1 Localization of region of interest

This part provides a robust initialization of the segmentation. The goal is to locate the center voxel of the organ in the target image. When this has been done a prediction mask is placed centered around the predicted organ center. The prediction mask later defines the region of interest where the convolutional neural network is initially applied. The use of an initialization method enables us to train more specialized, or regional, networks that only need to differentiate between a certain organ and the background. This means that the classification task that the network needs to perform is simplified and computationally less demanding networks can be applied.

The location of the organ center in the target image is obtained using a feature-based multi-atlas approach. Each atlas image is registered to the target using the method described in [7]. The method is computationally efficient compared to traditional intensity-based registration methods. More importantly though, this approach provides a robust



**Fig. 1.** Example of localization of region of interest for the Spleen. The green sphere is the estimated center point, the red mask describes the ground truth and the blue mask describes the estimated region of interest.

and reliable estimate for organ locations which have been demonstrated for many different settings and modalities. In our framework, the registration is performed individually for each organ and atlas image. Affine transformations are computed and then used to transform each organ center point from an atlas image to the target image. The median of the transformed center points provides us with an estimate of the center point for the region of interest in the target image. The reason for using the median, and not for instance the mean operator, is that it provides a robust estimate of the center point, that is, it is not affected by a few, spurious outliers.

The prediction mask is estimated using the ground truth segmentations of the atlas images. Let the ground truth segmentations be represented by a binary image of the same dimension as the atlas image  $G^{(l)}$ , where  $l$  is the image id, and  $G_{ijk}^{(l)} = 1$  if and only if voxel with index  $i, j, k$  in image with id  $l$  is foreground (or organ). Further, define  $D^{(l)}$  as the binary image formed by dilating  $G^{(l)}$  by a cube of size  $25 \times 25 \times 25$  voxels and translating it so that the center of the organ is located at the center of the image. The prediction mask  $P$  is then defined as the binary image where each element  $P_{ijk}$  is

$$P_{ijk} = \begin{cases} 1 & \text{if } \frac{1}{N} \sum_{l=1}^N D_{ijk}^{(l)} \geq \delta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where  $N$  is the number of atlas images and  $\delta$  is a threshold set to  $\delta = 0.5$  for the majority of the organs. Finally, the region of interest  $R$  is defined as the prediction mask centered around the estimated center point. An example of a localization of region of interest is shown in Figure 1.

## 2.2 Voxel classification using a convolutional neural network

The convolutional neural network is applied using a sliding window approach. For each voxel to be segmented two cubes of different resolutions centered around said voxel

are extracted and used as input features to the network. The network in return outputs a probability, denoted  $p_{ijk}$ , of the voxel being organ.

To speed up the process the network is not applied to every voxel in the area that is being segmented, denoted  $S$ . Instead, it takes steps of three in each dimension over  $S$ . The probabilities output by the network are then interpolated to every voxel in  $S$ . All voxels in  $S$  that have been assigned an interpolated probability neither close to zero nor close to one (voxels with a probability between 0.1 and 0.9 to be specific) will be classified by the network once more.

To reduce the dependency on the quality of the initial region of interest where the convolutional neural network is applied, a region growing algorithm is used. Call the set of voxels that should be segmented  $S$ . Further, call the set of voxels already classified by  $D$  and the set of voxels with an assigned probability larger than 0.5 by  $O$ . The region growing algorithm is then described by Algorithm 1. The use of a region growing algorithm means that even though the initial region may only cover part of the organ, a successful segmentation is still possible, see Figure 3.

```

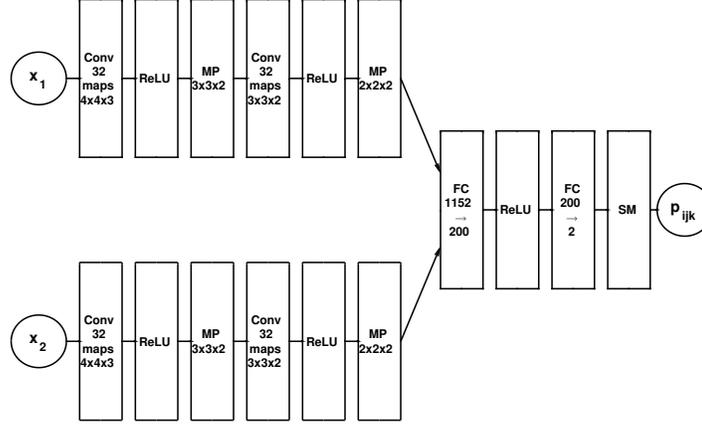
Initialize:
-  $S$  as the region of interest  $R$ 
-  $D$  as  $\emptyset$ 
-  $O$  as  $\emptyset$ .

while  $S \neq \emptyset$  do
|   - Classify voxels in  $S$ 
|   - Set  $D = D \cup S$ , and  $O$  as the set of voxels with an assigned
|     probability larger than 0.5
|   - Let  $O^+$  be the set  $O$  dilated by a cube of size  $12 \times 12 \times 4$  voxels
|   - Set  $S = O^+ \setminus D$ 
end
Output:  $O$ 

```

**Algorithm 1:** Region growing algorithm for efficient classification.

**Convolutional neural network setup.** The convolutional neural network used performs voxelwise binary classification. The input features for the network are two image cubes, one with a fine resolution similar to the original CT image and the other with a coarse resolution. The fine resolution input feature is meant to provide the network with local information ensuring local precision while the coarse resolution input feature is meant to ensure global spatial consistency. The inputs are processed in separate pipelines and the aggregated features from both pipelines are then merged for the last part of the network. A schematic of the convolutional neural network is shown in Figure 2.



**Fig. 2.** Structure of the convolutional neural network. Both type and size of each layer are shown. The following abbreviations are used. Conv: Convolutional layer, ReLU: Rectified Linear Unit, MP: Max Pooling, FC: Fully Connected and SM: Soft Max. Both inputs are cubes containing  $27 \times 27 \times 12$  voxels and are centered around the voxel being classified. Input  $x_1$  has as high resolution with voxels of size  $1 \times 1 \times 3 \text{ mm}^3$ , while input  $x_2$  is downsampled by a factor of five in each dimension.

**Implementation and training.** For the implementation of the convolutional neural network the framework Torch7 was used [5]. For each convolutional neural network the training and validation set were extracted from the region of interest calculated as described previously and the area around the part of the image describing the organ. This was done for each image in the training set. For the majority of the organs a balanced training set was used, meaning that there was an equal amount of foreground and background samples in the training set. However, since some of the organs are quite small this leads to a relatively small training set. Several methods, listed below, were used to deal with this problem.

1. For organs present in pairs, kidneys and adrenal glands, training samples from both the left and the right organ were used. Note that this does not pose a problem during inference since the initialization part of the method will separate the organs.
2. Expansion of the training set by adding slightly distorted CT images, transforming them using a random affine transformation similar to the identity transformation. The transformation matrix  $T$  was randomized as

$$T = \begin{pmatrix} 1 + \delta_{11} & \delta_{12} & \delta_{13} & 0 \\ \delta_{21} & 1 + \delta_{22} & \delta_{23} & 0 \\ \delta_{31} & \delta_{32} & 1 + \delta_{33} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where  $\delta_{ij}$  are independently and uniformly randomized numbers between  $-0.25$  and  $0.25$  for  $i = 1, 2, 3$  and  $j = 1, 2, 3$ .

3. Including a greater number of background samples than foreground samples in the training set. This leads to a larger but unbalanced training set.

The choice of what methods to use were empirically decided individually for each organ. The evaluation used for the decision was how well the network performed on the validation set. The networks were trained in mini batches using stochastic gradient descent with Nesterov’s momentum [15] and weight decay. The training parameters were set to: batch size 100, learning rate  $5 \cdot 10^{-3}$ , momentum weight 0.9, weight decay  $10^{-5}$ . The error function used was negative log likelihood. When an unbalanced training set was used the loss was multiplied by a factor  $k$  for foreground samples where  $k$  is the ratio between background and foreground samples. To avoid overfitting, dropout was applied during training [14]. The networks were trained for 10 epochs or more. The network that obtained the highest validation score was finally picked for the segmentation of the test images (see experimental section for different network and data settings).

### 2.3 Postprocessing

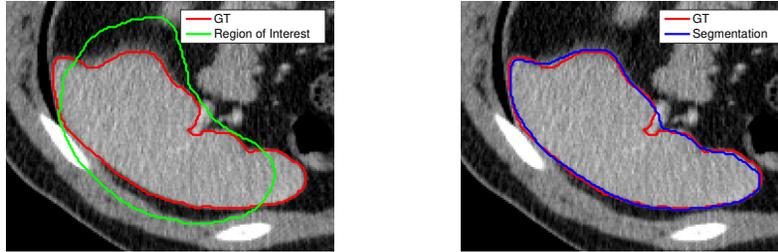
As a final step the probabilities from the convolutional neural network are thresholded, with a organ specific value estimated from data, in order to create a binary image. Everything but the largest connected component is set to zero producing the final segmentation.

## 3 Experimental Results

Our system was evaluated by submitting an entry to the MICCAI 2015 challenge “Multi-Atlas Labeling Beyond the Cranial Vault” in the free competition for organ segmentation in the abdomen [18]. In this challenge, there are 30 CT images coupled with manual segmentations of 13 organs, listed in Table 1. These 30 images and segmentations are available for method development and validation. Out of the 30 images 20 were used for training and 10 for validation.

In addition to these images training data from the VISCERAL challenge [6] was also used for training. The VISCERAL training data consists of 20 unenhanced whole body CT images and 20 contrast enhanced CT images over the abdomen and thorax. In these images organ ids 1, 2, 3, 4, 6, 8, 11, 12 and 13 were manually segmented. The unenhanced whole body CT images were excluded from the training set for organs with organ id 1, 2, 6, 8, 10, 11 and 12 since they differed too much from the enhanced CT images. All images were resampled to the same resolution of  $1 \text{ mm} \times 1 \text{ mm} \times 3 \text{ mm}$ . For the right kidney, a network trained on a training set formed by samples from both the right and the left kidney was used. For the stomach, the data set was expanded with distorted CT images and for the left adrenal gland an unbalanced data set was used with twice as many background samples as foreground samples.

The test set of the MICCAI challenge consists of 20 CT images. The submitted segmentations are evaluated by calculating the Dice coefficient for each organ. The final results are given in Table 1 with the currently two best competitors:



**Fig. 3.** Example of the resulting segmentation of the spleen for a CT slice. Note that even though the initial region of interest did not contain the entire organ the final result still does. This is due to region growing.

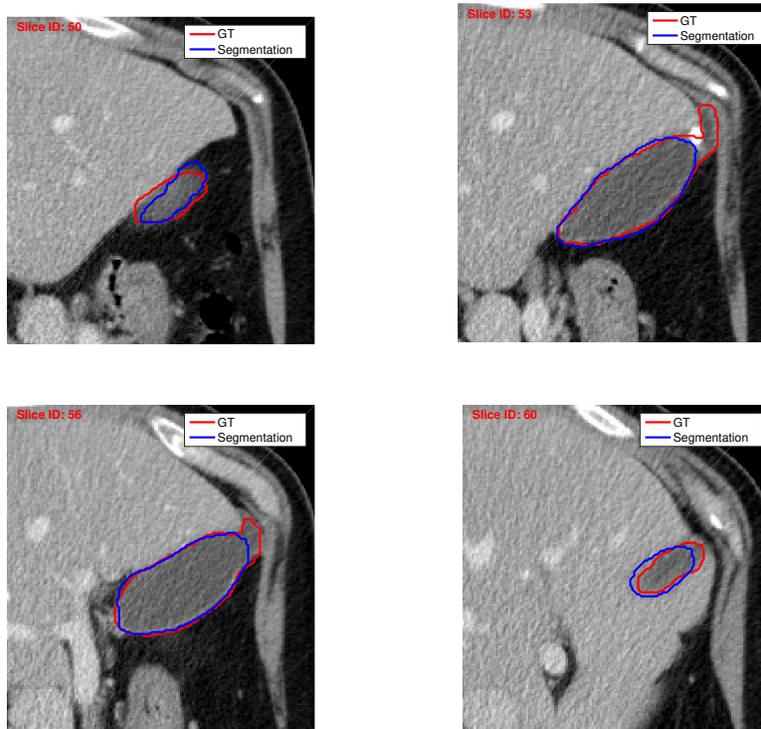
- **IMI** - algorithm name: *IMI\_deeds\_SSC\_jointCL* submitted by Mattias Heinrich at the Institute of Medical Informatics, Lübeck, Germany.
- **CLS** - algorithm name: *CLSIMPLEJLF\_organwise* submitted by Zhoubing Xu at the Vanderbilt University, Nashville, TN, USA.
- **Other** - this column contains results from other competitors. The score is only shown if they are the highest for that organ.

For each organ the estimation of the organ center took about 20 seconds and the classification with CNN took between 30 seconds and 5 minutes, depending on organ size. The CNN computations were performed on a GeForce GTX TITAN X GPU with Maxwell architecture

We note that for some organs that were not present in the VISCERAL data like Veins and Pancreas, we get lower results than IMI. The ideal solution to this problem would be to include more images in the training set. This, however, requires more manually segmented CT images which are not always easy to acquire. Another approach would be to train a network on several organs, and then fine tune the network weights for each specific organ. This could enable the network to learn higher order features that differentiate well between all organs in the CT image, not only between the organs located closest to the organ that is currently being segmented. This is left for future work.

## 4 Discussion

In Table 2 a comparison of the validation score and the test score of our method is presented. As can be seen from the table there is a large difference between validation and testing scores for some organ. This means that our networks do not generalize well to the test set for these organs which might be an indication of overfitting and that the input features and structure of our network is not ideal to learn high order information that generalize to all other CT images. However, since the validation data has not been



**Fig. 4.** Example of the resulting segmentation of the gallbladder. This is one of many images where our method achieved good result for the gallbladder, for this image we got a Dice coefficient of 0.90.

used for the actual training, only for the decision on when to stop the training, these differences might not be only due to overfitting. Instead it might be due to the existence of anatomical variations in the test set that differs too much from anything seen in the training and validation images for the network to perform well. A specific example of where our method performed badly on the test data, for an organ with good validation result, is shown in Figure 5. Here, the network has classified most of the right kidney correctly. However, it has also classified a lot of surrounding organs or tissue as right kidney as well.

The ideal solution to this problem would be to include more images in the training set. This however, requires more manually segmented CT images which are not always easy to acquire. Other approaches to solve this problem would be to train a network on several organs, and then fine tune the network weights for each specific organ. This could enable the network to learn higher order features that differentiates well between all organs in the CT image, not only between the organs located closest to the organ that is currently being segmented.

**Table 1.** Final results measured in Dice metric for organ segmentation in CT images. Our approach gives the best results for 3 out of the 13 organs. Here '-' means that one of the specified methods achieved best result.

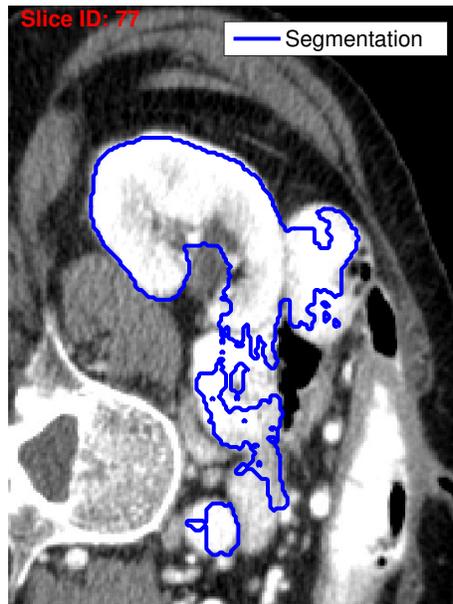
<i>Organ</i>	<i>IMI</i>	<i>CLS</i>	<i>Other</i>	<i>Our</i>
Spleen	0.919	0.911	<b>0.964</b>	0.930
Right Kidney	<b>0.901</b>	0.893	-	0.866
Left Kidney	0.914	0.901	<b>0.917</b>	0.911
Gallbladder	0.604	0.375	-	<b>0.624</b>
Esophagus	<b>0.692</b>	0.607	-	0.662
Liver	<b>0.948</b>	0.940	-	0.946
Stomach	<b>0.805</b>	0.704	-	0.775
Aorta	0.857	0.811	-	<b>0.860</b>
Inferior Vena Cava	<b>0.828</b>	0.760	-	0.776
Portal and Splenic Vein	0.754	0.649	<b>0.756</b>	0.567
Pancreas	<b>0.740</b>	0.643	-	0.602
Right Adrenal Gland	0.615	0.557	-	<b>0.631</b>
Left Adrenal Gland	<b>0.623</b>	0.582	-	0.583
Average	<b>0.790</b>	0.723	-	0.757

## 5 Conclusion

In this paper, an efficient system for abdominal organ segmentation was presented. Our approach first uses a robust localization algorithm for finding the region of interest. As a second step a convolutional neural network is applied performing voxelwise classification. The network takes two different resolutions of image data as input to ensure both global and local consistency. The method was evaluated by submitting an entry to the MICCAI2015 challenge “Multi-Atlas Labeling Beyond the Cranial Vault” in the free competition for organ segmentation in the abdomen. The entry achieved on par with state-of-the-art for a majority of the organs with a mean Dice coefficient of 0.757.

**Table 2.** Comparison of validation score and test score measured in Dice metric for organ segmentation in CT images.

<i>Organ</i>	<i>Validation</i>	<i>Test</i>
Spleen	0.944	0.930
Right Kidney	0.940	0.866
Left Kidney	0.928	0.911
Gallbladder	0.744	0.624
Esophagus	0.724	0.662
Liver	0.947	0.946
Stomach	0.823	0.775
Aorta	0.892	0.860
Inferior Vena Cava	0.823	0.776
Portal Vein and Splenic Vein	0.632	0.567
Pancreas	0.689	0.602
Right Adrenal Gland	0.600	0.631
Left Adrenal Gland	0.580	0.583
Average	0.790	0.757



**Fig. 5.** Example of the resulting segmentation of the right kidney for a CT slice from the test set. The final segmentation is marked in blue. This segmentation was one of the examples where the method performed badly.

## References

1. Tom Brosch, Lisa Y. W. Tang, Youngjin Yoo, David K. B. Li, Anthony Traboulsee, and Roger Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.
2. Chengwen Chu, Masahiro Oda, Takayuki Kitasaka, Kazunari Misawa, Michitaka Fujiwara, Yuichiro Hayashi, Yukitaka Nimura, Daniel Rueckert, and Kensaku Mori. Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 165–172. 2013.
3. Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*, pages 424–432. Springer International Publishing, Cham, 2016.
4. Dan Claudiu Ciresan, Alessandro Giusti, Luca Maria Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *MICCAI*, volume 2, pages 411–418, 2013.
5. Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.
6. Allan Hanbury, Henning Müller, Georg Langs, and Bjoern H. Menze. *The Future Internet: Future Internet Assembly 2013: Validated Results and New Horizons*, chapter Cloud-Based Evaluation Framework for Big Data, pages 104–114. Berlin, Heidelberg, 2013.
7. Fredrik Kahl, Jennifer Alvé, Olof Enqvist, Frida Fejné, Johannes Ulén, Johan Fredriksson, Matilda Landgren, and Viktor Larsson. Good features for reliable registration in multi-atlas segmentation. In *VISCERAL Anatomy3 Segmentation Challenge*, pages 12–17, 2015.
8. Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, and Ben Glocker. Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic Stroke Lesion Segmentation*, page 13, 2015.
9. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
10. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
11. Hyunjin Park, Peyton H. Bland., and Charles R. Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *Medical Imaging, IEEE Transactions on*, 22(4):483–492, April 2003.
12. Olaf Ronneberger, Philipp Fischer, and Tomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).
13. Holger Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim Turkbey, and Ronald Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9349 of *Lecture Notes in Computer Science*, pages 556–564. 2015.
14. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.
15. Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 1139–1147, 2013.

16. Zehan Wang, Kanwal K. Bhatia, Ben Glocker, Antonio Marvao, Tim Dawes, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert. Geodesic patch-based segmentation. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, volume 8673 of *Lecture Notes in Computer Science*, pages 666–673. 2014.
17. Robin Wolz, Chris Chu, Kazunari Misawa, Masamichi Fujiwara, Kazuo Mori, and Daniel Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *Medical Imaging, IEEE Transactions on*, 32(9):1723–1730, 2013.
18. Zhoubing Xu. Multi-atlas labeling beyond the cranial vault - workshop and challenge, 2016. [Online; accessed 10-January-2017].