3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images

Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, Pheng-Ann Heng

PII: S1361-8415(17)30072-5
DOI: 10.1016/j.media.2017.05.001
Reference: MEDIMA 1256

To appear in: *Medical Image Analysis*

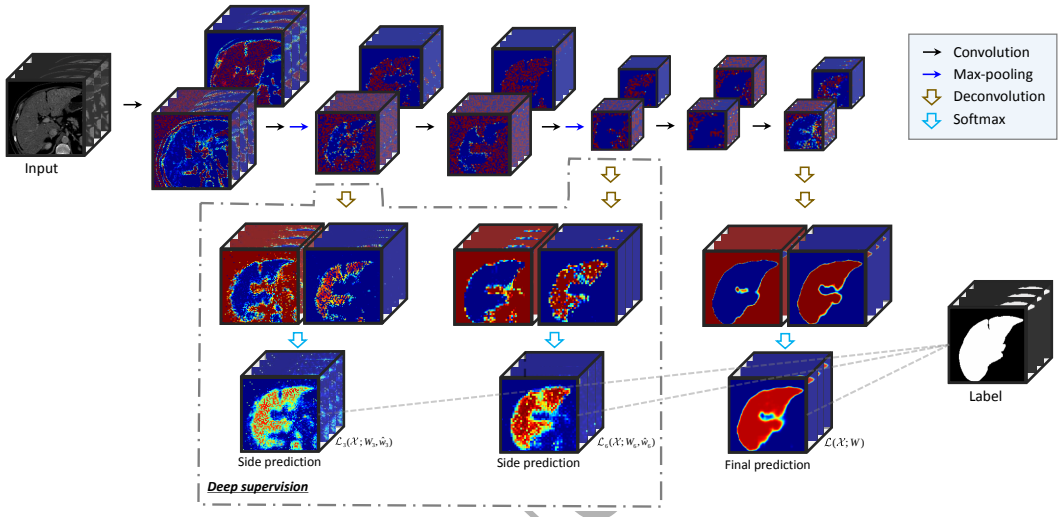
Received date: 26 January 2017
Revised date: 14 April 2017
Accepted date: 1 May 2017

Please cite this article as: Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, Pheng-Ann Heng, 3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images, *Medical Image Analysis* (2017), doi: 10.1016/j.media.2017.05.001

Highlights

- 3D fully convolutional networks for efficient volume-to-volume learning and inference
- Per-voxel-wise error backpropagation which alleviates the risk of overfitting on limited dataset
- A 3D deep supervision mechanism that simultaneously accelerates optimization and boosts model performance
- State-of-the-art performance on two typical yet challenging medical image segmentation tasks

Graphical Abstract



3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images

Qi Dou[a], Lequan Yu[a], Hao Chen[a], Yueming Jin[a], Xin Yang[a], Jing Qin[b,*], Pheng-Ann Heng[a]

[a] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China*

[b] *Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China*

Abstract

While deep convolutional neural networks (CNNs) have achieved remarkable success in 2D medical image segmentation, it is still a difficult task for CNNs to segment important organs or structures from 3D medical images owing to several mutually affected challenges, including the complicated anatomical environments in volumetric images, optimization difficulties of 3D networks and inadequacy of training samples. In this paper, we present a novel and efficient 3D fully convolutional network equipped with a 3D deep supervision mechanism to comprehensively address these challenges; we call it 3D DSN. Our proposed 3D DSN is capable of conducting volume-to-volume learning and inference, which can eliminate redundant computations and alleviate the risk of over-fitting on limited training data. More importantly, the 3D deep supervision mechanism can effectively cope with the optimization problem of gradients vanishing or exploding when training a 3D deep model, accelerating the convergence speed and simultaneously improving the discrimination capability. Such a mechanism is developed by deriving an objective function that directly guides the training of both lower and upper layers in the network, so that the adverse effects of unstable gradient changes can be counteracted during the training procedure. We also employ a fully connected conditional random field model as a post-processing step to refine the segmentation results. We have extensively validated the proposed 3D DSN on two typical yet challenging volumetric medical image segmentation tasks: (i) liver segmentation from 3D CT scans and (ii) whole heart and great vessels segmentation from 3D MR images, by participating two grand challenges held in conjunction with MICCAI. We have achieved competitive segmentation results to state-of-the-art approaches in both challenges with a much faster speed, corroborating the effectiveness of our proposed 3D DSN.

Keywords: Volumetric medical image segmentation, 3D deeply supervised

*Corresponding author at: harry.qin@polyu.edu.hk

networks, 3D fully convolutional networks, deep learning.

1. Introduction

Delineating important organs or structures from volumetric medical images, such as 3D computed tomography (CT) and magnetic resonance (MR) images, is of great significance for clinical practice especially with the proliferation of 3D images in diagnosis and treatment of many diseases. Accurate segmentation not only facilitates the subsequent quantitative assessment of the regions of interest but also benefits precise diagnosis, prediction of prognosis, and surgical planning and intra-operative guidance. For examples, liver segmentation from 3D abdominal CT scans is a crucial prerequisite for computer-aided interventions of living donor transplantations, tumor resection and minimal invasive surgery (Heimann et al. (2009a); Radtke et al. (2007); Meinzer et al. (2002)); volumetric cardiac MR image segmentation is indispensable for cardiovascular disease treatment including radio-frequency ablation and surgical planning of complex congenital heart disease (Peters et al. (2007); Pace et al. (2015); Atehorta et al. (2016)). In this paper, we take these two representative yet challenging segmentation tasks as examples but note that many other volumetric segmentation tasks share the common challenges with these two tasks.

Nowadays, the recognized golden standard segmentation results are obtained from experienced physicians and radiologists via their visual inspection and manual delineations. However, the annotation of volumetric images with hundreds of slices in a slice-by-slice manner is tedious, time-consuming and very expensive. In addition, the manual labeling is subjective, suffers from the low reproducibility and would introduce a high inter-observer variability, as the quality of the segmentation results could be significantly influenced by the operator's experience and knowledge. To the end, automated segmentation algorithms are highly demanded, especially if we would like to efficiently obtain accurate and reproducible segmentation results in day-to-day clinical practice.

Automated volumetric medical image segmentation is indeed a challenging task. The appearance and shape variations of the targeting objects are often significant among patients. The boundary between the targeting organs or structures and its neighboring tissues is usually ambiguous with limited contrast, which is in essence caused by their similar imaging-related physical properties, e.g., attenuation coefficients in CT imaging and relaxation times in MR imaging (Kronman & Joskowicz (2016)). To meet these challenges, various algorithms have been extensively studied in the past decades. Previous algorithms mainly utilized statistical shape modeling, level sets, active contours, multi-atlas and graphical models, with hand-crafted features. However, these hand-crafted features usually have too limited representation capability to deal with the large variations of appearance and shape. Later on, learning based methods have been explored to seek more powerful features, but it is still difficult for those methods to take full advantage of the 3D spatial information existing in the volumetric medical images to achieve satisfactory segmentation results.

Recently, convolutional neural networks (CNNs), leveraging their hierarchically learned highly representative features, have revolutionized the natural image processing (Krizhevsky et al. (2012); He et al. (2015); Long et al. (2015)), and also witnessed successful applications in medical image analysis domain (Albarqouni et al. (2016); Chen et al. (2016c); Moeskops et al. (2016a); Setio et al. (2016)). Deep learning based methods have been emerging as a competitive and important branch of alternatives to resolve the traditional medical image segmentation tasks. While deep CNNs have achieved remarkable succuss in 2D medical image segmentations (Xing et al. (2016); Chen et al. (2016d); Dhungel et al. (2015)), it is still a difficult task for CNNs to segment objects from 3D medical images owing to the following mutually affected challenges. First, the 3D medical images have much more complicated anatomical environments than 2D images, and hence 3D variants of CNNs with much more parameters are usually required to capture more representative features. Second, training such a 3D CNN often confronts various optimization difficulties, such as over-fitting, gradients vanishing or exploding, and slow convergence speed. Third, the inadequacy of training data in many medical applications makes capturing distinctive features and training a deep 3D CNN even harder.

In this paper, we present a novel and efficient 3D CNN equipped with fully convolutional architecture and a 3D deep supervision mechanism to comprehensively address these challenges of volumetric medical image segmentation. We call it 3D DSN for short. Our contributions are summarized as follows:

- First, we develop a 3D fully convolutional architecture with 3D deconvolutional layers to bridge the coarse feature volumes to the dense probability predictions for voxel-level segmentation tasks. This architecture is capable of eliminating redundant computations of patch-based methods and realizing volume-to-volume learning and inference. Moreover, the per-voxel-wise error back-propagation extensively enlarges the training database, and hence alleviates the risk of over-fitting on limited training data.

- We further propose a 3D deep supervision mechanism by formulating an objective function that directly guides the training of both upper and lower layers in order to reinforce the prorogation of gradients flows within the network and hence learn more powerful and representative features. Such a mechanism can simultaneously speed up the optimization process and improve discrimination capability of the model.

- We extensively validate our proposed 3D DSN on two typical yet highly challenging volumetric medical image segmentation tasks: (i) liver segmentation from 3D CT scans and (ii) whole heart and great vessels segmentation from 3D MR images, by participating two well-known and influential challenges: (i) *Segmentation of Liver Competition* from MICCAI 2007 and (ii) *Whole-Heart and Great Vessels Segmentation from 3D Cardiovascular MRI in Congenital Heart Desease* from MICCAI 2016. We have achieved better or comparable segmentation results compared with

many state-of-the-art approaches in both challenges, demonstrating the effectiveness and generalization capability of the proposed 3D DSN.

A preliminary version of this work was presented in MICCAI 2016 Dou et al. (2016a). In this paper, we have substantially revised and extended the original paper. The main modifications include elaborating proposed methods, analyzing underlying design principles, adding much more experiments, discussing advantages and limitations, and providing a more comprehensive literature review. To facilitate future researches and encourage further improvements, we make our implementation of the 3D DSN publicly available in our project webpage[1].

2. Related Work

While there is a large literature on automated volumetric medical image segmentation, we shall focus on the recently published CNN-based algorithms, which are closely relevant to our work. We shall also provide comprehensive reviews on automated liver and heart segmentations, as we take these two applications as examples in this paper.

2.1. CNNs for Volumetric Structure Segmentation

Convolutional neural networks have been demonstrating state-of-the-art performance on many challenging medical image analysis tasks, including classification (Sirinukunwattana et al. (2016); Yu et al. (2016); Jamaludin et al. (2016)), detection (Setio et al. (2016); Dou et al. (2016b); Shin et al. (2016)) and segmentation (Ciresan et al. (2012); Havaei et al. (2016b)) in recent years. CNN-based volumetric medical image segmentation algorithms can be broadly categorized into two groups, i.e., 2D CNN based and 3D CNN based methods.

The 2D CNN based methods usually segment the volumetric CT or MR data in a slice-by-slice manner (Havaei et al. (2016a); Roth et al. (2016); Moeskops et al. (2016b); Zikic et al. (2014)). For examples, Havaei et al. (2016a) proposed two-pathway shallow networks with various cascaded architectures for low/high grade glioblastomas segmentation in brain MR images. Roth et al. (2016) proposed spatial aggregation of holistically-nested networks for pancreas segmentation in CT scans. Another representative work was U-Net (Ronneberger et al. (2015)) which formed a 2D fully convolutional architecture (Long et al. (2015)) and was efficient for dense medical image segmentations.

Even though these 2D CNN based methods have greatly improved the segmentation accuracy over traditional hand-crafted features based methods, they are conceivably not optimal for volumetric medical image analysis as they cannot take full advantage of the special information encoded in the volumetric data. The 2.5D methods (Roth et al. (2014)) have been proposed to incorporate richer spatial information but were still limited to 2D kernels. To overcome this

[1] `http://www.cse.cuhk.edu.hk/~qdou/3d-dsn/3d-dsn.html`

shortcoming, 3D CNN based algorithms (Chen et al. (2016a); Kamnitsas et al. (2016)) have been recently proposed, aiming at extracting more powerful volumetric representations across all three spatial dimensions. For example, Kamnitsas et al. (2016) proposed a 3D network consisting of dual pathways and the training strategy exploited a dense inference technique on image segments to overcome computational burden. Further using a 3D conditional random field model, this framework has demonstrated state-of-the-art performance on lesion segmentation from multi-channel MR volumes with traumatic brain injuries, brain tumors and ischemic stroke.

Concurrent with our MICCAI paper (Dou et al. (2016a)), which we substantially revised in this extended version, several 3D volume-to-volume segmentation networks have been proposed, including 3D U-Net (Çiçek et al. (2016)), V-Net (Milletari et al. (2016)), I2I-3D (Merkow et al. (2016)) and VoxResNet (Chen et al. (2016b)). The 3D U-Net extended the 2D U-Net into a 3D version, which had an analysis path to abstract features and a synthesis path to produce a full-resolution segmentation. Shortcut connections were established between layers of equal resolution in the analysis and synthesis paths. The V-Net divided the architecture into stages and incorporated residual connections. V-Net was trained towards a novel Dice coefficient based objective function which aimed to deal with the class imbalance situation. The VoxResNet profoundly borrowed the spirit of 2D deep residual learning (He et al. (2015)) and constructed a very deep 3D network. Multi-modality input and multi-level contextual information were further leveraged to produce state-of-the-art brain segmentation results. The I2I-3D was the most similar work to ours and was proposed for vascular boundary detection. To localize small vascular structures, complex multi-scale interactions were designed via mixing layers which concatenated features in upper and lower layers, following $1 \times 1 \times 1$ convolution. The I2I-3D included auxiliary supervision via side outputs in a holistic and dense manner, sharing the spirit of Xie & Tu (2015). Differently, our method employs supervision in a sparse manner, which can greatly reduce the scale of network parameters and the computation workload. Furthermore, we implicitly incorporate multi-scale information by aggregating the side outputs and final outputs as the unary potential into a CRF model.

2.2. Liver Segmentation

Accurate liver segmentation is fundamental for various computer-aided procedures including liver cancer diagnosis, hepatic disease interventions and treatment planning (Heimann et al. (2009a); Campadelli et al. (2009)). The challenges of automated liver segmentation in 3D CT scans arise from the large inter-patient shape variation, low intensity contrast between liver and adjacent organs (e.g., stomach, pancreas and heart), and the existence of various pathologies (e.g., tumors, cirrhosis and cysts). A variety of automatic approaches have been proposed for liver segmentation, reflecting the importance as well as difficulty of this application. The Campadelli et al. (2009) have conducted a comprehensive review of CT-based liver segmentation techniques.

Driven by the gray scale distribution in contrast-enhanced 3D CT images, early automatic solutions were dedicated to region growing (Rusko et al. (2007)), active contours (Shang et al. (2011); Suzuki et al. (2010)) and clustering based approaches (Zhao et al. (2010)). By further considering the structural information, statistical shape models (Cootes et al. (1995); Heimann & Meinzer (2009)) became the most successful and popular branch, which incorporated shape priors (Heimann et al. (2007); Kainmüller et al. (2007); Wimmer et al. (2009)), intensity distributions (Kainmüller et al. (2007)), as well as boundary and region information (Wimmer et al. (2009)) to describe the features of liver and delineate its boundary. Meanwhile, some researches have employed graph cuts (Li et al. (2015); Linguraru et al. (2012)) or level set (Li et al. (2013)) techniques to segment the liver and its interior lesions from 3D CT images. Recently, learning based methods have been emerging as promising alternatives which constructed discriminative classifiers to label the target/non-target regions (Danciu et al. (2012); Freiman et al. (2011)), or sought sparse representations for powerful features (Al-Shaikhli et al. (2015)). A pioneer learning based work was Ling et al. (2008) which proposed a hierarchical model using marginal space learning. Recently, the CNNs have been employed for automatic liver segmentation (Christ et al. (2016)), which was based on a 2D CNN architecture with a 3D conditional random field model.

2.3. Whole Heart and Great Vessel Segmentation

Automatic volumetric cardiac image segmentation is crucial for clinical cardiac applications, such as quantitative evaluation of blood ejection, atrial fibrillation ablation and cardiovascular surgical planning for congenital heart disease (Frangi et al. (2001); Wang et al. (2009); Tobon-Gomez et al. (2015)). Subjecting to the large shape variations, low contrast of myocardium against surrounding tissues and branchy structure of vessels, developing automatic cardiac image segmentation methods presents to be a very challenging task (Petitjean & Dacher (2011)).

Driven by boundary information, early automatic solutions resorted to active contours (Jolly et al. (2001); Kaus et al. (2004)) and level sets boosted variants (Fritscher et al. (2005)). By incorporating explicit 2D/3D shape and texture prior knowledge of heart and vessels, statistical shape models (Van Assen et al. (2006); Koikkalainen et al. (2008); Peters et al. (2010)) and appearance models (Mitchell et al. (2001)) which could potentially tackle the boundary deficiency became the frequent choices for cardiac image segmentation. Another popular stream was multi-atlas segmentation which propagated the class labels of cardiac atlas images to unseen cases utilizing image registration and label fusion techniques (Rueckert et al. (2002); Zhuang et al. (2010)). However, the limited discrimination capability of hand-crafted features, the subjective landmark annotations in training and the sensitivities to initializations would set the performance bottleneck on aforementioned attempts. Learning based methods have been rapidly emerging as viable alternatives for this application. For quantitative functional analysis of heart, Zheng et al. (2008) explored marginal space learning to localize key anatomical landmarks, which guided a 3D shape model

to delineate the boundary of chambers. Based on compact and discriminative features, Zhen et al. (2016) proposed to learn a direct regression relationship between image appearance and four-chamber volume segmentation. The latest advancements for cardiac segmentation were dedicated to CNNs which hierarchically extract representations in a data-driven manner, for examples, Tran (2016) employed fully convolutional networks and Wolterink et al. (2016) utilized dilated convolutional neural networks.

3. Methods

The architecture of our proposed 3D DSN is illustrated in Fig. 1, where we take the 3D liver segmentation as an example. In order to achieve efficient end-to-end learning and inference, we first develop a 3D fully convolutional network which enables per-voxel-wise error back-propagation during the training procedure and directly outputting of an equal-sized prediction volume with the input volume during the testing procedure. More importantly, we propose a 3D supervision mechanism and seamlessly integrate it into the fully convolutional network to cope with the optimization difficulties when training such a deep network with limited training data, and hence to accelerate the convergence speed and improve the discrimination performance of the network. We employ a fully connected conditional random field model on top of the score volumes generated from 3D DSN to obtain the segmentation results.

3.1. 3D Convolutional Network

Considering that extracting feature representations across three-dimensional anatomical context is vitally important for volumetric medical image segmentation, we first implement a 3D CNN. Compared with its 2D counterparts, the 3D CNN is capable of encoding representations from volumetric receptive fields, and therefore extracting more discriminative features via richer 3D spatial information. The main components of the 3D CNN are the 3D convolutional layers and 3D sub-sampling (i.e., max-pooling) layers, which are successively stacked as a hierarchical architecture. The feature maps containing neuron activations in each layer are a set of 3D tensors; we refer them as feature volumes hereafter in this paper.

To generate a new feature volume in a convolutional layer, we establish a set of 3D kernels sweeping over the inputs, sum up the activations from these kernels, add a bias term and finally apply a non-linear activation function. The neurons have sparse interactions and the kernel weights are spatially shared, which can greatly reduce the number of parameters and hence alleviate the computational workload of the model. The 3D kernels are learned via the stochastic gradient descent in a data-driven manner, which is the key advancement of convolutional networks compared with traditional pre-defining transformations of hand-crafted features.

In a sub-sampling layer, the output responses from a convolutional layer are further modified by computing the summary statistic of nearby neurons. In our

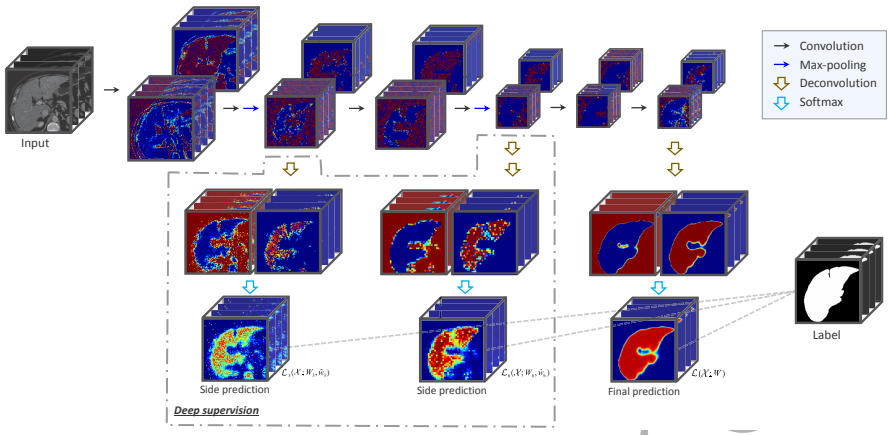


Figure 1: The architecture of the proposed 3D DSN (taking 3D liver segmentation as an example), with intermediate feature volumes, deep supervision layer predictions and last layer score volumes visualized in colormap.

3D max-pooling function, the maximum response within a small cubic neighborhood is selected out and proceeded to subsequent computations. After the pooling operation, the resolution of feature volumes are reduced corresponding to the pooling kernel size. Theoretically, the pooling contributes to make the learned features become invariant to local translations in 3D space, which is a very useful characteristic for image processing (Goodfellow et al. (2016)).

3.2. 3D End-to-End Learning for Volumetric Segmentation

In a classic CNN such as the AlexNet (Krizhevsky et al. (2012)) and the VGGNet (Simonyan & Zisserman (2014)), the last several layers are typically fully-connected ones, which restrict the network to take fixed-sized input image and generate non-spatial predictions (usually a probability vector with a length of the number of target classes). Patch-based methods (Ciresan et al. (2012)) are feasible to provide predictions with spatial information, but they are quite computation-intensive and therefore not practical for volumetric segmentation, which requires dense predictions for every voxel of the 3D images.

In this regard, we cast the fully-connected layer into convolutional layer by re-organizing the parameter weight matrix into high-dimensional convolution kernels (Long et al. (2015)). In this case, the entire network forms a fully convolutional architecture, where all layers are either convolutional or pooling, and both have no restriction on fixed-sized input. In other words, the network is able to input volumetric images with arbitrary sizes, and output spatially arranged classification probability volumes for the entire input images. Therefore, the fully convolutional network successfully eliminates the redundant computations due to overlappings in the patch-based methods.

While the fully convolutional architecture can predict score volumes for arbitrary-sized inputs, the outputs are usually quite coarse with reduced dimen-

sions compared with the original input image due to successive abstractions. In this case, the image voxels receive predictions at a stride corresponding to the setting of pooling layers in the network. However, the segmentation tasks require very dense predictions where each single voxel should obtain a class label. One straightforward way to achieve this is to interpolate the coarse score volumes into full-sized segmentation masks. But an obvious disadvantage of this approach is that it is difficult to determine the interpolation weights and inappropriate weights would introduce imprecise results, especially for the boundary regions.

We alternatively solve this problem using an effective and efficient method. We develop 3D deconvolutional layers to bridge the coarse feature volumes to the dense probability predictions. Specifically, we iteratively conduct a series of $3 \times 3 \times 3$ convolutions with a backwards strided output (e.g., stride of 2 for double size up-scaling). This deconvolution operation can be regarded as a reverse procedure of the convolutions in the forward pass with a corresponding stride. This strategy is quite effective to reconstruct representations from nearby neighborhoods and to up-scale feature volumes to the resolution of original input volumes. Furthermore, these deconvolutional kernels are built in-network and also trainable during the learning process.

Overall, the architecture forms a 3D variant of fully convolutional network which enables efficient end-to-end learning and inference, i.e., inputting a volumetric volume with a arbitrary size and directly outputting an equal-sized prediction score volume (see Fig. 1). To train such a network, the segmentation masks with the same size of the input volumes are employed as the ground truth labels. The optimized loss function is formulated as the sum of the negative log-likelihoods over all the spatial locations of the input images. When we randomly crop training patches from the whole image, we can regard all the spatial components within the input as a training mini-batch for performing stochastic gradient descent. For each iteration of parameter updating, the learning of the 3D network is formulated as a per-voxel-wise classification error back-propagation conducted in a volume-to-volume manner.

Besides the advantage of computation efficiency, the end-to-end learning also benefits the optimization procedure. Previous patch-based methods would restrict the loss to a randomly sampled subset of spatial locations whereas excluding some locations from the gradient computation. In contrast, our end-to-end training with per-voxel-wise error computation would dramatically enlarge the equivalent training database. To the end, the risk of serious over-fitting could be potentially alleviated, which is crucial for many medical applications facing the hardships of insufficiency of training data. Last but not least, the end-to-end network is also economical with regard to storage consumption, since the patches are cropped online and there is no need to pre-save training samples.

3.3. 3D Deep Supervision Mechanism

To segment the organ or structures from the complicated anatomical environments in volumetric medical images, we usually need relatively deep models to encode highly representative features. However, training a deep network is

broadly recognized as a difficult task. One notorious problem is the presence of gradients vanishing or exploding which would make the loss back-propagation ineffective and hamper the convergence of the training process (Glorot & Bengio (2010a)). Particularly, Bradley (2010) found that the back-propagated gradients would become smaller as they move from the output layer towards the input layer during the training. This would make different layers in the network receive gradients with very different magnitudes, leading to ill-conditioning and slower training. The training challenges could be severer in our volumetric medical image segmentation tasks due to the low inter-class voxel variations in medical images, the larger amount of parameters in 3D networks compared with 2D counterparts and the limited training data for many medical applications.

In order to counteract the adverse effects of unstable gradients changes, we propose to exploit explicit supervision to the training of hidden layers in our 3D fully convolutional network. Specifically, we first up-scale some lower-level and middle-level feature volumes using additional deconvolutional layers. Then, we employ the softmax function on these full-sized feature volumes and obtain extra dense predictions. For these branched prediction results, we calculate their classification errors (i.e., negative log-likelihood) with regard to the ground truth segmentation masks. These auxiliary losses together with the loss from the last output layer are integrated to energize the back-propagation of gradients for more effective parameter updating in each iteration.

We call the layers whose feature volumes are directly path-connected to the last output layer as the *mainstream network*. Let w^l be the weights in the lth ($l = 1, 2, ..., L$) layer of the mainstream network, we denote the set of weights in the mainstream network by $W = (w^1, w^2, ..., w^L)$. With $p\left(t_i \mid x_i; W\right)$ representing the probability prediction of a voxel x_i after the softmax function in the last output layer, the negative-log likelihood loss can be formulated as:

$$\mathcal{L}(\mathcal{X}; W) = \sum_{x_i \in \mathcal{X}} - \log p\left(t_i \mid x_i; W\right), \tag{1}$$

where $\mathcal{X}$ represents the training database and t_i is the target class label corresponding to the voxel $x_i \in \mathcal{X}$.

On the other hand, we call the layers which produce auxiliary dense predictions as the *branch networks*. The deep supervision is exactly introduced via these branch networks. To introduce deep supervision from the dth hidden layer, we denote the weights of the first d layers in the mainstream network by $W_d = (w^1, w^2, ..., w^d)$ and use $\hat{w}_d$ to represent the weights which bridge the dth layer feature volumes to dense predictions, and then the auxiliary loss for deep supervision can be formulated as:

$$\mathcal{L}_d(\mathcal{X}; W_d, \hat{w}_d) = \sum_{x_i \in \mathcal{X}} - \log p\left(t_i \mid x_i; W_d, \hat{w}_d\right). \tag{2}$$

Finally, we learn the weights W and all $\hat{w}_d$ using the back-propagation algorithm (LeCun et al. (1989)) by minimizing the following overall objective

function:

$$\mathcal{L} = \mathcal{L}(\mathcal{X}; W) + \sum_{d \in \mathcal{D}} \eta_d \mathcal{L}_d(\mathcal{X}; W_d, \hat{w}_d) + \lambda(\|W\|^2 + \sum_{d \in \mathcal{D}} \|\hat{w}_d\|^2), \tag{3}$$

where η_d is the balancing weight of $\mathcal{L}_d$, which is decayed during learning, and $\mathcal{D}$ is the set of indexes of all the hidden layers which are equipped with the deep supervision. The first term corresponds to the output predictions in the last output layer. The second term is from the deep supervision. The third term is the weight decay regularizations and λ is the trade-off hyperparameter. In each training iteration, the inputs to the network are large volumetric data and the error back-propagations from these different loss components are simultaneously conducted.

The effectiveness of the proposed deep supervision mechanism can be justified from the following two complementary perspectives. First, according to (Lee et al. (2014)), who first proposed to improve the convergence rate and discrimination capability of CNNs for image classification by supervising the training of hidden layers, the deep supervision can directly drive the low- and mid-level hidden layers to favor highly discriminative features towards explicit predictions. In addition, decomposed from these hidden layer features, representations in upper layers can more easily gain superior determinativeness and therefore further boost its generalization capability. Second, introducing such a deep supervision mechanism into a CNN can be considered as adding a kind of shortcut connections (Bishop (1995); Venables & Ripley (2013)) established from the loss to the weights in hidden layers to a CNN, viewing the deconvolutional layers as transformations. Such shortcut connections can improve the prorogation of gradient flows within the network so that the gradient vanishing problem can be greatly alleviated, and therefore obviously enhance the discrimination capability of the networks (He et al. (2015), Srivastava et al. (2015), Szegedy et al. (2015)).

3.4. Contour Refinement with Conditional Random Field

The contour of ambiguous regions can sometimes be imprecise if we only utilize probability thresholding on the score volumes obtained from the 3D DSN. To improve the accuracy of the segmentation results at these regions, we propose to employ a conditional random field (CRF) model (Krähenbühl & Koltun (2012)) to refine the segmentation masks. The model solves the energy function $E(y) = \sum_i -\log \hat{p}(y_i|x_i) + \sum_{i,j} f(y_i, y_j)\phi(x_i, x_j)$, where the first term is the unary potential indicating the distribution over label assignment y_i at a voxel x_i. To aggregate multi-scale information, the $\hat{p}(y_i|x_i)$ is initialized as the linear combination of the last output layer and the branch network predictions obtained from the 3D deeply supervised network:

$$\hat{p}(y_i|x_i) = (1 - \sum_{d \in \mathcal{D}} \tau_d)\, p(y_i|x_i; W) + \sum_{d \in \mathcal{D}} \tau_d\, p(y_i|x_i; W_d, \hat{w}_d). \tag{4}$$

The second term in $E(y)$ is the pairwise potential, where $f(y_i, y_j)$=1 if $y_i \neq y_j$, and 0 otherwise; the $\phi(x_i, x_j)$ incorporates the local appearance and smoothness

by employing the gray-scale value I_i and I_j and bilateral position s_i and s_j of the voxel x_i and x_j, as follows:

$$\phi(x_i, x_j) = \mu_1 \exp(-\frac{\|s_i - s_j\|^2}{2\theta_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\theta_\beta^2}) + \mu_2 \exp(-\frac{\|s_i - s_j\|^2)}{2\theta_\gamma^2}). \quad (5)$$

The constant weights τ_d in the unary potential and parameters $\mu_1, \mu_2, \theta_\alpha, \theta_\beta, \theta_\gamma$ in the pairwise potential were optimized using a grid search on the training set.

4. Experimental Datasets and Materials

We extensively validated the proposed 3D DSN with two challenging segmentation tasks, i.e., liver segmentation from abdomen 3D CT scans and heart segmentation from 3D MR images, by participating two well-known challenges held in conjunction with MICCAI. In this section, we will introduce the datasets employed in our experiments. In section 5, we shall report our extensive experiments for analyzing the effectiveness of the 3D deep supervision mechanism. In section 6, we shall comprehensively present the results of these two challenges.

4.1. Liver Segmentation Dataset

To validate our proposed method on the application of liver segmentation, we employed the SLiver07 (Heimann et al. (2009b)) dataset, which was from the *Segmentation of Liver Competition* held in conjunction with MICCAI 2007, and the grand challenge remained open afterwards. The dataset totally consisted of 30 CT scans with 20 training and 10 testing. All the CT images were acquired contrast-dye-enhanced in the central venous phase. Depending on the machine and protocol used, the pixel spacing varied from 0.55 to 0.80 mm in the transverse plane, and the slice distance varied from 1 to 3 mm. We normalized the intensities of CT images into the range of $[0, 1]$. During training, we randomly cropped patches of size $160 \times 160 \times 72$, and performed rotation augmentations of $[90°, 180°, 270°]$ in the transverse plane. Most of the CT scans were pathologic and included tumors, metastasis and cysts of different sizes. The referenced segmentation masks were defined as the entire liver tissue including all internal structures like vessel systems, tumors, etc., and were manually delineated by radiologist experts. The ground truths of the training set were released to competitors to develop their methods, and ground truths of the testing set are held out by the challenge organizers for independent evaluation.

4.2. Heart Segmentation Dataset

To validate our proposed method on the application of heart segmentation, we employed the dataset of MICCAI 2016 Challenge on *Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*, for short, the HVSMR challenge. The dataset overall consisted of 20 axial, cropped images with 10 training and 10 testing. The cardiovascular MR images were acquired in an axial view on a 1.5T scanner without contrast

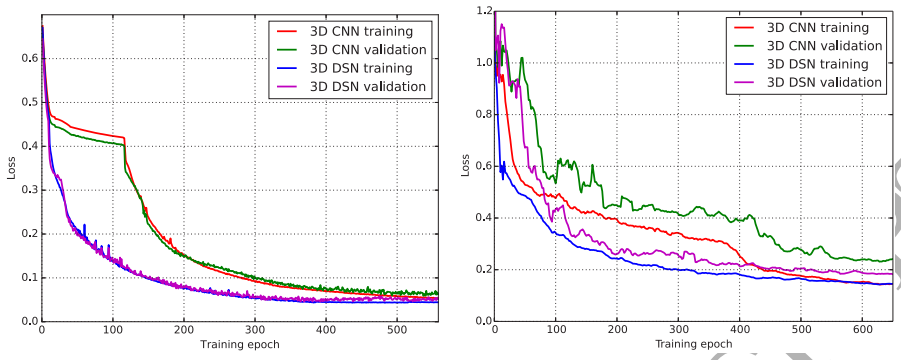


Figure 2: Comparison of learning curves of the 3D DSN and the pure end-to-end 3D CNN without the deep supervision mechanism. The left and right figures are from liver and heart segmentation datasets, respectively.

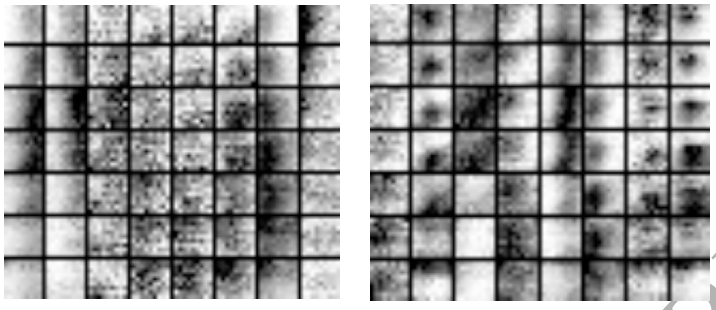
agent using a steady-state free precession pulse sequence. The image dimension and spacing varied across subjects with an average of $390 \times 390 \times 165$ and $0.90 \times 0.90 \times 0.85$ mm, respectively. All the MR images were normalized to have zero mean and unit variance. We utilized data augmentations including random rotations of $[90°, 180°, 270°]$ and flipping along the axial plane. Some subjects had congenital heart defects and some had undergone interventions. The task of the challenge was to segment the blood pool and myocardium from a 3D cardiovascular MR volume. The blood pool class included the left and right atria, left and right ventricles, aorta, pulmonary veins, pulmonary arteries, and the superior and inferior vena cava. Vessels (except the aorta) were extended only a few centimeters past their origin. The segmentations of the blood pool and ventricular myocardium were manually delineated by a trained rater, and validated by two clinical experts. The ground truths of the training set were released to competitors, and those of the testing are held out by the challenge organizers for independent evaluation.

5. Analysis of 3D Deep Supervision

In this section, we experimentally analyze the effectiveness of the 3D deep supervision mechanism as well as the end-to-end training strategy. To conduct the experiments, for each dataset, we split the images with ground truths masks into validation and training subsets in order to more clearly analyze the learning process and compare the segmentation results. We constructed a baseline model which was a pure 3D end-to-end network without any deep supervision. This baseline model had identical architecture as the mainstream network of 3D DSN.

5.1. Learning Curves

We first analyze the learning process of the proposed 3D DSN and the baseline model without deep supervision. As shown in Fig. 2, in all cases, the validation loss consistently decreases as the training loss goes down, demonstrating



(a) 3D CNN (b) 3D DSN

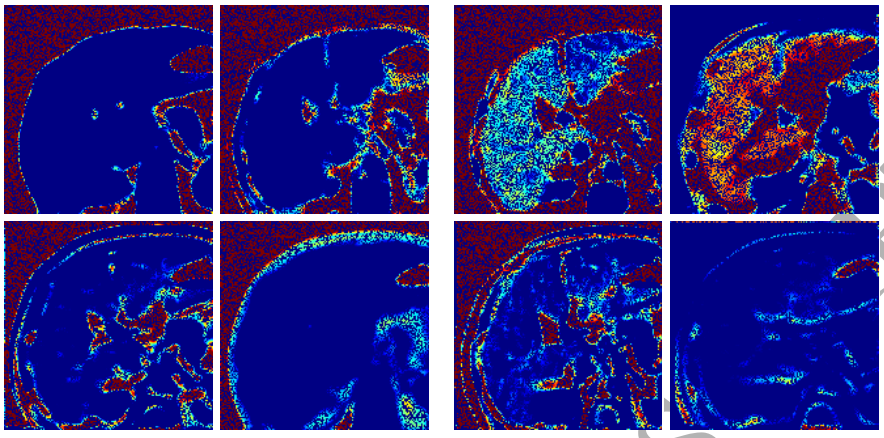
Figure 3: Visualization of the learned 3D kernels in the first layer of (a) 3D CNN and (b) 3D DSN, each column presents a single kernel of size $9 \times 9 \times 7$ expanded along the third dimension as seven 9×9 maps.

that no serious over-fitting is observed even with such small datasets. These results present the effectiveness of the voxel-to-voxel error back-propagation learning strategy benefiting from the 3D fully convolutional architecture. By regarding each single voxel as an independent training sample, the actual training database is dramatically enlarged and therefore the risk of over-fitting can be alleviated compared with traditional patch-based training schemes.

When comparing the learning curves of the 3D DSN and the pure 3D CNN, the 3D DSN converges much faster than the pure 3D CNN. Particularly during the early learning stage, the loss of 3D DSN reduces much faster than the pure 3D CNN. This trend is especially obvious in the left figure in Fig. 2, where the 3D DSN successfully conquers the difficulty of vanishing/exploding gradients in the beginning and achieves a steady decrease of loss, whereas the 3D CNN experiences a plateaus without effective update of parameters (Glorot & Bengio (2010b)). These results demonstrate the proposed 3D deep supervision mechanism can effectively speed up the training procedure by overcoming optimization difficulties through managing the training of both upper and lower layers in the network. Furthermore, it is also observed that the 3D DSN finally achieves lower training as well as validation loss, which is more discernible for the heart segmentation task. This corroborates that the 3D deep supervision mechanism can also improve the discrimination capability of the network. Since the heart segmentation task is a multi-class labeling problem which is more complex than binary classification, the superior efficacy of 3D deep supervision is more visible compared with the liver segmentation task.

5.2. Visualization of 3D Network Kernels and Features

Next, we visualize the intermediate results of the neural networks trained on the liver segmentation dataset to validate the effectiveness of the deep supervision on early layers in the training process. Fig. 3 visualizes the learned 3D kernels of the first convolutional layer with (a) from the pure end-to-end 3D CNN and (b) from the 3D DSN. Each column presents a single kernel sized



(a) 3D CNN (b) 3D DSN

Figure 4: Visualization of typical features in the first layer of (a) 3D CNN and (b) 3D DSN. Slices are extracted from the feature volumes for clear visualization.

$9 \times 9 \times 7$, and we expand it as seven 9×9 2D maps for clear visualization. Our networks in this layer employ eight 3D kernels, so here we visualize all the learned kernels of the layer. It is observed that the kernels of 3D DSN have learned clearer and more organized oriented patterns compared with the pure 3D CNN without any auxiliary supervision. In Fig. 4, we further visualize the set of corresponding feature volumes produced by these 3D kernels in the first convolutional layer, which are explicitly displayed in a slice-wise manner. We extract one slice from each feature volume and here we sample four feature volumes for visualization. We can find that the extracted features by the 3D DSN present less correlations than the pure 3D CNN, which can indicate a superior representative capability of the features from 3D DSN (Lee et al. (2014)). Here, we do not visualize the learned kernels for heart segmentation, because we employed small $3 \times 3 \times 3$ kernels for this task, which are unattainable to intuitively interpret due to the small size.

5.3. Qualitative Comparison of Heart Segmentation Results

Finally, we qualitatively evaluate the efficacy of the 3D deep supervision mechanism on the heart segmentation task. Fig. 5 visually compares the segmentation results obtained from the 3D DSN and pure end-to-end 3D CNN. The first three rows present results from three view directions, i.e., sagittal plane, transverse plane and coronal plane from top to down; and the last row presents 3D reconstructions of the volumetric results. The first column is the raw cardiac MR image; the second and third columns are results from 3D CNN and 3D DSN, respectively; the fourth column shows the ground truth segmentation mask. It is observed that the 3D CNN is able to produce acceptable results which can already delineate general boundaries for the great vessels of the heart, indicating

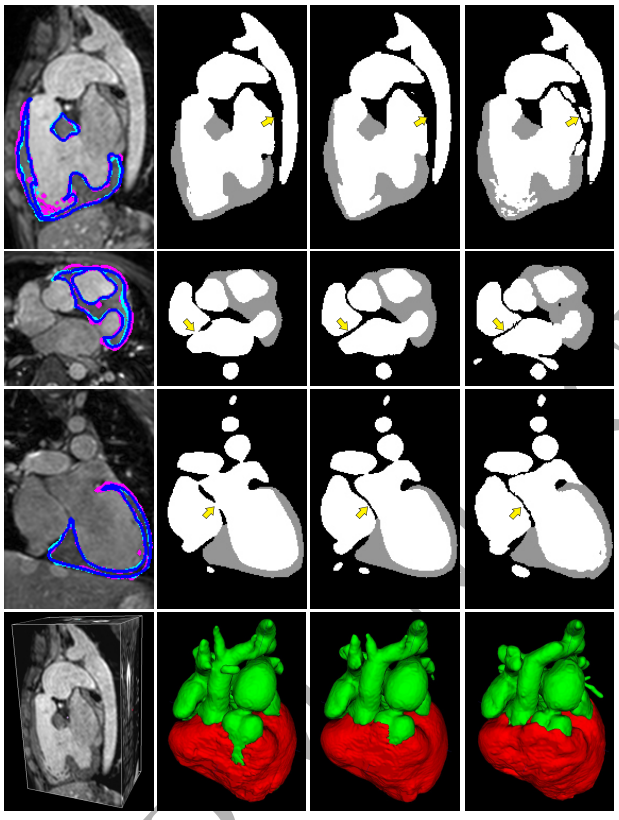


Figure 5: Qualitative comparison of heart segmentation results with and without 3D deep supervision. Columns from left to right are the raw MR images, pure 3D CNN results, 3D DSN results and the ground truth masks. The white and gray regions denote structures of the blood pool and myocardium, respectively. In the first column, we overlay the contours of myocardium with the pure 3D CNN results, 3D DSN results and the ground truth masks indicated in cyan, blue, and magenta colors, respectively. The last row presents 3D reconstructions of the data and segmentations with red for blood pool and green for myocardium.

the effectiveness of the 3D fully convolutional architecture. By leveraging deep supervision, the 3D DSN can generate more precise boundaries for the blood pool, see the yellow arrows in Fig. 5. In addition, the 3D DSN demonstrates sensible superior performance when segmenting the myocardium regions, which is the most challenging element in this application. For examples, observing the first column, the contours of 3D DSN results (blue lines) are more closer to the ground truth (magenta lines) than contours of the pure 3D CNN results (cyan lines). In addition, viewing the 3D reconstruction results, the 3D DSN presents more accurate segmentations for the myocardium which coincide well with the ground truth, whereas the pure 3D CNN misclassifying some myocardium tissues as blood pool or background.

6. Experimental Segmentation Results

6.1. Liver Segmentation from 3D CT Volumes

6.1.1. Network Architecture and Training Settings

Our 3D deeply supervised network for liver segmentation consisted of 11 layers, i.e., 6 convolutional layers, 2 max-pooling layers, 2 deconvolutional layers and one softmax output layer, in the mainstream network. To be specific, after the input layer, we first stacked two convolutional layers (i.e., conv1a and conv1b both with size $8@9\times9\times7$), and one max-pooling layer (i.e., pool1 with down-sampling stride of 2). Next, we employed another two convolutional layers (i.e., conv2a $16@7\times7\times5$ and conv2b $32@7\times7\times5$), and one max-pooling layer (i.e., pool2 with down-sampling stride of 2). Then, two more convolutional layers (i.e., conv3a $32@5\times5\times3$ and conv3b $32@1\times1\times1$) were connected. We designed relatively large kernel sizes for the down-sampling path in order to form a proper receptive field to recognize the liver region.

Since we max-pooled twice, we utilized two deconvolutional layers via which to obtain full-sized prediction score volume. A softmax layer was used to obtain the prediction probabilities. We injected the 3D deep supervision via two layers (i.e., pool1 and pool2), whose feature volumes were up-scaled for auxiliary classifiers. The numbers of deconvolutional layers in the branch networks were determined by the reduced size of hidden feature volumes. In our implementation, one and two deconvolutional layers followed the pool1 and pool2 layer, respectively.

We trained the network from scratch with weights initialized from Gaussian distribution ($\mu=0, \sigma=0.01$). The learning rate was initialized as 0.1 and divided by 10 every fifty epochs. The deep supervision balancing weights were initialized as 0.3 and 0.4, and then decayed as training going on. We cropped patches of size $160 \times 160 \times 72$ as input to the network and the training was stopped when the validation accuracy did not increase anymore.

6.1.2. Segmentation Evaluation Metrics

The challenge employed five evaluation metrics based on volumetric overlap and surface distances. The five measurements include volumetric overlap error (VOE[%]), relative volume difference (VD[%]), average symmetric surface distance (AvgD[mm]), root mean square symmetric surface distance (RMSD[mm]) and maximum symmetric surface distance (MaxD[mm]).

The volumetric overlap error and the relative volume difference are volumetric measurements given in percentage. Denoting our segmentation results by R and the ground truth masks by G, the VOE and VD are calculated as:

$$\begin{aligned} VOE(R,G) &= (1-\frac{|R\cap G|}{|R\cup G|})\cdot 100\%, \\ VD(R,G) &= \frac{|R|-|G|}{|G|}\cdot 100\%. \end{aligned} \tag{6}$$

The VOE calculates the ratio between the intersection and union of the produced results and reference masks. This metric is correlated with the Jaccard coefficient. A value of 0 represents a perfect segmentation and 100 if the results and reference have no overlap at all. The VOE is one of the most popular and important metric to evaluate the accuracy of segmentation results (Heimann et al. (2009b)). On the other hand, the VD non-symmetrically calculates the difference of volumes in the results and reference. A value of 0 represents that both volume sizes are equal, but does not imply that R and G are identical. Combined with other measurements, the VD reveals if a method tends to over- (positive number) or under- (negative number) segment the image.

The average symmetric surface distance, the root mean square symmetric surface distance and the maximum symmetric surface distance are surface based measurements given in millimeters. If a voxel has at least one non-object voxel within its 18-neighborhood, then this voxel is regarded as a surface voxel (Heimann et al. (2009b)). For each surface voxel of segmentation results R, the Euclidean distance to the closest surface voxel from the ground truth G is calculated, and vise versa. Let S(R) denote the set of surface voxels of R, the shortest distance of an arbitrary voxel v to $S(R)$ is defined as $d(v, S(R)) = min_{s_R \in S(R)} ||v - s_R||$. Based on this, the average symmetric surface distance is calculated as:

$$AvgD(R,G) = \frac{1}{|S(R)| + |S(G)|} \Big(\sum_{s_R \in S(R)} d(s_R, S(G)) \quad + \sum_{s_G \in S(G)} d(s_G, S(R)) \Big) \tag{7}$$

The AvgD is the most popular and important surface based metric in segmentation evaluations, which is of similar significance to the VOE for volumetric evaluation. In combination with AvgD, when using the squared Euclidean distance, the root mean square symmetric surface distance is calculates as:

$$RMSD(R,G) = \sqrt{\frac{1}{|S(R)| + |S(G)|}} \cdot \sqrt{\sum_{s_R \in S(R)} d^2(s_R, S(G)) \quad + \sum_{s_G \in S(G)} d^2(s_G, S(R))} \tag{8}$$

When using the maximum symmetric distance, the MaxD is calculated as:

$$MaxD(R,G) = \max \Big\{ \max_{s_R \in S(R)} d(s_R, S(G)), \max_{s_G \in S(G)} d(s_G, S(R)) \Big\} \tag{9}$$

This metric is also known as Hausdorff distance (Huttenlocher et al. (1993)), and it is sensitive to outliers because the maximum error is counted. For all the five measurements, lower absolute values indicate better segmentation performance.

6.1.3. Segmentation Results

On the SLiver07 dataset, we conduct a series of experiments to analyze the contributions of different components in the proposed method. We first compare with the results of a baseline pure end-to-end 3D network (the same configuration as reported in Section 5) and the proposed 3D DSN. We then add CRF (parameters of $\mu_1 = 5$, $\theta_\alpha = 20$, $\theta_\beta = 30$, $\mu_2 = 3$, $\theta_\gamma = 3$) as a post-processing step to these two networks.

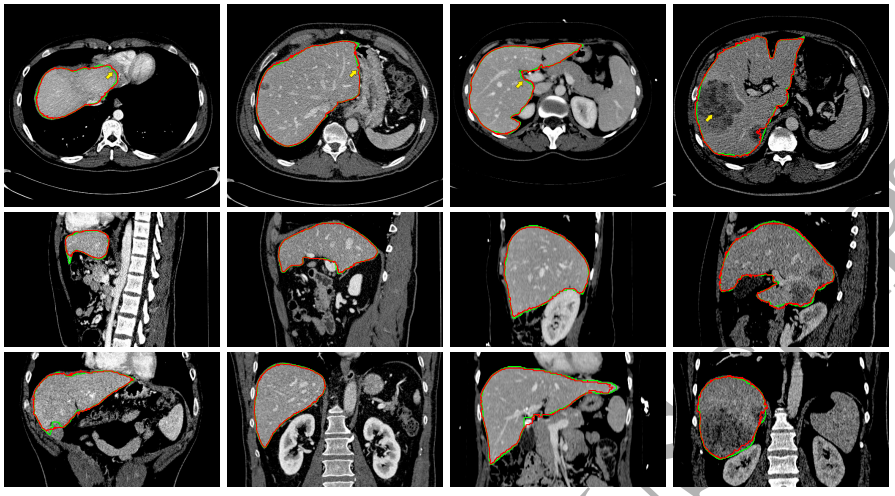
Table 1: Quantitative evaluation results of our methods under different settings.

Methods	VOE[%]	VD[%]	AvgD[mm]	RMSD[mm]	MaxD[mm]
3D-CNN	7.68±1.02	1.98±1.75	1.56±0.21	4.09±0.41	45.99±21.78
3D-DSN	6.27±0.89	1.46±1.56	1.32±0.18	3.38±0.24	36.49±19.65
3D-CNN+CRF	5.64±0.77	1.72±1.43	0.89±0.15	1.73±0.37	34.42±17.23
3D-DSN+CRF	**5.37±0.73**	**1.32±1.35**	**0.67±0.12**	**1.48±0.21**	**29.63±16.31**

In this regard, we totally get four settings and we call them 3D-CNN, 3D-DSN, 3D-CNN+CRF and 3D-DSN+CRF, respectively. All experiments were conducted on the training set using leave-one-out strategy.

Experimental results of various metrics are listed in Table 1. Comparing results of the 3D-CNN and the 3D-DSN (the first two rows), we observe that the volumetric overlap error of 3D-DSN reduced to 6.27% from 7.68% of 3D-CNN, and the distance measurements were also reduced by introducing the 3D deep supervision. This demonstrates that the deep supervision is able to enhance the discrimination capability of the network through strengthening discriminativeness of intermediate layers and alleviating gradients exploding/vanishing. Furthermore, based on the high-quality unary potential produced by the deep 3D networks, the CRF model further improves the segmentation accuracy and the surface distance based metrics see more significant improvements from the graphical model. For example, the average symmetric surface distance reduced from 1.32 *mm* to 0.67 *mm* after performing CRF on top of 3D-DSN. This post-processing step has potential significance for further processing such as reconstruction and visualization. Typical examples of liver segmentation results using our framework are shown in Fig. 6. Leveraging the high-level discriminative representations learned from rich 3D contextual information, our 3D-DSN+CRF method can successfully delineate the liver from adjacent anatomical structures with low intensity contrast (Fig. 6 (a)), conquer the large inter-patient shape variations (Fig. 6(b) and (c)), and handle the internal pathologies with abnormal appearance (Fig. 6(d)).

We also validated our method on the testing set and submitted our results to challenge organizers for evaluation. Table 2 compares our method with the top-ranking teams in the on-site competition (Kainmüller et al. (2007); Heimann et al. (2007)) as well as published state-of-the-art approaches (Al-Shaikhli et al. (2015); Wimmer et al. (2009)) on the current leaderboard. The method of ZIB-Charite (Kainmüller et al. (2007)) was based on a constrained free-form and statistical deformable model relying on the typical intensity distribution around the liver boundary and the neighboring anatomical structures. The method of MBI@DKFZ (Heimann et al. (2007)) first used an evolutionary algorithm to provide the initial parameters for a statistical shape model, and then utilized a deformable mesh which tried to equilibrium between internal and external forces. This method was proposed in the year 2007 and the same team later made a new submission in the year 2016 reporting better results. For fair comparison, we also include their new segmentation results here, denoting by MBI-DKFZ (2016) in Table 2. The new results were produced with



(a) (b) (c) (d)

Figure 6: Examples of segmentation results of our proposed method. The ground-truths are denoted in green, and our results are in red. Each column corresponds to a subject with three view planes, i.e., transverse, sagittal and coronal planes, from top to bottom.

Table 2: Comparison with different approaches on the liver segmentation testing set.

Teams	VOE[%]	VD[%]	AvgD[%]	RMSD[%]	MaxD[%]	Runtime
ZIB-Charite (Kainmüller et al. (2007))	6.09±2.02	-2.86±2.76	0.95±0.33	1.87±0.76	18.69±8.02	15 mins
MBI@DKFZ (Heimann et al. (2007))	7.73±1.81	1.66±3.02	1.39±0.39	3.25±1.28	30.07±9.70	7 mins
MBI-DKFZ (2016)	5.90±1.00	1.17±1.82	0.98±0.17	2.08±0.48	21.63±4.47	-
TNT-LUH (Al-Shaikhli et al. (2015))	6.44±0.66	1.53±1.67	0.95±0.08	**1.58±0.25**	**15.92±4.85**	-
LME Erlangen (Wimmer et al. (2009))	6.47±0.92	**1.04±2.69**	1.02±0.16	2.00±0.35	18.32±4.66	-
Ours (3D-DSN+CRF)	**5.42±0.72**	1.75±1.77	**0.79±0.14**	1.64±0.37	33.55±19.64	**1.5 mins**

Note: the - means that runtime was not reported.

an extended 3D statistical shape model, which utilized landmark-wise random regression forests for a non-local appearance modeling, then following an omni-directional landmark detection. The method of TNT-LUH (Al-Shaikhli et al. (2015)) embedded sparse representations of both region-based global information and voxel-wise local information into a level set formulation. The method of LME Erlangen (Wimmer et al. (2009)) relied on a combination of a nearest neighbor boundary appearance model for region information and a Parzen density estimation based shape modeling.

From Table 2, we can observe that our method has achieved an exceeding VOE of 5.42% and AvgD of 0.79 *mm*, which are the two most important and commonly used evaluation measurements (Heimann et al. (2009b)). The second leading value of VOE is 5.90% from MBI-DKFZ (2016) which is 0.48% behind ours. We also achieved satisfactory performance on the root mean square symmetric surface distance with 1.64 *mm*, which is the second highest performance following the 1.58 *mm* from TNT-LUH (Al-Shaikhli et al. (2015)). Our method did not perform well on the metric of maximum symmetric surface distance,

because the MaxD is very sensitive to outliers, and we did not adopt shape priors in our method. Overall, our method achieved better or comparable results when compared with these state-of-the-art algorithms.

Considering the potential requirements for intraoperative planning and guidance, we also recorded the time performance of our method. Our framework took about 1.5 minutes to handle a CT volume using GPU acceleration. In fact, our 3D network was very fast and most of the time was spent by the CRF model. Previous methods usually took several minutes or even longer to process one subject as reported in their papers, but some of them did not harness GPU for acceleration.

6.2. Heart Segmentation from 3D MR Volumes

6.2.1. Network Architecture and Training Settings

We then evaluate the proposed 3D DSN on the heart segmentation application. The network architecture employed were similar to the one used in liver segmentation with some slight differences adapted for special characteristics of the structures. Specifically, we constructed a 14-layer network stacking 7 convolutional layers, 3 max-pooling layers, 3 deconvolutional layers and one softmax output layer in the mainstream network. The detailed down-sampling path was input-conv1a-pool1-conv2a-conv2b-pool2-conv3a-conv3b-pool3-conv4a-conv4b. All the convolutional layers employed small kernels of $3\times3\times3$, considering the small structures of myocardium. For the number of feature volumes, conv1a had 32 kernels; conv2a and conv2b had 64 kernels; conv3a and conv3b had 128 kernels; conv4a and conv4b had 256 kernels. In order to form a competent receptive field for the blood pool, we utilized 3 max-pooling layers with a down-sampling stride of 2. In the upsampling path, we employed 3 deconvolutional layers to learn the dense predictions. To perform 3D deep supervision, we connected the layers of conv2b and conv3b to auxiliary classifiers.

The network was trained from scratch with weights initialized from Gaussian distribution ($\mu = 0, \sigma = 0.01$). Considering the large variance of the heart segmentation dataset, we utilized batch normalization (Ioffe & Szegedy (2015)) to reduce the internal covariance shift within the network's hidden neurons. The learning rate was initialized as 0.01 and decayed using the "poly" learning rate policy (Liu et al. (2015)). The deep supervision balancing weights were initialized as 0.2 and 0.4 and decayed during training procedure. We cropped patches of size $64\times64\times64$ as input to the network, considering consumption of the GPU memory, and the training was stopped when the validation accuracy did not increase anymore.

6.2.2. Segmentation Evaluation Metrics

The HVSMR challenge adopted seven evaluation criteria including the Dice coefficient (Dice), Jaccard coefficient (Jac), positive predictive value (PPV), sensitivity (Sens), specificity (Spec), average distance of boundaries (Adb[mm]) and Hausdorff distance of boundaries (Hdb[mm]), which are calculated for the structures of blood pool and myocardium, respectively.

Table 3: Comparison with different approaches on heart segmentation task. The evaluations of blood pool and myocardium are listed in top and bottom, respectively.

Methods	Dice	Jac	PPV	Sens	Spec	Adb[mm]	Hdb[mm]
Shahzad et al. (2016)	0.885±0.028	0.795±0.044	0.907±0.052	0.867±0.046	0.984±0.008	1.553±0.376	9.408±3.059
Wolterink et al. (2016)	0.926±0.018	0.863±0.030	0.951±0.024	0.905±0.047	**0.992±0.004**	**0.885±0.223**	**7.069±2.857**
Tziritas (2016)	0.867±0.047	0.768±0.068	0.861±0.062	0.889±0.108	0.972±0.014	2.157±0.503	19.723±4.078
Mukhopadhyay (2016)	0.794±0.053	0.661±0.071	**0.964±0.035**	0.680±0.081	0.996±0.004	2.550±0.996	14.634±8.200
3D U-Net (Çiçek et al. (2016))	0.926±0.016	0.863 ±0.028	0.940±0.028	0.916±0.048	0.989±0.005	0.940±0.193	8.628±3.390
Ours	**0.928±0.014**	**0.865±0.023**	0.934±0.024	**0.924±0.039**	0.988±0.003	1.017±0.181	7.704±2.892
Shahzad et al. (2016)	0.747±0.075	0.602±0.094	0.767±0.054	0.734±0.108	0.989±0.004	1.099±0.204	**5.091±1.658**
Wolterink et al. (2016)	**0.802±0.060**	**0.673±0.084**	0.802±0.065	**0.805±0.076**	0.990±0.004	**0.957±0.302**	6.126±3.565
Tziritas (2016)	0.612±0.153	0.457±0.149	0.666±0.164	0.571±0.150	0.985±0.008	2.041±1.022	13.199±6.025
Mukhopadhyay (2016)	0.495±0.126	0.338±0.110	0.546±0.134	0.462±0.142	0.980±0.007	2.596±1.358	12.796±4.435
3D U-Net (Çiçek et al. (2016))	0.694±0.076	0.536±0.089	0.798±0.076	0.618±0.092	0.992±0.004	1.461±0.397	10.221±4.339
Ours	0.739±0.072	0.591±0.090	**0.856±0.054**	0.653±0.089	**0.994±0.002**	1.035±0.240	5.248±1.332

The Dice and Jaccard coefficients are closely related with the VOE measurement described in Section 6.1.2. They also measure the spatial overlap between the segmentation results R and the ground truth masks G, with the calculations slightly different. Specifically, the Dice and Jaccard are defined as follows:

$$\begin{aligned} Dice(R,G) &= \frac{2|R \cap G|}{|R| + |G|}, \\ Jac(R,G) &= \frac{|R \cap G|}{|R \cup G|}. \end{aligned} \tag{10}$$

Larger values of the Dice and Jaccard coefficients indicate higher segmentation accuracy. In addition, three more ratios (i.e., PPV, Sens, and Spec) also belong to the volume based measurements. The PPV is the ratio of true positives to true positives plus false positives. The sensitivity represents the ratio of true positives to true positives plus false negatives. The specificity denotes the ratio of true negatives to true negatives plus false positives. For these three ratios, higher values indicate better segmentation performance.

The remaining two metrics are both used in the liver segmentation evaluation as detailedly described in Section 6.1.2. The measure Adb[mm] is the same as the AvgD[mm], which symmetrically calculates the average surface distance of segmentation results and ground truth masks. The measure Hdb[mm] is the same as the MaxD[mm], which counts the maximum distance between the results and ground truth surfaces. For both Adb[mm] and Hdb[mm], the lower the distance values, the better the segmentation performance.

6.2.3. Segmentation Results

Table 3 presents the heart segmentation results on the MICCAI2016 HVSMR testing dataset. The top part lists results of the blood pool segmentation and the bottom part shows the results of the myocardium segmentation. We compared our method with representative approaches from other participating teams in the challenge, which employed either traditional segmentation methods or machine learning based methods. Specifically, Shahzad et al. (2016) developed an automated algorithm by combining multi-atlases and level-sets; Tziritas (2016) utilized a 3D Markov random field model combined with substructures tracking. The other two belong to machine learning based methods with Mukhopadhyay

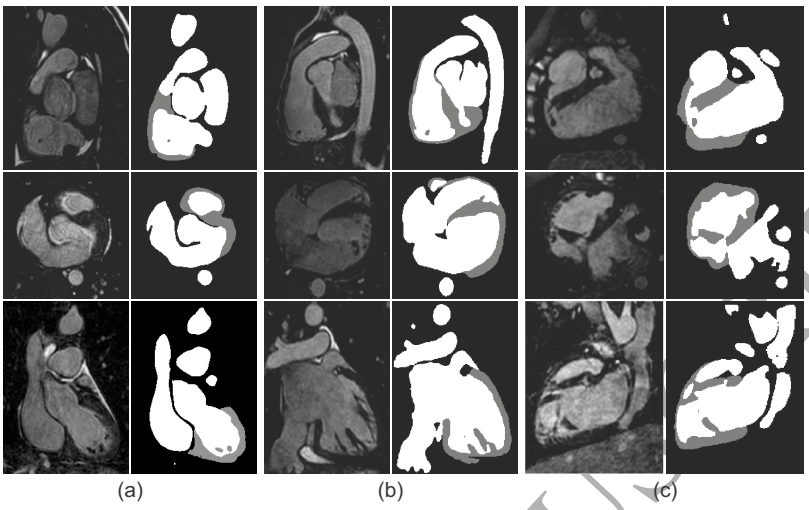


Figure 7: Typical heart segmentation results using our method. We present three view directions, with sagittal, transverse and coronal planes listed from top to down. The white and gray regions denote segmentations for blood pool and myocardium, respectively.

(2016) leveraging random forest variants and the Wolterink et al. (2016) utilizing 2D dilated convolutional networks (Yu & Koltun (2016)). In addition, we conducted comparison experiments using the 3D U-Net (Çiçek et al. (2016)) which also employs the 3D fully convolutional network. Specifically, we implemented the 3D U-Net architecture and obtained optimal segmentation results by carefully tuning the model on our HVSMR dataset. We submitted the results to the challenge evaluation system to get the scores listed in Table 3.

In the blood pool segmentation task, our method achieved Dice of 0.928 and Jaccard of 0.865, outperforming the other participating teams. The average distance of boundaries and Hausdorff distance of our method were also quite competitive, approaching the highest performance from the Wolterink et al. (2016). For the results of myocardium segmentation, our method presented the best performance on positive predictive value and specificity, with promising performance on average distance and Hausdorff distance of boundaries. When comparing with 3D U-Net, our proposed method achieved a higher performance on segmentation of both the myocardium and blood pool. Nevertheless, the 3D U-Net results of the blood pool are quite close to ours, while better than most of other baseline approaches. This observation can validate the effectiveness of 3D FCN on volumetric medical image segmentations. For time performance, our method takes around 1 minute to handle a MR volume. The Fig. 7 presents some typical heart segmentation results on the testing dataset, from top to down are views from the sagittal plane, transverse plane and coronal plane, respectively. For each subject, the left column are raw image data and the right are our segmentation results. We can observe that our method successfully

delineated the anatomical structures of myocardium and blood pool. Note that there exists a large variation in the testing dataset. For example, the case of Fig. 7 (c) comes with an inverse orientation from other cases. Under this challenging situation, our method can still discriminate the characteristics of both anatomical structures and produce accurate segmentation masks, with blood pool Dice of 0.903 and myocardium Dice of 0.646. Meanwhile, as reported in Table 3, the standard deviations of our method are usually smaller than those of other approaches, somehow demonstrating the stability and generalization capability of our model.

7. Discussion

We present a novel and efficient method for volumetric medical image segmentation, which is a fundamental yet challenging problem in medical image analysis. Extensive studies have been dedicated to developing various algorithms to address this challenging problem, including level sets, statistical shape models, multi-atlas based methods, graph cuts, and so on. Although remarkable achievements have been attained in past decades, the main shortcoming of these traditional methods is that most of them are developed based on hand-crafted features, which greatly limit their transferability and generalization ability. For example, the shape priors derived for liver segmentation may not be applicable for cardiac image segmentation; reciprocally, the atlas constructed for heart segmentation cannot be applied to liver segmentation. In addition, the intensity distribution gap between different imaging modalities is also a disincentive to generalize those traditional methods constructed based on statistical models of intensity distribution. In recent years, deep learning techniques, especially deep convolutional neural networks, have been emerging as a competitive alternative to resolve complicated medical image analysis tasks. In contrast to traditional methods, these techniques leverage the highly representative features learned from labeled training datasets. Such a data-driven learning process is readily generalizable among different datasets as well as different imaging modalities without many elaborations. In this regard, it is promising to construct more general algorithms for medical image analysis tasks based on deep learning techniques and the successful utilization of the proposed 3D DSN for two distinct volumetric medical image segmentation tasks with different imaging modalities demonstrates this compelling hallmark of these techniques.

One of the main challenges of harnessing CNNs for medical image analysis tasks is that, compared with natural image applications, medical applications usually have limited training data. Although the medical image datasets used in our two applications are not very large in subject-level, we still achieve competitive performance to state-of-the-art methods. This can be attributed to the advantage of the fully convolutional architecture. As we perform per-voxel-wise classification error back-propagation during the training, each single voxel in the volumetric data is treated as an independent training sample. Hence, the training database is extensively enlarged compared with traditional patch-based methods when viewing from the voxel-level rather than the subject-level. For

example, for an input training sub-volume of size 160×160×72, there are almost two million voxels whose classification errors are counted into the loss function. When we randomly cropping training sub-volumes from the whole image, all voxels in the image are stochastically considered during the training procedure, and therefore can alleviate the risk of over-fitting.

Another challenge of leveraging CNNs for volumetric medical image segmentation is that while we hope to utilize a relatively deeper network to capture more representative features for more accurate segmentation, we usually confront the optimization difficulties when a network goes deeper. To the end, we propose a deep supervision mechanism to resolve these difficulties by formulating an objective function that can guide the training of low- and mid-level layers in a direct and unmediated way and hence energize the prorogation of information flows within the network so that more representative features can be learned from the training data. Such a mechanism can further alleviate the problem of limited training data (from the perspective of subject-level) by tapping the potentials of the data to generate more powerful and generalizable features.

For the overall architecture of our 3D DSN, we choose to introduce auxiliary classifiers in a sparse manner rather than a dense manner as Merkow et al. (2016) did. Such a sparse manner can help manage the scale of the model, as more additional deconvolutional layers would bring in larger number of the parameters. Note that if a model is too complicated, its efficiency may be inhibited as it is difficult to train a model with massive parameters with limited training data. In addition, it would be not easy to adjust the balancing weights η for dense supervision during training.

In practice, we need to contemplate that different anatomical organs and structures have significant variations in sizes, shapes and textures, which may cause some slight adjustments of our network configurations. For example, the network's receptive field is mainly determined based on the size of the targeting object. In this case, we use slightly different network configurations to handle liver segmentation and heart segmentation in our paper. Note that these slight adjustments do not influence the generalization of the proposed 3D DSN. The two networks, in essence, employ the same techniques and share the same design principles.

The conditional random field is built on top of the score volumes obtained from 3D DSN, and therefore the CRF part is separable and flexible in practice. Based on our experimental results, the CRF is helpful to improve the liver segmentation results. However, cardiac image segmentation does not gain great improvements from this post-processing. One possible reason is that the substructures of the heart (i.e., myocardium and blood pool) are much more complex and less continuous than the liver tissues. For example, the myocardium has a surface structure encompassing the branchy vessels with small thickness. Another reason is that the liver segmentation is binary labeling whereas the heart segmentation comes with multi-class fashion. It was difficult to figure out a global set of parameters that make improvements on both myocardium and blood pool. We also tried to process the two classes separately, but was faced with the problem of how to deal with the overlapping regions. The work

of Kamnitsas et al. (2016) also reported similar challenging issues on their 3D extension of the fully connected CRF model. One possible solution for addressing this problem is to cast CRF into the recurrent layers and jointly train it with the 3D DSN Zheng et al. (2015).

8. Conclusion

We propose a 3D deeply supervised network (i.e., 3D DSN) for end-to-end volumetric anatomical structure segmentation from medical images. By upscaling the intermediate coarse feature volumes in-network, we can efficiently perform the dense segmentation in a volume-to-volume manner, where we directly obtain equal-sized output as the input data. Furthermore, we develop a 3D deep supervision mechanism by connecting hidden layer features to auxiliary classifiers. This strategy can not only address the optimization challenges when training deep 3D networks, but also improve the discriminative capability of networks. Finally, based on the high-quality score volumes obtained from 3D DSN, we employ a CRF model to refine the segmentation results. We extensively validate the proposed 3D DSN on two distinct applications and the results demonstrate the effectiveness and generalization of the proposed network. Last but not least, our 3D DSN is very efficient and has great potential to be used in clinical applications requiring timely interactions, such as intraoperative planning and guidance.

Acknowledgments

This work is supported by grants from Hong Kong Research Grants Council General Research Fund (Project no. CUHK 412513 and CUHK 14202514) and a grant from Guangdong Natural Science Foundation (Project no. 2016A030313047).

References

References

Al-Shaikhli, S. D. S., Yang, M. Y., & Rosenhahn, B. (2015). Automatic 3d liver segmentation using sparse representation of global and local image information via level set formulation. *arXiv preprint arXiv:1508.01521*, .

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, *35*, 1313–1321. doi:10.1109/TMI.2016.2528120.

Atehorta, A., Zuluaga, M. A., Ourselin, S., Giraldo, D., & Romero, E. (2016). Automatic segmentation of 4d cardiac mr images for extraction of ventricular chambers using a spatio-temporal approach. URL: http://dx.doi.org/10.1117/12.2217076. doi:10.1117/12.2217076.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Bradley, D. M. (2010). *Learning in modular systems*. Technical Report DTIC Document.

Campadelli, P., Casiraghi, E., & Esposito, A. (2009). Liver segmentation from computed tomography scans: A survey and a new algorithm. *Artificial Intelligence in Medicine*, *45*, 185–196.

Chen, H., Dou, Q., Wang, X., Qin, J., Cheng, J. C., & Heng, P.-A. (2016a). 3d fully convolutional networks for intervertebral disc localization and segmentation. In *International Conference on Medical Imaging and Virtual Reality* (pp. 375–382). Springer International Publishing.

Chen, H., Dou, Q., Yu, L., & Heng, P.-A. (2016b). Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, .

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., & Heng, P.-A. (2016c). Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis*, .

Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., & Zhou, S. K. (2016d). Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 487–495). Springer.

Christ, P. F., Elshaer, M. E. A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., DAnastasi, M. et al. (2016). Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 415–423). Springer.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3d u-net: Learning dense volumetric segmentation from sparse annotation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II* (pp. 424–432). Cham: Springer International Publishing.

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems* (pp. 2843–2851).

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, *61*, 38–59.

Danciu, M., Gordan, M., Florea, C., & Vlaicu, A. (2012). 3d dct supervised segmentation applied on liver volumes. In *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on* (pp. 779–783). IEEE.

Dhungel, N., Carneiro, G., & Bradley, A. P. (2015). Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 605–612). Springer.

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., & Heng, P.-A. (2016a). 3d deeply supervised network for automatic liver segmentation from ct volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 149–157). Springer.

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C., Shi, L., & Heng, P.-A. (2016b). Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE TMI*, *35*, 1182–1195.

Frangi, A. F., Niessen, W. J., & Viergever, M. A. (2001). Three-dimensional modeling for functional analysis of cardiac images, a review. *IEEE transactions on medical imaging*, *20*, 2–5.

Freiman, M., Cooper, O., Lischinski, D., & Joskowicz, L. (2011). Liver tumors segmentation from cta images using voxels classification and affinity constraint propagation. *International journal of computer assisted radiology and surgery*, *6*, 247–255.

Fritscher, K. D., Pilgram, R., & Schubert, R. (2005). Automatic cardiac 4d segmentation using level sets. In *International Workshop on Functional Imaging and Modeling of the Heart* (pp. 113–122). Springer.

Glorot, X., & Bengio, Y. (2010a). Understanding the difficulty of training deep feedforward neural networks. In *Aistats* (pp. 249–256). volume 9.

Glorot, X., & Bengio, Y. (2010b). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS* (pp. 249–256).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2016a). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, .

Havaei, M., Guizard, N., Chapados, N., & Bengio, Y. (2016b). Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 469–477). Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, .

Heimann, T., & Meinzer, H.-P. (2009). Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, *13*, 543–563.

Heimann, T., Meinzer, H.-P., & Wolf, I. (2007). A statistical deformable model for the segmentation of liver ct volumes. In *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge* (pp. 161–166).

Heimann, T., Van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G. et al. (2009a). Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, *28*, 1251–1265.

Heimann, T., Van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A. et al. (2009b). Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, *28*, 1251–1265.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, *15*, 850–863.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, .

Jamaludin, A., Kadir, T., & Zisserman, A. (2016). Spinenet: Automatically pinpointing classification evidence in spinal mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 166–175). Springer.

Jolly, M.-P., Duta, N., & Funka-Lea, G. (2001). Segmentation of the left ventricle in cardiac mr images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (pp. 501–508). IEEE volume 1.

Kainmüller, D., Lange, T., & Lamecker, H. (2007). Shape constrained automatic segmentation of the liver based on a heuristic intensity model. In *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge* (pp. 109–116).

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2016). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *arXiv preprint arXiv:1603.05959*, .

Kaus, M. R., von Berg, J., Weese, J., Niessen, W., & Pekar, V. (2004). Automated segmentation of the left ventricle in cardiac mri. *Medical image analysis*, *8*, 245–254.

Koikkalainen, J., Tolli, T., Lauerma, K., Antila, K., Mattila, E., Lilja, M., & Lotjonen, J. (2008). Methods of artificial enlargement of the training set for statistical shape models. *IEEE Transactions on Medical Imaging*, *27*, 1643–1654.

Krähenbühl, P., & Koltun, V. (2012). Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, .

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kronman, A., & Joskowicz, L. (2016). A geometric method for the detection and correction of segmentation leaks of anatomical structures in volumetric medical images. *International Journal of Computer Assisted Radiology and Surgery*, *11*, 369–380. URL: `http://dx.doi.org/10.1007/s11548-015-1285-z`. doi:`10.1007/s11548-015-1285-z`.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*, 541–551.

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

Li, C., Wang, X., Eberl, S., Fulham, M., Yin, Y., Chen, J., & Feng, D. D. (2013). A likelihood and local constraint level set model for liver tumor segmentation from ct volumes. *IEEE Transactions on Biomedical Engineering*, *60*, 2967–2977.

Li, G., Chen, X., Shi, F., Zhu, W., Tian, J., & Xiang, D. (2015). Automatic liver segmentation based on shape constraints and deformable graph cut in ct images. *IEEE Transactions on Image Processing*, *24*, 5315–5329.

Ling, H., Zhou, S. K., Zheng, Y., Georgescu, B., Suehling, M., & Comaniciu, D. (2008). Hierarchical, learning-based automatic liver segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

Linguraru, M. G., Richbourg, W. J., Liu, J., Watt, J. M., Pamulapati, V., Wang, S., & Summers, R. M. (2012). Tumor burden analysis on computed tomography by automated liver and tumor segmentation. *IEEE transactions on medical imaging*, *31*, 1965–1976.

Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, .

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

Meinzer, H.-P., Thorn, M., & Cárdenas, C. E. (2002). Computerized planning of liver surgeryan overview. *Computers & Graphics*, *26*, 569–576.

Merkow, J., Marsden, A., Kriegman, D., & Tu, Z. (2016). Dense volume-to-volume vascular boundary detection, . (pp. 371–379).

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

Mitchell, S. C., Lelieveldt, B. P., Van Der Geest, R. J., Bosch, H. G., Reiver, J., & Sonka, M. (2001). Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac mr images. *IEEE Transactions on medical imaging*, *20*, 415–423.

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., & Išgum, I. (2016a). Automatic segmentation of mr brain images with a convolutional neural network. *IEEE transactions on medical imaging*, *35*, 1252–1261.

Moeskops, P., Wolterink, J. M., van der Velden, B. H. M., Gilhuijs, K. G. A., Leiner, T., Viergever, M. A., & Išgum, I. (2016b). Deep learning for multi-task medical image segmentation in multiple modalities. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II* (pp. 478–486). Cham: Springer International Publishing. URL: http://dx.doi.org/10.1007/978-3-319-46723-8_55. doi:10.1007/978-3-319-46723-8_55.

Mukhopadhyay, A. (2016). Total variation random forest: Fully automatic mri segmentation in congenital heart disease. In *HVSMR 2016: MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*.

Pace, D. F., Dalca, A. V., Geva, T., Powell, A. J., Moghari, M. H., & Golland, P. (2015). Interactive whole-heart segmentation in congenital heart disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 80–88). Springer.

Peters, J., Ecabert, O., Meyer, C., Kneser, R., & Weese, J. (2010). Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. *Medical image analysis*, *14*, 70–84.

Peters, J., Ecabert, O., Meyer, C., Schramm, H., Kneser, R., Groth, A., & Weese, J. (2007). Automatic whole heart segmentation in static magnetic resonance image volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 402–410). Springer.

Petitjean, C., & Dacher, J.-N. (2011). A review of segmentation methods in short axis cardiac mr images. *Medical image analysis*, *15*, 169–184.

Radtke, A., Nadalin, S., Sotiropoulos, G., Molmenti, E., Schroeder, T., Valentin-Gamazo, C., Lang, H., Bockhorn, M., Peitgen, H., Broelsch, C. et al. (2007). Computer-assisted operative planning in adult living donor liver transplantation: a new way to resolve the dilemma of the middle hepatic vein. *World journal of surgery*, *31*, 175–185.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer.

Roth, H. R., Lu, L., Farag, A., Sohn, A., & Summers, R. M. (2016). Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II* (pp. 451–459). Cham: Springer International Publishing.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., & Summers, R. M. (2014). A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 520–527). Springer.

Rueckert, D., Lorenzo-Valdes, M., Chandrashekara, R., Sanchez-Ortiz, G., & Mohiaddin, R. (2002). Non-rigid registration of cardiac mr: Application to motion modelling and atlas-based segmentation. In *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on* (pp. 481–484). IEEE.

Rusko, L., Bekes, G., Nemeth, G., & Fidrich, M. (2007). Fully automatic liver segmentation for contrast-enhanced ct images. *MICCAI Wshp. 3D Segmentation in the Clinic: A Grand Challenge*, *2*.

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sánchez, C. I., & van Ginneken, B. (2016). Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, *35*, 1160–1169.

Shahzad, R., Gao, S., Tao, Q., Dzyubachyk, O., & van der Geest, R. (2016). Automated cardiovascular segmentation in patients with congenital heart disease from 3d cmr scans: Combining multi-atlases and level-sets. In *HVSMR 2016: MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*.

Shang, Y., Deklerck, R., Nyssen, E., Markova, A., De Mey, J., Yang, X., & Sun, K. (2011). Vascular active contour for vessel tree segmentation. *Biomedical Engineering, IEEE Transactions on*, *58*, 1023–1032.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, *35*, 1285–1298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, .

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, *35*, 1196–1206.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway networks, . *abs/1505.00387*.

Suzuki, K., Kohlbrenner, R., Epstein, M. L., Obajuluwa, A. M., Xu, J., & Hori, M. (2010). Computer-aided measurement of liver volumes in ct by means of geodesic active contour segmentation coupled with level-set algorithms. *Medical physics*, *37*, 2159–2166.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Tobon-Gomez, C., Geers, A. J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z. et al. (2015). Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE transactions on medical imaging*, *34*, 1460–1473.

Tran, P. V. (2016). A fully convolutional neural network for cardiac segmentation in short-axis mri. *arXiv preprint arXiv:1604.00494*, .

Tziritas, G. (2016). Fast fully-automatic segmentation of cardiac images using 3-d mrf model optimization and substructures tracking. In *HVSMR 2016: MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*.

Van Assen, H. C., Danilouchkine, M. G., Frangi, A. F., Ordás, S., Westenberg, J. J., Reiber, J. H., & Lelieveldt, B. P. (2006). Spasm: a 3d-asm for segmentation of sparse and arbitrarily oriented cardiac mri data. *Medical Image Analysis*, *10*, 286–303.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Wang, V. Y., Lam, H., Ennis, D. B., Cowan, B. R., Young, A. A., & Nash, M. P. (2009). Modelling passive diastolic mechanics with quantitative mri of cardiac structure and function. *Medical image analysis*, *13*, 773–784.

Wimmer, A., Soza, G., & Hornegger, J. (2009). A generic probabilistic active shape model for organ segmentation. In *Yang, G., Hawkes, D., Rueckert, D., Nobel, A., Taylor, C. (eds.) MICCAI 2009, PartII. LNCS, vol. 5762, pp. 26-33.*.

Wolterink, J., Leiner, T., Viergever, M., & Isgum, I. (2016). Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In *HVSMR 2016: MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1395–1403).

Xing, F., Xie, Y., & Yang, L. (2016). An automatic learning-based framework for robust nucleus segmentation. *IEEE transactions on medical imaging*, *35*, 550–566.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *ICLR*.

Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P. A. (2016). Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging,,* . doi:10.1109/TMI.2016.2642839.

Zhao, Y., Zan, Y., Wang, X., & Li, G. (2010). Fuzzy c-means clustering-based multilayer perceptron neural network for liver ct images automatic segmentation. In *Control and Decision Conference (CCDC), 2010 Chinese* (pp. 3423–3427). IEEE.

Zhen, X., Zhang, H., Islam, A., Bhaduri, M., Chan, I., & Li, S. (2016). Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression. *Medical Image Analysis*, .

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., & Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1529–1537).

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

Zhuang, X., Rhode, K. S., Razavi, R. S., Hawkes, D. J., & Ourselin, S. (2010). A registration-based propagation framework for automatic whole heart segmentation of cardiac mri. *IEEE transactions on medical imaging*, *29*, 1612–1625.

Zikic, D., Ioannou, Y., Brown, M., & Criminisi, A. (2014). Segmentation of brain tumor tissues with convolutional neural networks. *Proceedings MICCAI-BRATS*, (pp. 36–39).