Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method

Xiangrong Zhou[a)] and Ryosuke Takayama
Department of Intelligent Image Information, Graduate School of Medicine, Gifu University, Gifu 501-1194, Japan

Song Wang
Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

Takeshi Hara and Hiroshi Fujita
Department of Intelligent Image Information, Graduate School of Medicine, Gifu University, Gifu 501-1194, Japan

(Received 31 October 2016; revised 3 July 2017; accepted for publication 10 July 2017; published 31 August 2017)

Purpose: We propose a single network trained by pixel-to-label deep learning to address the general issue of automatic multiple organ segmentation in three-dimensional (3D) computed tomography (CT) images. Our method can be described as a voxel-wise multiple-class classification scheme for automatically assigning labels to each pixel/voxel in a 2D/3D CT image.
Methods: We simplify the segmentation algorithms of anatomical structures (including multiple organs) in a CT image (generally in 3D) to a majority voting scheme over the semantic segmentation of multiple 2D slices drawn from different viewpoints with redundancy. The proposed method inherits the spirit of fully convolutional networks (FCNs) that consist of "convolution" and "deconvolution" layers for 2D semantic image segmentation, and expands the core structure with 3D-2D-3D transformations to adapt to 3D CT image segmentation. All parameters in the proposed network are trained pixel-to-label from a small number of CT cases with human annotations as the ground truth. The proposed network naturally fulfills the requirements of multiple organ segmentations in CT cases of different sizes that cover arbitrary scan regions without any adjustment.
Results: The proposed network was trained and validated using the simultaneous segmentation of 19 anatomical structures in the human torso, including 17 major organs and two special regions (lumen and content inside of stomach). Some of these structures have never been reported in previous research on CT segmentation. A database consisting of 240 (95% for training and 5% for testing) 3D CT scans, together with their manually annotated ground-truth segmentations, was used in our experiments. The results show that the 19 structures of interest were segmented with acceptable accuracy (88.1% and 87.9% voxels in the training and testing datasets, respectively, were labeled correctly) against the ground truth.
Conclusions: We propose a single network based on pixel-to-label deep learning to address the challenging issue of anatomical structure segmentation in 3D CT cases. The novelty of this work is the policy of deep learning of the different 2D sectional appearances of 3D anatomical structures for CT cases and the majority voting of the 3D segmentation results from multiple crossed 2D sections to achieve availability and reliability with better efficiency, generality, and flexibility than conventional segmentation methods, which must be guided by human expertise. © *2017 The Authors. Medical Physics published by Wiley Periodicals, Inc. on behalf of American Association of Physicists in Medicine.* [https://doi.org/10.1002/mp.12480]

Key words: 2D semantic segmentation, anatomical structure segmentation, CT images, deep learning, fully convolutional network (FCN)

1. INTRODUCTION

Three-dimensional (3D) computed tomography (CT) images provide useful internal information about the human anatomy that can be used to support diagnosis, surgery, and therapy. The recognition and segmentation of anatomical structures play critical roles in the quantitative interpretation of CT images, and, conventionally, the image processing is accomplished through human interpretation and manual annotations by expert radiologists. However, human interpretation is often qualitative and subjective, with relatively high intra and interobserver variability. Manual annotations also have limited reproducibility and are very time-consuming. Therefore, automatic image segmentation would offer improved efficiency and reliability, as well as reducing the burden on radiologists.[1,2]

Fully automatic image segmentation, which transfers the physical image signal to a useful abstraction, is a crucial prerequisite for computer-based image analysis of 3D CT cases.[3] Nevertheless, this task is challenging for several reasons. First, there is a large variation in the appearance of anatomical structures in clinical image data (abnormal in most cases), and this is difficult to represent using mathematical rules.

Second, low intensity contrast and high image noise usually lead to ambiguous and blurred boundaries between different organ or tissue regions in CT images. These boundaries are difficult to identify using low-level digital signal/image processing. Third, different CT cases may cover different parts of the human body using different spatial image resolutions according to specific clinical requirements. It is difficult and costly to prepare models that are applicable to all possible CT cases and apply these models to CT image segmentation. Therefore, accurate CT image segmentation has become the bottleneck in many applications of computer-based medical image analysis and image interpretation.

CT image segmentation[3] covers a wide range of research fields. In this paper, we focus on simultaneous multiple organ segmentation across a wide range of the human body in 3D CT images, which is a major research topic in computer-aided diagnosis. Conventional approaches for multiple organ segmentation typically use pixel-based methods. In these approaches, image segmentations are divided into a number of functional modules, and numerous hand-crafted signal processing algorithms and image features are examined according to human experience and knowledge. Although some mathematical models[4–11] have recently been introduced, conventional CT image segmentation methods still attempt to emulate limited human-specified rules or operations in segmenting CT images directly. These methods can achieve reasonable segmentation results on CT images for a special organ type within a known narrow scan range. However, they may fail in many clinical cases that include anatomical structures that are seriously deformed and generally cannot deal with scan ranges that are unknown *a priori*. To further improve the accuracy and robustness of CT image segmentation, the developed segmentation methods are expected to handle a larger variety of ambiguous image appearances, shapes, and relationships of anatomical structures. It is difficult to achieve this goal by defining and considering human knowledge and rules explicitly. Instead, data-driven approaches using large sets of image data — such as a deep convolutional neural network (CNN) — are more appropriate for solving this segmentation problem.

Recently, deep CNNs have been applied to medical image analysis in several studies. Most of them have used deep CNNs for lesion detection or classification,[12–15] while others have embedded CNNs into conventional organ-segmentation processes to reduce the false positive rate in the segmentation results or to predict the likelihood of the image patches.[16–18] Studies of this type usually divide CT images into numerous small 2D/3D patches at different locations, and then classify these patches into multiple predefined categories. Deep CNNs can also be used to learn a set of optimized image features (sometimes combined with a classifier) to achieve the optimal classification rate for these image patches. However, the anatomical segmentation of CT images over a wide region of the human body is still challenging because of similarities in the images of different structures, as well as the difficulty of ensuring global spatial consistency in the labeling of patches in different CT cases.

This paper proposes a deep learning-based approach to the general multiple organ-segmentation problem in 2D/3D CT images. The initial idea of this approach was presented in a conference with a preliminary result.[19] Our method tackles three critical issues in CT image segmentation. First, efficiency and generality: our approach uses pixel-to-label learning to train all of the variable arguments together for general multiple organ segmentations. This is much more convenient to use and extend than conventional methods, which require specific models and algorithms to be prepared for each type of organ. Second, performance and complexity: our method tackles 3D image segmentation as an iteration of single 2D CNN, which is implemented as GPU-based parallel processing. This greatly speeds up the segmentation process compared with CPU-based systems. Third, applicability and flexibility: the core component of image segmentation uses a fully convolutional network (FCN), which is naturally adaptive to segmenting different content from different-size images that may cover arbitrary CT scan ranges (e.g., body, chest, abdomen). No CT image segmentation technique with this capability has previously been published.

2. METHODS

2.A. Overview

The basis of our method for CT image segmentation can be summarized as "multiple 2D proposals followed by 3D integration."[19] This comes from the way in which a radiologist interprets CT cases — observing many 2D sections and then reconstructing/imagining the 3D anatomical structure. Multiple organ segmentations on 2D CT sections are much more difficult than direct segmentation on a 3D CT volume, because 2D sectional fragments of 3D anatomical structures from different perspectives may appear to have little in common and are difficult to integrate using conventional modeling approaches. The reasons we use a 2D sectional image segmentation as the basic processing unit, rather than directly segmenting the 3D CT volumes, are to (a) learn a model with features in a pixel-to-label way that can successfully represent anatomical structures as completely as possible under the current computer hardware resources (NVIDIA graphics processing units (GPUs) in a Linux environment), (b) gain better segmentation performance by employing majority voting over multiple 2D segmentation results (increasing redundancy) on the same location in 3D, and (c) satisfy the needs of clinical medicine regarding 3D or 2D images over an arbitrary CT scan range (e.g., body, chest, abdomen). To model the large variance of 2D sectional image appearances, we train a deep CNN to encode the anatomical structures from a relatively small number of 3D CT images, and accomplish CT image segmentation using pixel-wise labeling and decoding the input image based on the trained deep CNN.

The proposed segmentation method is illustrated in Fig. 1. The input is a 3D CT image (a 2D case can be regarded as a

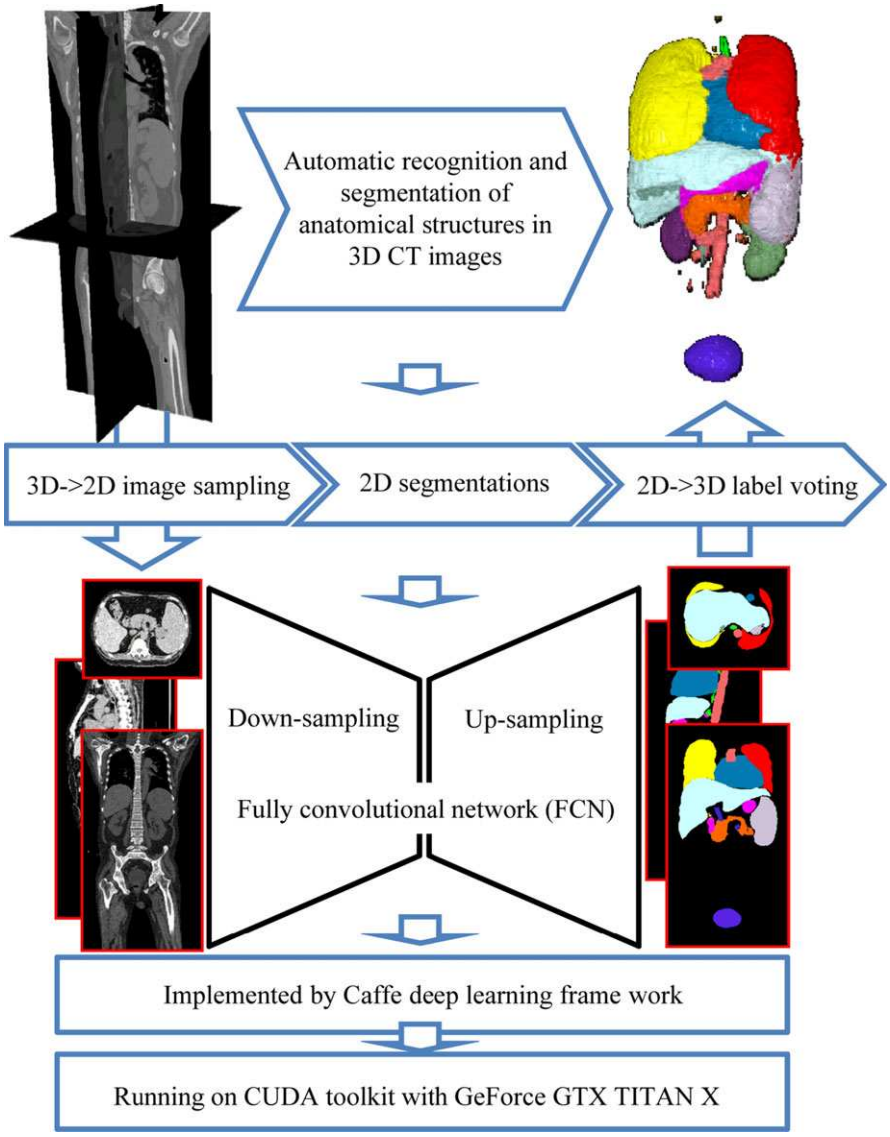


FIG. 1. Network of the proposed anatomical structure segmentation for 3D CT image. See Fig. 2 for the details of the FCN part.

degenerate 3D case with only one section), and the output is a label map of the same size and dimensions in which the labels are an annotated set of anatomical structures. Our segmentation process is repeated to sample 2D sections from a 3D CT image, pass them to the deep CNN for pixel-wise annotation, and stack the 2D labeled results back into 3D. Finally, the anatomical structure label at each voxel is determined based on majority voting from multiple 2D labeled results crossed at the voxel. The 2D segmentation uses an FCN[20] for the anatomical segmentation of 2D CT image sections. This FCN is trained on a set of 3D CT images, with human annotations as the ground truth. The processing steps in our segmentation are integrated into a single network under a simple architecture without the need for conventional image-processing algorithms such as smoothing, filtering, and level-set methods. The parameters of the network are learnable and optimized based on a pixel-to-label training scheme.

2.B. 2D CT image segmentation using FCN

FCNs have achieved state-of-the-art performance on natural image segmentation tasks and feature representation for dense classification.[20] This dense classification can also be used to predict the probability of each pixel belonging to a specific anatomical structure in a CT image. The architecture of FCNs includes two modules (down-sampling path and up-sampling path) that are integrated into a simple network trained in a pixel-to-label way. The motivation behind the FCN is that the down-sampling path extracts high-level abstraction information (target location and type), while the up-sampling path predicts the low-level (pixel-wise)

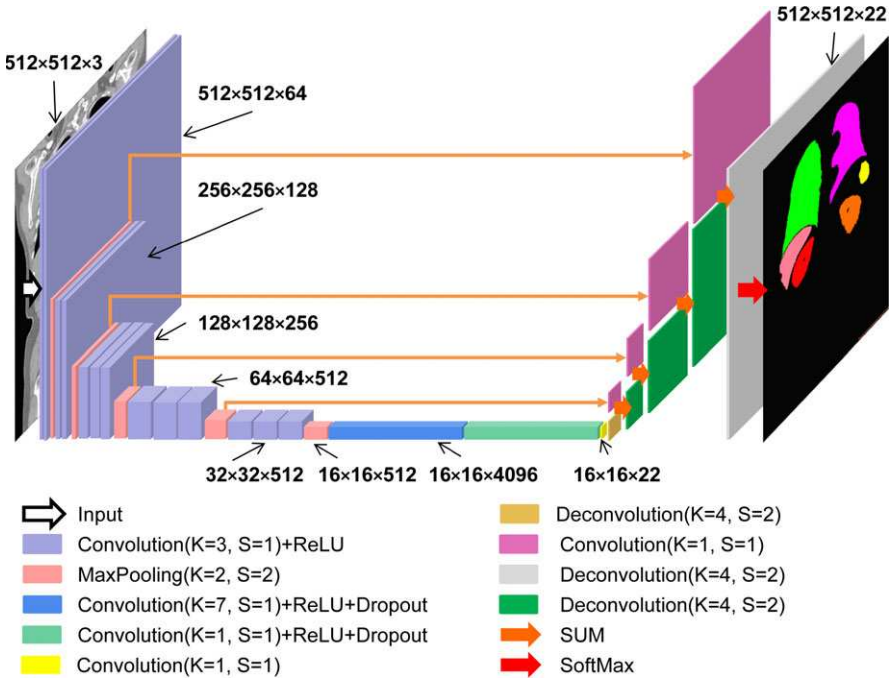


FIG. 2. Semantic image segmentation of 2D CT slice using a fully convolutional network (FCN).[20] (K kernel size, S: stride).

information (target shape and contour). The parameters in these two modules of the FCN are globally optimized in the training stage. The structure of the FCN used in the proposed method for 2D CT image segmentation is shown in Fig. 2.

The FCN predicts scores (probability of each label class) based on intensity–texture information within a given receptive field (e.g., an image patch with a predefined size). To avoid confusing similar patches that belong to different organs in CT images, our method uses a variable-size receptive field and regards the whole region of a 2D CT sectional image as one receptive field. This enables all information in a 2D CT section to be used directly to predict complex structures (multiple labels). Using the FCN architecture, our segmentation network provides the capability to adapt naturally to input CT images of any size and any scan range across the human body, producing an output with the corresponding spatial dimensions.

The down-sampling path of our FCN uses the VGG16 net structure[21] (16 3 × 3 convolution layers interleaved with five maximum pooling layers plus three fully connected layers), as shown in Fig. 2. We change the three fully connected layers of VGG16 to convolutional layers.[20] The final fully connected classifier is a 1 × 1 convolution layer whose channel dimension is fixed according to the number of labels (22 in our case). The up-sampling path contains five deconvolutional (backward-stride convolution) and convolutional layers. These have a skip structure that passes information lost in the lower convolution layers of VGG16 directly into the deconvolution process, enabling detailed contours to be recovered sequentially under a higher image resolution.[20] Rectified Linear Unit (ReLU) is used as the activation function in both the up- and down-sampling paths. A graph for easily visualizing the details of our FCN structure is presented in the Appendix.

2.C. 3D CT image segmentation by expanding FCN with 3D-2D-3D transformation

Each voxel in a 3D CT image can lie on different 2D sections that pass through the voxel with different orientations. Our idea of 3D CT image segmentation is to use the rich image information of the entire 2D section to predict the anatomical label of this voxel. The robustness and accuracy of this technique are increased by redundantly labeling this voxel on the multiple 2D sections with different orientations. We sample a 3D CT case over numerous sections (2D images) with different orientations, segment each 2D section using the FCN, and then assemble the output of the segmentation (i.e., labeled 2D maps) back into 3D. In this work, we select all the 2D sections in three orthogonal directions (axial, sagittal, and coronal body); this ensures that each voxel in the 3D image is located on three 2D CT sections. After the 2D image segmentation, each voxel is redundantly annotated three times, once for each 2D CT section. The annotated results for each voxel should ideally be identical, but may differ because of mislabeling by the FCN during the 2D image segmentation. The labels are fused using majority voting (selecting the mode of the three labels) to improve the stability and accuracy of the final decision (refer to Fig. 3). When there is no duplication among the three labels, our method selects the label of an organ type with the largest volume (the highest *prior* for the voxel appearances in CT images based on anatomical knowledge) as the output.

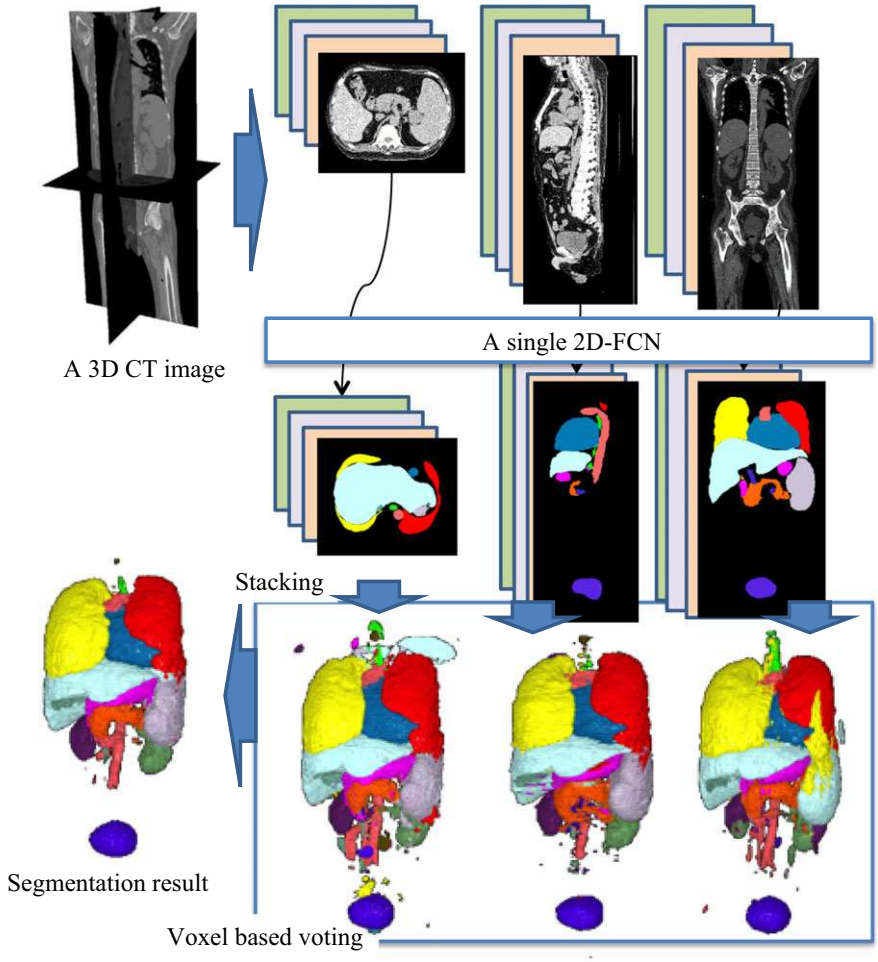


FIG. 3. 3D CT image segmentation using a single FCN with 3D-2D-3D transformation.

2.D. Network training based on pixel-to-label learning and transfer learning

The proposed network is trained using a set of 3D CT cases with human-annotated anatomical structures by pixel-to-label learning. We assume the labels of each voxel on any 2D CT section are identical, and do not invoke the majority voting step during the training process. We combine all the 2D CT sections (and their corresponding label maps) along the three body orientations to train the FCN. The training process repeats feed-forward computation and back-propagation to minimize the loss function, which is defined as the sum of the pixel-wise losses between the network prediction and the label map annotated by the human experts.

Because of the expensive and complicated acquisition process, there is a scarcity of 3D CT cases with accurate annotations. However, we can decompose 3D CT volumes into a number of 2D sectional images to greatly increase the number of training samples. The total number of 2D image samples (about 100,000 in our experiments) may still be insufficient for training a deep CNN (in computer vision tasks, CNN-based image segmentation and object detection are usually trained on the ImageNet dataset, which has about 10,000,000 labeled images). Transfer learning is a useful method of alleviating the problem of insufficient training data in the medical domain.[12] The learned parameters (features) in the lower layers of a deep CNN are general, whereas those in higher layers are more specific to different tasks. Thus, transferring the rich feature hierarchies with embedded knowledge in lower layers of a deep CNN learned from a huge number of natural images (such as the ImageNet dataset) should help to reduce the overfitting caused by the limited number of training CT scans and further boost performance.

In this study, we used an off-the-shelf model from VGG16,[21] which has been trained on the ImageNet ILSVRC-2014 dataset. Compared to our small-scale dataset (240 CT scans), ImageNet has a huge number of image samples with a large range of content. We initialized the layers in the down-sampling path with pretrained parameters from VGG16, and set the parameters of the remaining layers to small random numbers with zero mean. The whole network was then fine-tuned using our CT image dataset in a pixel-to-label way.

The fine-tuning was achieved sequentially by adding deconvolution layers.[20] Initially, a coarse prediction (using 32-pixel strides) was supplied to the modified VGG16

network with one deconvolution layer (called FCN32s). A finer training sample was then added after inserting one further deconvolution layer at the end of the network. This was done using skips that combine the final prediction layer and a lower layer with a finer stride in the modified VGG16 network. This fine-grained training was repeated with more network layers trained from the predictions of 16, 8, 4, and 2 strides on the CT images to build FCN16s, 8s, 4s, and 2s, respectively. The detailed network structure of FCN2s can be found in the Appendix.

2.E. Implementation details

Our network was developed based on the open-source library Caffe[22] and the reference implementation of FCN.[20] In the training stage, we used two optimization functions: stochastic gradient descent (SGD) with a momentum of 0.9 and ADAM[23] for comparison. A mini-batch size of 20 images, learning rate of 10^{-4}, and weight decay of 2^{-4} were used as the training parameters. In addition, we incorporated dropout layers (drop rate 0.5) and local contrast normalization (LCN) layers (local size 3, alpha = 5×10^{-5}, beta = 0.75) into the deconvolution layers to validate the performance. A workstation with the CUDA Library on a GPU (NVIDIA GeForce TITAN-X with 12 GB of memory) was used for network training and testing.

3. EXPERIMENTS AND RESULTS

3.A. Dataset

Our research was conducted with the approval of the Institutional Review Boards at Gifu and Tokushima Universities, Japan. We evaluated our method on a shared dataset produced by a research project entitled "Computational Anatomy."[24] This dataset consists of 640 3D volumetric CT scans that are used for the diagnosis of diverse types of lesions at Tokushima University Hospital. The anatomical ground truth (a maximum of 19 labels that show major organ types and interesting regions inside the human body) of 240 CT scans from 200 patients (167 patients with one CT scan, 24 patients with two CT scans, seven patients with three CT scans, and one patient with four CT scans) was also included in the dataset. The anatomical structures in the CT scans were first annotated manually by many members in different research groups; then, the annotated results were validated and refined intensively by medical experts on a committee of this research project to maintain a high quality. A semiautomatic method was developed to support the annotation task.[25] Our experimental study used all 240 CT scans with the ground truth, comprising 89 torso, 17 chest, and 134 abdomen or pelvis scans. The scan ranges of the CT images are shown in Fig. 4. The total number of target organs is listed in Table I. These CT image scanned by a multislice CT scanner (Aquilion from Toshiba Medical Systems Corporation, Otawara-shi, Japan) and reconstructed by different protocols and kernel functions, leading to different image resolutions (distributed from 0.625 to 1.148 mm with 1.0 mm slice thickness) and different image quality (specialized for lung or abdominal observations). Contrast media enhancements were applied in 155 CT scans. Although the size, density, texture, and spatial resolution of these CT images are not identical, we used them for training and testing directly, without any normalization. The down-sampling path of our network is based on the structure of VGG16, which requires an 8-bit RGB color image as input. Thus, we transformed the CT images (12-bit one-channel gray-level format) at the entrance of the network into 8-bit RGB color images. Actually, this transformation used a linear interpolation that converted each CT number in a CT image into an 8-bit (0–255) gray level and duplicated this gray level to each RGB channel at the same location in the CT image.

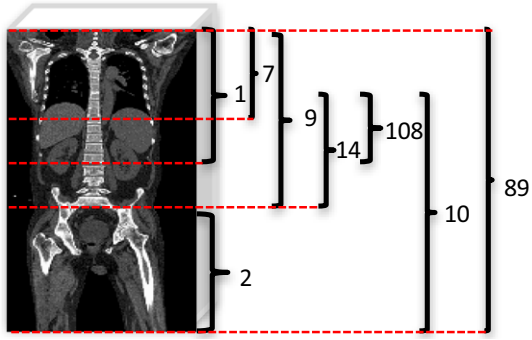


FIG. 4. Number of CT scans in our dataset that cover different ranges of human body, shown by the brackets in the coronal body direction. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE I. Number of target regions involved in 240 CT scans.

Target names	CT scans
Right lung	106
Left lung	106
Heart	106
Aorta	97
Esophagus	95
Liver	231
Gallbladder	228
Stomach and duodenum (2nd pos.)	230
Stomach and duodenum lumen	153
Contents inside of stomach and duodenum	137
Spleen	228
Right kidney	225
Left kidney	229
Inferior vena cava	223
Portal vein, splenic vein, and superior mesenteric vein	230
Pancreas	230
Bladder	108
Prostate	72
Uterus	32

3.B. Experimental setting

We randomly picked 5% of the samples from the torso, chest, and abdomen CT scans as the testing dataset (total of 12 CT scans) and used the remaining 228 CT cases to train the network. Multiple CT scans from the same patients were only used for training. We repeated this procedure three times to train three networks, applied them to three testing datasets, and obtained segmentation results for a total of 36 (3×12) CT scans without overlap.

In each training stage, a single network was trained using the ground-truth labels of the 19 target regions (Heart; right/left Lung; Aorta; Esophagus; Liver; Gallbladder; Stomach and Duodenum (lumen and contents); Spleen; right/left Kidney; Inferior Vena Cava; region of Portal Vein, Splenic Vein, and Superior Mesenteric Vein; Pancreas; Uterus; Prostate; and Bladder). The trained network was then applied to the 12 test samples. Three examples of the segmentation for a 3D CT case covering the human torso, chest, and abdomen are shown in Fig. 5.

3.C. Quantitative evaluations

The accuracy of the segmentation was evaluated for each organ type and each image. First, we measured the intersection over union (IU)[20] (also known as the Jaccard similarity coefficient) between the segmentation result and the ground truth. The mean and standard deviation of IU values per organ type are presented in Table II for all of the 684 training and 36 testing CT scans used in the experiments.

Generally speaking, a CT scan may contain different anatomical structures, and this information is unknown before the segmentation. We performed a comprehensive evaluation of multiple target segmentation results for all images in the test dataset by considering the variance of the organ number and volume. The measures (mean voxel accuracy, mean IU, and frequency-weighted IU) that are commonly used in semantic segmentation and scene parsing were employed in this study for the evaluations.[20] Let n_{ij} be the number of pixels in target i classified as target j, n_{cl} be the

TABLE II. Accuracy evaluations in terms of mean value and standard deviation (SD) of IUs for 19 target types between segmentation and ground truth in 684 (228×3) training and 36 (12×3) test CT scans.

	IU			
	Testing samples		Training samples	
Target name	Mean	SD	Mean	SD
Right lung	0.916	0.0286	0.916	0.096
Left lung	0.890	0.0449	0.898	0.089
Heart	0.817	0.0370	0.839	0.035
Aorta	0.585	0.1107	0.628	0.109
Esophagus	0.107	0.0624	0.065	0.053
Liver	0.882	0.0507	0.880	0.036
Gallbladder	0.424	0.2136	0.478	0.187
Stomach and duodenum (2nd_pos.)	0.454	0.1471	0.414	0.163
Stomach and duodenum lumen	0.389	0.2739	0.405	0.248
Contents inside of stomach and duodenum	0.120	0.1752	0.129	0.151
Spleen	0.767	0.0847	0.766	0.098
Right kidney	0.792	0.0659	0.768	0.106
Left kidney	0.792	0.0611	0.755	0.103
Inferior vena cava	0.435	0.1749	0.412	0.157
Portal vein, splenic vein, and superior mesenteric vein	0.126	0.1110	0.130	0.115
Pancreas	0.390	0.1217	0.360	0.119
Bladder	0.589	0.1325	0.543	0.195
Prostate	0.276	0.2028	0.305	0.170
Uterus	0.326	0.1587	0.207	0.147

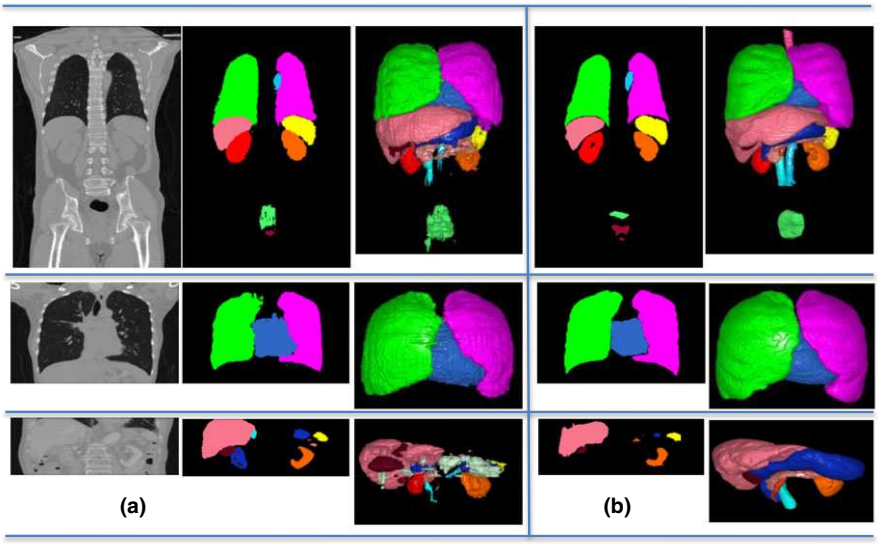


FIG. 5. (a): Three examples of segmentation in 3D CT case covering torso (upper), chest (middle), and abdomen (lower) regions along with segmented regions labeled with different colors for one 2D coronal CT slice (middle column) and 3D visualization based on surface-rendering method (right column). (b): Corresponding ground-truth segmentations for three cases.

total number of different targets in a CT case, and $t_i = \sum_j n_{ij}$ be the total number of pixels in target i. These measures are defined as:

- Mean voxel accuracy: $\left(\sum_i n_{ii}/t_i\right)/n_{cl}$ (1)
- Mean IU: $\left(\sum_i n_{ii}/\left(t_i + \sum_j n_{ji} - n_{ii}\right)\right)/n_{cl}$ (2)
- Frequency-weighted IU: $\left(\sum_k t_k\right)^{-1} \sum_i t_i n_{ii}/\left(t_i + \sum_j n_{ji} - n_{ii}\right)$ (3)

The evaluation results for the mean voxel accuracy and frequency-weighted IU were 87.9% and 80.5%, respectively, when averaged over all the segmentation results of the test dataset. These results mean that 87.9% of the voxels within the anatomical structures (constructed using multiple target regions) were labeled correctly, with an overlap ratio of 80.5% for the test dataset. We conducted the same evaluation on the training dataset, and found corresponding values of 88.1% and 79.9%.

3.D. Segmentation performance

We validated the performance of the proposed network by evaluating the segmentation results of FCN 8s. The network was trained over 160,000 iterations using the ADAM optimizer with the training protocol described above. The trained network was then applied to the test CT cases. Except for one gallbladder and one prostate, our network recognized and extracted all organs correctly. Because our segmentation targets cover a wide range of shapes, volumes, and sizes, either with or without contrast enhancement, and come from different locations in the human body, these experimental results offer an excellent demonstration of the capability of our approach to recognize anatomical structures in the types of CT images actually used in clinical medicine. Table II presents the mean per-target IUs between the segmentation results and ground truth in both the testing and training data. We found that the mean IUs of organs with larger volumes (e.g., lung) were comparable to those achieved by previous methods.[6–11] For some smaller organs (e.g., gallbladder) and stomach contents (which have not previously been reported), our segmentation did not produce particularly high IUs. The limited image resolution is likely to be the major cause of this poor performance for these organs. Our evaluation shows that the average segmentation accuracy of all targets over both the test and training CT images is approximately 80.5% and 79.9% in terms of the frequency-weighted IUs. Because the performance of deep learning is highly dependent to the amount of training data, we reduced the number of CT scans in the training stage from 95% to 75% and increased the number CT scans for testing from 5% to 25% by using the same dataset and experimental setting and used a fourfold cross-validation to evaluate the performance again. These additional experimental results demonstrated that the average segmentation accuracy of all targets over the 240 (4 × 60) test CT scans was approximately 78.3% in terms of the frequency-weighted IUs and 86.1% in terms of the mean voxel accuracy, which are comparable to the performance (80.5% and 87.9%) in the previous experiment. The IUs of most organ types showed similar values, except that the spleen, prostate, and uterus showed a large decrease in the accuracy of more than 10% in terms of the IU. These decreases in the performance were caused by the shortage of the training samples and may be improved by increasing the number of CT scans in the training stage. Thus, our approach can recognize and extract all types of major anatomical structures simultaneously, achieving a reasonable accuracy according to the organ volume in the CT images.

4. DISCUSSION

4.A. Transfer learning and training protocols

For comparison, we trained our network to "learn from scratch" by initializing all parameters of an FCN to small random numbers with zero mean. No convergence was observed within 80,000 learning iterations, and the network then failed to segment any of the targets in the test dataset. These results indicate that the size of our CT dataset with current training protocols is insufficient to train the proposed network successfully when starting from scratch. However, when we fine-tuned our network using VGG16,[21] which is pretrained using ImageNet, convergence was achieved after 22,000 iterations. The trained network fine-tuned from VGG16 in 80,000 learning iterations could segment multiple organs successfully in CT images from both the testing and training datasets. This demonstrates that comprehensive knowledge learned from large-scale, well-annotated datasets can be transferred to our network to accomplish the CT image segmentation task.

We also compared the performance of networks optimized by SGD and ADAM with the same training protocols described in Section 2.E. The segmentation results on the test data indicate that the network trained by ADAM offers slightly better performance (up by 0.3% in voxel accuracy, 0.15% in frequency-weighted IU) than that trained by SGD. Because the learning rate does not need to be tuned in ADAM and the default parameters are likely to achieve good results, we used this function as the default optimizer for our network in subsequent experiments.

The performance of the network may be affected by the number of training iterations. We compared the segmentation results on the test dataset given by networks after 80,000, 160,000, and 550,000 training iterations. We found that 160,000 iterations was sufficient to train the network. Further training iterations may improve the segmentation accuracy of some organ types, but could not improve the overall performance across all organ types.

4.B. Network structure

For comparison, we incorporated dropout layers with each deconvolution layer in the up-sampling path and retrained the

network. We found that the network performance with the test dataset decreased (down by 23% in voxel accuracy and 28% in frequency-weighted IU) after inserting these dropout layers. We also tried to incorporate LCN layers in the same way, but did not observe any significant improvement in performance. Based on these results, we do not include dropout and LCN layers in the up-sampling path of the proposed network.

The up-sampling path consists of five deconvolutional layers (FCN32s to FCN2s). We investigated the segmentation results in the test samples after applying each FCN layer to the network sequentially. We found that the frequency-weighted IUs increased monotonically (69.8%, 81.1%, 84.9% and 88.0% at FCN32s, 16s, 8s, and 4s, respectively), and no further improvement was observed by FCN2s. This result demonstrates that diminishing returns of gradient descent occurred from FCN4s in the training stage. A similar phenomenon was noted in the original FCN structures[20] and confirmed in our preliminary reports.[19,26]

Some alterations to the deep CNN architecture have recently been reported in the field of computer vision and medical image analysis.[27–29] The well-known SegNet[27] network achieved better segmentation performance than FCN,[20] especially for small targets in natural images. We replaced the FCN part of our network with SegNet and examined its performance in CT image segmentation. This experiment showed that the original SegNet implementation could only deal with predefined input image sizes. Even when all CT cases were re-scaled accordingly, the preliminary experimental results did not suggest better performance than the FCN. It is possible that a customized version of SegNet is needed for CT image segmentation. However, further investigation of the differences in performance offered by FCN and SegNet is beyond the scope of this paper. Our current results show that FCN is the best deep learning network for our CT image segmentation task.

4.C. Insights offered by the trained FCN

To infer the functions of different layers learned from the CT images, we investigated the output of several "key" middle layers (two score maps of the pooling layers, and the final layer of the down-sampling path) in the trained FCN. An example of the intermediate results when passing a chest CT image through the trained network is presented in Fig. 6. The low-level image features (edges) are filtered and transferred along the network to become concentrated in the high-level abstract region (organ-wise heat map). We hypothesize that the down-sampling path of the network learned a set of features from the training dataset to successfully represent each target appearance in the CT images. These representation results act as the "probabilistic atlas" proposed in conventional approaches to show the approximate range of each target region corresponding to an input image. For the up-sampling path, we compared the output of each deconvolution layer (two examples before and after the first deconvolution layer are shown in Fig. 6). The detailed contour of each target region was restored sequentially, and the image resolution of the outputs increased. We believe that the up-sampling layers learn the shape information of the target regions from the training dataset, and this information guides the deconvolution layers to selectively detect the edges in the images for each specific target. Compared to conventional segmentation methods, all of these functions are enveloped into one network and optimized globally in our approach.

4.D. Comparison between 2D and 3D segmentations

We compared the segmentation results before and after the 3D majority voting process. The average value of mean IUs of 19 targets in testing samples without 3D majority voting was 49.1%, 50.3%, and 48.4% by stacking 2D segmentation results only in axial, coronal, and sagittal direction. After the 3D majority voting, the average value of mean IUs was improved to 53% (refer to Table II). The mean IUs of each target given by the segmentation results in three body directions were close to the mean IUs given by the final 3D voting results. We found that the best IU value of the 2D segmentation results is also comparable to the final results in 3D (refer to Fig. 3). Although some mislabeling was observed in individual 2D sections, especially in the sagittal body direction (caused by symmetry of anatomy), our network still displayed the potential to support real-time 2D image interpretation by a radiologist tracing and interpreting what is shown on the screen.

A major concern for our approach is the advantage of using a 2D FCN instead of a 3D FCN, which seems to be more straightforward for 3D image segmentation. We expanded our FCN from 2D to 3D, and confirmed that the 3D FCN could not operate under our computer environment with the current CT dataset because of insufficient memory in the GPU (NVIDIA GeForce TITAN-X with 12 GB of memory).

Recently, a number of studies[30–34] have applied an FCN to 3D medical image processes, and 3D U-Net (a 3D extension of an FCN with a good reputation) showed effective 3D image segmentation.[30] Therefore, we tested 3D U-Net with our training/testing CT scans. The experimental results showed that the original 3D U-Net did not work directly for multiple organ segmentation of CT images because a typical input image matrix of 3D U-Net is 248 $\times$ 244 $\times$ 64 and cannot be extended any further owing to the memory limitations of the GPU (NVIDIA GeForce TITAN-X with 12 GB of memory). This ROI-sized input did not have a sufficient image resolution to represent the anatomical structures sufficiently for entire CT images (maximum size of 512 $\times$ 512 $\times$ 1141 for the image matrix in our dataset) by a down-sampling process.

This problem may be addressed by dividing the CT images into a number of 3D patches before training and testing. Under this consideration, cutting-edge technology[34] based on 3D U-Net using a two-stage, coarse-to-fine approach (two 3D U-Nets with a cascade structure, which we call Cascade U-Nets) to accomplish a multiple organ segmentation was proposed and showed state-of-the-art performance. We compared our method with this novel approach[34] based

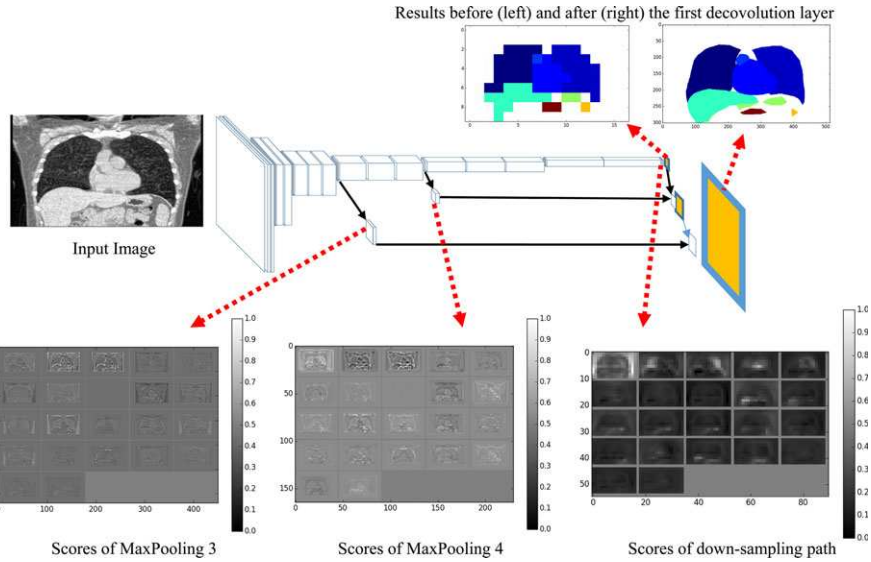


FIG. 6. Insight of learned FCN. Lower: outputs of MaxPooling layers and final result of down-sampling path; Upper: outputs before and after the first deconvolution layer. [Color figure can be viewed at wileyonlinelibrary.com]

on the same training (228) and testing (12) CT scans. The experiment showed the accuracies (IUs) in each target type on average of 12 testing CT scans obtained using our method were better than a single 3D U-Net with ROI-sized inputs, and still better than the Cascade U-Nets for nine target types, except for the other nine types of target (having a small volume or tube structures). The performance in terms of the frequency-weighted IUs of our method (80.5%) was comparable to the Cascade U-Nets (80.3%) for the test dataset.

Considering the difference between the two structures (two 3D U-Nets versus a single 2D FCN + voting), we must conclude that our 2D FCN is currently a realistic method for CT image segmentation.

4.E. Comparison to previous work

The previous studies most closely related to our work are those of Udupa et al.,[10,11] Wolz et al.,[8] Shimizu et al.,[7] Okada et al.,[9] and Lay et al.[6] Common to all these works is that they focused on multiple organ segmentation in CT images, as in our study. It is difficult to give a direct quantitative grading for all of these previous techniques, because the datasets, acquisition protocols, image resolutions, targets of interest, training and testing subdivisions, validation methods, and computer platforms were different. We directly compared the proposed method to a method (Okada et al.[9]) that was state-of-the-art among the conventional methods using the same experimental setting (228 CT scans for training of the models and 12 CT scans for testing). The training and testing processes used a computer with a CPU (Intel Core i7-2700K, 3.50 GHz, 16.0 GB memory). We confirmed that the method[9] only worked successfully for abdomen CT scans as its original design. As a result, the models for seven organ types were successfully constructed based on 93 contrast-enhanced abdomen CT scans from 93 patients within 228 training CT scans, and reasonable segmentation results for these organ types were obtained for seven contrast-enhanced abdomen CT scans within 12 testing CT scans. The time of organ segmentation for one abdomen CT scan (512 × 512 × 221 image matrix) was 40 min using multithreaded parallel computing. A comparison of the results with those of our proposed method for the seven testing CT scans is presented in Table III. The experimental results indicated that our method showed a better accuracy (mean value of IUs) for six organ types and worse accuracy for the gallbladder. The standard deviation of the IUs for all organ types of our proposed method was better than that of the conventional method.[9] This experimental results demonstrated the advantage of our proposed FCN-based method with regards to the accuracy (higher mean value of IUs), robustness (stability with a lower standard deviation of IUs), generality (one model for all organ types), and usability (ability to adapt to CT scans of different portions).

The advantage of the proposed network is that it combines the model generation and image segmentation steps within the same network using pixel-to-label learning. This provides more opportunities to use sufficient features suitable for CT image segmentation. Experimental results indicate that most of the target organs considered by previous studies can be segmented successfully using our network without any additional modification. Some difficult organs such as the stomach were also recognized and extracted. Thus, the proposed single network has sufficient generality to enable the segmentation of anatomical structures in CT images. In addition, the simple structure and GPU-based implementation of our segmentation process is computationally efficient. The computing time for training an FCN is approximately 3 days. The multiple organ segmentation of one 3D CT scan with a 512 × 512 × 221 matrix takes 3 min 43 s. The efficiency in terms of system development and improvement is much better than that of previous studies that attempted to incorporate

TABLE III. Accuracy comparison between the proposed method (FCN) and a conventional method[9] in terms of the IUs between the segmentation and the ground truth based on 228 training and 12 test CT scans. We only show the IUs of seven organ types along with mean and standard deviation (SD) for seven testing abdomen CT cases for which the conventional method[9] worked successfully.

	Spleen		Liver		Gallbladder		Right Kidney		Left Kidney		Inferior Vena Cava		Pancreas	
Case	FCN	Ref. [9]	FCN	Ref. [9]	FCN	Ref. [9]	FCN	Ref. [9]	FCN	Ref. [9]	FCN	Ref. [9]	FCN	Ref. [9]
1	0.880	0.922	0.937	0.957	0.632	0.896	0.885	0.888	0.857	0.910	0.705	0.756	0.553	0.694
2	0.745	0.760	0.747	0.638	0.000	0.000	0.837	0.277	0.776	0.705	0.565	0.819	0.169	0.115
3	0.810	0.044	0.926	0.921	0.764	0.819	0.919	0.931	0.855	0.120	0.610	0.517	0.434	0.279
4	0.782	0.327	0.933	0.913	0.004	0.078	0.851	0.884	0.804	0.876	0.205	0.153	0.074	0.836
5	0.892	0.316	0.903	0.892	0.668	0.922	0.887	0.898	0.858	0.707	0.595	0.563	0.509	0.133
6	0.861	0.941	0.934	0.923	0.011	0.000	0.833	0.873	0.849	0.912	0.681	0.676	0.442	0.617
7	0.879	0.001	0.869	0.682	N/A	N/A	0.900	0.827	0.905	0.010	0.630	0.501	0.536	0.005
Mean	0.836	0.473	0.893	0.846	0.346	0.453	0.873	0.797	0.843	0.606	0.570	0.569	0.388	0.383
SD	0.057	0.399	0.069	0.129	0.377	0.469	0.033	0.232	0.042	0.381	0.168	0.220	0.189	0.328

human specialist experience into complex algorithms for segmenting different organs. Although labeling the anatomical structures in the training samples still takes time, this burden can be reduced using bespoke and advanced semiautomatic algorithms.[25]

The initial study of this work was presented in a conference paper[19] that validated our idea of "2D FCN with 3D voting" for CT image segmentation with a preliminary experiment.[26] In this work, we improved the segmentation performance of this idea by refining the network structure and training method and reported more experimental results, including the exploration of the relation between the network parameters and the resulting performance and the comparisons with related works based on the same dataset. The performance of the network proposed in this work and its advantages over that presented in previous work were demonstrated for the first time in this paper with detailed descriptions of the methodology.

The drawback of our network is the poor accuracy (IUs) when segmenting smaller structures. This will be improved in future by using a larger training dataset and new network structures.[34] We will also expand the proposed network to other imaging modalities such as FDG-PET and MR images. We also plan to use more 2D slices for 2D FCN training and learn a weighted 3D voxel voting by sampling slices from more arbitrary directions in each 3D CT case.

5. CONCLUSIONS

For the automatic segmentation of anatomical structures (multiple organs) in CT images with different scan ranges, we proposed a single network trained by pixel-to-label learning. This network was applied to segment 19 types of targets in 240 3D CT scans, demonstrating highly promising results. Our work is the first to tackle anatomical segmentation (with a maximum of 19 targets) on scale-free CT scans (both 2D and 3D images) through a single deep neural network.

Compared with previous work, the novelty and advantages of our study are as follows. (a) Our approach uses voxel-to-voxel labeling with pixel-to-label global optimization of parameters, which has the advantage of better performance and flexibility in accommodating the large variety of anatomical structures in different CT cases. (b) Our method can automatically learn a set of image features to represent all organ types collectively using a 2D FCN with majority voting (a simple structure for both model training and implementation) for image segmentation. This approach leads to more robust image segmentation that is easier to implement and extend based on current hardware resources. Image segmentation using our approach has more advantages than previous methods in terms of usability (it can be used to segment any type of organ), adaptability (it can handle 2D or 3D CT images over any scan range), and efficiency (it is much easier to implement and extend). The proposed approach could also be extended as a general solution for more complex anatomical structure segmentation in other image modalities, which present fundamental problems in medical physics (e.g., MR and FDG-PET imaging).

ACKNOWLEDGMENTS

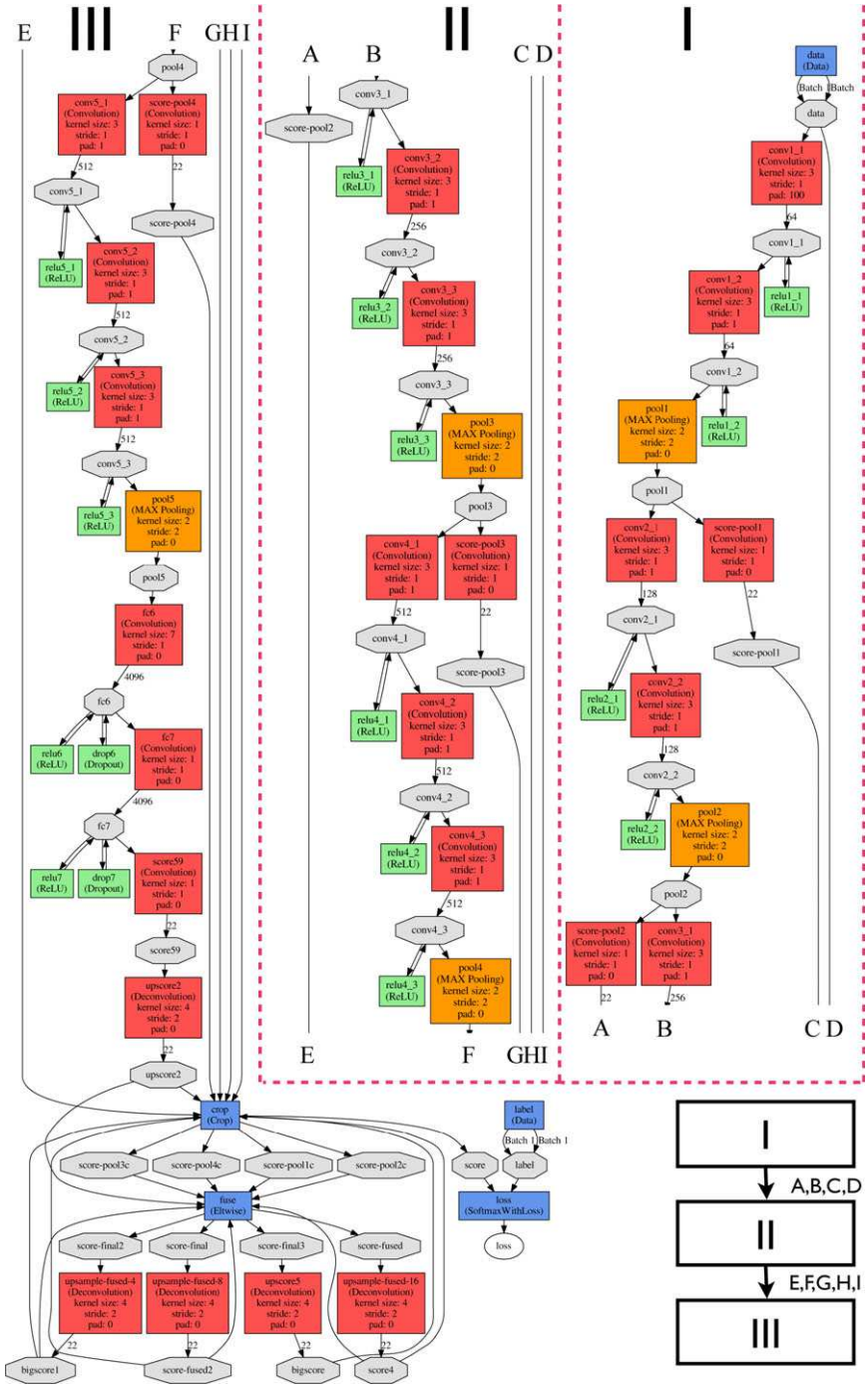
The authors would like to thank all the members of the Fujita Laboratory in the Graduate School of Medicine, Gifu University, for their collaboration. We especially thank Dr. Okada of Tsukuba University for providing binary code and Dr. Roth of Nagoya University for testing systems based on our dataset. We would like to thank all the members of the Computational Anatomy[24] research project, especially Dr. Ueno of Tokushima University, for providing the CT image database. This research was supported in part by a Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 26108005), and in part by a Grant-in-Aid for Scientific Research (C26330134), MEXT, Japan.

CONFLICT OF INTEREST

The authors have no COI to report.

APPENDIX

A graph that shows the network structure of FCN2s for the image segmentation of CT images.



[a]Author to whom correspondence should be addressed. Electronic mail: zxr@fjt.info.gifu-u.ac.jp.

REFERENCES

1. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*. 2007;31:198–211.
2. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection diagnosis, and treatment of cancer. *Annu Rev Biomed Eng*. 2013;15:327–357.
3. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. *Biomed Eng*. 2000;2:315–333.
4. Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: a review. *Med Image Anal*. 2009;13:543–563.
5. Xu Y, Xu C, Kuang X, et al. 3D-SIFT-Flow for atlas-based CT liver image segmentation. *Med Phys*. 2016;43:2229–2241.
6. Lay N, Birkbeck N, Zhang J, Zhou SK. Rapid multi-organ segmentation using context integration and discriminative models. *Proc IPMI*. 2013;7917:450–462.
7. Shimizu A, Ohno R, Ikegami T, Kobatake H, Nawano S, Smutek D. Segmentation of multiple organs in non-contrast 3D abdominal CT images. *Int J Comput Assist Radiol Surg*. 2007;2:135–142.
8. Wolz R, Chu C, Misawa K, Fujiwara M, Mori K, Rueckert D. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Trans Med Imaging*. 2013;32:1723–1730.
9. Okada T, Linguraru MG, Hori M, Summers RM, Tomiyama N, Sato Y. Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors. *Med Image Anal*. 2015;26:1–18.
10. Bagci U, Udupa JK, Mendhiratta N, et al. Joint segmentation of anatomical and functional images: applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images. *Med Image Anal*. 2013;17:929–945.
11. Sun K, Udupa JK, Odhner D, Tong Y, Zhao L, Torigian DA. Automatic thoracic anatomy segmentation on CT images using hierarchical fuzzy models and registration. *Med Phys*. 2016;43:1882–1896.
12. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
13. Ciompi F, de Hoop B, van Riel SJ, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Med Image Anal*. 2015;26:195–202.
14. Teramoto A, Fujita H, Yamamuro O, Tamaki T. Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. *Med Phys*. 2016;43:2821–2827.
15. Näppi JJ, Hironaka T, Regge D, Yoshida H. Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography. *Proc SPIE*. 2016;9785:97852B-1–97852B-8.
16. Brebisson D, Montana G, "Deep neural networks for anatomical brain segmentation," Proc. CVPR Workshops; 2015, 20–28.
17. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. *Proc SPIE*. 2015;9413:94131G-1–94131G-8.
18. Cha KH, Hadjiiski L, Samala RK, Chan HP, Caoili EM, Cohan RH. Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. *Med Phys*. 2016;43:1882–1896.
19. Zhou X, Ito T, Takayama R, Wang S, Hara T, Fujita H. Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting. Proc. 2nd Workshop on Deep Learning in Medical Image Analysis, MICCAI 2016, LNCS 10008, 111-120; 2016.
20. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proc CVPR*. 2015;3431–3440.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Proc ICLR*. 2015; arXiv:1409.1556.
22. Deep learning framework. http://caffe.berkeleyvision.org.
23. Kingma DP, Ba JL. ADAM: a method for stochastic optimization. *Proc ICLR*. 2015; arXiv:1412.6980.
24. Computational Anatomy for Computer-aided Diagnosis and Therapy. http://www.comp-anatomy.org/wiki/.
25. Watanabe H, Shimizu A, Ueno J, Umetsu S, Nawano S, Kobatake H. Semi-automated organ segmentation using 3-dimensional medical imagery through sparse representation. *Trans Jpn Soc Med Biol Eng*. 2013;51:300–312.
26. Zhou X, Ito T, Takayama R, Wang S, Hara T, Fujita H. First trial and evaluation of anatomical structure segmentations in 3D CT images based only on deep learning. *Med Image Inform Sci*. 2016;33:69–74.
27. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation; 2015, arXiv. http://arxiv.org/abs/1511.00561.
28. Lerouge J, Herault R, Chatelain C, Jardin F, Modzelewski R. IODA: an input/output deep architecture for image labeling. *Pattern Recog*. 2015;48:2847–2858.
29. Hoo-Chang S, Orton MR, Collins DJ, Doran SJ, Leach MO. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Trans Pattern Anal Mach Intell*. 2013;35:1930–1943.
30. Çiçek Ö, Abdulkadir A, Lienkamp SS, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *MICCAI 2016*. Vol. 9901. Athens, Greece: LNCS; 2016:424–432.
31. Yang L, Zhang Y, Guldner IH, Zhang S, Chen DZ. 3D Segmentation of glial cells using fully convolutional networks and k-terminal cut. *MICCAI 2016*. Vol. 9901. Athens, Greece: LNCS; 2016:658–666.
32. Christ PF, Elshaer MEA, Ettlinger F, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. *MICCAI 2016*. Vol. 9901. Athens, Greece: LNCS; 2016: 415–423.
33. Dou Q, Chen H, Yu L, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans Med Imaging*. 2016;35:1182–1195.
34. Roth HR, Oda H, Hayashi Y, et al. Hierarchical 3D fully convolutional networks for multi-organ segmentation; 2017, arXiv. https://arxiv.org/abs/1704.06382.