An application of cascaded 3D fully convolutional networks for medical image segmentation

Holger R. Roth[a,*], Hirohisa Oda[a], Xiangrong Zhou[b], Natsuki Shimizu[a], Ying Yang[a], Yuichiro Hayashi[a], Masahiro Oda[a], Michitaka Fujiwara[c], Kazunari Misawa[d], Kensaku Mori[a,*]

[a]*Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan*
[b]*Gifu University, Yanagido, Gifu, Japan*
[c]*Nagoya University Graduate School of Medicine, Nagoya, Japan*
[d]*Aichi Cancer Center, Kanokoden, Chikusa-ku, Nagoya, Japan*

Abstract

Recent advances in 3D fully convolutional networks (FCN) have made it feasible to produce dense voxel-wise predictions of volumetric images. In this work, we show that a multi-class 3D FCN trained on manually labeled CT scans of several anatomical structures (ranging from the large organs to thin vessels) can achieve competitive segmentation results, while avoiding the need for handcrafting features or training class-specific models.

To this end, we propose a two-stage, coarse-to-fine approach that will first use a 3D FCN to roughly define a candidate region, which will then be used as input to a second 3D FCN. This reduces the number of voxels the second FCN has to classify to ∼10% and allows it to focus on more detailed segmentation of the organs and vessels.

We utilize training and validation sets consisting of 331 clinical CT images and test our models on a completely unseen data collection acquired at a different hospital that includes 150 CT scans, targeting three anatomical organs (liver, spleen, and pancreas). In challenging organs such as the pancreas, our cascaded approach improves the mean Dice score from 68.5 to 82.2%, achieving the highest reported average score on this dataset. We compare with a 2D FCN method on a separate dataset of 240 CT scans with 18 classes and achieve a significantly higher performance in small organs and vessels. Furthermore, we explore fine-tuning our models to different datasets.

Our experiments illustrate the promise and robustness of current 3D FCN based semantic segmentation of medical images, achieving state-of-the-art results.[1].

Keywords: fully convolutional networks, deep learning, medical imaging, computed tomography, multi-organ segmentation

*rothhr@mori.m.is.nagoya-u.ac.jp or kensaku@is.nagoya-u.ac.jp

[1]Our code and trained models are available for download: `github.com/holgerroth/3Dunet_abdomen_cascade`

1. Introduction

Recent advances in fully convolutional networks (FCN) have made it feasible to train models for pixel-wise segmentation in an end-to-end fashion (Long et al., 2015). Efficient implementations of 3D convolution and growing GPU memory have made it possible to extent these methods to 3D medical imaging and train networks on large amounts of annotated volumes. One such example is the recently proposed 3D U-Net (Çiçek et al., 2016), which applies a 3D FCN with skip connections to sparsely annotated biomedical images. Alternative architectures for processing volumetric images have also been successfully applied to 3D medical image segmentation (Milletari et al., 2016; Chen et al., 2016; Dou et al., 2017). In this work, we show that a 3D FCN, like 3D U-Net, trained on manually labeled data of several anatomical structures (ranging from the large organs to thin vessels) can also achieve competitive segmentation results on clinical CT images, very different from the original application of 3D U-Net using confocal microscopy images. We furthermore compare our approach to 2D FCNs applied to the same images.

Our approach applies 3D FCN architectures to problems of multi-organ and vessel segmentation in a cascaded fashion. A FCN can be trained on whole 3D CT scans. However, because of the high imbalance between background and foreground voxels (organs, vessels, etc.) the network will concentrate on differentiating the foreground from the background voxels in order to minimize the loss function used for training. While this enables the FCN to roughly segment the organs, it causes particularly smaller organs (like the pancreas or gallbladder) and vessels to suffer from inaccuracies around their boundaries.

To overcome this limitation, we learn a second-stage FCN in a cascaded manner that focuses more on the boundary regions. This is a coarse-to-fine approach in which the first-stage FCN sees around 40% of the voxels using only a simple automatically generated mask of the patient's body. In the second stage, the amount of the image's voxels is further reduced to around 10%. In effect, this step narrows down and simplifies the search space for the FCN to decide which voxels belong to the background or any of the foreground classes; this strategy has been successful in many computer vision problems (Viola and Jones, 2004; Li et al., 2016). Our approach is illustrated on a training example in Fig. 1.

1.1. Related work

Multi-organ segmentation has attracted considerable interest over the years. Classical approaches include statistical shape models (Cerrolaza et al., 2015; Okada et al., 2015), and/or employ techniques based on image registration. So called multi-atlas label fusion (Rohlfing et al., 2004; Wang et al., 2013; Iglesias and Sabuncu, 2015) has found wide application in clinical research and practice. Approaches that combine techniques from multi-atlas registration and machine learning are also common place and have been successfully applied to multi-organ segmentation in abdominal imaging (Tong et al., 2015; Oda et al., 2016). However, a fundamental disadvantage of image registration based methods is there extensive computational cost (Iglesias and Sabuncu, 2015). Typical methods need hours of computation time in order to complete on single desktop machines (Wolz et al., 2013).

The recent success of deep learning based classification and segmentation methods are now transitioning to applications of multi-class segmentation in medical imaging. Recent examples of deep learning applied to organ segmentation include (Roth et al., 2017; Zhou et al., 2016b; Christ et al., 2016; Zhou et al., 2016a). Many methods focus on the segmentation of single organs like prostate (Milletari et al., 2016), liver (Christ et al., 2016), or pancreas (Roth et al., 2015, 2016b).

Multi-organ segmentation in abdominal CT has also been approached by works like (Hu et al., 2017; Gibson et al., 2017). Most methods are based on variants of FCNs (Long et al., 2015) that either employ 2D convolutional layers in a slice-by-slice fashion Roth et al. (2016b); Zhou et al. (2016b); Christ et al. (2016); Zhou et al. (2016a), 2D convolutions on orthogonal (2.5D) cross-sections (Roth et al., 2015; Prasoon et al., 2013), and 3D convolutional layers (Milletari et al., 2016; Chen et al., 2016; Dou et al., 2017; Kamnitsas et al., 2017). A common feature of these novel segmentation methods is that they are able to extract the features useful for image segmentation directly from the training imaging data, which is crucial for the success of deep learning (LeCun et al., 2015). This avoids the need for *hand-crafting* features that are suitable for detection of individual organs.

1.2. Contributions

Due to the automatic learning of image feature and in contrast to previous approaches of multi-organ segmentation where separate models have to be created for each organ (Oda et al., 2016; Tong et al., 2015), our proposed method allows us to use the same model to segment very different anatomical structures such as large abdominal organs (liver, spleen), but also vessels like arteries and veins. Furthermore, other recent FCN-based methods that applied in medical imaging in cascaded/iterative fashion were often constrained to using rectangular bounding boxes around single organs (Roth et al., 2017; Zhou et al., 2016b) and/or performing slice-wise processing in 2D (Christ et al., 2016; Zhou et al., 2016a).

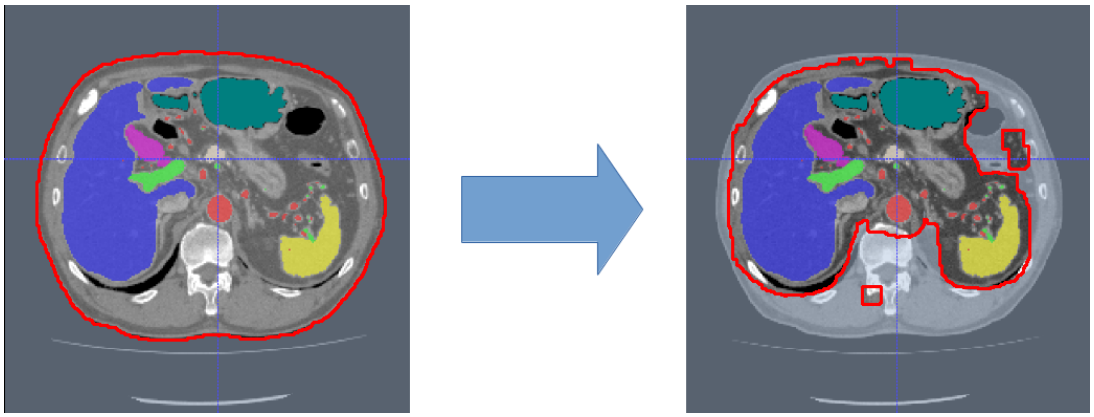


Figure 1: Cascaded 3D fully convolutional networks in a coarse-to-fine approach: the first stage (left) learns the generation of a candidate region for training a second-stage FCN (right) for finer prediction. Outlined red area shows candidate region C_1 used in first stage and C_2 used in second stage. Colored regions denote ground truth annotations for training (best viewed in color).

2. Methods

Convolutional neural networks have the ability to solve challenging classification tasks in a data-driven manner. Given a training set of images and labels $\mathbf{S} = \{(I_n, L_n), n = 1, \ldots, N\}$, I_n denotes the raw CT images and L_n denotes the ground truth label images. Each L_n contains K class labels consisting of the manual segmentations of the foreground anatomy (e.g. artery, portal vein, lungs, liver, spleen, stomach, gallbladder, and pancreas) and the background for each voxel in the CT

image. Our employed network architecture is the 3D extension by Çiçek et al. (2016) of the U-Net proposed by Ronneberger et al. (2015). U-Net, which is a type of fully convolutional network (FCN) (Long et al., 2015) was originally proposed for bio-medical image applications, utilizes deconvolution (Long et al., 2015) (or sometimes called up-convolutions (Çiçek et al., 2016)) to remap the lower resolution feature maps within the network to the denser space of the input images. This operation allows for denser voxel-to-voxel predictions in contrast to previously proposed sliding-window CNN methods where each voxel under the window is classified independently making such architecture inefficient for processing large 3D volumes. In 3D U-Net, operations such as 2D convolution, 2D max-pooling, and 2D deconvolution are replaced by their 3D counterparts (Çiçek et al., 2016). We use the open-source implementation of 3D U-Net[2] based on the Caffe deep learning library (Jia et al., 2014). The 3D U-Net architecture consists of analysis and synthesis paths with four resolution levels each. Each resolution level in the analysis path contains two $3 \times 3 \times 3$ convolutional layers, each followed by rectified linear units (ReLU) and a $2 \times 2 \times 2$ max pooling with strides of two in each dimension. In the synthesis path, the convolutional layers are replaced by deconvolutions of $2 \times 2 \times 2$ with strides of two in each dimension. These are followed by two $3 \times 3 \times 3$ convolutions, each of which has a ReLU. Furthermore, 3D U-Net employs shortcut (or skip) connections from layers of equal resolution in the analysis path to provide higher-resolution features to the synthesis path (Çiçek et al., 2016). The last layer contains a $1 \times 1 \times 1$ convolution that reduces the number of output channels to the number of class labels K. This architecture has over 19 million learnable parameters and can be trained to minimize a weighted voxel-wise cross-entropy loss (Çiçek et al., 2016). A schematic illustration of 3D U-Net is shown in Fig. 2.

2.1. Loss function: adjustments for multi-organ segmentation

The voxel-wise cross-entropy loss is defined as

$$\mathcal{L} \;=\; \frac{-1}{N} \sum_{k=1}^{K} \left(\sum_{x \in S_k} \log \left(\hat{p}_k(x) \right) \right), \tag{1}$$

where $\hat{p}_k$ are the *softmax* class probabilities

$$\hat{p}_k(x) \;=\; \frac{\exp(x_k(x))}{\sum_{k'=1}^{K} \exp(x_{k'}(x))}, \tag{2}$$

N are the total number of voxels x, S_k is the set of voxels within one class in L_n, and $k \in [1, 2, \ldots, K]$ indicates the ground truth class label. The input to this loss function is real valued output predictions $x \in [-\infty, +\infty]$ from the last convolutional layer.

However, in most cases minimizing this loss will instantly make the network converge to classifying every voxel as background. This is because of the large dominance of the background class in the images. In order to combat this large data imbalance between foreground/background voxels and differently sized organs and vessels, we apply a voxel-wise weight λ_k to this loss function (Eq. 1). In this work, we choose λ_i such that $\sum_{k=1}^{K} \lambda_k = 1$, with

$$\lambda_k = \frac{1 - N_k/N_C}{K - 1}, \tag{3}$$

[2] `http://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html`

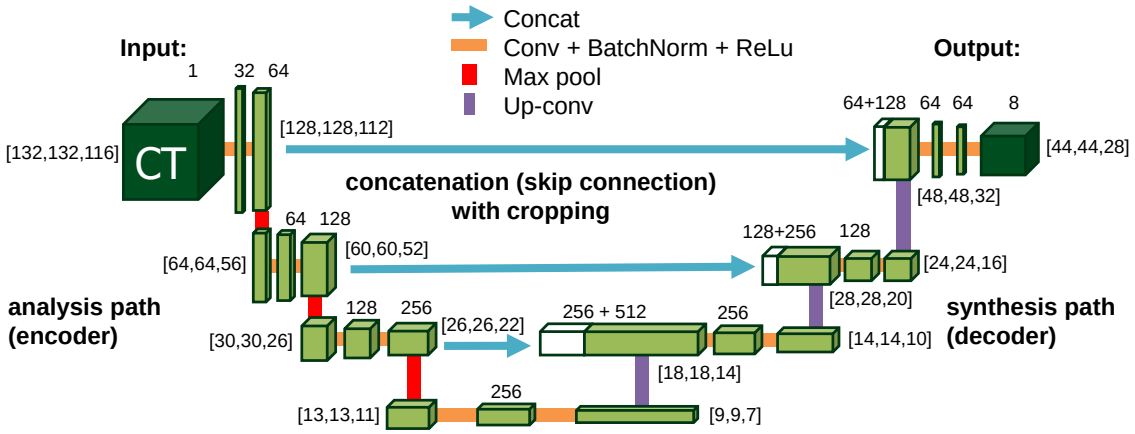


Figure 2: The architecture of 3D U-Net (Çiçek et al., 2016), a type of fully convolutional network. It applies an end-to-end architecture using only valid convolutions (*Conv*) with no padding and kernel sizes of $3 \times 3 \times 3$. Rectified Linear units (*ReLU*) are used as activation functions. This results in a smaller output size than input size and requires cropping of when mapping lower level feature maps of the analysis path to the synthesis path of the network via concatenation (*Concat*). Max-pooling (*Max pool*) is used to reduce the resolution of feature maps, while up-convolutions (*Up-conv*) are used for up-sampling the feature maps back to higher resolutions. The number of extracted feature maps is noted above each layer. We show the input and output size of feature maps at each level of the network. These parameters are kept constant for all experiments performed in this study. Batch normalization (*BatchNorm*) is used throughout the network for improved convergence (Ioffe and Szegedy, 2015).

where N_k is the number of voxels in each class S_k, and N_C is the number of voxels within a candidate region C_1 or C_2. The weights λ_i help to balance the common voxels (i.e., background) with respect to such smaller organs as vessels or the pancreas by giving more weight to the latter.

Now, the weighted cross-entropy loss can be written as:

$$\mathcal{L} = \frac{-1}{N} \sum_{k=1}^{K} \lambda_k \left(\sum_{x \in S_k} \log \left(\hat{p}_k(x) \right) \right), \tag{4}$$

We use the loss formulation in Eq. 4 for all experiments in this paper.

2.2. Coarse-to-fine prediction

In our experiments, the input to the network is fixed to a given size $N_x \times N_y \times N_z$, mainly influenced by considerations of available memory size on the GPU. In training, sub-volumes of that given size are randomly sampled from the candidate regions within the training CT images, as described below. To increase the field of view presented to the CNN and reduce informative redundancy among neighboring voxels, each image is downsampled by a factor of 2. The resulting prediction maps are then resampled back to the original resolution using nearest neighbor interpolation (or linear interpolation in case of the probability maps).

1st Stage. In the first stage, we apply simple thresholding in combination with morphological operations (hole filling and largest component selection) to get a mask of the patient's body. This mask can be utilized as candidate region C_1 to reduce the number of voxels necessary to compute the network's loss function and reduce the amount of input 3D regions shown to the CNN during training to about 40%.

2^{nd} Stage. After training the first-stage FCN, it is applied to each image to generate candidate regions C_2 for training the second-stage FCN (see Fig. 1). We define the predicted organ labels in the testing phase using the argmax of the class probability maps. All foreground labels are then dilated in 3D using a voxel radius of r in order to compute C_2, resulting in a binary candidate map.

When comparing the recall and false-positive rates of the first-stage FCN with respect to r for both the training and validation sets, $r = 3$ gives good trade-off between high recall (>99%) and low false-positive rates (∼10%) for each organ on our training and validation sets (see Fig. 6).

Our overall multi-stage training scheme is illustrated in Fig. 3

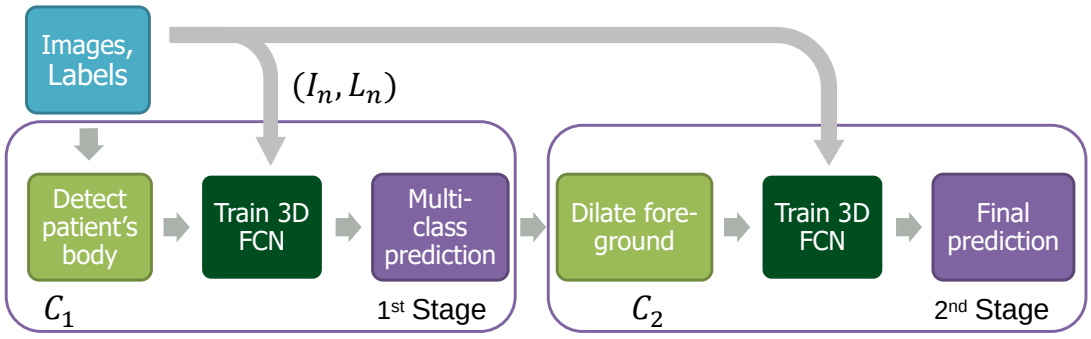


Figure 3: Flowchart of our multi-stage cascaded training scheme.

2.3. Training

The network iteratively adjusts its parameters by stochastic gradient descent. Batch normalization is used throughout the network for improved convergence and we utilize random elastic deformations in 3D during training to artificially increase the amount of available data samples and increase robustness, similar to (Çiçek et al., 2016). Hence, we randomly sample deformation fields from a uniform distribution with a maximum displacement of ±4 and a grid spacing of 32 voxels (see Fig. 4). Furthermore, we applied random rotations between $-5°$ and $+5°$, and translations of -20 to 20 voxels in each direction at each iteration in order to generate plausible deformations during training. Each training sub-volume is randomly extracted from C_1 or C_2 in both stages.

2.4. Testing

The CT image is processed by the 3D FCN using a tiling strategy (sliding-window) (Çiçek et al., 2016) as illustrated in Fig. 5. For greater speed, we use non-overlapping tiles in the first stage and investigate the use of non-overlapping and overlapping tiles in the second. When using overlapping tiles (with a 4× higher sampling rate of each voxel x), the resulting probabilities for the overlapping voxels are averaged:

$$p(x) = \frac{1}{R} \sum_{r=1}^{R} p_r(x). \tag{5}$$

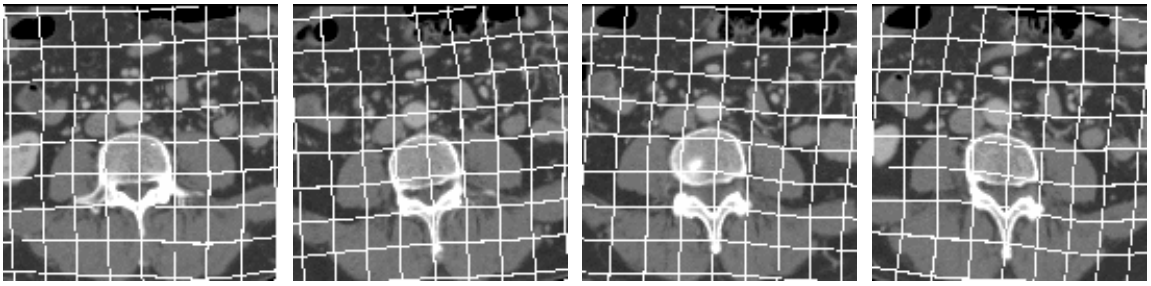


Figure 4: Axial cross-section through the same patient CT image at various examples of plausible random deformation during training. A deformed grid pattern is overlaid in order to better illustrate the applied deformation. At each iteration, the random deformation is computed on the fly.

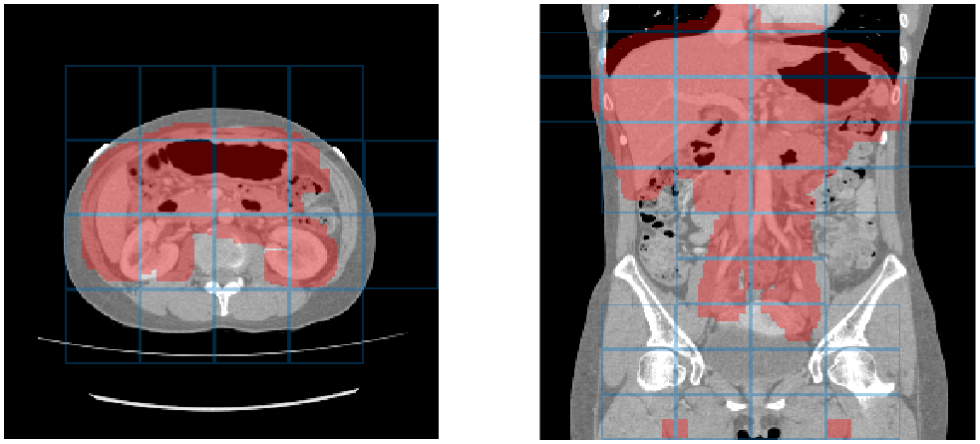
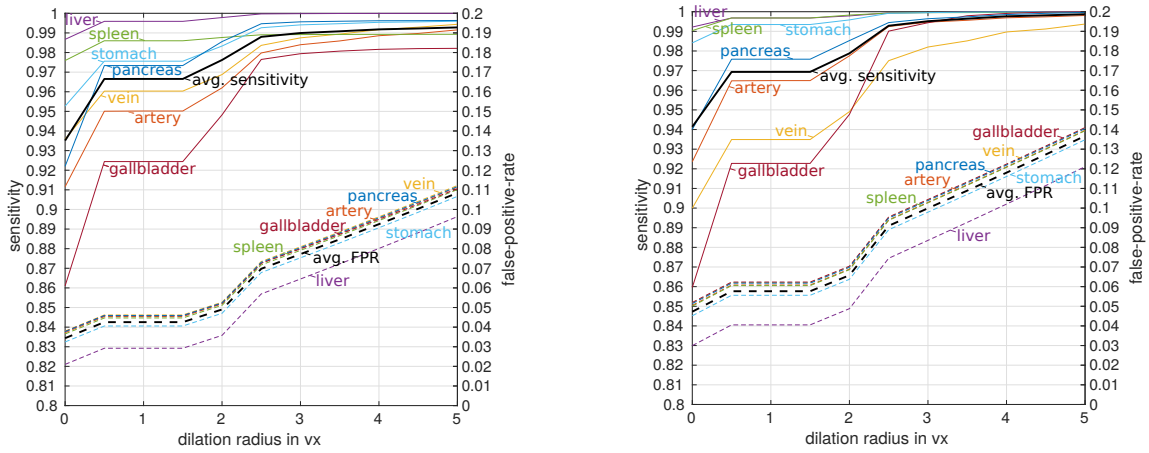


Figure 5: The non-overlapping tiling approach on second stage candidate region C_2. Note that the grid shows the output tiles of size $44 \times 44 \times 28$ (x, y, z-directions). Each predicted tile is based on a larger input of $132 \times 132 \times 116$ that the network processes.

3. Experiments & Results

3.1. Training and validation

Our dataset includes 331 contrast-enhanced abdominal clinical CT images in the portal venous phase used for pre-operative planning in gastric surgery. Each CT volume consists of $460 - 1177$ slices of 512×512 pixels. The voxel dimensions are [0.59-0.98, 0.59-0.98, 0.5-1.0] mm. A random split of 281/50 patients is used for training and validating the network, i.e., determining when to stop training to avoid overfitting. In order to generate plausible deformations during training, we sample from a normal distribution with a standard derivation of 4 and a grid spacing of 32 voxels, and apply random rotations between $-5°$ and $+5°$ to the training images. No deformations were applied during testing. We trained 200,000 iterations in the first stage and 115,000 in the second. Table 1 summarizes the Dice similarity scores for each organ labeled in the 50 validation cases. On average, we achieved a 7.5% improvement in Dice scores per organ. Small, thin organs such as arteries especially benefit from our two-stage cascaded approach. For example, the mean Dice



(a) Training (b) Validation

Figure 6: Sensitivity and false-positive-rate (FPR) as a function of dilating prediction maps of first stage in training (a) and validation (b). We observe good trade-off between high sensitivity (>99% on average) and low false-positive-rate (∼10% on average) at dilation radius of $r = 3$.

score for arteries improved from 59.0 to 79.6% and from 54.8 to 63.1% for the pancreas. The effect is less pronounced for large organs, like the liver, the spleen, and the stomach. Fig. 7 shows an example result from the validation set and illustrates the tiling approach. The 3D U-Net separates the foreground organs well from the background tissue of the images.

3.2. Testing

Our test set is different from our training and validation data. It originates from a different hospital, scanners, and research study with gastric cancer patients. 150 abdominal CT scans were acquired in the portal venous phase. Each CT volume consists of $263 - 1061$ slices of 512×512 pixels. Voxel dimensions are [0.55-0.82, 0.55-0.82, 0.4-0.80] mm. The pancreas, liver, and spleen

Table 1: **Validation set:** Dice similarity score [%] of different stages of FCN processing

Stage 1: Non-overlapping

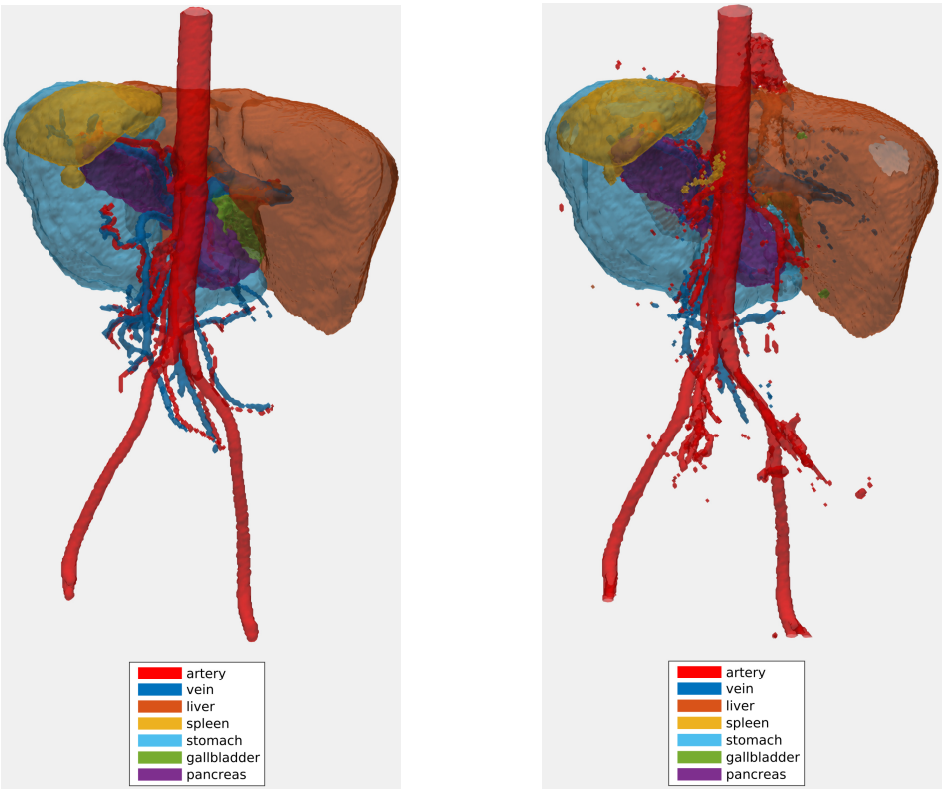
Dice	artery	vein	liver	spleen	stomach	gallbladder	pancreas	Mean
Mean	59.0	64.7	89.6	84.1	80.0	69.6	54.8	71.7
Std	7.8	8.6	1.7	4.7	18.3	14.1	11.0	9.5
Median	59.8	67.3	90.0	85.2	87.5	73.2	57.2	74.3
Min	41.0	34.5	84.4	70.9	8.4	13.8	23.5	39.5
Max	75.7	76.0	92.6	91.4	94.8	86.8	72.0	84.2

Stage 2: Non-overlapping

Dice	artery	vein	liver	spleen	stomach	gallbladder	pancreas	Mean
Mean	79.6	73.1	93.2	90.6	84.3	70.6	63.1	79.2
Std	6.5	7.9	1.5	2.8	17.3	15.9	10.7	8.9
Median	82.3	74.6	93.5	91.2	90.9	77.3	64.5	82.1
Min	62.9	33.3	88.9	82.3	10.9	13.0	32.4	46.2
Max	87.0	83.2	95.6	95.1	96.3	89.4	81.8	89.8

Stage2 vs Stage1

Dice	artery	vein	liver	spleen	stomach	gallbladder	pancreas	Mean
Mean	20.61	8.41	3.60	6.42	4.22	0.93	8.26	7.49
Std	-1.24	-0.68	-0.18	-1.97	-0.97	1.78	-0.35	-0.52
Median	22.57	7.34	3.42	6.00	3.44	4.15	7.31	7.75
Min	21.83	-1.20	4.47	11.35	2.44	-0.75	8.85	6.71
Max	11.28	7.21	3.06	3.70	1.52	2.67	9.74	5.60



(a) Ground truth (b) Stage 2 - N/OL

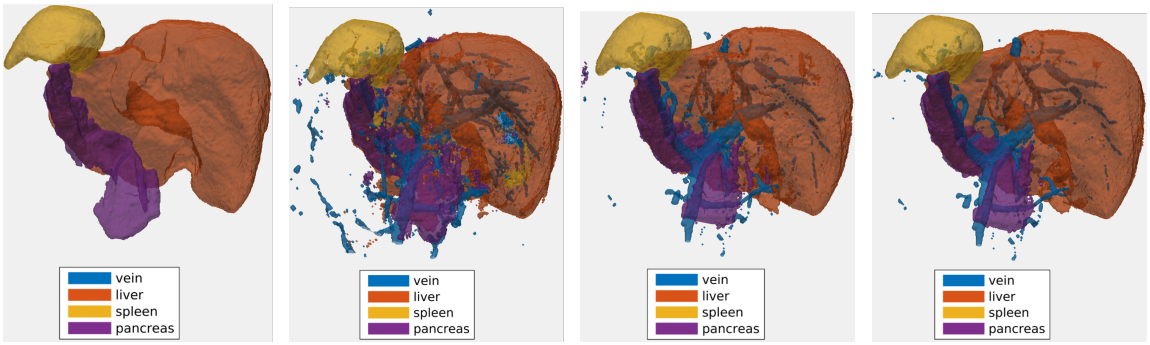
Figure 7: Example of the validation set with (a) ground truth and (b) the corresponding non-overlapping (N/OL) segmentation result. The posterior to anterior view is shown to visualize the inner organs.

were semi-automatically delineated by three trained researchers and confirmed by a clinician. Figure 8 shows surface renderings for comparison of the different stages of the algorithm. A typical testing case in the first and second stages is shown using non-overlapping and overlapping tiles. Dice similarity scores are listed in Table 2. The second stage achieves the highest reported average score for pancreas in this dataset with 82.2% ±10.2%. Previous state of the art on this dataset was at 75.1% ±15.4% while using leave-one-out-validation (Oda et al., 2016).

The testing dataset provides slightly higher image quality than our training/validation dataset. Furthermore, its field of view is more constrained to the upper abdomen. This likely explains the improved performance for liver and pancreas compared to the validation set in Table 1.

Table 2: **Testing on unseen dataset:** Dice similarity score [%] of different stages of FCN processing.

	Stage 1: Non-overlapping			**Stage 2: non-overlapping**			**Stage 2: Overlapping**		
Dice	**liver**	**spleen**	**pancreas**	**liver**	**spleen**	**pancreas**	**liver**	**spleen**	**pancreas**
Mean	93.6	89.7	68.5	94.9	91.4	81.2	95.4	92.8	82.2
Std	2.5	8.2	8.2	2.1	8.9	10.2	2.0	8.0	10.2
Median	94.2	91.8	70.3	95.4	94.2	83.1	96.0	95.4	84.5
Min	78.2	20.6	32.0	80.4	22.3	1.9	80.9	21.7	1.8
Max	96.8	95.7	82.3	97.3	97.4	91.3	97.7	98.1	92.2



(a) Ground truth (b) Stage 1 (c) Stage 2: N/OL (d) Stage 2: OL

Figure 8: Surface renderings: (a) ground truth segmentation, (b) result of proposed method in first stage, second-stage results using (c) non-overlapping (N/OL), and (d) overlapping (OL) tiles strategy. The posterior to anterior view is shown for better visualization of the pancreas.

3.3. Comparison to other methods

Even though direct comparison is difficult due to the differences in datasets, training/testing evaluation schemes, and segmented organs, we try to indicate how well our model performed with respect to recent state-of-the-art methods in Table 3. In particualar, we provide a comparison to recent methods on two different datasets: (1) our own *in-house* dataset for pancreas segmentation, acquired at Nagoya University Hospital, Japan, and consisting of 150 CT images; and (2) the publicly available *TCIA Pancreas-CT* dataset of 82 patient images[3] (Roth et al., 2016a). For comparison with (2), we use the same 4-fold cross-validation (CV) split as in (Roth et al., 2015, 2017).

Our results on dataset (1) achieves the highest reported performance in testing. On the other hand, our results on the public dataset (2) are comparable to other recent works that developed methods especially targeting this dataset and focusing on pancreas segmentation alone (Roth et al., 2017; Zhou et al., 2016b).

[3] `https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT` (Roth et al., 2016a) hosted by TCIA (Clark et al., 2013).

Table 3: **Comparison to other methods.** We list other recent segmentation work performed on the same/similar datasets and organs and based on atlas-based segmentation propagation using global affine (Wang et al., 2014), local non-rigid registration methods (Wolz et al., 2013) and in combination with machine learning (ML) (Tong et al., 2015). We also list a method using regression forest (RF) and graph cut (GC) (Oda et al., 2016), and two other methods utilizing 2D FCNs (Roth et al., 2017; Zhou et al., 2016b). Validation of other methods was performed using either leave-one-out-validation (LOOV) or cross-validation(CV). Best performance is shown in **bold**.

Method	**Subjects**	**Approach**	**Validation**	**Organs**	**Dice [%]**	**Time [h]**
			(1) In-house dataset			
Proposed	150	3D FCN	Testing	Liver	**95.4± 2.0**	0.07
				Spleen	**92.8± 8.0**	
				Pancreas	**82.2± 10.2**	
Tong et al. (2015)	150	Global affine + ML	LOOV	Liver	94.9 ± 1.9	0.5
				Spleen	92.5 ± 6.5	
				Pancreas	71.1 ± 14.7	
Wang et al. (2014)	100	Global affine	LOOV	Liver	94.5 ± 2.5	14
				Spleen	92.5 ± 8.4	
				Pancreas	65.5 ± 18.6	
Wolz et al. (2013)	150	Local non-rigid	LOOV	Liver	94.0 ± 2.8	3
				Spleen	92.0 ± 9.2	
				Pancreas	69.6 ± 16.7	
Oda et al. (2016)	147	RF + GC	LOOV	Pancreas	75.1 ± 15.4	3
			(2) TCIA Pancreas-CT dataset			
Proposed	82	3D FCN	4-fold CV	Pancreas	76.8 ± 9.4	0.07
Roth et al. (2017)	82	2D FCN	4-fold CV	Pancreas	81.3 ± 6.3	0.05
Zhou et al. (2016b)	82	2D FCN	4-fold CV	Pancreas	**82.4± 5.7**	n/a

3.4. Direct comparison to 2D FCN networks

Furthermore, we implement the method of Zhou et al. (Zhou et al., 2016a, 2017) and apply it to the same dataset. This method employs a combination of three 2D FCNs trained on the orthogonal planes of the images. The results of each model are then fused by majority voting. This dataset consists of 240 3D CT scans with 18 manually annotated organs. A split of 228/12 cases was used for our training/testing as in (Zhou et al., 2016a, 2017). A direct comparison can be seen in Table 4. It can be observed that our 3D FCN approach has a clear advantage for the smaller, thinner organs (like aorta, esophagus, gallbladder, inferior vena cava, portal vein, and prostate) but only performs comparable to the 2D FCNs when aiming at the larger organs (like lungs, liver, kidneys). Furthermore, a slightly higher overall performance can be observed for the average of all organ/vessel predictions when using the proposed cascaded 3D FCN approach.

3.5. Computation

Training on 281 cases can take 2-3 days for 200-k iterations on a NVIDIA GeForce GTX TITAN X with 12 GB memory. However, in testing, the processing time for each volume was 1.4-3.3 minutes for each stage, depending on the size of the candidate regions; and 1.6-4.4 minutes using overlapping tiles in the second stage. In order to achieve optimal GPU memory usage in training, we keep the input subvolume size at $N_{\mathrm{x}} \times N_{\mathrm{y}} \times N_{\mathrm{z}} = 132 \times 132 \times 116$, resulting in an output size of $44 \times 44 \times 28$ for each class output channel as in (Çiçek et al., 2016).

Table 4: Direct comparison of the proposed cascaded 3D FCN approach against a 2D FCN approach using a majority voting scheme as in (Zhou et al., 2016a, 2017). 18 different anatomical structures are compared using the Dice similarity score [%]. Significantly better performance is shown in bold ($p < 0.05$, Wilcoxon signed-rank test).

Label (Dice)	**2D (Zhou et al., 2016a)**	**3D (stage1)**	**3D (stage2)**	**p-value**
right lung	**94.6 ± 2.8**	90.1 ± 4.0	91.9 ± 3.6	0.028
left lung	92.8 ± 3.8	87.8 ± 5.3	93.2 ± 4.0	0.959
heart	**91.2 ± 4.0**	70.9 ± 11.8	86.2 ± 5.6	0.016
aorta	76.0 ± 11.8	50.7 ± 5.4	**82.3 ± 5.9**	0.038
esophagus	24.6 ± 16.7	0.0 ± 0.0	**51.9 ± 5.3**	0.011
liver	**94.3 ± 3.3**	90.2 ± 3.5	93.6 ± 2.7	0.049
gallbladder	47.5 ± 39.9	9.1 ± 11.6	**58.4 ± 33.2**	0.011
stomach and duodenal	68.0 ± 19.1	58.2 ± 15.2	61.9 ± 13.4	0.070
stomach and duodenal (air)	**64.0 ± 32.2**	52.4 ± 27.4	48.8 ± 26.6	0.001
stomach and duodenal (not air)	8.5 ± 15.6	1.1 ± 1.7	**20.7 ± 20.0**	0.000
spleen	86.7 ± 14.5	81.2 ± 12.1	86.6 ± 6.6	0.326
right kidney	92.2 ± 2.1	80.9 ± 10.6	90.8 ± 6.7	0.918
left kidney	90.2 ± 4.0	82.8 ± 8.3	86.1 ± 11.1	0.179
inferior vena cava	63.6 ± 19.1	59.8 ± 15.2	**70.6 ± 17.5**	0.007
portal vein	33.6 ± 29.5	30.0 ± 21.3	**56.0 ± 16.6**	0.002
pancreas	55.4 ± 19.3	47.5 ± 16.3	**71.0 ± 13.9**	0.001
prostate	1.2 ± 2.3	0.0 ± 0.0	**38.0 ± 29.3**	0.008
bladder	78.8 ± 13.9	57.2 ± 14.0	71.6 ± 20.8	0.213
Mean	65.0	52.8	**69.3**	0.004
Std	33.4	32.2	26.1	
Min	0.0	0.0	1.1	
Max	97.7	90.2	97.3	

4. Fine-tuning to other datasets

One advantage of deep learning based models is their ability to transfer learned features across dataset domains (Shin et al., 2016). To this end, we trained a general FCN model employing the 3D U-Net architecture (Çiçek et al., 2016) on the large dataset of CT scans including the major abdominal organ labels of Section 3.1. This model can then be fine-tuned to other (smaller) datasets aiming at more detailed classification tasks or different field of views. For this purpose, we utilize separate training, fine-tuning, and testing datasets. As mentioned above, the general training set consists of 280 clinical CT images with seven abdominal structures (artery, vein, liver, spleen, stomach, gallbladder, and pancreas) labeled.

We then fine-tune on a much smaller dataset consisting only of 20 contrast enhanced CT images from the Visceral Challenge dataset[4] (Jimenez-del Toro et al., 2016), but with substantially more anatomical structures labeled in each image (20 in total). This fine-tuning process across different datasets is illustrated in Fig. 9 with some ground truth label examples used for pre-training and fine-tuning. In fine-tuning, we use a 10 times smaller learning rate. We furthermore test our models on a completely unseen data collection of 10 torso CT images with 8 labels, including organs that were not labeled in the original abdominal dataset, e.g. the kidneys and lungs. A probabilistic output for kidney (not in the pre-training dataset) from our model is shown in Fig. 10.

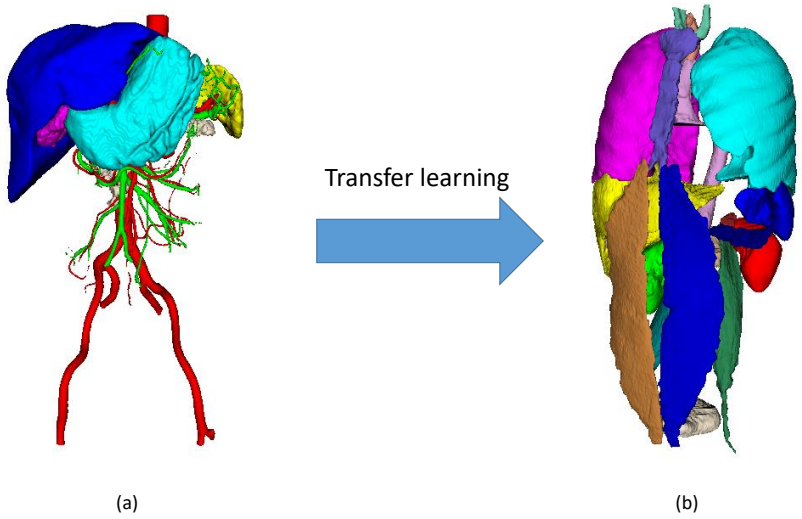


Figure 9: We fine-tune our model via transfer learning from 8 anatomical structures in the abdomen (a) to 20 anatomical structures in the whole torso (b). We show some typical ground truth labels that are used for training on both datasets.

[4] http://www.visceral.eu/benchmarks/anatomy3-open/

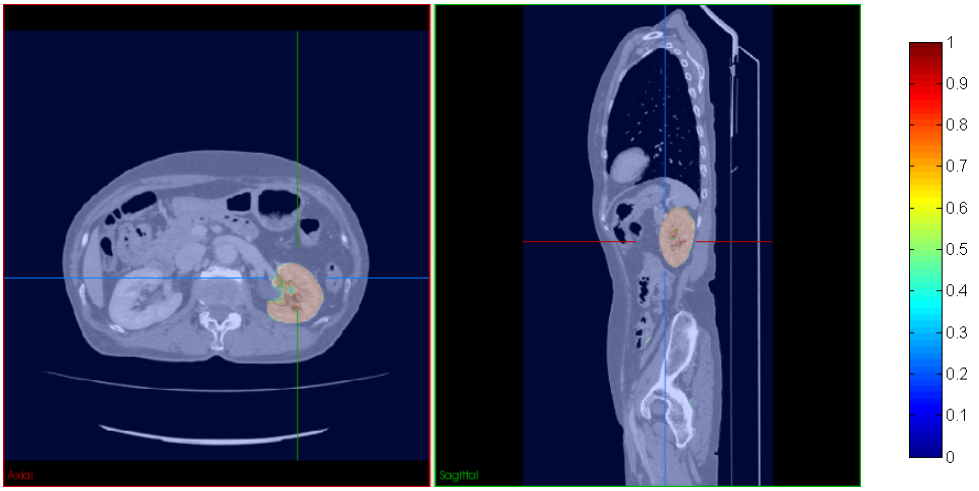


Figure 10: Automated probability map for left kidney after transfer learning.

Table 5: **Testing on unseen whole torso dataset:** Dice scores [%] for each segmented organ.

Dice	**r. lung**	**l. lung**	**liver**	**gall**	**spleen**	**r. kidney**	**l. kidney**	**pancreas**	**Avg.**
scratch	96.2	96.3	94.0	74.9	91.0	87.6	84.1	32.0	82.0
fine-tuned	96.4	96.6	94.9	76.3	90.1	90.5	88.5	33.0	83.3

4.1. Fine-tuning results

In testing, we deploy our fine-tuned model using a non-overlapping tiling approach as in previous sections. An automated segmentation result on the unseen test dataset by our fine-tuned model is shown in Fig. 11. Our fine-tuned approach provides a Dice score of right lung, left lung, liver, gall bladder, spleen, right kidney, left kidney, and pancreas are 0.96, 0.97, 0.95, 0.77, 0.90, 0.90, 0.88, and 0.36, respectively (summarized in Table 5). The relatively lower score for pancreas is due to several outlier cases on this dataset. These outliers are likely caused by variations of contrast enhancement across the datasets and the higher variability of the pancreas shape and intensity profile compared to other organs across different patients.

Our approach and results, however, illustrate the generalizability and robustness of our models across different datasets. Fine-tuning can be useful when the amount of training examples for some target organs are limited. In this case, transfer learning achieves slight improvements over learning from scratch, especially in the kidneys (see Table 5). It should be noted that for this particular application, data augmentation already gives a good performance when learning models from scratch.

5. Discussion

The cascaded coarse-to-fine approach presented in this paper provides a simple yet effective method for employing 3D FCNs in medical imaging settings. No post-processing was applied to any of the FCN outputs. The improved performance stemming from our cascaded approach is especially visible in smaller, thinner organs, such as arteries and veins, particularly when compared

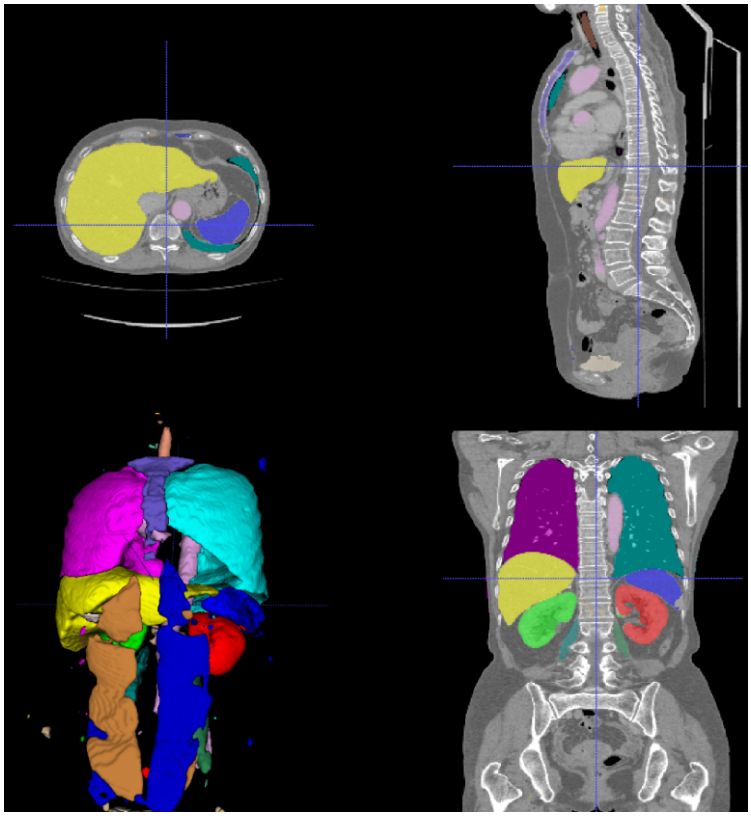


Figure 11: Multi-organ segmentation result. Each color represents an organ region on the unseen whole torso test set.

to other recent FCN approaches using 2D FCNs (Zhou et al., 2016a). Our results and recent literature indicate that 2D FCNs and especially the combination of orthogonally applied 2D FCNs (Zhou et al., 2016a; Roth et al., 2017; Zhou et al., 2016b) might be sufficient for larger and mid-sized organs. In fact, the combination of 2D FCNs even slightly outperforms our 3D approach for some organs. On the other hand 3D convolutional kernels are important for distinguishing the thin (vessel-like) and small organs as can be seen in the improved performance of our approach. When compared to other cascaded approaches using 2D FCNs that focus on single organs (Roth et al., 2017; Zhou et al., 2016b), we perform similar to the state of the art. Our findings are also consistent with (Roth et al., 2017; Zhou et al., 2016b) that show that cascaded approaches are useful for applying deep learning methods to medical image segmentation. Note that we used different datasets (from different hospitals and scanners) for separate training/validation and testing. These experiments illustrate our method's generalizability and robustness to differences in image quality and populations. Running the algorithms at half resolution allows efficient training on a single GPU. In contrast, using the same field of view for each subvolume with the original resolution would require $8\times$ more memory with the current architecture and would force us to reduce the amount of context visible to the 3D FCNs. In this work, we utilized 3D U-Net for the segmentation of CT scans. However, the proposed cascaded approach in principle should also work well for

other 3D CNN/FCN architectures and 3D image modalities. Exploration of other loss functions such as the Dice score (Milletari et al., 2016; Li et al., 2017) could help further in dealing with the class imbalance issue. We used Caffe's stochastic gradient descent solver (Jia et al., 2014) for all experiments in this work. Alternative optimizers could further improve training performance (Kingma and Ba, 2014).

In the future, prediction results from different models could be combined in order to achieve the best overall performance. Furthermore, additional anatomical constraints could be included in order to guarantee topologically correct segmentation results (BenTaieb and Hamarneh, 2016; Oktay et al., 2017). With growing amounts of available GPU memory, the need for computing overlapping sub-volume predictions as in this work will be reduced as it will be come possible to reshape the network to accept arbitrary 3D input image sizes (Long et al., 2015).

6. Conclusion

In conclusion, we showed that a cascaded deployment of volumetric fully convolutional networks (3D U-Net) can produce competitive results for medical image segmentation on a clinical CT dataset while being efficiently deployed on a single GPU. An overlapping tiles approach during testing produces better results with only moderate additional computational cost. The proposed method compares favorably to recent state-of-the-art work on a completely unseen dataset. Our results indicate that 3D convolutional features are advantageous for detecting smaller organs and vessel. A promising future direction might be hybrid approaches that combine 2D and 3D FCN-type architectures at multiple scales. We have made our code, pre-trained models, and fine-tuned models available for download[5] in order to allow further applications and fine-tuning to different datasets.

Acknowledgments This paper was supported by MEXT KAKENHI (26108006, 26560255, 25242047, 17H00867, 15H01116) and the JPSP International Bilateral Collaboration Grant.

Conflict of interest statement: The authors declare that they have no conflict of interest.

References

BenTaieb, A., Hamarneh, G., 2016. Topology aware fully convolutional networks for histology gland segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 460–468.

Cerrolaza, J. J., Reyes, M., Summers, R. M., González-Ballester, M. Á., Linguraru, M. G., 2015. Automatic multi-resolution shape modeling of multi-organ structures. Medical image analysis 25 (1), 11–21.

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

[5] `https://github.com/holgerroth/3Dunet_abdomen_cascade`

Christ, P. F., Elshaer, M. E. A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., DAnastasi, M., Sommer, W. H., Ahmadi, S.-A., Menze, B. H., 2016. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3D conditional random fields. In: MICCAI. Springer, pp. 415–423.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. Springer, pp. 424–432.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging 26 (6), 1045–1057.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3d deeply supervised network for automated segmentation of volumetric medical images. Medical Image Analysis.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B. R., Pereira, S. P., Clarkson, M. J., Barratt, D. C., 2017. Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal ct with dense dilated networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 728–736.

Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F., Kong, D., 2017. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. International journal of computer assisted radiology and surgery 12 (3), 399–411.

Iglesias, J. E., Sabuncu, M. R., 2015. Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24 (1), 205–219.

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

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp. 675–678.

Jimenez-del Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., Kontokotsios, G., Langs, G., Menze, B., Salas Fernandez, T., Schaer, R., Walleyo, A., Weber, M., Dicente Cid, Y., Gass, T., Heinrich, M., Jia, F., Kahl, F., Kechichian, R., Mai, D., Spanier, A., Vincent, G., Wang, C., Wyeth, D., Hanbury, A., 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. IEEE transactions on medical imaging 35 (11), 2459–2475.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, K., Hariharan, B., Malik, J., 2016. Iterative instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3659–3667.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 348–360.

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

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

Oda, M., Shimizu, N., Karasawa, K., Nimura, Y., Kitasaka, T., Misawa, K., Fujiwara, M., Rueckert, D., Mori, K., 2016. Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In: MICCAI. Springer, pp. 556–563.

Okada, T., Linguraru, M. G., Hori, M., Summers, R. M., Tomiyama, N., Sato, Y., 2015. Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors. Medical image analysis 26 (1), 1–18.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Guerrero, R., Cook, S., de Marvao, A., O'Regan, D., Kainz, B., Glocker, B., Rueckert, D., 2017. Anatomically constrained neural networks (acnn): Application to cardiac image enhancement and segmentation. arXiv preprint arXiv:1705.08302.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 246–253.

Rohlfing, T., Brandt, R., Menzel, R., Maurer, C. R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Roth, H. R., Farag, A., Turkbey, E. B., Lu, L., Liu, J., Summers, R. M., 2016a. Data citation: Data from pancreas-ct. the cancer imaging archive. http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBUAccessed: March 20, 2018.

Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., Summers, R. M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 556–564.

Roth, H. R., Lu, L., Farag, A., Sohn, A., Summers, R. M., 2016b. Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 451–459.

Roth, H. R., Lu, L., Lay, N., Harrison, A. P., Farag, A., Sohn, A., Summers, R. M., 2017. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. arXiv preprint arXiv:1702.00045.

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.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J. V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. Medical Image Analysis 23 (1), 92–104.

Viola, P., Jones, M. J., 2004. Robust real-time face detection. International journal of computer vision 57 (2), 137–154.

Wang, H., Pouch, A., Takabe, M., Jackson, B., Gorman, J., Gorman, R., Yushkevich, P. A., 2013. Multi-atlas segmentation with robust label transfer and label fusion. In: Information processing in medical imaging: proceedings of the... conference. Vol. 23. NIH Public Access, p. 548.

Wang, Z., Bhatia, K. K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K., Rueckert, D., 2014. Geodesic patch-based segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 666–673.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Transactions on Medical Imaging 32 (9), 1723–1730.

Zhou, X., Ito, T., Takayama, R., Wang, S., Hara, T., Fujita, H., 2016a. Three-dimensional ct image segmentation by combining 2D fully convolutional network with 3D majority voting. In: LABELS workshop. Springer, pp. 111–120.

Zhou, X., Takayama, R., Wang, S., Hara, T., Fujita, H., 2017. Deep learning of the sectional appearances of 3d ct images for anatomical structure segmentation based on an fcn voting method. Medical Physics.

Zhou, Y., Xie, L., Shen, W., Fishman, E., Yuille, A., 2016b. Pancreas segmentation in abdominal ct scan: A coarse-to-fine approach. arXiv preprint arXiv:1612.08230.