Published in final edited form as:

IEEE Trans Biomed Eng. 2020 October ; 67(10): 2710–2720. doi:10.1109/TBME.2020.2969608.

Iterative Label Denoising Network: Segmenting Male Pelvic Organs in CT from 3D Bounding Box Annotations

Shuai Wang [Member, IEEE],

Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Qian Wang,

Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

Yeqin Shao,

School of Transportation and Civil Engineering, Nantong University, Jiangsu 226019, China

Liangqiong Qu,

Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Chunfeng Lian,

Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Jun Lian,

Department of Radiation Oncology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA

Dinggang Shen* [Fellow, IEEE]

Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea.

Abstract

Obtaining accurate segmentation of the prostate and nearby organs at risk (e.g., bladder and rectum) in CT images is critical for radiotherapy of prostate cancer. Currently, the leading automatic segmentation algorithms are based on Fully Convolutional Networks (FCNs), which achieve remarkable performance but usually need large-scale datasets with high-quality voxel-wise annotations for full supervision of the training. Unfortunately, such annotations are difficult to acquire, which becomes a bottleneck to build accurate segmentation models in real clinical applications. In this paper, we propose a novel weakly supervised segmentation approach that only needs 3D bounding box annotations covering the organs of interest to start the training. Obviously, the bounding box includes many non-organ voxels that carry noisy labels to mislead the

segmentation model. To this end, we propose the label denoising module and embed it into the iterative training scheme of the label denoising network (LDnet) for segmentation. The labels of the training voxels are predicted by the tentative LDnet, while the label denoising module identifies the voxels with unreliable labels. As only the good training voxels are preserved, the iteratively re-trained LDnet can refine its segmentation capability gradually. Our results are remarkable, i.e., reaching ~94% (prostate), ~91% (bladder), and ~86% (rectum) of the Dice Similarity Coefficients (DSCs), compared to the case of fully supervised learning upon high-quality voxel-wise annotations and also superior to several state-of-the-art approaches. To our best knowledge, this is the first work to achieve voxel-wise segmentation in CT images from simple 3D bounding box annotations, which can greatly reduce many labeling efforts and meet the demands of the practical clinical applications.

Keywords

Image Segmentation; Bounding Box Annotation; Weakly Supervised Learning; Fully Convolutional Network (FCN); Pelvic Organ; CT

I. INTRODUCTION

PROSTATE cancer, a potentially risky disease for all men, is estimated with 174,650 new cases in the US during 2019 [1]. Nowadays, external beam radiation therapy (EBRT) for prostate cancer is one of the standard and also primary ways to treat prostate cancer patients. In EBRT, an essential step is to accurately delineate the prostate and the nearby organs at risk (e.g., bladder and rectum) in CT image for the maximum killing of cancer cells while protecting healthy cells. However, manual delineation is laborious even for experienced oncologists and often suffers from large inter-operator variability. To address this problem, many studies have been reported to develop automatic organ segmentation methods for male pelvic CT [2]–[6]. Especially, in recent days, deep-learning - based methods powered by massive amounts of voxel-wise annotations have shown remarkable performance in medical image segmentation [7]–[9].

However, one of the major challenges in training a powerful deep-learning-based segmentation model is the requirement of the large dataset with high-quality voxel-wise annotations. Such detailed delineation costs high in manual efforts and poses difficulties in collecting enough training data in practice. Hence, weak yet low-cost annotations (e.g., image-level labels, bounding boxes or scribbles of the tumors and organs) are often employed, which may yield comparably accurate results concerning the costly voxel-wise annotations [10]–[13].

Among different forms of weak annotations, bounding box annotation (e.g., covering an organ in the image) is user-friendly and cheap to acquire. Note that the 3D bounding box annotations in this work, as illustrated by Fig. 1, require oncologists to place a cube to fully and compactly enclose the target organs in the images, which can be implemented by visualizing the volumetric images from three different views and marking the first/last slice of the organ for each view. While it takes an experienced oncologist 10 - 12 minutes to manually delineate the whole prostate boundary [14], it just takes approximately 15 seconds

to place the 3D bounding box for the prostate. This low-cost annotation facilitates the collection of large datasets and is more suitable for real clinical applications.

There have been some studies to perform semantic segmentation by combining deep learning with box annotations [11], [15], [16]. Unfortunately, these approaches cannot be directly applied to pelvic organ segmentation in CT images with only 3D bounding box annotations available. The existing methods may suffer from two reasons. First, most of them can handle only 2D bounding boxes rather than 3D ones. Second, compared with natural images [15], [16] or MR images [11] in early works, pelvic CT images come with low contrast and blurred organ boundaries, which is challenging even to fully supervised segmentation with high-quality voxel-wise annotations.

If simply assuming all voxels inside a 3D bounding box to be belonging to the target organ, the learned feature representations are insufficient to well separate organ and non-organ voxels. The reason is that the segmentation model can be fooled by many voxels in the bounding box yet not belonging to the target organ. To this end, we argue that it is essentially important to refine the voxel-wise labels in the annotated 3D bounding boxes and suppress the supervision from the misleading voxels. Some boosting approaches have been proposed to iteratively filter out weak or noisy samples [17], [18]. Classical boosting approaches such as AdaBoost [19] and Gradient Boosting [20] learn ensembles of classifiers fully supervised by high-quality annotations. However, these approaches may be sensitive to significant label noise since the sample selection is directly under the guidance of the labels with high noise [18], [21]. Combining deep learning techniques with boosting strategy, some deep learning boosting methods are proposed [22]–[26]. Unfortunately, these approaches usually require high-quality labels for training or rely on sophisticated sample selection. The limited access to the clean labels and the poorly trained model from highly noisy data make them not suitable to address our problem.

Different from the mentioned works, we start from 3D bounding box annotations and design a Fully Convolutional Network (FCN) to predict the voxel-wise labels in the training images. Then, we propose a novel label denoising module to identify the training voxels with noisy labels. Those unreliable voxels are then excluded from subsequent training. The above procedure is iteratively called, leading to the continuously updated label denoising network (LDnet) that can eventually well segment the target organ.

To our best knowledge, this is the first work to attain voxel-wise pelvic organ segmentation of male CT from 3D bounding box annotations yet without using any other prior. The main contributions of this work are three-fold: 1) We propose a novel deep learning method, along with an iterative training scheme, to achieve voxel-wise segmentation, given only the 3D bounding box annotations for training in the beginning; 2) We design a novel label denoising module by measuring the predicted label stability across consecutive training iterations, which can effectively identify and suppress the noisy training labels; 3) Our work obtains remarkable segmentation performance on a challenging pelvic CT dataset. Specifically, our method reaches ~94% (prostate), ~91% (bladder), and ~86% (rectum) of the Dice Similarity Coefficients (DSCs) compared to the case of fully supervised learning upon high-quality voxel-wise annotations.

II. RELATED WORKS

A. Learning-Based Image Segmentation

Medical image segmentation is a critical step for many real clinical applications, such as computer-assisted diagnosis, image-guided surgery, and radiological treatment [27]–[29]. For automatic image segmentation, learning-based methods have attracted increasing attention due to their remarkable performance. Different from multi-atlas-based methods and deformable model-based methods [30], learning-based methods get rid of time-consuming registration and sensitive shape initialization. Some early works have adapted shallow learning techniques, such as Random Forests [31], Support Vector Machine [32], [33], and Logical Regression [34], to assign a single label/value to each voxel relying on hand-crafted features. To enhance the representation power, Kallel et al. [35] imitated human visual attention in image interpretation by using possibilistic knowledge modeling for blind medical image segmentation. Gao et al. [36] used the detected boundary based on a regression forest as the shape prior to guide prostate segmentation. Bianchi et al. [37] extracted low-level static and dynamic context features to improve the detection of mild traumatic brain injury lesions.

Recently, deep learning has become a hot topic in semantic segmentation which can automatically learn complex representations from data by itself [8], [38]. Convolutional Neural Networks (CNNs) [39] can be trained end-to-end from raw voxels to outputs. They achieve higher accuracy than many traditional learning methods that rely on hand-crafted features [40], [41]. However, they suffer from inefficiency for testing when a sliding patch is necessary to make a dense prediction for each voxel. FCNs [42] (replacing those fully connected layers with convolution layers) can make dense segmentation for the entire image by one feedforward propagation. For medical image segmentation, efficient FCNs such as U-net [7], [43], DCAN [44] and V-net [45] have been proposed and achieved outstanding performance in different tasks. Unfortunately, the great success of FCNs depends heavily on the quality and quantity of the training data, since massive and accurate voxel-wise annotations are essential for full supervision to train the models. Such voxel-wise annotations require expensive manual efforts, which are often not practically applicable. Moreover, the delineation may suffer large inter-operator variability. Thus, insufficient voxel-wise annotations become a bottleneck to build FCN-based solution to supervised segmentation.

B. Weakly Supervised Image Segmentation

To reduce the burden of acquiring voxel-wise annotations for training, some learning-based methods have explored different forms of weak annotations, such as image-level labels [46], points [47], [48], scribbles [49], or bounding boxes [11], [15], [50], to achieve voxel-wise segmentation.

In this work, we will build an automatic deep model for the CT segmentation of male pelvic organs, while only the 3D bounding box that compactly covers the organ is delineated for training. An important issue to build a robust voxel-wise segmentation model based on the weak annotations is to identify true voxel-wise labels in training. For example, STC [46]

utilized simple images downloaded from [Flickr.com](https://www.flickr.com/) to obtain single class supervision information and then achieved the segmentation of complex images. Boxsup [15] adopted a region proposal method to generate a set of segmentation candidates based on the bounding box annotations for updating the deep network. DeepCut [11] also did a pre-refinement, instead of using bounding boxes directly, to initialize the training of the model, and observed impressive accuracy improvement in final segmentation.

III. OUR METHOD

In this section, we present our proposed segmentation method that is trained from 3D bounding box annotations only. The whole pipeline is shown in Fig. 2, in which the label denoising network (LDnet) is iteratively trained to achieve the voxel-wise segmentation. Specially, to start training LDnet, a pre-refinement step is first used to extract initial training labels from the 3D bounding box annotations. Then, LDnet is trained (and further re-trained in subsequent iterations) under the supervision of the training voxels and their labels. Given a trained LDnet, we can predict the labels for the training voxels, from which a noise map is generated by the label denoising module. Then, those training voxels with unreliable labels are identified and suppressed from supervising the re-training of LDnet in the next iterations. The iteration training of LDnet loops until the predicted labels of the training voxels become stable. In the test stage, a test image will only need pass through the final LDnet to produce the segmentation result. Please note that the LDnet can follow arbitrary FCN-based segmentation architectures, such as U-net [7], DenseNet [51], DeepLabV3+ [52], and V-net [45]. Here, we adopt ResNet34 [53] as the backbone for U-net [7].

A. Label Update Using Iterative Learning

While 3D bounding box annotations just provide extremely coarse coverage for male pelvic organs with large shape variation, iterative refinement of the labels of the training voxels can identify the voxels that are trustful to supervise the training of the segmentation network. Iterative learning is widely used in weakly supervised segmentation methods, such as GrabCut [54], Boxsup [15], and DeepCut [11]. For male pelvic organs with blurred boundaries in CT images, we observe that iterative learning is effective. Specifically, the initial labels are used to supervise the training of the deep model in the first iteration. Then the labels of the training voxels are re-estimated by the tentatively trained model. The updated labels, in turn, supervise the re-training of the model in the next iteration.

B. Label Denoising Module

Assuming that a voxel-wise label mask has been generated by a model trained from 3D bounding box annotations, there must be high errors (or noisy labels) in the tentative label mask, especially near the blurred organ boundaries. Therefore, we design a label denoising module for the tentatively predicted label masks, such that the confidence of the voxel-wise labels can be assessed. Specially, we generate the noise map for the tentatively predicted label mask of each training image and use it for the next iteration of training of the segmentation network. The noise map reflects the stability of the predicted labels across consecutive training iterations, as we argue that a voxel label keeps consistent if the voxel is highly reliable to supervise the training.

Given a 2D slice, if the predicted labels of this slice vary significantly across consecutive iterations, the slice and its tentatively predicted labels are not stable, implying that the slice should be excluded from subsequent training. Otherwise, the segmentation network will possibly be misled by the noisy labels on this slice. Inspired by the above observation, we let L and L_p denote the predicted label masks of the 3D training image V at the current and previous training iterations, respectively. For the i -th slice V^i we utilize DSC as the metric to evaluate the stability of the predicted label masks. In this way, we generate the corresponding Noise as:

$$W^i = \begin{cases} O, & D(L^i, L_p^i) \geq \theta, \\ C^i \oplus S[K^i], & \text{otherwise}, \end{cases} \quad (1)$$

where O denotes a matrix of logical zeros, $D(\cdot, \cdot)$ is the DSC value between the predicted label masks of the slice under consideration over two consecutive iterations, θ ($= 0.95$ in our case) is an empirical threshold to determine whether the predicted labels are stable. C^i is the binary organ contour reconstructed from L^i , $S[\cdot]$ is a binary structuring element with morphological kernel size K^i , and $\oplus$ denotes morphological dilation. In Eq. 1, '1' indicates that the predicted label of the corresponding voxel in the current iteration is not stable while '0' indicates the high stableness.

Eq. 1 implies that, if the predicted mask of a slice is already stable, all voxels on this slice are regarded reliable for the next training iteration (i.e., '0' in the noise map W). On the contrary, the predicted labels on this slice are noisy especially near the potential boundaries (i.e., '1' in W). The range of the noisy region is controlled by the kernel size of K^i according to

$$K^i = \lfloor \lambda(1 - D(L^i, L_p^i))(1 + E(i, m)) \rfloor, \quad (2)$$

where i and m are the indices (after 0-to-1 normalization) of the i -th slice and the middle slice of the 3D bounding box, $E(i, m)$ denotes the distance from the i -th slice under consideration to the m -th slice, $\lfloor \cdot \rfloor$ is the round-down function, and λ is a hyper-parameter to control the noisy region.

To interpret the above, smaller DSC value implies greater instability and thus larger noisy region, i.e., $(1 - D(\cdot, \cdot)) \in [0, 1]$ in Eq. 2 (as indicated in Fig. 4(a)). Larger $E(\cdot, \cdot)$ value indicates the shorter distance to the hard-to-segment organ border, also corresponding to larger noisy region, i.e., $(1 + E(\cdot, \cdot)) \in [1, 2]$ in Eq. 2 (as indicated in Fig. 4(b)). In our implementation, K^i is guaranteed to be odd by multiplying by 2 and then adding 1 to balance the width of the noisy regions in the foreground (organ) and background (non-organ); λ is set to 5 based on the sizes of the organs (i.e., prostate, bladder, and rectum) in our dataset.

The noise map and the label mask are corresponding with each other. Before starting the next iteration, we first output the tentatively predicted label for each training voxel based on the current segmentation model, and then update the corresponding noise map using Eq. 1 and Eq. 2 based on the current predictions as well as the outputs of the previous iteration.

C. LDnet Architecture

The detailed architecture of our LDnet for generating the (tentative) label mask is shown in Fig. 3, where the prostate is used as an example of the target organ. Because there are overlaps between the 3D bounding box annotations of neighboring organs, we segment three target organs (i.e., prostate, bladder, and rectum) separately. To derive the final segmentation results for all three organs, we use the weighted max-voting method proposed in [5] to deal with the overlapping of the predictions.

The whole architecture of LDnet owns two sub-networks: the slice-level classification sub-network and the voxel-wise segmentation sub-network with denoising loss.

1) Slice-Level Classification for Localization Prior: Even though no voxel-wise labels are available in the beginning, the slice-level labels can still be easily inferred from 3D bounding box annotations. That is, the slice partially covered by a 3D bounding box is “with” the organ, or “without” vice-versa. Intuitively, the network trained for the slice-level classification task can generate class activation maps (CAMs), where high activation helps identify the target organ from the background [55]–[57]. To this end, we implement the slice-level classification sub-network to encode this localization prior, which facilitates the segmentation task. Specifically, we replace the expanding path of U-net [7] with two fully connected layers (the first layer with 1000 channels and the last layer with two channels) for slice-level classification. The feature maps of the last two convolutional layers are first converted to 2D matrices via the Global Average Pooling (GAP) operation. Then, the two matrices are fused by the element-wise sum operation after being up-sampled to the same size as the input image. The resulting matrix is scaled to [0,1] and serves as the localization prior heatmap for segmentation.

2) Voxel-Wise Segmentation with Denoising Loss: The segmentation sub-network performs voxel-wise segmentation by using ResNet34 [53] as the backbone for U-Net [7]. Three consecutive axial slices with their corresponding localization prior are input, to determine the label mask of the middle of the three input slices. Based on the predicted labels on the training slice and the corresponding noise map, we introduce a novel denoising loss to train the segmentation network. Denote the tentatively input label mask of a certain training image as L_n and its corresponding noise map as W_n (n for iteration index). The label denoising loss is defined as:

$$L_{DE} = - \frac{2 \sum_{ij} (1 - W_n^{ij}) L_n^{ij} \hat{L}^{ij}}{\sum_{ij} (1 - W_n^{ij}) (L_n^{ij})^2 + \sum_{ij} (1 - W_n^{ij}) (\hat{L}^{ij})^2}, \quad (3)$$

where L_n^{ij} denotes the predicted label of the j -th voxel in the i -th slice, W_n^{ij} is the corresponding value in the noise map, and $\hat{L}^{ij}$ is the predicted label of the current training iteration.

D Pre- and Post-Processing

While label denoising and LDnet are coupled, we adopt pre-refinement before the first training iteration, as well as post-refinement after every iteration.

1) Pre-Refinement: The quality of the training labels is the basis of training a good LDnet. Though iterative learning is used to refine the prediction, the final performance is still sensitive to the initialization not only in segmentation accuracy but also in convergence speed. Hence, we perform a necessary pre-refinement step for each organ using GrabCut [54] before the first iteration of LDnet. The pre-refinement results are treated as the supervision to train the model of the first iteration.

2) Post-Refinement: At the end of every iteration, we also design post-refinement to improve the quality of the LDnet outputs further. Specially, two cues directly derived from 3D bounding box annotations are used: (1) there are only background voxels outside the 3D bounding box as we require the box covers the entire organ; and (2) at least one voxel per slice in the 3D bounding box belongs to the target organ. Therefore, two strategies are used for post-refinement:

- All voxels outside the 3D bounding box are set to the background.
- If there is no organ voxel detected in a certain slice covered by the 3D bounding box, the prediction labels of this slice will roll back to the previous iteration.

Although the post-refinement may only correct the labels of a small amount of training data, it is helpful especially to the regions that are difficult to segment, such as the apex and base of the prostate. Note that pre- and post-refinement are performed only in the training phase. In the test phase, neither pre-/post-refinement nor 3D bounding box annotation is needed for the input image.

IV. EXPERIMENTAL SETUP

A. Dataset and Evaluation Metrics

Dataset—There are 313 planning CT images collected from the North Carolina Cancer Hospital for experiments. These images are scanned from different patients with prostate cancer. Oncologists have carefully delineated three pelvic organs (i.e., prostate, bladder, and rectum). Please note that voxel-wise annotations are not available in our training phase, but only used for performance evaluation. In our implementation, the only known annotations for training are the 3D bounding boxes of the organs, which are indicated by the coordinates of the cuboid border that fully enclose the target organs. Our training is based on the axial slices only, though there is no restriction upon the slice orientation in our method design. The in-plane resolutions of the images range from 0.932 *mm* to 1.365 *mm*, with the inter-slice thickness ranging from 1 *mm* to 3 *mm*. We have resampled all the images and their corresponding delineation to the same resolution with $1 \times 1 \times 1 \text{ mm}^3$ for better processing.

Metrics—Following [6], [30], we adopt two metrics for evaluation, i.e., DSC (%) and Average Symmetric Surface Distance (ASD, *mm*). And to better evaluate the model generalization capability, 5-fold cross-validation is performed for all methods. Specifically,

we randomly split the whole dataset into five folds. In each round, one fold is for testing and the remaining four folds are for training. Five rounds are performed to guarantee that each fold is used for testing in turn. We report the average performance over all five folds for comparison.

B Implementation Details

All networks are implemented using Keras [58] with TensorFlow [59] as the backend. In the training phase, the inputs of the network are the patches randomly cropped from three sequential slices (one slice before and one slice after the middle slice) along the axial plane with the size $128 \times 128 \times 3$, and the output is the corresponding label patch of the middle slice. In the test phase, the overlapped patches extracted from test images with a stride $16 \times 16 \times 1$ are input into the trained model. The overlapped predictions are averaged and then used as the final segmentation results. Compared with single-slice-based networks, our implementation employs pseudo-3D context information. Compared with 3D networks, our implementation costs fewer computing resources. For each patch, we perform min-max normalization and subtract the mean patch computed on the training data. With an initial learning rate of 10^{-3} , a reducing factor (10^{-1}) is multiplied after five epochs with no improvement in validation performance. The batch size is 12. The number of epochs for each iteration is 100, and iteration will be terminated while all predicted training labels become stable.

C Comparative Methods

To our best knowledge, our work is the first to implement voxel-wise segmentation from the training data with 3D bounding box annotations only. Hence, we compare our method with the 3D variants of state-of-the-art 2D approaches that can handle bounding box annotations.

- The first one is FCN_{BB} [11], which treats all voxels in 3D bounding boxes as being belonging to the target organ and is trained without using iterative learning. FCN_{BB} can thus be considered as the baseline to compare.
- The second one is FCN_{GC}, which is similar to FCN_{BB} but includes pre-refinement of GrabCut [54]. With refined initialization for training, FCN_{GC} should perform better than FCN_{BB}.
- The last one is FCN_{GCI}, which is the iterative version of FCN_{GC}. Specifically, FCN_{GCI} starts training by using the pre-refinement results of GrabCut [54]. The labels re-estimated by the tentatively trained model are processed through post-refinement, and then used to supervise the re-training of the model in the next iteration. Note that FCN_{GCI} adopts the initialization and the iterative training strategy that are similar to Boxsup [15] and DeepCut [11].
- Moreover, we train a full supervision network FCN_{FS} using high-quality voxel-wise annotations, which provides the ceiling performance for our method to compare.

In ablation studies, we modify our method and implement FCN_{DE} by removing the proposed classification sub-network to verify the effectiveness of the localization prior. To make a fair comparison, all methods adopt the same FCN architecture as the backbone. Generally,

FCN_{BB}, FCN_{GC}, FCN_{GCI}, FCN_{DE}, and ours are trained using the training dataset with 3D bounding box annotations while FCN_{FS} is trained using voxel-wise annotations. In the testing, the predicted results of each method on the test dataset are compared with the ground-truth voxel-wise annotations to evaluate the performance.

V. RESULTS

In Table I, we report the segmentation performance of all compared methods in DSC (%) and ASD (*mm*). In Fig. 5 and Fig. 6, examples of the segmentation results of individual methods are shown.

A. Compared with Existing Bounding Box Based Approaches

In this section, we compare our method with other approaches as mentioned in Sec. IV-C, including BB, FCN_{BB}, GC, FCN_{GC}, and FCN_{GCI}. From Table I, we can see that our method achieves the best performance with large margins in terms of both DSC (%) and ASD (*mm*) for the segmentation of all three organs. Specially, the proposed method has an increase of 5.11 (%), 8.72 (%), and 13.00 (%) in terms of DSC and a decrease of 0.76 (*mm*), 2.08 (*mm*), and 2.51 (*mm*) in terms of ASD over the best competing method (i.e., FCN_{GCI}), for the prostate, bladder, and rectum segmentation, respectively. In Fig. 5 and Fig. 6, we provide visualization of representative segmentation results for all methods.

Because the 3D bounding box can only provide a very rough delineation of the organ, if treating all voxels in the bounding box as being belonging to the organ, it will introduce massive noisy labels and mislead the training. Especially when the boundaries are blurred, the trained model may barely learn anything to segment the organ foreground from the non-organ background. The above argument can be verified by the poor results of FCN_{BB} (i.e., 58.90%, 60.59%, and 42.82% in DSC for the segmentation of the prostate, bladder, and rectum).

Benefiting from the pre-refinement using GrabCut [54], FCN_{GC} can learn from a better prepared training dataset with fewer noisy labels (in terms of the performance of GC). Hence, the DSCs of FCN_{GC} have increased by 12.53 (%), 13.19 (%), and 17.48 (%) compared with FCN_{BB} for prostate, bladder, and rectum, respectively, which are very remarkable improvements for the challenging pelvic organ segmentation of CT images. FCN_{GCI} is with further performance improvements benefiting from the iterative refinement of the training voxels. However, it is still hindered by the noisy labels, especially near the boundary. On the contrary, our proposed method effectively alleviates this problem and achieves better performance.

B. Compared with High-quality Voxel-wise Annotation Based Approach

Using 3D bounding box annotations instead of high-quality voxel-wise annotations can greatly reduce the cost in terms of both finance and human effort. However, the effort would not be useful in real applications if the loss in segmentation accuracy was too high. Therefore, we verify the segmentation performance of our method by comparing it with the fully supervised approach FCN_{fs}, which is trained using high-quality voxel-wise annotations and delivers the ceiling accuracy. As indicated in Table I, our method reaches ~94%, ~91%,

and ~86% of the DSC performance achieved by FCN_{fs} for the segmentation of the prostate, bladder, and rectum, respectively. Moreover, while manually delineating the whole organ boundary in the CT image of each patient takes about 10 - 12 minutes [14], marking a 3D bounding box for each organ takes only about 15 seconds. That is, our proposed method can achieve competitive performance compared with the fully supervised approach and also make data collection efficient and simple.

C. Importance of Pre- and Post-Refinement

To improve the quality of labels for training, a pre-refinement step and a post-refinement step are designed. The comparisons between the results of FCN_{BB} and FCN_{GC} in Sec. V-A have verified the importance of the pre-refinement. In Table II, we report the DSC increase on the training data after performing post-refinement over training iterations of three methods (FCN_{GCI}, FCN_{DE}, and our proposed method) using iterative learning for three organs. We can see that, after post-refinement, the quality of training labels is unproved. Even though some improvements are relatively slight, it is still helpful to enhance the model capability of dealing with hard samples or outliers, which are especially abundant around the challenging organ boundary.

D. Ablation Studies

There are mainly three techniques for ablation studies, namely label denoising module, slice-level classification sub-network, and iterative learning. To verify their respective effectiveness, we compare the segmentation performance before and after removing one of these modules. All reported results in this section are shown in Table III.

With VS. Without Label Denoising—To verify the effectiveness of label denoising, we compare the results of FCN_{DE} with FCN_{GCI}, while the only difference between them is whether the noise maps are used in the training phase. As shown in Table III, there are 3.90 (%) DSC increase for the prostate, 8.05 (%) for the bladder, and 11.32 (%) for the rectum in the last iteration of FCN_{DE} with respect to FCN_{GCI} when introducing label denoising. We can also observe that both DSC and ASD of FCN_{DE} are better than those of FCN_{GCI} in each iteration, which indicates the effectiveness of our proposed label denoising.

With VS. Without Slice Classification—To maximize the value of simple 3D bounding box annotations, we add the slice classification sub-network in our method to exploit organ localization prior for better performance. To verify its effectiveness, we compare the results of our method with its simplified version FCN_{DE} without the slice classification subnetwork. From Table III, we can see that the slice classification sub-network can play a positive role in almost all organs, metrics, and iterations.

With VS. Without Iterative Learning—Iterative training is an effective mechanism to capture the organ shape gradually. Boxsup [15] and DeepCut [11] that achieve significant performance in early works all benefit from iterative learning. To verify its effectiveness in our case, we compare the DSC and ASD values of three methods (FCN_{GCI}, FCN_{DE}, and our proposed method) using iterative learning. As shown in Table III, DSC and ASD are

iteratively unproved for all methods and organs, except for ASD of FCN_{GCI} in rectum segmentation.

In Fig. 7, we show several typical segmentation examples for the prostate, bladder, and rectum over iterations. We can see that the predicted boundaries are refined over iterations. And with the help of label denoising and slice classification, our method can achieve improved segmentation results compared with the ground truth.

We also show the influence of the number of iterations in Fig. 8. For the segmentation of all three organs, the DSC values are not very sensitive to the number of iterations when it reaches three. That is mainly because the predicted training labels have stabilized and the unidentified noisy labels hinder further performance improvements.

VI. DISCUSSION AND CONCLUSION

To our best knowledge, this is the first work to implement voxel-wise segmentation from training data with 3D bounding box annotations only. The proposed method can effectively harness 3D bounding box annotations to train FCNs for the challenging male pelvic organ segmentation in CT images. The experiments presented in this paper confirm the necessity of identifying noisy labels in weakly supervised segmentation of low-contrast CT images and provide new insight on how to address this problem. Our results are superior over previous methods, and particularly reach ~94%, ~91%, and ~86% of the DSCs achieved when training from high-quality voxel-wise annotations for the prostate, bladder, and rectum, respectively. Note that the proposed method is user-friendly in acquiring training annotations, and simultaneously greatly reduces the manual effort.

Even though our method is with great potential in real applications, there is still a challenging problem that needs to be solved for better promotion. That is, our method may not work perfectly to accurately segment some slices near the organ border due to: 1) the slices near the organ border themselves are more difficult to segment compared with the middle of the organs; 2) these parts are more sensitive to noisy labels, hard to converge, making the network easy to overfit in iterative learning. In future work, we will try to address this problem by combining our method with a small amount of user interaction. And we expect to achieve similar performance to fully supervised approaches using voxel-wise annotations while minimizing the cost of data collection as much as possible.

Acknowledgments

This work was supported in part by NIH grant CA206100.

REFERENCES

- [1]. A. C. Society, "Cancer facts & figures 2019," Atlanta: American Cancer Society, 2019.
- [2]. Lu C et al. "An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy," *Medical Image Analysis*, vol. 15, no. 5, pp. 772–785, 2011. [PubMed: 21646038]

- [3]. Mahapatra D and Buhmann JM, "Prostate mri segmentation using learned semantic knowledge and graph cuts," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 756–764, 2013. [PubMed: 24235297]
- [4]. Chang Y et al. "Accurate pelvis and femur segmentation in hip ct with a novel patch-based refinement," *IEEE Journal of Biomedical and Health Informatics*, 2018.
- [5]. He K et al. "Pelvic organ segmentation using distinctive curve guided fully convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 585–595, 2019. [PubMed: 30176583]
- [6]. Wang S et al. "Ct male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation," *Medical Image Analysis*, 2019.
- [7]. Ronneberger O et al. "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [8]. Shen D et al. "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.
- [9]. Wang G et al. "Deepigeos: a deep interactive geodesic framework for medical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [10]. Wang S et al. "Computer aided endoscope diagnosis via weakly labeled data mining," in *2015 IEEE International Conference on Image Processing*. IEEE, 2015, pp. 3072–3076.
- [11]. Rajchl M et al. "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 674–683, 2016. [PubMed: 27845654]
- [12]. Wang S et al. "Computer-aided endoscopic diagnosis without human-specific labeling," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 11, pp. 2347–2358, 2016. [PubMed: 26890528]
- [13]. Jia Z et al. "Constrained deep weak supervision for histopathology image segmentation," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2376–2388, 2017. [PubMed: 28692971]
- [14]. Gao Y et al. "Learning distance transform for boundary detection and deformable segmentation in ct prostate images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2014, pp. 93–100.
- [15]. Dai J et al. "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.
- [16]. Khoreva A et al. "Simple does it: Weakly supervised instance and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 876–885.
- [17]. Miao Q et al. "Rboost: label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2216–2228, 2015. [PubMed: 26415187]
- [18]. Sabzevari M et al. "Vote-boosting ensembles," *Pattern Recognition*, vol. 83, pp. 119–133, 2018.
- [19]. Freund Y et al. "Experiments with a new boosting algorithm," in *Proceedings of the International Conference on Machine Learning*, vol. 96. Citeseer, 1996, pp. 148–156.
- [20]. Friedman JH, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [21]. Kakar P and Chia AY-S, "If you can't beat them, join them: Learning with noisy data," in *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, pp. 571–580.
- [22]. Huang W et al. "Dynamic boosting in deep learning using reconstruction error," in *International Joint Conference on Neural Networks*. IEEE, 2014, pp. 473–480.
- [23]. Ghosh A et al. "Robust loss functions under label noise for deep neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [24]. Veit A et al. "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 839–847.

- [25]. Han B et al. "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018, pp. 8527–8537.
- [26]. Ren M et al. "Learning to reweight examples for robust deep learning," *arXiv preprint arXiv:1803.09050*, 2018.
- [27]. Feng Q et al. "Segmenting ct prostate images using population and patient-specific statistics for radiotherapy," *Medical Physics*, vol. 37, no. 8, pp. 4121–4132, 2010. [PubMed: 20879572]
- [28]. Chen L et al. "Drinet for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2453–2462, 2018. [PubMed: 29993738]
- [29]. Gu Z et al. "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2019.
- [30]. Gao Y et al. "Accurate segmentation of ct male pelvic organs via regression-based deformable models and multi-task random forests," *IEEE Transactions on Medical Imaging*, vol. 35, no. 6, pp. 1532–1543, 2016. [PubMed: 26800531]
- [31]. Mahapatra D, "Analyzing training information from random forests for improved image segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1504–1512, 2014. [PubMed: 24569439]
- [32]. Zhan Y and Shen D, "Automated segmentation of 3d us prostate images using statistical texture-based matching method," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2003, pp. 688–696.
- [33]. Ricci E and Perfetti R, "Retinal blood vessel segmentation using line operators and support vector classification," *IEEE Transactions on Medical Imaging*, vol. 26, no. 10, pp. 1357–1365, 2007. [PubMed: 17948726]
- [34]. Van Ravesteijn VF et al. "Computer-aided detection of polyps in ct colonography using logistic regression," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 120–131, 2009. [PubMed: 19666332]
- [35]. Kallel IK et al. "An iterative possibilistic knowledge diffusion approach for blind medical image segmentation," *Pattern Recognition*, vol. 78, pp. 182–197, 2018.
- [36]. Shao Y et al. "Locally-constrained boundary regression for segmentation of prostate and rectum in the planning ct images," *Medical Image Analysis*, vol. 26, no. 1, pp. 345–356, 2015. [PubMed: 26439938]
- [37]. Bianchi A et al. "Dynamic low-level context for the detection of mild traumatic brain injury," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 145–153, 2014. [PubMed: 25073162]
- [38]. Litjens G et al. "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [PubMed: 28778026]
- [39]. Karpathy A et al. "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
- [40]. Wachinger C et al. "Deepnat: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, 2018. [PubMed: 28223187]
- [41]. Jia H et al. "Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging," *Neurocomputing*, vol. 275, pp. 1358–1369, 2018.
- [42]. Long J et al. "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [43]. Çiçek Ö et al. "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.
- [44]. Chen H et al. "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2487–2496.
- [45]. Milletari F et al. "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.

- [46]. Wei Y et al. "Stc: A simple to complex framework for weakly-supervised semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2314–2320, 2016. [PubMed: 28114002]
- [47]. Maninis K-K et al. "Deep extreme cut: From extreme points to object segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 616–625.
- [48]. Kervadec H et al. "Constrained-cnn losses for weakly supervised segmentation," Medical Image Analysis, vol. 54, pp. 88–99, 2019. [PubMed: 30851541]
- [49]. Tang M et al. "Normalized cut loss for weakly-supervised cnn segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1818–1827.
- [50]. Yang H et al. "Exploit bounding box annotations for multi-label object recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 280–288.
- [51]. Jégou S et al. "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–19.
- [52]. Chen L-C et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.
- [53]. He K et al. "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [54]. Rother C et al. "Grabcut: Interactive foreground extraction using iterated graph cuts," in ACM Transactions on Graphics, vol. 23, no. 3. ACM, 2004, pp. 309–314.
- [55]. Saleh F et al. "Built-in foreground/background prior for weakly-supervised semantic segmentation," in Proceedings of European Conference on Computer Vision. Springer, 2016, pp. 413–432.
- [56]. Diba A et al. "Weakly supervised cascaded convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 914–922.
- [57]. Wei Y et al. "Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection," in Proceedings of the European Conference on Computer Vision, 2018, pp. 434–450.
- [58]. Chollet F et al. "Keras," 2015.
- [59]. Abadi M et al. "Tensorflow: A system for large-scale machine learning," in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

Fig. 1:

We propose a novel method to recover voxel-wise segmentation for CT male pelvic organs from 3D bounding box annotations. Our method reaches ~94% (prostate, in red), ~91% (bladder, in green), and ~86% (rectum, in blue) of the Dice Similarity Coefficients (DSCs), using the fully voxel-wise annotations for training.

Fig. 2:

The pipeline of our proposed method learning from 3D bounding box annotations.

Fig. 3:
The detailed architecture of our label denoising network (LDnet). Only prostate segmentation is used as an example in this figure.

Fig. 4:

Two factors in determining the value of K : (a) greater instability ($1 - D(\cdot, \cdot)$) of the predicted labels across consecutive training iterations corresponding to larger K ; (b) shorter distance ($E(\cdot, \cdot)$) to the hard-to-segment organ border corresponding to larger K .

Fig. 5:

Typical results for male pelvic organ segmentation of all compared methods. Note that due to overlaps between bounding boxes of neighboring organs, the organs segmented by poorly performing FCN_{BB} may be severely occluded with each other in 3D visualization. (Red: prostate; green: bladder; blue: rectum)

Fig. 6:

Comparisons between the ground truth (red) and the automatic segmentation results of all compared methods (yellow). The numerical value in the blue block indicates the corresponding DSC value.

Fig. 7:
Representative segmentation examples over iterations in case of the prostate, bladder, and rectum, respectively. The numerical value in the blue block indicates the corresponding DSC.

Fig. 8:
Mean DSCs (%) of the proposed method for the prostate, bladder, and rectum segmentation with respect to the number of iterations.

TABLE I:

Segmentation accuracy for male pelvic organs in terms of DSC (%) and ASD (*mm*).

Metrics		FCN _{BB}	FCN _{GC}	FCN _{GCI}	FCN _{DE}	Proposed	FCN _{FS}
Prostate	DSC	58.90 ± 8.56	71.43 ± 7.58	75.20 ± 8.36	79.10 ± 6.90	80.31 ± 5.83	85.82 ± 7.07
	ASD	6.35 ± 1.95	3.83 ± 1.61	3.41 ± 1.72	2.74 ± 1.12	2.65 ± 0.90	1.79 ± 0.99
Bladder	DSC	60.59 ± 7.61	73.78 ± 7.96	76.75 ± 7.49	84.80 ± 7.74	85.47 ± 7.88	93.83 ± 4.97
	ASD	9.69 ± 1.97	5.70 ± 1.60	4.99 ± 1.46	2.79 ± 1.09	2.91 ± 1.13	1.26 ± 1.19
Rectum	DSC	42.82 ± 8.41	60.30 ± 8.75	61.06 ± 8.77	72.38 ± 7.05	74.06 ± 6.11	86.16 ± 5.18
	ASD	10.56 ± 1.89	5.84 ± 1.69	5.82 ± 1.59	3.84 ± 1.34	3.31 ± 0.88	1.98 ± 1.01

TABLE II:

DSC increase (%) on the training dataset after performing post-refinement over iterations.

Metrics		Iter. 1	Iter. 2	Iter. 3
FCN _{GCI}	Prostate	1.13	0.99	0.62
	Bladder	0.43	1.23	1.10
	Rectum	0.94	0.99	1.11
FCN _{DE}	Prostate	1.92	0.69	0.35
	Bladder	0.48	1.05	0.17
	Rectum	0.31	0.56	0.10
Proposed	Prostate	2.88	1.08	0.26
	Bladder	0.17	0.17	0.13
	Rectum	0.24	1.33	1.19

Accuracy improvement in terms of DSC (%) and ASD (*mm*) over iterations of FCN_{GCI}, FCN_{DE}, and our proposed method. The best result in each iteration is indicated in **bold**.

Table III:

Metrics		Prostate			Bladder			Rectum		
		Iter. 1	Iter. 2	Iter. 3	Iter. 1	Iter. 2	Iter. 3	Iter. 1	Iter. 2	Iter. 3
FCN _{GCI}	DSC	71.43	73.28	75.20	73.78	75.80	76.75	60.30	60.53	61.06
	ASD	3.83	3.66	3.41	5.70	5.20	4.99	5.84	5.77	5.82
FCN _{DE}	DSC	76.92	78.64	79.10	82.95	84.11	84.80	69.87	71.11	72.38
	ASD	2.98	2.81	2.74	3.25	2.95	2.79	4.01	3.85	3.84
Proposed	DSC	77.20	79.33	80.31	83.89	84.26	85.47	69.33	73.06	74.06
	ASD	2.91	2.85	2.65	3.00	2.89	2.74	4.15	3.50	3.31