Semi-Supervised Segmentation of Radiation-Induced Pulmonary Fibrosis from Lung CT Scans with Multi-Scale Guided Dense Attention

Guotai Wang,

School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

Shuwei Zhai,

School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

Giovanni Lasio,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Baoshe Zhang,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Byong Yi,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Shifeng Chen,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Thomas J. Macvittie,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Dimitris Metaxas,

Department of Computer Science, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA.

Jinghao Zhou,

Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore MD 21201, USA.

Shaoting Zhang

School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

Corresponding authors: Guotai Wang; Shaoting Zhang, guotai.wang@uestc.edu.cn; zhangshaoting@uestc.edu.cn.

Abstract

Computed Tomography (CT) plays an important role in monitoring radiation-induced Pulmonary Fibrosis (PF), where accurate segmentation of the PF lesions is highly desired for diagnosis and treatment follow-up. However, the task is challenged by ambiguous boundary, irregular shape, various position and size of the lesions, as well as the difficulty in acquiring a large set of annotated volumetric images for training. To overcome these problems, we propose a novel convolutional neural network called PF-Net and incorporate it into a semi-supervised learning framework based on Iterative Confidence-based Refinement And Weighting of pseudo Labels (I-CRAWL). Our PF-Net combines 2D and 3D convolutions to deal with CT volumes with large inter-slice spacing, and uses multi-scale guided dense attention to segment complex PF lesions. For semi-supervised learning, our I-CRAWL employs pixel-level uncertainty-based confidence-aware refinement to improve the accuracy of pseudo labels of unannotated images, and uses image-level uncertainty for confidence-based image weighting to suppress low-quality pseudo labels in an iterative training process. Extensive experiments with CT scans of Rhesus Macaques with radiation-induced PF showed that: 1) PF-Net achieved higher segmentation accuracy than existing 2D, 3D and 2.5D neural networks, and 2) I-CRAWL outperformed state-of-the-art semi-supervised learning methods for the PF lesion segmentation task. Our method has a potential to improve the diagnosis of PF and clinical assessment of side effects of radiotherapy for lung cancers.

Keywords

I. INTRODUCTION

LUNG cancer is the leading cause of cancer deaths around the world. In 2018, the global number of new cases and deaths were 2.09 million and 1.76 million, respectively [1]. Radiation therapy is one of the most effective and widely used treatment for cancer patients [2]. With the development of such treatment technologies, lung cancer death rate dropped 45% from 1990 to 2015 among men, and 19% from 2002 to 2015 among women [3]. However, about half of all cancer patients who receive radiation therapy during their course of illness will suffer from Radiation-Induced Injuries (RII) to the hematopoietic tissue, skin, lung and gastrointestinal (GI) systems [2]. To prevent, mitigate or treat the RII plays an important role in improving the quality of radiation therapies.

For lung cancer, the most common RII is the radiation-induced Pulmonary Fibrosis (PF) [4], i.e., inflammation and subsequent scarring of lung tissues caused by radiation, which could lead to breathing problems due to lung damage and even lung failure [5]. Observation and assessment of the PF lesions using Computed Tomography (CT) imaging is critical for diagnosis and treatment follow-up of this disease [6]. For an accurate and quantitative measurement of PF, it is desirable to segment the PF lesions from 3D CT scans. The segmentation results can provide detailed spatial distribution and accurate volumetric measurement of the lesions, which is important for treatment decision making, PF progress modeling, treatment effect assessment and prognosis prediction.

As manual segmentation of lesions from 3D images is time-consuming, labor-intensive and faced with inter- and intra-observer varabilities, automatic PF lesion segmentation from CT images is highly desirable [7]. However, this is challenging due to several reasons. Firstly, with different severity of the disease, PF lesions have a large variation of size and shape. A small lesion may only contain few pixels, while a large lesion can occupy a lung segment. The irregular shapes make it difficult to use a statistical model for the segmentation task [8]. Secondly, the lesions have a complex spatial distribution, and can be scattered in different segments of the lung. Thirdly, PF lesions often adhere to lung structures including vessels, airways and the pleura, and other lesions like lung nodules and pneumonia lesions with similar appearance may also exist. These factors, alongside with the low contrast of soft tissues in CT images, make it hard to delineate the boundary of PF lesions. Fig. 1 shows two examples of lung CT scans with PF, and it demonstrates the difficulties of accurate segmentation.

Recently, Convolutional Neural Networks (CNNs) have been increasingly used for automatic medical image segmentation [9]. By automatically learning features from a large set of annotated images, they have outperformed most traditional segmentation methods using hand-crafted features [7], such as for recognition of lung nodules [10], [11], lung lobes [12] and COVID-19 infection lesions [13], [14]. However, to the best of our knowledge, CNNs for PF lesion segmentation have rarely been investigated so far.

Besides the above challenges, existing CNNs may obtain suboptimal performance for the PF lesion segmentation task due to the following reasons. First, lung CT images usually have anisotropic 3D resolutions with high inter-slice spacing and low intra-slice spacing. Existing CNNs using pure 2D convolutions or pure 3D convolutions have limited ability to learn effective 3D features from such images, as 2D CNNs [8], [15]–[17] can only learn intra-slice features, and most 3D networks [18]–[21] are designed with isotropic receptive field in terms of voxels. When dealing with 3D images with large inter-slice spacing, they have an imbalanced physical receptive field (in terms of mm) along each axis, i.e., the physical receptive field in the through-plane direction is much larger than that in in-plane directions, which may limit effective learning of 3D features. Second, existing CNNs often use position invariant convolutions without spatial awareness, which makes it difficult to handle objects with various position and size. Attention mechanisms have recently been proposed to improve the spatial awareness [17], [22], [23], but their obtained spatial attention does not match the target region well, and their performance on PF lesions has not been investigated.

What's more, current success of deep learning methods for segmentation relies highly on a large set of annotated images for training [9]. For 3D medical images, acquiring pixel-level annotations in segmentation tasks is extremely time-consuming and difficult, as accurate annotations could be only provided by experts with domain knowledge [24]. For the PF lesion segmentation task, annotation of a CT volume could take several hours, and the complex shape and appearance of PF lesions further increase the efforts and time needed for annotation, which makes it difficult to annotate a large set of 3D pulmonary CT scans for training.

To deal with these problems, we propose a novel semi-supervised framework with a novel 2.5D CNN based on multi-scale attention for the segmentation of PF lesions from CT scans with large inter-slice spacing. The contribution is three-fold. First, we propose a novel network for PF lesion segmentation (i.e., PF-Net), which employs multi-scale guided dense attention to deal with lesions with various size and position, and combines 2D and 3D convolutions to achieve balanced physical receptive field along different axes to better learn 3D features from medical images with anisotropic resolution. Second, a novel semi-supervised learning framework using Iterative Confidence-based Refinement And Weighting of pseudo Labels (I-CRAWL) is proposed, where uncertainty estimation is employed to assess the quality of pseudo labels of unannotated images in both pixel level and image level. We propose Confidence-Aware Refinement (CAR) based on pixel-level uncertainty to refine pseudo labels, and introduce confidence-based image weighting according to image-level uncertainty to suppress low-quality pseudo labels. Thirdly, we apply our proposed method to radiation-induced PF lesion segmentation from CT scans, and extensive experimental results show that our method outperformed state-of-the-art semi-supervised methods and existing 2D, 3D and 2.5D CNNs for segmentation. As far as we know, this is the first work on PF lesion segmentation based on deep learning, and our method has a potential to reduce the annotation burden for large-scale 3D image datasets in the development of automatic segmentation models with high performance.

II. RELATED WORKS

A. CNNs for Medical Image Segmentation

CNNs have achieved state-of-the-art performance for many medical image segmentation tasks [9]. Most widely used segmentation CNNs are inspired by U-Net [15], which is based on an encoder-decoder structure to learn features at multiple scales. UNet++ [16] extends U-Net with a series of nested, dense skip pathways for a higher performance. Attention U-Net [22] introduced an attention gate using high-level features to calibrate low-level features. Spatial and channel "Squeeze and Excitation" (scSE) [17] enables a 2D network to focus on the most relevant features for better performance.

Typical networks for segmentation of 3D volumes include 3D U-Net [18], V-Net [19] and HighRes3DNet [20]. They assume that the input volume has an isotropic 3D resolution to learn 3D features, and are not suitable for images with large inter-slice spacing. To better deal with such images, a 3D anisotropic hybrid network that uses a pre-trained 2D encoder and a decoder with anisotropic convolutions was proposed in [25]. Jia et al. [26] designed a pyramid anisotropic CNN based on decomposition of 3D convolutions. The nnU-Net [27] automatically configures network structures and training strategies, where different types of convolution kernels can be adaptively combined for a given dataset. In [28], a lightweight CNN combining inter-slice and intra-slice convolutions was proposed to for segmentation of CT images. In [29], 2D and 3D convolutions were combined in a single network for Vestibular Schwannoma segmentation from images with anisotropic resolution. However, these networks have limited ability to segment lesions with various size and position.

B. Segmentation of Lung CT Images

CNNs have been widely used for segmentation of lung structures from CT images [7]. In [12], cascaded CNNs with non-local modules [23] were proposed to leverage structured relationships for pulmonary lobe segmentation. In [30], a CNN was combined with freeze-and-grow propagation for airway segmentation. For lung lesions, a central focused CNN [11] was proposed to segment lung nodules from heterogeneous CT images, and Fan et al. [13] employed reverse attention and edge attention to segment COVID-19 lung infection. Wang et al. [14] developed a noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions by learning from non-expert annotations. Despite the large amount of works on lung structure and lesion segmentation so far, there is a lack of deep learning models for the challenging task of radiation-induced pulmonary fibrosis segmentation.

C. Semi-Supervised Learning

To reduce the burden for annotation, semi-supervised learning methods have been increasingly employed for medical image segmentation by using a limited number of annotated images and a large amount of unannotated images [24]. Existing semi-supervised methods mainly have two categories. The first category is based on pseudo labels [13], [31], [32], where a model trained with annotated images obtains pseudo labels for unannotaed images that are then used to update the segmentation model. Lee [31] used such a strategy for classification problems, and Bai et al. [32] updated the pseudo segmentation labels and network parameters alternatively and used Conditional Random Field (CRF) to refine the pseudo labels. Fan et al. [13] progressively enlarged the training set with unlabeled data and their pseudo labels for learning. However, this method ignores the quality of pseudo labels, which may limit the performance of the learned model.

The second category is to learn from annotated and unannotated images simultaneously, and they often consist of a supervised loss function for annotated images and an unsupervised regularization loss function for all the images. The regularization can be based on teacher-student consistency [33], [34], transformation consistency [35], multi-view consistency [36] and reconstruction-based auxiliary task [37]. Adversarial learning [38] also regularizes the segmentation model by minimizing the distribution difference between segmentation results of annotated images and those of unannotaed images. However, adversarial models are hard to train, and the complex size and shape of PF lesions make it difficult to capture the true distribution of lesion masks when only a small set of annotated images are available.

III. METHOD

In this section, we first introduce our proposed Pulmonary Fibrosis segmentation Network (PF-Net) with multi-scale guided dense attention, and then describe how it is used in our I-CRAWL framework for semi-supervised learning.

A. PF-Net: A 2.5D Network with Multi-Scale Guided Dense Attention

Our proposed PF-Net is illustrated in Fig. 2. It employs an encoder-decoder backbone structure that is commonly used and effective for medical image segmentation [15], [16],

[19]. The encoder contains five scales, and each is implemented by a convolutional block followed by a max-pooling layer for down-sampling. In each block, we use two convolutional layers each followed by a Batch Normalization (BN) [39] layer and a parametric Rectified Linear Unit (pReLU), and a dropout layer is inserted before the second convolutional layer. The decoder uses the same type of convolutional blocks as the encoder. We extend this backbone in the following aspects:

1) 2.5D Network Structure: To deal with the anisotropic 3D resolution with high inter-slice spacing and low intra-slice spacing, we combine 2D (i.e., intra-slice) and 3D convolutions so that the network has an approximately balanced physical receptive field along each axis. Let S represent the number of scales in the network ($S = 5$ in this paper), we use 2D convolutions and 2D max-poolings for the first M scales in the encoder, and employ 3D convolutions and 3D max-poolings for the last $S - M$ scales in the encoder. Each resolution level in the decoder contains the same type of 2D or 3D convolutional blocks as in the encoder. We use trilinear interpolation for upsampling in the decoder.

As our lung CT images have a resolution around $0.3 \times 0.3 \times 1.25$ mm, i.e., the in-plane resolution is about four times of the through-plane resolution, we set $M = 2$ so that the 2D max-pooling layers in the first two scales make the resulted feature maps have a near-isotropic 3D resolution, as shown in Fig. 2. Note that the first and S-th scales have the highest and lowest spatial resolution, respectively. We use F_s^e and F_s^d to denote the feature maps obtained by the last convolutional block at scale s in the encoder and decoder, respectively. Note that $F_S^e = F_S^d$ in the bottleneck block.

2) Multi-Scale Guided Dense Attention: To better deal with PF lesions with various position and size, we use multi-scale attention to improve the network's spatial awareness, and propose dense attention to leverage multi-scale contextual information for the segmentation task. Specifically, at each scale s of the decoder, we use a convolutional layer to get a spatial attention map $\tilde{P}_s$, and use $\tilde{P}_s$ as a high-level attention signal to guide the learning in all lower scales of the decoder. The input of the decoder at scale s is a concatenation of three parts: F_s^e from scale s of the encoder, an upsampled version of the decoder feature map F_{s+1}^d and $\tilde{P}_{s+1}^s \oplus \tilde{P}_{s+2}^s \oplus \ldots \oplus \tilde{P}_S^s$, where $\oplus$ is the concatenation operation and $\tilde{P}_{s+1}^s$ is the upsampled version of $\tilde{P}_{s+1}$ so that it has the same spatial resolution as F_s^e. Therefore, a lower scale accepts the attention maps from all the higher scales as input, which is referred to as Multi-Scale Dense Attention (MSDA). The decoder thus takes advantage of multi-scale contextual information that enables the network to pay more attention to the target region.

To better learn the spatial attention, we propose Multi-Scale Guided Attention (MSGA) to explicitly supervise $\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_S$ at different scales. Let P_s denote the softmaxed output of $\tilde{P}_s$ and Y denote the ground truth (i.e., one-hot probability map) of a training sample X. Unlike a common deep supervison strategy that upsamples $\tilde{P}_s$ or P_s at different scales to the same spatial resolution as Y [40]–[42], we down-sample Y to obtain multi-scale ground

truth for spatial attention, which makes the loss calculation more efficient and can directly supervise the spatial attention maps at different scales. Let Y_s denote the down-sampled ground truth at scale s. The multi-scale loss function for a single image is:

$$\ell\,(\mathscr{P}, Y) = \sum_{s=1}^{S} \alpha_s L_s(P_s, Y_s) \tag{1}$$

where $\mathscr{P} = \{P_1, P_2, ..., P_S\}$. $L_s()$ is a base loss function for image segmentation, such as the Dice loss [19]. α_s is the weight of $L_s()$ at scale s.

B. Semi-Supervised Learning using I-CRAWL

Generating pseudo labels for unannotated images has been shown effective for semi-supervised segmentation [13], [31], [32]. However, the pseudo labels often contain some incorrect regions, and low-quality pseudo labels can largely limit the performance of the learned segmentation model. To prevent the learning process from being corrupted by inaccurate pseudo labels, we propose an Iterative Confidence-based Refinement And Weighting of pseudo Labels (I-CRAWL) framework for semi-supervised segmentation.

Assume that the entire training set consists of one subset $\mathscr{D}^L$ with N^l labeled images and another subset $\mathscr{D}^U$ with N^U unlabeled images. For an image $X_i^L \in \mathscr{D}^L$, its ground truth label Y_i^L is known, while for an image $X_i^U \in \mathscr{D}^U$, its ground truth label is not provided, and we use Y_i^U to denote its estimated pseudo label. As the pseudo labels may have a large range of quality, we introduce an image-level weight $w_i^U \in [0, 1]$ for each pair of X_i^U, Y_i^U for learning, and define w_i^U based on the confidence (or uncertainty, inversely) of Y_i^U. The weight for a labeled image X_i^L from $\mathscr{D}^L$ can be similarly denoted as w_i^L, and we set $w_i^L = 1$ as the corresponding label Y_i^L is clean and reliable. Therefore, the labeled subset can be denoted as $\mathscr{D}^L = \left\{ \left(X_i^L, Y_i^L, w_i^L \right), ... \right\}$ $(i = 1, 2, ..., N^L)$, and the unlabeled subset with pseudo labels can be denoted as $\mathscr{D}^U = \left\{ \left(X_i^U, Y_i^U, w_i^U \right), ... \right\}$ $(i = 1, 2, ..., N^U)$.

Our I-CRAWL is illustrated in Fig. 3, and it is an iterative learning process with K rounds. Each round has four steps: 1) inference for unannotated images with uncertainty estimation, 2) confidence-aware refinement of pseudo labels, 3) confidence-based image weighting, and 4) network update where the current pseudo labels and image-level weights are used to train the network. These steps are detailed as follows.

1) Inference for Unannotated Images with Uncertainty Estimation: With the network parameter θ_{k-1} obtained in the last round, in round k, we first use θ_{k-1} to predict provisional pseudo labels for unannotated images in $\mathscr{D}^U$ and the associated uncertainty estimation. Note that in the first round (i.e., $k = 1$), the initial network parameter θ_0 is obtained by pre-training with the annotated images in $\mathscr{D}^L$.

With θ_{k-1}, we employ the Monte Carlo (MC) Dropout [43] that has shown to be an effective method for estimation of epistemic uncertainty caused by the lack of training data [34], [36]. MC Dropout feeds an image X_i^U into the network R times with random dropout, which leads to R predictions, i.e., R foreground probability maps [43]. The average of these R foreground probability maps is taken as the provisional probability map P_i^U, a binarization of which leads to a provisional pseudo segmentation label Y_i^U.

At the same time, the statistical variance of the R foreground probability maps is taken as the uncertainty map V_i^U, which gives voxel-level uncertainty. As uncertainty information can indicate potentially wrong segmentation results [34], [44], we treat the pseudo labels with high uncertainty (i.e., low confidence) values as unreliable labels for images in $\mathscr{D}^U$, and propose a Confidence-Aware Refinement (CAR) method to improve the pseudo labels' quality.

2) Confidence-Aware Refinement of Pseudo Labels: Given the provisional pseudo label Y_i^U with the uncertainty map V_i^U for an unannotated image $X_i^U \in \mathscr{D}^U$, and let x denote a voxel, we split the voxels in Y_i^U into three sets according to the status: high-confidence foreground voxels $\mathbb{F} = \left\{ \mathbf{x} \mid Y_{ix}^U = 1, V_{ix}^U \leq \epsilon \right\}$, high-confidence background voxels $\mathbb{B} = \left\{ \mathbf{x} \mid Y_{ix}^U = 0, V_{ix}^U \leq \epsilon \right\}$ and undetermined voxels $\mathbb{U} = \left\{ \mathbf{x} \mid V_{ix}^U > \epsilon \right\}$, where $\epsilon = 0.05$ is a small threshold value. For undetermined voxels in $\mathbb{U}$, we refine their labels according to a contextual regularization considering intervoxel connections and softened probabilities for these voxels.

Our CAR for pseudo label refinement has two steps: probability map softening and contextual regularization. First, we soften the foreground probabilities for uncertain voxels, which will degrade the influence of the network's prediction for these voxels in the following contextual regularization step. Let p denote a foreground probability value and u denote the corresponding uncertainty value, the softening function is:

$$g(p, u) = 0.5 + (p - 0.5) \times (1 - u) \tag{2}$$

where the softened foreground probability get closer to 0.5 when u is larger. Let $\dot{P}_{ix}^U$ denote the softened foreground probability of voxel $\mathbf{x}$, and is obtained by:

$$\dot{P}_{i\mathbf{x}}^U = \begin{cases} P_{i\mathbf{x}}^U & \text{if } \mathbf{x} \in \mathbb{F} \cup \mathbb{B} \\ g\left(P_{i\mathbf{x}}^U, V_{i\mathbf{x}}^U \right) & \text{if } \mathbf{x} \in \mathbb{U} \end{cases} \tag{3}$$

Then, we use contextual regularization taking $\dot{P}_i^U$ as input to refine the pseudo labels, which is implemented by a fully connected Conditional Random Field (CRF) [45]. For simplicity, we denote X_{ix}^U, Y_{ix}^U and $\dot{P}_{ix}^U$ as $x_{\mathbf{x}}$, $y_{\mathbf{x}}$ and $\dot{p}_{\mathbf{x}}$, respectively. The energy function of CRF is:

$$E\left(Y_i^U\right) = \sum_{\mathbf{x}} \phi(y_{\mathbf{x}}) + \sum_{\mathbf{x}, \mathbf{y}} \psi\left(y_{\mathbf{x}}, y_{\mathbf{y}}\right) \tag{4}$$

where $\phi(y_{\mathbf{x}}) = -y_{\mathbf{x}}\log(\dot{p}_{\mathbf{x}}) - (1 - y_{\mathbf{x}})(1 - \dot{p}_{\mathbf{x}})$ constrains the output to be consistent with the softened probability map, and this constraint for the low-confidence voxels in $\mathbb{U}$ is weak. The second term in Eq. (4) is a pairwise potential that encourages the label's contextual consistency:

$$\psi\left(y_{\mathbf{x}}, y_{\mathbf{y}}\right) = \mu\left(y_{\mathbf{x}}, y_{\mathbf{y}}\right)\left[w_1\exp(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma_\alpha^2} - \frac{\|x_{\mathbf{x}} - x_{\mathbf{y}}\|^2}{2\sigma_\beta^2}) \right. \\ \left. + w_2\exp(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma_\gamma^2})\right] \tag{5}$$

where $\mu(y_{\mathbf{x}}, y_{\mathbf{y}}) = 1$ if $y_{\mathbf{x}} \neq y_{\mathbf{y}}$ and 0 otherwise. Minimization of Eq. (4) leads to a refined pseudo label Y_i^U for X_i^U.

3) Confidence-based Image Weighting of Pseudo Labels: With the new pseudo label Y_i^U, we further employ the confidence to update its image-level weight w_i^U to suppress low-quality pseudo labels at the image level. We first define an image-level uncertainty $\mathbf{v}_i$ based on uncertainty map V_i^U:

$$\mathbf{v}_i = \frac{\sum_x V_{ix}^U}{\sum_x Y_{ix}^U + \eta} \tag{6}$$

where $\mathbf{v}_i$ is the sum of voxel-level uncertainty normalized by the segmented lesion's volume in the image. $\eta = 10^{-5}$ is a small number for numerical stability. Let v_{max} and v_{min} denote the maximal and minimal values of $\mathbf{v}_i$ among all the unannotated images, we map $\mathbf{v}_i$ to the range of [0, 1]:

$$\mathbf{v}_i' = \frac{\mathbf{v}_i - \mathbf{v}_{min}}{\mathbf{v}_{max} - \mathbf{v}_{min}} \tag{7}$$

Finally, the image-level weight w_i^U is defined as:

$$w_i^U = \left(1.0 - \mathbf{v}_i'\right)^{1/\gamma} \tag{8}$$

where $\gamma \geq 1.0$ is a hyper parameter to control the non-linear mapping between the image-level uncertainty v_i' and the weight. We do not use $\gamma < 1.0$ as it will lead the weight of most samples to be very small (close to 0.0). In contrast, $\gamma > 1.0$ leads the weight for most samples close to 1.0, and only samples with a high uncertainty will be strongly suppressed. In the experiment, we set $\gamma = 3.0$ according to the best performance on the validation set.

4) Model Update with Batch Training: With the refined pseudo labels and image-level weights obtained above, we train the network based on $\mathscr{D}^L$ and the current pseudo labels for images in $\mathscr{D}^U$, where each image is weighted by w_i^L or w_i^U in the segmentation loss function. The weighted loss for the entire training set is:

$$\mathscr{L} = \sum_{i=1}^{N^L} w_i^L \, \ell \left(\mathscr{P}_i^L, Y_i^L \right) + \sum_{i=1}^{N^U} w_i^U \, \ell \left(\mathscr{P}_i^U, Y_i^U \right) \tag{9}$$

where $\ell()$ is defined in Eq. (1). $\mathscr{P}_i^L$ and $\mathscr{P}_i^U$ are the multi-scale predictions obtained by PF-Net for an annotated image and an unannotated image, respectively.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setting

1) Data: Thoracic CT scans of 41 Male Rhesus Macaques with radiation-induced lung damage were collected with ethical committee approval. Once irradiated, each individual underwent serial CT scans for assessment of PF around every 30 days in 3 to 8 months. 133 CT scans with PF were used for experiments of the segmentation task. The CT scans have a slice thickness of 1.25 mm, with image size 512×512 and pixel size ranging from 0.20 mm $\times$ 0.20 mm to 0.38 mm $\times$ 0.38 mm. We randomly split the dataset at individual level into 86 scans from 27 individuals for training, 15 scans from 4 individuals for validation, and 32 scans from 10 individuals for testing. Manual annotations given by an experienced radiologist were used as the segmentation ground truth. In the training set, we used 18 scans with annotations as $\mathscr{D}^L$ and the other 68 scans as unannotated images $\mathscr{D}^U$ for semi-supervised learning, and also investigated the performance of our method with different ratios of annotated images. For preprocessing, we crop the lung region and normalize the intensity to [0, 1] using a window/level of 1500/−650.

2) Implementation and Evaluation Metrics: Our PF-Net[1] and I-CRAWL framework were implemented in Pytorch with PyMIC[2] [14] library on a Ubuntu desktop with an NVIDIA GTX 1080 Ti GPU. The channel number parameter N in PF-Net was set as 16. Dropout was only used in the encoder. The dropout rate at the first two scales of the encoder was 0 due to the small number of feature channels, and that for the last three scales was 0.3, 0.4 and 0.5, respectively. We set the base loss function $L_s()$ as Dice loss [19]. PF-Net was trained with Adam optimizer, weight decay of 10^{-5}, patch size of $48 \times 192 \times 192$ and batch size of 2.

The round number for our I-CRAWL was $K = 3$, and we use round 0 to refer to the pre-training with annotated images. In each of the following round, the pseudo labels of unannotated images acquired by CAR were kept fixed, and we used Adam optimizer to train the network for tens of thousands of iterations until the performance on validation

---

[1] https://github.com/HiLab-git/PF-Net
[2] https://github.com/HiLab-git/PyMIC

set stopped to increase. The learning rate was initialized as 10^{-3} and halved every 10k iterations. Uncertainty estimation was based on 10 forward passes of MC dropout. We used the SimpleCRF library[3] to implement fully connected CRF [45]. Following [46], image intensity was rescaled from [0, 1] to [0, 255] before the image was sent into the CRF, and the CRF parameters were: $w_1 = 3$, $w_2 = 10$, $\sigma_a = 10$, $\sigma_\beta = 20$ and $\sigma_\gamma = 15$, which was tuned based on the validation set. γ in Eq. (8) was 3.0, and the performance based on different γ values is shown in Table III.

For quantitative evaluation of the segmentation, we used Dice score, Relative Volume Error (RVE) and the 95-th percentile of Hausdorff Distance (HD_{95}) between the segmented PF lesions and the ground truth in 3D volumes. Paired t-test was used to see if two methods were significantly different.

B. Performance of PF-Net

In this section, we investigate the performance of our PF-Net for pulmonary fibrosis segmentation only using the 18 annotated images for training. The results of semi-supervised learning will be demonstrated in Section IV-C.

1) Comparison with Existing Networks: Our PF-Net was compared with three different categories of network structures: 1) 2D CNNs including the typical 2D U-Net [15] and more advanced attention U-Net [22] that leverages spatial attention to focus more on the segmentation target; 2) 3D CNNs including 3D U-Net [18], 3D V-Net [19] and scSE-Net [17] that combines a 3D U-Net [18] backbone with spatial and channel "squeeze and excitation" modules; and 3) existing networks designed for dealing with volumetric images with large inter-slice spacing (i.e., anisotropic resolution): nnU-Net [27] that automatically configures the network structure so that it is adapted to the given dataset, AH-Net [25] that transfers features learned from 2D images to 3D anisotropic volumes, and VS-Net [29] that uses a mixture of 2D and 3D convolutions with spatial attention. Quantitative evaluation results of these methods are shown in Table I. We found that 3D U-Net [18] achieved a Dice score of 57.19%, which was the lowest among the compared methods. 2D U-Net [15], Attention U-Net [22], 3D V-Net [19] and 3D scSE-Net [17] achieved similar performance with Dice score around 60%. The nnU-Net [27], AH-Net [25] and VS-Net [29] designed to deal with anisotropic resolution generally performed better than these 2D and 3D networks. Among these existing methods, VS-Net achieved the best Dice that was 67.45%. Our PF-Net achieved Dice, RVE and HD_{95} of 70.36%, 27.96% and 10.87 mm, respectively, where the Dice and HD_{95} were significantly better than those of the other compared methods.

Fig. 4 shows a qualitative comparison between these networks, where (a) and (b) are from two individuals, and axial and coronal views are shown in each case. In Fig. 4(a), it can be seen that 2D U-Net [15], 3D V-Net [19] and AH-Net [25] lead to obvious under-segmentation, as highlighted by blue arrows in the first row. VS-Net [29] performs better than them, but the result of our PF-Net is closer to the ground truth than that of

---

[3] https://github.com/HiLab-git/SimpleCRF

VS-Net. From the coronal view of Fig. 4(b), we can observe that the result of 2D U-Net [15] lacks inter-slice consistency. 3D V-Net [19] achieves better inter-slice consistency, but it has a poor segmentation in the upper and lower regions of the lesion, as indicated by the blue arrow in the third column. AH-Net [25], VS-Net [29] and our PF-Net that consider anisotropic resolution obtain better performance than the above networks purely using 2D or 3D convolutions. What's more, Fig. 4 shows that the lesions have complex and irregular sizes and shapes, and the proposed PF-Net is able to achieve more accurate segmentation in these cases than AH-Net [25] and VS-Net [29].

2) Ablation Study: To investigate the effectiveness of each component of our PF-Net, we set the baseline as a naive 2.5D U-Net that extends 2D U-Net [15] by using pReLU, replacing 2D convolutions with 3D convolutions at the three lowest resolution levels, and adding dropout layers to each convolutional block. To justify the choice of using 2D convolutions at the first two resolution levels and 3D convolutions at the other resolution levels, we set M to 0–5 respectively. Note that $M = 0$ and $M = 5$ correspond to pure 2D and pure 3D networks, respectively. Comparison between these variants are listed in the first section of Table II, which shows that the performance increases as M changes from 0 to 2, and decreases when M is 3 and larger. This is in line with our motivation to set M to 2 due to the fact that the in-plane resolution is around four times of the through-plane resolution. Thus, we use baseline ($M = 2$) in the following ablation study.

The proposed PF-Net is referred to as baseline + MSDA + MSGA, where MSDA is our proposed multi-scale dense attention and MSGA is the proposed multi-scale guided attention. We compared the baseline and our PF-Net with: 1) Baseline + non-local [23], where the non-local is a self-attention block inserted at the bottleneck (scale 5) of the baseline network, and it was not used at the lower scales with higher resolution due to memory constraint; 2) Baseline + deep supervision, where the baseline network was only combined with a typical deep supervision strategy as implemented in [40], [41]; 3) Baseline + MSGA, without using MSDA; 4) Baseline + MSGA + MSDA°, where MSDA° is a variant of MSDA and it refers to $\tilde{P}_s$ is only sent to its next lower scale $s - 1$, rather than all the lower scales in the decoder of PF-Net.

Quantitative evaluation results of these variants are listed in the second section of Table II. It shows that compared with the baseline ($M = 2$), deep supervision improved the Dice score from 67.87% to 68.57%, but our MSGA was more effective with Dice of 69.32%. Using MSDA° could further improve the performance, but it was less effective than our MSDA. Table II demonstrates that our PF-Net was better than the compared variants, and it significantly outperformed the baseline in terms of Dice and HD_{95}. Fig. 6 presents a visualization of multi-scale attention maps of PF-Net. It shows that attention maps across scales are consistent with each other and they change from coarse to fine as the spatial resolution increases.

C. Results of Semi-supervised Learning

With PF-Net as the segmentation network structure, we further validate our proposed I-CRAWL for semi-supervised training. In Section IV-C1 and IV-C2 we used 18 annotated

and 68 unannounced images, i.e., the annotation ratio was around 20%. In Section IV-C3, we experimented with different annotation ratios including 10%, 20% and 50%.

1) Hyper-Parameter Setting: First, to investigate the best value of hyper parameter γ in Eq. (8) that controls the weight of unannotated images, we measured the model's performance on the validation set at the first round of I-CRAWL with γ ranging from 1.0 to 5.0. They are compared with "Initial" that refers to model pre-trained on the annotated images and "No weighting" that denotes treating all unannotated images equally without considering the quality of pseudo labels. Quantitative measurements shown in Table III demonstrate that the best γ value was 3.0, and its corresponding Dice score was 68.75%, which was better than 66.62% obtained by "Initial" and 67.83% obtained by "No weighting". Therefore, we set $\gamma = 3.0$ in the following experiments.

2) Uncertainty and Confidence-Aware Refinement: Fig. 5 shows a visual comparison between a standard fully connected CRF [45] and our proposed CAR that leverages confidence (uncertainty) for refinement of pseudo labels. The first row of each subfigure shows an unannotated image and the pseudo label with uncertainty obtained by the CNN, and the second row shows the updated pseudo labels. In Fig. 5(a), the initial pseudo label has a large under-segmentated region, which is associated with high values in the uncertainty map, i.e., low confidence. The standard CRF only fixed the pseudo label moderately. In contrast, with the help of confidence, our CAR largely improved the pseudo label's quality by recovering the under-segmented region. In Fig. 5(b), the initial pseudo label has some over-segmentation in airways, and the uncertainty map indicates the potentially wrong segmentation in the corresponding regions well. We can observe that CAR outperformed CRF in removing the over-segmented airways.

We also compared MC Dropout [43] with two other uncertainty estimation methods: entropy minimization and Bayesian network [47]. A visual comparison of them is shown in Fig. 7. We found that in spite of longer time required than the other methods, MC Dropout could generate more calibrated uncertainty estimation. As shown in Fig. 7, the uncertain region obtained by entropy minimization and Bayesian network [47] are mostly located around the border of the segmentation output, and uncertainty map obtained by MC dropout can better indicate under- and over-segmentation regions, which is highlighted by red and yellow arrows in Fig. 7. For quantitative comparison, we applied CAR as a post processing method to the validation set in the first round of I-CRAWL, where these uncertainty estimation methods were used respectively. Results in Table IV show that using MC dropout for CAR improved the prediction accuracy from 66.62% to 69.05% in terms of Dice, which outperformed using the other two uncertainty estimation methods.

Fig. 8 shows a visualization of pseudo labels and uncertainty as the round increases. It can be observed that the initial pseudo label at round 0 has a large under-segmented region with high uncertainty. The pseudo label becomes more accurate and confident as the round increases.

3) Ablation Study of I-CRAWL: For ablation study of I-CRAWL, we set the PF-Net trained only with the annotated images as a baseline, and it was compared with: 1) IT

that refers to naive iterative training, where in each round pseudo label of an unannotated image is reset to the prediction given by the network without refinement, 2) IT + CRF that uses standard fully connected CRF [45] to refine pseudo labels, 3) IT + CAR denoting that our confidence-aware refinement is used to update pseudo labels in each round, and 4) our I-CRAWL (IT + CAR + IW) where IW denotes our confidence-based image weighting of pseudo labels.

The performance of these methods at different rounds are shown in Fig. 9. Note that round 0 is the baseline, and all the methods based on iterative training performed better than the baseline. However, the improvement obtained by only using IT is slight. Using CRF or CAR to refine the pseudo labels at different rounds achieved a large improvement of Dice, and our CAR considering the voxel-level confidence of predictions outperformed the naive CRF. Weighting of pseudo labels based on image-level confidence helped to obtain more accurate result, and our I-CRAWL outperformed the other variants. Fig. 9 also shows that the improvement from round 0 to round 1 of I-CRAWL is large, but the model's performance does not change much at round 2 and 3.

Table V shows quantitative comparison between the baseline and variants of I-CRAWL at the end of training (round 3). It can be observed that IT's performance was not far from the baseline, with an average Dice of 70.87% compared with 70.36%. Using CRF and CAR improved the average Dice to 72.13% and 72.71%, respectively, showing the superiority of CAR compared with CRF. I-CRAWL improved the average Dice to 73.04%, with HD_{95} value of 7.92 mm in average, which was better than the other variants.

4) Comparison with Existing Methods: I-CRAWL was compared with several sate-of-the-art semi-supervised methods for medical image segmentation: 1) Fan et al. [13] that uses a randomly selected propagation strategy for semi-supervised COVID-19 lung infection segmentation, 2) Bai et al. [32] that uses CRF to refine pseudo labels in an iterative training framework, corresponding to "IT + CRF" described previously, 3) Cui et al. [33] that is an adapted mean teacher method, and 4) UA-MT [34] that is uncertainty-aware mean teacher. For all these methods, we used our PF-Net as the backbone network.

We investigated the performance of these methods with different ratios of labeled data: 10%, 20% and 50%. For each setting, the baseline was learning only from the labeled images, and the upper bound was "full supervision" where 100% training images were labeled. The results are shown in Table VI. It can be observed that with only 10% images annotated, the baseline performed poorly with an average Dice of 54.73%. Our I-CRAWL improved it to 67.18%, which largely outperformed the other methods. When 20% images were annotated, our method improved the average Dice from 70.36% to 73.04% compared with the baseline, while the mean teacher-based methods did not bring much performance gain. When 50% images were annotated, our method also outperformed the others, and it was comparable with full supervision (74.02±13.77% versus 74.58±10.99% in terms of Dice, with p-value > 0.05).

V. DISCUSSION AND CONCLUSION

Our PF-Net combines 2D and 3D convolutions to deal with anisotropic resolutions, and we set $M = 2$ as the in-plane resolution is four times of the through-plane resolution in our dataset. It may be set to other values according to the spacing information of different datasets. Multi-scale guided dense attention in PF-Net is important for dealing with PF lesions with various positions, shapes and scales. We noticed that Sinha et al. [48] also proposed a multi-scale attention, but it has key differences from ours. First, Sinha et al. [48] concatenated feature at different levels of the encoder to obtain a multi-scale feature, which is used as input for parallel attention modules at different scales. While PF-Net learns multi-scale attentions sequentially, where attention at a lower resolution level is used as input for all the higher resolution levels with dense connections. Second, Sinha et al. [48] used self-attention inspired by non-local block [23] that is computationally expensive with large memory consumption, while PF-Net uses convolution to obtain the attention coefficients at different spatial positions, which has higher memory and computational efficiency. In addition, for each attention module, the attention maps are calculated in two steps in [48], and a consistency between the two steps is imposed via an L2 distance of their encoded representations, which is called guided attention by the authors. In contrast, guided attention in PF-Net refers to supervising attention maps directly by the resampled segmentation ground truth.

Our I-CRAWL is a pseudo label-based method for semi-supervised learning. Despite that pseudo label has been previously investigated [13], [32], I-CRAWL is superior to these works mainly for the following reasons. First, the pseudo labels generated by a model trained with a small set of annotated images inevitably contain a lot of inaccurate predictions. Improving the quality of pseudo labels would benefit the final segmentation model. However, the method in [13] does not refine pseudo labels, and Bai et al. [32] refines pseudo labels by CRF without considering their confidence. Our method employs uncertainty estimation to find uncertain regions that are likely to be mis-segmented, and the confidence-aware refinement is more effective to refine these mis-segmented regions, leading to improved accuracy of pseudo labels. Second, the quality of pseudo labels of different images varies a lot, and it is important to exclude low-quality pseudo labels that may corrupt the segmentation model. However, Bai et al. [32] and [13] ignored this point and treated all the pseudo labels equally. In contrast, I-CRAWL uses image-level uncertainty information to highlight more confident pseudo labels and down-weight uncertain ones that are unreliable. Thus, the model is less affected by low-quality pseudo labels.

Uncertainty estimation plays an important role in our I-CRAWL framework. We found that the simple yet effective MC dropout performed better than alternatives including entropy minimization and Bayesian networks [47]. Despite that MC Dropout is slow for uncertainty estimation, it is used offline at the beginning of each round of our method and takes a short time compared with the batch training step. In the scenario of 20% annotated data, our uncertainty estimation takes 8.72 minutes (7.69s per 3D image), and CAR and image weighting take 12.98 minutes (11.45s per 3D image). In contrast, the model update with batch training takes around 4 hours for each round, i.e., the first three steps of I-CRAWL account for 8.29% of the entire runtime of each round, which could be further accelerated

by multi-thread parallel computing. Therefore, our uncertainty estimation and CAR require little extra time compared with naive iterative training and Bai et al. [32].

In conclusion, we present a novel 2.5D network structure and an uncertainty-based semi-supervised learning method for automatic segmentation of pulmonary fibrosis from anisotropic CT scans. To deal with complex PF lesions with irregular structures and appearance in CT volumes with anisotropic 3D resolution, we propose PF-Net that combines a 2.5D network baseline with multi-scale guided dense attention. To leverage unannotated images for learning, we propose I-CRAWL that is an iterative training framework, where a confidence-aware refinement process is introduced to update pseudo labels and a confidence-based image weighting is proposed to suppress images with low-quality pseudo labels. Experimental results with lung CT scans from Rhesus Macaques showed that our PF-Net outperformed existing 2D, 3D and 2.5D networks for PF lesion segmentation, and our I-CRAWL could better leverage unannotated images for training than state-of-the-art semi-supervised methods. Our methods can be extended to deal with other structures and human CT scans in the future.

Acknowledgments

REFERENCES

[1]. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, and Jemal A, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 0, pp. 1–31, 2018.

[2]. Baskar R, Lee KA, Yeo R, and Yeoh KW, "Cancer and radiation therapy: Current advances and future directions," International Journal of Medical Sciences, vol. 9, no. 3, pp. 193–199, 2012. [PubMed: 22408567]

[3]. Siegel RL, Miller KD, and Jemal A, "Cancer statistics, 2018," CA: A Cancer Journal for Clinicians, vol. 68, no. 1, pp. 7–30, 2018. [PubMed: 29313949]

[4]. Wilson MS and Wynn TA, "Pulmonary fibrosis: Pathogenesis, etiology and regulation," Mucosal Immunology, vol. 2, no. 2, pp. 103–121, 2009. [PubMed: 19129758]

[5]. Giuranno L, Ient J, De Ruysscher D, and Vooijs MA, "Radiation-induced lung injury (RILI)," Frontiers in Oncology, vol. 9, p. 877, 2019. [PubMed: 31555602]

[6]. Christe A, Peters AA, Drakopoulos D, Heverhagen JT, Geiser T, Stathopoulou T, Christodoulidis S, Anthimopoulos M, Mougiakakou SG, and Ebner L, "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images," Investigative Radiology, vol. 54, no. 20, pp. 627–632, 2019. [PubMed: 31483764]

[7]. Van Rikxoort EM and Van Ginneken B, "Automated segmentation of pulmonary structures in thoracic computed tomography scans: A review," Physics in Medicine and Biology, vol. 58, no. 17, p. R187, 2013. [PubMed: 23956328]

[8]. Wang G, Zhang S, Xie H, Metaxas DN, and Gu L, "A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning," Medical Image Analysis, vol. 19, no. 1, pp. 176–186, 2015. [PubMed: 25461336]

[9]. Shen D, Wu G, and Suk H-I, "Deep learning in medical image analysis," Annual Review of Biomedical Engineering, vol. 19, no. 1, pp. 221–248, 2017.

[10]. Xie Y, Xia Y, Zhang J, Song Y, Feng D, Fulham M, and Cai W, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT," IEEE Transactions on Medical Imaging, vol. 38, no. 4, pp. 991–1004, 2019. [PubMed: 30334786]

[11]. Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, Dong D, Gevaert O, and Tian J, "Central focused convolutional neural networks : Developing a data-driven model for lung nodule segmentation," Medical Image Analysis, vol. 40, pp. 172–183, 2017. [PubMed: 28688283]

[12]. Xie W, Jacobs C, Charbonnier J-P, and van Ginneken B, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2664–2675, 2020. [PubMed: 32730216]

[13]. Fan D-P, Zhou T, Ji G-P, Zhou Y, Chen G, Fu H, Shen J, and Shao L, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Scans," IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2626–2637, 2020. [PubMed: 32730213]

[14]. Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, Meng T, Li K, Huang N, and Zhang S, "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2653–2663, 2020. [PubMed: 32730215]

[15]. Ronneberger O, Fischer P, and Brox T, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[16]. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, and Liang J, "Unet++: A nested u-net architecture for medical image segmentation," in MICCAI workshop on DLMIA, vol. 11045, 2018, pp. 3–11.

[17]. Roy AG, Navab N, and Wachinger C, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," IEEE Transactions on Medical Imaging, vol. 38, no. 2, pp. 540–549, 2019. [PubMed: 30716024]

[18]. Abdulkadir A, Lienkamp SS, Brox T, and Ronneberger O, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 424–432.

[19]. Milletari F, Navab N, and Ahmadi S-A, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in International Conference on 3D Vision, 2016, pp. 565–571.

[20]. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, and Vercauteren T, "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, 2017, pp. 348–360.

[21]. Huang C, Han H, Yao Q, Zhu S, and Zhou SK, "3D U2-Net: A 3D universal U-Net for multi-domain medical image segmentation," in Medical Image Computing and Computer-Assisted Intervention, vol. 2, 2019, pp. 291–299.

[22]. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, Mcdonagh S, Hammerla NY, Kainz B, Glocker B, and Rueckert D, "Attention U-Net: Learning where to look for the pancreas," in MIDL, 2018, pp. 1–10.

[23]. Wang X, Girshick R, Gupta A, and He K, "Non-local Neural Networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[24]. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang J, Wu Z, and Ding X, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," Medical Image Analysis, vol. 63, no. 2018, p. 101693, 2019.

[25]. Liu S, Xu D, Zhou SK, Pauly O, Grbic S, Mertelmeier T, Wicklein J, Jerebko A, Cai W, and Comaniciu D, "3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes," in International Conference on Medical Image Computing and Computer-Assisted Intervention, no. 3, 2018, pp. 851–858.

[26]. Jia H, Xia Y, Song Y, Zhang D, Huang H, Zhang Y, and Cai W, "3D APA-Net: 3D Adversarial Pyramid Anisotropic Convolutional Network for Prostate Segmentation in MR Images," IEEE Transactions on Medical Imaging, vol. 39, no. 2, pp. 447–457, 2020. [PubMed: 31295109]

[27]. Isensee F, Jaeger PF, Kohl SA, Petersen J, and Maier-Hein KH, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021. [PubMed: 33288961]

[28]. Lei W, Wang H, Gu R, Zhang S, Zhang S, and Wang G, "DeepIGeoS-V2: Deep interactive segmentation of multiple organs from head and neck images with lightweight CNNs," in LABELS workshop of MICCAI, 2019, pp. 61–69.

[29]. Wang G, Shapey J, Li W, Dorent R, Demitriadis A, Bisdas S, Paddick I, Bradford R, Ourselin S, and Vercauteren T, "Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss," in International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 2, 2019, pp. 264–272.

[30]. Nadeem SA, Hoffman EA, Sieren JC, Comellas AP, Bhatt SP, Barjaktarevic IZ, Abtin F, and Saha PK, "A CT-based automated algorithm for airway segmentation using freeze-and-grow propagation and deep learning," IEEE Transactions on Medical Imaging, vol. 40, no. 1, pp. 405–418, 2021. [PubMed: 33021934]

[31]. Lee D.-h., "Pseudo-Label : The simple and efficient semi-supervised learning method for deep neural networks," in ICML 2013 Workshop : Challenges in Representation Learning (WREPL), 2013, pp. 1–6.

[32]. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, and Rueckert D, "Semi-supervised learning for network-based cardiac MR image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 10434, 2017, pp. 253–260.

[33]. Cui W, Liu Y, Li Y, Guo M, Li Y, Li X, Wang T, Zeng X, and Ye C, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in International Conference on Information Processing in Medical Imaging, 2019, pp. 554–565.

[34]. Yu L, Wang S, Li X, Fu CW, and Heng PA, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 605–613.

[35]. Li X, Yu L, Chen H, Fu C-W, and Heng P-A, "Transformation consistent self-ensembling model for semi-supervised medical image segmentation," IEEE Transactions on Neural Networks and Learning Systems, pp. 18–21, 2020.

[36]. Xia Y, Yang D, Yu Z, Liu F, Cai J, Yu L, Zhu Z, Xu D, Yuille A, and Roth H, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," Medical Image Analysis, vol. 65, p. 101766, 2020. [PubMed: 32623276]

[37]. Chen S, Bortsova G, García-Uceda Juárez A, van Tulder G, and de Bruijne M, "Multi-task attention-based semi-supervised learning for medical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 457–465.

[38]. Nie D, Gao Y, Wang L, and Shen D, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 370–378.

[39]. Ioffe S and Szegedy C, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning, 2015, pp. 448–456.

[40]. Dou Q, Chen H, Jin Y, Yu L, Qin J, and Heng P-A, "3D deeply supervised network for automatic liver segmentation from CT volumes," in International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 1, 2016, pp. 149–157.

[41]. Zhang Y and Chung AC, "Deep supervision with additional labels for retinal vessel segmentation task," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 83–91.

[42]. Wang G, Li W, Ourselin S, and Vercauteren T, "Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation," Frontiers in Computational Neuroscience, vol. 13, no. August, pp. 1–13, 2019. [PubMed: 30809141]

[43]. Gal Y and Ghahramani Z, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in International Conference on Machine Learning, 2016, pp. 1050–1059.

[44]. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, and Vercauteren T, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," Neurocomputing, vol. 338, pp. 34–45, 2019.

[45]. Krähenbühl P and Koltun V, "Efficient inference in fully connected CRFs with gaussian edge potentials," in Neural Information Processing Systems, 2011, pp. 109–117.

[46]. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, and Glocker B, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Medical Image Analysis, vol. 36, pp. 61–78, 2017. [PubMed: 27865153]

[47]. Jena R and Awate SP, "A bayesian neural net to segment images with uncertainty estimates and good calibration," in International Conference on Information Processing in Medical Imaging, 2019, pp. 3–15.

[48]. Sinha A and Dolz J, "Multi-Scale Self-Guided Attention for Medical Image Segmentation," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 1, pp. 121–130, 2021. [PubMed: 32305947]
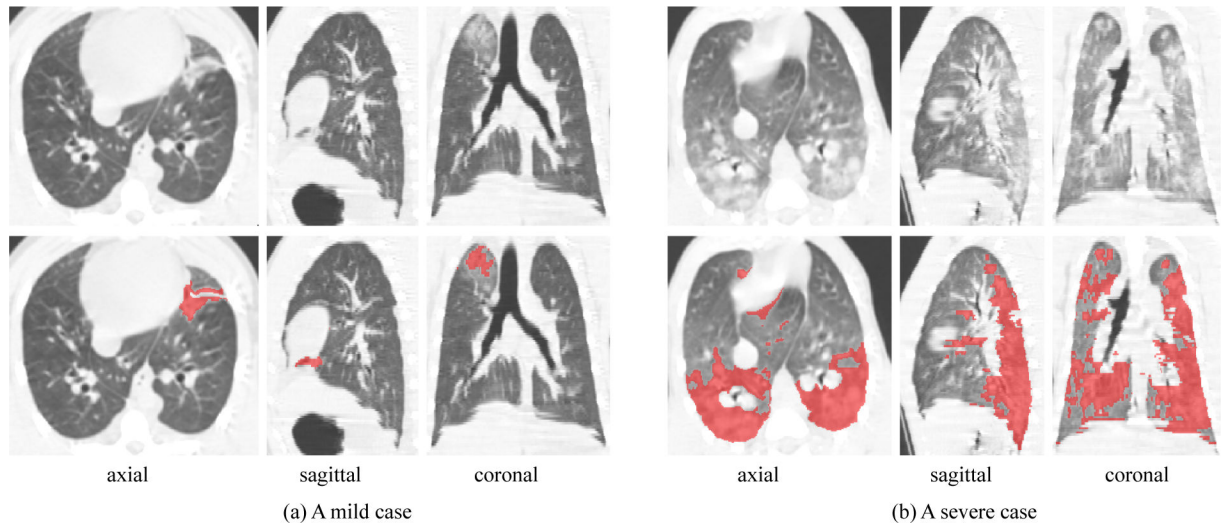
axial   sagittal   coronal   axial   sagittal   coronal

(a) A mild case   (b) A severe case

Fig. 1:

Examples of radiation-induced Pulmonary Fibrosis (PF) of the Rhesus Macaque. The first row shows lung CT images, and the second row shows manual segmentation results of PF lesions. Note that the ambiguous boundary, irregular shape and various size and position make the segmentation task challenging.
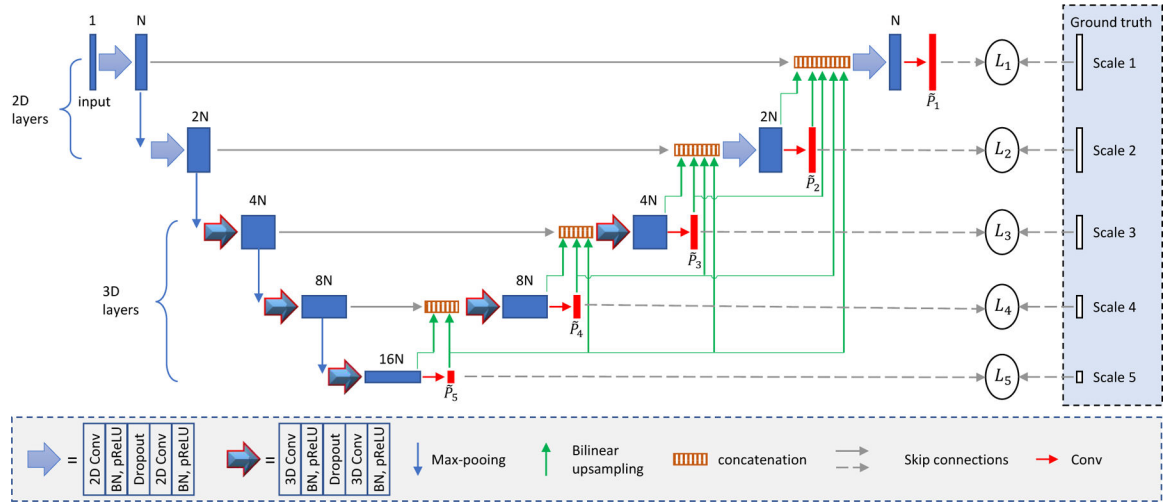
Fig. 2:

Structure of our proposed PF-Net. To deal with 3D images with large inter-slice spacing, the first two scales use 2D convolutions while the other scales use 3D convolutions. $\widetilde{P}_s$ is the predicted spatial attention at scale s, and it is sent to all lower scales with dense connection in the decoder, as shown by green lines.
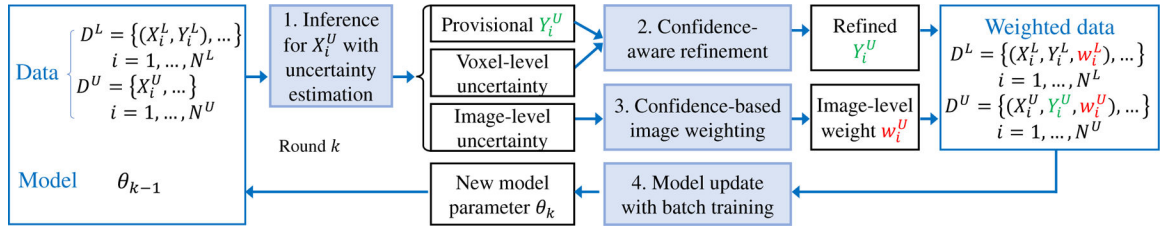
Fig. 3:

Our proposed I-CRAWL (Iterative Confidence-based Refinement And Weighting of pseudo Labels) framework for semi-supervised segmentation.
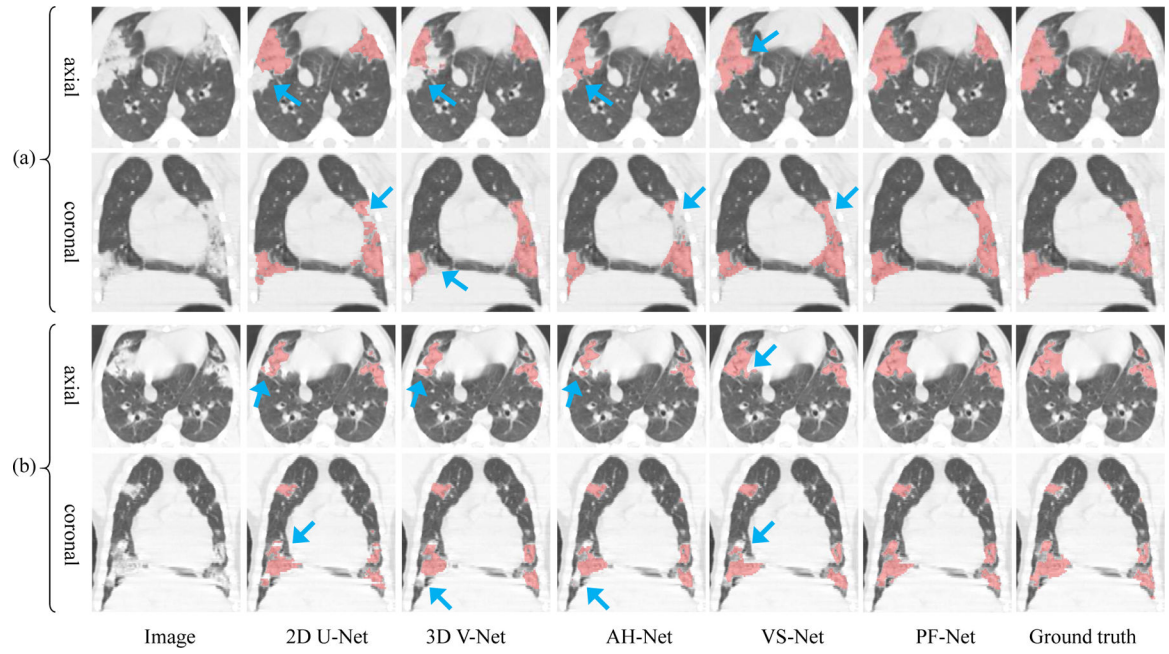
Fig. 4:

Qualitative comparison between different networks for PF lesion segmentation. (a) and (b) are from two different individuals. Blue arrows indicate some mis-segmented regions.
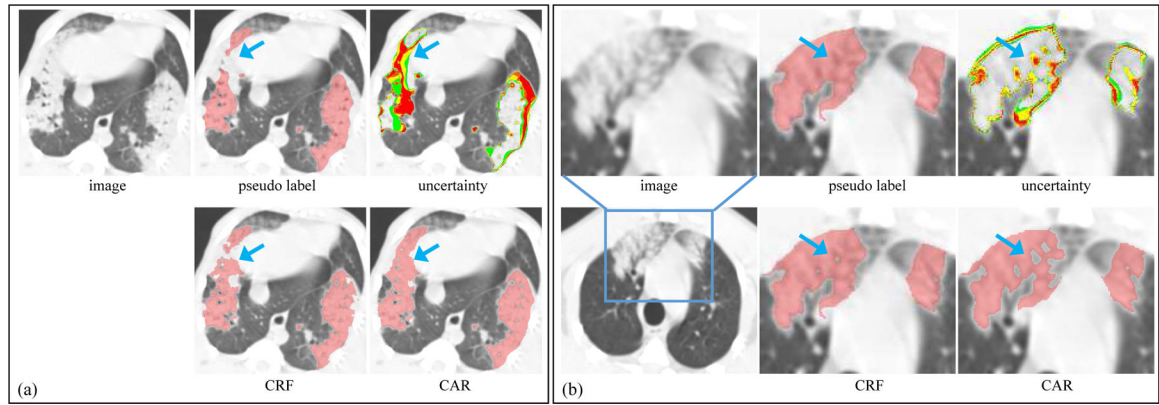
Fig. 5:
Visual comparison between CRF and our proposed Confidence-Aware Refinement (CAR) for update of pseudo labels. In the uncertainty map, dark green and red colors represent low and high uncertainty values, respectively. Blue arrows highlight some local differences.
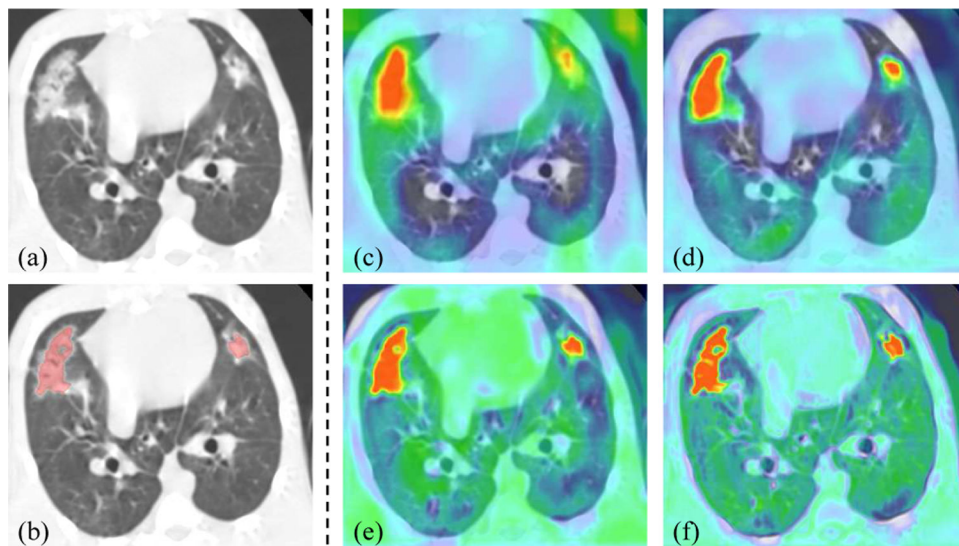
Fig. 6:

Visualization of multi-scale attention maps of PF-Net. (a)-(b) CT image and segmentation result. (c)-(f) Attention maps $\tilde{P}_5$, $\tilde{P}_4$, $\tilde{P}_3$ and $\tilde{P}_2$, respectively.

(a) Image (b) Network prediction (c) Ground truth

(d) Uncertainty (entropy min) (e) Uncertainty (Bayesian) (f) Uncertainty (MC dropout)
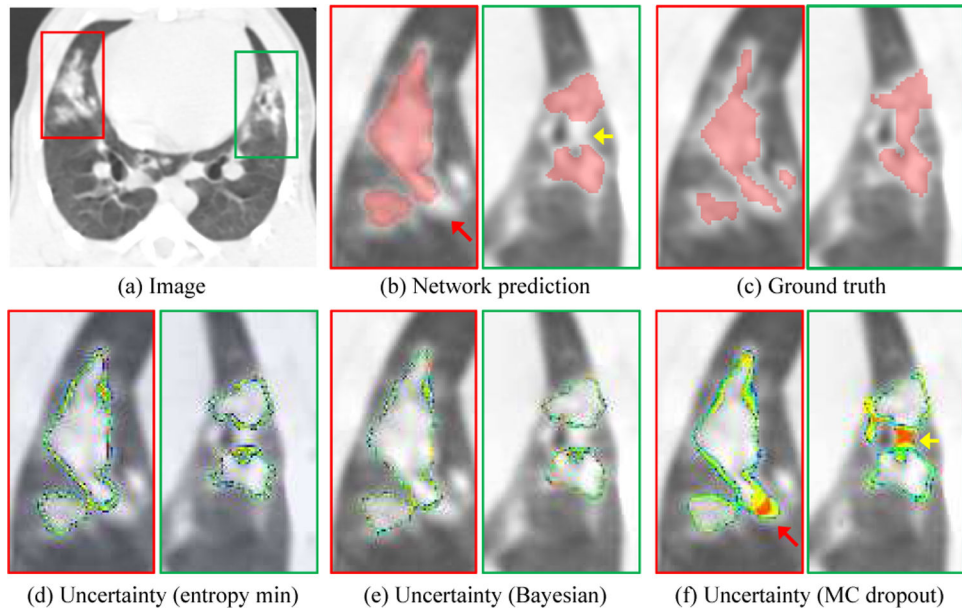
Fig. 7:
Visual comparison of uncertainty obtained by different methods. (b)-(f) show two sub-regions from (a), as indicated by red and green rectangles, respectively.

Image and ground truth	Round 0	Round 1	Round 2	Round 3
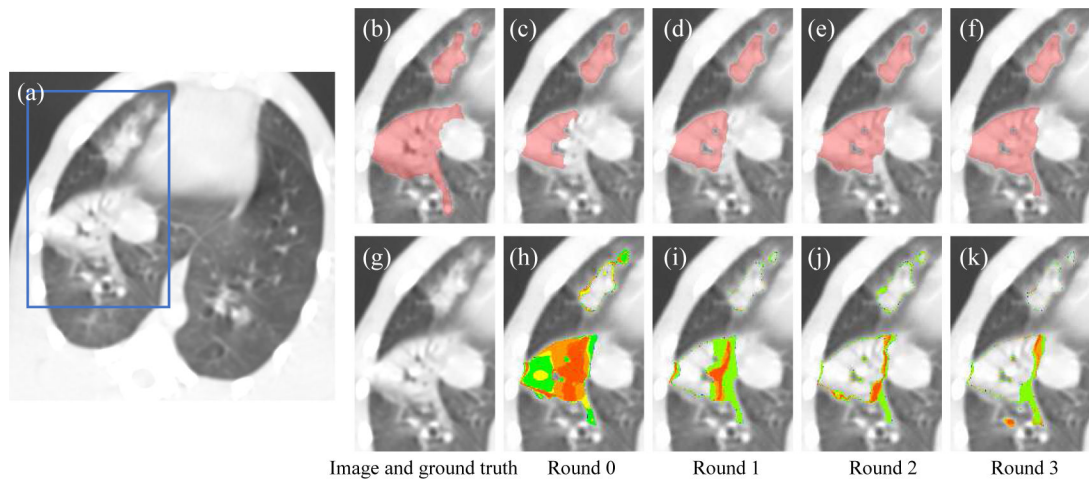
Fig. 8:
Change of uncertainty maps at different rounds of I-CRAWL. (a) input image. (g) sub-region of the input. (b) ground truth (just for reference here, not used during training). (c)-(f) and (h)-(k) show the pseudo labels and the corresponding uncertainty maps of I-CRAWL as the round increases.

(a) Dice

(b) HD$_{95}$
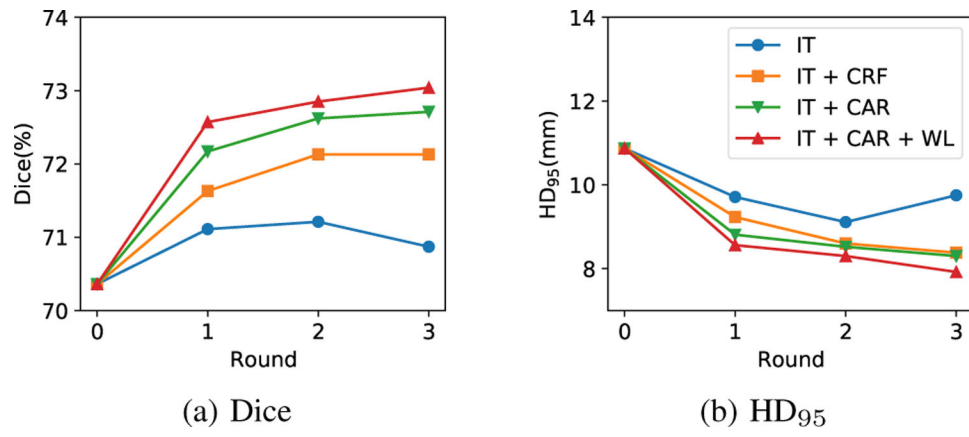
Fig. 9:

Comparison between variants of I-CRAWL at different training rounds for semi-supervised learning. Round 0 means the baseline that only learns from annotated images.

TABLE I:

Quantitative comparison of different networks for PF lesion segmentation.

	Dice (%)	RVE (%)	HD$_{95}$ (mm)
2D U-Net [15]	59.31±15.66	38.05±19.17	15.73±15.57
Attention U-Net [22]	59.22±14.55	31.93 ±20.47	18.64±15.58
3D U-Net [18]	57.19±13.47	40.09 ±22.45	18.93±12.98
3D scSE-Net [17]	59.54±12.24	41.95±19.59	19.18±15.57
3D V-Net [19]	60.37±13.33	35.03 ±27.95	19.22±14.65
nnU-Net [27]	64.51±15.25	33.19±21.54	18.46±15.70
AH-Net [25]	60.50±13.27	35.77±23.96	17.90±13.66
VS-Net [29]	67.45±9.70	30.03 ±24.78	18.35±17.67
PF-Net	**70.36±10.14** [*]	**27.96±21.72**	**10.87±9.28** [*]

[*] denotes significant difference from the others (p-values < 0.05).

TABLE II:

Quantitative comparison of different variants of baseline network and our PF-Net. Baseline (M) means 2D and 3D convolutions are used in the first M and the other convolutional blocks, respectively.

	Dice (%)	RVE (%)	HD$_{95}$ (mm)
Baseline ($M = 0$)	60.83±14.93	36.09±21.29	19.80±11.48
Baseline ($M = 1$)	62.42±13.01	33.74±24.50	18.87±12.03
Baseline ($M = 2$)	**67.87±9.96**	**30.79±20.37**	**15.31±15.88**
Baseline ($M = 3$)	65.98±12.27	31.45±21.84	16.92±12.66
Baseline ($M = 4$)	65.93±14.01	32.37±23.17	17.35±12.67
Baseline ($M = 5$)	63.11±10.90	32.71 ±22.35	17.60±12.28
+ non-local	68.63±15.49	29.82±22.27	14.80±13.99
+ deep supervision	68.57±10.30	30.73 ±21.24	17.56±19.22
+ MSGA	69.32±13.20	**27.55±21.67**	13.85±16.68
+ MSGA + MSDA°	70.25±10.09 *	28.09±19.67	12.28±12.12
+ MSGA + MSDA	**70.36±10.14** *	27.96±21.72	**10.87±9.28** *

The second section is based on Baseline ($M = 2$), and

* denotes significant improvement from it (p-value < 0.05).

Note that the last row is our PF-Net.

TABLE III:

Performance on the validation set based on different γ values in the confidence-based sampling weighting of pseudo labels in round 1 of I-CRAWL. "Initial" refers to model pre-trained on annotated images (round 0).

	Dice (%)	RVE (%)	HD$_{95}$ (mm)
Initial	66.62±8.53	23.91±15.14	11.11±6.00
No weighting	67.83±8.25	22.74±17.82	9.12±4.48
$\gamma = 1.0$	67.36±8.75	24.73±18.83	10.02±5.24
$\gamma = 2.0$	68.00±8.23	23.01±16.90	8.69±4.60[*]
$\gamma = 3.0$	**68.75±8.00**[*]	**22.42±17.11**	**8.97±4.45**[*]
$\gamma = 4.0$	68.54±8.64[*]	22.45±18.06	8.87±4.72[*]
$\gamma = 5.0$	66.86±8.74	23.44±17.95	8.75±4.51[*]

[*] denotes significant improvement from it (p-value < 0.05).

TABLE IV:

Comparison of different uncertainty estimation methods used in CAR for the validation set in the first round.

	Dice (%)	RVE (%)	HD$_{95}$ (mm)
Initial	66.62±8.53	23.91±15.14	11.11 ±6.00
CAR (entropy min)	67.68±10.64	22.73±17.38	10.54±6.39
CAR (Bayesian)	68.31±9.52	21.80±16.53	10.39±6.03
CAR (MC Dropout)	**69.05±9.41** *	**21.31±16.10**	**9.94±6.16** *

*
denotes significant improvement from Initial.

TABLE V:

Ablation study of our semi-supervised method I-CRAWL for PF lesion segmentation. IT: iterative training. CAR: Confidence-aware refinement. IW: Image weighting of pseudo labels.

	Dice (%)	RVE (%)	HD$_{95}$ (mm)
Baseline	70.36±10.14	27.96±21.72	10.87±9.28
IT	70.87±8.77	28.37±17.38	9.75±6.36
IT + CRF	72.13±8.97	27.38±20.39	8.38±4.75
IT + CAR	72.71±9.50 [*]	**26.05±18.79**	8.30±5.17
IT + CAR + IW	**73.04±10.14** [*]	26.12±20.37	**7.92±4.87** [*]
Full Supervision	74.58±10.99	23.07±15.79	7.03±4.09

[*] denotes significant improvement from the baseline (p-value < 0.05).

TABLE VI:

Quantitative comparison of different semi-supervised methods for PF lesion segmentation. R_{an}: Ratio of annotated images in the training set.

R_{an}	Method	Dice (%)	RVE (%)	HD$_{95}$ (mm)
10%	Baseline	54.73±16.34	45.83±24.99	25.79±18.71
	Fan et al. [13]	60.44±15.86	42.40±21.06	17.44±16.00
	Bai et al. [32]	65.03±18.01	32.98±23.16	16.82±14.82
	Cui et al. [33]	61.86±16.20	41.16±20.78	16.72±16.18
	UA-MT [34]	63.94±15.72	38.47±21.22	15.14±15.34
	I-CRAWL	**67.18±14.02**	**32.12±18.50**	**11.85±9.82**
20%	Baseline	70.36±10.14	27.96±21.72	10.87±9.28
	Fan et al. [13]	71.62±10.32	27.57±19.09	10.54±8.41
	Bai et al. [32]	72.13±8.97	27.38±20.39	8.38±4.75
	Cui et al. [33]	70.39±10.42	26.63±24.85	9.48±8.82
	UA-MT [34]	70.46±9.53	27.14±23.17	8.25±4.93
	I-CRAWL	**73.04±10.14**	**26.12±20.37**	**7.92±4.87** †
50%	Baseline	72.09±12.54	26.49±23.19	9.38±6.43
	Fan et al. [13]	72.69±11.41	26.46±17.71	9.19±6.45
	Bai et al. [32]	72.92±10.41	25.71±19.97 †	7.45±4.98 †
	Cui et al. [33]	72.72±12.44	24.57±19.96 †	8.11±5.76 †
	UA-MT [34]	73.40±11.48	24.67±19.08 †	7.86±4.56 †
	I-CRAWL	**74.02±13.77** †	**23.93±16.75** †	**7.24±5.03** †
Full Supervision		74.58±10.99	23.07± 15.79	7.03±4.09

†denotes there is no significant difference from full supervision (p-value > 0.05).