Segmentation of Blood Vessels and Pathological Regions from Computed Tomography Images using Convolutional Neural Networks

Submitted to

Graduate School of Information Science and Technology

Osaka University

January 2022

Yuki SUZUKI

List of Publications

As the first author

Journal

1. Suzuki, Y., Kido, S., Mabu, S., Yanagawa, M., Tomiyama, N., Sato, Y. Segmentation of Diffuse Lung Abnormality Patterns on Computed Tomography Images using Partially Supervised Learning. 2021. Advanced Biomedical Engineering (accepted for publication).

2. Suzuki, Y., Hori, M., Kido, S., Otake, Y., Ono, M., Tomiyama, N., Sato, Y. Comparative Study of Vessel Detection Methods for Contrast Enhanced CT: Effects of CNN Patch Size and Network Architecture. 2021. Advanced Biomedical Engineering, vol. 10, pp. 138–149.

International conference

3. Suzuki, Y., Yamagata, K., Yanagawa, M., Kido, S., Tomiyama, N., 2020. Weak supervision in convolutional neural network for semantic segmentation of diffuse lung diseases using partially annotated dataset, in: Medical Imaging 2020: Computer-Aided Diagnosis. Presented at the Medical Imaging 2020: Computer-Aided Diagnosis, International Society for Optics and Photonics, p. 113142A.

4. Suzuki, Y., Fung, G.S.K., Shen, Z., Otake, Y., Lee, O., Ciuffo, L., Ashikaga, H., Sato, Y., Taguchi, K., 2017. Projection-based motion estimation for cardiac functional analysis with high temporal resolution: a proof-of-concept study with digital phantom experiment, in: Medical Imaging 2017: Physics of Medical Imaging. Presented at the Medical Imaging 2017: Physics of Medical Imaging, SPIE, pp. 778–781.

5. Suzuki, Y., Okada, T., Hori, M., Yokota, F., Linguraru, M.G., Tomiyama, N., Sato, Y., 2013. Automated Segmentation and Anatomical Labeling of Abdominal Arteries Based on Multi-organ Segmentation from Contrast-Enhanced CT Data, in: Drechsler, K., Erdt, M.,

Linguraru, M.G., Oyarzun Laura, C., Sharma, K., Shekhar, R., Wesarg, S. (Eds.), Clinical Image-Based Procedures. From Planning to Intervention, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 67–74.

6. Suzuki, Y., Okada, T., Hori, M., Yokota, F., Linguraru, M.G., Tomiyama, N., Sato, Y., 2012. Automated anatomical labeling of abdominal arteries from CT data based on optimal path finding between segmented organ and aorta regions : A robust method against topological variability. Int J Comput Assist Radiol Surg, 7(Suppl. 1), S47-48.

Domestic conference (written in Japanese)

7. Suzuki, Y., Okada, T., Yokota M., Hori, M., Tomiyama, N., Sato, Y.: Automated anatomical labeling of abdominal arteries from CT data based on optimal path finding between segmented organ and aorta regions -Improvements in optimal path finding-, IEICE Technical Report, Vol. 112, No. 38, pp. 91-96, 2012 (鈴木 裕紀, 岡田 俊之, 横田 太, 堀 雅敏, 富山 憲幸, 佐藤 嘉伸: 大血管と臓器領域間の最適経路探索に基づく多時相三次元腹部 CT 画像からの血管名自動対応付け ～ 最適経路探索の改善 ～, 電子情報通信学会技術報告, Vol. 112, No. 38, pp. 91-96, 2012)

8. Suzuki, Y., Okada, T., Yokota M., Hori, M., Tomiyama, N., Sato, Y.: Automated anatomical labeling of abdominal arteries from CT data based on optimal path finding between segmented organ and aorta regions, IEICE Technical Report, Vol. 111, No. 389, pp. 401-406, 2012 (鈴木 裕紀, 岡田 俊之, 横田 太, 堀 雅敏, 富山 憲幸, 佐藤 嘉伸: 大血管と臓器領域間の最適経路探索に基づく多時相三次元腹部 CT 画像からの血管名自動対応付け, 電子情報通信学会技術報告, Vol. 111, No. 389, pp. 401-406, 2012)

As a co-author

Journal

9. Masumoto, N., Suzuki, Y., Cui, S., Wakazaki, M., Sato, M., Kumaishi, K., Shibata, A., Furuta, K.M., Ichihashi, Y., Shirasu, K., Toyooka, K., Sato, Y., Yoshida, S., 2021. Three-dimensional reconstructions of

haustoria in two parasitic plant species in the Orobanchaceae. Plant Physiology 185, 1429–1442.

10. Wataya, T., Nakanishi, K., Suzuki, Y., Kido, S., Tomiyama, N., 2020. Introduction to deep learning: minimum essence required to launch a research. Jpn J Radiol 38, 907–921.

11. Hata, A., Yanagawa, M., Yamagata, K., Suzuki, Y., Kido, S., Kawata, A., Doi, S., Yoshida, Y., Miyata, T., Tsubamoto, M., Kikuchi, N., Tomiyama, N., 2021. Deep learning algorithm for detection of aortic dissection on non-contrast-enhanced CT. Eur Radiol 31, 1151–1159.

12. Sakamoto, M., Hiasa, Y., Otake, Y., Takao, M., Suzuki, Y., Sugano, N., Sato, Y., 2020. Bayesian Segmentation of Hip and Thigh Muscles in Metal Artifact-Contaminated CT Using Convolutional Neural Network-Enhanced Normalized Metal Artifact Reduction. J Sign Process Syst 92, 335–344.

13. Nishi, M., Kanaji, S., Otake, Y., Harada, H., Yamamoto, M., Oshikiri, T., Nakamura, T., Suzuki, S., Suzuki, Y., Hiasa, Y., 2017. Quantitative comparison of operative skill using 2-and 3-dimensional monitors during laparoscopic phantom tasks. Surgery 161, 1334–1340.

International conference

14. Sakamoto, M., Hiasa, Y., Otake, Y., Takao, M., Suzuki, Y., Sugano, N., Sato, Y., 2019. Automated segmentation of hip and thigh muscles in metal artifact contaminated CT using CNN, in: International Forum on Medical Imaging in Asia 2019. Presented at the International Forum on Medical Imaging in Asia 2019, SPIE, pp. 124–129.

15. Kidoh, M., Shen, Z., Suzuki, Y., Ciuffo, L., Ashikaga, H., Fung, G.S.K., Otake, Y., Zimmerman, S.L., Lima, J.A.C., Higuchi, T., Lee, O., Sato, Y., Becker, L.C., Fishman, E.K., Taguchi, K., 2017. False dyssynchrony: problem with image-based cardiac functional analysis using x-ray computed tomography, in: Medical Imaging 2017: Physics of Medical Imaging. Presented at the Medical Imaging 2017: Physics of Medical Imaging, SPIE, pp. 449–455.

16. Okada, T., Linguraru, M.G., Hori, M., Suzuki, Y., Summers, R.M., Tomiyama, N., Sato, Y., 2012. Multi-organ segmentation in abdominal CT images. Annu Int Conf IEEE Eng Med Biol Soc 2012, 3986–3989.

Abstract

Semantic segmentation plays fundamental rolls in today's medical fields from both clinical and scientific viewpoints. Automating semantic segmentation is highly valuable and desired because manually performing segmentation is far too cost-ineffective and time-consuming. The segmentation targets in medical images can be divided into the following three categories. 1) Large anatomical structures such as abdominal organs, muscle, bones, lung, and brains. 2) Inter and intra structures of organs such as blood vessels and bronchi. 3) Pathological abnormalities such as tumors, hemorrhage, and aneurysms. The goal of this thesis is to address following two problems that found in 2) and 3) respectively. The problem found in category 2) is that there are few comparative studies of methods while there are numerous studies that propose new methods. The problem found in category 3) is that it is hard to collect large-scale dataset that is annotated by qualified experts because of its higher cost for annotation.

The first problem was addressed through a study on abdominal artery segmentation. Two methods that use convolutional neural network and one method that is based on analytical filter were compared through the experiments that used 30 cases of contrast enhanced abdominal computed tomography images. The experiments showed that the newer method (namely UNet) was the best in the training region of interests. However, the experiments also showed that the other two methods outperformed the newer method in other regions.

The second problem was addressed through a study on diffuse lung abnormality patterns segmentation. A new method that takes full advantage of partially annotated dataset is proposed. The experiments that used 372 patients of chest CT images were conducted to evaluate the proposed method. The experimental results showed that the proposed method improved segmentation accuracy by reducing the leakage of the segmentation. Further analysis of the results showed that the proposed method effectively utilized unannotated pixels, which were mostly comprised of healthy lung pixels, and improved the segmentation accuracy between normal lung pixels and other pixels with abnormal textures.

Contents

1 Introduction

1.1 Background

Semantic segmentation of medical images is of great importance in today's medical field. Medical images such as computed tomography (CT) image and magnetic resonance image (MRI) are scanned as part of the clinical routine dozens of times a day in a hospital and enormous amount of human resource is spent for processing them. Semantic segmentation is an image processing task of segmenting an image into semantic clusters, which can be individual organs, muscles, blood vessels, and pathological regions based on the context and the goal of the medical image. Clinically, semantic segmentation can be used for various tasks such as measuring sizes of certain structures to assess its functionalities and the visualization based on semantic segmentation is essential to grasp shapes and relationships of anatomical structures inside the body. Furthermore, semantic segmentation is a fundamental step for medical image analysis such as computational anatomy and computational physiology.

Manually performing semantic segmentation is possible but not the most reliable and cost-effective option. First, there is inter and intra operator variability in the quality and accuracy of manual segmentation. Second, labor-intensive nature of manual segmentation is problematic especially for medical images because medical images such as CT and MRI are three dimensional, which makes it more labor-

intensive, and outsourcing the task is usually not an option because of the patient privacy. Automating semantic segmentation can mitigate these problems and therefore highly valuable and desired.

1.2 Methods and challenges

One typical classification of the segmentation targets is to divide them into following three categories based on their anatomical characteristics. 1) Large anatomical structures that compose most of the body such as abdominal organs, muscles, bones, lung, and brains. 2) Inter and intra structures of organs such as blood vessels and bronchi. 3) Pathological abnormalities such as tumors, hemorrhage, and aneurysms.

1.2.1 Large anatomical structures

Large anatomical structures such as abdominal organs, muscles, bones, lung, brain and so on compose most of the body. Segmenting these regions is often performed as a first step of image processing. Most of these regions are immediately recognizable and searching through the image is not necessary because they are large and usually in certain positions of the body. However, segmenting these regions can be often hard because the image contrast is low, or the image is noisy. Furthermore, separating neighboring structures in tightly packed region such as thigh muscles can also be hard because of unclear margins between structures and it requires experiences to manually delineate each structure. Another factor that could affect segmentation

is deformations and changes in pixel values caused by pathological abnormalities.

Prime example of the segmentation tasks in this category is abdominal organ segmentation [1] and has been in one of the main focuses of medical image processing community. The segmentation tasks that perform subdivision of the organ include muscle [2], [3], lung lobe [4], and liver lobe [5], [6] segmentation. This category has been studied fairly well and segmenting healthy structure with high accuracy is now becoming possible unless there are abnormalities [7].

1.2.2 Inter and intra structures of organs

Inter and intra structures of organs include blood vessels and bronchi. Imaging protocols such as contrast enhanced CT help to recognize structures in this category but segmentation accuracy in this category is often limited by the image resolution because tiny structures such as arteries can easily be sub-voxel scales. As they branch away from major trunk such as aorta and trachea, their positional and topological variation increases. Blood vessels attract the interest of many researchers because of its variety and clinical importance. Blood vessels are not large in volume compared to the structures in the first category because they consist of tubular structures. However, manually segmenting them from start to end is painfully time-consuming because it requires tracking numerous branches that can span large area.

Blood vessels in different regions have different properties, therefore they often require specialized studies. Retinal vessels found in fundus

images [8] have crossovers because the images are two dimensional. Coronary arteries [9] and cerebral vessels [10] usually have static branching and running patterns. Pulmonary vessels tend to be straighter than vessels in other regions and are easy to distinguish from the surrounding tissues.

A number of new methods based on deep learning has been proposed but not many of them are focusing on comparing different methods. There are comparative studies of deep learning based methods on retinal vessel segmentation [11]–[14]. On the other hand, although there are comparative studies on other vessels [15], [16], they are not based on deep learning. Therefore, there is a lack of comparative studies on abdominal vessels.

1.2.3 Pathological abnormalities

This category includes abnormalities such as tumors, hemorrhage, and aneurysms. The variations in shapes and appearances tend to be much larger compared to the structures in previous two categories. Determining the type and area of the abnormality is hard and there are substantial amount of inter-operator variabilities even among qualified experts [17], [18]. For small objects such as lung nodules [19]–[21] or aneurysms [22], [23], segmentation can be substituted or preceded by object detections.

Primary example in this category is hepatic or renal tumor segmentation [24], [25]. One of the challenges in this category is that it is hard to collect large-scale dataset that is annotated by qualified experts because of its higher cost for annotation. Therefore, in this

category, taking full advantage of limited amount of annotated dataset is more important.

1.2.4 Methodological aspects

Semantic segmentation methods commonly used for medical images are listed here. Even though deep learning based methods are the main focus of the thesis as well as the current research scene, non-deep learning methods commonly used before the prevalence of deep learning are also listed here to help understand the characteristics of the segmentation tasks. These conventional methods are usually based on statistical or analytical priors designed and constructed by the researchers and the dataset.

Graph cut [26], [27] is a semi-automated segmentation method that typically works well for well-defined object with homogeneous pixel intensity values. Even though this is not a fully automated method, interactive update of the segmentation makes this method far more convenient than fully manual segmentation. Statistical atlas model [1], [28], [29] is a method that statistically models the position, appearance, and shape of the objects and is commonly applied to large anatomical structures. Multi atlas method is a method based on image registration between the atlases and the patient. Individual muscle segmentation [2] and parcellation of brain region [30], [31] are primary applications of this method and this method works robustly on objects that have no clear boundaries. Hessian based filters [32]–[34] are analytical image filters that enhances geometric shapes such as tubes and sheets based on their local appearances. They work well when their analytical models match actual image appearances.

Convolutional neural network (CNN) revolutionized semantic segmentation and most of newly proposed methods these days are based on CNN. While CNN contributed to the improvement of accuracy and the unification of the methods, it also created issues that have not been addressed well yet. First, despite the increase in the number of studies published in the field, studies that are focused on the comparison among methods are scarce. Second, although training CNNs necessitates a large-scale dataset, large-scale dataset is sometimes hard to obtain especially when it is a medical image dataset.

1.3 Research objectives

The purpose of this thesis is as follows.

1. Addressing the lack of comparison study on abdominal blood vessel segmentation compared to the abundance of new proposals of methods.
2. Proposing and testing a method that takes full advantage of existing dataset even when expert annotations are only partially available.

The first purpose is addressed through a study on abdominal artery segmentation in chapter 2. In this study, multiple methods are tested on abdominal arteries and their characteristics are compared. The second purpose is addressed through a study on diffuse lung abnormality patterns segmentation in chapter 3. In this study, it is hard to construct a large-scale dataset and new training method is proposed

to make effective use of limited amount of expert-annotated training dataset.

1.4 Thesis outline

In chapter 2, details of current landscape of vessel segmentation are described and experiments are conducted to study characteristics of multiple methods. This chapter is based on my previous publication [35] about renal artery segmentation. In chapter 3, difficulties and related methods of diffuse lung abnormality patterns segmentation are described and new method is proposed to tackle the challenges of this segmentation task. This chapter is based on my previous publication [36], [37]. Finally, summary of this thesis and future work is discussed in the chapter 4.

2 Abdominal Vessel Segmentation

2.1 Overview

Segmenting blood vessels is an important step in a wide variety of tasks in medical image analysis. Patch-based CNNs are often used for vascular detection, but the impact of patch size and choice of CNN architecture have not been addressed in detail in previous studies. In this study, we aim to investigate the impacts of patch size and CNN architecture on the accuracy of vascular detection from contract enhanced CT. We targeted the renal arteries as the primary focus of detection.

We conducted experiments involving 30 cases of contrast enhanced abdominal CT data. For the experiments, arteries in the pre-defined regions of interest were manually labeled to build a dataset of input CT images and ground truth labels. We repeated the experiments with four patch sizes and two patch-based 3D CNN architectures (U-Net like and a simple sequential model) to evaluate the differences. Moreover, a Hessian-based line enhancing method was included in the evaluation to compare the CNNs with a non-deep learning method.

The experimental results show that patch size has a significant impact on detection accuracy. U-Net like model had a peak accuracy at a certain patch size unlike the sequential model, which plateaued with

large patch sizes. Although both CNNs outperformed Hessian-based line enhancement by a large margin, Hessian-based line enhancement obtained good recall when enhancing vessel structures not included in the CNN training. Our experiments showed that different network architectures have different characteristics regarding their response to different patch sizes and vessel structures unseen during training.

2.2 Background

Blood vessels play a vital role in a wide variety of medical tasks such as surgical planning and diagnosis. For example, in organ transplants, a patient's vessel structure is one of the factors that determines the operability of the patient. In a partial resection operation (e.g., partial hepatectomy and nephrectomy), the optimal resection is designed by estimating blood vessel-dominant regions based on the vascular structure.

Understanding vascular structure is a challenging task because the structure has an elongated tubular structure with curves and branching points. Therefore, assisting humans with vessel recognition has been a main topic of the studies in the medical image analysis field. Common approaches in these studies are vessel enhancement [32], [33] and automated vessel segmentation [38]–[40]. Vessel structures are extracted from images in a broad range of image protocols and body parts, such as fundus images for diabetic retinopathy, cardiac CT for coronary stenosis, and head CT for aneurysms [41].

CNNs have been widely used in semantic segmentation since AlexNet [42] won an image classification contest by a large margin and a large number of methods and tools have been developed. In medical image analysis, researchers have begun to use CNNs for a variety of tasks such as classification and detection. A fully convolutional network [43] is typically used for segmentation tasks. The semantic segmentation of blood vessels is no exception to this trend, and a number of new techniques specifically for segmenting tubular structures have been proposed [44]–[48].

Patch-based networks are commonly used in the semantic segmentation of medical images. This is because the target object is often localized in a limited region of interest (ROI) and because 3D volumetric images are too big to fit in GPU memory, which is commonly used in the training of CNNs. Although a CNN is capable of learning its weight parameters through training, there are hyper parameters for the network that are not optimized by the training process. In patch-based networks, the patch size is one hyper-parameter that is often heuristically determined. Whereas patch-based CNN is often used for vascular segmentation, the impacts of patch size have not been addressed in detail so far. In many previous studies, a fixed patch size was used ([49]–[51] to name a few). Although a couple of papers [52], [53] have mentioned patch size in their studies, more detailed study on the impacts of patch size is needed to achieve better understanding of vascular segmentation using patch-based CNN.

Oda et al. [52] proposed a method that segments abdominal arteries in contrast enhanced CT and reported the segmentation accuracies for

different patch sizes. A limitation in their work is that they did not test a 3D network. They used 2.5D network that incorporates three planes, namely the axial, sagittal, and coronal planes. In their work, the 2.5D CNN performed better than a simple 2D CNN. However, the development of a 3D CNN that is capable of capturing a 3D structure remained unaddressed.

Yang et al. [53] proposed a method that segments catheters in 3D ultrasound images using techniques such as focal loss and dense sampling. In their experiments, a 3D U-Net was used, and they reported that changing patch size changed the segmentation accuracy. The limitation of in their work regarding patch size in the context of vascular segmentation is as follows. First, their target is a catheter, which has a tubular structure similar to that of blood vessels. However, unlike vessels, a catheter does not have branches nor large changes in diameter. Second, because they used 3D ultrasound, image characteristics such as field of view and background organs are different from those in other image modalities such as CT and MR.

The task of this work is to detect abdominal arteries in contrast enhanced 3D CT using patch-based 3D CNN. Our contribution is three-fold. First, we study the impact of the patch size and network architecture in the task of abdominal artery detection in contrast enhanced 3D CT through experiments. We determine the best-performing CNN architecture and its optimal patch size as well as how different CNN architectures are affected by patch size. Second, we study the CNN's behavior when detecting categories of vessels that are not included in the training data. Third, a typical non-deep learning method,

namely the Hessian-based multiscale line enhancement method [32], is compared with CNN-based methods.

2.3 Materials and Methods

2.3.1 Materials

In this work, we used 30 cases of contrast enhanced abdominal CT scanned at Osaka University Hospital. The average voxel size of the images is approximately $0.68 \times 0.68 \times 0.63\,\mathrm{mm}^3$. The ground truth data of vessel regions were prepared for three types of ROIs, which were defined for each case to observe the differences among methods in different regions. The three types of ROIs, as illustrated in Figure 2.1, are as follows.

- Renal ROI (red rectangles in Figure 2.1): A Bounding box that circumscribes both a kidney and the arterial tree stemming from the aorta. Two (left and right) renal ROIs were defined on each patient. The average size of each renal ROI was $142 \times 121 \times 171$ voxels and the standard deviation was $18.2 \times 7.41 \times 20.1$ voxels.
- Lung ROI (green rectangles in Figure 2.1): A $32 \times 32 \times 32$ voxels cube at the bottom of the right lung. No organs except for lung tissues and vessels were in the ROI.
- Spine ROI (blue rectangles in Figure 2.1): A $32 \times 32 \times 32$ voxels cube next to 12th thoracic vertebra with an intercostal artery at the center of the ROI. A portion of spinal bone tissue was included in the ROI.

The main ROI type was renal ROI, which was used for both the training of the CNNs and evaluation using cross-validation. The other two types of ROIs were used only for the evaluation. There are four contrast phases of the CT images (namely non-contrasted, early arterial, late arterial, and venous phases) in our contrast enhanced abdominal CT scan. The early arterial phase has the best contrast between arteries and other non-arterial regions such as veins in the renal ROI. Additionally, vessels in lung and spine ROIs are not very sensitive to the contrast phase because non-vascular regions in these ROIs are primarily air and bone respectively. Therefore, we used early arterial phase for all three types of ROIs. All arterial regions in the renal and spine ROIs and vessels in the lung ROIs were manually labeled by medical image researchers under an experienced radiologist's supervision.

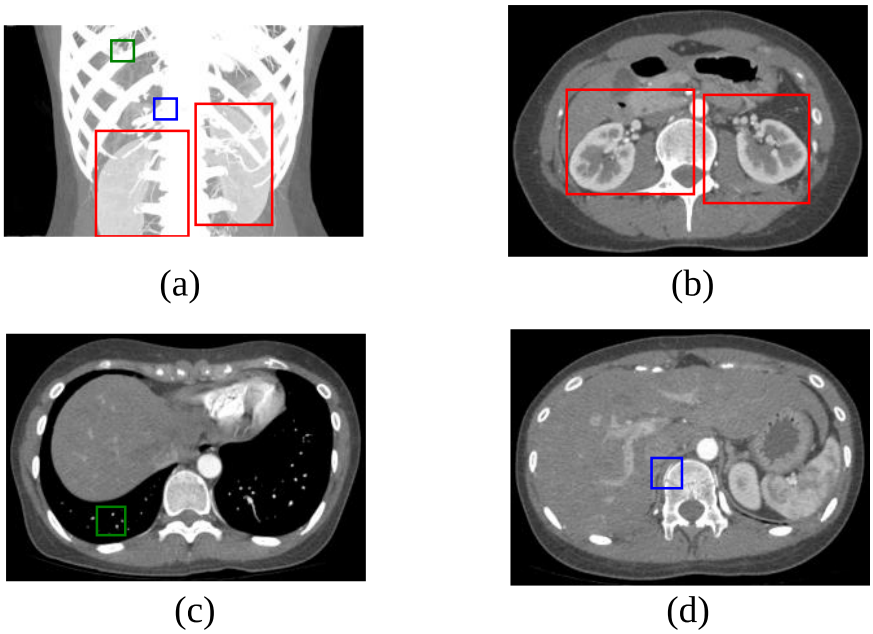


Figure 2.1: Placements of the three types of ROIs. Colored rectangles indicate ROI placements (Red: Renal ROI. Green: Lung ROI. Blue: Spine ROI.) (a) Maximum intensity projection (MIP) in the coronal plane. (b), (c), and (d) image slices with renal, lung, and spine ROIs respectively.

2.3.2 Methods

In this work, the following three methods were evaluated.

1. A U-Net [54] like (UN): a CNN that uses a U-Net like architecture, which is a network architecture commonly used in semantic segmentation of medical images.
2. Sequential network (SN): a CNN that uses a simple sequential network without skip connections.
3. Hessian-based line enhancement method (HM) [32].

The UN architecture used in the experiments is shown in Figure 2.2 (a). Our UN architecture is a 3D network using 3D convolution layers and 3D max pooling layers. The network has two max pooling layers for downsampling in the encoding part and two upsampling layers in the decoder part. The number of convolutional kernels is doubled at each convolutional layer before max pooling to avoid bottlenecks [55]. Batch normalization layers (BNs) are inserted between the convolutional layers and their activations.

The SN is a simple network shown in Figure 2.2 (b) that only consists of convolutional layers. Unlike UN, there are no skip-connections or downsampling/upsampling layers. The SN has the same number of convolutional layers as the UN.

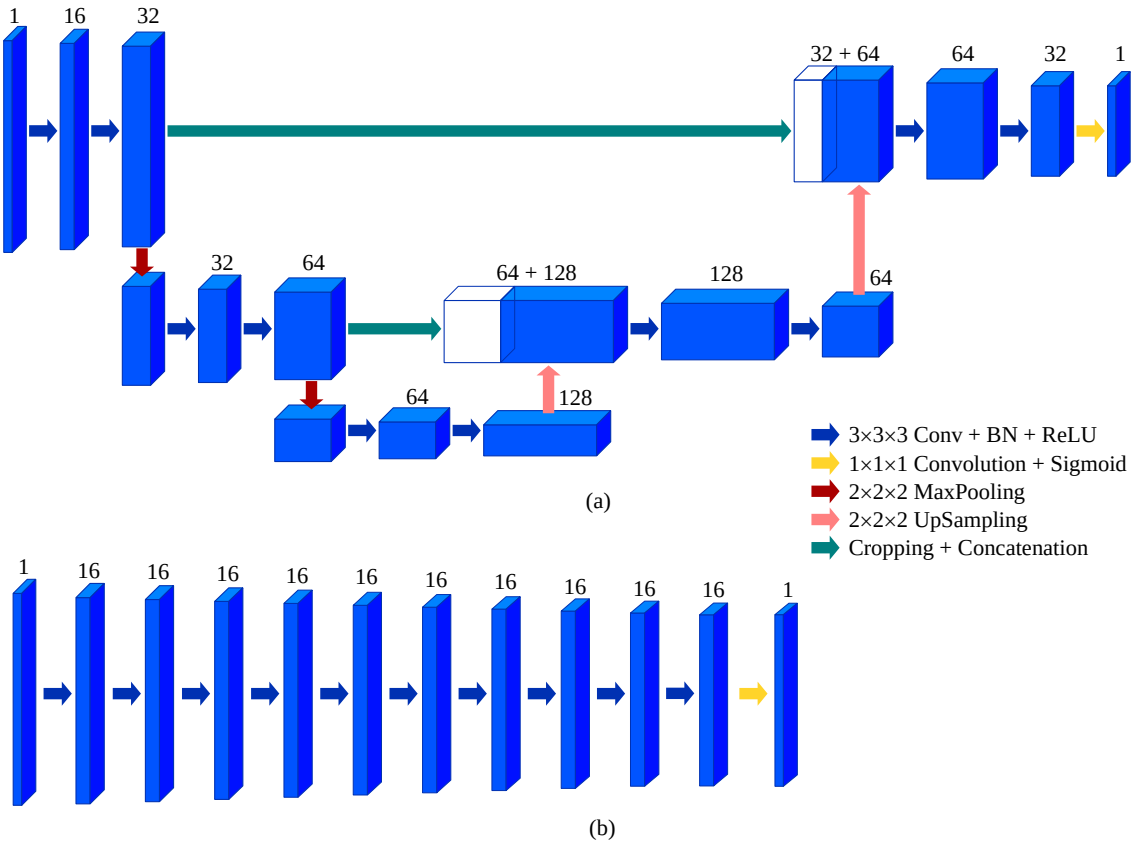


Figure 2.2: Network architectures used in the experiments. The blue boxes represent input, intermediate, and output 3D data with the number of channels indicated above. The white boxes represent concatenated data. (a) UN. (b) SN.

We used a patch-based method for detecting arteries with 3D CNNs, in which each patch is cubic. This is because the entire 3D volumetric images are too large to fit in the memory available in a GPU, which is essential for training of the CNNs. Furthermore, arterial regions and non-arterial regions are imbalanced in the entire volume and it is easier to rectify this volumetric imbalance using a patch-based method by extracting more patches that contain arteries from the training images than patches that do not contain arteries. With a patch-based method, an input volume is divided into 3D patches to be fed to the

network and the output patches are stitched together to reconstruct a full-sized output volume.

In our experiments, 3D patches were sampled using a sliding window with no overlap between output patches. Therefore, the step size of the sliding window is simply s_o, where s_o is the output patch size. In addition to the patches sampled by the sliding window, patches that contain the arterial center line were also sampled to mitigate the volumetric imbalance. A 3D patch was sampled at every voxel of the arterial center lines. Only patches sampled from renal ROIs were used for the training.

To study the impact of patch size, where the patch size denotes the edge length of each cubic patch, we repeated the experiment using four different patch sizes with the other parameters fixed. In our experiments, convolutional layers were applied without padding to avoid introducing false signals in the perimeter of the input, therefore convolutional layers reduce the output sizes. The minimum patch size applicable in the network is determined by the number of convolutional and pooling layers and it was 48 voxels in our experiments. The maximum patch size is bound by the amount of GPU memory and it was 96 voxels in our experimental environment, which is described later. Therefore, patch sizes used in our experiments ranged from 48 voxels to 96 voxels with a step size of 16, where the step size was heuristically determined to be small enough to study the impact of patch size.

In addition to the UN shown in Figure 2.2, shallower and deeper UNs were tested to observe the impact of the depth of the network. While

the UN has three levels of resolution, shallower and deeper UNs have two and four levels of resolution respectively. In other words, compared to the UN, shallower UN has one less pair of encoder and decoder and deeper UN has one more pair of encoder and decoder.

Early stopping was used to dynamically determine the number of training epochs. The validation set was extracted from the training set to calculate the validation loss. The early stopping algorithm monitored the validation loss and stopped the training process at the end of epoch in which the validation loss started to increase. After the early stopping, the best weight parameters (those that minimized the validation loss), were used for the evaluation.

The other hyper-parameters used in the experiments are the optimization method, which was Adam [56] with learning rate $\eta = 0.001$ (default) and batch size, which was 32. Binary cross entropy, which is defined as

$$H_{binary}(p, q) = -p\log q + (1 - p)\log(1 - q), \qquad (2.1)$$

where p is the ground truth label (0 or 1) and q is the prediction, was used as the loss function. The segmentation accuracies of the CNN methods were evaluated using experiments with 3-fold cross-validation. The training dataset was split at patient level so that the patches of one patient did not become split over different sets. The training and evaluation were conducted using a workstation with single NVIDIA TITAN RTX with 24GB of GPU memory. Training one UN took 461 minutes and training one SN 352 minutes on average.

In addition to the two CNNs, the HM [32] was also evaluated to compare the CNNs with a conventional method that does not require training thus does not require a labeled dataset. The gaussian standard deviation σ was adjusted for thin arteries in the HM to the values $\sigma = 1, \sqrt{2}, 2$ voxels.

Instead of volumetric evaluation metrics such as the Dice similarity coefficient [57] or Jaccard index [58], we chose metrics based on the similarity of the center lines of arteries for the following two reasons.

1. Because blood vessels are small in volume, volumetric similarity metrics are too sensitive to slight differences in the boundaries between manual traces and automatically extracted vessel regions.
2. When understanding vessel structure is the main purpose of detection, center line-based metrics are more appropriate than volumetric based metrics.

The output images of the methods were binarized by thresholding and a binary thinning algorithm [59] was applied to obtain the center lines of the extracted arteries. Likewise, the ground truth of the vessel center lines was generated by applying a binary thinning algorithm to the manually labeled vessels.

The area under the precision-recall curve [60] (AUPRC) was used to evaluate recall and precision of the results. The precision and recall are defined as

$$\begin{aligned} Precision &= TP/(TP + FP) \\ Recall &= TP/(TP + FN), \end{aligned}$$

where TP, FN and FP are defines as follows,

- The number of true positives (TP) is defined as the number of extracted center line voxels that are within two voxels of the center line of the ground truth.
- The number of false negatives (FN) is defined as the number of ground truth voxels that do not have an extracted centerline within a two-voxel proximity.
- The number of false positives (FP) is defined as the number of extracted center line voxels that are not within two voxels of the center line of the ground truth.

Although the arteries in the lung and spine ROIs were not used for the training, they are treated as true arteries in the evaluation. The Wilcoxon signed-rank test was used to calculate statistical significance.

Python 3.7 [61] and TensorFlow [62] were used to construct and train neural networks. Three-dimensional thinning of segmentation results [63] was implemented using Insight Toolkit [64].

2.4 Results

In this section, the results of the three methods for the three ROIs are shown. Hereafter, the UN and SN may have their patch size indicated in parentheses such as SN(64) and UN(80).

2.4.1 Evaluations for the renal ROI

Figure 2.3 shows box-and-whisker plots of the AUPRC comparing results for different patch sizes. As shown in Figure 2.3 (a), the SN hit a

plateau at a patch size of 64. There was no statistical significance among the models with patch sizes of 64, 80, and 96, although larger patch sizes only slightly increased average accuracy.

In the UNs, the model with a patch size of 80 had the best accuracy with statistical significance (Figure 2.3 (b)). Unlike the results with the SNs, the model with the largest patch size 96 had significantly worse accuracy than the model with a smaller patch size of 80.

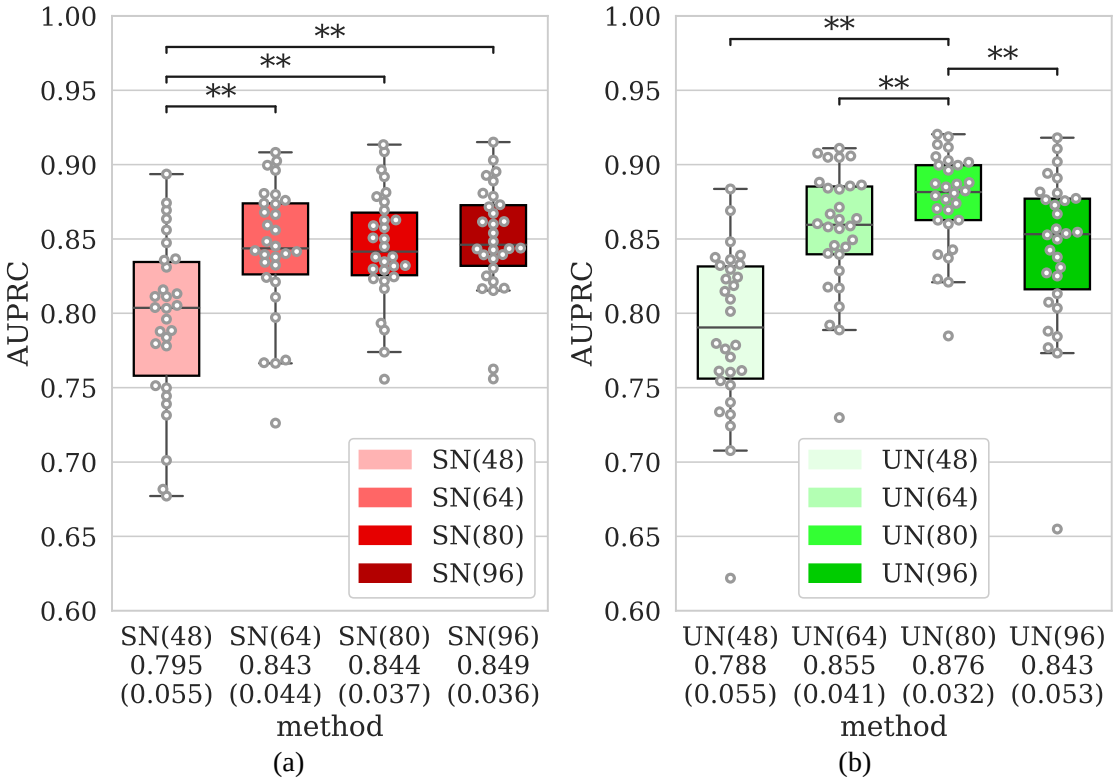


Figure 2.3: Box-and-whisker plots of the results for the renal ROI. The mean and standard deviation of each method is indicated below the method names. Statistical significances are indicated by ** (: $p < 0.01$.) (a) SN results. (b) UN results. In (b), statistical significance is only shown for a patch size of 80.

Figure 2.4 compares the results of the three methods. For each of the CNN methods, the patch size with the best result was chosen for the comparison. As shown in Figure 2.4 (a), the UN yielded in the best accuracy of the three methods.

In the precision-recall curve shown in Figure 2.4 (b), the UN obtained better results for both recall and precision. Although the HM had much lower precision, the best recall was better than those of the CNN-based methods.

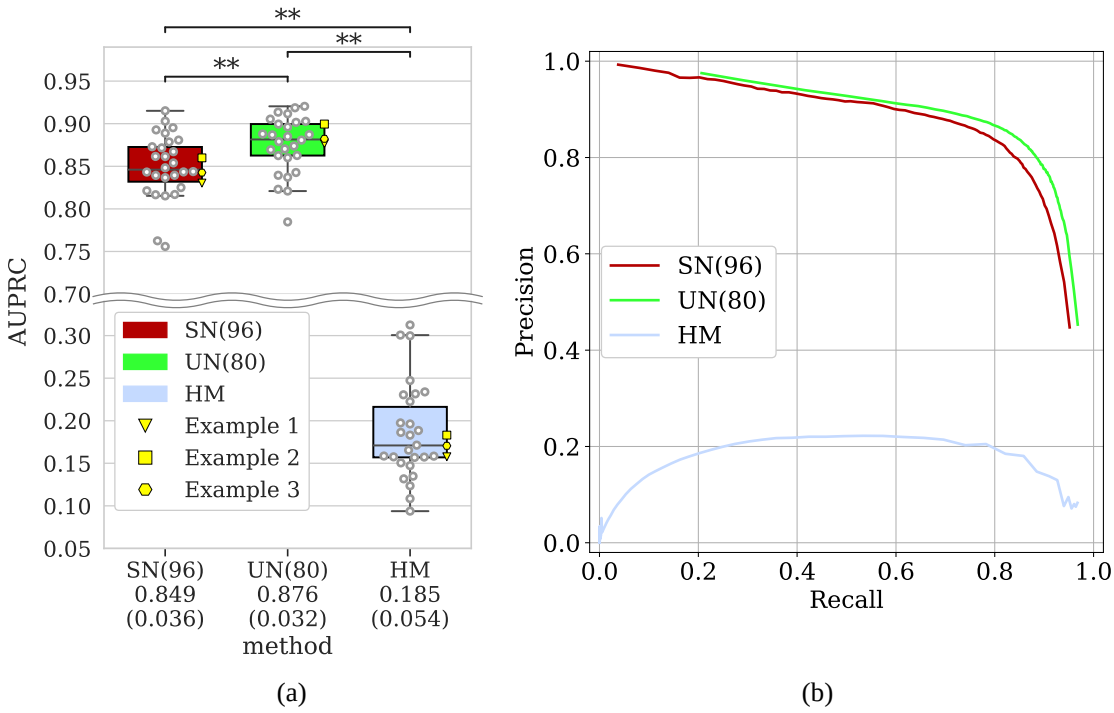
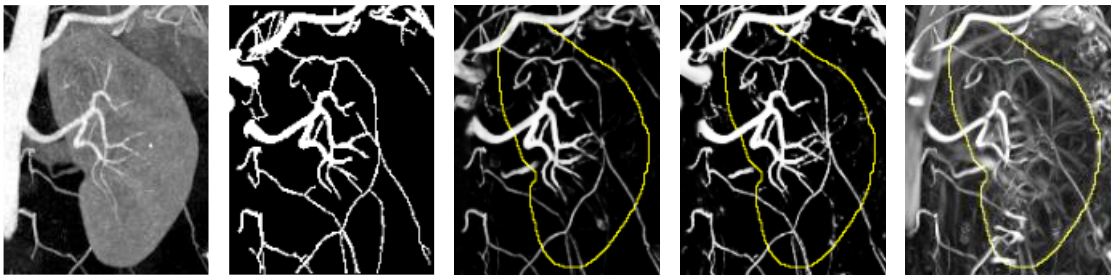
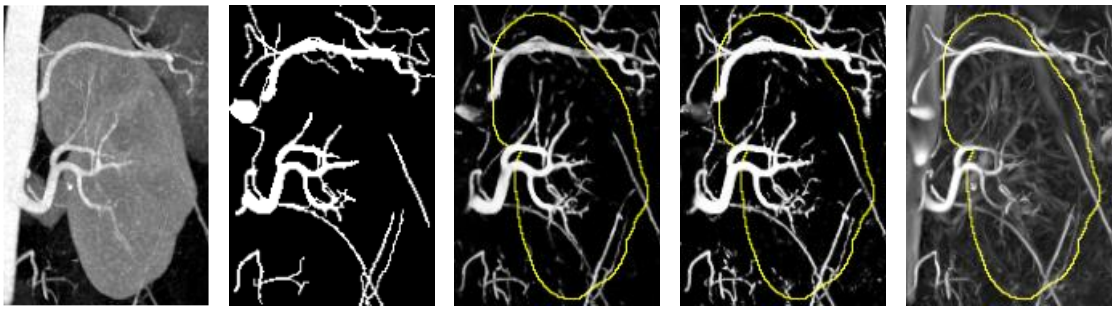


Figure 2.4: Comparison of the three methods for the renal ROI. (a) Box-and-whisker plot comparing the AUPRCs of the CNNs and HM. The AUPRC values of the representative examples shown in Figures 2.10 (a), 2.10 (b), 2.10 (c) are plotted as examples 1, 2, and 3. (b) Average Precision-Recall curves of the methods.

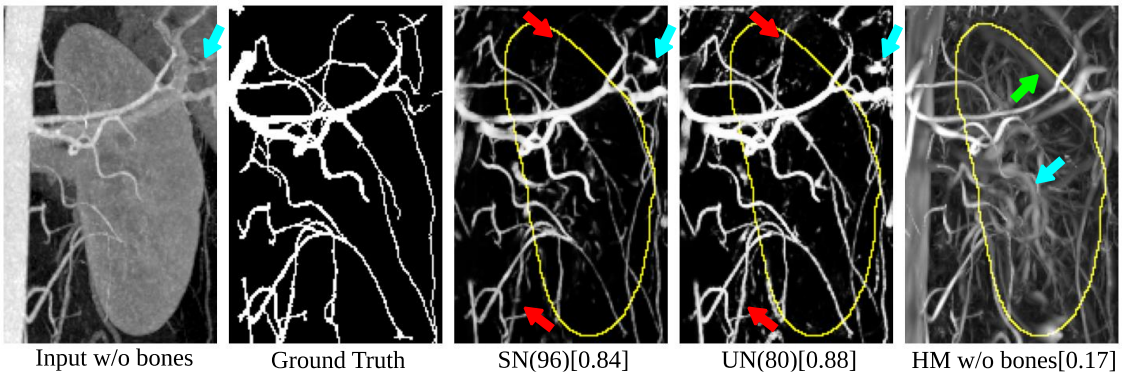
Figure 2.5 shows a qualitative comparison of the methods for the renal ROI results. The CNNs were able to enhance the arterial regions while suppressing the non-arterial regions, such as kidneys and veins. No major difference was observed the results of the between the results of the UN and SN. However, the UN results tended to have better accuracy for the thinner arteries as indicated by the red arrows. Some veins were mistakenly enhanced in the CNN models (cyan arrows). HM enhanced not only arteries but also veins (cyan arrow) and boundaries of renal cortex and medulla (green arrow). Note that HM w/o bones results in Figure 2.5 are shown with bone regions removed just before the visualization and no bone removal was performed in the quantitative evaluations. The HM w/o bones results are shown because the false positive detection of the ridges of the bones hides the vessels in the results and it is relatively easy to remove bones.



(a) Example 1.



(b) Example 2.



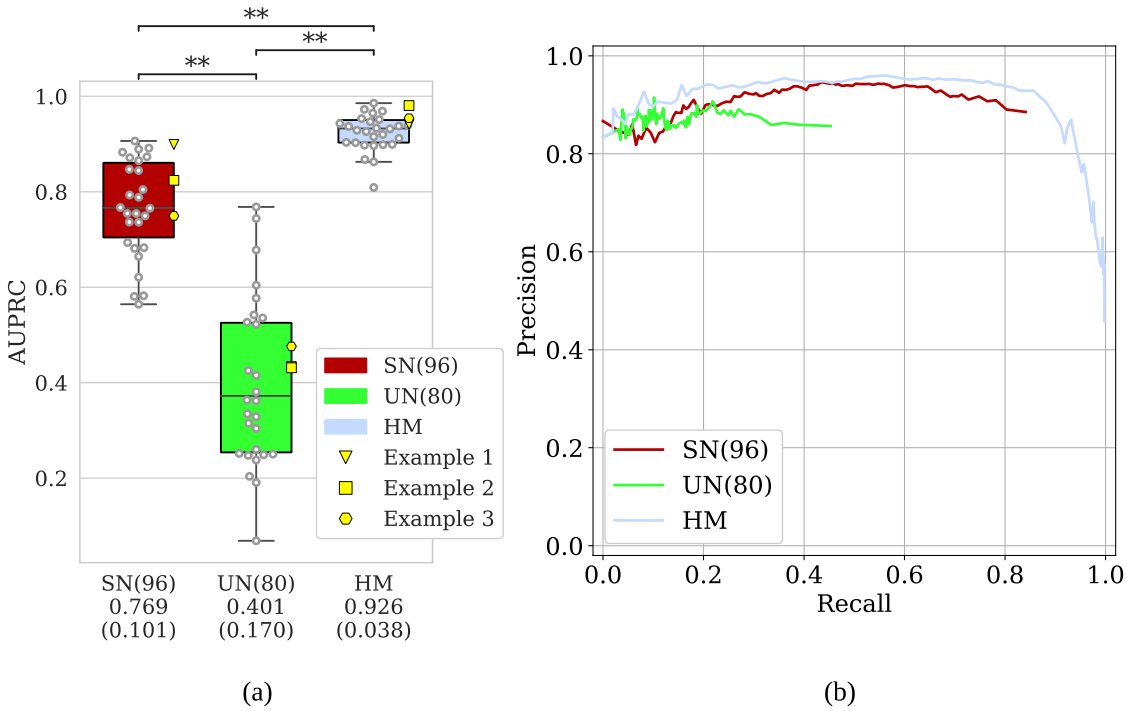
(c) Example 3.

Figure 2.5: MIPs of the renal ROI in the coronal plane for the Example 3 in Figure 2.3 (a). Original input image (Input w/o bones), ground truth (Truth), and results of the three methods (SN, UN, and HM w/o bones) are shown. Regarding the original input and HM result, bone regions were removed just before the MIP operation for better understanding of the results. Yellow lines indicate the contour of the kidney. AUPRCs are shown in brackets.

2.4.2 Evaluations for the lung ROI

As shown in Figure 2.6 (a), the HM had the best score in the lung ROI. As shown in Figure 2.6 (b), HM obtained better results both in recall and precision. For the lung ROI, UN yielded the worst AUPRC, and its recall score was much lower than that of other two methods.

In this evaluation, the lung vessels were defined as targets for detection. Therefore, detected lung vessels were counted as TP even though lung vessels are not a renal artery or any of the abdominal arteries.

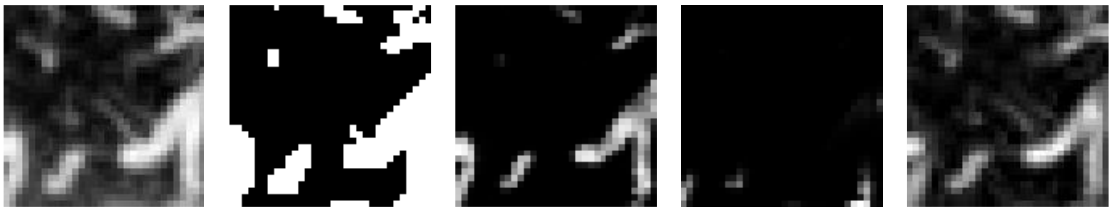


(a) (b)

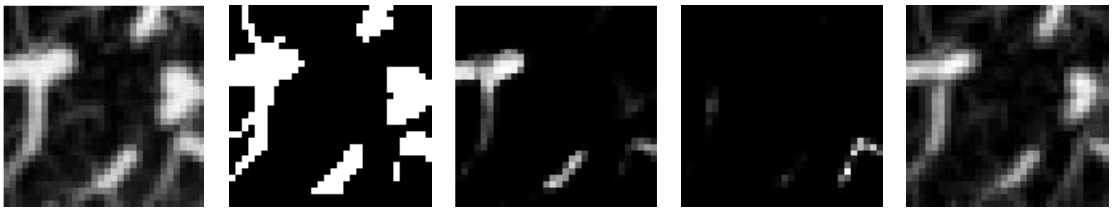
Figure 2.6: Quantitative evaluation of lung ROI. (a) Box-and-whisker plot comparing AUPRC of CNNs and HM. AUPRC values of representative examples shown in Figures 2.10 (a), 2.10 (b), 2.10 (c) are plotted as examples 1, 2, and 3, and the numbers correspond to the examples in Figures 2.10 (a), 2.10 (b), 2.10 (c). (b) Average Precision-Recall curves of the methods.

Figure 2.7 shows the qualitative comparisons for the lung ROI results. In accordance with the quantitative results, the UN removed most of the lung vessels whereas the SN retained some of the vessels in the results. As noted in the quantitative results, if lung vessels were not targets for detection, the UN would have had the best results because there was almost nothing to detect in the lung ROIs.

(a) Example 1.



(b) Example 2.



(c) Example 3.

Figure 2.7: Qualitative illustration of the results in the lung ROI. Results are shown as slab MIP in the axial plane. Original input image (Input), ground truth (Truth), and results of the three methods (SN, UN, and HM) are shown. AUPRCs are shown in brackets.

2.4.3 Evaluations for the spine ROI

As was the case with the evaluations for the lung ROIs, arteries in the spine ROIs are considered targets for detection. A notable difference between lung ROIs and spine ROIs is that the arteries (intercostal arteries) are more similar to renal arteries.

Figure 2.8 shows the quantitative comparisons for the spine ROI results. For the spine ROIs, the SN and UN obtained equally good results (Figure 2.8 (a)). The HM yielded low precision and there were too many false-positive enhancements.

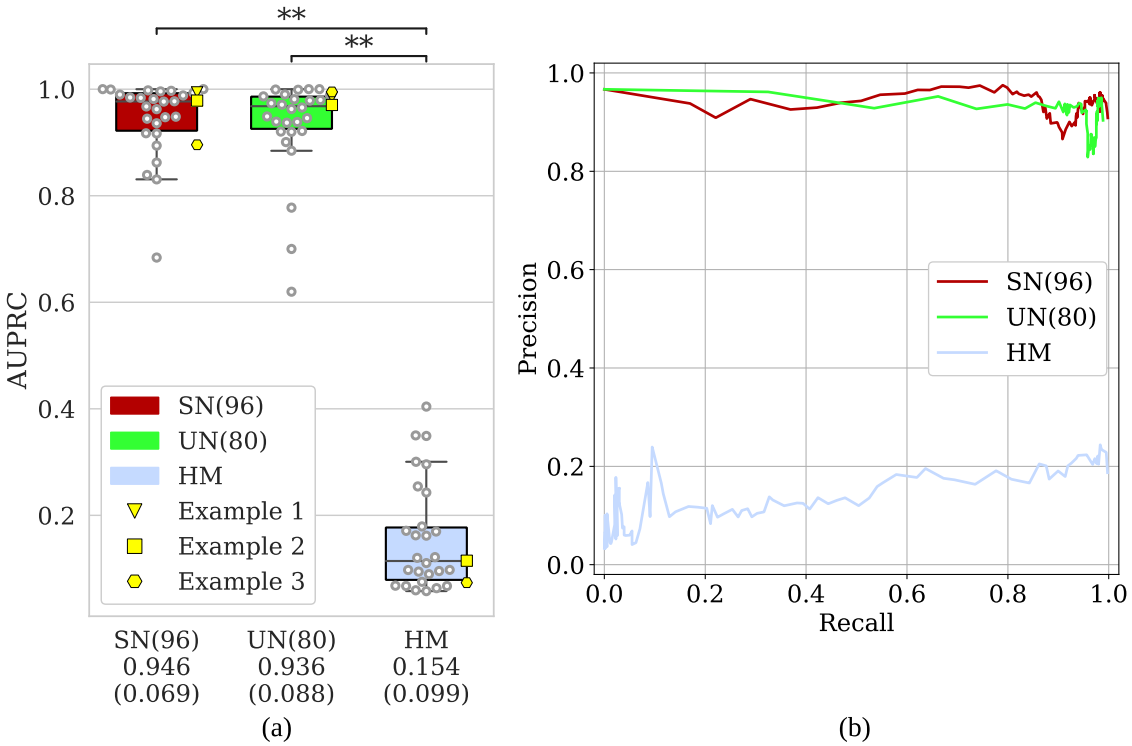
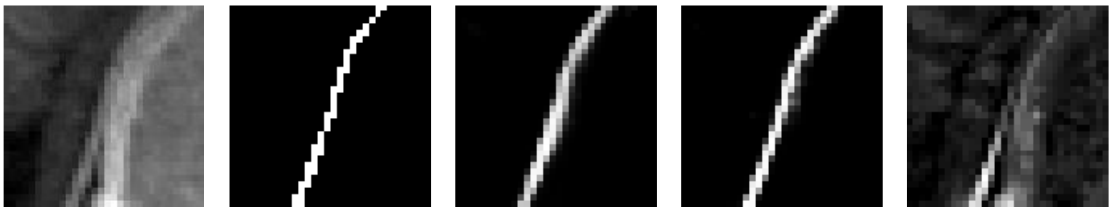


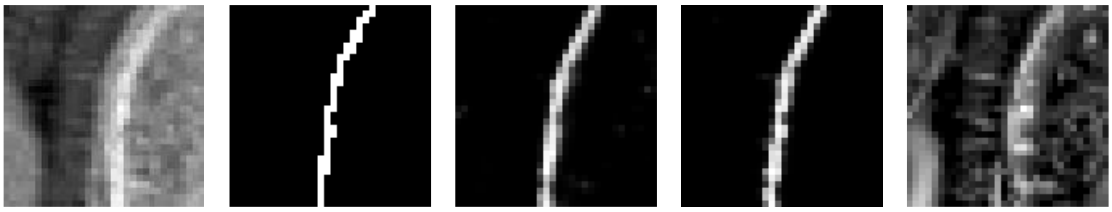
Figure 2.8: Quantitative evaluation of the spine ROI results. (a) Box-and-whisker plot comparing the AUPRC of the CNNs and HM. The AUPRC values of representative examples shown in Figures 2.10 (a), 2.10 (b), 2.10 (c) are plotted as examples 1, 2, and 3. (b) Average Precision-Recall curves of the methods.

Figure 2.9 (a) shows a quantitative comparison of the spine ROI results. In this example, CNNs achieved good accuracy on the intercostal artery with no false positive enhancement of the spinal bone. The HM had a strong response at the edge of the spinal bone. This is because cortical

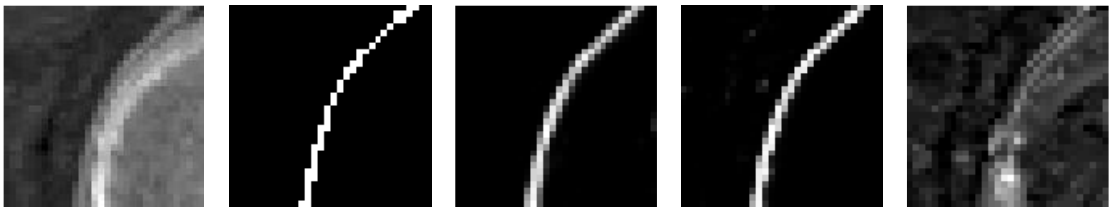
bones have an intensity distribution similar to the distribution that the HM is mathematically designed for.



(a) Example 1.



(b) Example 2.



(c) Example 3.

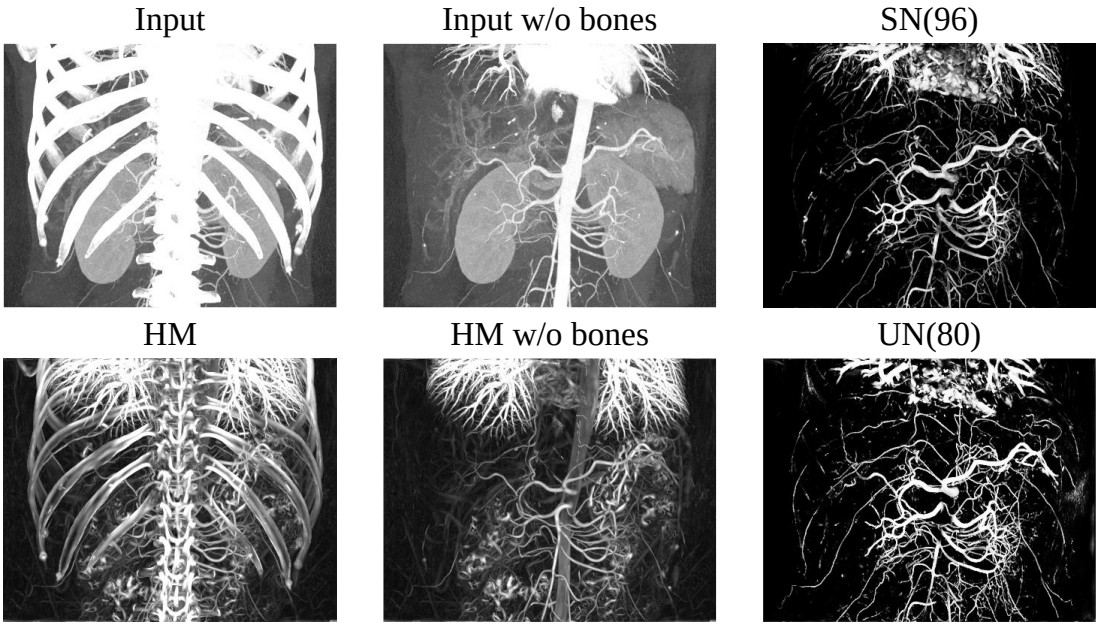
Figure 2.9: Qualitative illustration of results for the spine ROI. The ROIs are shown as slab MIPs in the axial plane. Original input image (Input), ground truth (Truth), and results of the three methods (SN, UN, and HM) are shown. AUPRCs are shown in brackets.

2.4.4 Overall image evaluations

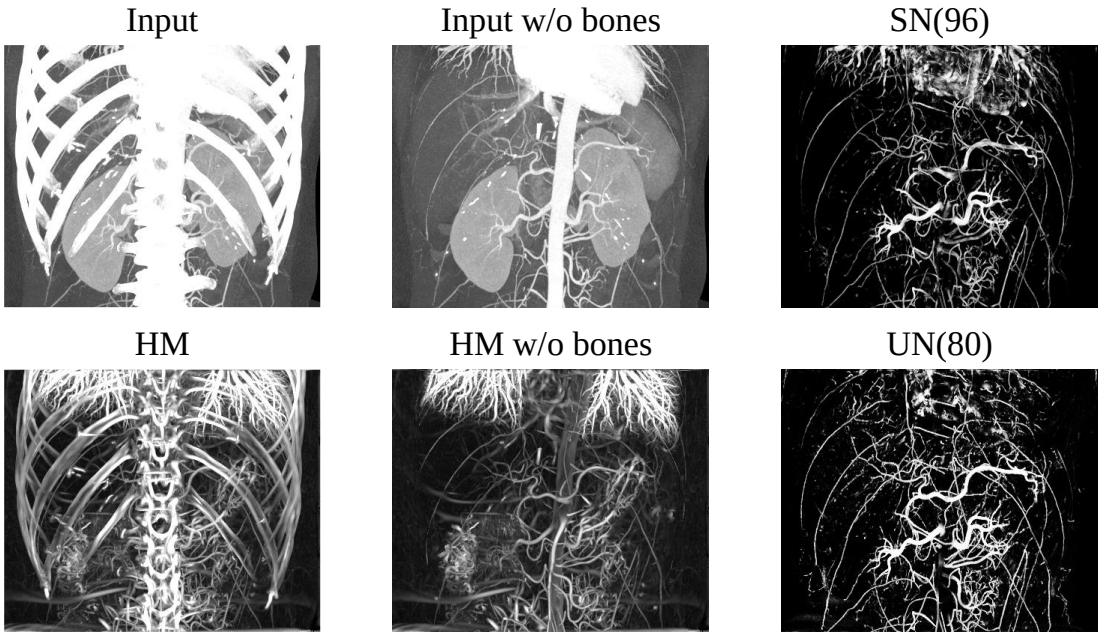
Figure 2.10 shows a qualitative comparison of the results of the entire image. The differences between the UN and the SN results are as follows.

1. In the lung region, the SN enhances more vessels than the UN.
2. Thin vessels, such as the intercostal artery, are better enhanced by the UN.

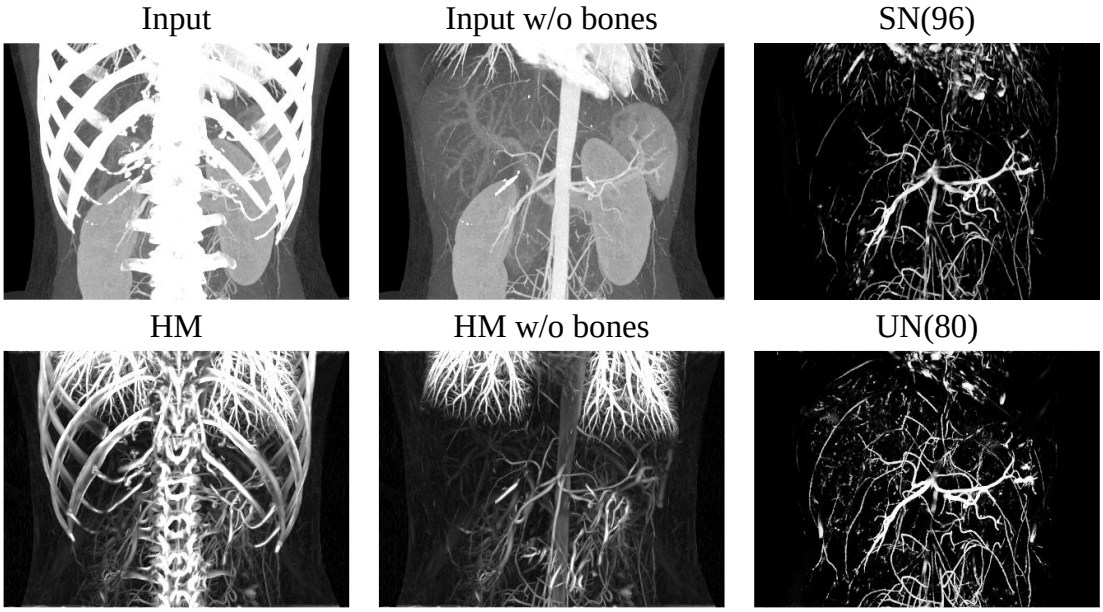
The HM has the best result for the lung vessels of the three methods, although many artifacts originating from bones and other non-arterial organs are included. The same bone removals as Figure 2.5 were applied to the HM results in Figure 2.10 only for the visualization, which has no effects on the any quantitative results. As shown in Figure 2.10, the HM w/o bones results were good when artifacts due to the bones were removed. However, abdominal arteries were not selectively enhanced; instead, abdominal veins and lung vessels were also enhanced.



(a) Example 1



(b) Example 2.



(c) Example 3.

Figure 2.10: Qualitative illustration of entire images of Example 1 in Figures 2.4, 2.6, 2.8. Here, 3D volumetric images are reduced to 2D images using MIP in the coronal plane. For better visualization, MIP image with bones masked out are also shown (Input w/o bones and HM w/o bones). Note that the bones were only masked for the visualization; the original images were used as the input for each method.

2.4.5 U-Net depth

Figure 2.11 shows the results of three UNs (UN (default setting), shallower UN, and deeper UN) in the renal ROI. Note that we were able to run deeper UN only with the patch size 96 because deeper UN has too many convolution and max pooling layers for smaller patch sizes. As shown in Figure 2.11, the UN (default setting) was better than shallower and deeper UNs in all patch sizes and both shallower UN and the UN (default setting) had the peak accuracy with patch size 80.

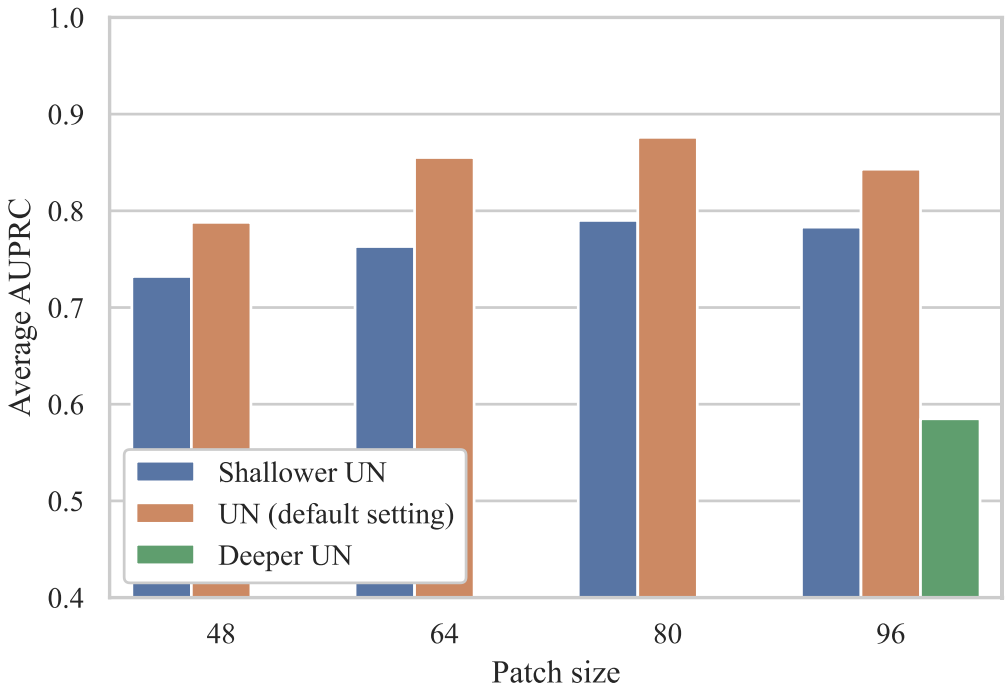
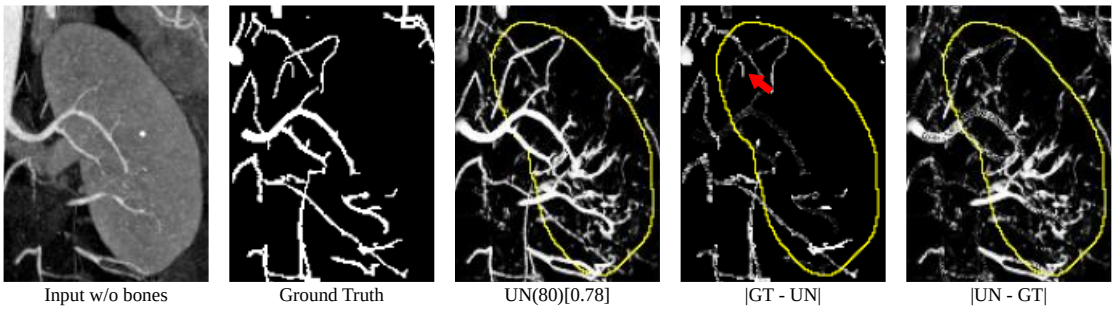


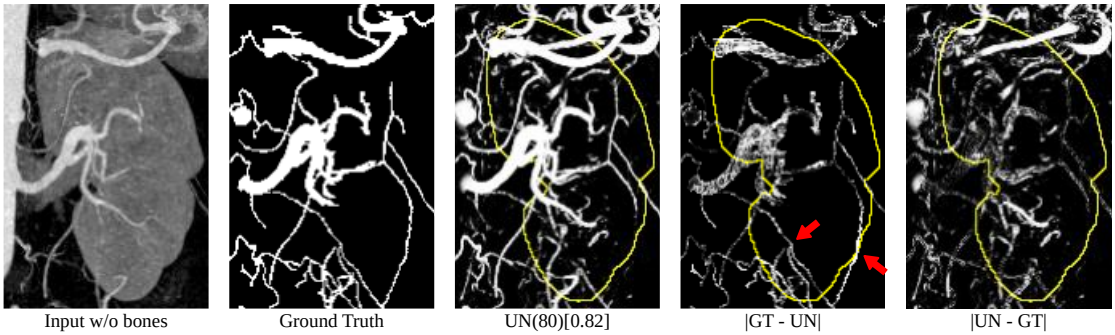
Figure 2.11: Average AUPRC of UN variants with different depths and patch sizes.

2.4.6 Worst cases

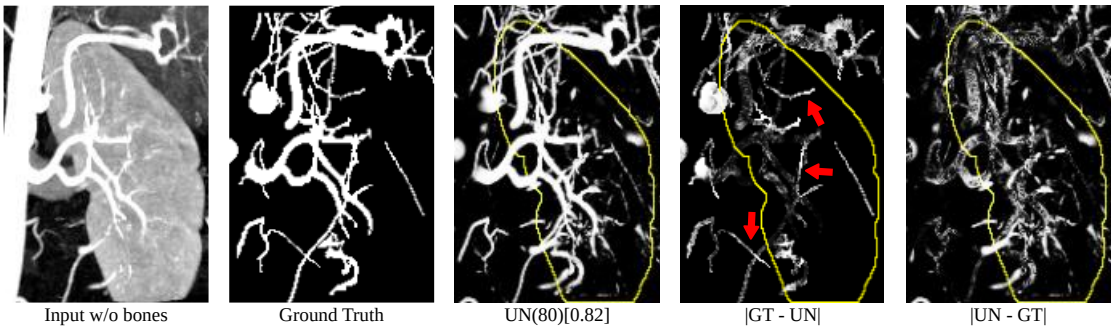
Although main goal of this study is to compare various methods, worst cases for UN(80) are listed in Figure 2.12 to investigate how to improve the segmentation accuracy in the future In Figure 2.12, the same visualization methods as Figure 2.5 were applied to the cases. In addition to the UN(80) results, post-processed images of UN(80) results are also shown to emphasize the differences between the ground truth and the results. The |GT - UN| results indicate false negative arteries missing in the UN(80) results while the |UN -GT| results indicate false positive arteries.



(a) The worst case.



(b) The second worst case.



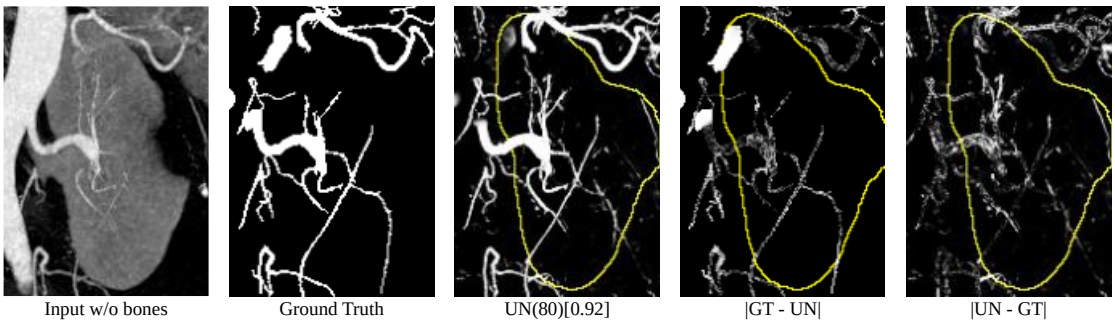
(c) The third worst case.

Figure 2.12: The worst cases for UN(80). Red arrows indicate notable false negative arteries.

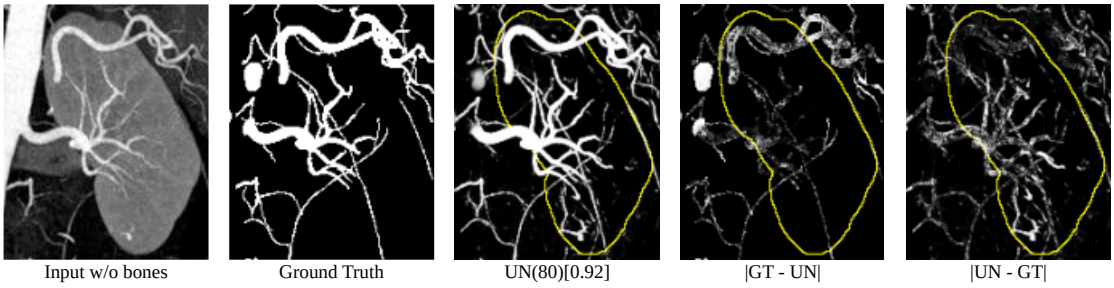
2.4.7 Best cases

The best cases for UN(80) are shown in Figure 2.13 to visualize the differences between easy cases and hard cases. Compared to the hard

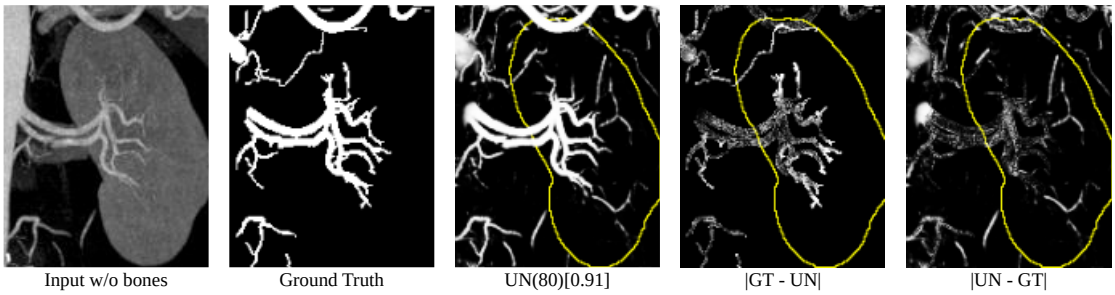
cases shown in Figure 2.12, the easy cases shown in Figure 2.13 tend to have less thin arteries. Although there are some thin arteries, most of them have relatively simple structure with simple surrounding tissues.



(a) The best case.



(b) The second best case.



(c) The third best case.

Figure 2.13: The best cases for UN(80). The same visualization methods as Figure 2.5 were used. Red arrows indicate notable false negative arteries.

2.5 Discussion and conclusion

2.5.1 Discussion

In this study, we experimentally investigated the impact of patch size and network architecture on vessel detection and discrimination accuracy in 3D CT data. Two patch-based 3D CNNs, that is, a U-Net-like architecture and a sequential architecture, were investigated. In addition to the CNNs, a conventional HM was also tested.

The characteristics of the two CNNs and HM are summarized below. First, the UN had the characteristics of high selectivity when extracting particular vessels such as abdominal and intercostal arteries, which have similar features to the renal arteries in the training data. Therefore, the UN showed the best results on abdominal artery extraction when trained on renal artery data, and it will be suitable for the selective extraction of vessels similar to those of the training data. The patch size was needed to be fine-tuned in UN. Second, the SN had similar characteristics to that of the UN, but its selectivity to particular vessels was less strong. The patch size needed to be tuned, but its impact on accuracy was less sensitive than in the UN. Third, HM produced many false positives in renal and spine ROIs, but it should be noted that HM showed the best performance in lung ROI and recall (the true positive rate) was better than or equivalent to the CNN methods in all three ROIs in our experiments. The advantages of the HM are that it does not require training data and it responds to vessels generally (while producing false positives), therefore it will be still useful as a general-purpose method when training dataset is not available.

As shown in Figure 2.3, the patch size had a significant influence on the detection accuracy of patch-based 3D CNNs. The optimal patch size depended on the CNN architecture. Therefore, it is necessary to adjust the patch size to achieve better accuracy in vascular detection.

The UN and SN had different responses to the input patch size. The UN was sensitive to patch size because it had a peak accuracy at a patch size of 80 whereas the SN was less sensitive to patch size with a plateau in accuracy. In previous papers, most studies [49]–[51] used a fixed patch size in a patch-based CNN, and the impact of patch size has not been investigated in detail. Among them, some papers [47], [53], [65], [66] have addressed the impact of patch size. Two of them [65], [66] showed that large patch sizes always improved accuracy. In these papers, however, the patch was not 3D, instead patch-based 2D or 2.5D CNNs were used. Yang et al. [53] addressed a 3D CNN and showed partly similar results to those of our work. However, the work addressed catheter detection from 3D ultrasound data. The catheters had a fixed width, and the impacts of patch size were not analyzed in detail. In contrast, our study presents detailed experimental results on the impact of the patch size in 3D, including the different characteristics of different CNN architectures. Patel et al. [47] reported that using large patch sizes increased the accuracy of cerebrovascular segmentation using 3D CNN with a multi-resolution strategy. However, they were not able to show optimal patch size with a peak or plateau in accuracy due to the hardware restriction.

As shown in Figure 2.4, network architecture significantly changes the detection accuracy. Although the UN had the best results for the renal

ROI, other methods had better or equivalent result in other regions. Because the SN had a better generalization ability, it can be used when only limited training data are available but generally applicable vessel extraction is needed. The HM responds to any vessels, but the responses largely depend on image contrast.

Although there are numerous new network architectures designed for semantic segmentation in medical images [67]–[71], in this study, we focused on two of the most fundamental network architectures. We believe it is still valuable to test these two architectures because testing fundamental architectures gives us insights to design or assess new architectures, which are typically derived from these fundamental architectures. Future work will include a further comparative study for these recent architectures.

As shown in Figure 2.11, deeper UN performed significantly worse compared to other two UNs in our experiments. Before conducting the experiments, we expected deeper UN performs better because deeper networks are able to capture larger context, which usually helps to distinguish objects. One potential explanation why deeper UN failed is that it is harder to train deeper UN because deeper UN has more layers and more trainable weight parameters. Another potential reason is that large context is not always as important as local appearance when it comes to recognizing arteries, which typically have elongated structure but are small in diameter. As shown in the result, the depth has more impact than the patch size in the accuracy. Therefore, in hyperparameter tuning, we recommend tuning the depth first and then

tuning other hyper parameters such as patch size for the U-Net like architectures.

Figure 2.12 shows the worst cases for UN(80) results. As indicated by the red arrows in Figure 2.12, some thin arteries were missing in UN(80) results. Although thin arteries are small in volume, missing them can be critical considering their clinical importance. One obvious approach to mitigate this problem is to increase the number of cases in the training dataset. However, this approach is costly because annotating thin vessels are labor-intensive and time-consuming. Another possible approach is to use synthetic dataset to pre-train the models [72]. This approach is much less costly compared to the first one. Using network architectures specialized for vessel segmentation [73] also should be considered.

Because of the limited amount of GPU memory available in our experimental environment, we were not able to conduct experiments with patch sizes larger than 96 voxels. However, as shown in the results, we expect larger patch sizes will result in similar or worse accuracy compared with the accuracies achieved by the best patch sizes in the experiments.

More recently, new methods using deep learning have been proposed for vessel segmentation. Contributions of newly proposed methods can be divided into the following two categories: network architecture and loss function. The methods proposing new network architectures [47], [70] and new loss functions [44], [46] both take advantage of the properties of the vessels. Objective comparison of these new methods is in the future work.

2.5.2 Conclusions and future work

The results of this study show that it is necessary to adjust the patch size to achieve better accuracy in vascular detection in 3D data, especially when the patch-based 3D UN was used. Although we cannot determine a conclusive procedure about the hyperparameter tuning, we recommend tuning the depth of the network first and then tuning patch size because the depth had larger impact on accuracy and the depth tuning seems more insensitive to the patch size variation. Regarding comparison among different methods, UN obtained the best result in the experiments when the task was to selectively extract specific vessels whose features were particularly matched to those of training data. However, the other methods (SN and HM) will be more suitable when the task is to extract vessels more generally.

In this study, we performed experiments on abdominal arteries. Future work will include applications to arteries in other domains and veins, which may have different properties and characteristics such as thickness and curvature.

3 Diffuse Lung Abnormality Patterns Segmentation

3.1 Overview

Computer-aided diagnosis methods that provide semantic segmentation of texture patterns of diffuse lung diseases (DLDs) on chest CT are extremely useful for detecting, identifying, and quantifying lung pathologies. While a fully annotated dataset is desired to build a semantic segmentation model, building such a dataset for DLDs is costly due to the requirements of manual segmentation and certified experts for annotation. Partially supervised learning (PSL) was recently proposed to take advantage of the partially annotated dataset and reduce the full annotation burden. Creating a partially annotated dataset is much inexpensive than creating a fully annotated dataset; therefore, PSL has great potential to build a semantic segmentation model that only requires a feasible amount of annotation.

In this study, a loss function that uses both annotated and unannotated pixels of the partially annotated dataset is proposed as a method of PSL. The proposed loss function is based on the cross-entropy loss and it uses unannotated pixels to penalize the leakage of the segmentation. A parameter that controls the balance between the two types of

supervision is introduced to the loss function to enable tuning and studying the proposed PSL. Experiments using chest CT images of 372 patients were used to investigate the effectiveness and characteristics of PSL for the segmentation of DLDs (consolidation, ground grass opacity, honeycombing, emphysema, and normal).

The experimental results show that the proposed PSL improved the mean Dice score from 0.76 to 0.79 and that a higher value of the balancing parameter increased the precision of the segmentation. Using the proposed PSL, which takes full advantage of the partially annotated dataset, we improved the accuracy of DLD segmentation. Furthermore, the experimental results clarified that the proposed PSL improved the precision of the models using unannotated pixels. Our implementation of the proposed PSL is available at https://github.com/yk-szk/psl-dld.

3.2 Introduction

Diffuse lung abnormality patterns are texture pattern of DLDs seen in medical images such as chest X-rays and CT. Accurately identifying the textual patterns of DLDs is a challenging task even for experienced radiologists [18], [74], and quantifying the volume and distribution of DLDs in a chest CT image is time-consuming because DLDs typically spread over a large area and numerous slices in the lung. Therefore, developing computer-aided diagnosis (CAD) system capable of promptly and objectively assessing the lung with DLDs is of great value. CAD tools for DLDs have been developed for many tasks including

classifying CT slices [75], classifying image patches [76], [77], semantic segmentation at pixel level [78] and prognostic prediction [79].

While machine learning is required for the development of a CAD system, it frequently necessitates the use of annotated datasets, which are not always readily available. Fully convolutional networks[43], [54] (FCNs) are commonly used for semantic segmentation tasks, and FCN models are trained using a fully annotated dataset, in which every pixel of the images in the dataset is annotated. Creating a fully annotated dataset for semantic segmentation is expensive because manual segmentation of images is time-consuming. It is more expensive to create a dataset for DLDs since it requires qualified professionals (e.g., radiologists) and it is difficult to distinguish between different DLD texture patterns with normal and unusual appearances. Therefore, creating a large-scale fully annotated dataset for DLDs is almost impossible.

In most cases, there is a trade-off between accuracy and the amount of annotated dataset used training for machine learning. Unsupervised learning is on one end of the trade-off because it does not require an annotated dataset. However, its accuracy is not as good as supervised methods because no supervision is provided during the training [80]. Fully supervised learning, on the other hand, is on the opposite end of the trade-off with the best performance of all learning methods. However, it requires a fully annotated dataset. Therefore, learning methods with adequate performance and minimum amount of annotation are sought to build a practical CAD system for DLDs. PSL [81] has recently been proposed to take advantage of a partially

annotated dataset. PSL is a promising technique for DLD segmentation because it only requires partial annotation while supervising effective training. Therefore, we focus on this learning technique in this study.

Other than PSL, several strategies have been developed to reduce the need for annotated dataset [82]. Unannotated data approaches like entropy minimization [83] and consistency training are used in semi-supervised learning [84]. Weakly supervised learning [85] allows semantic segmentation to be performed using weak annotation, such as image-level annotation[86], [87]. Transfer learning, also known as fine-tuning[89] is a strategy for a pre-training a model with a large-scale dataset to achieve a well-generalized model even with a limited target dataset [90], [91] for the application's domain. PSL and these approaches are not mutually exclusive and can be used together when appropriate.

PSL has been effectively used to various applications, including abdominal organ segmentation [92]–[94] and brain tissue segmentation[81], [95]–[97]. More recently, PSL begins to be applied to segmentation of tumors and other disease regions, which are more complicated than organ regions [98], [99]. However, no application for DLDs has been reported yet. Applying PSL to DLDs can be more effective than applying it to regular anatomical structures like organs and the brain because of the differences in characteristics between DLDs and regular anatomical structures. First, DLDs can appear anywhere in the lung, making it harder to split input images into the regions of interest (ROIs) and apply the specific models to each ROI. Second, a typical DLDs patient only has a couple of DLDs, unlike organ

or tissue segmentation tasks, where a patient has a complete set of organs without missing ones. This makes it harder to construct an ensemble model composed of partial models using subsets of the dataset because each model cannot learn DLD classes absent in the subsets.

Our contribution is two-fold. 1) Proposing a method of PSL that uses both annotated and unannotated pixels and applying it to DLDs to show the applicability of PSL. 2) Studying the characteristics of the proposed method by introducing a parameter that controls the balance of the supervision. We previously presented early results before [36], however in this study, we report additional evaluation results using samples of fully annotated images as well as tests with wider range of the balancing parameter.

The DLD patterns, including the healthy lung considered in this study, are consolidation (CON), ground-glass opacity (GGO), honeycombing (HCM), emphysema (EMP), and normal (NOR). Throughout this study, each DLD pattern is represented or superposed in the following colors (CON: cyan, GGO: yellow, HCM: red, EMP: green, NOR: brown.) The remainder of this study is organized as follows. In the materials and methods section, partially annotated datasets and PSL are explained. The results of the experiments involving chest CT images of 372 patients are shown in the results section. The discussion section discusses the advantages and characteristics of the proposed method found in the experiments.

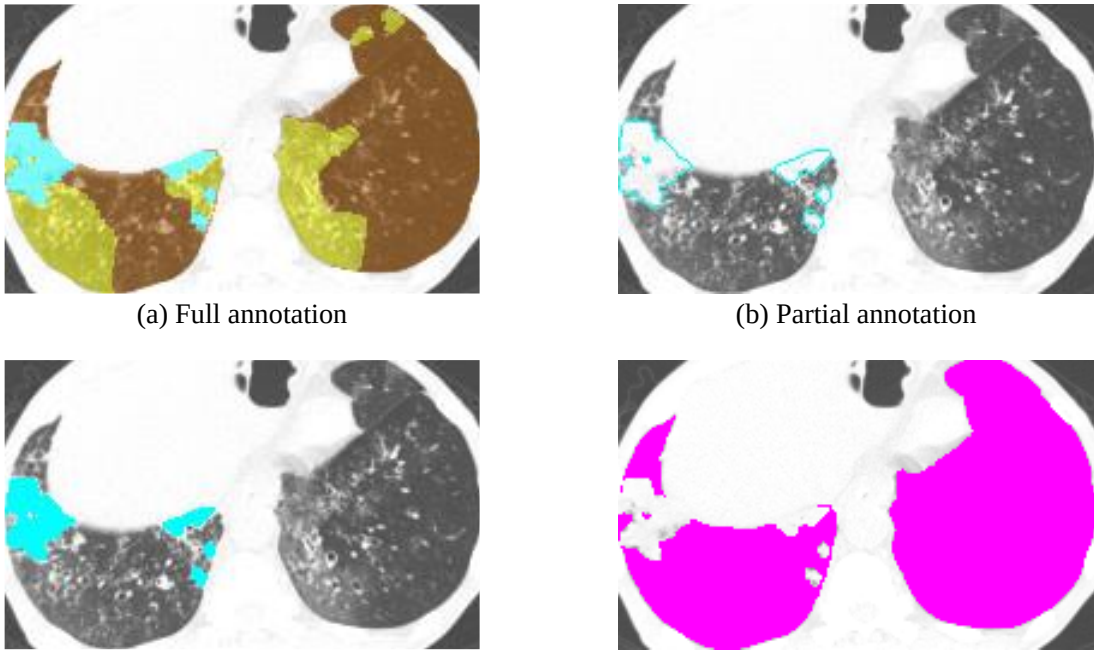
3.3 Materials and Methods

3.3.1 Materials

The dataset used in this study consisted of chest CT images with a matrix size of 512 × 512 taken in Yamaguchi University Hospital, Japan. CT images of 372 patients were collected. The pixel size mean and the standard deviation were 0.684 mm and 0.0517 mm, respectively, and slice thickness was 1 mm. In the experiments, no pixel size equalization was performed because the deviation in the pixel sizes was negligibly small.

Only one or a subset of the potential classes for an image are annotated in a partially annotated dataset. Figure 3.1 shows a comparison between full annotation and partial annotation. Figure 3.1 (a) shows a fully annotated image, in which every pixel in the lung is labeled with one of the possible classes $L_{positive} = \{CON, GGO, HCM, EMP, NOR\}$. Figure 3.1 (b) and (c) show partial annotation and annotated pixels of an image, in which just one class (CON) was chosen for the image and annotated while other regions with other classes (GGO or NOR) were left unannotated. In Figure 3.1 (b), the delineated region indicates the CON region, which derived annotated pixels shown in Figure 3.1 (c). Partial annotation is easier and takes less time than full annotation, because with partial annotation, annotators can focus only on one label, and the area required for annotation is much smaller. Figure 3.1 (d) shows an unannotated region of a partially annotated image. In this image, regions that are not CON are left unannotated, and therefore virtually annotated as ¬CON = {GGO, HCM, EMP, NOR} region. Note that

the fully annotated image shown in Figure 3.1 (a) is created only for illustration and a fully annotated image like this is not available for the training in this study. Although annotation can be a subset of possible classes, we only annotated one class per image in this study.



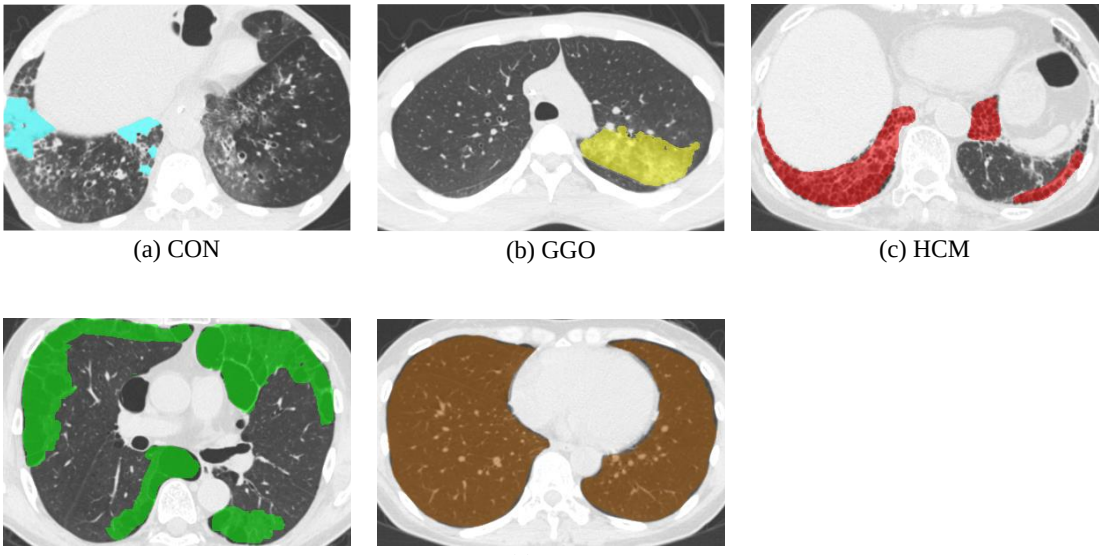
(a) Full annotation (b) Partial annotation

(c) Annotated pixels (d) Unannotated pixels

Figure 3.1: Comparison between fully and partially annotated images. (a) Full annotation with consolidation (CON), ground-glass opacity (GGO) and normal (NOR) regions are annotated in the lung. (b) Partial annotation with only the CON region is annotated (shown in cyan). (c) Annotated pixels (shown in cyan) derived from (b). (d) Unannotated pixels (shown in magenta), which were implicitly annotated as ¬CON in (b).

Our partially annotated dataset was created in the following steps. First, up to three slices and representing DLD texture pattern for each slice were chosen for the annotation for each CT scan by a board-certified radiologist. Second, three board-certified radiologists performed a

partial annotation of the chosen DLD for every image. Finally, pixels annotated by two or more radiologists were designated as annotated pixels, whereas pixels that were annotated by less than two radiologists were left as unannotated. Note that three radiologists performed annotations to avoid creating a biased dataset. In this study, unannotated pixels are considered annotated as one of the following labels $L_{negative} = \{\neg CON, \neg GGO, \neg HCM, \neg EMP, \neg NOR\}$. In addition to the DLD annotation, lung fields were manually segmented under the supervision of radiologists and training and evaluation were conducted specifically within the lung fields. Figure 3.2 shows examples of partially annotated images for each class of DLDs. Table 3.1 shows the statistics of the partially annotated dataset used in the experiments.



(a) CON (b) GGO (c) HCM

(d) EMP (e) NOR

Figure 3.2: Typical slices of each diffuse lung disease (DLD) class. Annotated labels are superposed in transparent colors.

Table 3.1: Dataset statistics. Note that the total number of cases in the table is more than the number of cases because multiple slices with different DLDs were extracted from a patient. The number of unannotated pixels of each class is the number of pixels of its corresponding negative labels. For example, the number unannotated pixels for CON is the number of ¬CON pixels)

	CON	**GGO**	**HCM**	**EMP**	**NOR**	**Total**
Cases	56	112	51	131	55	405
Slices	150	114	129	163	55	611
Annotated (positive) pixels ($\times 10^3$)	658	1,676	1,345	4,134	2,552	10,365
Unannotated (negative) pixels ($\times 10^3$)	3,878	2,648	2,405	2,710	117	11,759

3.3.2 Methods

The foundational idea of the proposed PSL is to use unannotated pixels to supervise the model loosely. For example, in Figure 3.1 (d), the unannotated region can supervise the model not to output the CON label for the region. This offers the proposed PSL an advantage over training methods that only use annotated pixels. Although unannotated pixels are less informative than annotated pixels, using unannotated pixels approximately doubles the number of pixels involved in the training, as shown in Table 3.1. Note that the total number of cases (405) in Table 3.1 is more than the number of patients (372) because multiple slices of a patient can be chosen for different types of DLD annotation. Figure 3.3 shows an overview of the proposed PSL. As

shown in Figure 3.3, two types of supervision, which use annotated and unannotated pixels, respectively, are applied to the model.

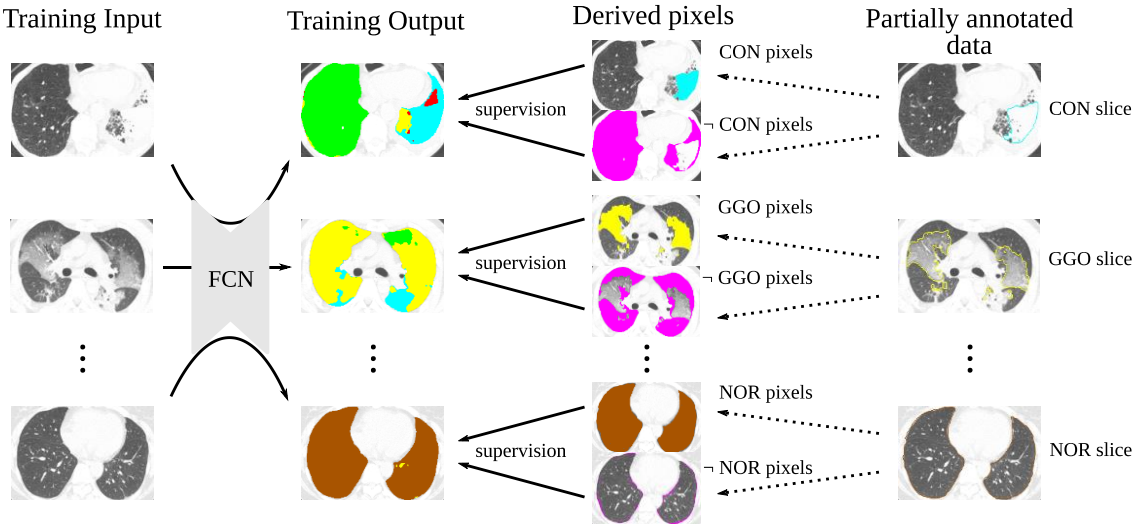


Figure 3.3: Overview of the training process of the proposed partially supervised learning.

Using unannotated pixels is realized by the loss function defined as

$$\mathcal{L}(l,\hat{\mathbf{y}}) = \begin{cases} (1-\lambda) * H(e(l),\hat{\mathbf{y}}), & l \in L_{\text{positive}} \\ \lambda * \dfrac{1}{H(e(l),\hat{\mathbf{y}})}, & l \in L_{\text{negative}} \end{cases} \tag{3.1}$$

$$H(p,q) = -\sum_{x} p\,(x)\log(q(x)),$$

where $l, \hat{\mathbf{y}}$, and $H(p,q)$ denote the ground-truth label, SoftMax output of the model, and cross-entropy for discrete probability distribution p and q, respectively. $e(l)$ is one-hot encoding function that works regularly for $l \in L_{positive}$ (e.g., $e(l_{CON}) = [1,0,0,0,0]$) while for $l \in L_{negative}$, it works so that unannotated pixels get encoded in the same way as the corresponding annotated pixels (e.g., $e(l_{CON}) = e(l_{\neg CON})$). This loss function is based on cross-entropy loss, and works regularly

for annotated pixels, whereas it penalizes unannotated pixels for getting predicted as corresponding annotated labels. λ is a balancing parameter which modifies the balance of the supervision used for the loss of unannotated pixels. The proposed method with the parameter $\lambda = 0$ is equivalent of the training method that only uses annotated pixels.

PSL can be used with any segmentation model and is not dependent on any network architecture. In this study, we used a slab U-Net shown in Figure 3.4 to take advantage of three-dimensional information of DLDs. The slab U-Net is a hybrid of 2D and 3D U-Nets[54], [55]. The input of the network is a 3D slab around an annotated slice with the shape of $6 \times 512 \times 512 \times 1$ and the output shape is $1 \times 512 \times 512 \times 5$, where elements represent the sizes along the z, y, x, and channel axis. In this study, x, y and z axes represent frontal, sagittal and longitudinal axes respectively. Shallower layers of the network consisted of 3D operations such as 3D convolution and 3D max pooling with no padding applied for z-axis, which reduces the data size along z-axis. Operations were switched to 2D ones after the data size along z-axis was reduced to one.

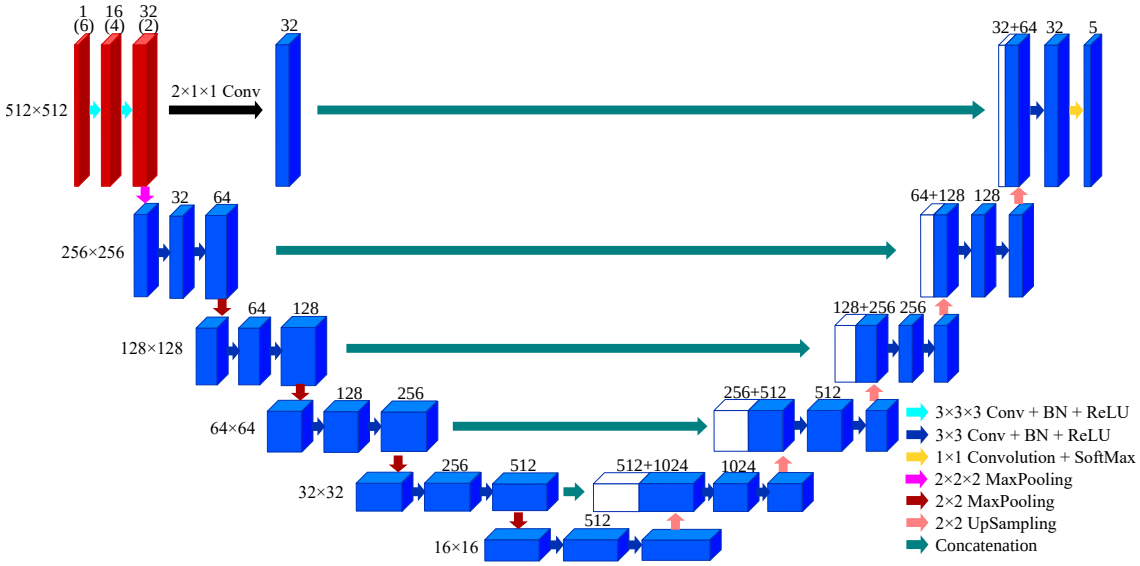


Figure 3.4: The slab U-Net. The red and blue boxes represent three-dimensional and two-dimensional data respectively. The number of channels is indicated by the numbers above the boxes and the parenthesized numbers above red boxes indicate the sizes along z-axis. The white boxes represent concatenated data. At each level of the network, image resolution is indicated on the left side.

3.4 Results

For training and evaluation, a five-fold stratified cross-validation was used. The stratified splitting strategy was adopted to avoid skewed results due to the dataset's uneven split and splitting was patient based to avoid data leakage. Each training subset had a validation subset to determine the number of epochs using early stopping. As data augmentation, geometric transformations (random flip around y-axis, random rotation around z-axis from -15 to 15 degrees, and random rescaling from 90% to 110%) are applied randomly to the input images on the fly during the training. These geometric transformations were applied primarily to increase the variation of body sizes and

orientations in the training dataset but also to help the network to learn more robust feature extractors [100]. Model network weights were optimized using Adam [56] optimizer with default parameters, and losses were computed using Equation 3.1. Our model was built with TensorFlow, and the loss function and network implementation may be seen at https://github.com/yk-szk/psl-dld. As evaluation metrics, Dice similarity score, precision, and recall are calculated for each slice. Dice similarity score (or Dice similarity coefficient) given by

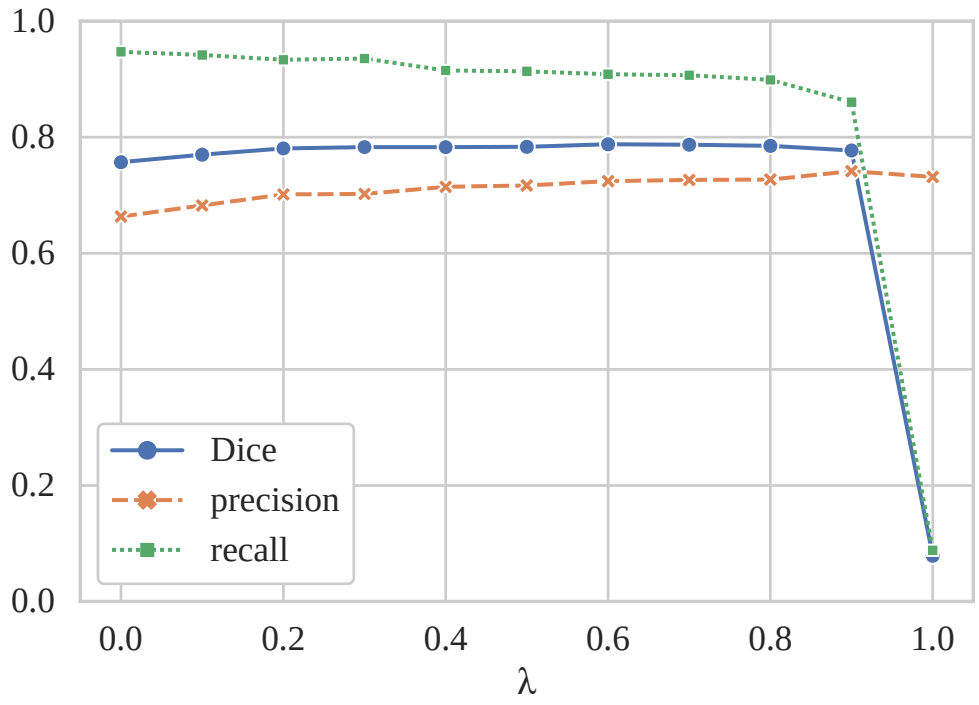
$$\begin{aligned} DSC &= \frac{2|X \cap Y|}{|X| + |Y|} \\ &= \frac{2TP}{2TP + FP + FN} \\ &= \frac{2Precision * Recall}{Precision + Recall}, \end{aligned} \tag{3.2}$$

where $|X|$ and $|Y|$ are the areas of the region X and Y and TP, FP, and FN denote the number of true positives, false positives, and false negatives respectively, is a metric for measuring the amount of overlaps between two regions and also known as F1 score. The training and evaluation were performed on an NVIDIA RTX 3090 with 24 GB of graphics processing unit memory.

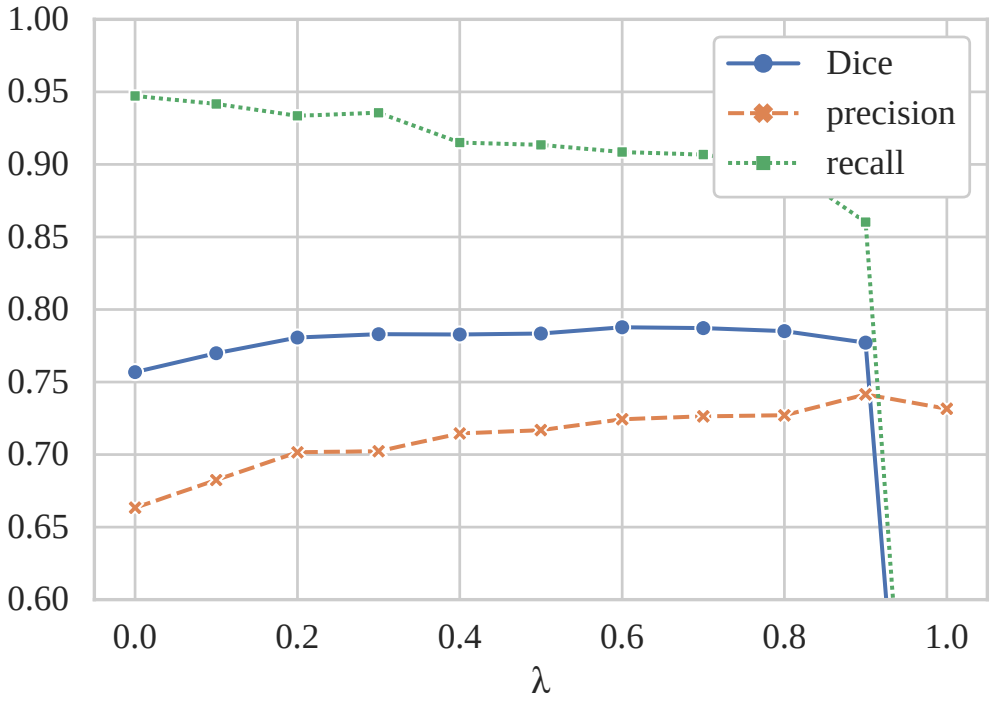
3.4.1 Evaluation on partial annotation

Figure 3.5 shows line plots of evaluation metrics (Dice score, precision, and recall) for the λ values ranging from 0 to 1 with step size of 0.1. The mean Dice score and recall for λ = 1 were very low (0.079 and 0.088 respectively) and plot points for these were trimmed off for the better visualization in Figure 3.5 (b). The segmentation accuracy for λ = 1 was very low because the models were not trained for typical texture

patterns at all. The distributions of the metrics can be seen in the box-whisker plots in Figure 3.6, in which, only λ = {0, .1, .6, .9} were shown for concise visualization. The Wilcoxon signed-rank test was used to test the statistical significances between λ = 0 and the other values. As shown in Figure 3.5, the proposed method with λ between 0 and 1 outperformed the results with λ = 0. Larger λ increases precision while decreasing recall. This is because larger λ penalizes an unannotated region for being segmented as annotated label. The parameter of λ = 0.6 balanced precision and recall well and achieved the best Dice score in our experiments.



(a) Line plots.



(b) Line plots focusing on λ = {0, 0.1, ..., 0.9}.

Figure 3.5: Line plots of Dice score, precision and recall

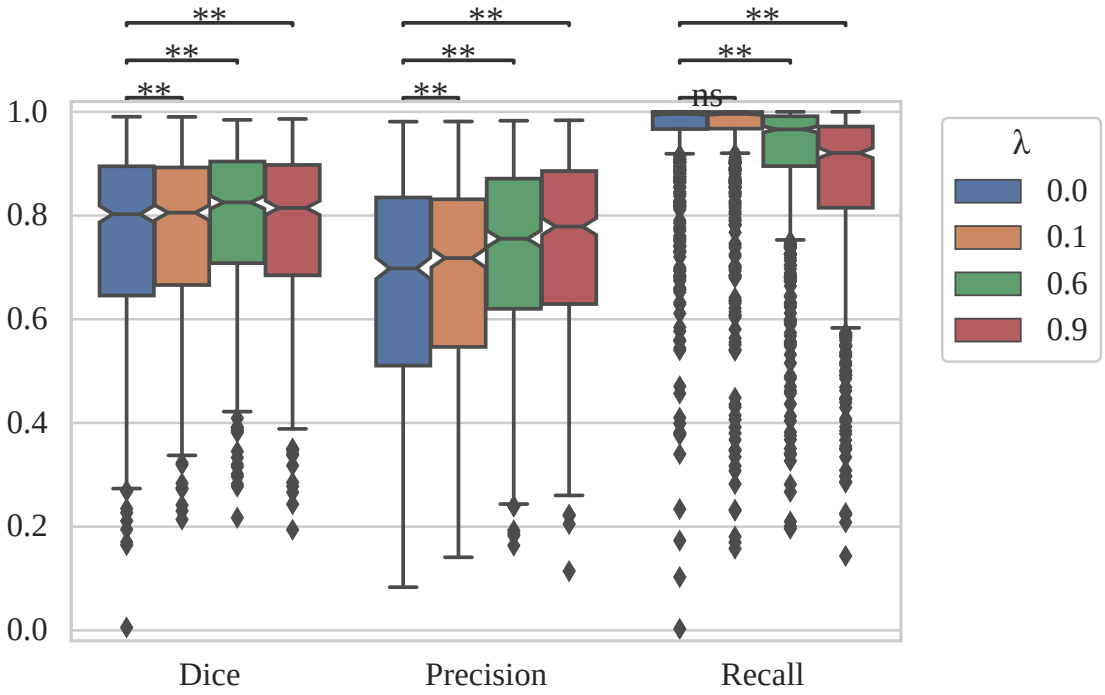


Figure 3.6: Box-whisker plots for parameter comparison. Statistical significances are indicated by ** (:$p < 0.01$).

The results of each DLD are shown in Figure 3.7 and Table 3.2, comparing the proposed method with λ = 0 and λ = 0.6. As shown in Figure 3.7, statistically significant improvement was observed with CON, GGO, HCM, and EMP. There was a statistically significant decrease in the Dice score with NOR, but the difference is smaller compared to other improvements, as shown in Table 3.2. The Dice score with NOR was reduced as not much room was left to improve the precision with NOR in the first place, while other DLDs improved their precision.

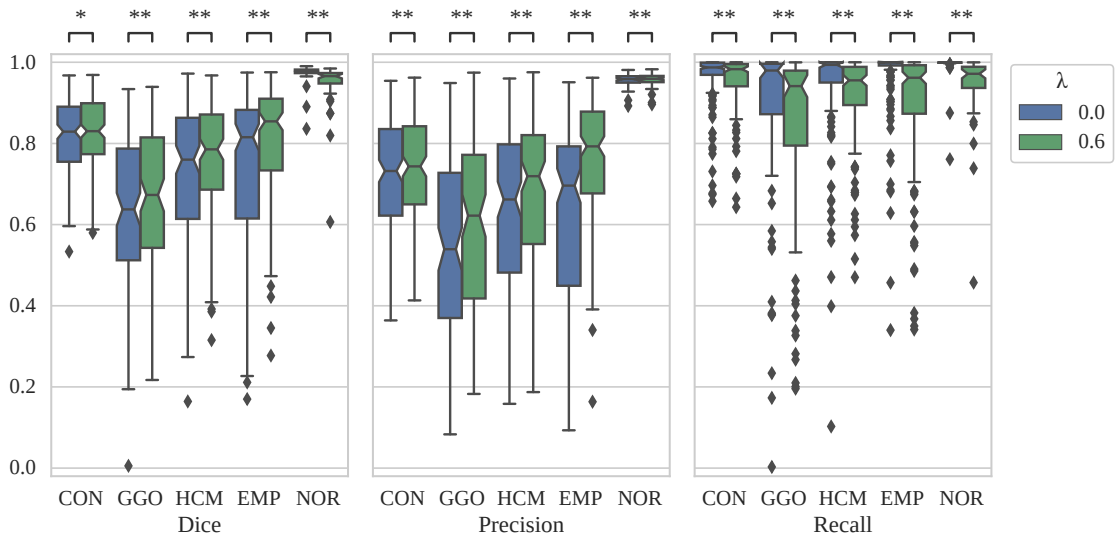


Figure 3.7: Comparison between λ = 0 and λ = 0.6 for each DLD. Statistical significances are indicated by * (:$p < 0.05$) and ** (:$p < 0.01$).

Table 3.2: Dice scores for each class.

	CON	**EMP**	**GGO**	**HCM**	**NOR**	**All**
mean (std)						
λ = 0	0.82 (0.09)	0.74 (0.19)	0.64 (0.19)	0.73 (0.16)	0.97 (0.02)	0.76 (0.18)
λ = 0.6	0.82 (0.09)	0.81 (0.14)	0.67 (0.17)	0.75 (0.15)	0.95 (0.06)	0.79 (0.15)
median						
λ = 0	0.83	0.82	0.64	0.76	0.98	0.80
λ = 0.6	0.83	0.85	0.67	0.79	0.97	0.83

Figure 3.8 shows confusion matrices of the results. Note that these confusion matrices were constructed using only annotated pixels because exact classes for unannotated pixels were unknown. Rows of the matrices were normalized; thus, diagonal elements of the matrix represent recall. Figure 3.8 (a) and (b) show similar patterns of misclassification. As diagonal elements of Figure 3.8 (a) and (b) show, the recall of the results with $\lambda = 0.6$ was lower than that of the results with $\lambda = 0$, which is also shown in Figure 3.7 and Table 3.2. Therefore, more analysis was conducted as follows.

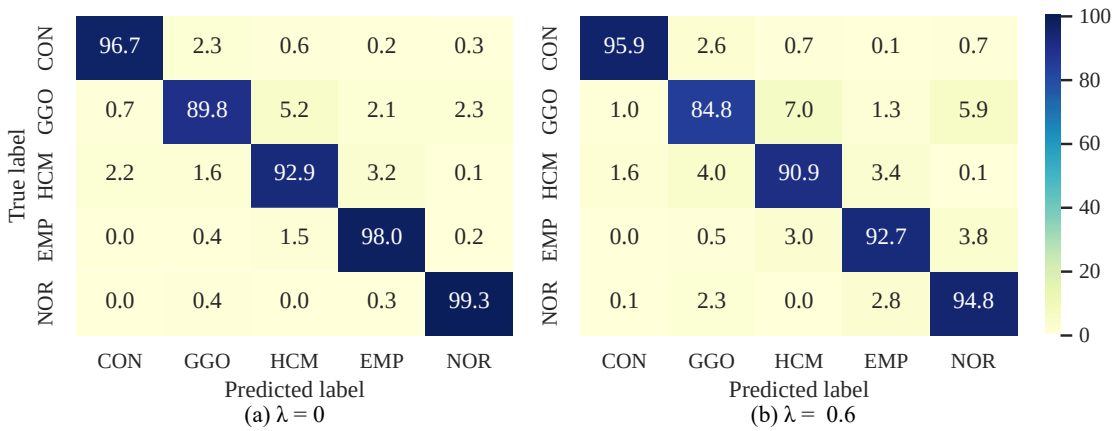
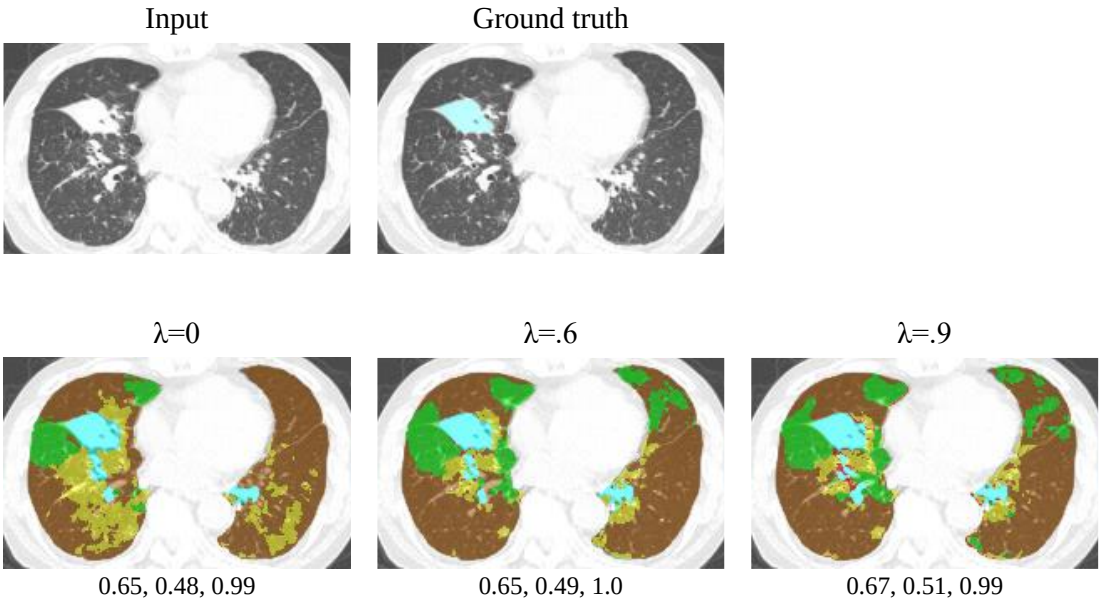
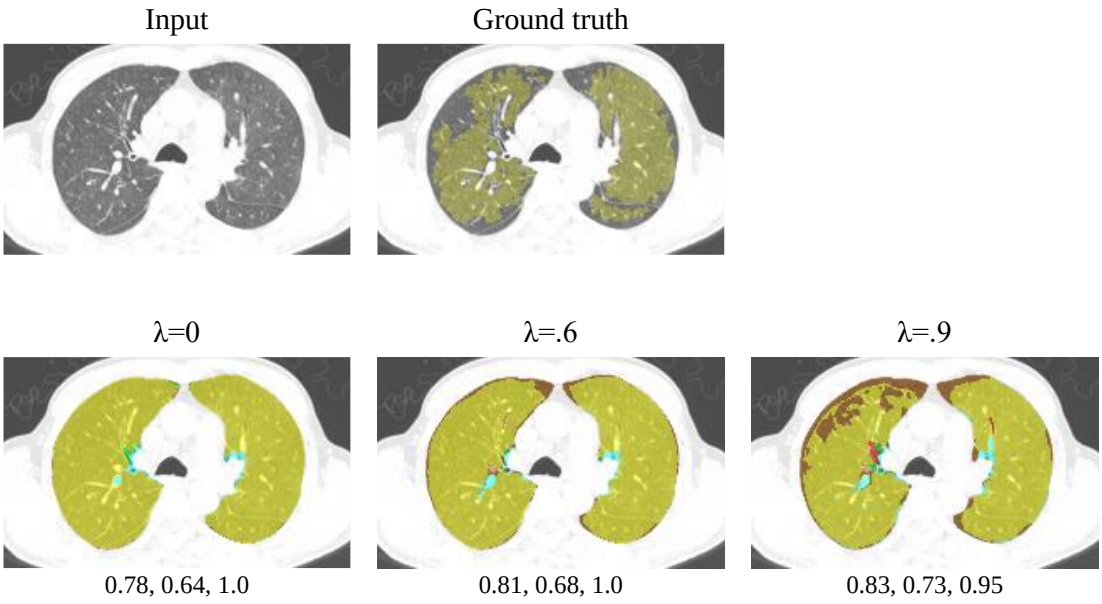


Figure 3.8: Confusion matrix constructed using only annotated pixels. Row-wise normalizations are applied.

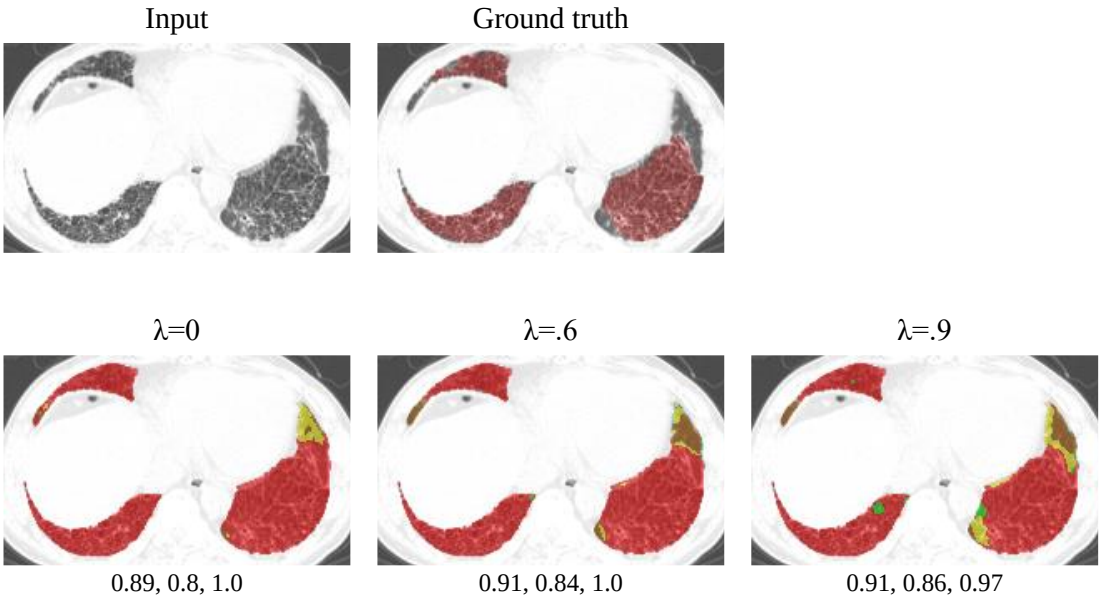
Figure 3.9 shows the typical segmentation results of each DLD. The differences in the Dice scores between the results of $\lambda = 0$ and $\lambda = 0.6$ were used to choose the typical slices. The slice with the median difference in Dice score was chosen as a typical slice for each DLD. As shown in Figure 3.9, the area of representative DLDs (especially GGO and EMP) was smaller in $\lambda = 0.6$ than $\lambda = 0$ and fewer pixels of representative DLDs were leaked over the non-representative DLD region, which was reflected as improvements in precision.



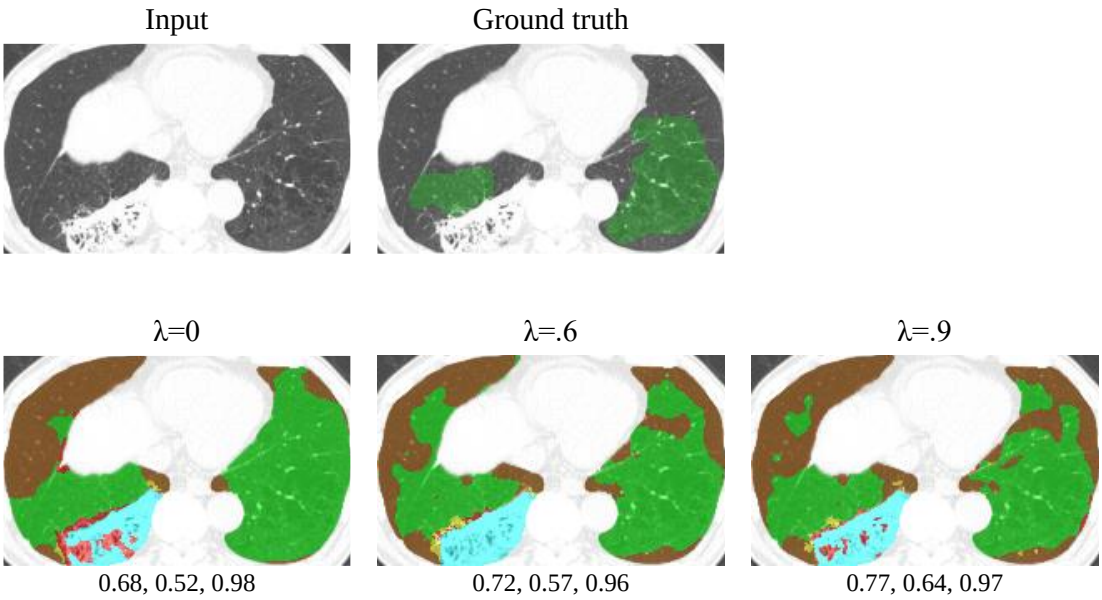
(a) Consolidation (CON).



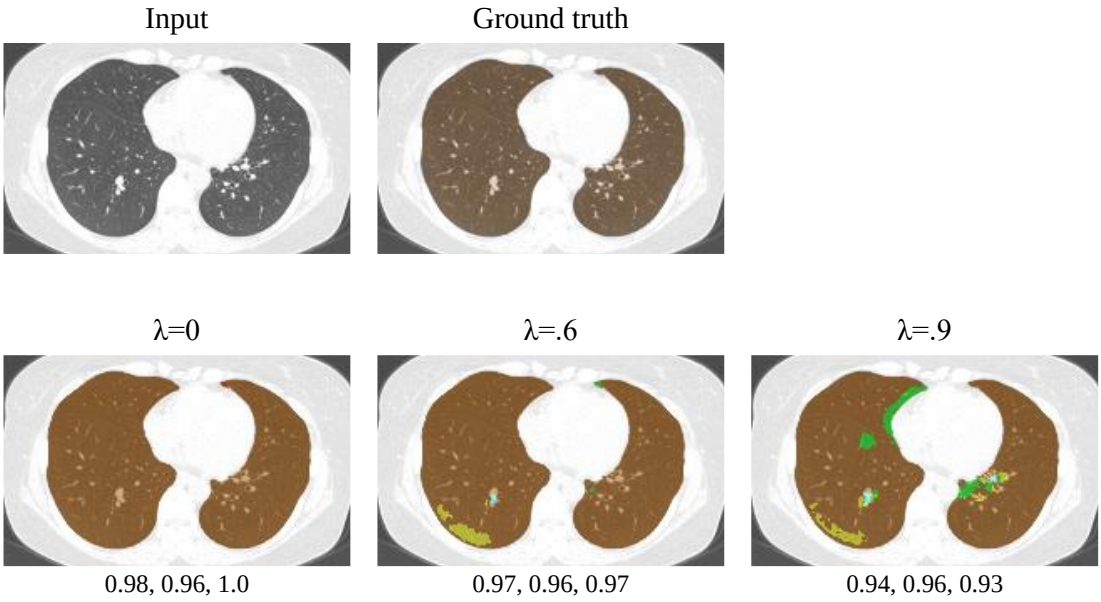
(b) Ground glass opacity (GGO).



(c) Honeycombing (HCM).



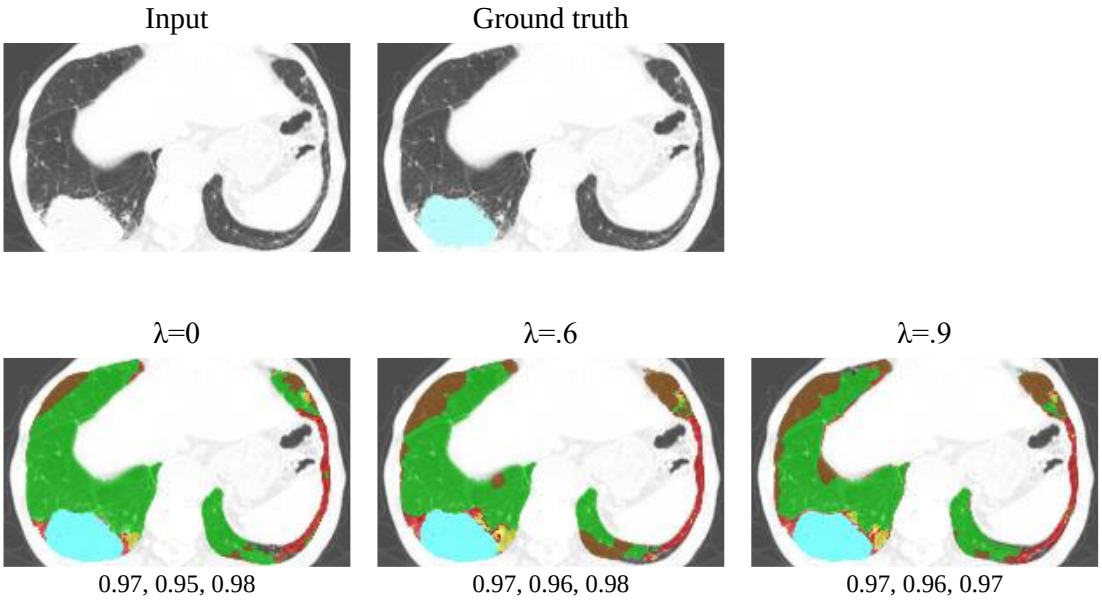
(d) Emphysema (EMP).



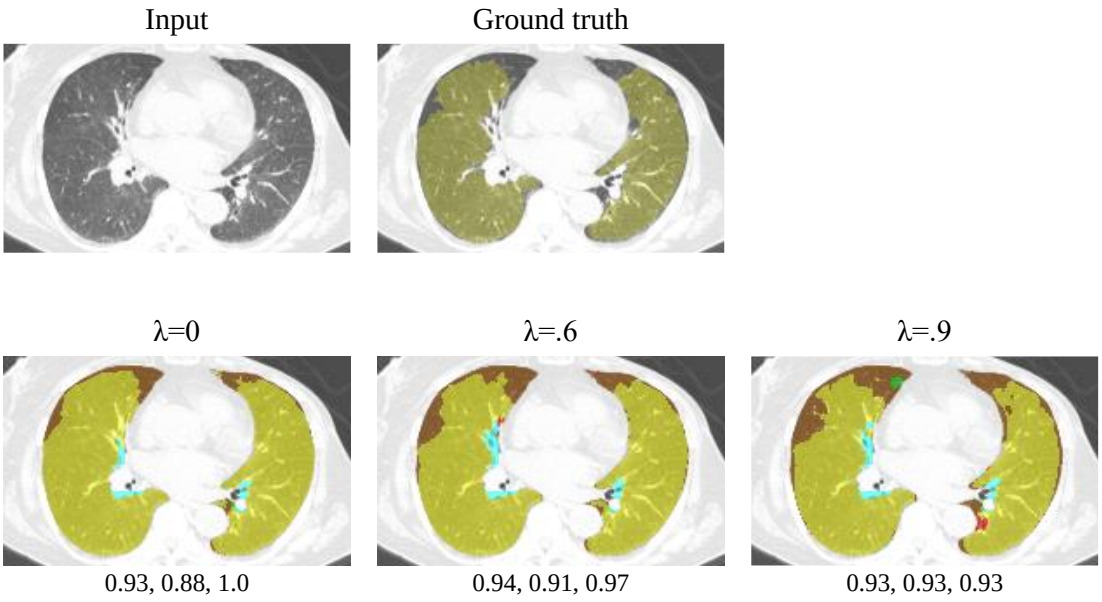
(e) Normal (NOR).

Figure 3.9: Examples of average segmentation results. The numbers below images indicate Dice score, precision, and recall respectively.

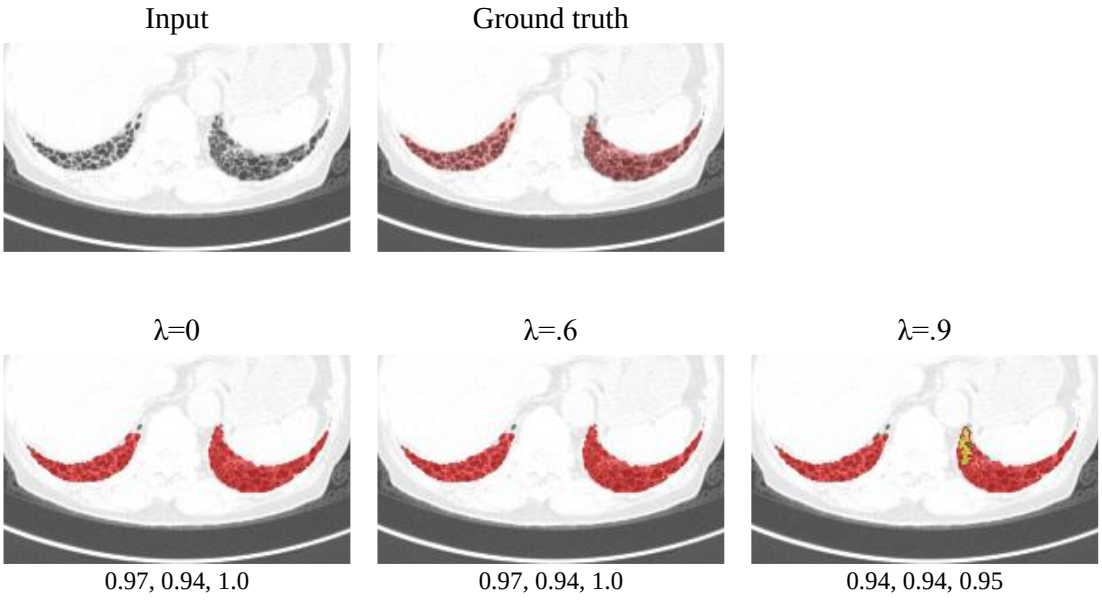
Figure 3.10 shows the best segmentation results of each DLD with λ = 0.6. Segmentation results with λ = 0 and λ = 0.9 are shown for references. In these best results, representative DLD patterns covered most of the lung fields. Results with λ = 0.9 tends to have small islands of non-representative DLDs.



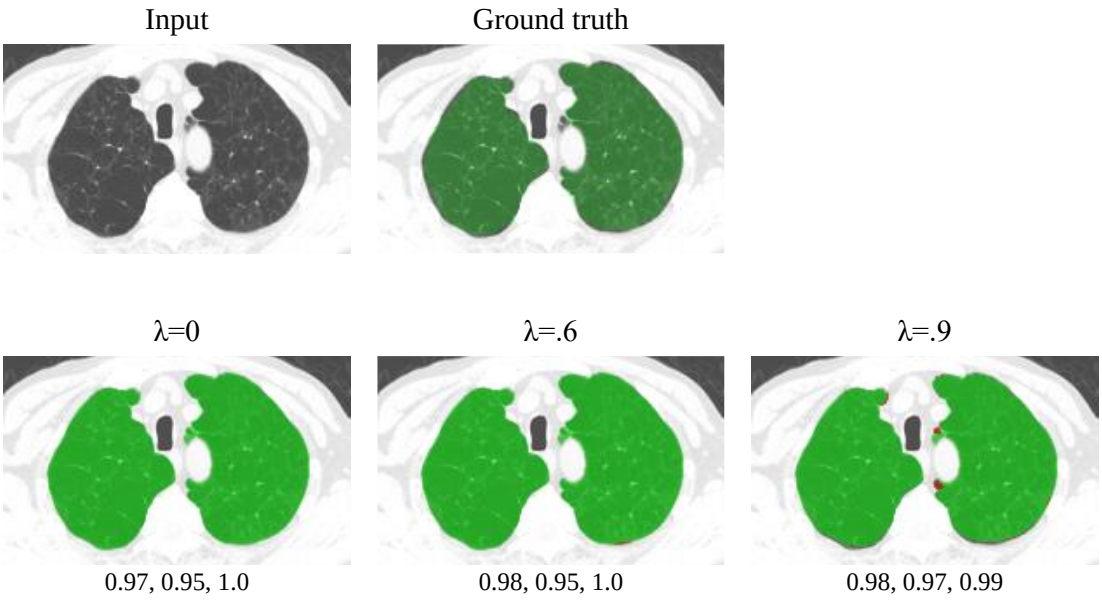
(a) Consolidation (CON).



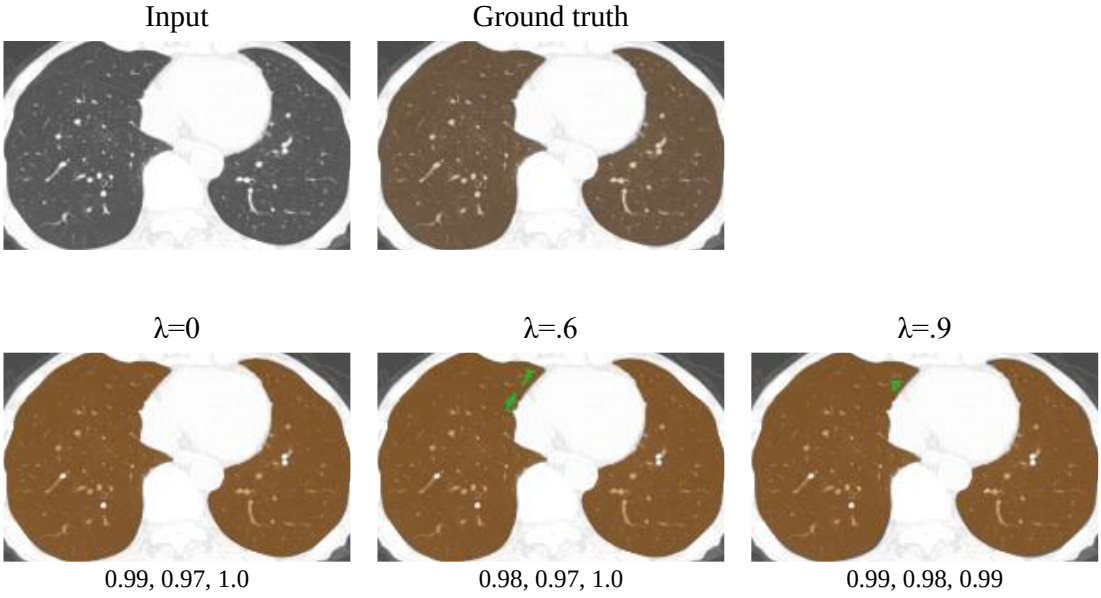
(b) Ground glass opacity (GGO).



(c) Honeycombing (HCM).



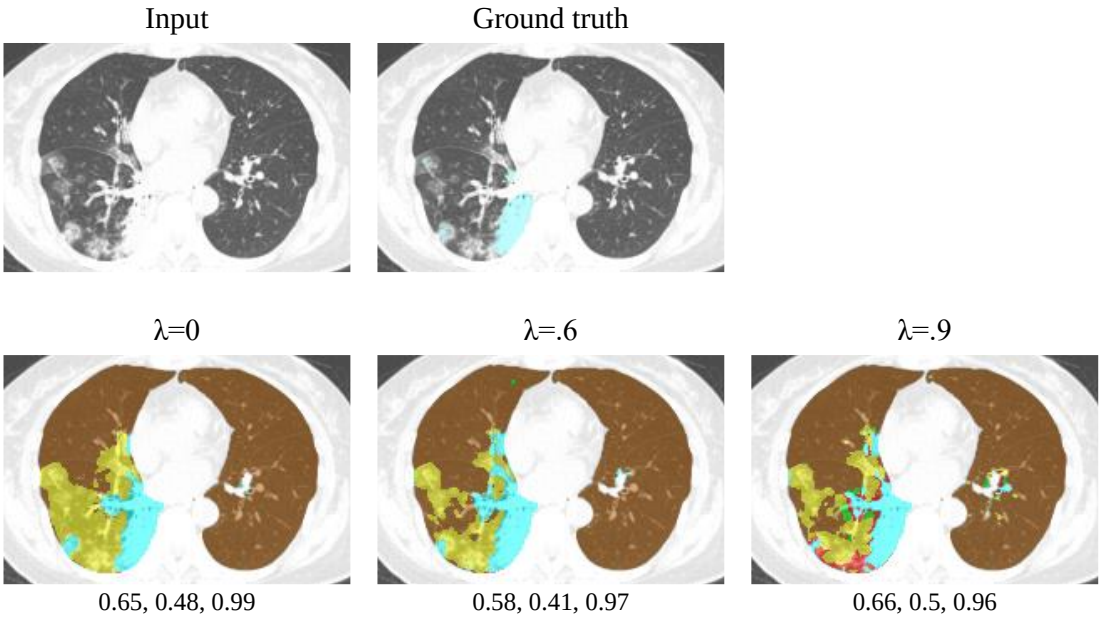
(d) Emphysema (EMP).



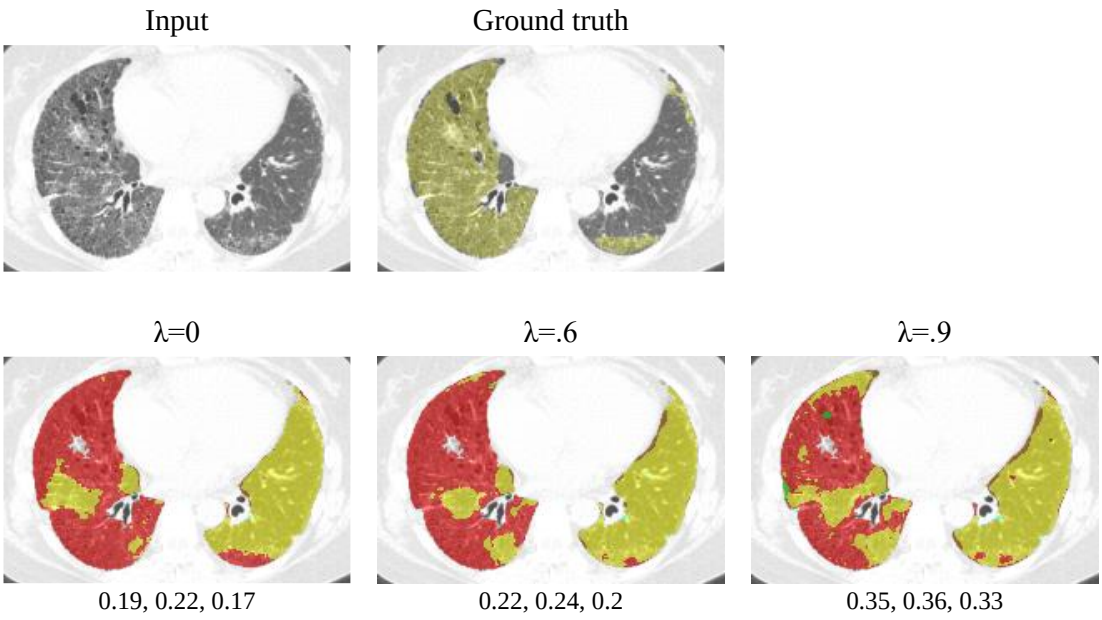
(e) Normal (NOR).

Figure 3.10: Examples of best segmentation results with λ = 0.6. The numbers below images indicate Dice score, precision, and recall respectively.

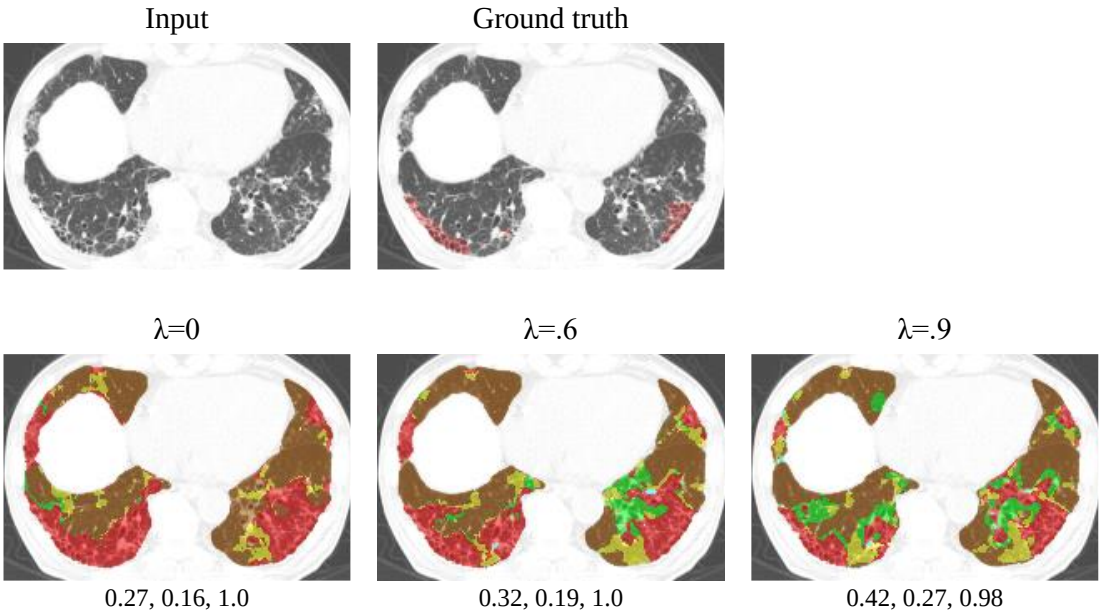
Figure 3.11 shows the worst segmentation results of each DLD with λ = 0.6. Segmentation results with λ = 0 and λ = 0.9 are shown for references. In these worst results, representative DLD patterns did not occupy big space in the lung fields. In Figure 3.11 (b), most of the GGO pixels were falsely segmented as HCM. Results with λ = 0.9 tends to have less false NOR pixels.



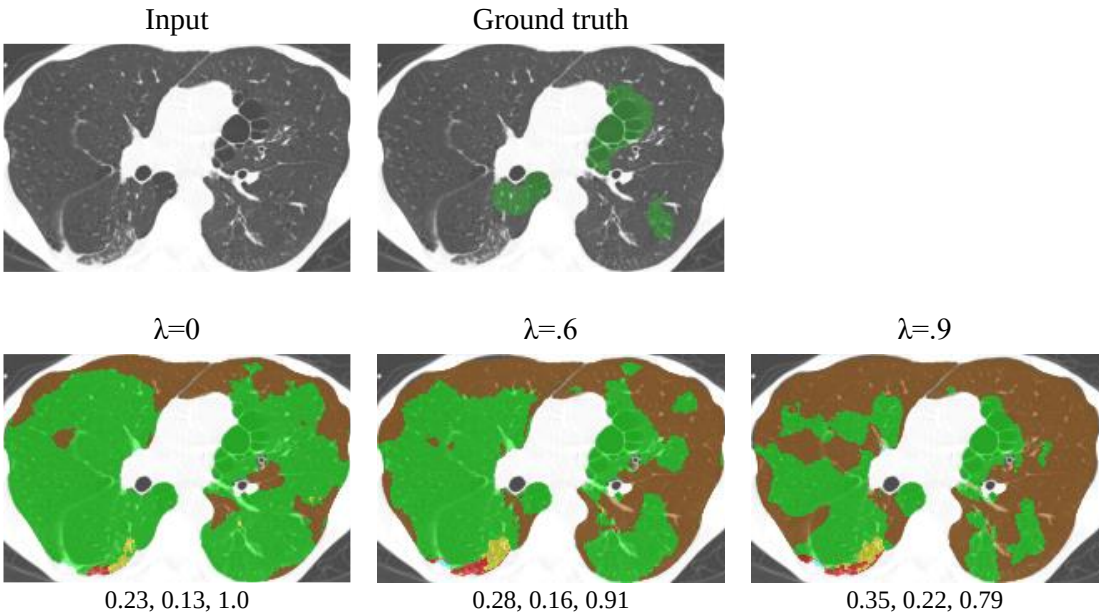
(a) Consolidation (CON).



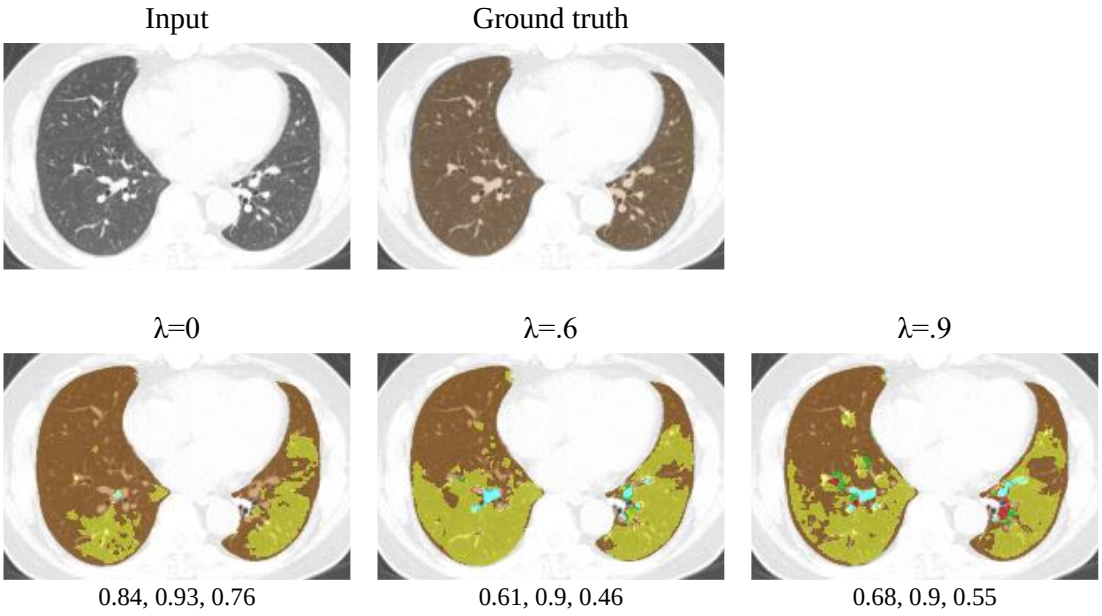
(b) Ground glass opacity (GGO).



(c) Honeycombing (HCM).



(d) Emphysema (EMP).



(e) Normal (NOR).

Figure 3.11: Examples of worst segmentation results with λ = 0.6. The numbers below images indicate Dice score, precision, and recall respectively.

Figure 3.12 shows the differences per slice in Dice score between the results with λ = 0 and other parameters. As shown in Figure 3.12, for GGO, HCM, and EMP, the proposed method with λ = .6 improved the segmentation accuracy for the most cases while for CON and NOR, the proposed method with λ = .6 worsen the accuracy for the most cases. Overall accuracy was improved because the differences of CON and NOR in the plots were smaller than those of GGO, HCM, and EMP.

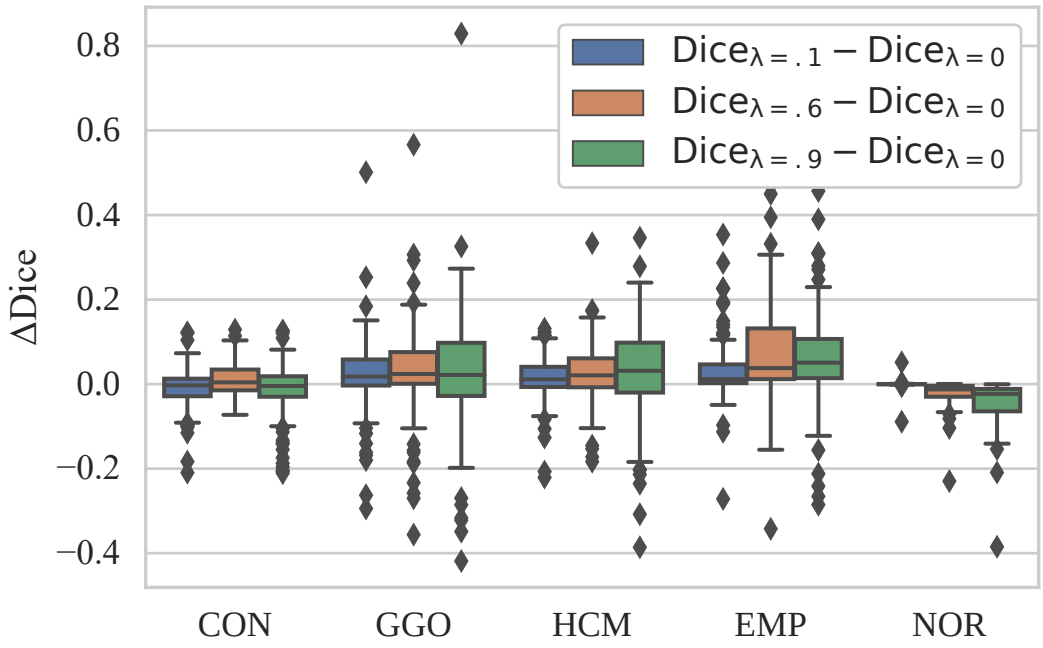


Figure 3.12: Box-and-whisker plots of the differences per slice in Dice score between the results with λ = 0 and other parameters.

3.4.2 Evaluation on samples of full annotation

In addition to the evaluation only using annotated pixels, we conducted additional evaluation using fully annotated images. For this evaluation, five slices were randomly chosen from each DLD and all pixels in the lung fields in the chosen slices were manually annotated. Figure 3.13 shows the confusion matrices for the segmentation results with λ = 0 and λ = 0.6. Unlike confusion matrices shown in Figure 3.8, these matrices were constructed using all pixels in the lung fields. As shown in the bottom row of Figure 3.13 (a), a significant number of ¬NOR pixels were falsely labeled as NOR with λ = 0, whereas in Figure 3.13 (b), there were much less false NOR pixels with λ = 0.6. As can be seen in the unannotated regions in Figures 3.2 and 3.9, most of the unannotated pixels consisted of NOR pixels, therefore, models trained

with λ = 0.6 was able to reduce the number of false NOR pixels. Table 3.3 shows the mean Dice scores. As shown in the table, λ = 0.6 was also optimal in this evaluation. Figure 3.14 shows line plots of evaluation metrics (Dice score, precision, and recall). The curves in Figure 3.14 are similar to the curves for the evaluation with partial annotation in Figure 3.5 with some fluctuation caused by the small number of samples (five per DLD) of fully annotated images.

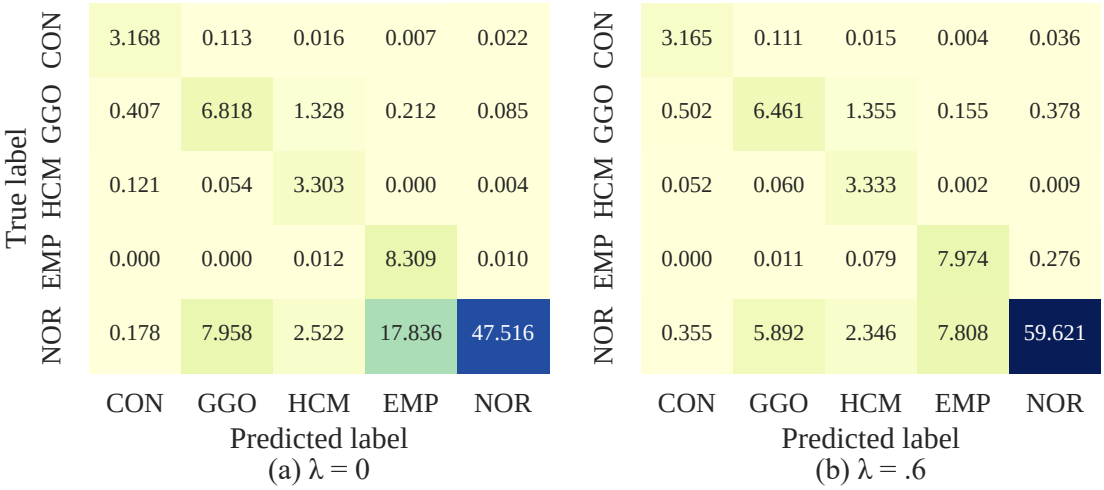
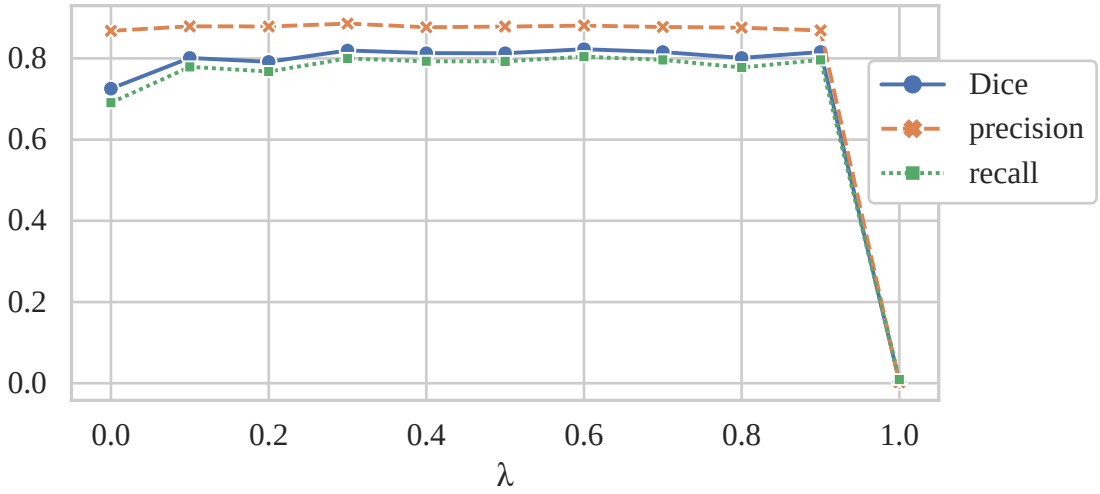


Figure 3.13: Confusion matrices calculated using fully annotated images. Note that matrices are normalized so that the sum of the elements in a matrix become 100.

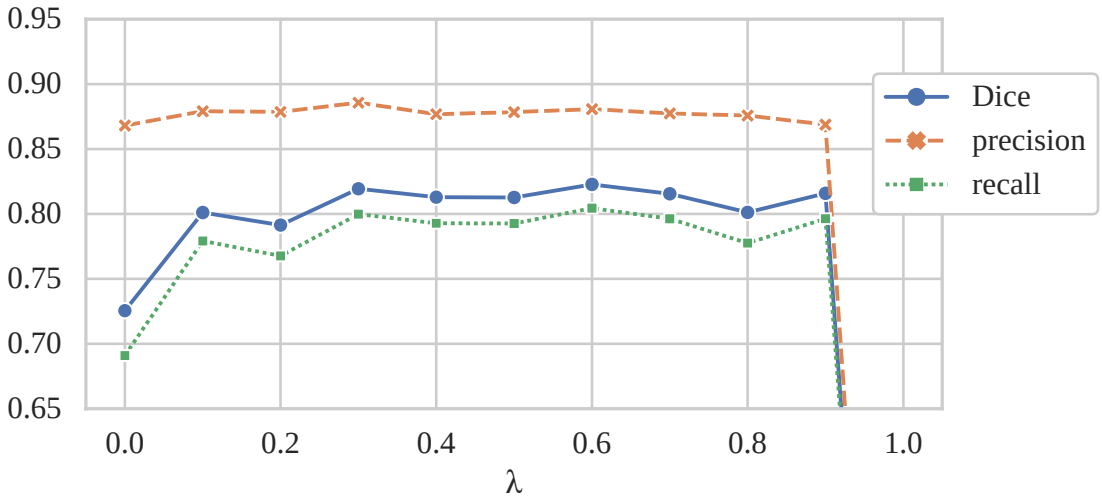
Table 3.3: Mean Dice scores on fully annotated images.

λ	Dice	precision	recall
0	0.726	0.868	0.691
0.1	0.801	0.879	0.779
0.2	0.791	0.879	0.768
0.3	0.819	**0.886**	0.8

0.4	0.813	0.877	0.793
0.5	0.813	0.878	0.793
0.6	**0.823**	0.881	**0.804**
0.7	0.816	0.877	0.796
0.8	0.801	0.876	0.778
0.9	0.816	0.869	0.796
1	0.004	0.002	0.009



(a) Line plots.



(b) Line plots focusing on λ = {0, 0.1, ..., 0.9}.

Figure 3.14: Line plots of Dice score, precision and recall calculated using fully annotated images.

3.5 Discussion

We developed a PSL approach and tested it on a partially annotated DLD dataset. The proposed method improved mean Dice score from 0.76 to 0.79 and from 0.73 to 0.82 in the evaluation using partially annotated dataset and fully annotated dataset respectively. These improvements were thanks to the improvements in precision as seen in Figure 3.5 and achieved with no extra cost for annotation. This means that training semantic segmentation models using the proposed method can improve accuracy over the training methods that only used annotated pixels. Because PSL only requires partial annotation, it is easier to use active learning technique [101] to construct or update dataset. Low precision shown in Figures 3.6 and 3.9 was likely due to

the lack of full annotation available in the training dataset and needs to be addressed in the future work.

We used partially annotated dataset in which only one class was annotated for an image in this study. This setting was chosen to test the proposed method in the hardest configuration. Moreover, it is common for publicly available datasets to have only one-class annotation and this study's dataset is similar to a dataset that is created by combining such datasets. However, if the goal is to develop a multi-class segmentation model, preparing dataset with at least a couple of fully annotated images is more reasonable for achieving high accuracy.

We conducted experiments with various values of balancing parameter λ. As shown in Figure 3.5, the supervision using unannotated pixels helps models to improve its precision. More specifically, the proposed method with optimal parameter significantly reduced the number of falsely labeled NOR pixels as shown in Figure 3.13. This was because most of the unannotated (negative) pixels were actually NOR pixels and training using these unannotated pixels helped the model to distinguish NOR and non-NOR pixels. In our experiment, $\lambda = 0.6$ was optimal regarding the Dice score but the parameter can be tuned based on the preference between precision and recall. To our best knowledge, while introduction of the balancing parameter to PSL was firstly introduced by our study [36], Shi et al. [98] also performed similar study that changed the balance of two terms in their loss function for multi-organ segmentation. In their experimental results, they had similar curve to ours with peak accuracy at the optimal parameter of 0.67, which was similar to our optimal parameter ($\lambda = 0.6$).

In our experiments, the parameter value $\lambda = 1$ did not work at all with almost zero Dice score. This is because of the property of our partially annotated dataset. Because annotations were performed for representative (or dominant) DLDs, DLD texture patterns with typical texture patterns and appearances were annotated. With $\lambda = 1$, annotated pixels, which have typical texture, were not used at all for the training, therefore models were not able to learn typical texture of DLDs and failed to segment most of the typical DLDs.

As shown in Figure 3.12, the proposed method worsened the segmentation accuracies for some cases. This is likely because of annotation errors in the dataset. Even though the annotation was performed carefully with multiple annotators, the annotation was still imperfect and the dataset contains some negative pixels that were actually positive pixels (i.e. some pixels of representative DLD were erroneously left unannotated). Two possible reasons below could explain why some cases had worse results with the proposed method. 1) Results were not accurately evaluated because of the erroneous evaluation dataset. 2) Using erroneous negative pixels for the training provides erroneous supervision to the models. Testing the second hypothesis is in the future work and it can be tested by artificially introducing errors in the annotation of the training dataset.

There are other possible approaches to realize multiclass segmentation using a partially annotated dataset. One approach is an ensemble model that combines multiple binary or subset classifier models using subsets of the dataset for the training of each sub-model. This approach, however, is expected to have less generalization ability than the

proposed PSL. Another possible approach is multi-task learning [102] that uses subsets of the dataset as different tasks and trains one model with multiple output layers. Comparing our method with other possible approaches [81], [82] is included in our future work.

We used the reciprocal of cross-entropy for the training using negative pixels. Figure 3.15 (a) shows the regular binary cross entropy (BCE) loss for positive pixels (y=1) and Figure 3.15 (b) shows the reciprocal of the binary cross entropy loss, which was applied for the negative pixels. As shown in Figure 3.15 (a), BCE smoothly decreases as the output ($\hat{y}$) gets close to one while there is rapid increase in BCE^{-1} near $\hat{y} = 1$ in Figure 3.15 (b). This asymmetry in the two types of loss can make it harder to balance these two types of loss. More symmetric losses such as label-set loss proposed in [81] can be used. Testing other possible losses such as label-set [81] is included in the future work.

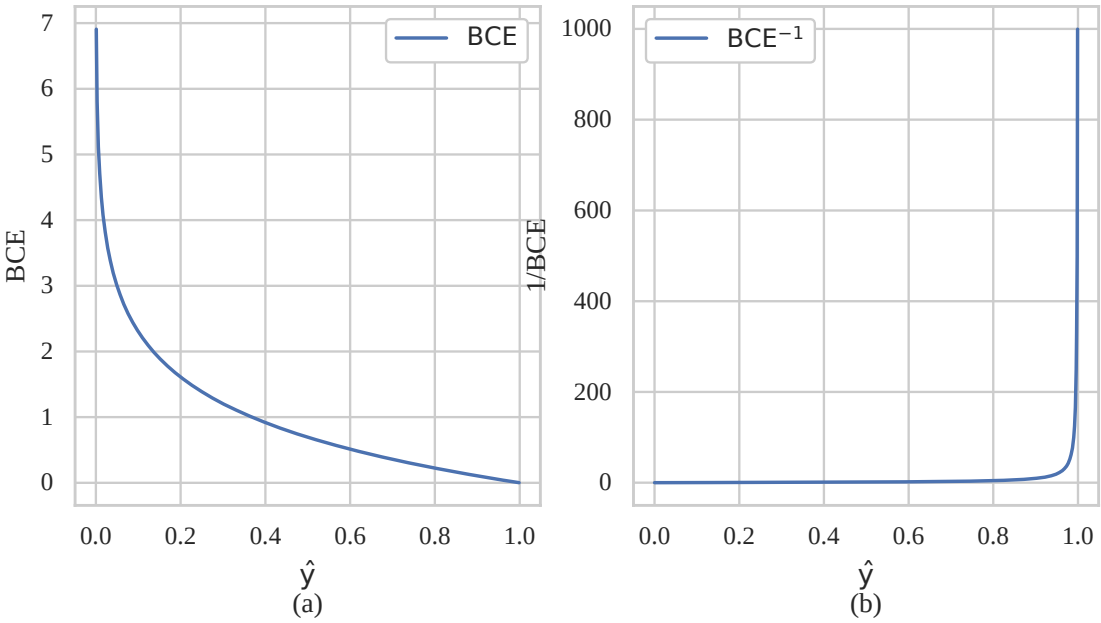


Figure 3.15: Plots of losses. (a) Binary cross entropy (BCE) loss for positive pixels (y = 1). (b) The reciprocal of binary cross entropy loss for negative pixels.

One of the similar approaches that utilize unannotated pixels is semi-supervised learning [103]. In semi-supervised learning, unannotated pixels are typically used for entropy minimization [83] or consistency training. The biggest difference between the proposed method and semi-supervised learning is that the proposed method utilizes unannotated pixels for negative training. Therefore, the proposed method is likely to outperform semi-supervised learning methods. However, combining partially supervised learning with semi-supervised techniques such as entropy minimization is possible and can improve the segmentation accuracy.

As shown in Figures 3.10 and 3.11, best results tend to have small DLD regions and worst results tend to have large DLD regions. Precision for

the worst results were low because of false NOR segmentation. As shown in Figure 3.13, GGO and EMP especially tend to have false NOR pixels. Therefore, NOR pixels that are similar to GGO and EMP need to be added to the dataset in the future. Using weak annotations such as scribble [99] can be useful to save annotation costs in combination with partial annotations and PSL.

Comparing levels of annotation is included in the future work. One-class per image annotation, which was used in this study, provides the least supervision with the lowest cost. On the other hand, all-class annotation, which annotate all pixels in the lung fields, provides the most supervision with the highest cost. Comparing above two and different levels of annotations between the two under fixed amount of cost (e.g. the time or budget for the annotation process) is needed to figure out the optimal level of annotation that maximizes segmentation accuracy under the limited amount of cost for the annotation.

Because one representative DLD pattern was chosen by a radiologist for each slice, our partially annotated dataset contains potential biases. First, representative patterns can be classified as wrong classes because they were chosen by only one radiologist. Second, there are more typical textures in the dataset than atypical textures of the classes because representative patterns were chosen. These need to be addressed in the future work to obtain less biased results.

One of the limitations of this study is that because the evaluation metrics were calculated mostly in annotated regions, the effects of the proposed PSL in unannotated regions were not fully studied other than the precision shown in Figure 3.7 and the evaluation only using a

fraction of the dataset shown in Figure 3.13 and Table 3.3. The lung tissue research consortium (LTRC) dataset [104] can be used for a more detailed study because the LTRC dataset provides fully annotated data.

A potential future direction is to combine PSL with other techniques such as transfer learning [90] to achieve even better accuracy. Another possibility is to combine other datasets in a similar domain such as interstitial lung diseases from the University Hospital of Geneva [105] and the LTRC dataset [104], to test the possibility of improving accuracy using a combination of similar but slightly different datasets.

4 Conclusion

4.1 Summary

This thesis served following two purposes.

1. Addressing the lack of comparison study on abdominal blood vessel segmentation compared to the abundance of new proposals of methods.
2. Proposing and testing a method that takes full advantage of exiting dataset.

Chapter 2 addressed the first purpose through the experiments using multiple methods of vessel segmentation. The experimental results showed that characteristics of the methods differed depending on the region. UN was the best method for the training ROI, whereas SN resulted better in non-training ROI such as lung ROI. HN generally was outperformed by the CNN methods but outperformed them in the lung ROI, where surrounding tissues are easily distinguishable from vessels.

In chapter 3, new method was proposed that addressed the second purpose. Although the proposed method did not improve the segmentation accuracy by a large margin, the proposed method can be useful considering that the improvement was achieved without additional cost for annotating new data. The experimental results

showed that the proposed method improved the precision of the segmentation.

4.2 Future work

As shown in the experimental results in chapter 2, segmentation errors were caused by thin vessels, therefore, investing in the improvement on thin vessels can be the next logical step. Training a CNN specialized to thin vessels can be an approach to improve the performance on thin vessels.

As shown in the results in chapter 3, the segmentation accuracy on some DLDs were still relatively low despite the improvement achieved by the proposed method. Possible future direction for this is to incorporate semi-supervised or unsupervised learning to utilize training dataset even further.

5 Acknowledgement

I would like to take this opportunity to express my deepest gratitude to those who patiently helped me. Without their generous help and support, it would have been hardly possible to complete this thesis.

First, I would like to thank Professor Yoshinobu Sato, Associate Professor Yoshito Otake, and Assistant Professor Mazen Soufi at Graduate School of Information Science, Nara Institute of Technology and Science (NAIST), for their insightful and constructive advice for this study. I would also like to thank all the members of the Imaging-based Computational Biomedicine Laboratory at NAIST. The experiences they gave me through the discussions and guidance they provided with have been and will always be irreplaceable part of me. In addition, I would also like to express my special thank Mrs. Fumiyo Tagawa, who is an administrative staff of the laboratory and made my life so much easier with her continuous help. Her cheerful and delightful personality made my time in the lab full of joy. I will be forever in her debt.

I would like to thank Professor Fumihiko Ino at Osaka University Graduate School of Information Science and Technology, who allowed me to be part of his laboratory despite the extra work they had to have because of me. If it were not for the long-term support from him, I would not be able to continue my work. Furthermore, I would like to

thank him, Professor Yasushi Yagi at The Institute of Scientific and Industrial Research, Osaka University and Hajime Nagahara at Osaka University Institute for Datability Science for their time and effort to review this thesis.

I was very fortunate to have medical doctors who were able to give me advice from clinical perspectives despite their very busy schedules. Professor Masatoshi Hori at Kobe University Graduate School of Medicine supplied me with CT images and always helped me with his deep clinical knowledge. Professor Shoji Kido and Noriyuki Tomiyama at Osaka University Graduate School of Medicine gave me an opportunity to work in their groups. Their help has been essential for me to keep my work going and stay clinically relevant.

Professor Shingo Mabu at Yamaguchi University greatly helped me with his expertise in machine learning and pulmonary texture analysis, despite the fact that he had no obligation to help me. His advice never missed the point and substantially improved the quality of this thesis.

I spent about 6 months at Johns Hopkins University as an internship student in 2016. I am sincerely grateful to Dr. Austin Reiter and Katsuyuki (Ken) Taguchi for accepting and welcoming me in their respective research groups. I learned a lot from them and experiences gained through the internship undoubtedly gave me the motivation and the grit to finish this thesis. Additionally, I would like to give special thanks to Dr. Robert Grupp, and Mr. and Mrs. Swann, who gave me immense support with my life in the US. My internship would not have been as fruitful as it was without them.

Finally, I thank my family and all who have been part of this great journey for their continuous support.

6 References

[1] T. Okada *et al.*, "Multi-Organ Segmentation in Abdominal CT Images," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2012, p. 10.1109/EMBC.2012.6346840, 2012, doi: 10.1109/EMBC.2012.6346840.

[2] F. Yokota *et al.*, "Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method," *Int J CARS*, vol. 13, no. 7, pp. 977–986, Jul. 2018, doi: 10.1007/s11548-018-1758-y.

[3] Y. Hiasa, Y. Otake, M. Takao, T. Ogawa, N. Sugano, and Y. Sato, "Automated Muscle Segmentation from Clinical CT Using Bayesian U-Net for Personalized Musculoskeletal Modeling," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1030–1040, Apr. 2020, doi: 10.1109/TMI.2019.2940555.

[4] J. Park *et al.*, "Fully Automated Lung Lobe Segmentation in Volumetric Chest CT with 3D U-Net: Validation with Intra- and Extra-Datasets," *J Digit Imaging*, vol. 33, no. 1, pp. 221–230, Feb. 2020, doi: 10.1007/s10278-019-00223-1.

[5] L. Wang *et al.*, "Benchmark on Automatic Six-Month-Old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge," *IEEE*

Transactions on Medical Imaging, vol. 38, no. 9, pp. 2219–2230, Sep. 2019, doi: 10.1109/TMI.2019.2901712.

[6] X. Yang *et al.*, "Segmentation of liver and vessels from CT images and classification of liver segments for preoperative liver surgical planning in living donor liver transplantation," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 41–52, May 2018, doi: 10.1016/j.cmpb.2017.12.008.

[7] J. Ma *et al.*, "Abdomenct-1k: Is abdominal organ segmentation a solved problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[8] K. Hu *et al.*, "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function," *Neurocomputing*, vol. 309, pp. 179–191, Oct. 2018, doi: 10.1016/j.neucom.2018.05.011.

[9] B. Kong *et al.*, "Learning tree-structured representation for 3D coronary artery segmentation," *Computerized Medical Imaging and Graphics*, vol. 80, p. 101688, Mar. 2020, doi: 10.1016/j.compmedimag.2019.101688.

[10] N. Li, S. Zhou, Z. Wu, B. Zhang, and G. Zhao, "Statistical modeling and knowledge-based segmentation of cerebral artery based on TOF-MRA and MR-T1," *Computer Methods and Programs in Biomedicine*, vol. 186, p. 105110, Apr. 2020, doi: 10.1016/j.cmpb.2019.105110.

[11] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods

on a new publicly available database," in *Medical Imaging 2004: Image Processing*, May 2004, vol. 5370, pp. 648–656. doi: 10.1117/12.535349.

[12] P. Vostatek, E. Claridge, H. Uusitalo, M. Hauta-Kasari, P. Fält, and L. Lensu, "Performance comparison of publicly available retinal blood vessel segmentation methods," *Computerized Medical Imaging and Graphics*, vol. 55, pp. 2–12, Jan. 2017, doi: 10.1016/j.compmedimag.2016.07.005.

[13] W. Wiharto and E. Suryani, "The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels," *Acta Inform Med*, vol. 28, no. 1, pp. 42–47, Mar. 2020, doi: 10.5455/aim.2020.28.42-47.

[14] M. R. K. Mookiah *et al.*, "A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification," *Medical Image Analysis*, vol. 68, p. 101905, Feb. 2021, doi: 10.1016/j.media.2020.101905.

[15] R. Phellan and N. D. Forkert, "Comparison of vessel enhancement algorithms applied to time-of-flight MRA images for cerebrovascular segmentation," *Medical Physics*, vol. 44, no. 11, pp. 5901–5915, 2017, doi: 10.1002/mp.12560.

[16] R. D. Rudyanto *et al.*, "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: The VESSEL12 study," *Medical Image Analysis*, vol. 18, no. 7, pp. 1217–1232, Oct. 2014, doi: 10.1016/j.media.2014.07.003.

[17] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *Eur*

Radiol, vol. 29, no. 3, pp. 1391–1399, Mar. 2019, doi: 10.1007/s00330-018-5695-5.

[18] J. Widell and M. Lidén, "Interobserver variability in high-resolution CT of the lungs," *Eur J Radiol Open*, vol. 7, p. 100228, Mar. 2020, doi: 10.1016/j.ejro.2020.100228.

[19] O. Ozdemir, R. L. Russell, and A. A. Berlin, "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1419–1429, May 2020, doi: 10.1109/TMI.2019.2947595.

[20] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, Jan. 2019, doi: 10.1016/j.patcog.2018.07.031.

[21] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, 2017, pp. 559–567. doi: 10.1007/978-3-319-66179-7_64.

[22] D. Ueda *et al.*, "Deep Learning for MR Angiography: Automated Detection of Cerebral Aneurysms," *Radiology*, vol. 290, no. 1, pp. 187–194, Jan. 2019, doi: 10.1148/radiol.2018180901.

[23] X. Dai *et al.*, "Deep learning for automated cerebral aneurysm detection on computed tomography images," *Int J CARS*, vol. 15, no. 4, pp. 715–723, Apr. 2020, doi: 10.1007/s11548-020-02121-2.

[24] P. Bilic *et al.*, “The Liver Tumor Segmentation Benchmark (LiTS),” *arXiv:1901.04056 [cs]*, Jan. 2019,Available: http://arxiv.org/abs/1901.04056

[25] N. Heller *et al.*, “The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge,” *Medical Image Analysis*, vol. 67, p. 101821, Jan. 2021, doi: 10.1016/j.media.2020.101821.

[26] C. Rother, V. Kolmogorov, and A. Blake, “"GrabCut": Interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004, doi: 10.1145/1015706.1015720.

[27] D. Freedman and T. Zhang, “Interactive graph cut based segmentation with shape priors,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, 2005, doi: 10.1109/CVPR.2005.191.

[28] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, “Automated abdominal multi-organ segmentation with subject-specific atlas generation,” *IEEE Trans Med Imaging*, vol. 32, no. 9, pp. 1723–1730, Sep. 2013, doi: 10.1109/TMI.2013.2265805.

[29] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, “Abdominal multi-organ segmentation from CT images using conditional shapelocation and unsupervised intensity priors,” *Medical Image Analysis*, vol. 26, no. 1, pp. 1–18, Dec. 2015, doi: 10.1016/j.media.2015.06.009.

[30] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, “A generative probability model of joint label fusion for multi-atlas based

brain segmentation," *Medical Image Analysis*, vol. 18, no. 6, pp. 881–890, Aug. 2014, doi: 10.1016/j.media.2013.10.013.

[31] Y. Huo *et al.*, "Consistent cortical reconstruction and multi-atlas brain segmentation," *NeuroImage*, vol. 138, pp. 197–210, Sep. 2016, doi: 10.1016/j.neuroimage.2016.05.030.

[32] Y. Sato *et al.*, "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," *Med Image Anal*, vol. 2, no. 2, pp. 143–168, 1998.

[33] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Med Image Comput Comput Assist Interv*, 1998, pp. 130–137. doi: 10.1007/BFb0056195.

[34] T. Jerman, F. Pernuš, B. Likar, and Ž. Špiclin, "Beyond Frangi: An improved multiscale vesselness filter," in *Medical Imaging 2015: Image Processing*, Mar. 2015, vol. 9413, pp. 623–633. doi: 10.1117/12.2081147.

[35] Y. Suzuki *et al.*, "Comparative Study of Vessel Detection Methods for Contrast Enhanced Computed Tomography: Effects of Convolutional Neural Network Architecture and Patch Size," *Advanced Biomedical Engineering*, vol. 10, pp. 138–149, 2021, doi: 10.14326/abe.10.138.

[36] Y. Suzuki, K. Yamagata, M. Yanagawa, S. Kido, and N. Tomiyama, "Weak supervision in convolutional neural network for semantic segmentation of diffuse lung diseases using partially annotated dataset," in *Medical Imaging 2020: Computer-Aided Diagnosis*, Mar. 2020, vol. 11314, p. 113142A. doi: 10.1117/12.2548930.

[37] Y. Suzuki, S. Kido, S. Mabu, M. Yanagawa, N. Tomiyama, and Y. Sato, “Segmentation of Diffuse Lung Abnormality Patterns on Computed Tomography Images using Partially Supervised Learning,” 2021.

[38] P. Bibiloni, M. González-Hidalgo, and S. Massanet, “A survey on curvilinear object segmentation in multiple applications,” *Pattern Recognit.*, vol. 60, pp. 949–970, 2016.

[39] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, “Blood vessel segmentation algorithms review of methods, datasets and evaluation metrics,” *Comput Methods Programs Biomed*, vol. 158, pp. 71–91, 2018.

[40] F. Zhao, Y. Chen, Y. Hou, and X. He, “Segmentation of blood vessels using rule-based and machine-learning-based methods: A review,” *Multimed. Syst.*, vol. 25, no. 2, pp. 109–118, Apr. 2019, doi: 10.1007/s00530-017-0580-7.

[41] S. Hanaoka *et al.*, “HoTPiG: A novel graph-based 3-D image feature set and its applications to computer-assisted detection of cerebral aneurysms and lung nodules,” *Int J CARS*, vol. 14, no. 12, pp. 2095–2107, Dec. 2019, doi: 10.1007/s11548-019-01942-0.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Neural Inf Process Syst*, 2012, pp. 1097–1105.

[43] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2015, pp. 3431–3440.

[44] C. Wang *et al.*, "Tubular Structure Segmentation Using Spatial Fully Connected Network with Radial Distance Loss for 3D Medical Images," in *Med Image Comput Comput Assist Interv*, 2019, pp. 348–356.

[45] A. Taha, P. Lo, J. Li, and T. Zhao, "Kid-net: Convolution networks for kidney vessels segmentation from ct-volumes," in *Med Image Comput Comput Assist Interv*, 2018, pp. 463–471.

[46] D. Keshwani, Y. Kitamura, S. Ihara, S. Iizuka, and E. Simo-Serra, "TopNet: Topology Preserving Metric Learning for Vessel Tree Reconstruction and Labelling," in *Med Image Comput Comput Assist Interv*, 2020, pp. 14–23.

[47] T. R. Patel *et al.*, "Multi-resolution CNN for brain vessel segmentation from cerebrovascular images of intracranial aneurysm: A comparison of U-Net and DeepMedic," in *Medical Imaging 2020: Computer-Aided Diagnosis*, 2020, vol. 11314, p. 113142W.

[48] F. Yu *et al.*, "Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images," in *Med Image Comput Comput Assist Interv*, 2019, pp. 714–722.

[49] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (R2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018,Available: http://arxiv.org/abs/1802.06955

[50] Q. Huang, J. Sun, H. Ding, X. Wang, and G. Wang, "Robust liver vessel extraction using 3D U-Net with variant dice loss function," *Computers in biology and medicine*, vol. 101, pp. 153–162, 2018.

[51] C. Wang, Z. Zhao, Q. Ren, Y. Xu, and Y. Yu, “Dense U-net based on patch-based learning for retinal vessel segmentation,” *Entropy*, vol. 21, no. 2, p. 168, 2019.

[52] M. Oda, H. R. Roth, T. Kitasaka, K. Misawa, M. Fujiwara, and K. Mori, “Abdominal artery segmentation method from CT volumes using fully convolutional neural network,” *Int J CARS*, vol. 14, no. 12, pp. 2069–2081, 2019.

[53] H. Yang, C. Shan, A. F. Kolen, and P. H. de With, “Automated catheter localization in volumetric ultrasound using 3D patch-wise U-net with focal loss,” in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1346–1350.

[54] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Med Image Comput Comput Assist Interv*, 2015, pp. 234–241.

[55] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation,” in *Med Image Comput Comput Assist Interv*, 2016, pp. 424–432. doi: 10.1007/978-3-319-46723-8_49.

[56] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014,Available: http://arxiv.org/abs/1412.6980

[57] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945, doi: 10.2307/1932409.

[58] P. Jaccard, “The Distribution of the Flora in the Alpine Zone,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

[59] T. C. Lee, R. L. Kashyap, and C. N. Chu, “Building Skeleton Models via 3-D Medial Surface Axis Thinning Algorithms,” *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 6, pp. 462–478, Nov. 1994, doi: 10.1006/cgip.1994.1042.

[60] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning*, Jun. 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[61] “Python.org,” *Python.org*. https://www.python.org/.

[62] “TensorFlow,” *TensorFlow*. https://www.tensorflow.org.

[63] H. Homann, “Implementation of a 3D thinning algorithm,” *The Insight Journal*, p. 181, Oct. 2007.

[64] “ITK | Insight Toolkit.” https://itk.org/.

[65] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,” *Biomedical optics express*, vol. 9, no. 7, pp. 3049–3066, 2018.

[66] M. Livne *et al.*, “A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease,” *Front. Neurosci.*, vol. 13, 2019, doi: 10.3389/fnins.2019.00097.

[67] H. R. Roth *et al.*, "A Multi-scale Pyramid of 3D Fully Convolutional Networks for Abdominal Multi-organ Segmentation," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2018*, 2018, pp. 417–425. doi: 10.1007/978-3-030-00937-3_48.

[68] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2019, pp. 358–368. doi: 10.1007/978-3-030-11726-9_32.

[69] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018, doi: 10.1109/TMI.2018.2845918.

[70] Y. He *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Med Image Anal*, vol. 63, p. 101722, Jul. 2020, doi: 10.1016/j.media.2020.101722.

[71] C. Wang *et al.*, "Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation," *Med Image Anal*, vol. 60, p. 101623, Feb. 2020, doi: 10.1016/j.media.2019.101623.

[72] G. Tetteh *et al.*, "DeepVesselNet: Vessel Segmentation, Centerline Prediction, and Bifurcation Detection in 3-D Angiographic Volumes," *Frontiers in Neuroscience*, vol. 14, p. 1285, 2020, doi: 10.3389/fnins.2020.592352.

[73] Q. Yan *et al.*, "Attention-Guided Deep Neural Network With Multi-Scale Feature Fusion for Liver Vessel Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2629–2642, Jul. 2021, doi: 10.1109/JBHI.2020.3042069.

[74] T. Watadani *et al.*, "Interobserver Variability in the CT Assessment of Honeycombing in the Lungs," *Radiology*, vol. 266, no. 3, pp. 936–944, Mar. 2013, doi: 10.1148/radiol.12112516.

[75] M. Gao *et al.*, "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, Jan. 2018, doi: 10.1080/21681163.2015.1124249.

[76] D. Bermejo-Peláez, S. Y. Ash, G. R. Washko, R. San José Estépar, and M. J. Ledesma-Carbayo, "Classification of Interstitial Lung Abnormality Patterns with an Ensemble of Deep Convolutional Neural Networks," *Sci Rep*, vol. 10, no. 1, p. 338, Jan. 2020, doi: 10.1038/s41598-019-56989-5.

[77] M. Negahdar, A. Coy, and D. Beymer, "An End-to-End Deep Learning Pipeline for Emphysema Quantification Using Multi-label Learning," *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2019, pp. 929–932, Jul. 2019, doi: 10.1109/EMBC.2019.8857392.

[78] M. Gao *et al.*, "Segmentation label propagation using deep convolutional neural networks and dense conditional random field," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1265–1268.

[79] T. Uemura, J. J. Näppi, C. Watari, T. Hironaka, T. Kamiya, and H. Yoshida, "Weakly unsupervised conditional generative adversarial network for image-based prognostic prediction for COVID-19 patients based on chest CT," *Medical Image Analysis*, vol. 73, p. 102159, Oct. 2021, doi: 10.1016/j.media.2021.102159.

[80] S. Mabu, M. Obayashi, T. Kuremoto, N. Hashimoto, Y. Hirano, and S. Kido, "Unsupervised class labeling of diffuse lung diseases using frequent attribute patterns," *Int J CARS*, vol. 12, no. 3, pp. 519–528, Mar. 2017, doi: 10.1007/s11548-016-1476-2.

[81] L. Fidon *et al.*, "Label-Set Loss Functions for Partial Supervision: Application to Fetal Brain 3D MRI Parcellation," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2021*, 2021, pp. 647–657. doi: 10.1007/978-3-030-87196-3_60.

[82] J. Peng and Y. Wang, "Medical Image Segmentation With Limited Supervision: A Review of Deep Network Models," *IEEE Access*, vol. 9, pp. 36827–36851, 2021, doi: 10.1109/ACCESS.2021.3062380.

[83] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, and S. Mougiakakou, "Semantic Segmentation of Pathological Lung Tissue With Dilated Fully Convolutional Networks," *IEEE J Biomed Health Inform*, vol. 23, no. 2, pp. 714–722, Mar. 2019, doi: 10.1109/JBHI.2018.2818620.

[84] O. Chapelle, B. Scholkopf, and A. Zien Eds., "Semi-Supervised Learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, Mar. 2009, doi: 10.1109/TNN.2009.2015974.

[85] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, Jul. 2020, doi: 10.1016/j.media.2020.101693.

[86] C. Wang *et al.*, "Weakly-supervised deep learning of interstitial lung disease types on CT images," in *Medical Imaging 2019: Computer-Aided Diagnosis*, Mar. 2019, vol. 10950, pp. 373–379. doi: 10.1117/12.2512746.

[87] I. Laradji *et al.*, "A Weakly Supervised Consistency-Based Learning Method for COVID-19 Segmentation in CT Images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2453–2462.

[88] H.-C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[89] S. Chen, K. Ma, and Y. Zheng, *Med3D: Transfer Learning for 3D Medical Image Analysis*. 2019.

[90] S. Mabu, A. Atsumo, S. Kido, T. Kuremoto, and Y. Hirano, "Investigating the Effects of Transfer Learning on ROI-based Classification of Chest CT Images: A Case Study on Diffuse Lung Diseases," *J Sign Process Syst*, vol. 92, no. 3, pp. 307–313, Mar. 2020, doi: 10.1007/s11265-019-01499-w.

[91] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource Transfer Learning With Convolutional

Neural Networks for Lung Pattern Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 76–84, Jan. 2017, doi: 10.1109/JBHI.2016.2636929.

[92] X. Fang and P. Yan, "Multi-Organ Segmentation Over Partially Labeled Datasets With Multi-Scale Feature Abstraction," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3619–3629, Nov. 2020, doi: 10.1109/TMI.2020.3001036.

[93] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," *arXiv:2011.10217 [cs]*, Nov. 2020,Available: http://arxiv.org/abs/2011.10217

[94] K. Dmitriev and A. E. Kaufman, "Learning Multi-Class Segmentations From Single-Class Datasets," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 9493–9503. doi: 10.1109/CVPR.2019.00973.

[95] N. Roulet, D. F. Slezak, and E. Ferrante, "Joint Learning of Brain Lesion and Anatomy Segmentation from Heterogeneous Datasets," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, May 2019, pp. 401–413.

[96] Y. Zhou *et al.*, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10672–10681.

[97] R. Dorent *et al.*, "Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets,"

Medical Image Analysis, vol. 67, p. 101862, Jan. 2021, doi: 10.1016/j.media.2020.101862.

[98] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Medical Image Analysis*, vol. 70, p. 101979, May 2021, doi: 10.1016/j.media.2021.101979.

[99] X. Liu *et al.*, "Weakly Supervised Segmentation of COVID19 Infection with Scribble Annotation on CT Images," *Pattern Recognition*, vol. 122, p. 108341, Feb. 2022, doi: 10.1016/j.patcog.2021.108341.

[100] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[101] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, 2017, pp. 399–407. doi: 10.1007/978-3-319-66179-7_46.

[102] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, p. 104037, Nov. 2020, doi: 10.1016/j.compbiomed.2020.104037.

[103] Y. Ouali, C. Hudelot, and M. Tami, "An Overview of Deep Semi-Supervised Learning," *arXiv:2006.05278 [cs, stat]*, Jul. 2020,Available: http://arxiv.org/abs/2006.05278

[104] B. Bartholmai, R. Karwoski, V. Zavaletta, R. Robb, and D. R. I. Holmes, "The Lung Tissue Research Consortium: An extensive open database containing histological, clinical, and radiological data to study chronic lung disease," *Insight journal*, 2006.

[105] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, no. 3, pp. 227–238, Apr. 2012, doi: 10.1016/j.compmedimag.2011.07.003.